

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

RANDOM SUBCLONING, PAIRWISE END
SEQUENCING, AND THE MOLECULAR
EVOLUTION OF THE VERTEBRATE
TRYPSINOGENS

by

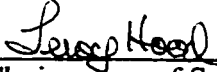
Jared C. Roach

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

1998

Approved by 
Chairperson of Supervisory Committee

Program Authorized
to Offer Degree Department of Immunology

Date December 18, 1997

UMI Number: 9826359

**Copyright 1998 by
Roach, Jared Carter**

All rights reserved.

**UMI Microform 9826359
Copyright 1998, by UMI Company. All rights reserved.**

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

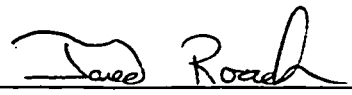
UMI
300 North Zeeb Road
Ann Arbor, MI 48103

© Copyright 1998

Jared C. Roach

Doctoral Dissertation

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 1490 Eisenhower Place, P.O. Box 975, Ann Arbor, MI 48106, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature  Road
Date 2 / 11 / 98

University of Washington

Abstract

**RANDOM SUBCLONING, PAIRWISE END
SEQUENCING, AND THE MOLECULAR
EVOLUTION OF THE VERTEBRATE
TRYPSINOGENS**

by Jared C. Roach

Chairperson of the Supervisory Committee: Professor Leroy Hood
Departments of Molecular Biotechnology and Immunology

Mathematical theory for random subcloning is presented and discussed in detail. The pairwise end sequencing strategy for mapping and sequencing is presented in a general form; specific examples are analyzed with the aid of computer simulations. The evolution of the vertebrate trypsinogen multigene family is discussed in the context of newly sequenced trypsinogen genes from the lamprey *Petromyzon marinus* and the tunicate *Boltenia villosa*.

TABLE OF CONTENTS

LIST OF FIGURES	iii
LIST OF TABLES	iv
GLOSSARY	v
LIST OF ABBREVIATIONS	vii
PREFACE	x
INTRODUCTION	1
CHAPTER 1. RANDOM SUBCLONING	5
1.1 Mapping And Sequencing Large Genomes	6
1.2 Sequence Walking	8
1.3 Overview Of Random Subcloning	13
1.4 Mathematical Model - Basic Formulation	17
1.5 Target Coverage (1)	22
1.6 Mathematical Model - The Beta Distribution	24
1.7 Gaps And Islands	28
1.8 The Number Of Clones In An Island	31
1.9 Island Length	36
1.10 Target Coverage (2)	38
1.11 Comparison With The Lander-Waterman Equations	41
1.12 Island Co-Dependency	45
1.13 Circular Targets	47
1.14 Simulations And Data	52
1.15 Examples	53
1.16 Random Closing Remarks	59
CHAPTER 2. PAIRWISE END SEQUENCING	71
2.1 The Double-Barrel Shotgun	71
2.2 Formulation	73
2.3 Computer Simulations	76
2.4 Raw Data Simulation	79
2.5 Perspective On Pairwise Strategies	81
2.6 Mathematical Models	83
2.7 Discussion	84
CHAPTER 3. VERTEBRATE TRYPSINOGEN EVOLUTION	95
3.1 Coincidental Evolution	96
3.2 Historical Perspective On Trypsinogen	97
3.3 The Compilation Of Trypsin And Trypsinogen Sequences	101
3.4 Cloning And Sequencing <i>Petromyzon Marinus</i> Trypsinogen	102
3.5 Lamprey Trypsinogens	103
3.6 Cloning And Sequencing <i>Boltenia Villosa</i> Trypsinogen	104
3.7 Tunicate Trypsinogen	105
3.8 Signal Sequences	107
3.9 Activation Peptides	108
3.10 Cystine Bridges	109
3.11 Insertions And Deletions	111
3.12 Intron/Exon Boundaries	112
3.13 Cationic And Anionic Trypsins	113
3.14 Multiple Sequence Alignments	114
3.15 Sequence Distances	117
3.16 Multidimensional Scaling	120

3.17 Phylogenies	126
3.18 Modes Of Trypsinogen Evolution	132
3.19 Expression Of Trypsin	136
3.20 Function Of Trypsin	139
3.21 Genesis Of Novel Genes	140
AFTERWORD	175
BIBLIOGRAPHY	176
APPENDIX A. DOUBLE-BARREL SHOTGUN ALGEBRA	198
A.1 One Clone	199
A.2 Two Clones	199
A.3 Three Clones	199
A.4 More Than Three Clones	211
APPENDIX B. ALTERNATIVE SEQUENCE DISTANCE METRICS	217
B.1 Sequence Distances	217
B.2 The Jukes-Cantor Model	220
B.3 A Peptide View of The Jukes-Cantor Model	222
APPENDIX C. PCR EVALUATION OF TRYPSINOGEN EXPRESSION	233
C.1 Methods and Reults	233
C.2 Discussion	234
APPENDIX D. GENOME RESEARCH 5:464-473	238
APPENDIX E. GENOMICS 26:345-353	248
APPENDIX F. JOURNAL OF MOLECULAR EVOLUTION 45(6):640-652	257

LIST OF FIGURES

<i>Figure</i>	<i>Page</i>
Figure 1.1. Schematic cartoon of a random subcloning project.	62
Figure 1.2. Schematic of a mathematical formulation for random subcloning.	63
Figure 1.3. Expected target coverage with respect to redundancy.	64
Figure 1.4. Relative error of the Lander-Waterman and "beta" models.	65
Figure 1.5. Computer simulations.	66
Figure 1.6. Expected closure costs.	67
Figure 1.7. Expected cost of reaching gapped project states.	68
Figure 1.8. Probability of completion with respect to redundancy.	69
Figure 1.9. Incremental cost of closing one gap.	70
 Figure 2.1. Model "double-barrel shotgun" assembly.	 89
Figure 2.2. Parameters from a 35 kb pairwise project with respect to redundancy.	90
Figure 2.3. Parameters from a 200 kb pairwise project with respect to redundancy.	91
Figure 2.4. Simulation of pairwise strategies employing a mix of insert sizes.	92
Figure 2.5. Simulation of a hybrid pairwise strategy.	93
 Figure 3.1. Cartoons of genomic trypsinogen organization.	 143
Figure 3.2. Stereoscopic view of the trypsin backbone.	144
Figure 3.3. Multiple alignment of chordate trypsin protein sequences.	145
Figure 3.4. Multidimensionally-scaled vertebrate trypsin sequence distances.	148
Figure 3.5. Stress vs. Dimensions.	149
Figure 3.6. Group I vs. Group II.	150
Figure 3.7. Allopatric division of the trypsinogen multigene family.	151
Figure 3.8. Group I vs. Group II (3-D).	152
Figure 3.9. Group I vs. Group II (alternative 3-D).	153
Figure 3.10. 5' vs. 3'.	154
Figure 3.11. Anionic vs. Cationic.	155
Figure 3.12. Multidimensionally scaled projection of the rodent group I trypsins.	156
Figure 3.13. Hypothetical phylogeny of the vertebrate trypsins.	157
Figure 3.14. Fitch-Margoliash phylogeny of forty-two vertebrate trypsins.	158
Figure 3.15. Fitch-Margoliash phylogeny of thirty-two vertebrate trypsins.	159
Figure 3.16. Coincidental statistics of thirty sequences.	160
Figure 3.17. Pseudogenes added to the vertebrate trypsin phylogeny.	161
Figure 3.18. Skewed Fitch-Margoliash phylogeny of the vertebrate trypsins.	162
Figure 3.19. A phylogeny of the rodent group I trypsins.	163
 Figure A.1. Topologies for one- and two-clone double-barrel configurations.	 214
Figure A.2. Topologies for three-clone double-barrel configurations.	215
Figure A.3. Probability of one scaffold with respect to insert:fragment ratio.	216
 Figure B.1. Trypsin site variability.	 230
Figure B.2. Trypsin sequence logo.	231
Figure B.3. An alternative Fitch-Margoliash trypsin phylogeny.	232

LIST OF TABLES

<i>Table</i>	<i>Page</i>
Table 2.1. Results from a raw data simulation of a pairwise strategy.	94
Table 3.1. Literature references for the chordate trypsinogens.	164
Table 3.2. Classification comments for the chordate trypsinogens.	167
Table 3.3. PCR Primers used in the analysis of the chordate trypsinogens.	169
Table 3.4. Signal peptides and activation peptides of the chordate trypsinogens.	170
Table 3.5. Cystine bridges of the chordate trypsinogens.	171
Table 3.6. Predicted isoelectric points and charges of the chordate trypsins.	172
Table 3.7. Genbank identification numbers for the human trypsinogen ESTs.	173
Table 3.8. Genbank accession numbers for the mouse trypsinogen ESTs.	174
Table C.1. Number of cDNAs sequenced by PCR from several human tissues.	236
Table C.2. Number of cDNAs sequenced by PCR from several mouse tissues.	237

GLOSSARY

Coincidental Evolution. A tendency of genes present in the same genome to evolve in a non-independent manner. The presence of coincidental evolution makes homologous genes within a genome more similar to each other than to homologous genes from other genomes. Many authors use the term “concerted evolution” as a synonym for coincidental evolution, which was originally defined by Hood et al. (1975).

Contig. An island consisting of at least two fragments.

Fitness. An organism’s ability to propagate.

Functional Genomics. The development and application of experimental approaches to assess gene function by making use of the information and reagents provided by structural genomics (Heiter and Boguski, 1997; McKusick, 1997).

Gap. A region of a target that is not represented in an island. Gaps are sometimes referred to as “oceans.”

Gene Product. The product of a gene. Most genes code for proteins, but some code for structural RNAs, and some affect the structure and/or regulation of the genome without being transcribed into a downstream product.

Gene. Information encoded in a segment of genomic DNA that affects the fitness of an organism. Semantically, it is useful to consider some genes as consisting of multiple gene segments, such as the TCR β gene.

Genome. The information encoded in the DNA of a cell. Every individual, with few exceptions, has a distinct genome. The genome can vary slightly from cell to cell within an organism. The term was coined in 1920 by Winkler as an elision of the words “gene” and “chromosome” (McKusick, 1997).

Genomicist. One who studies genomes.

Genomics. The study of genomes. By contrast, the term “genetics” refers to the study of inheritance. The term “genomics” was introduced by Roderick in 1986 (McKusick, 1997).

Homologous Genes. Genes that share a common ancestral gene. Orthology and paralogy are subcategories of homology. Genes may be homologous without necessarily being either paralogous or orthologous (see, for example, Tatusov et al., 1997).

Indel. An insertion or a deletion.

Island. A maximal set of fragments each of which is connected to all other island members by at least one path of overlapping fragments.

Isozymes. Enzymes that have identical (or nearly identical) biochemical properties.

Orphon. A term applied to gene segments separated from a complete functional gene locus. A typical example is a TCR V β segment on chromosome 9, unable to recombine with D β and C β segments to form a functional gene.

Orthologous Genes. Genes in different species that share a common function and evolved from a common ancestor. By definition, the split between such genes was caused by a speciation event. The result of speciation (Fitch, 1970).

Paralogous Genes. Multiple genes resulting from duplication within a particular genome. The result of gene duplication (Fitch, 1970).

Parameterization. A particular choice of parameters for a model (or a project). Parameters are variables that one can control, such as the choice of how many clones to analyze, or what length clones to choose.

Pseudogene. A gene that is no longer functional. The ancestral sequence was a functional gene that acquired one or more mutations that destroyed functionality. Typically, this would entail the acquisition of stop codons in an open reading frame. Point mutations are typically responsible for the creation of pseudogenes.

Relic. A fragment of a gene that was once functional. The semantic boundary between a pseudogene and a relic is fuzzy. Generally in order to be classified a relic, one or more recombinations or major deletions must have operated on the ancestral gene.

Repeat. A region of a genome that is nearly identical to another region of the same genome. It should be emphasized that in genomics terminology the word “repeat” does not imply 100% identity. The percent identity used to define a repeat is somewhat subjective.

Scaffold. An ordered and oriented list of islands. Also referred to as a “gapped island” (Port et al., 1995) or a “supercontig” (Lawrence et al., 1994).

Structural Genomics. The construction of genetic, physical, and transcript maps of genomes (Heiter and Boguski, 1997). Genomic sequence is considered to be the ultimate high resolution physical map of a genome. The definition can be rephrased as, “mapping and sequencing genomes” (McKusick, 1997).

Subclone. A clone of a fragment of a larger piece of DNA. The fragment has been genetically engineered into a vector that facilitates laboratory manipulation of the fragment.

Target. A genome or a subset of a genome that will be analyzed during the course of a project.

LIST OF ABBREVIATIONS

BAC. Bacterial artificial chromosome.

bp. Base pair.

cDNA. Complementary deoxyribonucleic acid.

DNA. Deoxyribonucleic acid.

ds. Double stranded.

EST. Expressed sequence tag.

HMM. Hidden Markov model.

kb. Kilobase pair.

mRNA. Messenger RNA.

OSS. Ordered shotgun sequencing.

PCR. Polymerase chain reaction.

RNA. Ribonucleic acid.

SMG. Sequence-mapped gap

ss. Single stranded.

STS. Sequence-tagged connector.

STS. Sequence-tagged site.

TCR. T-cell receptor.

YAC. Yeast artificial chromosome.

LIST OF VARIABLES

C . An arbitrary length (a partial target goal less than G).

D_k . The length of the k^{th} spacing from the left end of the target.

f . The fraction of the target that is covered.

f_g . The effective fractional coverage of the target provided by one fragment.

G . The length of a target (in bases). This abbreviation is most appropriate when the target is a genome, but has come to be used even when the target is something else, such as a subclone. I do not employ the convention of setting G equal to unity.

G_e . The effective length of the target.

g_m . The number of short spacings in the m^{th} island from the left end of the target.

I . The length of an insert.

L . The length of a clone (in bases). I do not employ the convention of setting L equal to unity.

l_m . The length of the m^{th} island from the left end of the target.

N . The number of permitted residues at a sequence site.

n . The total number of clones analyzed in a project, either by mapping or sequencing.

N_{gaps} . The number of gaps in a project.

N_{islands} . The number of islands in a project.

N_{long} . The number of long spacings in a project.

$N_{\text{singletons}}$. The number of single fragment islands in a project.

μ . The probability of a mutation per unit evolutionary time.

p_{gap} . The probability of a gap following a spacing.

R . Redundancy.

R_e . The effective redundancy.

S_k . The k^{th} spacing from the left end of the target.

τ . Evolutionary time.

T . The number of base pairs necessary to determine overlap during the assembly phase of a random subcloning project. In many practical cases $T \ll L$, and can be approximated as zero.

V . The length of the vector sequence.

z_m . The number of fragments in the m^{th} island from the left end of the target.

PREFACE

The doctoral thesis has a rich history. The tradition of the dissertation binds modern students back through the ages to the history and culture of the Renaissance Universities (Haskins, 1923). The richness of values governing the content of theses dictates certain compromises with respect to the choice of included material and its style of presentation.

My desire in this thesis is to provide knowledge in a understandable fashion to the widest possible audience: biologists, mathematicians, computer scientists, engineers, and those in allied fields. I hope to have made at least a portion of the text accessible to students at all levels. However, for much of this dissertation, I will assume at least a basic knowledge of molecular biology, such as that obtainable from *The Cartoon Guide to Genetics* (Gonick and Wheelis, 1991) or another introductory molecular biology textbook. The reader is encouraged to ignore or merely skim material that is either too basic or advanced. A more compact presentation of much of the material in this dissertation can be found in my original journal articles which are reproduced in the appendices.

My doctoral research has led me down many disparate paths, of which three have been significant enough to include in this dissertation, each as a separate chapter. Two of these chapters group naturally in the area of strategic genomics, while the third strikes out tangentially into the field of molecular evolution. Rather than attempt an integrated and detailed introduction to these three subjects simultaneously, I have kept the overall introduction, which follows, short and generally oriented. I include more detailed topical introductions at the beginning of each chapter. Also, in each chapter I employ several specialized terms appropriate to the subject matter; I encourage the reader to exploit the glossary when in doubt of the meaning of a particular term. Readers may also find the list of abbreviations and the list of variables useful as occasional references.

ACKNOWLEDGEMENTS

Collaboration is the cornerstone of modern research. One can lift a rock; many can move a mountain. The many include: Lee Hood, Roger Perlmutter, Chris Wilson, Dave Lewis, Ken Walsh, Andy Siegel, Steve Henikoff, Kai Wang, and innumerable technicians, graduate students, post-docs, professors, friends, and family, all who have contributed variously their labor, ideas, and encouragement. I am grateful for a grant from the Life & Health Insurance Medical Research Fund.

INTRODUCTION

The closing decades of the second millennium have brought forth a cascade of technology and knowledge that promises to bring forth greater changes than were brought by the twentieth century. Biological knowledge and biotechnology have perhaps lagged other cornerstones of change such as the microchip revolution, but may in the long run may have the greatest impact on the evolution of human life and society.

Biological study is increasingly a study of complex systems. It has always been, and it is becoming more so. Our capacity as scientists to analyze and understand complex systems is rapidly maturing. An organism, such as a human being, is a complex system itself composed of multiple complex systems. One of these systems is the genome. It is a complex task to understand the genome, and such understanding is but a small step towards understanding the organism.

However, progress is made of small steps. In this dissertation, I will discuss a few small steps towards understanding the genome. Scientific progress is often driven by a marriage of new technologies with a cutting edge problem. I will present and discuss two strategic technologies for genome study. I will then apply genomic knowledge to the analysis of vertebrate trypsinogen evolution, which is itself a model for the complexity of genome evolution.

The technologies for genome study I present here are *strategies* for genomic analysis. The emphasis underscores a worldwide change in the paradigm for genetic experiments. To date, most genetics have been done from a “bottom-up” perspective. Isolated problems in genetics were selected and analyzed in detail. Global genomic information was not sought. Any global information obtained was added piecemeal and without coordination to the body of scientific knowledge.

Now, increasing numbers of genomes are being sequenced in their entirety, with several large genome sequencing projects underway (Rowen et al., 1997). Emphasis is on a “top-down” perspective, with global information sought primarily.¹ Intent is that content

¹ Searching for a needle in a haystack is analogy for a “bottom-up” approach to genome analysis. The needle represents a sought-after gene such as that for Huntington’s Disease (i.e., HD) or breast cancer (e.g., BRCA1). The haystack is the genome. The search for

not only be applied to the study of isolated systems but also to systems that arise from complex interactions between multiple genes and proteins (e.g., DeRisi et al., 1997). An expectation is that a comprehensive global effort will not only be more thorough, but will also be more efficient than multiple “bottom-up” approaches.

Therefore, efficient strategies for genome sequencing are needed. The scale of genome projects is such that even minor improvements in strategy can effect major changes in cost, effort, and even feasibility.

In Chapter 1, I focus attention on random subcloning, which is a simple strategy for genome analysis. Until now, no adequate mathematical analysis of random subcloning has been available. It has been impossible to predict, other than empirically, project outcomes or costs. It has been impossible to compare complex strategies with basic standards. Here, I provide some fundamentals for addressing these issues.

In Chapter 2, I present pairwise end sequencing as an example of a more complex genome mapping and sequencing strategy. A robust mathematical analysis of pairwise end sequencing continues to elude researchers, so I analyze the strategy with the aid of computer simulations. Such simulations are a powerful way to answer the practical questions of project design and evaluation, even in the absence of a complete mathematical analysis.²

In Chapter 3, I depart from the purely theoretical themes of the first two chapters and address the evolution of the genome. I focus on the evolution of the trypsinogens, which are present in vertebrate genomes as a multicopy family tightly linked to the T-cell

each of these genes can be viewed as having been an independent search through the same haystack for different needles. A “top-down” approach to haystack analysis would be to sort through the entire haystack, one straw at a time, categorizing *everything* found in the haystack. Such an approach might be inefficient if just one needle is sought, but would become more efficient the more needles there were in the haystack. Identifying the approximately 75,000 genes in the human genome would require 75,000 independent bottom-up searches. Alternatively, one comprehensive search could be conducted. Providing this comprehensive search is one goal of the Human Genome Project. An excellent summary of the present state and future potential of genomic study has been provided by McKusick (1997). Estimates of the number of genes in the human genome have been tabulated by Fields et al. (1994). UW manuscript #1.

² For background on the power, utility, and limitations of computer simulations, the reader may wish to consider Galper and Brutlag (1993). The need for computer simulations in the Human Genome Project is highlighted by Koonin (1998).

receptor β locus. These genes display a panoply of evolutionary modalities, including striking examples of coincidental evolution. Coincidental evolution is the tendency of genes present in the same genome to evolve in a covariant manner. I also point out several difficulties that complicate evolutionary analysis of multigene families.

Strategies for genomic analysis are often analyzed with simple assumptions about the nature of multigene families and other repeated elements. These assumptions are often used to call into question the utility of proposed genomic strategies. Understanding the nature of repeats, including multigene families, and how they are formed is thus an important adjuvant to a discussion of structural genomics. With an understanding of the nature of evolution, we can begin to predict how similar repeats are likely to be, and this prediction, in turn, affects parameterizations of strategies and overall cost.

From a functional genomics point of view, the nature of repeat families tells us much not only about evolution, but also of the nature of complex biological systems interaction. Complex modern vertebrates are positioned at a pinnacle of billions of years of biological evolution on Earth.³ An important enabling feature of this evolution has been the ability of the evolutionary process to reuse and readapt previously constructed building blocks (Henikoff et al., 1997). Since the dawn of vertebrate evolution 600 million years ago, it may be that very few truly new genes have evolved (Holland and Garcia-Fernández, 1996). Rather, nature has adapted previously existing genes, sometimes in novel combinations, to new functions. Original genes are usually left untouched, with novel adaptation operating on duplicate copies of original gene family members. The resulting multiple isoforms of genes permit the establishment of complex systems with multiple similar but subtly different components. Evolutionary operations on gene families are likely to have been a major mode of evolution in the vertebrate subphylum.

The serine proteases, of which trypsin is a member, are one of the largest and most diverse gene superfamilies (Barrett and Rawlings, 1993). Their origins lie buried in the earliest stages of evolution – trypsin genes are present in eubacterial genomes (Rypniewski et al., 1994). Modern vertebrate serine proteases include chymotrypsin, elastase, the

³ All currently living species are positioned on their own respective “pinnacles of evolution.” The height of each pinnacle is the amount of evolution that has occurred to a species since the origin of life. If evolution is measured by time, all current pinnacles are equally high. The pinnacles that were reached by extinct species are lower. An introduction to evolutionary concepts can be found in Dawkins (1990 and 1996).

recently-evolved prostate specific antigen, blood clotting enzymes, several granzymes, and many other genes involved in immunological defense. The serine protease gene superfamily rivals the immunoglobulin gene superfamily in terms of its importance to both vertebrate evolution and vertebrate immune defense (Smyth et al., 1996; Hunkapiller et al., 1989). It is fascinating to ponder the close genetic linkage of the trypsin and T-cell receptor β gene loci.

The vertebrates almost certainly possess the most complex immune system of all living organisms. This has been made possible by the repeated use of duplications in multicopy gene families (see, for example, Ohno, 1978; Hunkapiller et al., 1989; Hood and Hunkapiller, 1991; Raport et al., 1996). These families include the immunoglobulins the serine proteases, and many others. There are, in fact, few genes of importance to immunology that are not members of multigene families. To study multigene families is to study the fabric of immunological complexity.

CHAPTER 1. RANDOM SUBCLONING

"You cannot see the wood for the trees."

English Colloquial Saying

Peer as we might, we find ourselves in a deep entangled forest, seeing about for only a short distance. What we see is vivid, in great detail, but it is not enough. Standing in one place we cannot grasp the whole. This is the state of genome research.

Imagine visiting the Louvre with a great magnifying glass, constrained to examine each painting one tiny fleck of paint at a time, unable to step back, unable to see the art itself. As you look at the Mona Lisa, with each glance, you are only able to see a spot one twentieth of a millimeter in diameter. The textures and colors of the cracked oil pigments leap out at you. You make detailed notes, recording each observation in a computer database. Your computer churns out statistics on pigment height, color, and crack length distribution. You become frustrated by your inability to see the Mona Lisa.

The human genome, packaged into a living cell a few microns in diameter, is paradoxically too large to see. It is true that chromosomes, the structural material encoding the genome, can be seen, but not the information itself. The word "genome" defines an informational concept. Chromosomes contain DNA. A genome contains information. DNA is an ink that forms words. Genes are sentences and stories, information conveyed by DNA. The genome is often analogized as a book of life (e.g., Wills, 1991).¹

It is not proper, however, to compare the genome to a book, for a book can be read by human eyes, while the genome cannot. A better analogy is that of a magnetic storage disk, able to release its information only with the aid of a mechanical interpreter, such as a computer. Within the confines of a living cell the genome is read with biological machines -

¹ Humans are by no means currently capable of understanding all of the informational content of any genome. Nor will such understanding come solely from genomics. Progress in all allied disciplines, particularly those focusing on protein function, is necessary to contribute to such understanding. Biological understanding is the result of a vast friendly collaboration between workers in an amazingly diverse array of fields.

polymerases, ribosomes, and a multitude of supporting enzymes and structures. Human observers use other machines – DNA sequencers – to read the genome. It is the limitations of these machines, and of the chemistries that underlie their workings, that place the genome researcher in the shoes of the magnifying-glass-toting art connoisseur.

Each machine can read only a small fraction of the genome at a time. The human genome contains three billion base pairs arrayed on twenty-four separate chromosomes. If this information were stored on a computer disk, it would occupy about a gigabyte of space. A single “sequence read” from a DNA sequencer permits the visualization of only a tiny amount of this data, currently about five hundred to one thousand base pairs.

1.1 MAPPING AND SEQUENCING LARGE GENOMES

Genomicists wish to obtain the entire sequence of the human genome. This is the goal of the Human Genome Project, which is currently a major international effort (McKusick, 1997; Rowen et al., 1997). However, the methodologies developed for the human genome will not end there, for many other genomes await. Nor will the human genome be the toughest genome ever to be sequenced. The genome of the lily *Fritillaria davisii*, for example, contains 3×10^{11} bp, while *Amoeba dubia* has 6.7×10^{11} bp (Wachtel and Tiersch, 1993; Li and Graur, 1991).

How to sequence these genomes is problematic. How to *best* sequence these genomes is even more problematic. Optimizing genome sequencing is not merely an academic question. The scale and expense of genome projects is such that slight differences in the efficiencies of various strategies can spell the difference between feasibility and impossibility. The first step in choosing the best strategy is to understand the consequences of pursuing any given strategy. We must ask ourselves, “What results can we expect to obtain given a certain level of expenditure of resources?” If we can decide upon our goals ahead of time, we should be able to estimate how much it will cost us to reach these goals.

Random subcloning is a popular strategy in use by genomicists. Random subcloning is simple, and is the easiest strategy to implement. The strategy iteratively generates sequences from random locations in a genome until, by chance, sequences from the entire genome have been obtained. At this point, data analysis algorithms are used to reconstruct the global picture from the random fragments. The “forest” is reconstructed from a random collage of local snapshots of the “trees.”

Random subcloning is not merely a sequencing strategy. The principle of breaking down large problems into small problems is not unique to sequencing. It is possible to analyze the genome without determining the details of its sequence. Such an analysis is termed “mapping.” A genome mapping project seeks to identify and order landmarks of the genome. A physical map uses structural landmarks, such as restriction sites, or the ends of clones. A genetic map uses informational landmarks, such as genes. The methodologies of genetic mapping are quite different from those of physical mapping, but when landmarks from the different methodologies can be correlated with each other, the maps can be integrated. I will not discuss genetic mapping further in this dissertation, and will focus on physical mapping and sequencing.

It is possible to use physical mapping techniques to analyze fragments of DNA that are much longer than the length of a sequence read. It is quicker and easier to build up a global picture with such techniques, as the fragments of the puzzle can be a thousand times larger than individual sequence reads. The trade-off for speed is that the resulting picture is fuzzier, much like an impressionist painting compared to a sharp color photograph. One advantage of physical maps is that their component subclones can be used as targets for sequencing. Since subclones are smaller than genomes, they present far more tractable problems. In practice, sequencing strategies are not applied directly to a target as large as the human genome, although they are to much smaller genomes, such as those of certain bacteria.²

An important and recurring debate among structural genomicists is when and in what proportions to use directed strategies as opposed to random strategies. Examples of both of these types of strategies are discussed further below, but the strategies can be characterized briefly: random strategies are cheap but necessarily redundant, with an exponentially decreasing return on investment as a project progresses; directed strategies are expensive but non-redundant, with a constant rate of return. This dichotomy immediately suggests compromise. It is possible to begin a project with a random strategy and switch to a directed strategy in a “finishing” phase. Timing this switch is an important economic choice.

² Sequencing and mapping can be interleaved. A promising strategy is described by Venter et al. (1996) and analyzed in detail by Siegel et al. (in preparation). Bacterial genomes sequenced to date include those described in Fleischmann et al. (1995), Fraser et al. (1995), Himmelreich et al. (1996), Bult et al. (1996), Kaneko et al. (1996), and Blattner et al. (1997).

Until 1993, when I was able to approach the mathematics of random subcloning, little was known about the expected outcomes of random subcloning projects given certain levels of resource expenditures and parameter choices. A simple equation for expected target coverage was communicated to the genomics community by Clarke and Carbon in 1976. In 1988, Lander and Waterman addressed several additional issues, and developed additional equations describing random subcloning projects, but these equations were valid only at low redundancies, and had little value for strategy determination. In fact, because of their flaws at high redundancies, these equations were often misleading. Most random strategies were thus conducted and evaluated empirically, but seldom could enough empirical data be collected to yield generalizable results.

The most fundamental question to ask of a random subcloning project is, “How many clones on average does one have to sequence or analyze before I completely cover my target?” Before we address this question, let us step back a bit, and discuss why a simple directed strategy, sequence walking, fails.

1.2 SEQUENCE WALKING

If one were presented with an unknown genome of 3×10^9 bp, it would seem that the easiest and most efficient way to sequence it would be to sequence all 3×10^9 bp in order, just one time, and be done with it. This strategy is known as “sequence walking” and it, in fact, works quite well, at least for very short genomic targets. The difficulties of sequence walking lie in the technical details of sequencing.

One cannot start a sequencing reaction, which will produce a sequence read, at any arbitrary point in the genome. Each sequencing reaction must be primed, and priming requires known sequence. A sequencing reaction can only start where the sequence is already known. The reaction can extend into an unknown region, but it must start in a known region. The required length of this known region is not set in stone, but it is about 20 bp. Each sequencing reaction starts with a primer, which is a short DNA oligonucleotide complementary to the known region.

Each primer must be unique. The need for unique primers adds to the expense of sequence walking.³ If the known region on the target genome to which the primer binds is

³ The commercial cost of primer synthesis is approximately \$0.70 per base, or \$15 for a typical sequencing primer (e.g., Operon sales literature, 1997). Due to significant

repeated in more than one site, the primer will bind at all these sites, and multiple sequencing reactions will occur simultaneously, producing nonsensical and useless sequence reads.⁴ There are 4^{20} different possible 20 bp primers (4^{20} is roughly 1×10^{12}). However, three additional factors must be considered. First, a primer will recognize not only a site that is exactly complementary, but also many other sites that are merely similar. Secondly, the composition of genomes is not completely random, in ways that increase the probability of a given sequence occurring more than once. Thirdly, not all primers are particularly good at initiating sequencing reactions, so many of them cannot be used.⁵ These factors limit the complexity of target templates used in sequencing reactions. It is hard to obtain sequence reads from templates that are longer than a few hundred kilobases. In addition to this, it is difficult to physically manipulate such templates in the laboratory.

With these factors in mind, it is not surprising that the quality of sequence data diminishes as the template length increases. So although sequence can be obtained from bacterial artificial chromosomes (BACs), the highest quality sequences are obtained from phage or plasmid templates, which are only a few thousand base pairs long.

Template complexity is only one problem. Consider now the basic strategy of sequence walking. The idea is to start at one end of the target and “walk” towards the other end.⁶ Each sequence read provides known sequence that can be used to generate a primer for the next sequence. If each sequence is 500 bp, and each primer is designed from the last 20 bp of the previous sequence read, then it would take 42 sequencing reactions to cross a twenty kilobase target. The redundancy R of such a project would be calculated as follows:

competition in the commercial primer synthesis field, this price is probably a good reflection of the actual cost of synthesizing primers. On-site primer synthesis can be done overnight; commercial synthesis incurs an additional delay due to shipping. These delays add to the net opportunity cost of sequence walking. On-site robotics drop the individual cost of primer synthesis, secondary to a large initial investment in equipment.

⁴ There are some clever tricks for doing multiple sequencing reactions in the same tube (i.e., Wiemann et al. 1996). However, none of these bypass the target sequence complexity issue discussed here.

⁵ For chemical reasons, primers are usually chosen so that their G-C content is about 50%. Additionally, the last base of a primer is usually chosen to be a guanine or a cytosine.

⁶ An obvious and common extension to the strategy is to start “walking” at both ends at meet at the middle.

$$R = \frac{\text{Total Sequence Obtained}}{\text{Target Length}} = \frac{nL}{G} \quad (1.1)$$

Here, and in general, n is the number of fragments sequenced in a project, L is the length of a fragment, and G is the length of the target. In this case R is 1.05, which is very close to the ideal of 1. Redundancy is one measure of the efficiency of a project. If this were the only measure of efficiency, the extremely low redundancy would be a compelling argument for sequence walking, at least on targets that were short enough so as not to be too complex for a sequencing reaction.

One might ask how a walking strategy is initiated. In many cases, sequencing targets are cloned DNA. A clone consists of unknown sequence embedded in a known “vector” sequence. The ends of the unknown target sequence are flanked by known vector sequence. It is thus a simple matter to initiate a walking strategy on cloned DNA. In other cases, such as mapping, a walking strategy can only be initiated from a previously analyzed region.

There is an additional problem that limits sequence walking. Regardless of the template complexity, some sequencing reactions will nevertheless “refuse” to work. When walking, each sequence read provides the data for priming the next, so if even one fails, the project halts.⁷ Reactions fail for many reasons, some of them unknown, but one cause might be the presence of tertiary structure such as a hairpin loop in the target DNA. Even if such failures are rare, the longer a template is, the more likely such a problem is to occur. Failure rates vary highly. If one primer fails, a nearby primer or alternative chemistry can be tried, often with success, but occasionally with repeated failure and always with project delay. In practice, few sequence walking projects tackle targets longer than 20 kb.

One of the most significant causes of the failure of a walking iteration is not so much that the reaction has failed to work, but rather that it has worked more than once. This happens when the target sequence contains sequences that are nearly identical to each other, known as repeats. Such repeats are common in metazoans and infrequent in other organisms. The presence of such a repeat brings a sequence walk to a halt as soon as the walk attempts an iteration from within a repeat. The only solution is to fragment the target

⁷ Even if the project is not terminated, a considerable amount of effort must be expended in order to continue walking, such as running sequencing reactions with alternative chemistries or using a different primer.

and initiate sequencing on a smaller subclone. If the target is going to have to be subcloned anyway, it would have been better to subclone it initially followed by detailed mapping of the subclones or, more likely, random sequencing.

Nevertheless, one of the largest headaches of sequence walking has nothing to do with chemistry, but rather administration.⁸ An administrative step must occur after every sequencing reaction in a sequence walking strategy.⁹ This step involves analyzing the data from the previous reaction, designing a new primer, synthesizing the primer, and then initiating a new sequencing reaction. Much of this administration can be implemented by computers and robots, but an unavoidable consequence is that each reaction must be done consecutively. Contemporary automated DNA sequencers have the capacity to run about fifty samples simultaneously. Well-equipped labs can process thousands of reactions per day. With a sequence walking strategy, all this capacity is wasted, as only one reaction can be done at a time.¹⁰ Keeping track of data, primers, clones, and targets can easily be the greatest challenge and cost of a walking strategy, far exceeding any savings in efficiency gained from an ultra-low redundancy.

Furthermore, an extremely low redundancy is not desirable. Errors occur in sequencing reactions at a rate of roughly 1%.¹¹ The desired accuracy of most projects is

⁸ This general observation holds for many things in life.

⁹ A detailed analysis of administrative costs is beyond the scope of the present work. From an operations research viewpoint, estimating such costs is one of the most difficult aspects of analyzing strategy costs. A series of preliminary forays into this difficult field can be found in two papers by Siegel et al. (1998b, in preparation). Many costs can be reduced by automation. However, the development costs, ultimate utility, and life cycle of robots are extremely hard to predict, or even in retrospect to calculate. Failed efforts at automation (i.e., "dud" robots) should be accounted for in such calculations. In the face of these difficulties, empiricism provides powerful insight into some of these hard to calculate costs. For example, all major sequencing labs have abandoned primer walking in favor of random subcloning. This suggests either mass delusion or a consensus that the overall opportunity cost of random subcloning is lower. However, even assuming that this global shift in strategies was originally fundamentally sound, it remains ever possible that improvements in alternative strategies might tip the opportunity cost balance in another direction. In future sections, I will point out a few potentials for such balance shifts.

¹⁰ In practice, many targets are analyzed in parallel, allowing the laboratory's full capacity to be used. However, this adds to the administrative complexity.

¹¹ The exact error rate varies as a function of position in the sequence read. Additional parameters include the sequencing chemistry, the template, the primer, the choice of automated sequencing machine (or absence thereof), and the choice of run parameters such as voltage and run time. For example, ABI sales literature claims production of 800

about 0.01%.¹² Therefore it is not sufficient to obtain a single sequence read across each region of the target. A bare minimum is to obtain one read from each strand of the target, bringing the theoretical ideal redundancy to 2. In practice, more than two reads are often needed for accuracy, particularly in places where the first two reads disagree. For targets that have highly similar repeats, even two sequence reads may not be enough to distinguish minor variations between the repeats. This presents significant challenges to a directed strategy or even to a low redundancy random strategy. A highly redundant random strategy will usually have relatively little problem resolving such repeats, often with no further need for experimentation. It can turn out to be less expensive to do more up-front random sequencing than to solve problems that arise from low redundancy data. Another approach to problem resolution is to use mapping data such as those provided by a pairwise project. Pairwise projects are described in Chapter 2. It should be noted that as advances in technology drop the error rate of sequence reads, then directed strategies become slightly more favorable.

Not all of the drawbacks to sequence walking apply to mapping projects. The technologies and chemistries involved in mapping projects are quite different. However, the problem of administrative overhead remains. Additionally, the problem of the “rare failure” becomes much worse. The iterative step in a sequence walking strategy that can fail is a sequencing reaction. The analogous step in a physical mapping project is the identification of a clone that overlaps a known map and extends into an unknown region.

bp of 98.5% accurate sequence in 8 hours on a PRISM 377 DNA Sequencer using a LongRead DNA cycle sequencing standard, beginning from base twenty of the read. Many genome centers maintain statistics on error rate as a function of sequence position, and will usually provide such statistics upon request. These statistics are constantly changing however, as centers implement incremental advances in technology. The estimate in the text of roughly 1% error is a ballpark estimate of current error rates. A detailed discussion of error rates is beyond the scope of the present work. One starting point for the acquisition of additional information is the web site of the National Human Genome Research Institute (www.nhgri.gov). Between 520 and 550 “high quality” bases are obtained for sequence reads from the *Pseudomonas aeruginosa* project underway at the University of Washington (Maynard Olson, personal communication). “High quality” is a statistic produced by the programs *PHRED* and *PHRAP* (Phil Green, author).

¹² The current standard for federally funded Genome Centers is 1 error in 10 kb (National Human Genome Research Institute guidelines). The error rate for genomic sequencing in the Hood lab from 1991 to 1994 ranged between 0.98 and 1.4 errors per 10 kb. For 1996 and 1997, the error rate was 0.16 to 0.20 errors per 10 kb (Lee Rowen, personal communication). Error rates are estimated based on discrepancies between overlapping clones derived from the same haplotype.

This identification process can also have a high failure rate. Most mapping walking projects are doomed to long-term failure, and are thus reserved for the generation of very short maps. An excellent review of the technology and difficulties of map walking has been provided by Stubbs (1992).

The considerable drawbacks of walking strategies have fueled the exploration of alternative strategies, including random subcloning.¹³

1.3 OVERVIEW OF RANDOM SUBCLONING

The forest-and-trees analogy can be rephrased in molecular biology terms: large DNA targets are intractable to direct analysis and must be broken down into smaller fragments before techniques such as restriction mapping or sequencing can be employed. Following detailed analysis of the many, a map or sequence of the target can be reconstructed.

As I move to a more technical description of random subcloning, I wish to be clear with my terminology. I mean different things by the word “fragment” when referring to mapping or sequencing. The use of the term “fragment” will allow me to discuss both methodologies concurrently and to develop a mathematical model applicable to both. Physical mapping requires fragmentation of the target. The resulting fragments are cloned into vectors. If the target was a clone, the new constructs are called “subclones.” I refer to the unknown DNA present in these clones or subclones as “fragments,” and this quite literally represents an actual physical fragment of the target. For sequencing projects I refer to each sequence read as a “fragment.” Although less literal, this definition allows me to maintain a precise analogy.

¹³ There is a constant interplay between cost, strategy choice, and advances in technology. For example, a recent advance in primer initiation chemistry may permit sequencing walking without the necessity of synthesizing a new primer for each walking iteration (Mugasimangalam et al., 1997). This method exploits “differential extension of nucleotide subsets.” Short primers from a presynthesized library upstream from target regions lacking a particular nucleotide, then extended without that nucleotide at low temperatures. Spurious priming sites are not extended due to the missing nucleotide. Subsequently the temperature is raised and the missing nucleotide added in order to complete a sequencing reaction. Early implementations of this technique entailed a compromise in sequence read length and an increased failure rate. However, optimization might eliminate such drawbacks. In any case, the use of this technique would still involve the administrative overhead of clone tracking through iterative steps.

Ideally, a direct strategy is pursued by analyzing a minimum number of fragments such that a minimum tiling path is followed. Walking, described above, is an example of a minimum-tiling-path strategy. In general, the determination of a minimum tiling path requires prior knowledge of the relation of each fragment to the original target. Such information is not easily available. Sequence walking obtains this data by iterative step-by-step sequencing. An alternative to sequence walking is to physically map a large number of subclones before sequencing any of them. If the number of subclones mapped is much larger than the number needed to define a minimum tiling path, it is usually possible to choose a path of subclones to sequence that approaches a minimum tiling.

In a prototypical random subcloning sequencing project, only one end of a subcloned fragment is sequenced. This is largely a matter of convenience, as the same primer, derived from known vector sequence, can be used for every sequencing reaction. The unknown DNA in a subclone is often much longer than a sequence read - perhaps a couple of thousand of base pairs compared to a read length of about 600 bp. Thus a tiling path for such a project will involve clone ends spaced less than one sequence read length apart.

The cost of producing a minimum tiling path map can be quite large. The exact costs of generating such "sequence-ready" maps are hard to determine, either empirically or theoretically. In all cases, however, these costs must be weighed against the alternative of using a less optimal map, or even no map at all. Without a map, fragments must be picked and analyzed at random. This limiting case is the strategy of random subcloning, also known as "shotgun sequencing."

Note that if one could analyze a single target molecule at a time, additional strategies would become available. One can imagine fragmenting a single DNA target molecule, and keeping track of each fragment and where it came from. One could analyze each fragment and then immediately reconstruct the target sequence. One would reach the ideal project redundancy of 1. It is actually possible in some circumstances to pursue such a strategy. Optical restriction mapping promises exactly this (see, for example, Anantharaman et al., 1997). Currently, it is not possible to sequence DNA in such a fashion.

During shotgun sequencing, fragments are generated from a vast number of identical target sequences, typically about a trillion. The resulting "library" from which the fragments are selected for further analysis is thus redundant. Individual fragments may

overlap in the sense that they mutually possess, in part or entirety, the same bit of target sequence. In particular, because of the effectively infinite number of fragmented target sequences, each fragment chosen at random from the resulting fragmented mixture is independent of all the other fragments. The locations from which these fragments arise can thus be considered to be uniformly distributed.¹⁴

One difficulty of random strategies is the problem of retrospectively determining from where a fragment came. If a fragment by chance happens to overlap known sequence, such as would occur on the boundary between the vector and the unknown target, the region of known DNA can be extended. If another fragment overlaps the first fragment, the known region can again be extended. The process of extending the known region of the target sequence is similar to that of sequence walking, but with the methodology reversed. In walking, one first identifies an unknown fragment overlapping known sequence, and then analyzes. In shotgunning, one first analyzes a very large number of fragments, and then finds one that overlaps known sequence.

The beauty of the “assembly” stage of a shotgun project is that it is not linear. Since every fragment has been analyzed up front, they all represent known sequence. The relationships of the known sequences to the original unknown target sequence are not known, but the sequences themselves are. Every fragment is a “seed” from which longer known sequence can be “grown” by identifying overlapping fragments. Every fragment is a starting point. Shotgun assembly works like a polymerization reaction, with eventual coalescence of all the fragments into one final assembled sequence. If enough fragments have been analyzed, this final assembly will continuously cover the target sequence, and the project will be over.

If not enough fragments have been analyzed, there will be gaps in the target. This is undesirable. On the other hand, analyzing more fragments than necessary to cover the target is also undesirable. Understanding the mathematics of shotgunning permits a judicious choice of the number of fragments to be analyzed up front. If not enough fragments are analyzed at first, a trial assembly can be made, and if gaps are discovered,

¹⁴ In practice, fragments from some locations are observed less often than others. The deviations from uniformity depend on the technique used to fragment the DNA (see, for example, Deininger, 1983). However, most modern techniques, such as shearing by HPLC, tend to be quite uniform. As long as the deviations in start site uniformity are small compared to the fragment length, a uniform distribution works well as an approximation.

more fragments can be analyzed until the gaps are closed. Alternatively these gaps can be closed by other strategies, such as walking. Determining the costs of gap closure of various strategies is important in choosing between alternative approaches to gap closure. It should be emphasized, however, that iterative strategies are undesirable, as they require extra administrative overhead.¹⁵ It is more desirable to analyze all necessary fragments at once and then make one final assembly than to analyze fewer fragments at first than necessary, assemble them, determine the need for more analysis, and then repeat the process, perhaps several times.

Assembly of analyzed fragments is a fascinating theoretical challenge. For random projects, no map exists to determine that two fragments lie adjacent to each other. Initially, decisions of adjacency must be made by pairwise comparison of fragments. Two fragments that overlap will share a portion of target sequence in common. By looking for these common sequences, adjacency can be detected. Sequence reads may contain errors, so fragments may be declared to overlap even if their common sequences do not match perfectly. Furthermore, even if two sequences match each other perfectly, or nearly so, the fragments from which they are generated may not overlap, as the target may contain two or more nearly identical sequences. Often higher-order comparisons, with checks for consistency between three or more sequences, are necessary to resolve ambiguities in assembly. Assembly is a task best done by computers. Algorithms for assembly, such as that of *PHRAP*, are constantly improving (Phil Green, author).

¹⁵ The costs of iterative steps are undesirable, but not necessarily prohibitive, depending on the strategy. An example of strategically desirable iteration is illustrated by the “sequence tagged connector” (STC) strategy of Venter et al. (1996). There are several key differences in the nature of the STC iterations and sequence walking iterations. First, STC iterations are mapping iterations that occur with low periodicity. A laboratory must cycle through one STC iteration for each BAC sequenced. If an iterative sequencing strategy were employed on a BACs completely sequenced during the course of an STC mapping and sequencing project, the complete BAC sequencing iterations would have to occur two to three orders of magnitude more frequently than the mapping iterations. Secondly, the failure rate of STC mapping iterations is predicted to be low (Siegel et al., in preparation). Furthermore, the “cost” of a rare failed STC mapping iteration will be low – in the worst case, a non-contiguous BAC will be completely sequenced, reducing the cost of future work. This cost will not be completely recouped due to inefficiencies introduced by the rogue BAC into the final sequenced tiling path. More likely than this scenario, however, would be an extremely early termination of complete BAC sequencing due to recognition of sequence inconsistencies. The STC strategy is a member of the pairwise end sequencing family of strategies, which I discuss at further length in Chapter 2.

As mentioned above, from a theoretical standpoint, random mapping and sequencing strategies can be treated identically. It is worth emphasizing, however, that they involve very different scales. Most physical mapping projects approach targets on the megabase scale or larger, such as entire genomes. These large targets are randomly fragmented into YAC, BAC, cosmid, or phage subclones ranging in size from tens of kilobases to several megabases. Analysis techniques include restriction mapping, STS content mapping, in situ hybridization, and many others.

Sequencing projects employ both smaller targets and smaller subclones. In particular, the targets of sequencing projects are often the subclones of mapping projects. Fragments of these subclones (i.e., subclones of subclones) are small enough to be employed as sequencing templates for automated DNA sequencing machines. A schematic diagram of a random subcloning project is shown in Figure 1.1.

In the increasingly automated modern scientific laboratory, an additional appeal of random subcloning is rooted in the absence of need for prior information about particular fragments. This allows projects to be undertaken with a great deal of “blind” automation and with a decreased need for highly trained human intervention. The main drawback of shotgun strategies is their dependence on overdetermination of information, with a need to generate several times as much raw data as an ideal directed strategy would. Accordingly, actual strategies may be a mix of both random and directed approaches, beginning with random and progressing to directed when the cost of choosing and sequencing directed subclones is judged to be less than the cost of continued shotgunning. Such decisions are predicated on the ability to determine such costs. Experience, simulations, and analytical models are the tools for this analysis.

In the following sections, I will present analytical models and simulations of random subcloning. I will illustrate these with empirical observations, where available.

1.4 MATHEMATICAL MODEL - BASIC FORMULATION

I assume a linear target of length G . In Section 1.13, I will modify my analysis to make it applicable to circular targets. Most sequencing targets consist of a single linear strand of DNA cloned into a circular vector, resulting in a circular construct similar to the original vector, just much bigger.

For purposes of this mathematical model these seemingly circular targets are properly considered to be linear. The reason for this is that the vector sequence is already known. No additional information is gained by analyzing fragments that arise exclusively from the vector. One method of avoiding analyzing vector sequence is to “screen” all fragments before analyzing them. A labeled probe made from vector sequences can be used to tag any fragments that contain vector sequences, allowing them to be excluded from further analysis. This is seldom done, as it involves extra labor and invites the possibility of error. In particular, unknown target sequence might be accidentally screened out by the labeled vector probe. Also, fragments that overlap both the vector and the unknown target would be screened out. This is usually undesirable, as such fragments can provide critical anchoring information during the project assembly phase.

More often, a sequencing project will not screen out vector sequences before fragment analysis begins. Thus any fragments that exclusively overlap vector will be sequenced regardless. This will represent wasted effort. However, fragments that overlap both the vector and the target contribute information to the final assembled sequence. Therefore the best way to set up a theoretical framework for the analysis of typical random shotgun sequencing projects is to ignore fragments that exclusively overlap target except for their presence as “wasted” sequencing reactions. Fragments that overlap the target by a base pair or more are not ignored, so the effective target length in such typical situations is increased by the length of a fragment (minus a base pair) on each of two ends.

Now, the total subclone length is the vector length V plus the target length:

$$\text{Subclone Length} = V + G \quad (1.2)$$

Fragments, each of length L , that overlap both the vector and the target will be included in assembly, so the total length of the linear target should be calculated as:¹⁵

¹⁵ Note that V is assumed to be greater than $2L-2$. If $2L-2 > V \geq L-1$, then V should be used in place of $2L-2$ in equation 3. If $V < L-1$, then the project should be analyzed using the results for circular targets presented in Section 1.13. In this final case, gaps in the vector sequence can be ignored, so the expected number of gaps should be modified as:

$$N_{\text{expected gaps}} = \left(\frac{G}{V + G} \right) N_{\text{predicted gaps}}$$

$$G' = 2L + G - 2 \quad (1.3)$$

Furthermore, if screening is not employed then not all analyzed fragments will be usefully included in the assembly, so the effective redundancy will be less than the actual redundancy:

$$R_{\text{effective}} = \left(\frac{G'}{V + G} \right) R_{\text{actual}} \quad (1.4)$$

I will be defining “gaps” in such a way that the uncovered target sequences, if any, at either end of the final assembled target are not counted as gaps. At the high redundancies necessary for project completion, these uncovered end regions are likely to be quite short. Furthermore, if the project is a typical shotgun sequence project as described above, these end regions will actually be vector, so it won't matter if they are not covered by analyzed fragments.

End regions do, however, become a concern if the target actually is physically linear, and not cloned into a vector. The only practical examples of this of which I am aware are genomic physical mapping projects targeted at a eukaryotic chromosomes. On average, the probability of a particular base pair of the target in such a project being covered by at least one fragment is:

$$P_{\text{coverage}} = 1 - \left(1 - \frac{L}{G} \right)^n \quad (1.5)$$

This probability drops precipitously near the ends of the target, however. In particular, the probability of the first or the last base pair being covered is:¹⁶

Other equations should be modified appropriately. None of these cases is likely, as vectors are usually longer than two fragment lengths.

¹⁶ The general equation for a base pair of a distance d away from the target end, with $0 < d < L$, is:

$$P_{\text{coverage}} = 1 - \left(1 - \frac{d}{G} \right)^n$$

These probability “edge effects” could become important if the fragment size was on the order of the target size. This does not occur in typical random subcloning projects, but a

$$P_{\text{coverage}} = 1 - \left(1 - \frac{1}{G}\right)^n \quad (1.6)$$

Therefore such projects need to employ special techniques to analyze the target ends. This will usually necessitate telomeric cloning, a discussion of which is outside the scope of this dissertation.

Another kind of project worth mentioning is one in which the target may consist of multiple linear segments. Most eukaryotic genomes contain multiple linear chromosomes, so this is the usual case for a eukaryotic mapping project. Universally, the length of each chromosome is much longer than the fragment length, so such projects can accurately be modeled using the present linear framework by setting the target length G to be the sum of all of the chromosome lengths. If the chromosomes are not present in stoichiometric amounts, appropriate considerations must be made for the underrepresented chromosome(s). This might occur if a mapping project were undertaken on an entire male genome, but most projects would seek to avoid this situation. With multiple chromosomes in a random subcloning project, both ends of each chromosome are likely to be uncovered, necessitating telomeric cloning or other approaches to the ends. It is also possible just to ignore the ends and allow their characterization, if necessary, to be carried out by a completely different project.

With the above considerations in mind, one can define the variables necessary for a mathematical analysis. For a given project, n fragments of constant length L are generated from the target and analyzed in some manner such that overlaps between fragments are detectable. This analysis would be sequencing for a sequencing project and might typically be restriction digestion for a physical mapping project. All fragments are generated from distinct identical copies of G .¹⁷ No fragments may start within $L-1$ bases of the last, rightmost base of G as such fragments would not be entirely contained within G . Thus the effective length G_e available for fragment start sites is $G-L+1$. I designate the starting, or

similar phenomenon occurs with genetic mapping of marker loci. Bishop et al. (1983) analyze this issue at length.

¹⁷ This statement is made to emphasize the claim that the fragments are independently and identically distributed. It is reasonable; the ratio of originally fragmented target molecules to randomly analyzed fragments is typically about $10^{12}:10^3$, or 10^9 . The probability that two fragments are derived from the same target molecule is negligible. Even if this were not the case, this mathematical model would likely remain quite valid.

leftmost, base pair of each fragment S_k , such that S_1 is the start site of the leftmost fragment, with $S_k \in [1, 2, 3, \dots, G_e]$. S_n begins the rightmost, or last fragment of G . The start site may be either the 5' or 3' base pair of the Crick strand of the fragment it begins, depending on fragment orientation relative to the target.

Because $G_e \gg 1$, we can consider the possible range of fragment start sites to be continuous, rather than the quantum entity that it is. In the present analysis, I will switch back and forth from discrete to continuous models without extensive justification. Continuous models tend to allow more elegant equations and derivations, while discrete formulations occasionally give more precise answers. The main difference between these two formulations is actually quite trivial. As noted above, $G_e = G - L + 1$. However, for a continuous model, $G_e = G - L$. In all cases $G \gg L$, so $G_e \approx G$. Nevertheless, without careful record keeping, the use of a continuous model will allow some asymptotic limits to converge towards slightly incorrect limits, such as $G+1$ rather than G . This is largely an aesthetic matter without much practical implication. Nevertheless, I believe that a failure of a mathematical model to converge towards an expected limit in extreme cases is a serious drawback. Therefore I will use discrete formulations when necessary to satisfy my own personal aesthetic appreciation for perfect asymptotic behavior. With this in mind, in the continuous model, the S_k are an ordered sample of n independently, identically, and uniformly distributed observations on the interval $(0, G_e)$. The formulation is drawn schematically in Figure 1.2.

An important assumption is that all the fragments are considered to be the same length. This is seldom, if ever, strictly the case. The actual length of fragments may vary by 20% or more. It turns out that this has very little effect on the utility of the mathematical model developed here. In some cases this can be demonstrated mathematically. This can be demonstrated quite easily for the equations for target coverage; I will leave this as an exercise for the reader. Also, the expected number of gaps in a project is unaffected by varying the fragment length. Siegel and Holst (1982) provide a formal proof of this for circular targets. Variations in fragment lengths can have subtle effects on the distributions of the number of gaps and the gap lengths, as well as a few other parameters. For practical purposes, these subtle effects are insignificant. Monte Carlo simulation, discussed further in Section 1.14, is another way to verify this assertion.

1.5 TARGET COVERAGE (1)

Two questions familiar to anyone who has taken a long family trip are “Are we there yet?” and “How much farther do we have to go?” For a random subcloning project, arrival means that there are no gaps in the coverage. The question of how much farther can be answered both in terms of the number of gaps that remain and in terms of what percentage of the target remains to be covered. It turns out that the number of gaps remaining turns out to be more useful for gauging the amount of additional effort necessary to complete a project. Nevertheless, the fractional target coverage is also an interesting quantity and, perhaps more importantly, is easy and fun to compute.

Consider Xeno’s paradox. Xeno shoots an arrow at a target. The arrow flies half way, covering half the distance. The arrow covers half the remaining distance, and then again half the remaining distance, and so on. Supposedly, the arrow never reaches its target. This is what happens during a random subcloning project. Actually, not quite, but let us examine the issue in more detail.

Each analyzed fragment in a random subcloning project can be considered to be generated sequentially. Each time a fragment is generated, it covers a random portion of the target. Each fragment is of length L . The fraction of the target covered by one fragment is thus L/G . A subsequent fragment will also randomly cover another region of the target of length L , on average proportionally distributed between already covered and uncovered target. Thus, on average, each additional fragment covers a fraction L/G of the remaining uncovered sequence. The actual amount of additional unknown sequence covered rapidly gets smaller and smaller, much as Xeno’s arrow will fly shorter and shorter distances with each iteration. Much as it would seem that the arrow can never reach its target, it might also seem that a shotgun project could never reach its goal of perfect coverage.¹⁴

It turns out that Xeno’s arrow will actually reach its goal, but for a different reason than a shotgun project will. Xeno’s arrow is helped along by the nature of time, infinity, and the convergence of a series towards its limit. Each iteration of the flight of Xeno’s arrow takes place in half the time of the previous iteration.¹⁵ Time is relentless, so the

¹⁴ Many researchers have also had this thought during the assembly of particularly difficult targets.

¹⁵ This assumes the arrow is not losing velocity. If it is, the arrow may very well never reach its target.

arrow hits its target. However, each generation of an analyzed fragment in a shotgun project requires the same amount of time and effort as was spent on the previous fragment. Time is not the solution to the shotgunner's dilemma.

Consider instead a blob of decaying uranium. As each half-life passes, half of the uranium decays. But not exactly. Only on average. Usually there are a vast number of uranium atoms in any blob, so an average is a fairly accurate estimate of the percent of atoms that decay during any given half-life. But consider now the case of an almost exhausted uranium blob, with only a few atoms left to decay. Quite by chance, all or none of them might decay. It is a stochastic process. Unlike Xeno's arrow driven by certainty, randomness takes charge. With only one atom left, it has a 50% chance of decaying in any given half-life. Eventually, it will decay. There are no paradoxes for blobs of uranium. The same is true for random subcloning. As a random process, eventually the target will be covered. There is a chance that it might not ever be covered, but this chance is infinitely small. These probabilities will be addressed in more detail later.

For now, let us return to the determination of expected target coverage. Because radioactivity has an exponential decay, and the analogy with random subcloning fits well, one expects to find an exponential "decay" equation for shotgunning. Assume for now that the target is a circle so that we can treat all base pairs identically. Our conclusions will also turn out to be excellent approximations for linear targets as well.

The probability that any given base pair is *not* covered is:

$$P_{\text{base not covered}} = \left(1 - \frac{L}{G}\right)^n \quad (1.7)$$

So the probability that a base *is* covered is:¹⁶

¹⁶ As a note of historical trivia, it is this equation that appears in Clarke and Carbon (1976). The first published use of the eponym "Clarke-Carbon" for equation (1.11) appeared in Waterman (1995). This equation without the eponym appeared in Lander and Waterman (1988). It also appeared, in a slightly different genomics context, in Lange and Boehnke (1982). However, before that, it may have appeared in a classroom lecture by Dr. Carbon (recollection communicated by a student). To my knowledge, the first derivation of equations (1.9) and (1.11) was provided by Robbins (1944 and 1945).

$$P_{\text{base covered}} = 1 - \left(1 - \frac{L}{G}\right)^n \quad (1.8)$$

On average, the number of covered bases will be equal to the number of base pairs times the probability that each one is covered:

$$\text{Expected Coverage} = G \left(1 - \left(1 - \frac{L}{G}\right)^n\right) \quad (1.9)$$

Equation (1.9) is often approximated. Recall that $L/G \ll 1$. Furthermore, for small x :

$$e^{-x} \approx 1 - x \quad (1.10)$$

Therefore, we can rewrite equation (1.9) as:

$$\text{Expected Coverage} = G \left(1 - \left(e^{-\frac{L}{G}}\right)^n\right) = G(1 - e^{-R}) \quad (1.11)$$

Equation (1.11) is often referred to in genomics circles as the “Clarke-Carbon” equation. Note that it has the form of an exponential decay, as we had anticipated. The Clarke-Carbon equation is diagrammed in Figure 1.3.

In future sections, I will return to the subject of coverage in somewhat more rigor and detail.

1.6 MATHEMATICAL MODEL - THE BETA DISTRIBUTION

Recall that the fragment start sites S_k are an ordered sample of n independently, identically, and uniformly distributed observations on the interval $(0, G_e)$. If one can specify the distribution of spacings between fragment start sites, then the number of gaps can be determined simply by counting the number of spacings greater than the length of a fragment. Furthermore, one can determine many other properties of interest by extending from this basic formulation. Before we go there, let us cover some basic definitions.

Let $D_k = S_{k+1} - S_k$ represent the distance between start sites, for $k=1,2,\dots,n-1$. D_0 represents the length of the uncovered target region before the first fragment, and equals S_1 . D_n is the distance $G_e - S_n$.

Assume that an overlap of length at least T is necessary and sufficient to detect adjacency of two fragments. This is clearly a simplification over the complex reality of fragment assembly.¹⁷ This assumption is very reasonable for shotgun sequencing projects, where $T \ll L$, so variations in T have little effect on the mathematical model. Some mapping projects necessitate both large and varying overlaps, and in these cases, some care must be taken in interpreting the mathematical model. Note that the use of this simple assumption is quite necessary. Without it, the mathematical model quickly becomes very complicated, and in many cases intractable. When in doubt as to whether this assumption holds, computer simulations make an excellent adjuvant to the mathematical model.

Redundancy, R , is defined as $\frac{nL}{G}$. For notational ease, I also define the effective fractional coverage f_G of the target provided by one fragment as $\frac{L-T}{G_e}$, and the effective redundancy R_e as nf_G .

By genomics conventions, an island is a maximal set of fragments each of which is connected to all other island members by at least one path of fragments overlapping by T or more. A contig is an island consisting of at least two fragments (Staden, 1980). In the genomics community, the term "contig" is occasionally used loosely as a synonym for "island." This is inconsistent with its original definition and can lead to linguistic imprecision; I discourage such usage. A contig consists of at least two overlapping fragments. An isolated fragment is not a contig; it is a singleton island.

In general, a target region not covered in any fragment is a gap. Adjacent islands are thus separated by gaps. The length of the gap between fragment start sites S_k and S_{k+1} is $D_k - L$, but a gap will only occur if $D_k > L - T$. If $D_k < L - T$, then the fragment that starts at S_k

¹⁷ Overlap is more appropriately expressed as a probability, not a certainty, and this "probability of overlap" is further affected when more than two fragments overlap at the same position. Also, repeated sequence elements in the target tend to decrease the probability of certain overlaps. These and other effects bring into question the use of the parameter T to model real projects. As long as $T \ll L$, or if an "effective" T can be defined, the current formulation will result in an adequate model. An example of a project that would probably not meet this constraint would be an STS content mapping project.

will extend at least T base pairs past the beginning of the fragment that starts at S_{k+1} , the overlap between these fragments will be detected, and no gap will occur in this interval. The two adjacent fragments will both be assigned to the same contig during assembly. Note that it is possible for the length of a gap to be negative. A negative gap length indicates that an overlap is present, but not detected. As mentioned previously, I am not counting the uncovered ends of the target as gaps.¹⁸

A simple geometric observation provides the basis and elegance for many of the equations derived from the mathematical model presented here: the domain space of the spacings D_k is the surface of the simplex $D_0 + D_1 + D_2 + \dots + D_n = G_e$, and their joint probability density is constant.¹⁹ Since $D_k \geq 0$, this surface will represent a line segment in one dimension ($n=1$), an equilateral triangle in two dimensions ($n=2$), a pyramid in three dimensions ($n=3$), and similar but hard to visualize symmetrical objects in higher dimensions. The n dimensional simplex is the shadow of the $n+1$ dimensional simplex. Executing a shotgun project with n fragments analyzed is exactly analogous to choosing a point at random from the n dimensional simplex. The Cartesian coordinates of this randomly chosen point represent $D_0, D_1, D_2, \dots, D_n$. The symmetrical nature of all of the spacings is immediately apparent. In particular, the two end spacings, D_0 and D_n , will have the same distribution as all the other spacings. This last point is sometimes hard to visualize.

Another useful analogy to shotgun sequencing is to imagine a circular piece of string representing the genome. Take a pair of scissors and make n cuts at random places in the string. You will now have n pieces of string. These pieces of string do not represent the analyzed fragments of a shotgun project, but rather the spacings between the fragment start sites.²⁰ Clearly, by symmetry, the probability distribution for the length of each piece of string will be identical. This will not change if you start with a linear piece of string and make $n-1$ cuts. This is easy to see by recognizing that the linear piece of string may very

¹⁸ This is purely a semantic issue; it is very easy to modify the equations to account for a definition that defines uncovered target ends to be gaps. It is not quite sufficient just to add 2 to the number of gaps, as there is a small probability that either D_0 or D_n will equal 0, resulting in no gap at one target end or the other. Also, in a typical shotgun sequencing projects such end "gaps," at moderate to high redundancies, will actually be within vector sequence (see Section 1.4).

¹⁹ To my knowledge, the first use of the simplex as an analogy for a problem of this nature was by Lévy (1939).

²⁰ Recall that all analyzed fragments have the same length.

well have just been the result of cutting a circular piece of string once. It doesn't matter if this first cut is in a non-random location, as long as all the other cuts are random.

These observations permit many probabilities of interest to be calculated by geometric considerations. Note again that in order for the model to work elegantly, G_e and D_k will usually be treated as continuous rather than discrete. This approximation is quite minor, given the scope of a genome, or even a cosmid, compared with the unit of divisibility: a base pair.

The probability distribution of a single coordinate of a randomly chosen point from a simplex is well known and characterized. This distribution is called the "beta distribution."²¹ I will use the beta distribution to obtain most of the results of interest to this present study. It should be noted that the beta distribution is a special case of a Dirichlet distribution. The Dirichlet distribution can characterize the *joint* distribution of *all* of the coordinates of a point on the simplex, rather than just one at a time. This can be important in some cases, because the length of one spacing can influence the length of another from the same project. For example, if we know that one of the spacings in a project is greater than half the length of the target, we know that none of the others are. In most cases however, I will be ignoring the correlation between spacing lengths. This is reasonable, since n is universally large for genome projects, so the correlation between any two given spacings is negligible.

An additional bookkeeping consideration should be mentioned. Strictly speaking, the beta distribution is defined on the interval $[0,1]$. However, the effective length of the target is not 1, but G_e . Some authors emphasizing mathematical purity address this issue by setting the genome size to 1, and allow the reader to retrospectively scale back results to the size of the genome. There is nothing wrong with such an approach. However, I prefer to maintain the proper proportionality throughout, as I this results in equations that are intuitively easier to grasp. Therefore, in what is to follow, I will make a few small deviations from notational orthodoxy for this purpose.

With these considerations, the density function for the beta distribution of the lengths of spacings between fragment start sites is:

²¹ In particular, this is a special beta distribution, i.e., $\text{Beta}(1,n)$. As a stylistic choice, I refer to the distribution employed in this paper as *the* beta distribution, rather than *a* beta distribution.

$$f_{D_k}(x) = n(1 - \frac{x}{G_e})^{n-1} \quad (1.12)$$

Now, the expectation of the beta distribution,

$$f_A(x) = n(1-x)^{n-1} \quad (1.13)$$

is easily verified to be:

$$E(A) = \frac{1}{n+1} \quad (1.14)$$

This makes intuitive sense, as we expect $n+1$ fragments when we make n cuts in a string of unit length. So we immediately have the expected value for the length of a spacing from equation (1.12):

$$E(D_k) = \frac{G_e}{n+1} \quad (1.15)$$

Notice that equation (1.15) specifies that the expected length of a spacing is equal to the effective target length divided by the number of spacings, which, for a linear target, is one greater than the number of fragments.

1.7 GAPS AND ISLANDS

As mentioned previously, a gap will occur following a fragment starting at S_k if and only if $D_k > L-T$. Thus, the probability of a gap following a given fragment is equivalent to the probability that $D_k > L-T$, which I will call p_{gap} :

$$\begin{aligned} p_{\text{gap}} &= \int_{L-T}^{G_e} f_{D_k}(x) dx \\ &= \int_{L-T}^{G_e} n(1 - \frac{x}{G_e})^{n-1} dx \quad \left(\text{now let } y = \frac{x}{G_e} \right) \\ &= n \int_{f_g}^1 (1-y)^{n-1} dy \\ &= (1-f_g)^n \end{aligned} \quad (1.16)$$

Recall that my definition of “gap” precludes a gap from occurring in either the first or the last spacing. Therefore this “gap probability” does not apply to D_0 or D_n , although

this does not alter the fact that these spacings are distributed identically to the other spacings.

Emphasizing the assumption that the lengths of the spacings between each S_k are independent, the distribution for the total number of gaps in a project is binomial. Again, this assumption is reasonable when there are a large number of spacings, the usual case for genome projects. In truth, there is a slight deviation from binomiality, as the occurrence of one gap will tend to inhibit the occurrence of others. Likewise, the absence of a gap in a short spacing will tend to promote the probability of a longer spacing with a gap. These effects will cause the actual distribution to have the same mean as the binomial distribution outlined here, but a smaller variance. The effects can be accurately modeled with a Dirichlet distribution.²²

In a given project there are $n-1$ opportunities for a gap to occur, one between each pair of adjacent fragment start sites. This gives us a binomial distribution for the number of gaps:

$$\begin{aligned} P(N_{\text{gaps}} = x) &\approx \binom{n-1}{x} p_{\text{gap}}^x (1 - p_{\text{gap}})^{n-1-x} \\ &= \binom{n-1}{x} (1 - f_G)^{nx} (1 - (1 - f_G)^n)^{n-1-x} \end{aligned} \quad (1.17)$$

One immediately has the probability of project closure as:

$$\begin{aligned} P(N_{\text{gaps}} = 0) &\approx (1 - p_{\text{gap}})^{n-1} \\ &= [1 - (1 - f_G)^n]^{n-1} \end{aligned} \quad (1.18)$$

There is no particular need to approximate this equation, but we can if we want, again by noting that for small x , $e^{-x} \approx 1 - x$. I will use this approximation twice in a row. So, continuing from equation (1.18),

²² The specific Dirichlet distribution is $D(1,1,1, \dots, 1)$.

$$\begin{aligned}
P(N_{\text{gaps}} = 0) &\approx (1 - e^{-R})^{n-1} \\
&\approx (e^{-e^{-R}})^{n-1} \\
&\approx e^{-ne^{-R}}
\end{aligned}
\tag{1.19}$$

The foregoing approximation is made primarily to draw a parallel with Siegel (1979), who provides an alternative derivation of equation (1.19) in more rigorous detail.

For the binomial distribution in equation (1.17), we have the expected number of gaps in a project as:

$$\begin{aligned}
E(N_{\text{gaps}}) &= (n-1)p_{\text{gap}} \\
&= (n-1)(1 - f_G)^n
\end{aligned}
\tag{1.20}$$

This is an exact equation. There is no particular reason to approximate it, except to illustrate parallels with other models, such as that of Lander and Waterman (1988). With this in mind, one could write equation (1.20) as follows:

$$E(N_{\text{gaps}}) \approx ne^{-R} \tag{1.21}$$

As noted above, the distribution for the number of gaps can be made exact with a Dirichlet distribution, which amounts to a summation of appropriate areas of an $n+1$ dimensional simplex. This somewhat awkward but nevertheless elegant distribution is provided by Stevens (1939) and in slightly different form by Flatto and Konheim (1962).²³ Stevens' distribution is approximated by equation (1.17).

The number of islands will be one greater than the number of gaps, as each gap separates two adjacent islands. To write this definition as an equation,

$$N_{\text{islands}} = N_{\text{gaps}} + 1 \tag{1.22}$$

The expected number of islands is therefore:

²³ A perusal of the literature would be incomplete without a glance at Fisher (1940), in which there is some discussion of Stevens (1939). To my knowledge, the first genomics use of the equation of Flatto and Konheim (1962) was by Lange and Boehnke (1982).

$$E(N_{\text{islands}}) = 1 + (n-1)(1-f_G)^n \quad (1.23)$$

1.8 THE NUMBER OF CLONES IN AN ISLAND

Any fragment start site spacing longer than $L-T$ will create a gap. All other spacings will not. This divides the spacings into two categories: those that form gaps and those that do not. Now the order of spacings is arbitrary, so all possible orders of gap-forming spacings amongst non gap-forming spacings are equally likely. This observation will permit us to determine the distribution of the number of clones in an island as well as the island length.

Let z_m specify the number of fragments in the m^{th} island in a project. Now, the total number of fragments in a project is n and the total number of islands is N_{islands} , so the expected number of fragments z_m in an arbitrary island is clearly:

$$E(z_m | N_{\text{islands}}) = \frac{n}{N_{\text{islands}}} \quad (1.24)$$

This simple result is easy to obtain and whets our appetite for things to come. It also will permit us to verify that the mean of the distribution that I derive meets this expectation.

To obtain the probability distribution of z_m , formally divide the spacings $\{D_k | k=1,2,\dots, n-1\}$ into two subsets: those $D_k > L-T$ and those $D_k \leq L-T$. The number of spacings in the first subset is N_{gaps} . I will refer to these spacings as long spacings. Since there are $n-1$ total spacings, the number of spacings in the second subset is $n-1-N_{\text{gaps}}$. I will refer to these as short spacings.

Each island is bounded by two long spacings. An exception may occur for the island that begins with the first fragment, starting at S_1 . This is because D_0 might be a short spacing. Likewise, the last island ending with the fragment starting at S_n will be bounded on the end by a short spacing if D_n is short.

The number of fragments in an island is equal to one plus the number of short spacings between its two bounding long spacings.²⁴ The shorter the spacings are in an

²⁴ If the last spacing is short, then the number of fragments in the last island will be one plus the number of short spacings following its initiating long spacing. Regardless of whether or not the first spacing is short, the number of fragments in the first island will

island, the more “piled up” will be the fragments in that island. This will result in multiple coverage of areas of the island. Multiple coverage is more common at higher redundancies. In fact, the average multiplicity of coverage is exactly equal to the redundancy.²⁵

Now, all orderings of long and short spacings are equally likely, as the D_k are exchangeable. The probability distribution for z_m can be analyzed combinatorically (see approaches to similar problems by Whitworth, 1897b; also Baticle, 1935), but to maintain simplicity one may employ a continuous approximation analogous to that employed previously to model spacing length in equation (1.12). This approximation will be good as long as the number of long spacings is small compared to the total number of spacings. This will be the case at higher redundancies when there are few singleton islands. The approximation should also be acceptable at lower redundancies.

During the actual assembly of a shotgun project, even at high redundancies, some analyzed fragments never get assembled into any contigs. This will be true even after the project is completed and there are no gaps. The reason for the continued existence of these singleton islands is that they represent “orphaned” fragments. These fragments may represent extremely poor quality sequence reads, a mislabeled or mishandled clone, contamination from another project, or contamination from a vector organism such as *E. coli*.²⁶ Since these singleton islands exist at high redundancies, it can be quite deceiving to compare the average number of fragments in islands from an actual project to the expected number of fragments predicted by a mathematical model. It is more informative to examine the distributions of the number of fragments in the larger islands. An excellent way to do this is graphically, by a bar graph, for example. I would recommend the inclusion of such a graphical comparison tool in any shotgun assembly computer program, particularly as it is relatively simple to program. A valuable use of such a tool would be to detect the presence of a significant number of orphaned clones in a project by noting a deviation from the expected number of singleton islands without deviations in other areas. An overall decrease in the number of clones in islands from their predicted numbers might suggest a problem

be equal to the number of spacings preceding the first long spacing other than D_0 . I ignore these minor effects in the main discussion, but they may be accounted for, if desired, at the cost of a little algebra.

²⁵ This should not be surprising.

²⁶ Now that the complete genome sequence of *E. coli* is known (Blattner et al., 1997), this last source of orphaned fragments should be a bane of the past.

with detecting overlaps. Comparisons of reality to expectations can be extremely valuable in troubleshooting problems during the course of a project.

To continue with our task at hand, we seek to know the distribution of the number of short spacings that lie between two long spacings. We will assume that there are plenty of short spacings, so that we can treat this number as a continuous variable. We immediately recognize that we are presented with the same problem of determining the lengths of pieces of string after random cuts have been made in an original piece. The cuts are the long spacings. Recall that since all orderings of spacings are equally likely, these cuts can be considered to randomly (i.e. independently and identically) distributed. The length of the string is equal to the number of short spacings. We can thus employ the beta distribution to model the number of short spacings bounded by two long spacings. There is thus a curious methodological symmetry between determining the distribution of the lengths of the spacings and determining the distribution of the number of spacings in an island.

The long spacings, or gaps, are uniformly distributed over the continuous interval $[0, n-1-N_{\text{gaps}}]$. As before, we will need to scale the beta distribution from its defined domain of $[0,1]$, this time by a factor of $n-1-N_{\text{gaps}}$. The conditional probability density for the number of short spacings bounded by two long spacings is therefore:

$$f(x | N_{\text{gaps}}) \approx N_{\text{gaps}} \left(1 - \frac{x}{n-1-N_{\text{gaps}}} \right)^{N_{\text{gaps}}-1} \quad (1.25)$$

This is written explicitly as an approximation because it represents a continuous approximation of a discrete phenomenon. Note also that it is conditioned on the number of gaps in an assembled project. This is not a problem when comparing the state of an actual project to its expected state based on modeling, as the number of gaps in the actual project will be known. It is a problem when using the model as a predictive theoretical tool for planning a project. In this case, this distribution can be evaluated approximately by using the expected number of gaps (equation (1.20)), or at the cost of a little extra algebra it can be evaluated more precisely by employing a probability weighted summation over all possible values for N_{gaps} . At the extreme end of precision, a combinatorial approach could be used, but this would be straying quite far away from the ideal of elegance in equations.

Note that it is the number of short spacings in an island that is beta distributed. The number of fragments in an island is one greater than the number of short spacings confined by the two long spacings that bound the island. Recall that there is a fragment associated with the terminating long spacing of an island. So with z_m as the number of fragments in an island, $z_m - 1$ is the number of short spacings in that island. The conditional probability density for z_m is therefore:

$$f_z(x | N_{\text{gaps}}) \approx N_{\text{gaps}} \left(1 - \frac{x-1}{n-1-N_{\text{gaps}}} \right)^{N_{\text{gaps}}-1} \quad (1.26)$$

An additional reason to write equation (1.26) as an approximation is that here the “edge effects” of the first and last island are ignored.²⁷

Recalling equation (1.14), which gives the expected value of a beta distribution, one can calculate the expected number of clones in an arbitrary island (conditioned on the number of gaps):

²⁷ An anonymous reviewer of Roach (1995) suggested the following equation for the number of fragments in an island:

$$P(z \geq x | N_{\text{gaps}}) = \prod_{b=1}^{N_{\text{gaps}}} \left(1 - \frac{x-1}{n-b} \right)$$

This discrete form of equation (1.26) is precisely analogous to the continuous *Beta*(1, n) distribution used in the text. Considering the availability of computers, there is no reason not to use this discrete form in place of the easier-to-manipulate equation used in the main body of the text. Note that although the discrete equation is an equality, it still requires conditioning on N_{gaps} , which will be influenced to a small extent by “edge effects.” The expected value of this equation can be calculated as:

$$E(z | N_{\text{gaps}}) = \sum_{x=1}^{n-N_{\text{gaps}}} \left[\prod_{b=1}^{N_{\text{gaps}}} \left(1 - \frac{x-1}{n-b} \right) \right] = \frac{n}{N_{\text{gaps}} + 1}$$

Working the algebra of this last equality can be amusing.

$$E(z_m | N_{\text{gaps}}) \approx \frac{n-1-N_{\text{gaps}}}{N_{\text{gaps}}+1} + 1 = \frac{n}{N_{\text{islands}}} \quad (1.27)$$

This equation was anticipated by equation (1.24).

The fraction of singleton islands expected in a project can be obtained by integrating the probability density in equation (1.26) over the range $x \in [1, 2)$; the remaining islands will be contigs. As mentioned above, due to the approximation of continuity, equation (1.26) is most valid at high redundancies, where there are few singleton islands. Therefore, if just the number of singletons is sought, then it is more accurate to calculate this more simply as the probability that a spacing is long and is immediately followed by another long spacing times the total number of non-end spacings (i.e. excluding D_0 , D_1 , and D_n). Additionally, if D_1 is long, a singleton will occur starting at S_1 , and if D_{n-1} is long, a singleton will occur starting at S_n . This results in:

$$E(N_{\text{singletons}}) \approx (p_{\text{gap}})^2(n-2) + 2p_{\text{gap}} \quad (1.28)$$

The distribution of singletons can be well approximated, if desired, with binomial considerations, or by making use of the discrete equation given in the last footnote.

Some motivation exists to predict the length of the longest island resulting from a project, as it is a readily identifiable feature of a work in progress. In particular, a failure to achieve islands of predicted length is often an indication of a technical inability to detect overlaps, and thus points to a problem that needs to be addressed. Whitworth (1897a) shows that for a given project, if the islands are ordered by increasing number of fragments, the expected number of fragments in the x^{th} smallest island is:

$$E(\text{number of fragments in the } x^{\text{th}} \text{ smallest island} | N_{\text{gaps}}) = 1 + \frac{n-1-N_{\text{gaps}}}{N_{\text{gaps}}} \sum_{i=1}^x \frac{1}{N_{\text{gaps}} - i + 1} \quad (1.29)$$

This expected value may be substituted in equation (1.32) below, and enables the prediction of the longest expected island for a project. A couple of points should be addressed, however. First, because the expectation is conditioned on N_{gaps} , to be useful in a predictive manner, a probability weighted summation would have to be employed.

However, since equation (1.29) will be evaluated by a computer anyway, this extra computation will perhaps not be extremely tedious. Secondly, Whitworth's equation does not address higher moments, such as the variance. There is likely to be high variance in this statistic, at least for the longest island. Thus, perhaps the best use of equation (1.29) is as a curiosity. It is nevertheless valuable to bring to light Whitworth's historic contribution to this field.

1.9 ISLAND LENGTH

The distribution of the number of clones in an island enables the determination of the distribution of the length of that island. Each island is the union of one or more fragments starting at base pairs $S_k, S_{k+1}, S_{k+2}, \dots$, and S_{k+z_m-1} . The total length l_m of an island with S_k beginning its first fragment is the sum of the spacings between its fragment start sites plus the entire length of the last fragment in the island (see Figure 1.2):

$$l_m = \begin{cases} L + \sum_{x=k}^{k+z_m-2} D_x & \text{if } z_m > 1 \\ L & \text{if } z_m = 1 \end{cases} \quad (1.30)$$

Now, spacings are exchangeable in that the joint distribution of all D_k is unchanged under any permutation of subscripts. Or rephrased, the lengths of the spacings are independent of their order. Expected island length conditioned on z_m is therefore:

$$E(l_m|z_m) = L + E(D_k)(z_m - 1) \quad (1.31)$$

By assuming that z_m is equal to its average value, one may approximate expected island length as:

$$\begin{aligned}
E(l_m) &\approx L + E(D_k)(E(z_m) - 1) \\
&\approx L + G_e \left(\frac{1}{n+1} \right) \left(\frac{n}{E(N_{\text{islands}})} - 1 \right) \\
&= L + G_e \left(\frac{1}{n+1} \right) \left(\frac{n}{1 + (n-1)(1-f_G)^n} - 1 \right)
\end{aligned} \tag{1.32}$$

This approximation is most valid when the relative variance of z_m is small, the usual case for genome projects. The accuracy of this approximation can be improved with the aid of a computer by summing equation (1.31) over all possible values of z_m , rather than employing the expected value of z_m . Note that the use of $E(D_k)$ as calculated above constitutes an additional approximation, as not all spacings can be included in islands. To account for this, a modification to $E(D_k)$ must be made. Based on evidence from computer simulations, however, equation (1.32) appears to offer enough accuracy for most purposes. The modification to $E(D_k)$ is described in the following paragraph.

Spacings greater than $L-T$ form gaps, so are not included in the subset of spacings that may be included in the length of an island. To proceed, one must eliminate these spacings from the distribution of D_k (equation (1.12)) by truncating and normalizing. The expected value of this truncated distribution is:²⁸

²⁸ The form of the last algebraic expression in equation (1.33) was suggested to me by the anonymous reviewer mentioned in the last footnote. The reviewer felt this expression most effectively brought out the subtle difference between it and $E(D_k) = G/(n+1)$.

$$\begin{aligned}
E(D_k | D_k \leq L - T) &\approx \frac{\int_0^{L-T} xn(1 - \frac{x}{G})^{n-1} dx}{\int_0^{L-T} n(1 - \frac{x}{G})^{n-1} dx} && \text{(now let } y = \frac{x}{G}\text{)} \\
&= G \frac{\int_0^{f_G} y(1-y)^{n-1} dy}{\int_0^{f_G} (1-y)^{n-1} dy} \\
&= G \frac{\left[\frac{(1-y)^{n+1}}{n+1} - \frac{(1-y)^n}{n} \right]_0^{f_G}}{\left[-\frac{(1-y)^n}{n} \right]_0^{f_G}} && (1.33) \\
&= G \left(\frac{1 - (1 + nf_G)(1 - f_G)^n}{(n+1)[1 - (1 - f_G)^n]} \right)
\end{aligned}$$

Note that the only reason that this equation is written as an approximation is the assumption of continuity. Thus, it probably would also have been reasonable for me to have written it as an equality.²⁹

1.10 TARGET COVERAGE (2)

I introduced the Clarke-Carbon equation in Section 1.5. I now consider a couple of alternative ways to derive this equation. The advantage of considering these different methodologies is primarily to gain insight. In addition, this will give me an opportunity to address the mathematical history of coverage problems. This will permit a digression on the higher moments of the coverage, such as the variance. It will also allow us to interpret situations where the apparent coverage is greater than the target length.

Perhaps the most obvious way to calculate target coverage is by multiplying the number of islands by their expected length:

²⁹ The reader may have noticed that I spend some effort discussing whether or not these equations are exact or approximations, and why. These were points of misunderstanding by a few anonymous reviewers of Roach (1995), so I felt it prudent to spend the extra effort here to elucidate. An example at a ludicrous extreme makes it clear that equation (1.33) is an approximation. The reader is encouraged to explore equation (1.33) with $G=3$, $L=2$, $T=1$, and $n=2$.

$$\begin{aligned}\text{Coverage} &= E\left(N_{\text{islands}} E(l|N_{\text{islands}})\right) \\ &\approx E(N_{\text{islands}})E(l)\end{aligned}\quad (1.34)$$

In order to demonstrate rough equivalence with the Clarke-Carbon equation, we can continue to make rough approximations by substituting equations (1.23) and (1.32) into equation (1.34). Equation (1.23) can be approximated as follows:

$$\begin{aligned}E(N_{\text{islands}}) &\approx (1 + (n-1)(1-f_G)^n) \\ &\approx 1 + n(1-f_G)^n \\ &\approx 1 + ne^{-R}\end{aligned}\quad (1.35)$$

and equation (1.34) can be approximated as:

$$\begin{aligned}E(l) &\approx L + G_e\left(\frac{1}{n+1}\right)\left(\frac{n}{1 + (n-1)(1-f_G)^n} - 1\right) \\ &\approx \frac{G}{n}\left(\frac{n}{1 + ne^{-R}} - 1\right)\end{aligned}\quad (1.36)$$

Combining equations (1.35) and (1.36) gives us the Clarke-Carbon equation:

$$\begin{aligned}\text{Coverage} &\approx E(N_{\text{islands}})E(l) \\ &\approx (1 + ne^{-R})\left(\frac{G}{n}\right)\left(\frac{n-1-ne^{-R}}{1 + ne^{-R}}\right) \\ &= G\left(1 - \frac{1}{n} - e^{-R}\right) \\ &\approx G(1 - e^{-R}) \quad \left(\text{with } \lim_{n \rightarrow \infty} \frac{1}{n} = 0\right)\end{aligned}\quad (1.37)$$

The coverage may also be calculated by subtracting the sum of the gap lengths from the total target length. This would entail the following:

$$\text{Coverage} \approx G - (N_{\text{islands}} - 1)\left[E(D_k|D_k \geq \text{Gap length}) - L\right] \quad (1.38)$$

This calculation will be saved as an exercise for the reader. The necessary integral can be evaluated similarly to that of equation (1.33).

Note that an excess of negative gap lengths will result in clonal coverage in apparent excess of the total target length.³⁰ This is most apparent when T is large. This situation has been known to occur in some physical mapping projects. An apparent coverage in excess of the target length is a poor prognosticator for the efficiency of overlap detection.³¹ If the “actual” coverage is desired, the length of a gap should be calculated as $D-L$ for $D>L$; alternatively, T can be set equal to zero.

Let us consider the issue of coverage a little more generally. Problems of coverage have intrigued mathematicians for some time. Perhaps the first “useful” application of such mathematics occurred in World War II. In addition to providing the genesis for the discipline of operations research, World War II stimulated interest in coverage problems and their relation to strategic bombing.³² The exact location of bomb and shell hits was largely a random process, so the percent of the target area affected by explosions could be calculated using an approach similar to the one that I used in Section 1.5. Robbins (1944) dealt with this problem more explicitly and rigorously.³³

In brief, let a shotgun strategy be executed such that p_x is the probability of coverage of base pair x by any given fragment. The probability that the base pair is covered by at least one fragment is thus $1-(1-p_x)^n$. The expected value and higher moments of f are calculated by Robbins, with:

$$E(f) = \frac{1}{G} \sum_{x=1}^G [1 - (1 - p_x)^n] \quad (1.39)$$

When $p_x = \frac{L}{G}$ for all x , this expected value is approximated by equation (1.11). The exact manner of coverage is not important. That is, p_x times G equals the number of base pairs sequenced in each of the n coverage iterations, regardless of whether or not the base pairs are contiguous. Therefore this equation will hold for the pairwise projects discussed

³⁰ This is something that will not be predicted from the simple application of the Clarke-Carbon equation and is one of the advantages of the present methodology.

³¹ Other conditions, such as extreme library contamination, could also produce this effect. In any case, it is not a good sign.

³² I mention operations research here because operations research methodology has much to offer genomics. See, for example, Siegel et al. (1997a, 1997b, and 1998).

³³ Robbins was unaware that he was working on a genomics problem. Thus by employing his equations in this context, we are in essence beating swords into plowshares.

in the next chapter. For linear targets p_x is not constant, and falls off near the edges, as discussed in Section 1.5. Despite this, unless L is a significant fraction of G , the Clarke-Carbon equation remains an adequate approximation to that of Robbins (equation (1.39)).

Note that if coverage is determined by the method of Robbins, or by the Clarke-Carbon equation, the expected island length can be calculated directly, rather than using the approach of Section 1.9. This is done merely by dividing the coverage by the expected number of islands. However, such an approach removes any insight into the distribution of island lengths, which could otherwise be obtained.³⁴

1.11 COMPARISON WITH THE LANDER-WATERMAN EQUATIONS

In 1988, in a watershed paper, Lander and Waterman published a model which formed a cornerstone of strategic genomic analysis. Their main concern in this model was to provide a mathematical model for the early physical mapping efforts underway at that time. Many of these efforts were done at low redundancies with the goal of building partial fragmented maps. The equations were not intended to model shotgun sequencing. As a result, the equations were valid at low redundancies, but not at high redundancies. Unfortunately, many subsequent workers have misinterpreted these results and applied them erroneously to high redundancy situations such as more advanced maps or to shotgun sequencing.³⁵ This has led to some remarkably incorrect claims about the state, or expected state, of completion of several projects.

The Lander-Waterman (L-W) equations were reworked with more painstaking detail by Port et al. (1995); I will use this more recent paper as a reference for the comments that follow. Two relevant L-W results are:

³⁴ This tabulation can be approached by replacing the expectations in equation 1.32 with their respective distributions.

³⁵ Surprisingly, there is neither discussion of the limits of the accuracy nor simulations for the equations in Lander and Waterman (1988). It is particularly unfortunate that several of the figures in this paper graph equations into moderate to high redundancies where inaccuracies occur, making it easier for a casual reader to be misled.

(I.i) The expected number of islands:

$$\begin{aligned} E(N_{\text{islands}}) &= ne^{-R\left(1-\frac{T}{L}\right)} \\ &\approx ne^{-R} \quad \left(\text{with } \lim_{T \rightarrow 0} \frac{T}{L} = 0\right) \end{aligned} \quad (1.40)$$

(I.v) The expected length of an island:

$$\begin{aligned} E(l) &= L \left[\left(\frac{e^{R\left(1-\frac{T}{L}\right)} - 1}{R} \right) + 1 - \frac{T}{L} \right] \\ &\approx L \left[\left(\frac{e^R - 1}{R} \right) + 1 \right] \quad \left(\text{with } \lim_{T \rightarrow 0} \frac{T}{L} = 0\right) \end{aligned} \quad (1.41)$$

I provide the limits of automatic overlap detection (i.e., $T=0$) to make the following discussion clearer.

The most striking thing about the L-W equations is their behavior as the redundancy grows. To wit:

$$\lim_{R \rightarrow \infty} E(N_{\text{islands}}) = 0 \quad (1.42)$$

and

$$\lim_{R \rightarrow \infty} E(l) = \infty \quad (1.43)$$

Clearly, the expected number of islands at high redundancy is one single island that covers the target. The length of this island should be equal to the target length. An island length that approaches infinity, or that even exceeds the target length is ludicrous.³⁶ Therefore the

³⁶ Note that the *sum* of the lengths of two or more islands can exceed the target length if the overlap parameter T is large. No single island can exceed the target length. Also, it is nonsensical for T to be greater than L , so in no case should the sum of island lengths be greater than $2G$. Even this would be ludicrous.

L-W equations are not accurate at high redundancy. Let us determine how high one must go before they reach their limit of accuracy.

In the present model, let us approximate equation (1.23) as follows:

$$\begin{aligned} E(N_{\text{islands}}) &= 1 + (n-1)(1-f_G)^n \\ &\approx 1 + ne^{-R} \end{aligned} \quad (1.44)$$

Now if $ne^{-R} \gg 1$, then we can claim that the L-W equation will approximate the current model. This will occur when $n \gg e^R$.

Likewise, in the present model,

$$\begin{aligned} E(l) &\approx \frac{G}{n} \left(\frac{n}{1 + ne^{-R}} - 1 \right) \quad (\text{now let } ne^{-R} \gg 1) \\ &\approx \frac{G}{n} \left(\frac{1}{e^{-R}} - 1 \right) \\ &\approx \frac{G}{n} (e^R - 1) \end{aligned} \quad (1.45)$$

Again, the models are almost exactly equivalent as long as $n \gg e^R$. This defines an upper bound as a function of n (or R) for the accuracy of the L-W equations. This bound occurs at redundancies in excess of threefold, depending on the exact parameterization of the project. A graph of the percent relative error of the Lander-Waterman and the current (“beta”) models relative to average island lengths from simulated projects is shown in Figure 1.4. The “beta” model is better at all redundancies, and considerably better at high redundancies.³⁷

There are several key assumptions that limit the L-W equations to low redundancy. First, the fragment start points (S_k) are assumed to follow a Poisson process. This is a deceptively facile assumption to make. In particular, a Poisson distribution looks very much like a beta distribution, so it would seem legitimate to expect only small differences between the two models. However, the problem with using a Poisson distribution lies with

³⁷ Neither model is particularly good when $n < 10$, mostly due to a failure of continuous approximations and growing “edge effects.” This is moot because such cases can be analyzed discretely and, furthermore, nobody at the lab bench is interested in random projects with less than ten clones.

certain assumptions of independence. A Poisson distribution assumes that the distance between two start sites is completely independent of the distance between any other two start sites. Thus the use of a Poisson model precludes analyzing any problems that involve fragment dependence, such as those discussed in the next section.

A consequence of using the Poisson distribution is that the exact number of clones in a project can not be prespecified.³⁸ This problem is noted by Port et al. (1995). As a result, one immediately has a feeling of uneasiness while using the model. From a biological point of view, this parameter (n) is precisely the parameter over which the investigator has the most knowledge and control.

The most damaging assumption for the L-W equations is that an island must be bounded on its right by a long spacing ($D > L - T$). In fact, there is almost always an island that does not meet this assumption. The reader is asked to visualize this island before reading the answer in the footnote.³⁹

If the last spacing D_n is short, then the L-W equations will undercount the number of islands by one. This is not a very big deal if the number of predicted islands is large, but as the number of islands shrinks, an extra island becomes a significant factor. In fact, at even moderate redundancies, the L-W equations will predict less than one island in a project. This is clearly ludicrous.⁴⁰ Nor can the problem be dismissed by suggesting that a miscount by a single island is a minor deviation. Project closure is defined by the arrival at a state of one island. If one cannot predict when this state occurs, then this most important question becomes unanswerable.

³⁸ Suppose n was prespecified. Then there would be a random and independent distance between each clone start point. The sum of these distance would not necessarily add up to G . It would be close, but not quite. A few clones would have to be added to or subtracted from the mathematical model, bringing it out of synch with the reality of the project. It is not immediately clear, however, that this particular objection will cause major consequences to the accuracy of any derived equations.

³⁹ The rightmost island in a project is bounded by the spacing D_n , which at high redundancies is more likely to be short than long.

⁴⁰ An individual at a scientific meeting once gave a lecture in which they claimed that, because the Lander-Waterman equations predicted less than one island for their then-underway genome mapping project, they were therefore nearly certain to have no gaps in their project. The importance of making the limitations of mathematical models clear to the end user cannot be overemphasized.

Furthermore, the prediction of island length is profoundly altered by the miscount of islands. Island length is, in essence, calculated as the reciprocal of the number of islands times the coverage. Thus, the difference of a miscount of one in the estimated number of islands, say between 0.1 and 1.1, results in a tenfold overestimate of expected island length. This miscount of one island is noted in Port et al. (1995), but no solution is offered, nor are the bounds of island length accuracy determined.

It might be naively suggested that the L-W equations can be “fixed” merely by adding one to the number of islands predicted. Unfortunately, the L-W equations are only off by *exactly one* in the limit as redundancy tends towards infinity. Below this limit, they are off by slightly less than one. Therefore, the determination of island length and project closure probabilities would remain open questions. Thus, what appears to be a subtle problem is actually somewhat difficult to fix, and requires that a Poisson model be discarded so that the interdependence of fragment spacings can be accounted for. What appear to be mere “edge effects” shake the foundation of the model. Nor does it turn out that increasing the genome size will diminish these “edge effects.” The reason for this is that in real projects the number of fragments n increases proportionally with the genome size to keep the redundancy R constant. In order for the “edge effects” to be negligible, the number of islands must be large, which occurs only at low redundancies, regardless of the genome size.⁴¹

One additional note: The L-W equations were developed for linear targets. Nevertheless, as we shall see in Section 1.13, it turns out that they are more accurate when applied to circular targets (i.e., bacterial genomes).⁴²

1.12 ISLAND CO-DEPENDENCY

One must be careful when considering more than one island from a given project, as their lengths are not independent. For example, if there are two islands in a project, it

⁴¹ It might be argued that I have shrugged off some “edge effects” during the course of the development of the “beta” model. This was done with due consideration for their effects on the relevant variables. The reader is encouraged to verify that the impact of these particular “edge effects” on calculations of interest to genomics is negligible.

⁴² Some intuition for this at first surprising result can be gained by re-reading the last few paragraphs. In some ways, a circular target *is* an infinite target, and has no “edge effects.”

might be that either one is greater than half the length of the target, but it is certain that both of them are not (barring undetected overlap). There are several calculations where it is no longer sufficient to assume that the sizes of the spacings, gaps, and islands are independent of each other. These calculations are somewhat more arcane than what has been presented hitherto, and tend to have more of a niche utility. Nevertheless, from time to time they can be useful. The calculations also provide a powerful illustration of the mathematical model constructed here, as it is extremely difficult to accurately adapt other models to the same ends.

For example, one may wish to consider the probability that a project contains at least i islands of length greater than a certain critical length C . This sort of calculation may be useful in evaluating performance of fragment-assembly algorithms or in planning projects with limited goals. Such a limited goal might be to sequence or map until at least one contig greater than a fixed length has been obtained, perhaps as part of an exploratory or preliminary survey of a target or genome.

Let N_{long} be the number of long spacings. Let g_m be the number of short spacings contained between adjacent long spacings. Now,

$$l_m \approx L + E(D)g_m \quad (1.46)$$

Thus:

$$P(l_m > C) \approx P\left[g_m > \frac{C - L}{E(D)}\right] \quad (1.47)$$

The number of short spacings preceding the first long spacing is g_0 ; the number following the last long spacing is $g_{N_{\text{long}}}$.⁴³ Now, since the distribution of long spacings is uniform among the set of all spacings, the distribution of g is defined on the simplex

⁴³ Since either D_0 or D_n may be long, N_{long} may be up to two greater than N_{gaps} . N_{gaps} is nevertheless a good approximation to N_{long} . This is true at low redundancies where $N_{\text{gaps}} \gg 2$. At high redundancies $n \gg N_{\text{long}}$, so it is unlikely that either D_0 or D_n is long. If greater accuracy is desired, an appropriate summation can be made.

$$\frac{(g_0 + g_1 + g_2 + \dots + g_{N_{\text{long}}})}{(n + 1 - N_{\text{long}})} = 1 \quad (1.48)$$

with all points of the simplex equiprobable. The reader should be by now familiar with the previous invocations of this analogy. To continue, we might employ the Dirichlet distribution instead of the beta distribution. However, rather than to invoke its full complexity here, I will borrow only a result from Stevens (1939), who was the first to address this particular aspect of the distribution.

Let R be the number of islands exceeding length C , let $c = \frac{C - L}{E(D)(n + 1 - N_{\text{long}})}$, and let k be the greatest integer less than $\frac{1}{c}$. Then one has directly from Stevens:

$$P(R \geq i | N_{\text{gaps}}) \approx \sum_{j=1}^k (-1)^{j-1} \frac{(N_{\text{gaps}} + 1)!}{(N_{\text{gaps}} + 1 - j)!(j - i)!(i - 1)!} \frac{(1 - jc)^{N_{\text{gaps}}}}{j} \quad (1.49)$$

In particular, the probability of having a contig in a project greater than half the length of the target can be approximated (e.g. $c = \frac{1}{2}$) as follows:

$$\begin{aligned} P(\text{one contig} > \frac{L}{2}) &\approx \sum_{v=0}^n P(N_{\text{gaps}} = v) P(\text{one contig} > \frac{L}{2} | N_{\text{gaps}} = v) \\ &\approx \sum_{v=0}^n \binom{n-1}{v} (1 - f_G)^v (1 - (1 - f_G)^n)^{n-1-v} \frac{v+1}{2^v} \end{aligned} \quad (1.50)$$

At moderate to high redundancies, only the first few terms of this last sum are necessary, as the subsequent terms rapidly diminish. In any case, this formula is best evaluated by a computer.

1.13 CIRCULAR TARGETS

The equations presented in the foregoing sections were designed explicitly with linear targets in mind. It is somewhat simpler, however, to write similar equations for circular targets. The choice to begin with equations for linear targets was motivated by the fact that the vast majority of targets are linear. Even targets that are seemingly circular, such as cosmids and BACs, need to be treated as linear due to the presence of the vector

sequence. In current practice, only bacterial genomes would be modeled as circular targets.⁴⁴

I will briefly recap the discussion of the “beta” distribution analysis, with appropriate modifications for circular targets. Note that $G_e=G$, so that $S_k \in [1, G]$. $D_k = S_{k+1} - S_k$ is the distance between start sites, for $k=1, 2, \dots, n-1$. D_n is the distance between S_n and S_1 . We therefore have the underlying beta distribution (cf. equation (1.12)):⁴⁵

$$f_{D_k}(x) = \begin{cases} (n-1)(1 - \frac{x}{G})^{n-2} & n > 1 \\ \delta(G) & n = 1 \end{cases} \quad (1.51)$$

The expected spacing length for the circle is (cf. equation (1.15)):

$$E(D_k) = \frac{G}{n} \quad (1.52)$$

The gap probability for the circle is (cf. equation (1.16)):

$$p_{\text{gap}} = (1 - f_G)^{n-1} \quad (1.53)$$

The gap distribution for the circle is (cf. equation (1.17)):

$$\begin{aligned} P(N_{\text{gaps}} = x) &= \binom{n}{x} p_{\text{gap}}^x (1 - p_{\text{gap}})^{n-x} \\ &= \binom{n}{x} (1 - f_G)^{(n-1)x} (1 - (1 - f_G)^{n-1})^{n-x} \end{aligned} \quad (1.54)$$

The approximation for the circle closure probability is (cf. equation (1.18)):

⁴⁴ Organellar genomes are also circular, but due to their small size they are seldom targets for random subcloning.

⁴⁵ $\delta(n)$ is the Dirac delta function. In this case, it implies that with only one fragment in a project, then the sole spacing length is G .

$$\begin{aligned}
 P(N_{\text{gaps}} = 0) &\approx (1 - p_{\text{gap}})^n \\
 &= [1 - (1 - f_G)^{n-1}]^n
 \end{aligned}
 \tag{1.55}$$

The expected number of gaps in a circle is (cf. equation (1.20)):

$$E(N_{\text{gaps}}) = n(1 - f_G)^{n-1} \tag{1.56}$$

Note that the calculation in equation (1.56) for the expected number of gaps in a circular target is exact. This is because there are no “edge effects” for a circle.⁴⁶

The number of islands will equal the number of gaps, unless there are zero gaps, in which case there will still be one island. So (cf. equation (1.22)),

$$N_{\text{islands}} = \begin{cases} N_{\text{gaps}} & \text{if } N_{\text{gaps}} > 0 \\ 1 & \text{if } N_{\text{gaps}} = 0 \end{cases} \tag{1.57}$$

and the expected number of islands can be calculated as (cf. equation (1.23)):

$$\begin{aligned}
 E(N_{\text{islands}}) &= 1 \cdot P(N_{\text{gaps}} = 0) + \sum_{x=1}^n xP(N_{\text{gaps}} = x) \\
 &= P(N_{\text{gaps}} = 0) + \sum_{x=1}^n xP(N_{\text{gaps}} = x) + 0 \\
 &= P(N_{\text{gaps}} = 0) + \sum_{x=0}^n xP(N_{\text{gaps}} = x) \\
 &\approx (1 - (1 - f_G)^{n-1})^n + n(1 - f_G)^{n-1}
 \end{aligned}
 \tag{1.58}$$

A few subtle changes must also be accounted for in order to determine the distribution of the number of clones in an island. The number of long spacings is N_{gaps} . The number of short spacings is $n - N_{\text{gaps}}$. Now the number of fragments in an island is equal to one plus the number of short spacings between its bounding long spacing(s), or

⁴⁶ David Gordon and Phil Green independently brought this to my attention.

simply n if there are no long spacings. The conditional probability density for z_m is therefore (cf. equation (1.26)):⁴⁷

$$f_z(x | N_{\text{gaps}}) \approx \begin{cases} (N_{\text{gaps}} - 1) \left(1 - \frac{x-1}{n - N_{\text{gaps}}} \right)^{N_{\text{gaps}}-2} & N_{\text{gaps}} > 1 \\ \delta(n) & N_{\text{gaps}} \leq 1 \end{cases} \quad (1.59)$$

The expected value of z_m is therefore (cf. equation (1.27)):

$$E(z_m | N_{\text{gaps}}) \approx \begin{cases} \frac{n}{E(N_{\text{gaps}})} = \frac{n}{E(N_{\text{islands}})} & N_{\text{gaps}} > 0 \\ n = \frac{n}{E(N_{\text{islands}})} & N_{\text{gaps}} = 0 \end{cases} \quad (1.60)$$

The definition of island length can be expressed mathematically as (cf. equation (1.30)):

$$l_m = \begin{cases} L + \sum_k^{k+z_m-2} D_x & z_m > 1 \\ L & z_m = 1 \\ G & \end{cases} \quad \begin{matrix} N_{\text{gaps}} > 0 \\ N_{\text{gaps}} = 0 \end{matrix} \quad (1.61)$$

We can approximate expected island length as (cf. equation (1.32)):

⁴⁷ $\delta(n)$ is the Dirac delta function. In this case, it implies that if the number of gaps is one or zero, then the number of clones in an island is certain to be n .

$$\begin{aligned}
E(l_m) &\approx L + E(D_k)(E(z_m) - 1) \\
&\approx L + E(D_k) \left(\frac{n}{E(N_{\text{islands}})} - 1 \right) \\
&\approx L + G \left(\frac{1}{n} \right) \left(\frac{n}{(1 - (1 - f_G)^{n-1})^n + n(1 - f_G)^{n-1}} - 1 \right) \\
&\approx G \left(\frac{1}{n} \right) \left(\frac{n}{(1 - (1 - f_G)^{n-1})^n + n(1 - f_G)^{n-1}} - 1 \right) \\
&\approx G \left(\frac{1}{n} \right) \left(\frac{n}{(1 - (1 - f_G)^{n-1})^n + n(1 - f_G)^{n-1}} \right) \\
&\approx \frac{G}{(1 - (1 - f_G)^{n-1})^n + n(1 - f_G)^{n-1}} \\
&\approx \frac{G}{E(N_{\text{islands}})}
\end{aligned} \tag{1.62}$$

Pick whichever successive approximation you are most comfortable with. The last approximation can also be made for the linear case. This approximation is very intuitive, and can be arrived at quickly as a "back-of-the-envelope" scribble, perhaps multiplied by $(1-f_G)^n$, or maybe $(1-e^{-R})$.⁴⁸ Again, the use of $E(D_k)$ constitutes an additional approximation, as not all spacings can be included in islands. To account for this, a modification to $E(D_k)$ must be made (similar to that done in the linear case). As usual, the accuracy of this approximation can be improved with the aid of a computer by summing over all possible values of z_m , rather than employing the expected value of z_m , and by taking into account the alternative cases in the distributions fed into equation (1.59).

The literature provides an exact formula for the expected number of fragments needed for closure of a circular target (Flatto and Konheim, 1962). Begging "edge effects," this equation can also be applied to the line. It is, for the limit as $T \rightarrow 0$, and where B is the greatest integer smaller than $\frac{G}{L}$:

⁴⁸ At high redundancies the coverage $1-e^{-R}$ is very close to one, and can be so approximated.

$$E(n \text{ needed for closure}) = 1 - \sum_{k=1}^B (-1)^k \frac{(1 - k \frac{L}{G})^{k-1}}{(k \frac{L}{G})^{k+1}} \quad (1.63)$$

One can also obtain a reasonable estimate for this value from the probability of project closure, equation (1.18), by assuming that the redundancy required to obtain a 50% chance of closure is roughly equal to the expected redundancy necessary for closure.

Some results for the circle are also available for the case of a varying parameter, $L-T$, where its distribution is known (Siegel and Holst, 1982). These results include the distribution for the number of gaps and its corollary, the probability of project closure. The utility of considering such cases is evident, since in actuality both the lengths of the clones and the amount of overlap necessary for detection will vary. The effects do not have to be considered separately, but can be combined into a single new parameter (L') equal to $L-T$. Such equations will nevertheless get complicated very quickly, and it may be that rather than pursue such a route, computer simulations would be a superior option. Unless, and perhaps even if, L' varies greatly, the assumption of a constant L' is reasonable.

Variations in fragment length should not cause much concern for the average genomicist. Siegel and Holst (1982) provide a proof that the expected number of gaps is dependent only on the expected fragment length, conditioning on the number of fragments. This proof is provided for coverage of a circle. Variation of the expected number of gaps on a finite line due to variation of fragment length is thus expected only due to "edge effects" and is predicted to be small. Computer simulations confirm this prediction (data not shown).⁴⁹

1.14 SIMULATIONS AND DATA

A large number of Monte Carlo simulations of projects can be generated quickly with a computer. They provide a useful comparison to the mathematical models, and can either support them or point out areas of weakness. The *JASON Report*, an independent review of the U.S. Department of Energy's contribution to the Human Genome project, calls for solicitation and support of detailed Monte Carlo computer simulations of the complete mapping and sequencing process (MITRE Corporation, 1997). Such simulations

⁴⁹ Although the expected number of gaps remains constant with varying fragment lengths, the distribution of island sizes will change. The probability of project closure will also be affected, but except in extreme cases these effects will be slight.

have been very useful for modeling other large-scale scientific efforts, encompassing areas such as particle physics, astronomy, and oceanography. The *JASON Report* has been summarized by Koonin (1998).

Computer simulations nicely demonstrate the accuracy of the present model, as shown in Figure 1.5. Computer simulations are particularly valuable, as many approximations necessary for mathematical tractability are easily incorporated into simulations. For example, fragment lengths and other variables can be modeled as distributions rather than a constant. An exploration, allowing such parameters to vary, can help validate the approximations of mathematical model, such as the “beta” model described here. In particular, modeling the fragment length as a square pulse of 100 bp width, rather than as a constant, has almost no effect on the statistics of Figure 1.5 (data not shown).

The model presented here also agrees with experimental data, where such data is available. There are few compilations of robust statistical data taken during intermediate project assembly points, particularly because compiling a statistically significant set of such data is burdensome. Nevertheless, experience in our laboratory is that cosmid sequencing projects require redundancies around sevenfold for closure (Rowen and Koop, 1994). This is also what the mathematical model predicts. Results from other laboratories support this view (Davison, 1991; Bodenteich et al., 1994; Martin-Gallardo et al., 1994).

1.15 EXAMPLES

I will briefly present a few examples to illustrate the use of a few of the equations developed in previous sections.

1.15.1 MAPPING THE HUMAN GENOME WITH YACs

Suppose one wishes to map the human genome with restriction digested YACs. The target G , in this case the entire human genome, has a length of 3×10^9 bp. The average YAC fragment length L has an average size of around 2.5×10^5 bp. The minimum detection overlap T for restriction mapping is around 3×10^4 bp. These numbers are approximate, and would be subject to the exact implementation of the strategy.

If a project were to undertake the analysis of 2.4×10^5 YACs, one would obtain a redundancy of twenty. From the Clarke-Carbon equation (1.9), there would be an average

of 6.18 uncovered bases. From equation (1.18), the probability of closure would be 99.46%. From equation (1.21), the expected number of gaps would be 5.44×10^{-3} . From equation (1.32), the average island length would be 2.98×10^9 bp. This would be a robust project.

1.15.2 SHOTGUN SEQUENCING A BAC

Consider the task of shotgunning a BAC with a target length G of 1.5×10^5 bp. Imagine sequencing it to a redundancy of seven, which is typical for shotgunning cosmids. Assume a typical sequence read length L of 650 bp. Assume T to be 20 bp.

Sevenfold redundancy will entail sequencing 1.62×10^3 fragments. The Clarke-Carbon equation predicts an average of 135 uncovered bases. The probability of closure will be 17.40%. The expected number of gaps will be 1.75. The average island length will be 5.48×10^4 . From equation (1.50), the probability of obtaining an island greater than half the target length will be 78.21%.

If we were to sequence a 3.5×10^4 bp cosmid target to a redundancy of sevenfold, our probability of closure would be 70.40%. Obtaining analogous project goals for a BAC requires higher redundancies than for a cosmid.

1.15.3 THE CHOICE OF PHAGE CLONES, BACs, OR COSMIDS AS SEQUENCING TARGETS

Planners of high-throughput genome sequencing projects are faced with the decision of what type of clone to use as targets for shotgun sequencing. One can address the sequencing costs associated with different choices of clones with the aid of the equations presented here. To consider a simple example, hypothesize that a mapping protocol has produced an approximation to a minimum tiling path of sequencing targets, with adjacent targets overlapping by 5 kb. Assume that each clone is sequenced to the expected redundancy needed for closure, and then any remaining gaps are closed by directed sequencing. Continue to assume $L=650$ bp and $T=20$ bp. Consider an overall project goal of sequencing a 100 Mb genome (but the results can easily be scaled to an arbitrary genome size).

If the targets are λ clones with $G=20$ kb, then the expected number of fragments for λ closure is 177 (equation (1.63)), which is a redundancy of 5.75. The expected number of

gaps per λ clone (equation (1.20)) will be 0.50. The average gap length will be 126 bp.⁵⁰ Considering the 5 kb target overlaps, one needs to sequence 6667 target λ clones to span 100 Mb. The total number of fragments sequenced will be 1.18×10^6 . Each gap has a one third chance of being covered by sequence from an overlapping λ clone, so a total of $(0.50) \times (6667) \times (0.66) = 2231$ gaps need to be closed.

If the targets are cosmids with $G=35$ kb, then the expected number of fragments for cosmid closure is 345, which is a redundancy of 6.42. The expected number of gaps per cosmid will be 0.58. The average gap length will be 99 bp. Considering the 5 kb target overlaps, one needs to sequence 3334 target cosmids to span 100 Mb. The total number of fragments sequenced will be 1.15×10^6 . Each gap has a 17% chance of being covered by sequence from an overlapping cosmid, so a total of $(0.58) \times (3334) \times (0.83) = 1598$ gaps need to be closed.

If the targets are BACs with $G=150$ kb, then the expected number of fragments for BAC closure is 1869, which is a redundancy of 8.10. The expected number of gaps per BAC will be 0.69. The average gap length will be 66 bp. Considering the 5 kb target overlaps, one needs to sequence 690 target BACs to span 100 Mb. The total number of fragments sequenced will be 1.29×10^6 . Each gap has a 3.5% chance of being covered by sequence from an overlapping BAC, so a total of $(0.69) \times (690) \times (0.965) = 460$ gaps need to be closed.

Assuming the cost of generating all three maps is equal, then clearly cosmids are a better choice than λ clones, as there are fewer overall sequence reads with fewer gaps to be closed. If BACs are employed instead of cosmids, then the required number of sequence reads increases by 1.38×10^5 , but the number of gaps to be closed by directed sequencing drops by 1138. So if the price of closing a gap by directed sequencing is less than 121 times the price of a single random sequence read, then a cosmid strategy would be cheaper. In reality, the cost of constructing a BAC map will be less than constructing a cosmid map, which will bias the choice of sequencing targets towards BACs.

⁵⁰ See the discussion in Section 1.10 for approaches to calculating gap length. A useful quick approximation for the expected gap length is:

$$E(\text{gap length}) = \frac{Ge^{-R}}{E(N_{\text{gaps}})}$$

Note that the above analysis assumes that the projects stop at the expected redundancy for closure and then move to directed gap closure. The choice of when to stop is another important parameter in strategy choice.⁵¹ Consider the next example.

1.15.4 WHEN TO STOP RANDOM SUBCLONING AND START DIRECTED GAP CLOSING

From the last example, one can recognize that there may be an optimal juncture at which to stop random subcloning and start directed sequencing. This juncture will depend on the relative costs of random and directed sequencing. Let us consider several possible cost ratios for the case of BAC sequencing: 10, 20, 75, and 1000. A directed to random cost ratio of 10 might represent a scenario with negligible primer synthesis costs and completely automated clone selection and primer design for gap closures. A ratio of 1000 might represent a scenario with high personnel and primer synthesis costs.⁵²

The expected costs to close a BAC are graphed in Figure 1.6. The parameters from the previous example are employed. As the cost of directed sequencing rises relative to an arbitrarily fixed random sequence read cost, clearly the overall project cost rises as well. The optimal stopping point for random sequencing occurs at a progressively higher redundancy as the cost of directed gap closure increases. Note that this example assumes that when a fixed number of sequence reads is obtained, random sequencing stops regardless of the number of gaps, and then gaps are closed by a directed strategy.⁵³

⁵¹ Knowing when to stop is important in many things: cars, genomics, and life activities in general.

⁵² A December, 1997, price quote from a commercial sequencing company gives approximately \$25 as the price of generating a single sequence read and approximately \$600 as the cost of closing a gap (Genome Systems, price quote). Costs in a genome center setting are considerably lower: approximately \$8-\$10 to generate a single read (Stephen Lasky and Maynard Olson, personal communications).

⁵³ An additional complication to consider is that some gaps are harder to close by a directed strategy than others. Short gaps that are spanned by known sequencing templates are straightforward to close with a simple walking iteration. Gaps without templates must generally be closed by first generating a PCR sequencing template, which involves extra cost and can increase the sequencing error rate. At high redundancies, particularly with a pairwise strategy (discussed in Chapter 2), all gaps are highly likely to be spanned by known sequencing templates, so the approximation of constant cost per directed gap closure is reasonable. However, at low redundancies, the average cost of a directed gap closure may rise due to increasing frequency of gaps not spanned by known templates. Figure 1.6 can be modified for this effect by assigning different gap closure costs to each type of gap and assigning the relative frequencies for gap type according to a model for the strategy in use. In this case, both the clone length and the sequence read length must

A slight decrease in expected cost might be obtained by altering the strategy in a manner that involves iterative feedback between assembly and shotgun sequencing. This would entail shotgun sequencing until a fixed number of gaps (e.g., 2) was obtained, and then closing the gaps in a directed manner. The administrative costs in executing a project in this fashion are slightly more intangible than described above. Nevertheless, this is an extremely common actual implementation of the shotgun strategy, with the number of sequence reads obtained in batches of several hundred between assemblies. Shotgunning stops when one of these incremental assemblies reduces the number of gaps to a point at which directed sequencing can begin.

1.15.5 THE RISE IN COST AS THE TARGET LENGTH IS INCREASED

A higher redundancy is needed to reach a state of expected closure for a longer target. Thus all other things being equal, shotgunning each clone in a minimum tiling path of mapped subclones of a target is cheaper than shotgunning the whole target. This is analogous to arguments that a divide-and-conquer strategy for STS-mapping the human genome chromosome by chromosome is cheaper than mapping all markers simultaneously.⁵⁴ However, creating such a tiling path has a cost. Additionally, the resulting “minimum” tiling path invariably has considerable overlap, increasing the total sequencing redundancy by perhaps 30%. Thus the decision to “divide and conquer” or to “brute force” the target must be made by comparing the decrease in shotgunning cost per base pair for shorter targets with the increase in mapping cost. This is a common dilemma in structural genomics. A typical example is a decision whether to subclone a BAC into cosmids or to directly shotgun the BAC. Another example would be the choice of subcloning a bacterial genome or shotgunning it directly.

be incorporated, as well as any mapping data, such as that obtained from a pairwise project. Incorporating these details is not difficult, but does depend on the exact strategy parameterization.

⁵⁴ This is discussed by Lange and Boehnke (1982). Note that these authors mistakenly exaggerate the cost of mapping 24 discrete linear chromosomes relative to the cost of mapping one hypothetical chromosome with length equal to the sum of the lengths of the 24 chromosomes. Presumably, this results from a specification in their computer simulations that the extreme telomeric ends of the chromosomes must be within a specified distance of a marker. A more thorough treatment of this issue is provided by Bishop et al. (1983).

We can determine the exact expected redundancy for closure of a circular target by employing the equation of Flatto and Konheim (1962). Alternatively, we can approximate the expected redundancy for closure by numerically integrating the weighted derivative of equation (1.55) (or, for a linear target, equation (1.18)). Such numerical integration is best carried out by a mathematical analysis package such as *Mathematica 3.0* (Wolfram Research), which is what I used for this purpose.⁵⁵ With respect to the expected redundancy for closure of a circular target, and for parameterizations of interest to genomics, the relative error of the numerical integration with respect to the Flatto-Konheim summation was between one and three percent. The use of numerical integration allows one to calculate probabilities not provided by Flatto and Konheim. For example, one can compute the expected redundancy necessary for a project to reach a state of three or fewer gaps as follows:

$$E(n) = \int_W^{\infty} n \frac{\partial(P(N_{\text{gaps}} \leq 3))}{\partial n} dn \quad (1.64)$$

Here, W is the minimum number of clones with which it is possible to span the target (i.e., the number of clones necessary for onefold redundancy). If the lower bound W is not used, then semantic difficulties arise as to exactly what is meant by a project with three gaps. Technically, most projects with three clones also have three gaps. We do not wish to include such cases in our integration, so we reasonably but somewhat arbitrarily specify that a completed project must have at least onefold redundancy. The probability of a project having three or fewer gaps can be approximated by summing appropriate parameterizations of equation (1.54) for the circle or equation (1.17) for the linear case:

$$P(N_{\text{gaps}} \leq 3) = P(N_{\text{gaps}} = 0) + P(N_{\text{gaps}} = 1) + P(N_{\text{gaps}} = 2) + P(N_{\text{gaps}} = 3) \quad (1.65)$$

The resulting equation becomes quite bulky but is easily handled by *Mathematica*. By employing this technique, I have calculated the expected redundancy to reach a state of three or fewer gaps for a variety of target sizes, as well as the expected redundancy to reach a state of six or fewer gaps (Figure 1.7). The Flatto-Konheim predicted redundancies for

⁵⁵ It should also be noted that the Flatto-Konheim summation employs alternating differences of ratios of very small numbers. This requires a numerical precision of approximately 400 digits for some parameterizations of relevance to genomics. The sum does not converge quickly, and cannot be approximated by a truncation.

closure are shown for reference. Note that this calculation is invariant to scale; G , L , and T can be altered proportionately with no change in redundancy costs to reach an expected number of gaps.

One can see that per-base-pair costs rise roughly logarithmically with respect to target length. This occurs when either closure or a fixed number of gaps is sought. However, if one allows projects to stop at a stage with a number of gaps proportional to the target length, then costs will rise at a slower rate, and in fact are almost invariant, at least with respect to this choice of parameters. For example, sevenfold redundancy will take a 600 kb project to a state of six gaps, while the same redundancy will take a 300 kb project to a state of three gaps. This observation tends to support the conclusion that longer is better.

1.16 RANDOM CLOSING REMARKS

The general approach to modeling finite genomes presented here may be useful when applied to other mapping and sequencing strategies, such as those based on random transposon insertion. Such strategies represent a genomics implementation of the well established theory of “coverage processes.” Hall (1988) provides a nice entry into some of the relevant mathematics literature. Solomon (1978) provides a good overview of the problem of random arcs on the circumference of a circle.

The equations derived here have many applications.⁵⁶ To begin with, a strategist is interested in the amount of work necessary to complete a project. This can be expressed as the probability of project closure at a given redundancy (equation (1.18)). These results are consistent with the expected redundancy needed for closure (equation (1.63)). I have combined results from these two equations in Figure 1.8.

Figure 1.8 highlights some of the most useful results to arise from the “beta” model for random subcloning. The figure shows that longer targets have a higher cost in redundancy to close. This means that, all other factors being equal, it is cheaper to shotgun two halves of a target separately than to do both at once. A practical application of this

⁵⁶ A reader interested in quick access to some simple Java implementations of some of the equations presented in this paper might wish to explore Andrei Grigoriev’s web pages at www.embl-heidelberg.de/~toldo/JaMBW and www.mpimp-berlin-dahlem.mpg.de/~andy/calc/mapcalc.html

observation might be to partition a multiple chromosome genome into its individual chromosomes before commencing a shotgun project (but see also Section 1.15.5).

In addition, Figure 1.8 shows that, in general, fewer longer fragments are more desirable than proportionally more shorter fragments. There are situations in sequencing projects where longer reads can be obtained, but at higher cost. The trade-off of increased cost per read versus decreased redundancy needed for closure can be analyzed. The payoff is more than linear as fragment length increases. This result, in particular, is completely unanticipated by the Clarke-Carbon formula, which predicts the same amount of coverage at a given redundancy, regardless of fragment length.⁵⁷

At high shotgun redundancies, the cost of directed sequencing is roughly constant per gap, no matter how long the gap is. This is because at high redundancies gaps are almost universally shorter than a sequence read length. Efforts to close such a gap will be equally expensive to administrate and execute whether the length of the gap is one base pair or L - T base pairs.⁵⁸

Should one choose to close gaps by continuing random subcloning, there will be an exponentially increasing cost in redundancy to close gaps as a project proceeds. Choosing whether and at which point to stop shotgunning and begin directed sequencing is a fundamental economic question. For this purpose it is useful to calculate the incremental redundancy cost of shotgun projects per gap expected to be closed. This cost can be compared with the cost of directed sequencing to determine if and when directed methodology is appropriate for a project. A graph of gap closure cost is shown in Figure 1.9.⁵⁹ The more gaps there are in a project, the cheaper it is to close them by shotgun sequencing. The cost of closing gaps rises exponentially with each successive gap closed.

⁵⁷ To rephrase: there are fewer gaps in projects with longer fragments, but these gaps are longer. Thus the total uncovered area remains constant as long as redundancy stays the same.

⁵⁸ A gap can be closed by primer walking along an existing template known to cross the gap. See Chapter 2 for discussion of one method to determine template positioning relative to a gap. In the absence of an existing sequencing template, a gap can be closed with PCR methodology.

⁵⁹ There is a subtlety here worth elucidating. Before a project starts, the redundancy cost of a project with x expected gaps is easily calculated from equation (1.20). However, once a project is underway, and a preliminary assembly has been made, the information gained from that assembly affects the prediction of how many gaps will be present in a future state of the project. In most cases this effect will be minor, particularly if the actual

One potential objection to any mathematical model for DNA cloning is that there may be regions of the target that are nearly impossible to clone. For example, some regions of the HIV genome are genetically unstable and thus absent from subclone libraries (Jon Anderson, personal communication). From a mathematical point of view, this will result in a very large local deviation from the uniformity of fragment start site distribution. Such large variations in uniformity tend to be target idiosyncratic and very difficult to model. This by no means dooms the utility of mathematical models. In fact, such cases are precisely where mathematical models can be of great service. Unexpected deviations in the actual target parameters from predicted values of these parameters serve as an indication of subcloning problems. Once detected by comparison with the mathematical model, such problems can then be addressed and corrected. Also, once knowledge is obtained of an unclonable or unanalyzable region, the model can be appropriately modified to reflect the new constraints.

There must be a constant interplay between theory and practice. Each serves to refine the other. Neither exists alone, nor is valuable without the other.

number of gaps is close to the expected number of gaps. However, the effect cannot be predicted ahead of time. Therefore the cost of closing a gap in Figure 1.9 is calculated by determining the redundancy necessary for an uncommenced project to reach a state with an expected number of gaps that is one less than the expected number of gaps of a project executed at the redundancy graphed on the abscissa.

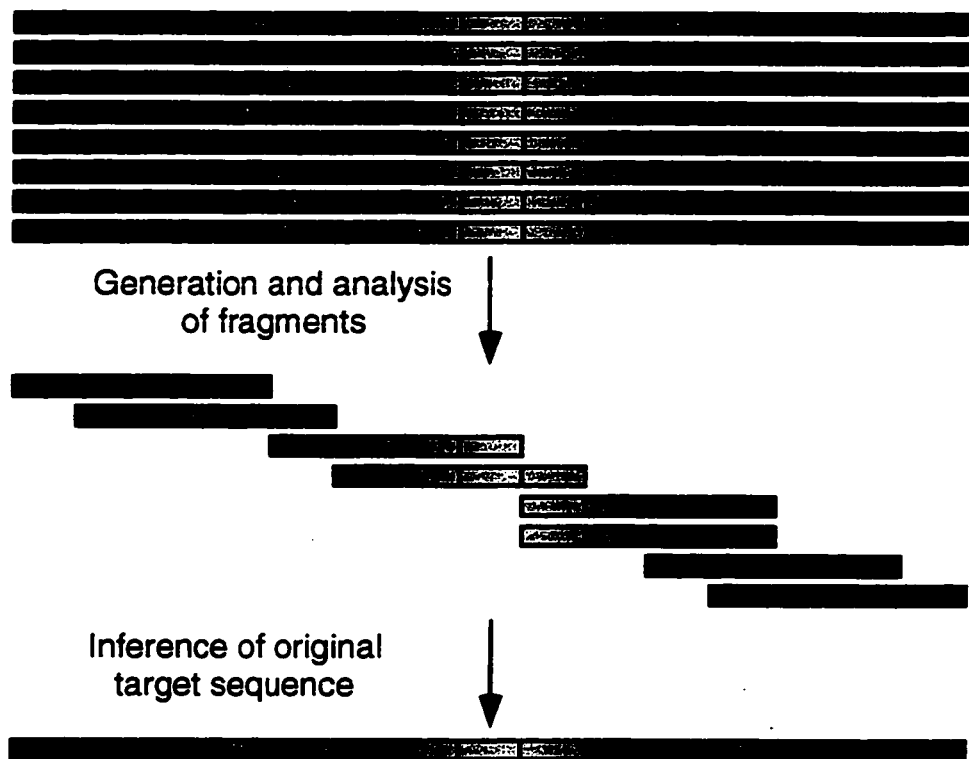


Figure 1.1. A schematic cartoon of a random subcloning project. Fragments of multiple identical copies of a target sequence (here a rainbow of colors) are analyzed and then reassembled based on bits of overlapping sequence identity.

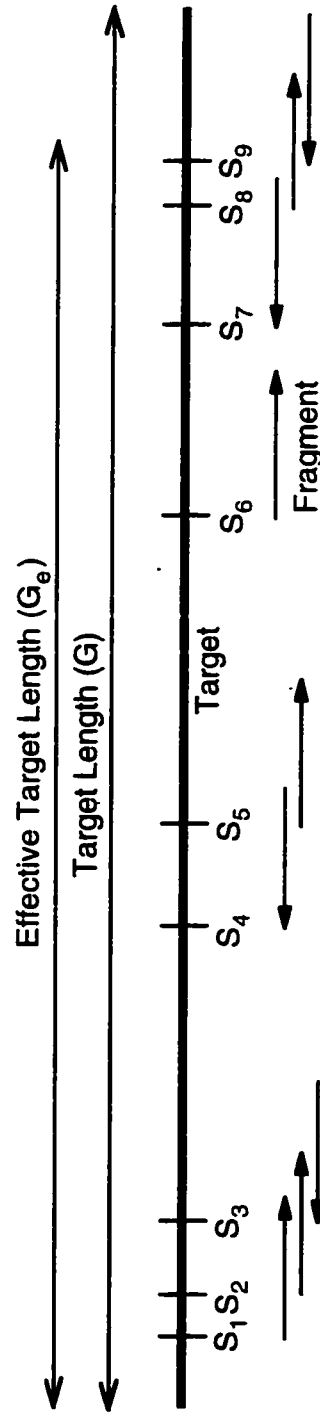


Figure 1.2. A schematic of a mathematical formulation for random subcloning. Note that the choice of orientation for the target is arbitrary. Following this arbitrary choice, the fragment orientation is ignored; the start site (S_k) of fragment k is the leftmost fragment end, whether that end is the 3' or the 5' end of the fragment.

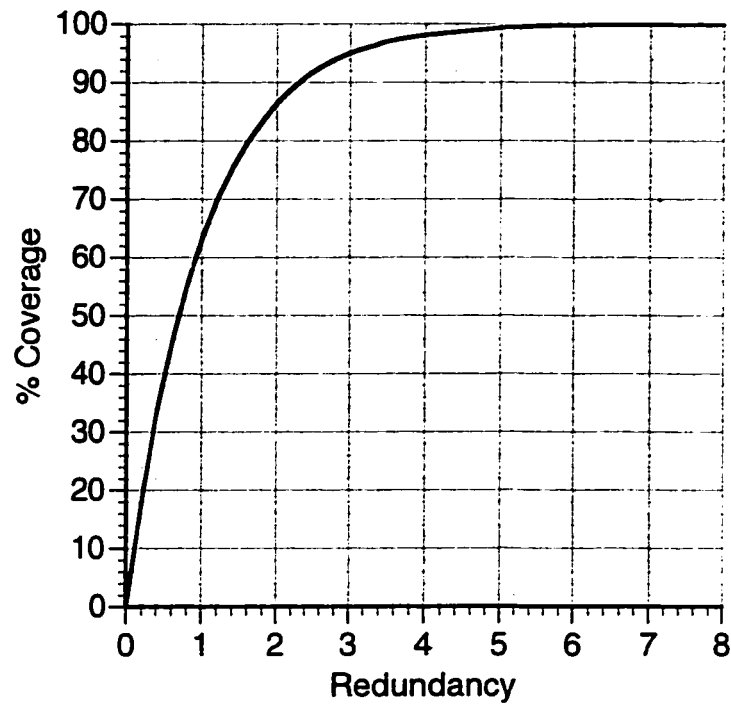


Figure 1.3. Expected coverage of a target with respect to redundancy (the Clarke-Carbon equation: Expected Fraction of Target Covered = $1 - e^{-R}$).

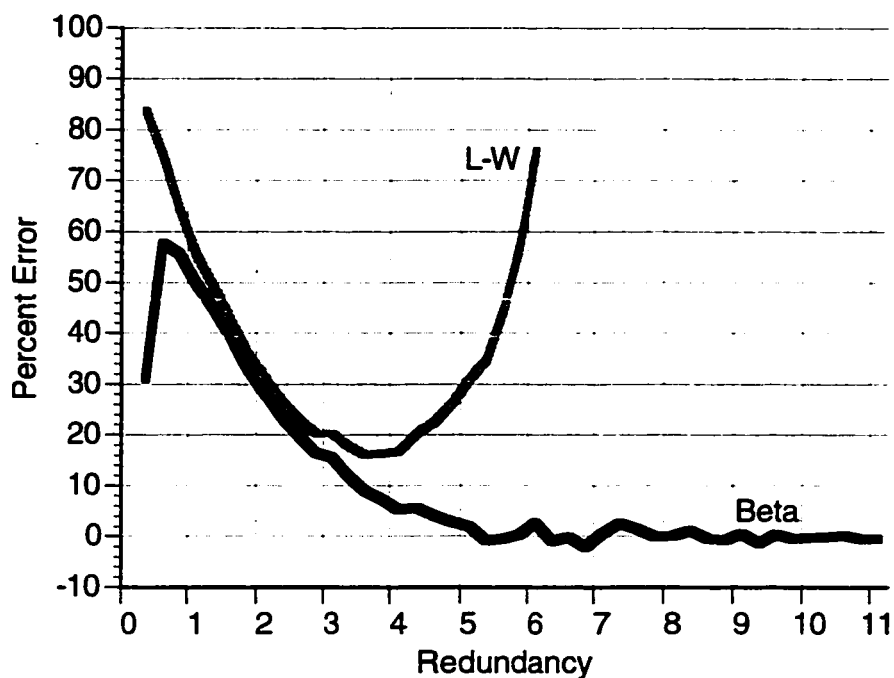


Figure 1.4. Relative error of the Lander-Waterman and “beta” models. The predicted average island lengths are plotted with respect to island lengths generated from simulations ($L=600$, $T=20$, $G=40000$). Each simulation data point is the mean value of the mean island length from 1000 project simulations. The small high-order variability is due to randomness in the simulations.

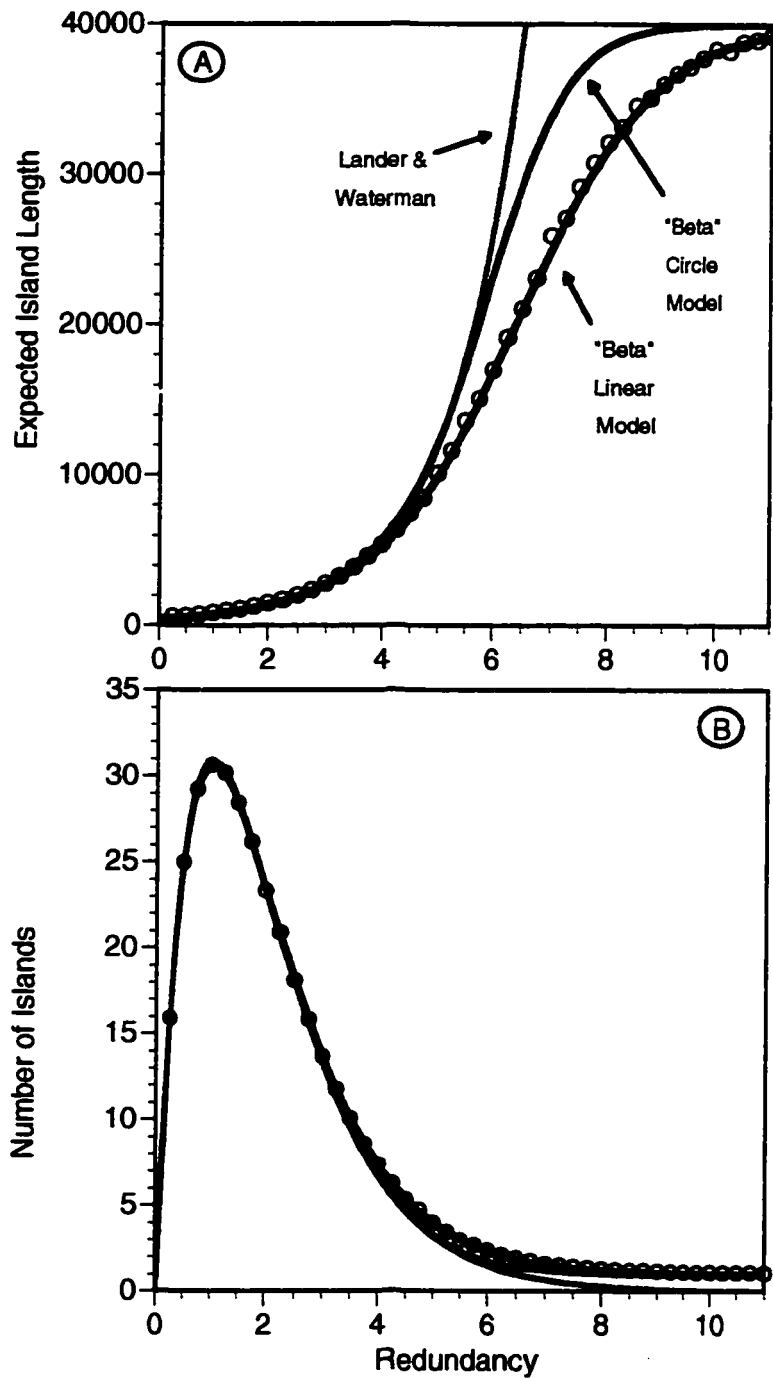


Figure 1.5. A graph of data points from computer simulations. Theoretical curves from the Lander-Waterman and "beta" models are provided for reference. Each data point represents the average of 1000 independent Monte Carlo simulations. (O) simulated data from a project with a linear target. ($G=40$ kb; $L=500$ bp; $T=20$ bp)

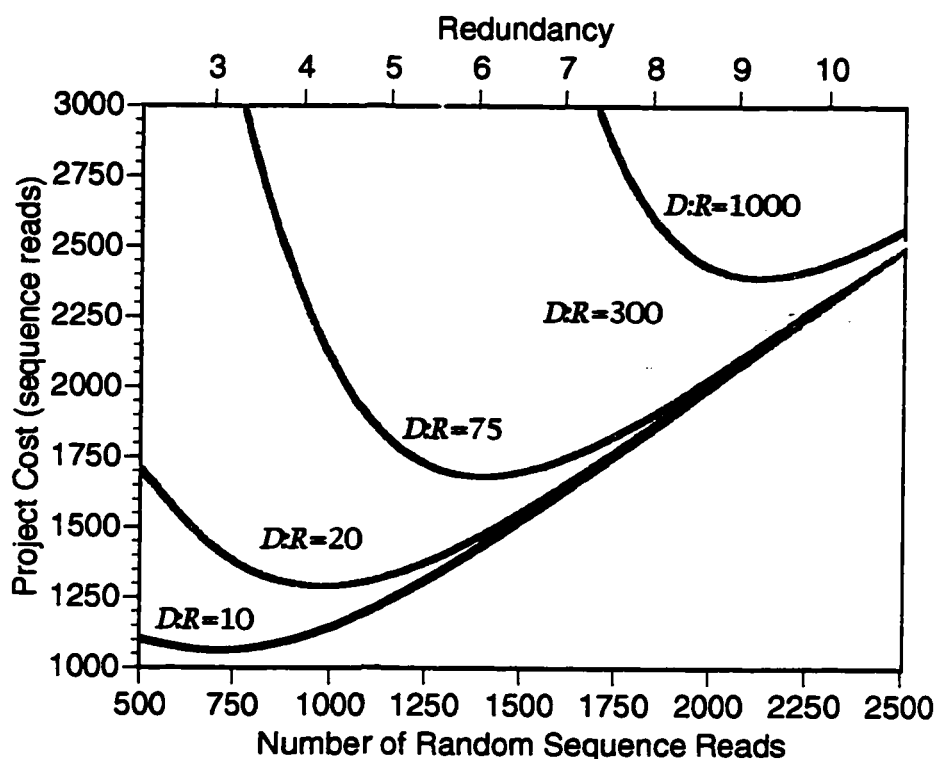


Figure 1.6. The expected cost of closure for a BAC shotgun sequencing project with directed finishing. The ratio $D:R$ is the relative cost of closing one gap versus executing a single random sequence read. Cost is standardized to the cost of a random sequence read [$\text{Cost} = n + (D:R)E(N_{\text{gaps}})$]. The abscissa represents the number of random reads obtained before directed sequencing begins. Note that all curves are asymptotic to a purely random strategy that has produced closure. ($G=150$ kb; $L=650$ bp; $T=20$ bp)

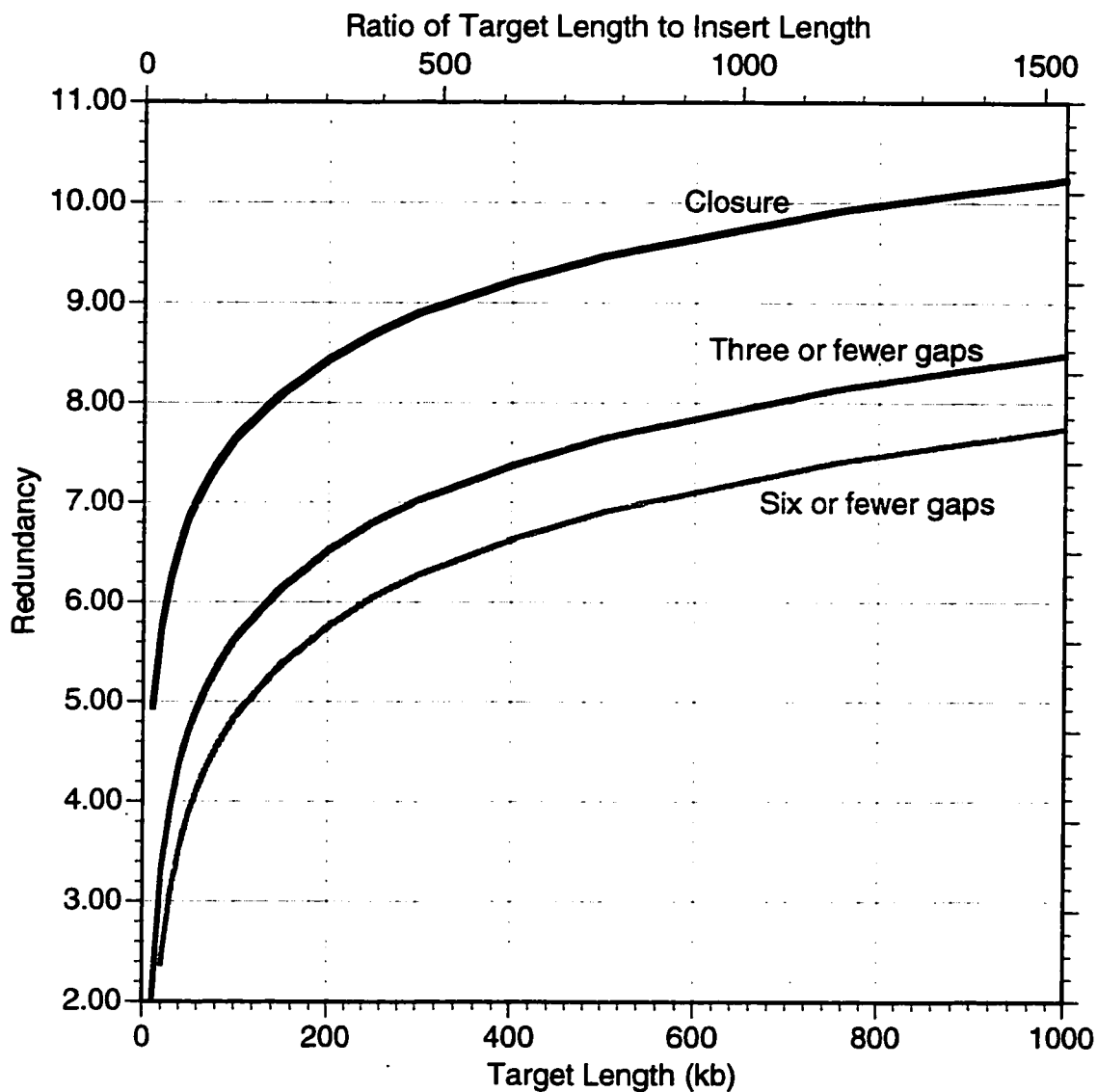


Figure 1.7. The expected cost of reaching a project state of three or fewer gaps graphed versus target size. The cost for six or fewer gaps is also graphed. These costs were calculated by numerical integration (see text). The expected costs of closure calculated by the equation of Flatto and Konheim are shown as a reference. A circular target is assumed. ($L=650$ bp; $T=20$ bp)

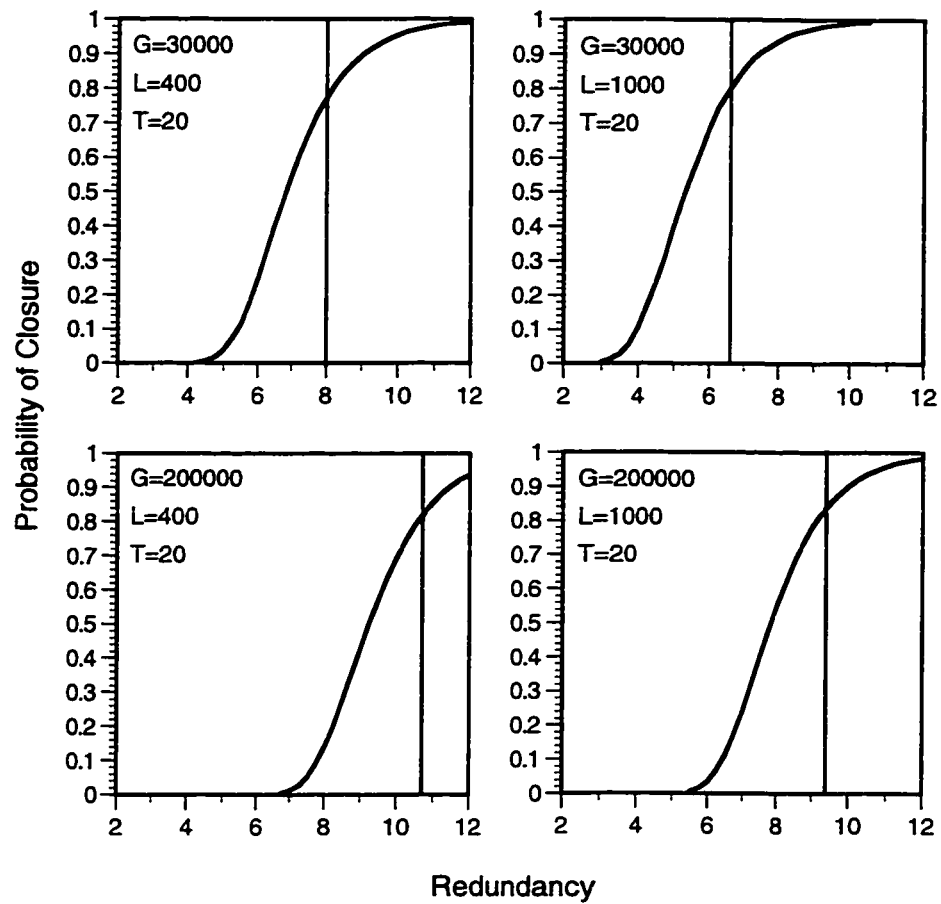


Figure 1.8. The probability of project completion with respect to redundancy, calculated using the exact equation of Stevens (1939). This equation is approximated in the text by equation (1.18). Four parameterizations are shown. The vertical lines intersect the expected redundancy necessary for closure, calculated using the exact equation of Flatto and Konheim (1962), with their parameter α set to $(L-T)/G$.

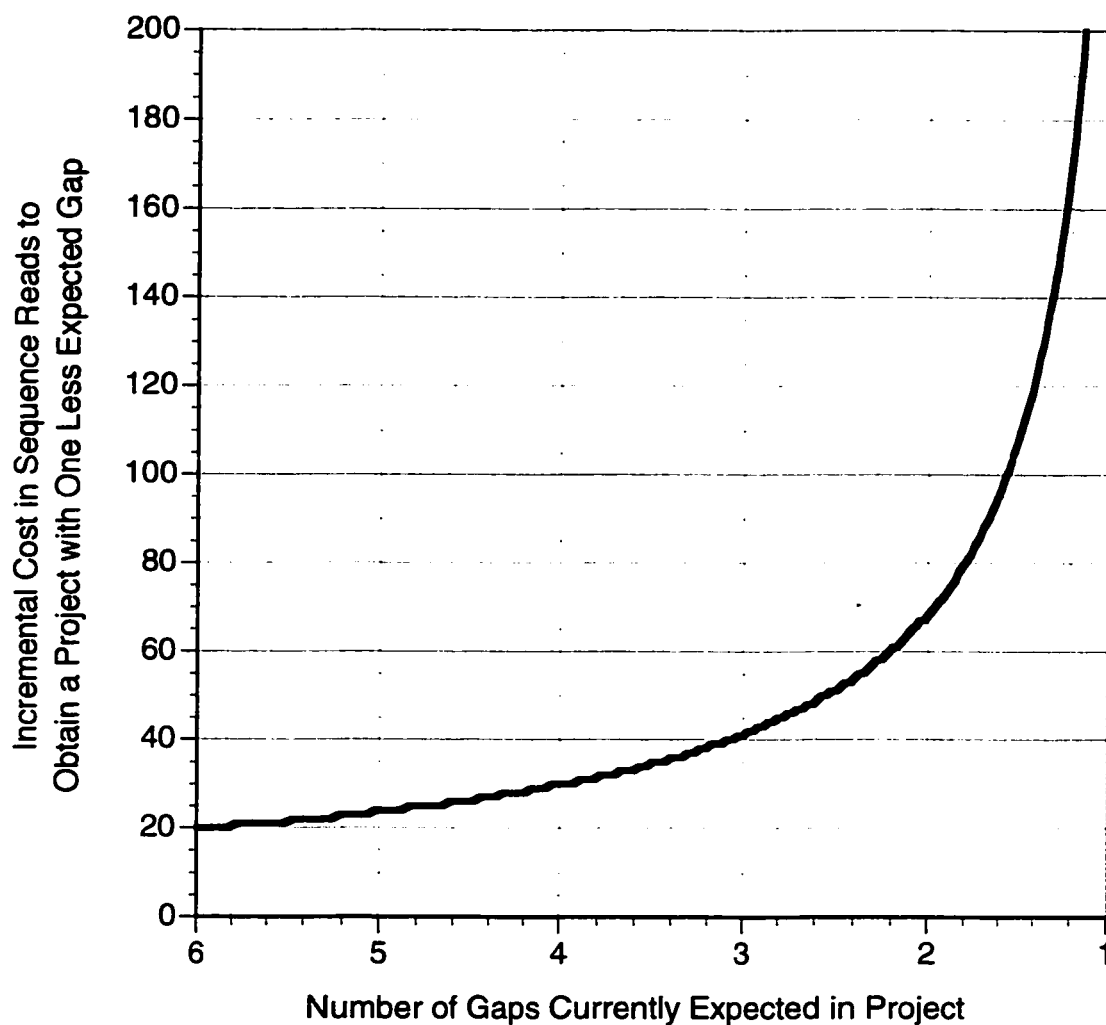


Figure 1.9. The incremental cost of closing one gap. This is calculated from the number of expected gaps in a project with no knowledge of a prior state of that project (see equation (1.20)). Note that it is impossible to plan a project with zero expected gaps, as gaps always remain a small but finite possibility. ($G=40$ kb; $L=500$ bp; $T=20$ bp)

CHAPTER 2. PAIRWISE END SEQUENCING

"Let the praises of God be in their mouth: and a two-edged sword in their hands."

The Book of Common Prayers

Random subcloning is a simple tool for mapping and sequencing DNA. In the previous chapter I provided a detailed analysis of mathematical models for random subcloning. More sophisticated strategies exist. With the maturation of the science of genomics, a wide variety of clever and innovative strategies have been developed (e.g., Evans, 1991; Burland et al., 1993; Li and Tucker, 1993; Kasai et al., 1992; Siemieniak et al., 1991). This plethora of strategies is a welcome delight to the researcher, for it adds to the armamentarium of tools available for genetic analysis. However, it brings with it the sometimes difficult decision of choosing which strategy is best.

Mathematical modeling, simulations, and experience provide the data upon which a strategy decision is made. In the previous chapter, I concentrated on a mathematical model supported by simulations as a tool for evaluating random subcloning strategies. However, it is not always possible to develop a sufficiently accurate mathematical model for a strategy. This is particularly true for the more complex strategies. In such cases, however, it is often possible to use simulations to acquire the necessary data. In this chapter I will describe a promising strategy for which a useful mathematical model has not been obtained and illustrate the use of simulations to provide data necessary for strategic decision making.

2.1 THE DOUBLE-BARREL SHOTGUN

Large-scale genomic projects are typically divided into two phases: first mapping, and then sequencing. A common strategy is to produce a rough map of approximately 40 kb completeness (terminology of Olson and Green, 1993), which is the level of cosmids.

Cosmids provided by such a mapping effort can then be employed in sequencing strategies, often using a shotgun approach.¹

After a genome has been sequenced, a map becomes useless. This is because the information in the map is redundant with information in the genomic sequence.² Mapping information is a subset of sequencing information. A map is only valuable insofar as it can reduce the cost of subsequent sequencing.³ For this reason, some have sought to combine mapping and sequencing. Because sequencing provides data that can be useful mapping information, it can make sense to begin sequencing before mapping is completed, so that the early sequence information can be used to lower the final cost of mapping. Taking this principle to its extreme, one can imagine a strategy that uses sequence data exclusively to build a map, bypassing completely the need for a separate mapping phase. Pairwise end-sequencing provides data that is particularly useful for this sort of approach. A variation of this strategy has recently been proposed as the method of choice for sequencing the human genome (Venter et al., 1996). The strategy in its pure form, dubbed “double-barrel shotgun sequencing,” is described here.

The double-barrel shotgun is a complete integration of mapping and sequencing, with a fine-scale map arising automatically from sequence data as a project proceeds. The strategy I describe retains the simplicity of random shotgun approaches, but, due to the fine-scale map produced, eliminates the need for more than minimal overdetermination of target sequence. It can half the final sequencing redundancy necessary to complete a project compared with a pure shotgun strategy. Its primary process of “scaffold building,” described below, is highly automatable and requires neither iterative steps nor intervention from highly trained individuals. At a low sequence redundancy this strategy can achieve target-spanning maps. However, since sequence accuracy is largely a function of its

¹ More recently, BACs have become a preferred sequencing template, necessitating maps of only about 100 kb completeness.

² One exception is that the map can be used as a partial check on the accuracy of the sequence. Information used to construct maps can also be used as a partial check on the integrity of a clone library. Mapping data can be used to detect chimeric and deleted clones. Thus, the actual economics of how much mapping data (and what type) to acquire is more complex than I can cover in detail here.

³ In the interim period (perhaps indefinite) between mapping and sequencing, a map can have considerable value. This interim is becoming shorter with the worldwide increase in sequencing capacity and speed.

redundancy, after the initial scaffold-building phase, a directed sequence-finishing phase will be necessary to complete the sequence. The scaffold constructed in the first phase is ideal for choosing templates for sequencing in the final phase.

Recall that DNA is double stranded. Typically a sequence is read from only one strand of a clone. A sequence read is currently 600 to 1000 bp long; a clone can be as long as 10 kb. Thus much of a clone in a pure random sequencing project remains unsequenced. Some technical and economic reasons for this were discussed in Sections 1.2 and 1.3. It is advantageous to use the same primer for every sequencing reaction, so this primer must be derived from the end of the vector sequence. However, the vector attaches to both ends of the unknown cloned DNA, forming a circle. As a result, there are two ends of vector sequence from which two separate standard primers can be derived.⁴ Therefore a clone can be sequenced at both ends. For a double barrel strategy, it is best to choose clones that are at least as long as twice the length of a sequence read. Therefore, from a sequencing perspective, one will obtain two fragments⁵ from the same clone.

From the most simplistic viewpoint, double-barrel shotgunning is merely a way to obtain twice the number of sequence reads from a set clones, thereby halving the cost of clone isolation over the course of a project. This is indeed a major advantage, but the true beauty of this approach lies in the use of the knowledge of the pairwise correlation of fragments. For each pair of fragments, not only is their separation distance known, but also their relative orientation. Together, these data enable map construction, empowered tremendously by the orientation data, and aided to some extent by the distance data.

A depiction of an executed pairwise strategy is shown in Figure 2.1.

2.2 FORMULATION

A project begins with a target of length G . The length of unknown cloned inserts is designated I . The Monte Carlo simulations presented here, except where indicated, assume a constant insert length. Practically, insert lengths seldom are less than kb or exceed 10 kb; it is this range that I will focus my attention on.

⁴ In most cases, these primers would be the m13-forward and m13-reverse primers.

⁵ The term “fragment” was defined in Section 1.3.

For these simulations I assume the sequence read length L to be a constant 400 bp.⁶ The number of inserts successfully sequenced in a project is denoted n . Since inserts are sequenced at both ends, the total number of sequence reads will be $2n$ and the total amount of sequence determined will be $2nL$. The redundancy of sequence data, denoted R_s , is defined to be $2nL/G$. Most of my results depend primarily on redundancy and only secondarily on sequence read length or quantity. Thus many short sequence reads are roughly equivalent to proportionally fewer long reads, and my choice of 400 bp for L is not critical. Note also that in mapping projects, redundancy is usually defined as the total length of all subcloned inserts analyzed. In the formulation presented here this quantity is denoted R_m and defined to be nL/G . The use of R_m permits comparison of double-barrel mapping results with other mapping techniques, such as restriction mapping.

After all insert end sequences have been determined, data can be analyzed and sequences can be assembled into islands and contigs. With a pairwise sequencing strategy, assembly of contigs is facilitated by knowledge of the pairwise orientation of sequences derived from the same insert. At low redundancies, it will not necessarily be possible to determine a single non-degenerate map for a project, as there may be sequence islands for which order or orientation is not determined. For a map to be finished, there must exist a path of bridging inserts between any two sequence islands, either directly or indirectly through other islands. Until enough redundancy is present to overcome this potential problem, there may be multiple coexisting and possibly overlapping contigs of clones. In order to address and discuss this issue, I define an ordered and oriented list of sequence islands to be a “scaffold.” Since a scaffold consists of one or more overlapping subcloned inserts, it could also be legitimately called an island, and would be if our discussion centered solely on mapping issues. Here, I reserve the term “island” to denote a set of overlapping sequence reads.⁷

⁶ This choice of sequence read length is already antiquated. The simulations presented here were initially run in early 1993. A more typical sequence read length today would be 600 bp. A motivating factor behind the choice of a short sequence read length was to present a “worst case scenario” to demonstrate the power of the double-barrel shotgun even under adverse conditions. The desirability of the strategy improves as sequence read length improves.

⁷ Port et al. (1995) refer to scaffolds as “gapped islands,” and sequence islands as “block islands.”

Beyond a certain point, as redundancy increases, the number of both islands and scaffolds will decrease, ultimately resulting in a single scaffold. Such scaffolds usually contain the entire target and as such are termed “complete.” A complete scaffold usually contains vector sequence as well, but for statistical purposes is considered to be equal to the length of the target. The longest scaffold resulting from a project is termed the “maximum” scaffold. Gaps in sequence data internal to a scaffold have previously been termed “sequence-mapped gaps,” or SMGs (Edwards and Caskey, 1991). For a complete scaffold, the size of the SMGs determines the *completeness* of the physical map, in the sense of Olson and Green (1993).

For my computer simulations, I assume that a mutual overlap of length T is necessary and sufficient to detect overlap between two sequence reads. This overlap T was set at 30 bp, but the effect of choosing a different T would be slight, particularly because $T \ll L$. Assigning T any value between 1 and 50 does not noticeably alter my results (data not shown).

For most projects, a target sequence will have been fragmented along with its vector (i.e. YAC, BAC, cosmid, phage). To minimize the sequencing of vector, one might employ target sequence as a probe to pick positive inserts, or vector sequence as a probe to screen out vector. I present here only simulations of the first strategy, which I also find to be representative of other strategies (data not shown).⁸ To this end, I assume that any insert that contains at least 40 bp of target sequence is a candidate for inclusion in a project. One advantage of this “positive screening” approach is that a few inserts will overlap vector sequence, and can be used to anchor the ends of some scaffolds to the vector.⁹ However, this effect is slight, especially with longer target lengths.

My analysis centers on two target lengths: 35 kb and 200 kb. I chose 35 kb as a representative length for cosmids. I chose 200 kb as a representative length for a BAC or YAC, to demonstrate the feasibility of using pairwise data to facilitate sequencing targets of

⁸ Screening is a labor intensive process. Therefore, as sequencing cost continue to drop, while screening costs might actually rise, it is more likely that projects of the future will not bother to screen, but rather accept the slight increased cost in redundancy due to sequencing the vector as well as the target.

⁹ If a screening approach were to be used, a “negative” screen would be more likely, as it is easier to implement. Also, “positive” screens generally have higher false positive and false negative rates.

that size or larger. All computer simulation data points represent the average of 100 determinations.

2.3 COMPUTER SIMULATIONS

Complete scaffolds are often an ideal project endpoint, so I focused on determining optimal methods for their derivation. I have also characterized the expected values for certain parameters, including average SMG size and total scaffold length. To these ends, I employed computer simulations which I in turn supplemented with a raw data simulation based on a highly redundant random shotgun project. I will discuss the computer simulations first.

The results of the computer simulations are presented in Figure 2.2 for 35 kb targets and in Figure 2.3 for 200 kb targets. In general, the number of scaffolds rises sharply at low redundancies and then declines at higher redundancies. The sharp rise occurs because each pair of insert sequence data added to a project at low redundancy has a high probability of forming a new scaffold. At higher redundancies, inserts begin to merge scaffolds and the number of scaffolds drops. For most projects, a single scaffold will form at a sequence redundancy between twofold and threefold. Slightly greater sequence redundancies were necessary to achieve single scaffolds of 200 kb targets than of 35 kb targets. Nonetheless, when 10 kb inserts were used, a single scaffold was always obtained at a redundancy less than twofold. In general, fewer scaffolds resulted when longer insert lengths were used. This is a result of longer inserts having a higher probability of spanning greater distances between sequence islands, and emphasizes the value of using as long an insert length as possible, which maximizes R_m .¹⁰

At high redundancies complete scaffolds are always obtained, as seen from the graphs of average maximum scaffold length (Figure 2.2 and Figure 2.3). For example, when 1.2 kb inserts are used for a 200 kb target, complete scaffolds are obtained around sevenfold redundancy. However, to obtain an improvement over traditional random shotgun sequencing strategies, complete scaffolds should be obtained at lower redundancies. This was clearly possible when longer insert lengths were employed. For

¹⁰ This result is intuitive, at least in retrospect. One of the more useful contributions of the work presented in this chapter was the demonstration that “longer is better,” at least with respect to pairwise insert length.

example, redundancies of twofold were sufficient to ensure complete scaffolds when 10 kb inserts were simulated. When a project resulted in a single scaffold, this scaffold was also complete, or nearly so (data not shown).

I did not notice significant differences in redundancies necessary to achieve analogous results for either 35 kb (Figure 2.2), 200 kb (Figure 2.3), or even 1 Mb targets (data not shown). This suggests that sequencing effort scales roughly linearly to results, and not exponentially, even with relatively large targets. This rough linearity stems from the use of pairwise data, and indirectly from the high mapping redundancy R_m .¹¹

The number of SMGs in maximum scaffolds increases, then decreases, as sequence redundancy increases. The initial increase is due to both the increasing length of the maximum scaffold, enabling it to contain more gaps, and to the division of large gaps into smaller gaps as sequence islands bisect them. The subsequent decrease in SMGs is due to additional sequence data closing gaps. Roughly speaking, the largest number of SMGs tends to occur in complete scaffolds that have been obtained with a minimum of sequence redundancy. Thus, at the redundancies between twofold and fourfold that we envision as reasonable for pairwise projects, a significant number of SMGs are likely to result.

For many projects a complete target sequence is desired, with no gaps fragmenting continuity. For other projects, such as gene finding, complete sequence is not a priority, but gap characterization may be of interest. In general, project design should aim for gaps no longer than a single sequence read, or at most two reads. A gap that is a sequence read long can be closed by one directed sequence from either end of the gap using the spanning insert as a template. Double stranded coverage can be obtained by a single read from each direction. A gap that is two reads long can be covered by sequence walking with one walking iteration. If gaps longer than two read lengths occur in a project, it is likely that a cost-benefit analysis will dictate continuing the random phase.

¹¹ Note that $R_m = R_s \frac{I}{2L}$. Therefore, for 10 kb inserts and 400 bp read lengths, R_m will be 25 when R_s is 2. Plugging a redundancy of 25 into the equations presented in Chapter 1 will give the reader an intuitive feel for exactly how powerful the double-barrel shotgun can be as a mapping tool. This back-of-the-envelope approach also brings home the value of using large insert lengths.

The simulations demonstrate that large gaps occur as expected at very low redundancies, but at redundancies above 1.5 average gap length tends to be less than a single sequence read length. More importantly, for all projects with sequence redundancies above twofold, the *maximum* observed gap length tended to be less than 800 bp, requiring at most two sequence read lengths to close. Occasionally longer gaps occur. For example, at a redundancy of 2.5 with a 35 kb target, 100 simulations of a project employing 2 kb inserts contained one gap greater than 800 bp in 17 cases, and two such gaps in a single case. Above twofold redundancies, there were no significant differences in SMG length resulting from alternative choices of insert size. In consequence, an occasional project will require continued random sequencing after a complete scaffold is obtained in order to eliminate long gaps.

Long insert lengths are not always convenient sequencing templates.¹² For this reason, I sought a strategy that minimized the need for longer inserts, and explored strategies that employed mixtures of insert sizes. In general, I found that benefits derived from large inserts could be obtained even when they represented a small fraction of the total number of inserts sequenced. In particular, I simulated strategies that employed a mixture of 2 kb and 10 kb inserts (Figure 2.4). For these simulations I held redundancy constant at 2.25 and assumed a 200 kb target. I found no significant differences between projects utilizing entirely 10 kb inserts and those that used only 15% 10 kb inserts.

I also envisioned strategies that mix pairwise data with data derived from a single strand only, such as might be obtained with m13 templates. A relatively small fraction of pairwise data suffices for the formation of complete scaffolds which are largely composed of single strand data (Figure 2.5). With a mixture of 60% single strand data, 30% 2 kb pairwise insert data, and 10% 10 kb pairwise insert data, a maximum scaffold was reached before threefold redundancy for a 35 kb target. This simulation addresses a practical question, for sequencing reactions will occasionally fail, which implies that most pairwise projects will be supplemented with a cohort of widowed sequences.

My simulations met the $1-e^{-R}$ expectation of the Clarke-Carbon formula (data not shown). Therefore, at any given redundancy R_s , target coverage will be the same for either a traditional shotgun or a pairwise sequencing strategy. I emphasize that increased target

¹² Discussed further in Section 1.2.

coverage is not an advantage of pairwise strategies. The advantage of pairwise strategies lies in their ability to map and not in more efficient placement of random sequences. At the redundancies of about 2.5 necessary to build complete scaffolds, target coverage will be about 92%.

The computer simulations presented here hold both sequence read length L and insert lengths I constant. In actual projects, such as that represented by the raw data simulation presented below, these parameters will vary. I have incorporated variations into several additional computer simulations (data not shown), particularly by allowing I to vary as a squarewave centered on a target value. No significant differences in predicted results were noticed when I was allowed to vary. Variations in L also have no significant effect, as long as redundancy remains constant (data not shown).

Another assumption of the computer simulations was that all target fragments are equiprobable. The accuracy of this approximation is dependent on the fragmentation method (see, for an out-of-date example, Deininger, 1983). For most cases this approximation is quite valid, for the regions of fluctuation in fragmentation probability tend to be smaller than the length of the inserts. See a more detailed discussion in Section 1.3.

2.4 RAW DATA SIMULATION

I wished to verify that results from my computer simulations accurately modeled real projects. Such projects utilize raw sequence data and might employ templates with significant repeat elements. In addition, I was interested in determining the ease of assembling scaffolds by hand.¹³ To this end, I designed a simulation built around a cosmid from the human T-cell receptor β locus that had previously been sequenced to a high redundancy using traditional shotgun sequencing.

This cosmid, designated A1-4, had been sequenced using a random shotgun strategy to a final redundancy of 8.4 (Koop et al., 1993). This cosmid consists of a 35343 bp target cloned into a 8213 bp vector. The target is notable in that it contains several repeats, including two 8.4 kb homologous elements. Their identity ranges from 85% to over 99% when 400 bp sliding windows are used for analysis. For this reason, the cosmid

¹³ Computer programs are yet to be developed that can maximally utilize pairwise data.

A1-4 was judged to represent a significant challenge for assembly (Lee Rowen, personal communication). The sequences used for the original assembly of A1-4 were derived primarily from single-stranded M13 templates and were sequenced with either Sequenase® or *Taq* cycle sequencing protocols.¹⁴

For my pairwise assembly simulation I chose a subset of these 678 sequences that might represent typical data from a pairwise project. To this end I planned for a 2.25 final redundancy R_s . I wished to pursue a strategy that employed a mix of long and short templates, so simulated 88 2.5 kb inserts and 22 7 kb inserts. I determined the start locations of these fragments with a random number generator. The length of the fragments was modified randomly with a squarewave to simulate uncertainty in fragment length, as might occur if such fragments were size selected by banding on an agarose gel. Sequence reads of the proper orientation were then chosen from the A1-4 data set to represent the pairwise end sequences of my hypothetical fragments. The closest raw sequences to my randomly generated fragment endpoints were used, although no sequence was used twice. The final range of short fragment lengths was 1738-3418 bp (2375 +/- 287 s.d.), while the range of long fragments was 5312-8245 bp (6819 +/- 781 s.d.). For my initial assembly, all sequences longer than 400 bp were clipped to 400 bp in order to demonstrate that long sequence reads are not necessary for the success of pairwise assemblies. In addition, a few sequences were shorter, although no sequence was less than 250 bp. My final redundancy R_s was thus slightly less than 2.25. I judged my protocol for sequence selection to be a reasonable approximation of what might be likely to result from an actual pairwise project.

I then assembled these pairwise sequences into a single scaffold. Before and during this assembly I was blind to the nature of the repeats in A1-4 other than that it was a “difficult” cosmid. Additionally I was blind to the exact length of the fragments, other than that they were either “long” or “short.” Sequence contigs were assembled with the software package DNA*® (Madison, WI). Scaffolds were assembled by sliding pieces of paper on a large table and were ultimately merged into a single scaffold (Figure 2.1). This assembly took about a day, illustrating that it would be a task best relegated to software implementation. Following the generation of this scaffold, each sequence contig was edited by hand for maximum accuracy. At this point, for editing purposes, the ends of sequences extending beyond 400 bp were used. Some of these sequences, although often low-

¹⁴ The Sequenase® protocol is now obsolete.

accuracy, served to verify the ordering and orienting of the contigs within the scaffold. Additionally, they helped improve the overall accuracy of the sequence data.

The results of this raw data simulation compared favorably with the averages predicted by my computer simulations (Table 1). 89% of the target sequence was represented in this scaffold. The remaining unknown sequence was contained in 17 SMGs. Sequence accuracy was 99.9%. All but one of the 44 errors were present in regions covered by only a single strand. This suggests that double-strand coverage is capable of obtaining extremely high accuracy, which could be obtained for these regions by sequencing opposite strands. The exact lengths of the 17 SMGs were unknown, but could be estimated. Ten of these SMGs were spanned by the low quality ends of sequence reads present in my data set. This data was insufficient for base calling, but allowed the estimation of gap lengths to within a few base pairs. The lengths of the remaining SMGs could be estimated based on the lengths of the fragments that spanned them. As subsequently verified, all 17 SMGs were less than 800 bp and all but two were less than 300 bp.

2.5 PERSPECTIVE ON PAIRWISE STRATEGIES

Pairwise knowledge was first used extensively during the sequencing of the HPRT locus (Edwards et al., 1990). The strategy itself was elucidated by Edwards and Caskey (1991). Smith et al. (1994) describe an approach in which the sequence islands in a scaffold can be employed as landmarks in a physical mapping project. Such landmarks were termed “mapped and sequenced tags,” or MASTs.

One notable example of a pairwise strategy has been designated “ordered shotgun sequencing” (OSS). OSS was proposed by Chen et al. (1993). OSS is characterized by a low-redundancy pairwise approach that produces multiple unlinked scaffolds which form the basis for further directed sequencing. The genome-wide strategy described in Venter et al. (1996) bears many similarities to OSS. A few preliminary simulations and a review of pairwise strategies was provided by Richards et al. (1994).

Notable recent implementations of pairwise projects include scaffold construction from the 115 kb *sigL* locus of *Bacillus subtilis* (Fabret et al., 1996), the identification of a MHC class I-like gene linked to hereditary haemochromatosis (Feder et al., 1996), the

identification of a candidate gene for Branchio-Oto-Renal syndrome (Abdelhak et al., 1997), and the complete sequencing by OSS of a 135 kb YAC (Chen et al., 1996).

It would seem that the advantages of pairwise strategies are overpowering. There appear to be no drawbacks. By executing a random strategy in a pairwise manner, one gains all the data of a traditional shotgun strategy, plus additional mapping data. This mapping data is acquired at no additional cost. In fact, the cost is slightly less, as fewer templates need to be prepared in a pairwise project. There is, however, one consideration which favors a pure shotgun approach.

The template best suited for sequencing is derived from the virus m13. This template is single stranded (ss), and can only be sequenced from one direction, preventing a pairwise strategy. Thus, for a pairwise strategy to be executed, one of two solutions must be employed. The usual approach is to use double-stranded (ds) plasmids as templates in place of ss m13. The problem with this approach is that with current methodology there is usually a small compromise in sequence read length. This creates a difficult cost-benefit decision. Should one pay a small cost in decreased sequence read length in order to benefit from the advantages of the double-barrel shotgun? In order to answer this question an exact dollar amount needs to be assigned to the components of the calculation, but once this is done the equations presented in this and the previous chapter enable the decision to be made. In most cases it is likely that the double-barrel benefits will outweigh any cost associated with a slight decrease in sequence read length.¹⁵

The second pairwise solution is the Janus strategy described by Burland et al. (1993). In this solution, after a single read is obtained from a ss m13 clone, the clone is transformed into a ds m13 plasmid, and the opposite strand is read. This permits acquisition of higher quality data, but at a considerable increase in clone handling and isolation costs. Therefore this strategy is not economically viable unless few clones are chosen for pairwise sequencing, leaving the majority of the reads as orphaned data. It is

¹⁵ In practice, it is extremely difficult to discover the actual dollar cost involved for any of the various costs, such as isolating a clone or sequencing it, and even more difficult to estimate the value of benefits such as a decrease in the difficulty of assembly due to the presence of pairwise data. Nevertheless, reasonable attempts can be made to estimate costs, particularly by standardizing on "laboratory operations" in place of dollars (see Siegel et al. 1996 and 1997). Genome Systems quotes the cost of subcloning a cosmid into m13 clones as \$1500 and the cost of generating a single sequence read as approximately \$25 (price quote, 12/97).

conceivable that such a strategy might be economically competitive with a plasmid-based pairwise strategy, as Figure 2.5 demonstrates that only a small amount of pairwise data is necessary for complete scaffold building. However, for now the high cost of converting ss m13 clones has sidelined this strategy.

It is unclear with current sequencing methodologies exactly how much is lost from the sequence read length when double-stranded in place of single-stranded templates are used. The effect appears to be rather slight, however, which will tend to bias a decision towards a plasmid-based pairwise project.

2.6 MATHEMATICAL MODELS

Little progress has been made towards the development of a mathematical model for the double-barrel shotgun strategy. An attempt was made by Port et al. (1995) to extend the approach of Lander and Waterman (1988) to pairwise data, but in addition to suffering from the drawbacks discussed in Section 1.1.1, these authors were limited to considerations of algorithmically “greedy” definitions of scaffolds.¹⁶ It is unlikely that further progress can be made with this approach.

A very simple attempt to predict the expected number of SMGs and scaffolds was offered by Edwards and Caskey (1991). Their approach was to apply the Lander-Waterman equations independently to the inserts and the sequence reads, and to assume that the number of SMGs was equal to the gaps in the sequence islands, with the number of scaffolds equaling the number of gaps in the insert islands. This approach fails to take into account any of the topological intricacies of scaffolds. It can be recommended only for its simplicity, which gives some rough insight into the number and kind of gaps likely to be present in a pairwise project.

It is my intuition that a potentially useful model for pairwise projects may eventually be developed by treating the pairwise-characterized inserts as analogs to polymer building

¹⁶ In short, the greedy algorithm blinds itself to certain intricate topological interconnections of scaffolds that consist of three or more clones. These interconnections are more likely with longer inserts, so one drawback of the greedy approach is a failure to predict the advantages of using longer inserts.

blocks in solution and applying mathematical analyses originally designed for gelling reactions. The scaffold-building process can be thought of as a gelling process.

As an alternative to the possible elegance of a gelling analogy, brute mathematical force may be utilized. For each number of inserts n , all possible topologies can be explored. The probability of each topology can be determined together with the characteristics of that topology, such as average scaffold length. The enumeration of topologies quickly becomes difficult. In Appendix A, I present an analysis for $n=\{1,2,3\}$. I have not had the patience to work out the $n=4$ case. In actual projects, n will be on the order of hundreds to thousands. The limitations of the brute force approach should be obvious.¹⁷ One result described in Appendix A is applicable to all n — increasing the insert length results in an increased scaffold length and an increased probability of obtaining a single contig. This is consistent with the results of the simulations.

2.7 DISCUSSION

The work on pairwise end sequencing presented here focuses on *utility*. I was primarily interested in determining the minimum amount of sequence redundancy necessary to reach satisfactory endpoints for projects that might actually be implemented in a laboratory. A scaffold that equals or exceeds target length is an ideal endpoint for a random strategy. Such a “maximum scaffold” is an ideal starting point for a directed strategy. I determined that it is possible to achieve such scaffolds at sequence redundancies around twofold.

A key factor in producing scaffolds at twofold redundancy is the choice of insert lengths. I found that the longer an insert is, the more useful it is. This is in considerable contrast with a misconception that the ideal insert length is three times the sequence read length.¹⁸ Nevertheless, there is a practical upper limit to useful insert size. This limit depends on three factors. First, it is difficult to routinely clone large fragments. Secondly, longer inserts have correspondingly more sequence complexity, which tends to degrade the

¹⁷ A computer algorithm could be designed to enumerate and evaluate pairwise topologies. This could surely work for cases above $n=3$. However, the problem appears to be quite “hard” in the computer sense of the word. This would imply that even a computer would have a hard time “brute-forcing” calculations at practical choices of n .

¹⁸ This misconception was common prior to 1995.

quality of the raw data. Thirdly, assembly becomes more difficult with longer fragments, as the absolute uncertainty of the length between pairwise ends tends to increase. These limitations vary in stringency depending on available technology and resources. Thus the optimal choice of fragment size may vary from one laboratory to another. However, given the option, fragment sizes should be chosen as large as possible. It should be noted that in addition to their advantages in scaffold-building, large fragments are also extremely useful in detecting and resolving repetitive elements in target sequence.¹⁹

The use of a mixture of small and large inserts gains most of the advantages that would occur with the sole use of large inserts (Figure 2.2 and Figure 2.3). This is true even when the large inserts represent a relatively small fraction of the total. Generally speaking, the total length of all the inserts should be chosen to maximize the mapping redundancy R_m . If for technical reasons an insert library is constructed of a single intermediate size, a slightly higher sequence redundancy can be used to ensure completeness. The exact balance between redundancy and insert lengths will depend on the laboratory and should be determined on a case-by-case basis with the aid of computer simulations.

Double-barrel shotgun sequencing has many advantages over traditional shotgun sequencing. Notably the mapping redundancy R_m for single-barrel sequencing is $2nL=R_s$. The mapping redundancy for double-barrel sequencing is nI , which should be several times greater than R_s . This creates a high-redundancy mapping situation which permits efficient low-redundancy sequencing. Pairwise strategies are not confined to low-pass sequencing and are equally valuable at high redundancies, particularly for sequence assembly. For these reasons, I feel that all random strategies should employ pairwise data, at least with the goal of generating complete scaffolds as a basis for further sequencing. Such sequencing can either continue to be random or switch to directed approaches.

I expect that most projects will move to directed sequencing after a complete scaffold is obtained. This “gap closure” phase will entail obtaining sequence for SMGs as

¹⁹ A detailed discussion is beyond the scope of the present work. As a general rule, a repeat cannot be properly analyzed if it is longer than the size of the mapping fragment, or smaller than the uncertainty in the position of the markers in the map. The positional uncertainty of finished sequence is very close to zero, so repeats can be efficiently detected if only if they are shorter than the mapping fragment size. For traditional shotgun sequencing this is L , but for double-barrel sequencing it is I .

well as reverse sequence from regions of single-strand coverage. The templates localized during scaffold construction are ideal substrates for such directed sequencing. One gap closure methodology is to sequence a PCR product spanning the gap. If the entire target sequence is not needed, only gaps of interest need be filled. For example, the entire target may not be needed once a gene is localized in a gene-finding effort. Likewise, if a gap of known size is clearly bounded by the 5' and 3' ends of a known element, such as an Alu repeat, the gap need not necessarily be sequenced.

For any reasonably large project, computational tools will be necessary to assemble and analyze scaffolds. I expect that such software will evolve in the future, as the advantages of pairwise assembly drive the market for assembly software. On the other hand, without assembly software many of the advantages of pairwise sequencing are tempered. Thus pairwise sequencing will not become a universal tool until such a time as good software becomes available to the community at large. Currently no available software tools use pairwise data to aid the assembly algorithm, although several are capable of displaying pairwise data or using it to verify accuracy.

Building a first generation software tool should be straightforward. One simple assembly algorithm is a four step process. First, assemble individual sequences into islands, blind to their pairwise nature. Second, order the resulting sequence islands by linking together sequences with their mates from opposite ends of the inserts. Third, check for inconsistencies, remove suspect pairs of sequences, and iterate the process. Finally, make rough estimates of gap distances based on insert lengths and on low quality ends of sequence reads. This algorithm was successfully employed by hand to assemble the cosmid A1-4, which I believe to have been a robust test of its efficiency. Improvements on such an algorithm can be made, but even an implementation as straightforward as this would be tremendously useful to workers in the field. The finished sequence from pairwise algorithms is more robust and accurate than that of traditional algorithms. Each paired sequence offers a positional check on its mate, allowing a majority of misplaced sequences to be immediately located following an assembly. For example, without this check I would have misplaced several sequences during my raw data simulation of the A1-4 assembly.

One concern occasionally raised to pairwise sequencing is that the exact lengths of the pairwise inserts is uncertain. Such uncertainty arises because fragments are typically

band purified on an agarose gel and not subsequently characterized.²⁰ In extreme cases, particularly at low redundancies, such uncertainty might result in indeterminate island order within a scaffold.²¹ However, in my raw data simulation of cosmid A1-4, I found that a redundancy R_s of 2.25 was more than enough to avoid any such problems. Thus, a knowledge of exact insert lengths would have contributed little to this project. This will be generally true in practice, for all SMGs are highly likely to be less than a sequence read in length. Therefore, rather than measuring their size, one should simply sequence across them. In the rare event that the gap was not closed after one or two iterations of sequence walking, the lengths of the fragments could at that point be determined.

Many modifications to the basic pairwise strategy have been proposed. For example, Burland et al. (1993) suggest sequencing only one end of inserts initially, and then only sequencing opposite of ends of clones that are likely to produce new information, such as those ends from inserts that extend beyond contig edges. This strategy makes sense if the cost of sequencing an opposite clone end is considerably greater than the cost of sequencing an initial end, as in the Janus strategy. It can also, however, alter the independent and random accumulation of sequence information. For example, one can intentionally avoid sequencing a fragment that lies in region already covered to a high depth in previous sequence reads. However, this comes at a high cost in clone isolation and clone-tracking administration. Additionally, an increased use of orphan sequences will prolong the achievement of a complete scaffold. I recommend sequencing both ends of all inserts, at least until complete scaffolds are obtained.

Another modification to the basic pairwise strategy is to halt a project before it reaches the complete scaffold stage. This is the approach of OSS.²² An extreme example is presented by Smith et al. (1994), in which cosmid clones are entirely mapped before their

²⁰ It is possible to more accurately measure the lengths of fragments, but this involves extra effort. Conveniently, such effort is not needed as the length information obtained is redundant with information which will be gained during the sequence finishing phase. In specific cases, if deemed worthwhile, one could measure the length of a fragment retrospectively.

²¹ Such a scaffold would fail to meet the strict definition of a scaffold, which requires that all islands of a scaffold be ordered and oriented.

²² It is also the approach of Venter et al. (1996), but in the Venter case, the BAC end sequences are not used for sequence finishing, so there is a driving incentive to halt the pairwise sequencing early.

ends are sequenced. I see no advantage in halting pairwise projects before complete scaffolds are achieved. The extra sequence redundancy necessary to achieve complete scaffolds is relatively small compared with the labor that is otherwise necessary to assemble unlinked scaffolds into a complete map.

Pairwise strategies can effectively handle megabase targets. My simulations demonstrate that sequence redundancies between twofold and threefold are more than adequate to span such targets with complete scaffolds (data not shown). By permitting direct shotgun sequencing, double-barrel strategies eliminate the need to use intermediate subclones of large mapping vectors such as BACs or YACs. This elimination of cosmid subcloning and mapping can represent a significant increase in the efficiency of genomic sequencing efforts. I particularly recommend double-barrel shotgun sequencing for small bacterial and viral genomes. Pairwise strategies were used extensively during the sequencing of the *Haemophilus influenza* genome, the first genome ever to be completely sequenced (Fleischmann et al., 1995). Pairwise strategies were again used for the *Mycoplasma genitalium* genome (Fraser et al., 1995). The speed with which these genomes were sequenced stunned the genomics community.²³

Low redundancy pairwise strategies are particularly useful for gene finding, as they provide most of the sequence data from a target region, which can then be utilized in similarity or feature identification searches. High accuracy sequence is not needed, nor is complete coverage. Regions of interest can be singled out for subsequent special attention facilitated by the structured nature of the scaffold.

To summarize: pairwise end sequencing can be characterized as mapping at high redundancy, but sequencing at low redundancy. It generates complete scaffolds more economically and more quickly than traditional shotgun sequencing. The advantages of this strategy include its simplicity and the absence of any need for clone mapping other than that which results as an incidental by-product of sequencing. It is capable of handling relatively large repeats or complex templates. Its utility includes STS generation, gene finding, low- and high-pass sequencing, and ultra-fine-scale mapping.

²³ By contrast, the *E. coli* genome project did not use pairwise data and took years (rather than months) to complete (Blattner et al., 1997). To be fair, other factors influenced the slow rate of *E. coli* sequencing.

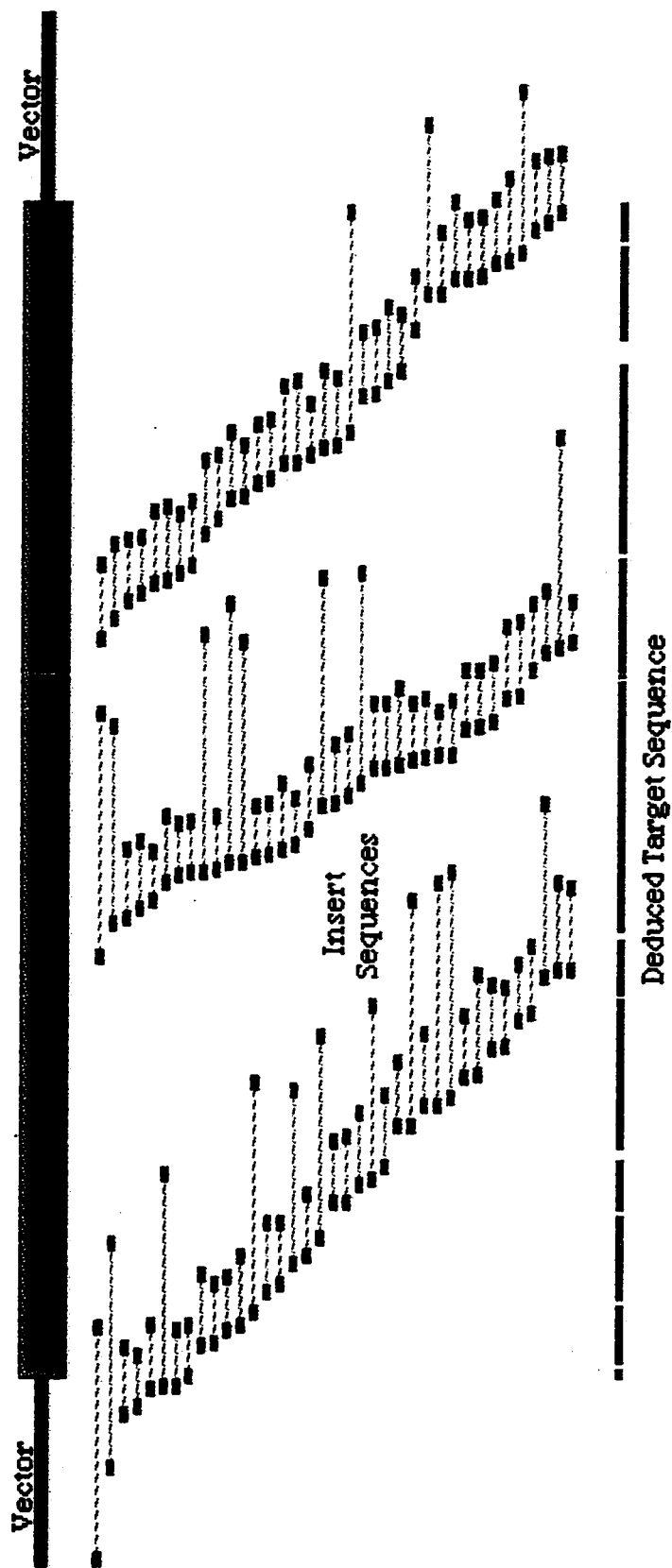


Figure 2.1. A model "double-barrel shotgun" assembly. A 2.25 sequence redundancy produces eighteen contigs which span ninety percent of an original target cosmid at 99.9% accuracy. Contig orientation and order are determined as shown. All but one gap are less than 400 bp; the remaining is 751 bp. More statistics are presented in Table 2.1.

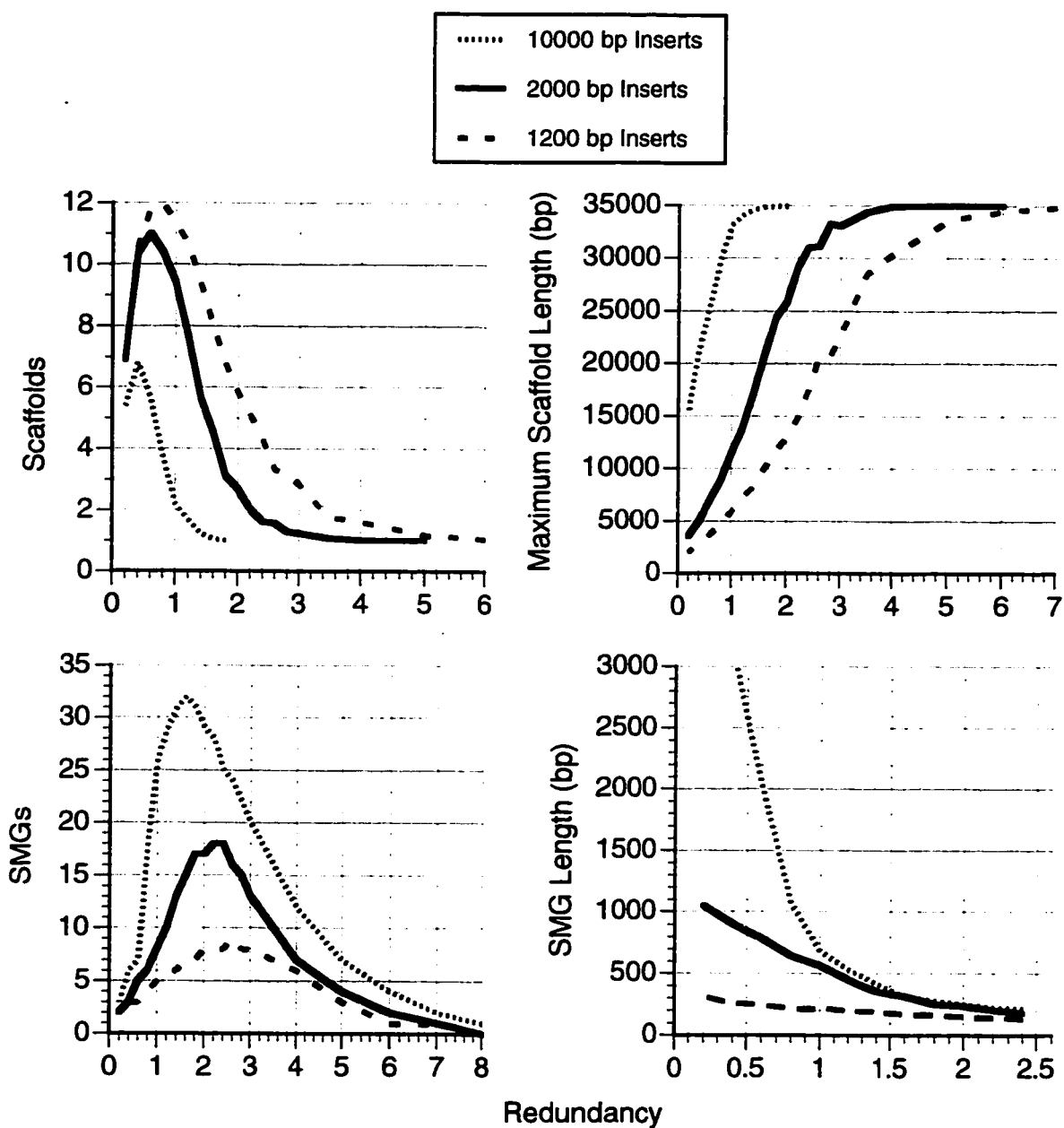


Figure 2.2. Parameters from a 35 kb pairwise project evaluated as a function of sequence redundancy. ($L=400$; $T=30$)

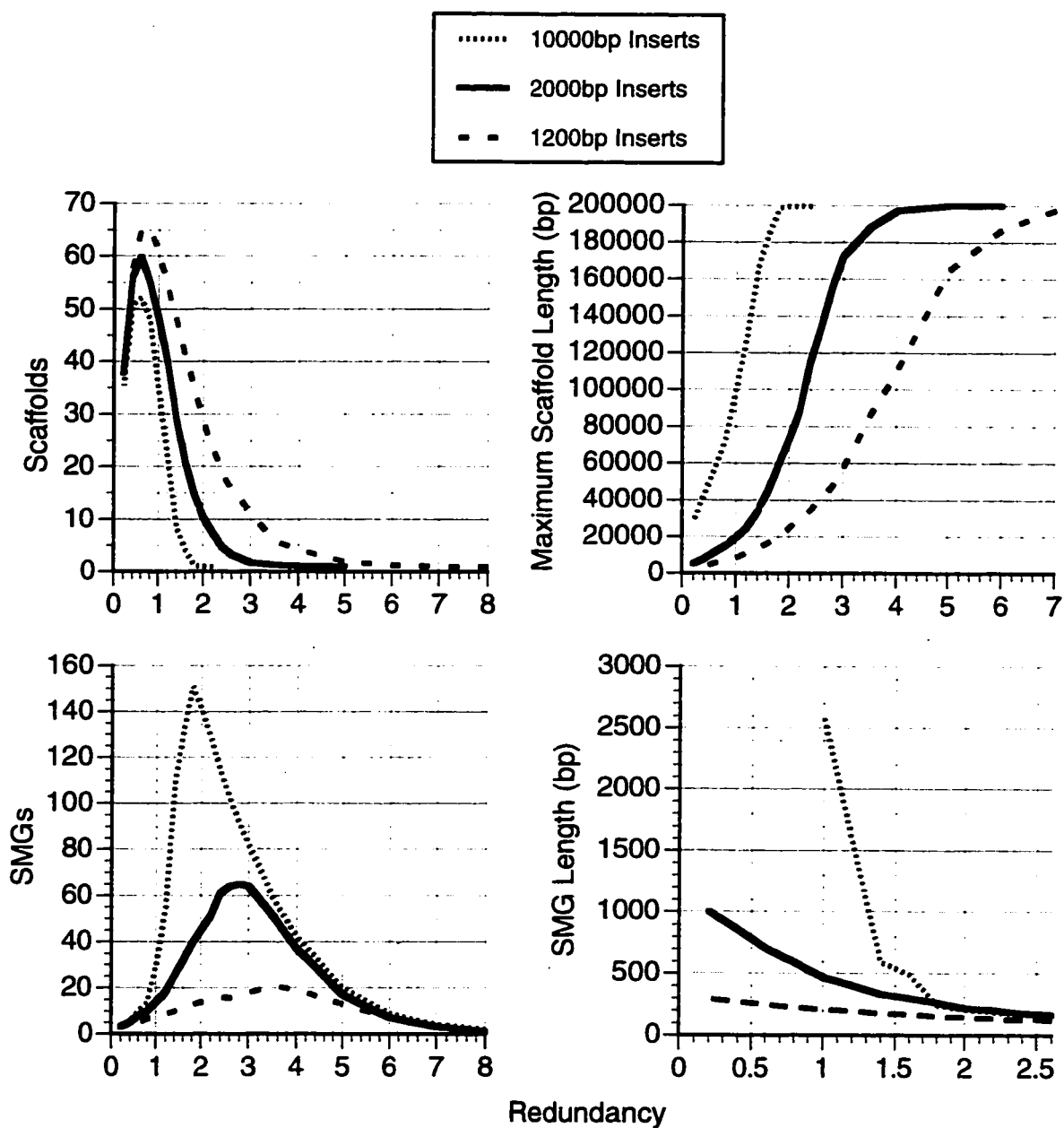


Figure 2.3. Parameters from a 200 kb pairwise project evaluated as a function of sequence redundancy. ($L=400$; $T=30$)

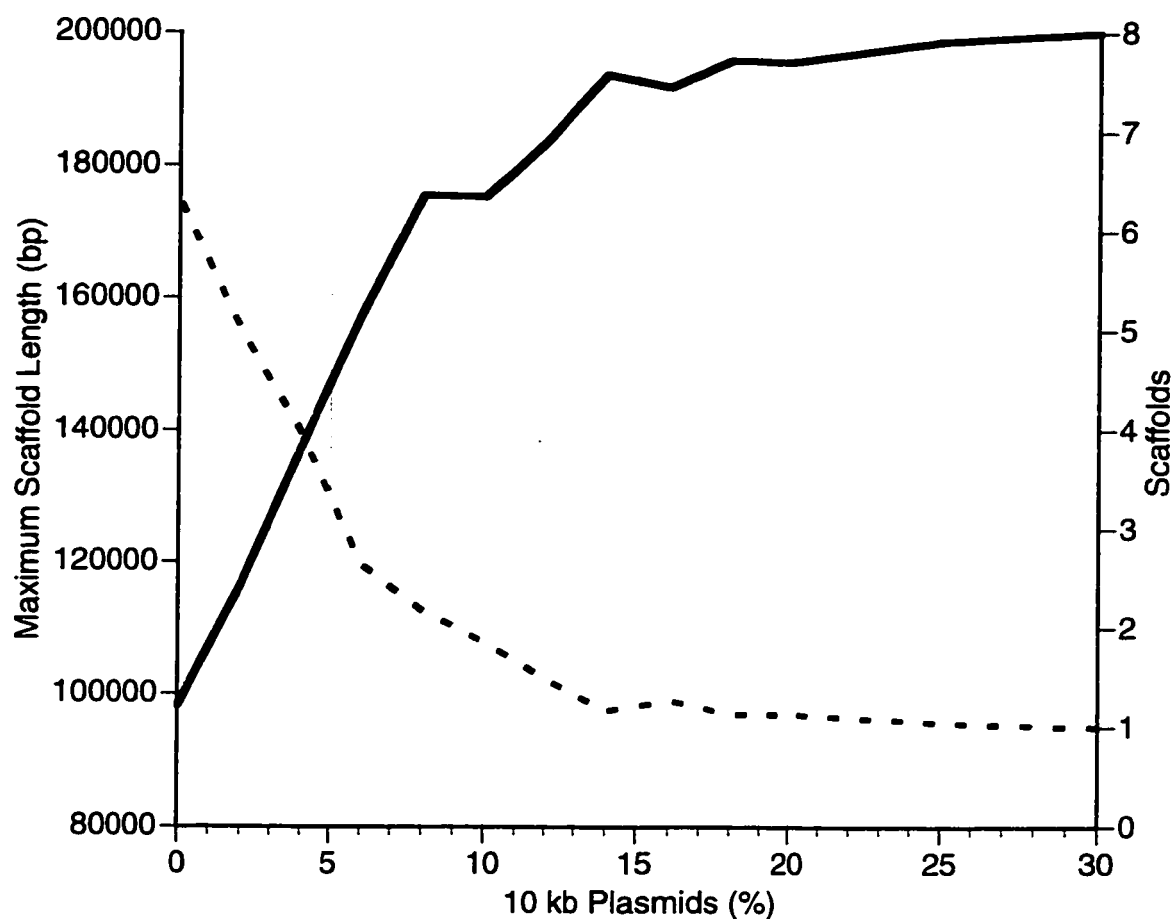


Figure 2.4. Pairwise strategies employing a mix of insert sizes were simulated. Here, a mix of 2000 bp and 10000 bp inserts were simulated at a constant sequence redundancy (2.25). As seen, a small proportion of larger inserts produces results comparable to those achieved when only large inserts are used. At 2.25 redundancy, complete scaffolds can be obtained with only a 15% mix of longer insert lengths. A 200 kb target was assumed ($L=400$; $T=30$). Maximum scaffold length, solid line; Scaffolds, dashed line.

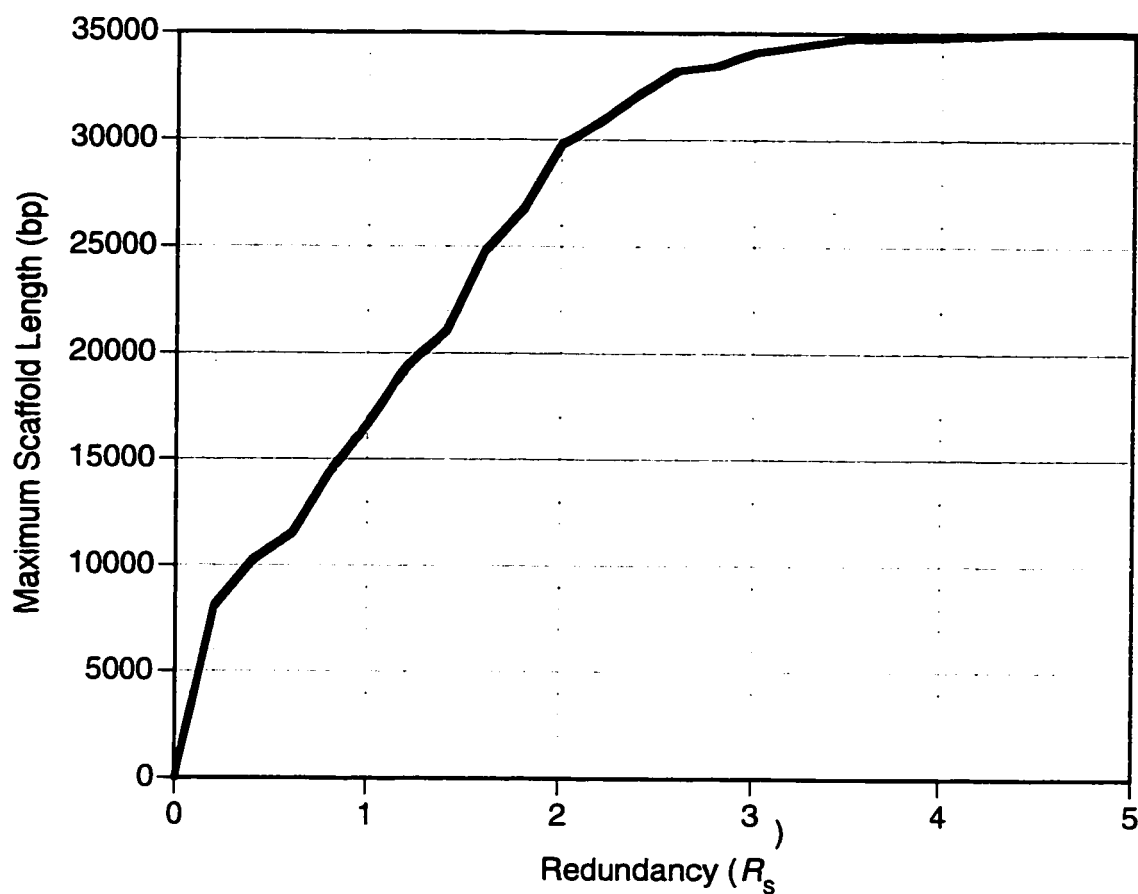


Figure 2.5. A hybrid strategy employing a combination of single strand and pairwise data was simulated. A mix of 60% single strand, 30% 2000 bp, and 10% 10000 bp data was employed. With this approach a complete scaffold can be obtained at less than threefold redundancy. A 35 kb target was assumed. ($L=400$; $T=30$)

Table 2.1. Results from a raw data simulation of a pairwise strategy employed on a 35343 bp cosmid which had previously been shotgun sequenced to ninefold redundancy. For this simulation, 2.25 sequence redundancy was derived from the end sequences of a mixture of hypothetical subclone inserts, 20% approximately 7000 bp in length, and 80% approximately 2500 bp in length. These results agree with results from my computer simulations and suggest the practicality of obtaining complete scaffolds in the range of twofold redundancy. The computer simulation column depicts average values from 100 independent determinations.

	<u>Computer Simulation</u>	<u>Raw Data Simulation</u>
Number of Scaffolds	1.02	1
Scaffold Length (bp)	35267	35343
Number of SMGs	21	17
Average SMG Length (bp)	169	223
% Target Covered	90.1	89.2

CHAPTER 3. VERTEBRATE TRYPSINOGEN EVOLUTION

"He [...] saw them subject to chances, the complications of existence, and saw them vividly, but then had to find for them the right relations, those that would bring them out."

Henry James

The motivation for studying evolution is circular: one studies evolution to understand genomes; one studies genomes to understand evolution. Each axis of study fuels the other, and both are goals in their own right. An attempt to distinguish the two fields of inquiry might employ the argument that the study of evolution fulfills a philosophical need to quest for origins, while the study of genomes fulfills a practical need to understand and manipulate biological systems. But any attempt to make such a distinction must ultimately fail, one cannot study one without the other.

I approached the study of the molecular evolution of the vertebrate trypsinogens with several interrelated aims. First, I wished to document the natural history, or phylogeny, of the trypsinogens. Secondly, I wished to elucidate the mechanisms of evolution that produced this phylogeny. Thirdly, I wished to produce cloning reagents that could be used to study syntenic relationships. Lastly, I wished to gain insight into possible functions of trypsinogen beyond its established role in digestion. An unexpected dividend of my studies was the identification of a novel gene recently evolved from one of the trypsinogen genes.

The basic elements of evolution are random variation and natural selection. Understanding evolution involves understanding the details of random variation of DNA and the natural selection acting on gene products. Genome analysis through genomic sequencing provides a major axis for understanding evolution. Sequencing provides direct observation of the endpoints of evolutionary processes. These endpoints are DNA sequences either within genes, or between them. The examination of many evolutionary endpoints is the primary approach to acquiring data on evolutionary processes.

The comparison of two or more sequences is the basis for such evolutionary analysis. Comparison reveals similarities which can have arisen through one of three mechanisms: random chance, sharing a common ancestor, or sharing a common selective pressure. Consideration of sequences in a phylogeny can reveal information about these three processes. Likewise, understanding these three processes can improve efforts to reconstruct phylogenies.

My efforts in this chapter will focus first on reconstructing the phylogeny of the vertebrate trypsinogens and on the complications that prevent the reconstruction of this phylogeny with high confidence. I will then employ my partial reconstruction of the trypsinogen phylogeny to draw conclusions about the nature of the evolutionary mechanisms that have operated on the trypsinogens. The most important of these conclusions is that the vertebrate trypsinogens have been subject to coincidental evolution.

3.1 COINCIDENTAL EVOLUTION

Multigene families do not necessarily demonstrate expected phylogenetic relationships. Imagine a multigene family that consists of two genes, YFG1 and YFG2. This gene family originated from the duplication of a single original ancestral YFG gene. Consider a phylogeny in which YFG1 and YFG2 are present in the genome of an ancestral species that undergoes a speciation event to produce two descendant species, such as *Mus musculus* and *Xenopus laevis*. One would expect the *Mus* YFG1 to be more similar to the *Xenopus* YFG1 gene than to the *Mus* YFG2 gene. This is because the two YFG1 genes share a more recent common ancestor than either one shares with a YFG2 gene. However, it is commonly observed that, in this type of situation, the *Mus* YFG1 gene more resembles the *Mus* YFG2 gene than it does the *Xenopus* YFG1 gene. This phenomenon is called coincidental evolution.

Coincidental evolution was first observed in the mid-1960's with the application of DNA reannealing techniques to the repetitive sequences present in eukaryotic genomes. Edelman and Gally (1970) noticed that repetitious strands from one species more readily reannealed to one another than they did to homologous strands from another species. A similar phenomenon was noted by Brown et al. (1972) while studying the rRNA intergenic spacer regions of two species of *Xenopus*. Many other examples have since been noted; Li (1997) provides a review.

The history of the terminology of coincidental evolution is tangled. The first general term describing this process was “coincidental evolution” (Hood et al., 1975). A later term, “concerted evolution,” due to Zimmer et al. (1980) has gained popularity in the literature, but will not be used here, as the word “concerted” connotes teleological conspiracy, which is not desirable.¹ The earlier terms, “horizontal evolution” (Brown et al., 1972) and “coevolution” (Edelman and Gally, 1970), refer specifically to direct transfer of genetic information, and not to the plethora of mechanisms that can account for the phenomenon of coincidental evolution. “Coevolution” also carries the connotation of a non-random driving force, which is a more specific connotation than desired. “Coincidental” carries a connotation of randomness, which is also a more specific connotation than desired, but in an opposite manner than “co-” or “concerted.” Although not perfect in connotations, “coincidental evolution” does have historical precedence as the first term defined with enough generality to cover all possible mechanisms and manifestations of the phenomenon. If an opportunity again arises to revise the nomenclature, perhaps the term “covariant evolution” will be proposed and accepted as having the desired meaning and connotations.

Several mechanisms can account for coincidental evolution. Horizontal transfer of genetic information has a sudden and dramatic effect on the similarity of two genes. Horizontal transfer can be accomplished by unequal crossing over, gene conversion, or possibly other mechanisms. Common selective pressures on function can produce a slow but inexorable convergence of similarity. A similar effect can be produced by a common bias in the mechanism of random variation, such as a nucleotide usage bias, or exposure to an environmental mutagen. However, it is unlikely that biases in random variation play a detectable role in vertebrate coincidental evolution. An overview of the mechanisms of coincidental evolution is provided by Li (1997).

3.2 HISTORICAL PERSPECTIVE ON TRYPSINOGEN

Trypsin from the bovine pancreas was one of the first proteases isolated with sufficient purity and enough quantity for precise biochemical studies (Northrup et al., 1948). This bovine trypsin was one of the first proteins to be sequenced (Walsh and Wilcox, 1970). Three-dimensional trypsin and trypsinogen protein structures were early

¹ Webster’s Dictionary defines “concerted” with the phrase “mutually contrived,” which has led to the following tongue-in-cheek terminology proposal: “contrived evolution.”

conquests of X-ray crystallography (Sweet et al., 1974; Kossiakoff et al., 1977). Thus, over a period of several decades, the details of the sequence, structure, and mechanism of action of trypsin were worked out (reviewed by Male et al., 1995). The early availability of data on trypsin and other serine proteases helped fuel the birth and development of the field of molecular evolution.

Because trypsin was of early interest to molecular biology, many trypsin protein sequences were obtained before it became easier to sequence DNA. Now that DNA sequencing is far cheaper than protein sequencing, new molecular sequences determined are DNA sequences. DNA sequences contain more information than primary protein sequences. This is due to the degeneracy of the genetic code that transfer RNAs employ during mRNA translation.² In some of my analyses, I have used algorithms that operate on DNA sequences, but for the most part I have employed algorithms that operate on protein sequences. This has allowed me to utilize all available data on trypsinogen, including the earliest protein sequences determined.

Often, interest in a particular gene is driven by interest in a disease caused by a defect in that gene. However, trypsinogen research has seldom been driven by such an interest. The only known disease to result from a defect in a trypsinogen gene is hereditary pancreatitis. The relationship between trypsinogen and hereditary pancreatitis was only recently discovered (Whitcomb et al., 1996). Hereditary pancreatitis affects only a few thousand people worldwide, but is very likely to be related to trypsinogen.

A clinical condition ascribed to a defect in trypsin has also been described: trypsinogen deficiency disease (TDD). There have been only six reported cases of TDD, none of them more recent than 1967 (Farber et al., 1943; Townes, 1972). Due to the multicopy nature and multi-chromosomal positioning of the trypsinogen genes, it seems unlikely that TDD is due to a molecular defect in a trypsinogen gene locus, but rather to another defect, such as an aberrant enterokinase gene. This cannot be verified unless another case of TDD is discovered.

² Currently, few algorithms that operate on DNA sequences fully exploit the information present in inferred coding regions (but see, for example, Zhang et al., 1997). This has led some to imply that primary protein sequence has utility beyond that of the corresponding DNA sequence. This is an incorrect implication, but it does serve to point out inadequacies of current DNA algorithms.

The trypsinogen genes and proteins, now and historically, have been studied primarily due to interest in trypsin as a model for protein structure and function. For example, one type of effort building from the trypsin knowledge base has been the design of proteins with novel functions (e.g., Corey and Craik, 1993). Trypsinogen genes have recently received renewed attention from a genomics viewpoint, following the serendipitous discovery of trypsinogen genes within the human T-cell receptor (TCR) β locus. This discovery was the result of an early large-scale genomic sequencing effort specifically directed at the TCR β locus (Rowen et al., 1996).

The topology of the genomic organization of the TCR β locus in humans, mice, and chickens is shown in Figure 3.1. As can be seen, the syntenic relationship of the trypsinogen genes and the TCR β locus is maintained in the mouse and chicken genomes (Lee Rowen, Genbank AE000522; Kai Wang, personal communication). In each of these three species the organization of the trypsinogen genes varies. These variations in organization highlight the dynamic nature of their evolution.

In all three cases, there is a physical separation of two groups of trypsinogen genes: one towards the 3' end of the TCR β locus, and one towards the 5' end of the TCR β locus (Figure 3.1). This separation defines the nomenclature for trypsinogen grouping, first introduced in Roach et al. (1997). Group I trypsinogens are those found 3' in the TCR β locus, and group II trypsinogens are those found 5' in the TCR β locus. This definition is supported by a logical grouping of trypsinogens from sequence distance and other considerations, and is discussed at length in later sections of this chapter.

As a rule of thumb, group II trypsinogens appear 5' to TCR V β gene segments, while group I trypsinogens appear 3' to TCR V β gene segments. In the human there are no functional group II trypsinogens, but there are two trypsinogen relics (T1 and T2) and a pseudogene (T3), all of which are immediately 5' to the entire TCR β locus.³ These are all derived from group II trypsinogens. The interval between V β 29S1 and D β 1 in humans contains 3 functional trypsinogen genes (T4, T6, and T8) and two trypsinogen pseudogenes (T5 and T7). All of these trypsinogens are group I. The 3' end of the human TCR β locus is duplicated on chromosome 9, where a fourth functional group I trypsinogen (T9) is found 3' to the orphon V β 29S2 gene segment.

³ Human T1 is not a true relic; see Section 3.21.

A similar organizational grouping of the trypsinogens is found in the mouse. However, the mouse has several functional group II trypsinogens. Two pseudogenes, two relics, and three functional group II genes lie 5' to the mouse TCR β locus (T1-T7).⁴ Five relics and eight functional group I trypsinogens lie between mouse V β 18S1 and D β 1 (T8-T20).

The organization of the trypsinogens in the chicken differs somewhat. There are two families of V gene segments in the chicken TCR β locus: V β 1 and V β 2. All the V β 1 segments appear 5' to the V β 2 segments. Each V segment is tandemly linked to a trypsinogen gene. There are three known chicken group I trypsinogens, in tandem with the three known V β 2 segments, with the last trypsinogen located between the most 3' V β 2 segment and D β 1. There are approximately 6 group II trypsinogens and V β 1 segments, tandemly linked with opposite orientations. Characterization of the chicken TCR β locus is not complete (Kai Wang, personal communication).

Linkage of the TCR β locus to trypsinogen genes has not been established in other organisms, but I postulate that the two will always occur together. The synteny of trypsinogen and TCR β may be due to an important functional synergism of the two loci or merely due to random association. If a random rearrangement was originally responsible for the syteny of the two loci, then there may be a mechanism that prevents successful chromosomal rearrangements that split this synteny. With trypsinogen genes internal to the TCR β locus, it may be that splitting rearrangements destroy the TCR β locus, and significantly decrease the fitness of the resulting mutant. Alternatively, there may be no such constraining selective force. In this case, the two loci remain syntenic merely because a random event has not separated them. I will return to this discussion in Section 3.18.

The tandemly repeated nature of trypsinogens has been observed in other organisms. Three repeated trypsinogens are present in the pufferfish genome, two in tandem and one with opposite orientation (Kai Wang, personal communication). Two or more repeated trypsinogens are linked in tandem in the lamprey genome. The *Drosophila* genome has four tandem trypsinogen genes in alternating orientations, but presumably not linked to a TCR locus (Davis, 1985), as the immune receptor loci are hypothesized to have first evolved in the chordate lineage.

⁴ Mouse T1 is not a true relic; see Section 3.21.

Intrigued by the evolutionarily conserved genomic organization of the trypsinogens, I sought to expand my knowledge of trypsinogen gene sequences and their modes of evolution. Already, a tremendous amount of information was available. Over the past decades, a large number of protein, cDNA, and genomic DNA sequences have been determined for either trypsin or trypsinogen from a variety of vertebrate species. In many cases, several different sequences representing different isozymes had been obtained from the same species.

In an effort to acquire molecular data from all vertebrate classes and to increase representation from within some classes, I obtained additional cDNA and genomic trypsinogen sequences from the lamprey *Petromyzon marinus*, while my colleague Kai Wang obtained sequences from the pufferfish *Takifugu rubripes* and the frog *Xenopus laevis* (Roach et al., 1997). These sequences allowed me to study the gross outlines of trypsinogen evolution across the entire vertebrate phylogeny.⁵ These sequences all arose from a common set of trypsinogen sequences present in the ancestral vertebrate species that lived approximately 600 million years ago.

To give focus to a phylogenetic study, it is helpful to study at least one sequence that is equally distant from all of the other sequences in the study, but not so distant that it has lost most or all of its relatedness. Such a sequence is referred to as a phylogenetic outgroup. No vertebrate sequence could serve this purpose. Therefore, in an effort to obtain an outgroup, I sequenced a trypsinogen cDNA from a urochordate, the tunicate *Boltenia villosa*. The urochordates, together with the hemichordates (acorn worms), the cephalochordates (amphioxus), and the vertebrates, form the phylum Chordata. Recently, another laboratory has sequenced two trypsinogens from the urochordate *Botryllus schlosseri* (Pancer et al., 1996).

3.3 THE COMPILATION OF TRYPSIN AND TRYPSINOGEN SEQUENCES

Previously published trypsin and trypsinogen sequences were culled from Genbank and SwissProt using a variety of text and homology based searches (Gish and States, 1993; Altschul et al., 1990). The homology searches were used to rule out the possibility

⁵ Many years from now, when complete trypsinogen sequences have been determined from several hundred vertebrate species, a detailed phylogeny may replace the gross outlines described in this paper.

that a trypsin or trypsinogen sequence might be present in a database under a different name. However, no such “misabeled” vertebrate sequences were identified.

Most known vertebrate trypsin and trypsinogen sequences are described in the literature. A table of many previously published sequences with original references can be found in Rypniewski et al. (1994). An expanded listing of literature references of all known chordate trypsinogens is presented in Table 3.1; a summary of additional data for these trypsinogens is presented in Table 3.2.

Two vertebrate trypsinogen cDNA sequences present in Genbank was not included in any of my analyses: those of *Pleuronectes platessa* and *Dissosthicus mawsoni*.⁶ The *Pleuronectes* sequence was a direct submission to the database and has never been documented in any publication. It fails to meet basic criteria for inclusion in the trypsinogen gene family. The *Pleuronectes* active site sequence GSRDACNGD differs from the almost-absolutely conserved trypsin consensus of GGDSCQGD. The glutamine to asparagine alteration occurs at one of four key “pocket specificity” residues which are absolutely conserved in all trypsins (Hedstrom et al., 1992). The specificity pocket more resembles that of granzyme A more than any other serine protease pocket. The carboxy-terminus of the *Pleuronectes* sequence is neither the correct length nor homologous to the other trypsin carboxy-termini. The *Pleuronectes* sequence contains an insertion near residue 150 as well as several other discrepancies that are difficult to reconcile with inclusion in the trypsin family. These discrepancies are clearly due to mistaken gene identification. Inclusion of this rogue sequence would have distorted the phylogenies that I derive for the trypsinogens. This distortion is similar to the effect produced by inclusion of a too-distantly related outgroup (see Section 3.17). The *Dissosthicus* sequence was originally isolated by PCR and was most likely identified as a trypsinogen based on its resemblance to the *Pleuronectes* sequence (Chen et al., 1997).

3.4 CLONING AND SEQUENCING *PETROMYZON MARINUS* TRYPSINOGEN

Poly(A)-mRNA was prepared from the dissected gut of a *Petromyzon marinus* ammocoete (a gift of James Seeley, Hammond Bay Biological Station, MI). The mRNA was reverse transcribed and cloned as cDNA into the λ -ZAP directional cloning vector

⁶ Genbank X56744 and U58835.

(Stratagene). Additionally, RT-PCR was performed on the poly(A)-mRNA with the trypsin specific primers TRYF and TRYR. The PCR primers used in this study are tabulated in Table 3.3. The sequence of the 387 bp TRYF-TRYR PCR product was consistent with trypsin and was used to probe the *Petromyzon* gut cDNA library. Thirty-one positive plaques were sequenced at their 5' ends with the m13 reverse primer. Several of these plaques hybridized weakly to the probe, and were picked due to the possibility that the probe might have hybridized to something unexpected but interesting.

Of these 31 plaques picked for sequencing, 17 were trypsinogen, 4 were chymotrypsinogen, one was similar to the chitotriosidase precursor cDNA, one was similar to the oligosaccharyl transferase STT3 subunit, and the other eight cDNAs were not positively identified. Seven of the cDNAs identified by 5' end sequencing as trypsinogen were completely sequenced by primer walking on both strands with the following primers: LT2, LT3R, LT3, LT4R, LT5R, XTA5, and TRYR.⁷ Based on contig assemblies, the seven completely sequenced and ten partially sequenced lamprey trypsinogen cDNAs fell into at least five clusters, indicating the presence of at least five different expressed lamprey trypsinogen isozyme genes (or alleles).

3.5 LAMPREY TRYPSINOGENS

The lamprey trypsinogen cDNA sequences contain all of the important sequence features expected of a trypsinogen. They possess overwhelming similarity to the known trypsinogens. In particular, they possess the three absolutely conserved cystine bridges present in all serine proteases, the four key trypsin “pocket specificity” residues, the three residues of the catalytic triad, a signal peptide (described in Section 3.8), an activation peptide (described in Section 3.9), a stabilizing trypsin amino-terminus, and a conserved calcium binding site. The conserved residues that are characteristic of the vertebrate trypsinogens are detailed Section 3.14. The overwhelming similarity of the lamprey trypsinogens to all other vertebrate trypsinogens, coupled with the origin of the cDNA library from the gut, permitted the lamprey sequences to be classified as trypsinogen with certainty.

⁷ Genbank AF011352 and AF011898-AF011901.

One significant exception to the preponderance of similarity of the lamprey trypsinogens to all other trypsinogens is that the lamprey trypsinogen activation peptides all end with a histidine residue. This novel property is discussed in Section 3.9.

I designated the five lamprey trypsinogen clusters: A1, with nine cDNAs (two completely sequenced); A2, with three cDNAs (one completely sequenced); A3, with one completely sequenced cDNA; B1, with three cDNAs (two completely sequenced); and B2, with one completely sequenced cDNA. The untranslated 3' tails of the three A-cluster trypsinogens are 92.0-96.0% identical. The untranslated 3' tails of B1 and B2 are 85.9% identical. The A- and B-cluster tails could not be aligned with each other. The coding regions of the A-cluster sequences are 98.9-99.7% identical at the nucleotide level. The coding regions of B1 and B2 are 97.7% identical at the nucleotide level. The nucleotide identity between the coding sequences of the A and B clusters is 92.5-93.2%.

Different lamprey trypsinogen genes (or alleles) can clearly be over 99% identical across regions longer than a single sequence read. This high similarity of multiple trypsinogen genes is observed in other species (Wang et al., 1995). The lamprey trypsinogen genes are probably encoded by highly similar tandem repeats, as is the case in humans, mice, and chickens. Two of the lamprey B cluster trypsinogens have been observed to be linked in tandem following double-barrel shotgun sequencing of a genomic cosmid clone (data not shown).

The lampreys diverged from the other members of the vertebrate subphylum early in vertebrate evolutionary history, so the presence of tandemly repeated trypsinogens in lampreys is strongly suggestive that this general organization of the trypsinogens has been maintained throughout vertebrate history. The presence of highly similar repeats is known to facilitate mechanisms of horizontal transfer of genetic information, such as unequal crossing over and gene conversion (Li, 1997). Therefore the vertebrate trypsinogens have been prime candidates for horizontal gene transfer throughout their history. Such horizontal transfer would result in coincidental evolution (see Section 3.1 and Section 3.17).

3.6 CLONING AND SEQUENCING *BOLTENIA VILLOSA* TRYPSINOGEN

Poly(A)-mRNA was prepared from the dissected gut of a specimen of *Boltenia villosa* (a gift of William Moody, University of Washington, WA). The mRNA was

reverse transcribed and cloned as cDNA into the λ -ZAP directional cloning vector (Stratagene). Additionally, RT-PCR was performed on the poly(A)-mRNA with degenerate primers designed to amplify serine proteases: H and S. The PCR primers used in this study are tabulated in Table 3.3. The H and S primers correspond to conserved sequences of the serine protease active site. Similar primers are described by Kang *et al.* (1992) and Wiegand *et al.* (1993). A resulting H-S PCR product of approximately 350 bp was agarose-gel isolated and used as a probe to screen the cDNA library. This band failed to sequence due to polyclonality and was judged to be a diverse mixture of serine-protease-derived products. Twenty-three positive plaques were picked and sequenced. Seven cDNAs identified by 5' end sequencing as trypsinogen were completely sequenced by primer walking on both strands with the following primers: TUN2F1, TUN2R1, TUN2F2, TUN2R2, TUN19F1, TUN19R1, and TUN2R3. All seven of the tunicate trypsinogen cDNAs I sequenced appeared to represent the same allele, as I could distinguish no sequence variation between them.⁸ Of the other sequences, five were chymotrypsinogen, two were ribosomal proteins, one was actin, one was glutathione S-transferase, one was from the mitochondrial 16S RNA, and six were not positively identified. The chymotrypsinogen sequences were identified based on the presence of a methionine at the position that is 192 according to the bovine chymotrypsinogen numbering system, as well as overall similarity.

3.7 TUNICATE TRYPSINOGEN

The *Boltenia* trypsinogen sequence was identified based on its presence in a gut derived cDNA library and its sequence similarity to the other chordate trypsinogens. The *Boltenia* trypsinogen cDNA sequence contains all but one of the important sequence features expected of a trypsinogen. These were mentioned in Section 3.5 for the lamprey trypsinogens, and are further detailed in Section 3.14.

The *Boltenia* trypsin appears to lack the residues forming the calcium binding site found in all vertebrate trypsins. The two known trypsinogens from the tunicate *Botryllus schlosseri* also appear to lack these residues. Therefore, the tunicate trypsinogens may bind calcium at an alternative site, much as the *Streptomyces griseus* trypsin does (Read and James, 1988). It has been suggested that calcium binding confers stability against thermal

⁸ Genbank AF011897.

or chemical denaturation (Martin, 1984). Additionally, a requirement for calcium may ensure that trypsin is only active extracellularly, since intracellular calcium concentrations are extremely low (Kretsinger, 1976). There is still uncertainty as to the exact location, function, and importance of calcium binding sites in active trypsin (Read and James, 1988; Smalås et al., 1994).

I identified only one trypsinogen isozyme in *Boltenia*. This raises the possibility that trypsinogen is single copy in *Boltenia*, rather than encoded by multiple repeats, as is likely in the other chordates. *Boltenia* feeds continuously, rather than periodically with meals. Therefore, *Boltenia* may have a decreased need for dynamic control of trypsinogen expression. It is conceivable that maintenance of multiple trypsinogen isozyme genes in vertebrates is driven by selective pressure for dynamic control of expression by gene dosage. If this pressure were absent in *Boltenia*, it might explain a single-copy trypsinogen.

However, it must be borne in mind that two different trypsinogens were identified by Pancer et al. (1996) in the tunicate *Botryllus schlosseri*. This would seem to contradict the above scenario for *Botryllus*, suggesting that if *Boltenia* lost multiple trypsinogens, this loss was recent. It is perhaps more likely that *Boltenia* does indeed possess multiple trypsinogens, and merely that I failed to identify the additional isozymes. This could have occurred if only one isozyme was dominantly expressed in the organism that I used to generate my *Boltenia* gut cDNA library. *Boltenia* and *Botryllus* both belong to the order Stolidobranchia of the class Ascidiacea, but to different families within this order. It is unclear when the two families diverged.

It is interesting to note that the *Botryllus* and *Boltenia* trypsins are quite divergent. They share only 37% amino acid identity. They are as identical to the trypsin from the crayfish *Astacus fluviatilis*, at 34%-38% identity, as they are to each other. Considered together, the tunicate and crayfish trypsins possess eleven indels with respect to the vertebrate trypsinogens. Of these eleven, four are shared by *Boltenia* and *Botryllus*, three are shared by *Astacus* and *Boltenia*, three are shared by *Astacus* and *Botryllus*, while at one site they all differ.

There are three possibilities to explain the large divergence of the tunicate trypsins. First, the trypsins from the two species may have different functions, and thus be subject to

different selective pressures. If this is the case, then one should expect that there are additional trypsins to be found in *Boltenia* and *Botryllus*. Secondly, the two families of tunicates may have diverged from each other very early in chordate evolution. Thirdly, there may have been a dramatic increase in the rate of evolution following a recent divergence of the two tunicate families.

3.8 SIGNAL SEQUENCES

Most, if not all, trypsinogens are secreted proteins. The secretion process begins with the translocation of the nascent polypeptide across the membrane of the endoplasmic reticulum. Residues at the amino-terminus of the trypsinogen polypeptide form a signal which is recognized by the signal recognition particle which serves as a chaperone for entry into the endoplasmic reticulum. A review of this process has been provided by Rapoport (1990). The signal sequences are cleaved by a signal peptidase within the endoplasmic reticulum. The resulting protein is properly called a trypsinogen; before cleavage the polypeptide is referred to as a "pretrypsinogen."

An alignment of chordate pretrypsinogen signal sequences and activation peptides is shown in Table 3.4. Only in the case of the canine have the sites been determined experimentally (Carne and Scheele, 1982). All of the other signal peptidase cleavage sites in Table 3.4 were predicted with the program PSORT (Nakai and Kanehisa, 1992), which implements the algorithm of von Heijne (1986).

The chordate pretrypsinogen signal sequences are highly conserved and conform to the general rules for eukaryotic signal sequences (von Heijne, 1985). These rules define a central hydrophobic region bounded by a charged amino-end and a polar carboxy-end. The vertebrate pretrypsinogen signal sequences are 15 or 16 residues in length. The majority are exactly 15 residues long and are therefore easy to align.

The short length of these sequences limit the statistical significance of any conclusions to be drawn from the alignment. Wang et al. (1995) have suggested that slight differences in signal sequences between "anionic" and "cationic" pretrypsinogens might bias targeting towards different cellular locales or influence the rate of secretion. The leucine-isoleucine-leucine sequence present at positions 5-8 of several anionic pretrypsinogens is suggested to fill this role. I feel that it is more likely that this similarity

stems from the common ancestry that these group I pretrypsinogens share. There is no need to invoke differential selection to explain trends observed from the alignment of the pretrypsinogen signal sequences.

Human trypsinogen T9 is expressed in two alternatively spliced forms, originally dubbed trypsinogen III and IV. Wiegand et al. (1993) identified trypsinogen IV by PCR, and this observation has since been confirmed by the addition of two ESTs to Genbank (AA088815 and AA045553). Although trypsinogen IV was originally designated “brain trypsinogen,” it does not appear to be specifically expressed in brain tissue. Trypsinogen IV uses a unique first exon, and so lacks the signal sequence found in all other trypsinogens. An intracellular role has been suggested for its function.

3.9 ACTIVATION PEPTIDES

Before a trypsinogen gains full enzymatic activity it must undergo a proteolytic cleavage that removes its activation peptide (Neurath and Dixon, 1957). Trypsin is a protease that specifically cleaves polypeptides after the basic residues lysine or arginine. Most trypsinogen activation peptides end with a lysine or arginine, so trypsin is capable of catalyzing the activation of trypsinogen. The enzyme enterokinase is also capable of cleaving the trypsinogen activation peptide. Enterokinase has a very high specificity for the highly acidic trypsinogen activation peptide (Maroux et al., 1970). Due to the presence of the acidic residues in the activation peptide, trypsin has a low specificity for this site, but nevertheless a greater specificity for cleavage after the activation peptide than anywhere else in trypsin (Abita et al., 1969). This system of cleavage specificities lays the groundwork for the exquisite regulation of trypsinogen activation within the vertebrate digestive system.

Following activation, the newly freed amino-terminus of the trypsin enzyme tucks itself into an internal pocket of the globular protein. This “isoleucine-valine-glycine-glycine” sequence stabilizes the conformation of the enzyme, raising its catalytic rate constant by several orders of magnitude (Huber and Bode, 1978; Morgan et al., 1972).⁹ This sequence is shown in orange in Figure 3.2. The amino-terminus sequence of *Xenopus*

⁹ The second order rate constant for the reaction of trypsinogen with diisopropylphosphorofluoridate is 0.041 liter mol⁻¹ min⁻¹; the second order rate constant for trypsin is 300 liter mol⁻¹ min⁻¹ (Morgan et al., 1972).

I and the two *Botryllus* trypsins is isoleucine-isoleucine-glycine-glycine, but this difference seems too minor to alter its function.

The activation peptides of the chordate trypsinogens are shown in Table 3.4. The key feature of trypsinogen activation peptides is a cluster of at least three anionic residues preceding a lysine or arginine. However, the lamprey activation peptide has only two penultimate anionic residues, while the tunicate has just one. Many of the Osteichthyes trypsinogens and one of the *Xenopus* trypsinogens have three anionic residues, while the higher vertebrates tend to have four or more such residues, suggesting a progressive increase in selective pressure for such residues during the course of vertebrate evolution.

Strikingly, none of the activation peptides for the lamprey trypsinogens end in a lysine or arginine residue. All lamprey trypsinogen activation peptides end in a histidine. Thus it seems unlikely that lamprey trypsin is capable of autocatalyzing its own activation, as trypsin is not capable of cleaving after a histidine residue. This suggests that lampreys rely exclusively on enterokinase for trypsinogen activation. The life cycle of the lamprey may explain selective pressure for greater control of digestive enzyme activation. For example, adult lampreys will go months to years without eating. The lamprey chymotrypsinogen activation peptide ends in an arginine and so could be activated by trypsin.¹⁰ This would allow lamprey enterokinase to function as a master control switch for digestion, allowing for little or no basal digestive enzyme activation.

3.10 CYSTINE BRIDGES

Acquisition and loss of cystine bridges is a rare evolutionary event and thus a useful phylogenetic marker. It is, however, unclear exactly how useful they are. Each cysteine residue alone is highly conserved, and the added knowledge that a bridge links two does not necessarily provide much additional information to phylogeny inference. Some bridges will be more conserved than other bridges, and the rare gain or loss of such a bridge will be particularly informative, much as would be a change in an active site residue. These issues aside, consideration of cystine bridges will strengthen the already strong case that the

¹⁰ The Genbank accession numbers for the lamprey chymotrypsinogen ESTs are AA618645-AA618648.

human T3 trypsinogen pseudogene is descended from a group II trypsinogen, in contrast to all of the functional human trypsinogens.

Statistical models for evolution, such as those discussed in Appendix B, are necessary for the implementation of many approaches to phylogeny inference, such as maximum likelihood. However, parsimony can be employed without such models, and shines when only rare events are used as the basis for inference. A backbone phylogeny of the serine proteases can be developed by considering the parsimonious addition and loss of cystine bridges (De Haën et al., 1975). Cystine bridges can characterize protein superfamilies. For example, the relatedness of members of a growth factor superfamily have been characterized with the aid of the consideration of cystine bridge topology (Murray-Rust et al., 1993).

For the trypsinogens, assignment of cystine bridges can be made from considerations of the homology of cysteine residues. Each cysteine residue can be assigned to a bridge based on its position in an alignment of the primary amino acid sequences of the serine proteases. Since the serine proteases differ in length, the absolute position of each conserved residue will vary between sequences. Therefore, it is useful to adapt a standard numbering system for the conserved residues of the serine proteases. By convention of the serine protease research community, this numbering system is that of chymotrypsinogen. A description of the chymotrypsinogen numbering system can be found in Zwilling and Neurath (1981). For this dissertation, I will place the letter “C” before numbers utilizing the chymotrypsinogen system.

There are six cystine bridges in most vertebrate trypsinogens (Kauffman, 1965). Of these, three are absolutely conserved in all serine proteases. The bacterial and crayfish trypsins lack all three of the “optional” vertebrate bridges (Titani et. al., 1983; Kim et al., 1991). The tunicate trypsin gains one bridge (between residues C136 and C201, designated C136/C201; Table 3.5). The lamprey trypsin gains another two bridges (C22/C157 and C127/C232) to reach the vertebrate standard. Curiously, all group I human trypsins have lost the C127/C232 bridge. Furthermore, human trypsin T4 has also lost the C136/C201 bridge. Thus, a progressive increase of cystine bridges is seen during the course of vertebrate trypsin evolution. The human lineage shows a subsequent decrease. This demonstrates that the consideration of trypsinogen cystine bridges for parsimony inference is particularly appropriate, as they are neither absolutely conserved nor extremely labile. If

they were absolutely conserved, they would provide no differentiating information. If they were labile, independent events might masquerade as descendants from a common ancestor.

The group II human trypsinogen T3 pseudogene possesses eleven of the twelve cysteine residues necessary to make the six cystine bridges labeled in Table 3.5. Thus, the functional precursor to this pseudogene had all six bridges, in contrast to all functional human trypsins. The loss of the cystine bridges from the group I human trypsins is therefore recent, as it occurred after the mammalian divergence, and in only one of the two major branches of the trypsinogen phylogeny.

3.11 INSERTIONS AND DELETIONS

Evolutionarily conserved insertions and deletions are expected to be rare events, and thus serve as good markers for tracking gene family phylogenies over large time scales.

The tunicate shares a single residue insertion at position 21 in common with rat trypsin IV. This event is most likely a coincidence, as it is found in no other vertebrates. A second tunicate insertion occurs near residues 45-51. This coincides precisely with the boundary between exons two and three. The crayfish shares this insertion (Titani et al., 1983). Of the vertebrates, only the lampreys have an insertion at this site, but the lamprey insertion consists of only two residues. This is consistent with a progressive loss of residues at this site during early vertebrate evolution, with five lost prior to the Agnathan divergence, and another two lost prior to the elasmobranch divergence. These residues appear to be part of a surface loop (Figure 3.2). They are thus unlikely to have much functional significance, other than possibly a role in determining substrate specificity. A third tunicate insertion of five residues occurs near positions 115-119. The tunicate also has a deletion of three residues around position 223. Neither of these events is observed in any other trypsin, consistent with the 600-700 million year period of independent evolution since the urochordate/vertebrate split.

The tunicate, lamprey, dogfish, and all but one of the Osteichthyes trypsins lack a residue at position 130 that is found in all other vertebrate trypsinogens. A residue is present at this position in salmon trypsin III. Therefore, most likely, there were at least two

trypsinogen isozymes present in the common Osteichthyes/tetrapod ancestor, one of which gained a residue at position 130. Both variations were maintained by the Osteichthyes, but the insertion became the exclusive variant for the tetrapods, perhaps due to coincidental evolution and/or gene copy number contraction and expansion (Hood et al., 1975). Note that rat trypsin V lacks a residue near position 130. This most likely represents an independent deletion event, especially considering that this gene appears to have undergone rapid evolution in recent times. This recent “burst” of evolution is discussed further in Section 3.18. An alternative multiple alignment with only a minor loss in alignment score permits the rat V deletion to align precisely with the deletion noted in Osteichthyes. It is conceivable that these deletions descend from a common ancestor. However, in order for this hypothesis to be supported, the deletion would have to be present in tetrapod trypsins yet to be sequenced.

3.12 INTRON/EXON BOUNDARIES

Point mutations that alter codons are not the only means of introducing variability into genes. Nevertheless, as discussed in Appendix B, most statistical models for evolution focus exclusively on point mutations. One potentially important mechanism for variability is junctional sliding. Junctional sliding refers to the reassignment of a splice acceptor or donor site for an intron. Junctional sliding has been referred to by other names, such as “intron shifting” or “intron sliding,” but this has resulted in confusion with a mechanism that reassigns both acceptor and donor sites simultaneously. The frequency, if any, of intron sliding is under debate (Stoltzfus et al., 1997). Junctional sliding is well documented (e.g., Mayo et al., 1985; Higashimoto and Liddle, 1993). Junctional sliding has been postulated to play a role in the development of certain insertions and deletions between members of the serine protease family (Craik et al., 1983).

Intron/exon boundaries for the vertebrate trypsinogens, where known, are shown in the multiple alignments as vertical lines (Figure 3.3 and Table 3.4). These are determined from the available genomic sequences for human, mouse, chicken, and lamprey. Of note is the absolutely conserved location of the boundary between exons four and five which occurs near the active site serine. The 1/2 and 3/4 exon boundaries are also highly conserved, so sliding of intron boundaries does not appear to have been a major mode of evolution for the vertebrate trypsinogens. A notable exception is the 2/3 exon boundary,

which occurs immediately adjacent to a position of inserted residues in lampreys and tunicates (see Section 3.9).

3.13 CATIONIC AND ANIONIC TRYPSINS

It has been known for some time that vertebrate trypsinogens occur in at least two different isoforms, termed “cationic” and “anionic” (discussed by Le Huerou et al., 1990). Most species appear to express one or more representatives of each of these isoforms. Whether a trypsin is cationic or anionic is determined by its isoelectric point. The predicted and experimental isoelectric points for the chordate trypsins are presented in Table 3.6.

There are no “key” residues at specific sites that are characteristic of either the anionic or cationic isoform groups. In other words, neither the cationic nor the anionic trypsin sequences possess highly significant conserved residues that the other isoform group does not also possess. Rather, net charge is governed by variations in a number of highly variable surface residues. This phenomenon is discussed by Smalås et al. (1994).

No difference in functional role has been demonstrated between the cationic and anionic trypsins, although a possible difference in substrate specificity has been proposed (Fletcher et al., 1987). The trypsins do, however, vary in their catalytic efficiencies for certain substrates as well as in their stabilities at a particular pH or temperature (Smalås et al., 1994). The trypsins also differ in their susceptibility to inhibitors (Read and James, 1988).

It is unclear that there is a selective advantage for an organism to have multiple trypsins with different isoelectric points. Such an advantage, if any, may be as simple as a need for different trypsin isozymes to have different substrate specificities in order to most efficiently digest a wide variety of foods. If this were the case, one would expect organisms with diverse diets to have more trypsin isozymes than organisms with restricted diets. This hypothesis will have to wait to be tested until more complete sets of trypsin sequences from particular species are available.

An alternative hypothesis to explain a selective advantage for two groups of trypsin isozymes is that there are two very distinct functions carried out within an organism by the trypsins, with one function the task of the anionic trypsin(s) and the other function the task

of the cationic trypsin(s). However, if this were the case, one would predict a more clear grouping of the isoforms, including specifically conserved residues critical to the unique task of the particular group. Also, one would predict a selective pressure towards optimal pIs for each of the two tasks, resulting in a bimodal distribution of the trypsin pIs. However, this is not seen. The predicted isoelectric points of the vertebrate trypsins span the pI spectrum continuously from 4.4 to 8.3 (Table 3.6). Note that measured isoelectric points depend not only on the net charge, but also on the distribution of the charge, whereas the charge predictions do not take this into account (Smalås et al. 1994). Also, based on data available to date, the lampreys and tunicates possess only biochemically anionic trypsins, suggesting no absolute need for a chordate to have trypsins of two different charges.

The relevance of the cationic and anionic groupings of the vertebrate trypsins to their phylogenetic origins is discussed in Section 3.16.

3.14 MULTIPLE SEQUENCE ALIGNMENTS

In order for a comparison between two sequences to be made, they must be aligned. In the general case, alignment can be quite difficult. However, for trypsin, it is extremely easy.

The potential difficulty in sequence alignment is uncertainty in which positions of the sequences correspond to each other. The sequences may start or end at different positions. They may be different lengths. One sequence may have insertions or deletions with respect to the other. The similarity between the sequences may not be sufficient to recognize across the whole length of the sequences, but may be confined to one or a few small regions. Thus multiple alignment is often extremely difficult. To address this difficulty, a number of algorithms have been developed (e.g., *HMMER*, *CLUSTAL W*, and several others).¹¹ However, none of these algorithms are necessary for the trypsinogens. Vertebrate trypsin sequences can be aligned manually.

¹¹ *CLUSTAL W* is described by Thompson et al. (1994); *HMMER* is described by Eddy et al. (1995). Overviews of multiple sequence alignment algorithms can be found in several sources, such as Gusfield (1997) or Waterman (1995).

The trypsinogen multigene family possesses a number of features that make it particularly amenable to multiple sequence alignment. These include a cleavage site following an activation peptide, six cysteine residues necessary to form the three absolutely conserved cystine bridges, four active-site pocket-specificity residues embedded in conserved sequences, three catalytic residues embedded in conserved sequences, and several other highly conserved sequences. These conserved sequences are spread throughout the length of the protein, allowing members of the multigene family to be easily aligned, as regions of low similarity are inevitably flanked by conserved residues.¹² Variations in sequence length between two conserved residues can be recognized as insertions or deletions.

There is considerable variation in overall sequence length between the various subfamilies of the serine protease gene family, including some of the invertebrate trypsins, as discussed in Section 3.7. However, there is very little variation in overall length within the vertebrate trypsinogens. Insertions and deletions have been rare during vertebrate trypsinogen evolution. Those that do exist are either one or two residues long. Both the amino- and carboxy-termini of trypsin are conserved, allowing for no uncertainty in aligning the protein ends.

The multiple alignment used for most of my analyses was performed at the amino acid level. This was necessary, as there are two vertebrate trypsinogens for which only amino acid sequence exists: the dogfish and pig sequences.¹³ These sequences represent key nodes in the vertebrate trypsinogen phylogeny so it would be limiting to restrict an analysis solely to known nucleic acid sequences. The nucleic acid sequences were incorporated into these alignments as hypothetical translations.

In some cases complete pretrypsinogen sequences are not available. For example, the protein sequence of the pig and dogfish trypsinogens do not include signal sequences. Therefore, I have limited my formal analysis to multiple alignments of the portions of the sequences coding for the mature trypsin peptide. Also, activation peptides vary in length,

¹² A series of papers by Hedstrom et al. (1992, 1994a, 1994b) describe many of the key functional constraints on trypsin residues and provide a review of relevant literature.

¹³ Until recently, the bovine cationic trypsinogen was also known only by its protein sequence. However, in 1994, Okajima made a direct submission of the identical "cattle" cDNA sequence to Genbank. Although not a vertebrate trypsin, the crayfish trypsin sequence is also only known at the amino acid level (A00951).

which would lead to ambiguity in precise alignment and distance calculations for complete trypsinogen or pretrypsinogen sequences (see Section 3.9). However, the last residue of the activation peptide, which is always a lysine, arginine, or histidine, can be unambiguously included in multiple alignments. The last residue of the activation peptide has therefore been included in my alignments, as it provides a small amount of additional phylogenetic information.

A multiple alignment of most of the known vertebrate trypsins is shown in Figure 3.3. Vertebrate trypsins that are not shown are nearly identical to one of the displayed trypsins. A sequence logo for the vertebrate trypsinogens is presented in Appendix B (Figure B.2). Sequence logos provide a graphical method of viewing the information content of the conserved residues in a multiple alignment. The multiple alignment and the corresponding sequence logo highlight the sequence features that are characteristic of the vertebrate trypsinogens.

All trypsins contain six absolutely conserved cysteine residues, which are necessary to build the three cystine bridges observed in all serine proteases (see Section 3.10). These cysteines are at positions C42, C58, C168, C182, C191, and C220.

All serine proteases, including the trypsins, contain three key catalytic residues: a histidine at position C57, an aspartate at position C102, and a serine at position C195. Each residue is positioned in highly conserved sequence contexts (reviewed in Zwilling and Neurath, 1981).

All trypsins contain four key “pocket specificity” residues: aspartate, glutamine, glycine, and glycine at positions C189, C192, C217, and C227 (Ken Walsh, personal communication). These four residues distinguish the trypsins from all other serine proteases. Hedstrom et al. (1994) provide a more extensive discussion of the sequence characteristics that determine the catalytic specificity of trypsin.

Mature trypsin sequences always begin with one of two nearly identical sequences: IVGG or IIGG at positions C16-C19 (see Section 3.9). Most serine proteases begin with similar sequences (Zwilling and Neurath, 1981).

All vertebrate trypsins possess a calcium binding site on a “calcium loop,” characterized by the residues glutamate, asparagine, valine, glutamate, and glutamate at

positions 56, 58, 61, 63, 66 (Bode and Schwager, 1975). The negatively charged acidic residues chelate the positively charged calcium ion. The role of calcium binding in trypsin was discussed in Section 3.7.

3.15 SEQUENCE DISTANCES

Once sequences have been aligned, evolutionary distances between them can be determined.¹⁴ There are many methods for calculating evolutionary distance. Most of these are algorithms that operate on a pair of sequences. Some, based on maximum likelihood, operate on an entire data set. Maximum-likelihood methods are favored, but are computationally intensive, and so may not be possible with large data sets (Felsenstein, 1983). Maximum-likelihood algorithms operating on nucleic acid sequences are more advanced than those that operate on protein sequences.¹⁵ Both because of lack of computational resources and lack of implemented protein algorithms, most of my trypsin analyses were done with pairwise distance methods.

For multidimensional scaling, described in Section 3.16, and phylogeny construction and jackknifing, described in Section 3.17, I used distances derived from the program *Protdist*, part of the *PHYLIP* package (Felsenstein, 1993). *Protdist* was executed with the “Dayhoff” algorithm, which utilizes Dayhoff’s PAM 001 matrix (Dayhoff, 1979). I chose this simple algorithm for this purpose for its ease of use and speed of execution. The distances used for the phylogeny in Appendix B (Figure B.3) and the statistics of coincidental evolution (Figure 3.16) were calculated with the algorithm described in Appendix B. In no case did I observe a qualitative difference in results when different distance algorithms were employed.

Distances calculated between pseudogenes were not generally considered. The model for distance calculation assumes that all genes are evolving under the same selective constraints. Pseudogenes evolve with a near lack of selective pressure, so it would be inappropriate to include them in distance calculations with functional genes. It is true that

¹⁴ It is often best to consider alignment and phylogeny simultaneously (see, for example, Vingron and von Haeseler, 1997). However, in the case of trypsinogen, with its unambiguous alignment, there is no need for this complication.

¹⁵ The two current protein maximum likelihood programs are *PROTML*, which is part of the *MOLPHY* package (Adachi and Hasegawa, authors), and *PAML* (Yang, 1997).

the assumption of uniform selective pressure is violated even by the functional trypsinogens. However, differences in selective pressure between functional genes will be negligible compared to a complete absence of selection. After construction of a phylogeny, pseudogenes can be assigned topological locations based on similarity or identity comparisons.

The differences in selective pressure operating on the functional trypsinogens may prevent accurate reconstruction of a phylogeny. This would occur if the rates of evolution in different branches of the phylogeny were skewed to such an extent as to warp the tree topology. This may indeed have happened, as discussed in Section 3.17.

Sequences that are distantly related to each other but subject to common selective constraints will still resemble each other. As the divergence time grows, the calculated distance between two sequences will approach the maximum distance dictated by the selective constraints. Such selective constraints play an important role in trypsin evolution (Read and James, 1988).

The aligned vertebrate trypsin amino acid sequences contain 228 sites (Section 3.14). As discussed in Appendix B, 115 of these sites are highly or absolutely conserved. Most of the variation from sequence to sequence occurs at the other 113 sites. Thus, almost exactly fifty percent of the vertebrate trypsin sequence is highly conserved. This suggests that two infinitely diverged vertebrate trypsin sequences will still share high identity.

It is relatively easy to calculate the expected identity of two “infinitely” diverged vertebrate trypsin sequences. As sequences approach large divergence times, converging mutations become as common as diverging mutations, obscuring the actual divergence time. The expected identity at infinite divergence time can be calculated from equation (B.8) by setting the divergence time, τ , to infinity. The resulting expectation is 53%. It turns out that several vertebrate trypsin sequences are nearly this far apart from each other. For example, Cod I and Rat V share 56% identity. This suggests that many vertebrate trypsin sequences are indeed very far apart from each other phylogenetically, despite their apparent similarity. The same statement can be made more generally of all trypsins. For example, the least identity between two chordate trypsins is 34%, between *Botryllus* I and Mouse T8. Several authors have debated the implications of similar observations (Hartley, 1970 and 1979; Hewett-Emmett et al., 1981).

If the vertebrate trypsinogens are truly confined by selection to share at least 53% identity plus or minus some variance, then there must be an explanation for the lower identity observed between vertebrate trypsins and trypsins from non-vertebrates. There are two possible explanations for this. The first is that changes in selective pressure account for the difference. The second is that one or more extremely rare events operated to create specific changes in the trypsin sequences. Indels might serve such a role.

If one imagines a multidimensional structure-space, representing all possible sequences, and a subset of that space representing all possible functional trypsins, then there may be several regions of that space which are separated by large mutational distances, with few functional sequences providing a mutational “pathway” connecting functional regions of the trypsin-space. This “sequence-space” concept was introduced by Eigen (1988) for application to viral phylogenies. Vertebrate trypsins may occupy one particular area of the trypsin function space, confined to that area by selection, only able to mutate out of that area if a rare mechanism of random variation operates on them. I propose that such a mechanism operated to effect the separation of the vertebrate and non-vertebrate trypsinogens, perhaps in conjunction with altered selective constraints.

If such a rare event (or events) separated the vertebrate trypsins from the other trypsins of the living kingdoms, then it will be hard to develop statistical models to estimate the sequence distances between them. Statistical models are discussed in Appendix B. It may be that parsimony is more suited for analysis across such great distances. A parsimonious analysis of rare events, such as indels, or gain and loss of highly conserved residues such as those in certain cystine bridges, may ultimately provide the topology not only of trypsin evolution, but of serine protease evolution in general.

There is one rare event in particular, which if it had occurred, would be of great interest. There are two sets of codons that can code for serine: TCN and AGY. It takes two point mutations to convert a codon of one set into a codon of the other set. The first of these two point mutations would alter the serine residue. Therefore, if a serine protease employed a codon of one set to code for its active site serine at C195, then that protein could not point mutate its active site to a codon of the other set without becoming non-functional after the first point mutation (Brenner, 1988). A switch of the codon for serine C195 from one set to the other in the vertebrate phylogeny would represent a rare and

intriguing event. However, this event is not observed. All trypsins, including the vertebrate trypsins, employ a TCN codon for serine C195.

One additional problem impedes distance calculations between trypsins. As mentioned above, the aligned vertebrate trypsin amino acid sequences are only 228 residues long. The corresponding nucleic acid sequences are three times that long, with 684 nucleotides. However, accurate reconstruction of the topologies of complex phylogenies is hypothesized to require about 2,000 sites, even in “easy” cases (Hillis, 1996). Thus one expects some uncertainty in a distance-based topology generated for the trypsin phylogeny. Such uncertainties can be evaluated with bootstrapping or jackknifing, as discussed in Section 3.17. Recent advances in tree-construction algorithms may ease the recovery of correct tree topologies for short sequences (Tandy Warnow, personal communication).

3.16 MULTIDIMENSIONAL SCALING

Once pairwise sequence distances have been calculated, the relationships between the sequences can be explored. Before embarking on a phylogenetic analysis, other techniques are useful for investigating sequence relationships. In particular one seeks to determine if the trypsins fall naturally into certain clusters based on their distance relationships. For example, one is interested in determining if clusters based on distance relationships correspond to “anionic” or “cationic” clusters. One is also interested in determining the sequence-distance clustering relationships of the group I and group II trypsinogens, which occupy different syntenic relationships with respect to the TCR V β gene segments, as discussed in Section 3.2.

An introduction to the subject of clustering is provided by Everitt (1993). In many cases it is not possible to provide rigorous statistical support for or against alternative clusterings of data. Therefore, one of the main goals of cluster analysis is to provide hypotheses which must be confirmed with external data. There may be many alternative hypotheses. For example, there are 2.2×10^{12} ways to cluster the 42 known vertebrate trypsin sequences into 2 groups. This can be calculated as follows (Liu, 1968):

$$N(n, g) = \frac{1}{g!} \sum_{i=0}^g (-1)^{g-i} \binom{g}{i} i^n \quad (3.1)$$

In equation (3.1), n is the number of sequences; g is the number of groups; N is the number of possible groupings.

One of the more useful methodologies of cluster analysis is termed “multidimensional scaling.” This methodology can be used to convert data from high-dimensional distance matrices, which cannot be visualized graphically, into two or three dimensional plots. Two and three dimensional plots can be visualized, and may illuminate important distance relationships and clusterings. Visual inspection of these plots can often permit one or a few hypotheses for clusterings to be selected from the myriad of alternatives. A introduction to the subject of multidimensional scaling is provided by Everitt and Dunn (1991); Cox and Cox (1994) provide additional details.

Other techniques of cluster analysis can be employed to suggest relationships between proteins. For example, Yee and Dill (1993) demonstrate the use of “minimal spanning trees” and “hierarchical clustering” to analyze the structural relatedness of globular proteins, including the serine proteases. However, both of these techniques are particularly susceptible to bias from coincidental evolution. Avoidance of this bias was a major factor in my selection of multidimensional scaling as a technique to analyze the trypsinogen sequences.

Multidimensional scaling is rarely used for phylogenetic analyses, but its use is increasing. For example, Suyama et al. (1997) employ multidimensional scaling to investigate the three-dimensional “structure profile” distances for the globins. Multidimensional scaling has the advantage of being free of a key assumption about phylogenetic relationships. This assumption is that sequences evolve independently after diverging. This is synonymous with assuming that coincidental evolution does not occur. The assumption of independence is fundamental to all current phylogeny programs, such as those in the *PHYLIP* package. As a result, extensive coincidental evolution will confound the topological reconstructions of such programs. Multidimensional scaling can help clarify the topology of the real phylogeny and, in the process, provide evidence for the occurrence of coincidental evolution.

Multidimensional scaling of the vertebrate trypsin distances is shown in Figure 3.4. Only 32 of the 42 known sequences are utilized; each of the remaining 10 sequences is

nearly identical to one of the included sequences. Note that multidimensional scaling is invariant to orthogonal transformations, which include rotations and magnifications. Therefore the rotational positioning of the axes is arbitrary, as are the units of distance. The informational content of the plot lies in the relative distances between points. If a “molecular clock” hypothesis held, as discussed in Appendix B, then these distances could be interpreted as units of time.

Two notes of caution should be sounded before reaching conclusions based on multidimensionally scaled plots. The first is potential existence of alternative minima; the second is the potential for overzealous dimensional reduction.

Figure 3.4 represents a global minimization but gives little insight into alternative possible local minima. Given the high variance of distance data from short sequences, and the large number of points, alternate minima may represent equally valid depictions of the data. However, I am unaware of programs that explore alternative local minima. Therefore, I would welcome the addition of such options in relevant computer programs, perhaps implemented with a simulated-annealing algorithm. I currently have no good method for evaluating alternative minima, so will not discuss it further.

The simplified subset of trypsin data is inherently 31-dimensional, one dimension less than the number of plotted sequences. Therefore, informational content is lost as the data is “compressed” into a lower-dimensional space. This loss in informational content can be characterized by the “stress” statistic, which is based on the squared differences between the original and the scaled distances (Cox and Cox, 1994). The stress for the multidimensionally-scaled plot in Figure 3.4 is graphed in Figure 3.5. The stress of the two-dimensional plot in Figures 3.4 is below 25%, and thus low enough to indicate that this plots adequately portrays the underlying structure of the data (Everitt and Dunn, 1991). The utility of this plot is supported by the gradual rise in stress through two dimensions, with a large increase occurring only between two dimensions and one dimension. This suggests that although one dimension is not enough to portray the data, two dimensions is sufficient. It is nevertheless interesting to examine the data in three dimensions. Views of a three-dimensional scaling of the data in Figure 3.4 are provided in Figure 3.7 and Figure 3.8.

An additional heuristic exists that can be used to gauge the validity of multidimensional scaling. The Euclidean pairwise distance matrix that is produced from the original data can be employed as input data for a phylogeny construction algorithm. If the “scaled” distances produce the same phylogeny as the original data, then one is reassured that scaling has not grossly altered the data. For the vertebrate trypsinogen data, this exercise produces a phylogeny which is identical in topology and nearly identical in branch lengths to the original phylogeny (data not shown). Phylogenies are discussed in greater detail in Section 3.17.

Having considered these cautionary notes, one can generate several possible hypotheses for clustering the trypsins are suggested by the multidimensionally-scaled plot in Figure 3.4. In particular, there seems to be a natural division of the trypsins into two groups, one at the top of the plot, and one at the bottom of the plot. This forms the basis of a hypothesis that the vertebrate trypsins cluster naturally into two groups. I propose that these two groups correspond to the two groups of trypsinogens defined in Section 3.2. I propose that the origin of the cluster is a schism of the ancestral vertebrate trypsinogen multigene locus into two separate multigene loci that were maintained by all descendant species of the ancestral species. I also propose that this schism occurred after the Agnathans diverged from the ancestral vertebrate lineage, so that the Agnathan trypsins belong to neither of the two groups of trypsin. This hypothesis is displayed in color in Figure 3.6, which, other than the added colors, is a reproduction of Figure 3.4. In this and in all subsequent figures, blue indicates group I trypsinogens, and red indicates group II trypsinogens.

The two possible general mechanisms for the schism of an ancestral trypsinogen multigene family are shown in Figure 3.7. The first mechanism is the insertion of a foreign gene into the locus, dividing it. The second mechanism is the duplication of the locus to the opposite side of a foreign gene. There are a number of specific molecular mechanisms that could account for either general mechanism (see, for example, Li, 1997). For the remainder of this chapter, I will refer to both of these mechanisms as the “division” of the trypsinogen multigene family, incorporating the possibility of division by duplication into this meaning. This division is analogous to the *allopatric* division of a species. Allopatry is a term used in population genetics to refer to the division of a species by a geographic barrier, such as a mountain range. Such a barrier can result in speciation, which is the division of a single

ancestral species into two descendant species. By analogy, I will refer to the division of the trypsinogen multigene family as an “allopatric” division.

Three-dimensional multidimensional scaling of the trypsin data supports the hypothesis of a single major division of the trypsinogen multigene family (Figure 3.8 and Figure 3.9). The three-dimensional views, in particular, support the hypothesis that the lamprey sequences belong to neither group I nor group II. The view in Figure 3.9 strikingly illustrates the division of the two groups along a central plane.

The syntenic positions of several trypsinogen genes with respect to the TCR β locus are known. This allows these sequences to be assigned to either group I or group II with certainty. Additionally, each of the rat sequences can be assigned to a group with near certainty, due to the extreme similarity they share with orthologous mouse genes.¹⁶ The trypsinogens with known syntenies are colored in the plot in Figure 3.10. This data is absolutely consistent with the hypothesis suggested in the preceding paragraphs, but leaves a few sequences unassigned, other than based on the clustering suggested by multidimensional scaling (Figure 3.6).

The grouping of the remaining sequences is supported by a consideration of their isoelectric points. The isoelectric points of the trypsins are indicated on the plot in Figure 3.11; the actual values for the isoelectric points are tabulated in Table 3.6. It is immediately apparent that the group I trypsins are mostly anionic, while the group II trypsins are mostly cationic. Thus, there is a concordance of three different data types with independent components that supports the hypothesis of exactly two natural groupings of the vertebrate trypsins, with the lamprey sequences not belonging to either. These are: one, syntenic evidence; two, multidimensionally-scaled distance data; and three, isoelectric point evidence.

Consideration of isoelectric points is particularly valuable for grouping the dogfish and Osteichthyes trypsins, for which no syntenic data is available. Without this data, an alternative grouping of the trypsinogens might gain consideration. From the

¹⁶ It is difficult to prove orthology. Additionally, the term is somewhat meaningless for members of a multigene family. My use of the word “orthology” in this instance is meant to imply recent divergence from an ancestral gene with the same syntenic relationship to the TCR β locus. The orthology of the rodent trypsinogens is discussed further in Section 3.17.

multidimensional scaling plots, it is possible visualize the trypsins grouped into about four clusters (Figure 3.4). In addition to division of the plotted points by a horizontal line, one can also imagine division by a vertical line separating the fish sequences from the remaining sequences. If one divided the trypsins in this manner, that might lend credence to an alternative hypothesis that there were two or more independent major duplications or divisions of the trypsinogen multigene family. However, the assignment of isoelectric points for the fish trypsins is absolutely consistent with their division in to exactly two groups. This supports my hypothesis of a single major division of the trypsinogen multigene family.

Significantly, the lamprey trypsins are strongly anionic. This suggests that they do not belong to group II. Additionally, as noted above, they do not cluster with either group I or Group II in three-dimensional multidimensionally-scaled plots (Figure 3.8 and Figure 3.9). Together, these data support the hypothesis that the division of the trypsinogens occurred after the divergence of the Agnathans.

Several discrepancies between isoelectric point data and phylogenetic expectations can be noted. Mouse T4, rat IV and rat V, which phylogenetically should be “cationic,” have predicted anionic charges. Human I and *Xenopus* 51, which phylogenetically should be “anionic,” have cationic charges. Thus isoelectric points merely correlate with phylogenetic grouping, rather than mirror it. For this reason, I feel that there is little utility in designating trypsins as “cationic” or “anionic,” and suggest a re-evaluation of this nomenclature.

The lack of absolute correlation of the isoelectric points with phylogenetic group is not a major obstacle to the “two group” hypothesis. First of all, all three discordant rodent sequences — mouse T4, rat IV and rat V — are already known to be group II based on synteny, as discussed above. The same is true of the discordant human T8 trypsinogen. This leaves only the discordant *Xenopus* 51 trypsinogen to be explained, but its similarity with *Xenopus* I coupled with its unambiguous position in the multidimensionally scaled plots leave little doubt as to its group I assignment. It remains possible that *Xenopus* 51 is a group II trypsin that has undergone extensive coincidental evolution, but if this were the case, it would have little impact on the conclusions of the “two group” hypothesis. It would, in fact, support the contention of Section 3.18 that coincidental evolution has played a major role in vertebrate trypsinogen evolution.

There is a possible explanation for the discordance of the human T8 isoelectric point. Recall from Section 3.2 that humans possess no functional group II trypsins. However, as mentioned in Section 3.13, it may be that both anionic and cationic isoforms have a functional niche, with the jawed vertebrates requiring both. In this scenario, human trypsin I, which is biochemically cationic but phylogenetically group I, would fill a crossover role by filling a niche vacated by the missing group II trypsinogens. Also since rat C and mouse T7 are cationic, in this scenario there might be little selective pressure on mouse T4, rat IV, and rat V to remain cationic, so their charge could have “drifted.”

I did not incorporate pseudogenes into my multidimensionally scaled plots. The use of pseudogenes might skew the structure of the data. Such skewing effects can be severe for phylogenies, as shown in Section 3.17. However, skewing due to distorted distances for multidimensional scaling can be surprisingly mild (Everitt and Dunn, 1991). This may be the case for vertebrate trypsin data. Several pseudogenes are incorporated into the data for Figure 4 of Appendix F (Roach, 1997). As expected, the group II pseudogenes cluster with the functional group II genes, and the group I pseudogenes cluster with the functional group I genes.

In an effort to further explore the utility of multidimensional scaling, I multidimensionally scaled the distances for the rodent group I trypsinogens, including several sequences that were not included in the simplified vertebrate data set (Figure 3.12). In this case, multidimensional scaling provides little insight beyond what can be obtained from a traditional phylogenetic analysis (see Section 3.17). A difference between the rodent group I data set and the vertebrate data set is that all of the group I rodent trypsins are very closely related. Multidimensional scaling may be most useful for examining distant relationships. The stress for the multidimensionally-scaled plot in Figure 3.12 is graphed in Figure 3.5.

3.17 PHYLOGENIES

A consideration of the data presented in the previous section leads one to conclude that the trypsinogen multigene family underwent an allopatric division about 500 million years ago, during the Ordovician or Silurian Periods. One thus predicts the existence of two groups of trypsinogen in all of the jawed vertebrates. All of the members of each group should be more related to each other than any are to members of the other group, as each

group shared a more recent ancestor. This assumes that no coincidental evolution has occurred (see Section 3.1). With these assumptions, a hypothetical phylogeny can be sketched (Figure 3.13). This figure also assumes a roughly constant rate of evolution. This figure can be compared to the computed phylogenies discussed below. Such comparisons highlight the impact of coincidental evolution on the vertebrate trypsinogens.

A phylogeny of the vertebrate trypsinogens can be computed from a pairwise distance matrix, calculated as discussed in Section 3.15. Such a phylogeny, for forty-two sequences, is shown in Figure 3.14. This phylogeny can be recalculated with fewer sequences. This can be done with little loss of information, as there are several sequences that are nearly identical to each other: mouse T11, T2, T15, and T16; mouse T8 and T9; mouse T4 and T5; chicken I and 38; salmon I and II; cod I and X; lamprey B1 and B2; lamprey A1 and A2. Not only is the resulting phylogeny less cluttered, but it also demands fewer computational resources to calculate.

A thirty-two sequence phylogeny is shown in Figure 3.15. The branches of the phylogeny are colored to correspond with the group to which the sequences at the terminus of the branch belong.

This phylogeny is striking in two major respects. First, it fails to support a molecular clock hypothesis. This is most striking for several of the rodent trypsins (Rat IV, Rat V, and Rat Cationic). Since mice possess nearly identical homologs to all the known rat trypsinogens, these rate variations must have occurred before the mouse/rat divergence. The largest intra-species rate variation observed in this data set occurs between the Rat C and Rat V branches of the phylogeny. In Figure 3.15, the ratio of the Rat V to Rat C branch lengths is 4.62. This represents an average rate difference. Rate differences in any given period after the divergence of these sequences from a common ancestor may have been larger or smaller. Therefore, since the mammalian radiation, rates of evolution may have differed by as much as an order of magnitude between different isozyme loci within a species. This is consistent with “bursts of sudden evolution” at particular loci, perhaps due to gene conversion events.

The second striking aspect of the phylogeny in Figure 3.15 is that it fails to reproduce the topology of trypsinogen evolution predicted in Figure 3.13. Notably, neither the group I nor the group II trypsinogens form a single clade. In particular, all of the

mammalian sequences appear to be more related to each other than they are to any other sequences, with the possible exception of chicken 29. The most likely explanation for this is that coincidental evolution has operated on the vertebrate trypsinogens.

There is an alternate explanation that could explain the deviations of Figure 3.15 from the expected topology of vertebrate trypsinogen evolution. Random variations in sequences can arbitrarily cause two sequences to appear to be more similar to each other than would be expected. In extreme cases, this might mimic coincidental evolution. This is likely to be the reason that the group II chicken sequences form a clade with the group I *Xenopus* sequences. During the jackknifing of the phylogeny in Figure 3.15, *Xenopus* I, *Xenopus* 51, and chicken I formed a clade twenty-two times. However, *Xenopus* I and *Xenopus* 51 formed a monophyletic clade more often, thirty-one times. Although this clade does not show up in the main tree, its significance underscores the lability of the “chicken-group-II and frog-group-I” clade. This particular clade is therefore more likely to be due to randomness in the sequences, and not coincidental evolution or independent duplications.

Note that if only one representative of each vertebrate class is considered, the resulting phylogeny more resembles a star than a tree. A likely explanation for this is that many of the vertebrate trypsins have reached an equilibrium distance from each other, as discussed in Section 3.15. Over large divergence times, chaotic fluctuations away from equilibrium are to be expected. This produces deviations from a perfect star phylogeny, notably small topological deviations in early branchings. The differences in branch length from the center of the star vary due to alterations in evolutionary rate. Some length variation is also expected due to the stochastic incidence of accepted mutations.

One can differentiate between random convergence and covariance, the statistical hallmark of coincidental evolution. For example, one can test the hypothesis that sequences from one group will co-vary with sequences from the other group that belong to the same vertebrate class. This can be done by comparing the distribution of all such distances with the distribution of distances between sequences that not only belong to different isozyme groups but also belong to different vertebrate classes (Figure 3.16). If there were no coincidental evolution, one would expect these distributions to be identical. However, these distributions are very highly significantly different, demonstrating that coincidental

evolution has had a major impact on vertebrate trypsinogen evolution.¹⁷ This finding is discussed in Section 3.18.

Bias must be considered as an explanation for the differences in these distributions. In general, these statistics are immune to many types of bias. For example, if one group has evolved more slowly than the other, the average distance between the two groups will remain invariant to whether or not the comparison is made within a class or between two classes.

One type of sampling bias that can mimic coincidental evolution is if a sequence that has been subject to an atypical rate of evolution belongs to a class that has known sequences from only one group. For example, only the group I dogfish trypsin is known. If this sequence has evolved at an atypical rate, then it might confuse the analysis. For example, if it evolved very slowly, then between-class comparisons would be biased high, making them more distinct from within-class comparisons, and creating the illusion of coincidental evolution. If they evolved more quickly, the opposite would happen, and coincidental evolution would be harder to detect. In general, this effect can occur when there is an imbalance between the number of between-class comparisons and the number of within-class comparisons involving a particular sequence. This bias can be corrected by normalizing the weight of each comparison so that each sequence is involved in the same net weight of between-class comparisons and within-class comparisons (data not shown). In this case, classes with only one known group, such as the elasmobranchs, must be disregarded. This is because, for such a case, there are zero within-class comparisons. Zero cannot be normalized. The unweighted distribution also shows the highly significant difference seen in Figure 3.16, suggesting that this type of sampling bias is not present in the dataset of known trypsinogens. The effect of this weighting is to equalize the number of sequences in each group for a given class by adding “ghost” sequences identical to known sequences.

Oversampling of similar sequences is another type of bias. For example, consideration of some sequences might support the hypothesis of coincidental evolution,

¹⁷ Very few non-mammalian within-class comparisons are available. If the mammalian sequences are excluded from this analysis, the coincidental evolution effect is of low statistical significance. Therefore, it is possible that the entire effect of coincidental evolution occurred in the mammalian lineage. More non-mammalian sequences are needed to clarify this point.

while consideration of other sequences might refute the hypothesis. For example, the mouse trypsinogens are highly sampled in the set of known vertebrate trypsinogens. If a series of recent gene conversion events homogenized the mouse group I sequences with the mouse group II sequences, then consideration of the mouse sequences will, correctly, support coincidental evolution. However, if coincidental evolution has not operated on other classes, and sequences from these classes are undersampled, then the apparent role of coincidental evolution in evolution will be artificially amplified. This type of bias is difficult to eliminate by normalization. One can minimize its effects by employing as many representative sequences from as many vertebrate classes as possible. This was one of my motivations for obtaining trypsinogen sequences from several different vertebrate classes, especially where none were known beforehand. The bias can also be minimized by excluding all but one of a set of recently diverged sequences. Consideration of recently diverged sequences will amplify the effects of any evolutionary events that have occurred prior to their divergence. Thus it is more appropriate to use the subset of thirty-two sequences shown in Figure 3.15 than the entire set of forty-two known vertebrate trypsinogens.

Pseudogenes can be incorporated into phylogenies. This incorporation must be done with care. Pseudogenes are not subject to the same selective constraints as functional genes (see Section 3.15), so should not initially be incorporated into a dataset of functional genes used to build a phylogeny. However, the topology of pseudogene divergences can be estimated after the phylogeny is built (Figure 3.17). For this figure, the insertions of the pseudogene branches were estimated by tabulating the nucleic-acid sequence identities of the best diagonals for pairwise comparisons with pseudogenes computed with the default dotplot from the program *MegAlign* (DNA*®, Madison, WI). Only the three highest identities for each pseudogene were considered. The topology of the insertion point of each pseudogene branch was then positioned so as to minimize the least-squares differences in the proportions of the resulting three branch lengths from the proportions of the three dotplot-diagonal identities. Because of the difficulty of estimating the exact divergence times of pseudogenes, no attempt was made to estimate their branch lengths.

This method of assigning a topology to pseudogene divergence points is somewhat arbitrary. It does, however, highlight numerous independent events that have spawned pseudogenes during the course of vertebrate trypsinogen evolution. Figure 3.17 also highlights the recent divergence of several functional trypsinogens and pseudogenes. In

particular, the close relationship of rat V and mouse T3 can be seen, which allows rat V to be assigned as orthologous to mouse T3, and definitively to group II (see Section 3.16).

Outgroups are useful when constructing phylogenies. In particular, they can be useful in calibrating a molecular clock, if one exists (Li and Graur, 1991). They also serve to root a phylogenetic tree. However, both of these functions can be hijacked if the chosen outgroup is so distant that it has reached equilibrium distance from all other sequences in the phylogeny. As discussed in Section 3.15, the tunicate sequences have indeed reached equilibrium distances from the rest of the vertebrate trypsinogens. The differences in sequence distances between any two vertebrate-urochordate comparisons is expected to be random. Inclusion of a urochordate sequence in a vertebrate trypsinogen phylogeny should therefore result in a random topological insertion of the chordate branch. As a result, urochordate sequences are poor outgroups.

An example of the type of skewing that can result from inclusion of too-distant outgroups is seen in Figure 3.18. In this figure, all three tunicate trypsins and the crayfish trypsin are utilized, but the effect on skewing will be similar if only one of these is employed. The random nature of the insertion point of the invertebrate outgroup can be best appreciated by consideration of the jackknife values for the larger clades. For example, the clade distal to salmon III is equivalent in both Figure 3.15 and Figure 3.18. This clade is found in 69% of the jackknives executed with just the vertebrate data set, but only 43% of the jackknives when the invertebrate outgroup is included. The topology of the chicken I sequence with respect to the *Xenopus* sequences is also altered. These effects are a result of the random insertion of the invertebrate outgroup.

Another type of error that can skew a phylogeny is the inclusion of a sequence that has been subject to markedly different selective pressures from the other sequences in the phylogeny. This would be the case if one of these sequences served a markedly different function. This effect is discussed by Fitch (1970). Including a pseudogene, as discussed above, would cause a similar skewing. For example, the *Pleuronectes* “trypsinogen” sequence is not a trypsin and therefore will not be subject to the same selective pressures as the trypsins (see Section 3.3). Rypniewski et al. (1994) include the *Pleuronectes* sequence in their phylogeny, quite possibly causing just such a distortion. Independently, Male et al. (1995) include the *Pleuronectes* sequence in a phylogeny, again contributing to a skew. Rypniewski et al. and Male et al. may have missed such skewing as a result of the high

divergence that exists between trypsinogen groups and species classes. Uncertainty introduced in these phylogenies by the high divergence of the trypsins may have also masked skewing caused by extreme outgroups.

Even after the exclusion of the invertebrate outgroups, it is difficult to construct a trypsinogen phylogeny with high confidence in its topology or divergence times. Vastly differing evolutionary rates obscure divergence times. The presence of coincidental evolution and maximally diverged sequences obscures topology. The low jackknife values for many of the nodes of the phylogeny in Figure 3.15 are a result of these effects.

However, the bootstrap values are more consistent for vertebrate classes, and reach 100% for certain clades, such as the Osteichthyes group I trypsins and the rodent group I trypsins. Therefore, molecular trypsin data may have some utility for the analysis of recent evolutionary events. To this end, I have constructed a phylogeny of the rodent group I trypsins. All of these trypsins have known DNA sequences. Furthermore, there are only ten of them, so they can be readily analyzed with nucleic-acid maximum-likelihood algorithms. A phylogeny of the rodent group I trypsins is shown in Figure 3.20. This phylogeny suffers from having representative sequences of only two species, but nevertheless has high bootstrap values at its nodes. The dog anionic trypsin is the least diverged sequence from the rodent trypsins, so I attempted to use it as an outgroup for this phylogeny. However, dog anionic trypsin rooted randomly during jackknifing, so I excluded it. A more detailed rodent trypsin phylogeny will have to await the determination of more sequences, such as those from the hamster.

3.18 MODES OF TRYPSINOGEN EVOLUTION

The trypsinogens are a multigene family, and should evolve as one. There is a large literature on multigene family evolution, reviewed by Li (1997). Multigene family evolution is frequently characterized by coincidental evolution. Horizontal transfer of genetic information in multigene families is analogous to the sexual transfer of alleles within a population, and this has led to the observation that the principles of population genetics may be better suited for the analysis of multigene families than traditional single-gene phylogenetic models.

From the information presented in the preceding sections, it is clear that trypsinogen does not evolve as a classical single locus gene with a constant rate. Vertebrate trypsinogen evolution has been dynamic and multimodal. The trypsinogen genes have been evolving covariantly, as a population, and not solely independently, as individuals.

In most vertebrate species, there are two groups of trypsinogen isozymes at separate genomic locations, each coded for by tandem repeats. Tandemly repeated genes exchange information with each other far more frequently than with genes at different loci (Li, 1997). Thus one expects significant genetic exchange within a trypsinogen group, but not necessarily between groups. One would expect trypsinogen genes within a group within a species to be highly identical to each other, with intra-group variation determined by an equilibrium between diverging mutations and converging horizontal genetic transfer. These converging events, such as unequal crossing-over and gene conversion, are a powerful force for homogenization.

Coincidental evolution can also operate on genes that are not tandemly repeated. Any component of coincidental evolution due to selective pressure is unlikely to be influenced by gene location. Crossing-over is unlikely to play a role, particularly in the case of the trypsinogen/TCR locus, as crossing-over would destroy the functional utility of the TCR β locus. However, gene conversion can still operate on separated genes, whether they are located on the same or different chromosomes. This has been well studied in yeast, and is likely to be true for all metazoans, as reviewed by Petes and Hill (1988), and Petes et al. (1991). Nevertheless, gene conversion is less frequent between separated genes than between adjacent genes. Therefore, the homogenization force of horizontal gene transfer is less powerful between separated members of a multigene family.

The division of the trypsinogen multigene family around the time of the elasmobranch divergence was perhaps the most significant event of vertebrate trypsinogen evolution.¹⁸ Following the divergence of the two groups of trypsinogens, they tended not to exchange genetic information with each other, acting largely as separately evolving gene

¹⁸ The dating of the divergence is not absolute. The divergence is possibly very old, predating the vertebrates or even the chordates. If this is the case, the current similarity of the two groups of trypsins in the jawed vertebrates as compared to the lampreys and tunicates (see Section 3.16) would have to be explained by coincidental evolution.

families. However, statistically significant exchange did occur, as documented in Figure 3.16.

The class Mammalia illustrates this point. All of the mammalian trypsinogens are more closely related to each other than to trypsinogens of other classes, with the possible exception of the chicken group I trypsinogen (Figure 3.14 & Figure 3.15). It is likely that this convergence will be more readily observed in other vertebrate classes as more sequences are obtained.

The quantitative contribution to coincidental evolution of gene conversion as opposed to selective pressure is very difficult to determine. Gene conversion dramatically and suddenly homogenizes sequences, so even if its effects are rare compared to mutations selected by species-specific pressures, the effect of gene conversion will predominate. Given the magnitude of this particular coincidental effect, it is difficult to ascribe the overall coincidental effect to selection. Gene conversion almost certainly played a major role in the coincidental evolution of the group I and group II trypsinogens.

The molecular traces of such gene-conversion events, if present, have been largely obliterated by subsequent mutation. In order to observe adjacent covariant point mutations, which are the hallmarks of gene conversion, sequences separated by very short evolutionary distances must be obtained. Currently, no such data exists for trypsinogen sequences. However, such data has permitted the observation of gene conversion in several immunoglobulin gene superfamily loci. This includes the mouse Class I major histocompatibility locus (Pease et al., 1993), and the chicken immunoglobulin lambda locus (Reynaud et al., 1985). The rate of gene conversion at immune receptor loci may be greater than the genome-wide average. Enhanced recombinogenicity at these loci may be a result of aberrant activity in the germline of the RAG1/RAG2 recombination machinery that is normally active only in somatic cells (Hagmann, 1997).

One possible explanation for an enhanced rate of gene conversion for the group I trypsinogens would be unique to these trypsinogens, if it exists. During T-cell development, TCR β locus episomes will contain several copies of Group I trypsinogen genes. This can only occur for genes intercalated between immunoglobulin-receptor gene segments. Only the group I trypsinogens are known to be in such a position. If an episome containing a trypsinogen somehow formed in or recombined with germline DNA, there

could be a striking and sudden change in, transposition of, or even novel generation of a trypsinogen locus. There is no data to suggest that such episomes form in germ cells. However, such formation could be extremely rare and still have a major impact over evolutionary time scales.

As noted in Section 3.13, the general biochemical features within a trypsinogen group are maintained in the absence of absolutely conserved characteristic residues. It may be that general selective pressure for a positive or negative charge accounts for this phenomenon. Alternatively, it may be that the principles of population genetics can account for this phenomenon. Each element of sequence potentially coding for a charged residue can be considered to be a locus at which any of several “alleles” may be present.¹⁹ Different “alleles” would code for different charged residues, or the absence of one. Each trypsinogen group would be characterized by different “allele” frequencies, with negatively charged “alleles” predominating in the group I trypsinogens and positively charged “alleles” predominating in the group II trypsinogens. However, even if this is the case, it seems unlikely that such a model could completely explain the observed trypsinogen sequences. Coalescence theory predicts the extinction of alleles for which there is no selective advantage. In the absence of differential selective pressure, new alleles would show no charge bias for either group I or group II, so over time the two groups of trypsinogen would converge with respect to their net charges. This would be accelerated by gene conversion. The extinction of differentially charged alleles coupled with the introduction of unbiased alleles would erase biochemical differences over evolutionary time scales. Therefore, since these differences have persisted, it seems likely that there is a selective pressure that helps maintain the biochemical differences between group I and group II trypsinogens. Quantitating this pressure would be very difficult.

An additional modality of evolution may operate on the trypsinogens. The repeat structure of the loci provides for the maintenance of a large pseudogene reservoir. The numerous human and mouse pseudogenes are diagrammed in Figure 3.1. These pseudogenes can evolve rapidly, free of evolutionary constraint, as digestive function is provided by the functional trypsinogens of a locus. It is conceivable that, rarely, these pseudogenes will back mutate to functionality, perhaps jump-started by a gene conversion

¹⁹ As an addition to pun-derived terminology, I propose the word “trypsinogene” to designate these trypsinogen “alleles.”

event. They may also provide a genetic reservoir of material for recombination events with functional trypsinogens.

All of the mechanisms of horizontal genetic transfer can produce large changes in evolutionary distance with a single event. For example, gene conversion events in yeast can alter up to 12 kb of contiguous sequence (Borts and Haber, 1997; Petes et al., 1991). Unequal crossing-over, episomal recombination, and pseudogene re-activation could also cause similarly large and sudden evolutionary changes.

The several gross and evolutionarily sudden events described in the preceding paragraphs can cause the apparent rate of evolution of trypsinogen to vary highly between species and even between loci within a species. Such events are complemented by a background of intron junctional sliding, as well as nucleotide transitions, transversions, insertions, and deletions. Taken together, these modes of evolution will significantly confound efforts to build reliable models for trypsinogen evolution. This will, in turn, preclude the construction of reliable phylogenies that span large evolutionary distances, such as those separating classes. Phylogenies spanning more than one phylum are particularly problematic, as described in Section 3.15 and Section 3.17.

Determination of species relationships from phylogenies based on multigene family data is difficult (Cilia et al., 1996; Hollingshead et al., 1994). During the evolution of Animalia, trypsinogen genes are likely to have been consistently present in genomes as one or more multigene loci. For example, several insect species are known to have multiple trypsinogens (Davis et al., 1985). Therefore, it is hard to recommend trypsinogen data for the reconstruction of unknown phylogenies, or for use in the determination of the relatedness of populations, such as is often called for in ecological conservation efforts. Genes definitely known to be single-copy provide the most appropriate data for such studies.

3.19 EXPRESSION OF TRYPSIN

The hypothesis that trypsinogens of different groups serve different functions would be supported if the trypsinogens displayed differential expression. If the two groups of trypsinogen showed a marked difference in the distribution of tissues in which they were expressed, that would suggest tissue-specific functions. Differences in the dynamics of

pancreatic expression might suggest differential regulation in the face of different substrate specificities or activities. Studies of trypsinogen expression can help sort out these possibilities. Additionally, such studies can verify that a genomic sequence is transcribed, spliced, and translated. A genomic sequence may appear to code for a functional gene, but may in actuality be a pseudogene if it is not expressed due to a dysfunction in its regulatory sequences. This possibility cannot be ruled out until a gene product is observed.

Human trypsinogen T6 appears to be a functional gene from its genomic sequence. An analysis of all trypsinogen cDNA sequences found in the EST database reveals six human T6 cDNAs: five from pancreas tumors and one from a normal adult male pancreas (Table 3.7). Additionally, I have observed a PCR product representing a processed human T6 mRNA (Appendix C). Therefore, human T6 is not a pseudogene. However, human T6 has not been noticed in previous studies, such as the cloning efforts for human T4, T8, and T9 (Emi et al., 1986; Tani et al., 1990). Additionally, only these three trypsins have been identified in pancreatic juices (Scheele et al., 1981). There are three possible explanations for the failure of these studies to detect human T6. First, it may be expressed in relatively low quantities. Secondly, it may have been confused with one of the other expressed trypsinogens. Thirdly, the individuals used for analysis may have been heterozygous or homozygous for a deletion of the human T6 gene. This deletion genotype is known and has an allelic frequency of approximately 46 percent (Lee Rowen, Genbank AF009664; Seboun et al., 1989).

There are 64 human T4, 94 human T8, and 14 human T9 cDNAs in the EST database.²⁰ All 64 human T4 and all 94 human T8 cDNAs are pancreatic in origin.²¹ Of the

²⁰ Spliced genomic sequences of human T4, T6, T8, and T9 were used as BLASTN queries of Genbank on 11/9/97. All results with a p-value greater than or equal to 0.05 to any of the four queries were retained. Each retained sequence was then locally dynamically aligned to each of human trypsinogens T4, T6, T8, T9, and chymotrypsinogen. The scoring for alignment was as follows: match, 1; mismatch, 2; gap, 4. All sequences with a maximum score below 40 were discarded, as were sequences that maximally matched chymotrypsinogen. Sequences close to this cutoff were verified by visual examination of dotplot alignments; no false positives or false negatives were detected. Sequences not discarded were identified as particular isozymes based on their highest alignment score. All T6 and T9 sequences were verified by visual examination of the alignments to rule out false positives. A modified version of my program *CrossMatcher* (available from my web site) was used for dynamic alignment.

²¹ One of the T4s is derived from "fetal liver-spleen" and two of the T8s are derived from "ovarian tumor." These descriptions are consistent with a pancreatic origin.

14 human T9 cDNAs, seven are of the alternatively-spliced form described as “brain trypsinogen” by Wiegand et al. (1993), one is the normally-spliced variation, and six do not include exon one, so cannot be assigned to a splice variant. The seven alternatively-spliced human T9 cDNAs are derived from pancreas, colon, pregnant uterus, and fetal heart. The normally-spliced human T9 cDNA is from pancreas. It is difficult to perform a clean dissection or biopsy of abdominal organs without minor contamination from pancreatic tissue, whether from the pancreas itself, or from pancreatic heterotopias.²² Therefore, trypsinogen ESTs found in the colon are likely pancreatic in origin. In summary, with the exception of two sequences from fetal tissue, all trypsinogen ESTs to date are likely to be derived from pancreatic tissue. This argues against trypsinogen performing a significant function in any tissue other than the pancreas.

It seems unreasonable to refer to the alternatively spliced form of human T9 as “brain trypsinogen,” as it seems no less pancreatic than any of the other isozymes. The original isolation by Wiegand et al. (1993) employed PCR, so their sequence is not present in the EST database, but in the main body of Genbank. This alternatively spliced form of human T9 appears to possess no signal peptide capable of transporting the nascent peptide into the lumen of the endoplasmic reticulum. Therefore, this form of human T9 may be targeted to another cellular location, perhaps into the cytoplasm, to play an unknown role.

Trypsinogen is expressed in minute amounts in tissues other than the pancreas. Wiegand et al. (1993) use PCR to detect trypsinogen expression in the brain. I have also employed PCR to survey the expression of trypsinogen in various tissues, as described in Appendix C. Although PCR is not a good measure of relative abundance of cDNA isozymes, it is a very sensitive measure of the presence of particular cDNAs. Trypsinogen cDNAs can be detected by PCR in most, if not all, tissues. This implies that trypsinogen may be expressed at a low level in all cells, or by cells that are present in low numbers in all tissues. Lymphocytes are present in low numbers in all tissues, so it is conceivable that they are the source of PCR-detectable trypsinogen expression. If so, this would be

²² The difficulty of obtaining gut tissue free of pancreatic contamination may be more than an issue of precision dissection. As many as 14% of cadavers have pancreatic heterotopias somewhere in their digestive tract that are detectable by careful histological examination (Ravitch, 1973; Thoeni and Gedgaudas, 1980). It is likely that there are many more pancreatic rests that are too small to be distinguished during a histological dissection, possibly even with some isolated cells.

consistent with an immunological function for trypsinogen, as might be predicted from its syntenic relationship with the TCR β locus.

There are fewer ESTs for mouse than for human trypsinogens. These consist of 24 sequences representing mouse trypsinogens T7, T8, and T9 (Table 3.7).²³ Additionally, mouse T8, T9, T10, and T11 have been observed as PCR products (Appendix C). The mRNA for mouse T20 has been cloned and is in the database. Mouse T4, T5, T12, T15, and T16 are apparently functional as genomic sequences, but their products have yet to be observed. Not enough murine expression data has been obtained to draw any significant conclusions from these observations.

Within the pancreas, the different trypsinogen isozymes are differentially regulated. This is suggested by the relative abundances of the isozyme ESTs in Table 3.7 and the cloned PCR products described in Appendix C. However, these relative abundance numbers should not be interpreted as relative levels of expression, as EST databases do not necessarily reflect tissue abundance, and PCR is subject to the effects of differential amplification. However, it has been well established that the trypsinogen isozymes are differentially expressed (e.g., Schick et al., 1984). Control at the translational level is clearly an important factor in regulating trypsinogen expression (Pinsky et al, 1985; Steinhilber et al., 1988). Additionally, mRNA stability is likely to play a role (Carreira et al, 1996). The relative contribution of transcriptional regulation has not been worked out. It is likely that trypsinogen expression is controlled at all potential checkpoints, permitting maximum physiological control over an enzymatic activity that must be precisely regulated. High gene dosage may play an important role in this control, as discussed in the next section.

3.20 FUNCTION OF TRYPSIN

The only known function of trypsin involves the digestion of food. The presence of multiple trypsin isozymes within an organism raises the possibility that they may perform different functions, as discussed in the preceding sections. The tight linkage of the trypsins to the TCR β locus raises the possibility that one or more trypsinogen isozymes may play

²³ The EST database was analyzed on 10/24/97 with the same methodology of the human trypsinogen EST search.

an immunological role. Their deletion from a functional TCR β locus seems inconsistent with such a role in mature T-cells, but does not rule out other immunological functions. Even the deletion of trypsinogens from functional TCR β loci may not necessarily indicate that trypsin has no role in such cells, for a percentage of peripheral T-cells maintain an unrecombined TCR β locus.²⁴ It is possible that T-cells that express trypsinogen might represent a functionally significant subclass.

Many serine proteases are known to play roles in immune defense (see, for example, Müller et al., 1994). Such roles include antigen processing and cytotoxic proteolysis. However, in the absence of concrete evidence to the contrary, I feel that the null hypothesis for trypsinogen is that it performs no function other than alimentary digestion. Large quantities of trypsin are needed on short notice for digestion, and one facet of gene regulation could involve high gene dosage. An alimentary selective pressure for high gene dosage may be sufficient to explain the maintenance of multiple trypsinogen genes in a genome.

Whether or not there is a need for multiple trypsinogen gene loci is unclear. Since the humans have no functional group II trypsinogens, it would seem that the group II locus is dispensable. However, it remains possible that group II trypsins play a role, perhaps non-alimentary, in vertebrates other than humans. This putative role would either not be necessary in humans or, more likely, be subsumed by a novel serine protease, perhaps ancestrally derived from a duplicated trypsinogen gene. To further complicate the picture, humans have acquired a novel trypsinogen locus by means of the group I translocation to chromosome 9. Any functional significance of this translocation is unclear.

3.21 GENESIS OF NOVEL GENES

The mechanisms of gene creation are of great evolutionary interest. Multigene families play an important role in the creation of new genes (Li, 1997; Henikoff et al., 1997). Duplicated members of a multigene family have a redundant function, so are free to vary without deleterious effects on the phenotype of an organism.

²⁴ The exact percentage is unclear, but is likely to be between 1% and 50%. Haars et al. (1986) and Seboun et al. (1992) present differing viewpoints.

At least once during the course of vertebrate evolution the trypsinogen multigene family spawned a novel gene. At the time that the human and mouse TCR β loci were first sequenced, it was thought that human trypsinogen T1 and mouse trypsinogen T1 were orthologous pseudogenes. It was noted that they had in-frame coding sequences, but there was no evidence that these sequences were expressed. Additionally, it was clear from sequence analysis that, if they were expressed, they could not be functional trypsinogens. They do not possess the key catalytic serine C195, which rules them out as serine proteases. They have no significant identity at the untranslated nucleotide level, and are identifiable as trypsinogen descendants only through consideration of their hypothetical translations. For these reasons, they were judged to be pseudogenes or relics, and assigned a trypsinogen isozyme label consistent with their position in the trypsinogen/TCR locus (Rowen et al., 1996).

However, recent additions to the EST database have included sequences corresponding to both human and mouse T1. Their hypothetically spliced and translated genomic sequences are 66% identical at the amino acid level and align with no indels. The mouse and human proteins are clearly orthologous. They are highly conserved and are likely to have an important function. At this point, there is not enough data to speculate on what this function might be. Furthermore, since these genes have a function other than tryptic activity, a re-evaluation of the nomenclature will be necessary. It will no longer appropriate to refer to them as "trypsinogen" T1.

The creation of novel function in "trypsinogen" T1 is a clear demonstration of the creation of a new gene from a multigene family. Other cases may exist. For example, although not trypsinogens, both the *Dissostichus* and *Pleuronectes* "trypsinogens" most likely are direct descendants from an ancestral trypsinogen gene. The divergence of these cold-adapted fish trypsinogens may have occurred during vertebrate evolution, but the date of this event cannot be determined. There are several other vertebrate-specific serine proteases that closely resemble trypsin. These include kallikrein and prostate-specific antigen. These proteins may have diverged from the trypsins during the course of vertebrate evolution. Considering how little is known of vertebrate genomes, the trypsinogen genes could conceivably have spawned many novel vertebrate-specific genes.

The trypsinogens may even be grandparents, in the sense that a gene derived from a trypsinogen ancestor has in turn become ancestral to another novel gene. The antifreeze

gene found in an Antarctic fish, *Dissostichus mawsoni* evolved by replacing the center region of a putative trypsinogen gene with a repeated nine-nucleotide element encoding hydrophobic residues (Chen et al., 1997; Logsdon and Doolittle, 1997). It retains the trypsinogen signal sequence and 3' untranslated region. The resulting gene bears no resemblance to a serine protease, but clearly evolved from the putative trypsinogen gene. However, as discussed above, the putative *Dissostichus* trypsinogen gene is a homolog of the *Pleuronectes* "trypsinogen," and is not a true trypsinogen. However, it is likely to have a trypsinogen gene as its ancestor, thus participating as both a child and a parent in a multi-generational spectrum of trypsinogen-derived genes.

Evidence suggests that there may have been a major phase of gene duplication at the dawn of vertebrate evolution (Holland and Garcia-Fernández, 1996). The trypsinogens may have participated in gene duplications at this time, and given rise to many of their apparent heirs, such as "trypsinogen" T1, kallikrein, prostate-specific antigen, and *Pleuronectes* "trypsinogen." If this were the case, one would expect all vertebrates to have essentially the same complement of genes. This hypothesis can be tested when complete vertebrate genomes are available from species representing a variety of vertebrate classes.

The interest and value of studying multigene family evolution is likely to grow as more sequences become available. Studying such families will reveal much about the dynamics and mechanisms of evolution.

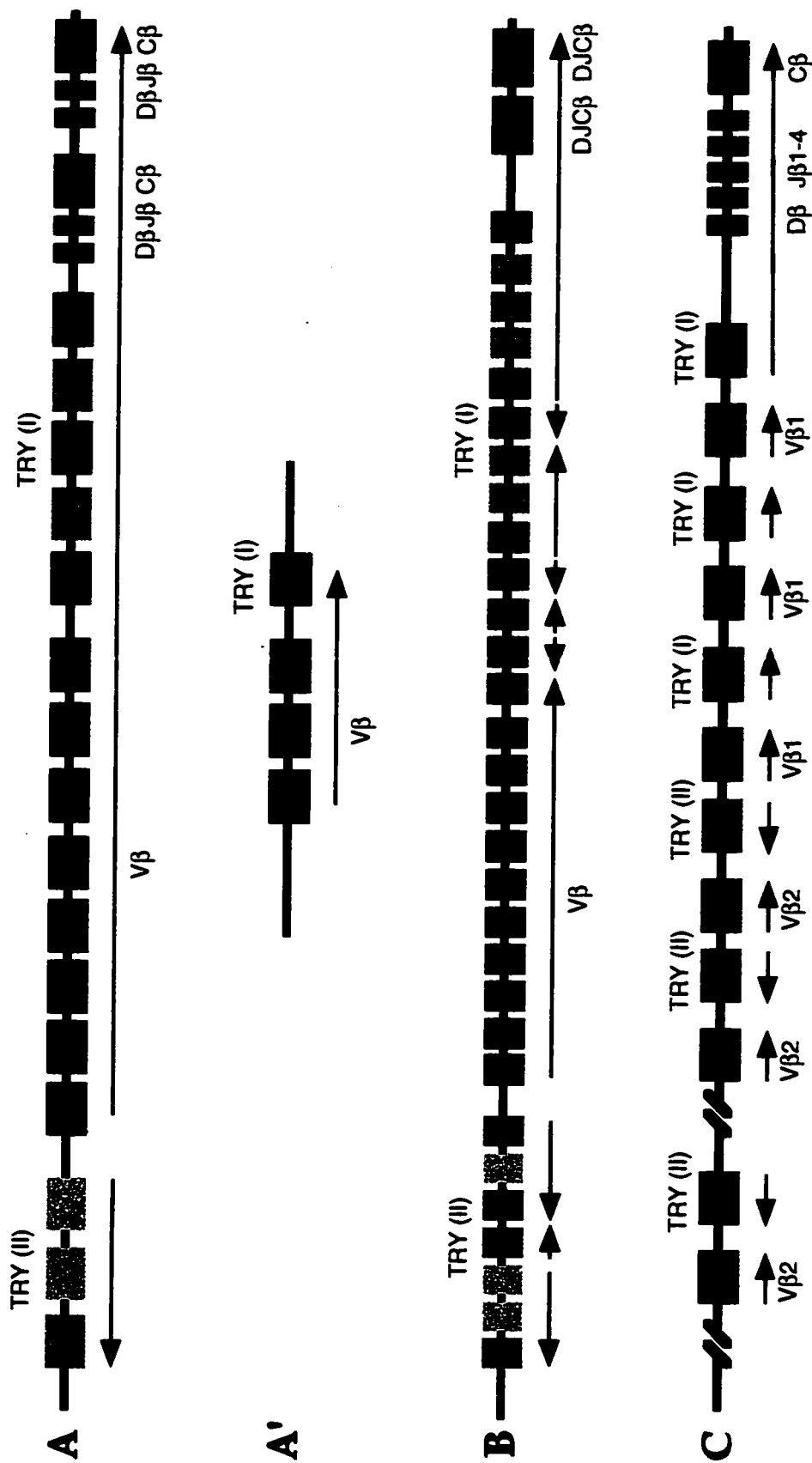


Figure 3.1. Cartoons of the topology of genomic trypsinogen organization. A, human TCR/TRY locus; A', partial translocation of human TCR/TRY locus to chromosome 9; B, murine TCR/TRY locus; C, chicken TCR/TRY locus. Blue, group I trypsinogen; Light blue, group I pseudotrypsinogen; Red, group II trypsinogen; Pink, group II pseudotrypsinogen; Orange, novel gene descended from an ancestral group II trypsinogen; Green, TCR gene segments. Not all human and mouse Vβ gene segments are shown. Mapping of the chicken TCR/TRY locus is incomplete. Arrows indicate transcriptional direction. Cartoons are not to scale.

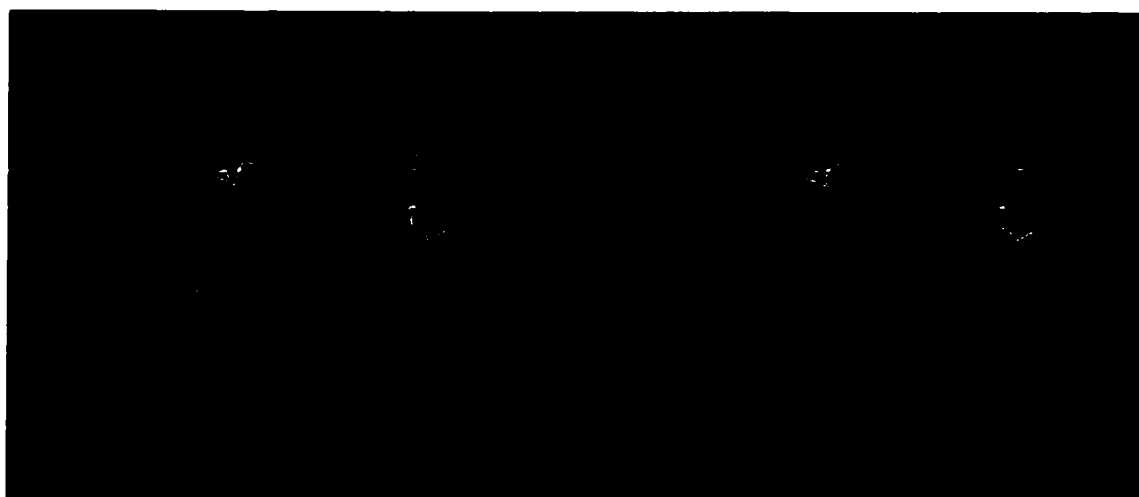


Figure 3.2. A stereoscopic view of the trypsin backbone. The *MAGE* program (David Richardson, author) was employed to visualize rat trypsin II (PDB designation 1ANE). The "IVGG" amino-terminus is tucked into the interior at the top of this view (residues 16-19 of 1ANE; orange). The three catalytic residues are shown in green (residues 57, 102, and 190), with the artificial substrate benzyldiamine (pink) in the active pocket. Cystine bridges are shown in yellow. A surface loop that has been subject to indels during vertebrate evolution is shown in magenta (residues C59-C62).

	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)	(31)	(32)	(33)	(34)	(35)	(36)	(37)	(38)	(39)	(40)	(41)	(42)	(43)	(44)	(45)	(46)	(47)	(48)	(49)	(50)	(51)	(52)	(53)	(54)	(55)	(56)	(57)	(58)	(59)	(60)	(61)	(62)	(63)	(64)	(65)	(66)	(67)	(68)	(69)	(70)	(71)	(72)	(73)	(74)	(75)	(76)	(77)	(78)	(79)	(80)	(81)	(82)	(83)	(84)	(85)	(86)	(87)	(88)	(89)	(90)	(91)	(92)	(93)	(94)	(95)	(96)	(97)	(98)	(99)	(100)
Greyfish	X	I	V	G	G	T	D	A	V	L	G	E	F	F	Y	Q	L	S	F	Q	E	F	S	F	H	F	C	G	A	S	I	Y	N	E	N	E	A	I	T	A	G	H	C	V	Y	G	D	S	G	L	Q	I	V	A	G	E	L	D	M	S	V	N	E	G	S	E	Q	T	I	T	V	S	K	I	L	H	E	N	F													
B. villosa	K	I	V	G	E	Q	A	G	A	I	P	Y	Q	A	R	L	Q	Y	S	A	G	S	I	R	C	G	S	L	I	S	E	T	Y	V	L	C	A	A	H	C	Q	S	A	W	K	I	V	L	G	L	Y	A	S	N	A	D	N	E	A	G	V	Q	T	F	N	V	N	A	Q	T	P	N	S	D	Y																	
B. schlosseri	K	I	I	G	S	S	A	S	N	G	Q	F	P	S	I	I	F	Q	K	S	G	S	F	F	C	G	T	I	T	P	N	R	V	L	S	A	A	B	C	E	Q	-	-	-	N	L	V	G	L	T	V	T	G	T	A	R	N	S	G	V	T	I	S	V	S	G	K	T	V	H	P	Q	Y																			
Lamprey A1	H	I	V	G	S	E	C	A	A	H	S	Q	P	W	Q	V	S	L	-	I	G	Y	H	F	C	G	S	L	I	S	Q	W	V	S	A	A	B	C	Y	Q	T	A	S	R	I	S	V	R	I	G	E	H	N	I	F	V	N	E	G	T	E	Q	I	Q	A	S	K	A	I	Q	H	P	Q	Y																		
Lamprey B1	H	I	V	G	E	C	A	A	H	S	Q	P	W	Q	V	S	L	-	I	G	Y	H	F	C	G	S	L	I	S	E	S	W	V	S	A	A	B	C	Y	Q	T	A	S	R	I	S	V	R	I	G	E	H	N	I	F	V	N	E	G	T	E	Q	I	Q	A	S	K	A	I	R	H	P	Q	Y																		
Dogfish	K	I	V	G	E	C	P	K	H	A	A	P	W	T	V	S	L	-	V	G	Y	H	F	C	G	S	L	I	A	P	G	W	V	S	A	A	B	C	Y	Q	-	-	R	R	I	Q	V	R	L	G	E	H	D	I	S	A	N	E	G	T	E	Y	I	D	S	S	M	V	I	R	H	P	N	Y																		
Pufferfish	K	I	V	G	E	C	R	K	N	S	V	A	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	V	N	E	N	W	V	S	A	A	B	C	Y	K	-	-	S	R	V	V	R	L	G	E	H	I	R	V	N	E	G	T	E	Q	F	I	S	S	S	R	V	I	R	H	P	N	Y																			
Cod I	K	I	V	G	E	C	T	K	H	S	Q	A	H	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	V	S	K	D	W	V	S	A	A	B	C	Y	K	-	-	S	V	L	R	V	R	L	G	E	H	I	R	V	N	E	G	T	E	Q	F	I	S	S	S	V	I	R	H	P	N	Y																			
Cod X	K	I	V	G	E	C	T	R	H	S	Q	A	H	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	V	S	K	D	W	V	S	A	A	B	C	Y	K	-	-	S	V	L	R	V	R	L	G	E	H	I	R	V	N	E	G	T	E	Q	F	I	S	S	S	V	I	R	H	P	N	Y																			
P. magellanicus	K	I	V	G	E	C	S	P	S	Q	P	H	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	V	N	E	N	W	V	S	A	A	B	C	Y	K	-	-	S	R	V	E	V	R	M	G	E	H	I	R	V	T	E	G	E	Q	F	I	S	S	S	R	V	I	R	H	P	N	Y																				
Salmon I	K	I	V	G	E	C	K	A	Y	S	Q	T	H	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	V	N	E	N	W	V	S	A	A	B	C	Y	K	-	-	S	R	V	E	V	R	L	G	E	H	I	K	V	T	E	G	S	E	Q	F	I	S	S	S	R	V	I	R	H	P	N	Y																		
Salmon II	K	I	V	G	E	C	K	A	Y	S	Q	P	H	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	V	N	E	N	W	V	S	A	A	B	C	Y	K	-	-	S	R	V	E	V	R	L	G	E	H	I	K	V	T	E	G	S	E	Q	F	I	S	S	S	R	V	I	R	H	P	N	Y																		
Salmon III	K	I	V	G	E	C	R	K	N	S	A	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	S	T	W	V	S	A	A	B	C	Y	K	-	-	S	R	I	Q	V	R	L	G	E	H	N	I	Q	V	T	E	G	S	E	Q	F	I	D	S	V	K	I	M	H	P	S	Y																				
Chicken P1	K	I	V	G	Y	T	C	P	E	H	S	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	S	Q	W	V	L	S	A	A	B	C	Y	K	-	-	S	S	I	Q	V	K	L	G	E	Y	N	L	A	Q	D	S	E	Q	T	I	S	S	S	K	V	I	R	H	S	G	Y																		
Chicken P29	K	I	V	G	Y	T	C	P	E	H	S	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	S	Q	W	V	L	S	A	A	B	C	Y	K	-	-	S	S	I	Q	V	R	L	G	E	Y	N	I	D	V	Q	E	D	S	E	V	R	S	S	V	I	R	H	P	K	Y																			
Chicken P38	K	I	V	G	Y	T	C	P	E	H	S	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	S	Q	W	V	L	S	A	A	B	C	Y	K	-	-	S	S	I	Q	V	K	L	G	E	Y	N	L	A	Q	D	S	E	Q	T	I	S	S	S	K	V	I	R	H	S	G	Y																		
Xenopus I	K	I	V	G	G	T	C	A	K	N	A	V	P	Y	Q	V	S	L	-	A	G	Y	H	F	C	G	S	L	I	S	Q	W	V	S	A	A	B	C	Y	K	-	-	S	R	I	Q	V	R	L	G	E	H	N	I	A	L	N	E	G	T	E	Q	F	I	D	S	K	V	I	K	H	P	N	Y																		
Xenopus II	K	I	G	G	A	T	C	A	K	S	S	V	P	Y	I	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	T	Q	W	V	S	A	A	B	C	Y	K	-	-	A	S	I	Q	V	R	L	G	E	H	N	I	A	L	S	E	G	T	E	Q	F	I	S	S	S	K	V	I	R	H	S	G	Y																	
Dog A	K	I	V	G	Y	T	C	E	N	S	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	S	D	Q	W	V	S	A	A	B	C	Y	K	-	-	S	R	I	Q	V	R	L	G	E	Y	N	I	D	V	L	E	G	N	E	Q	F	I	N	S	A	K	V	I	R	H	P	N	Y																	
Dog C	K	I	V	G	Y	T	C	S	R	N	S	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	S	D	Q	W	V	S	A	A	B	C	Y	K	-	-	S	R	I	Q	V	R	L	G	E	Y	N	I	A	V	S	E	G	G	E	Q	F	I	N	A	A	K	I	R	H	P	R	Y																	
Rat I	K	I	V	G	Y	T	C	P	E	H	S	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	N	D	Q	W	V	S	A	A	B	C	Y	K	-	-	S	R	I	Q	V	R	L	G	E	H	N	I	N	V	L	E	G	E	Q	F	I	N	A	A	K	I	K	H	P	N	Y																		
Rat II	K	I	V	G	Y	T	C	E	N	S	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	N	D	Q	W	V	S	A	A	B	C	Y	K	-	-	S	R	I	Q	V	R	L	G	E	H	N	I	N	V	L	E	G	N	E	Q	F	I	N	A	A	K	I	K	H	P	N	Y																		
Rat C	K	I	V	G	Y	T	C	E	N	S	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	N	D	Q	W	V	S	A	A	B	C	Y	K	-	-	S	R	I	Q	V	R	L	G	E	H	N	I	N	V	L	E	G	N	E	Q	F	I	N	A	A	K	I	K	H	P	N	Y																		
Rat C	K	I	V	G	Y	T	C	E	N	S	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	N	D	Q	W	V	S	A	A	B	C	Y	K	-	-	S	R	I	Q	V	R	L	G	E	H	N	I	N	V	L	E	G	N	E	Q	F	I	N	A	A	K	I	K	H	P	N	Y																		
Rat IV	K	I	V	G	Y	T	C	P	K	H	L	V	P	Y	Q	V	S	L	-	A	G	Y	H	F	C	G	S	L	I	S	D	Q	W	V	S	A	A	B	C	Y	K	-	-	S	R	I	Q	V	R	L	G	E	H	N	I	D	V	E	G	G	E	Q	F	I	D	A	A	K	I	R	H	P	S	Y																		
Rat V	R	I	V	G	Y	T	C	E	H	S	V	P	Y	Q	V	S	L	-	A	G	S	H	I	C	G	S	L	I	T	D	Q	W	V	L	S	A	A	B	C	Y	H	-	-	P	Q	L	Q	V	R	L	G	E	H	N	I	H	V	L	E	G	G	E	Q	F	I	D	A	E	K	I	R	H	P	E	Y																	
Bovine A	K	I	V	G	Y	T	C	A	E	N	S	V	P	Y	Q	V	S	L	-	A	G	Y	H	F	C	G	S	L	I	N	D	Q	W	V	S	A	A	B	C	Y	K	-	-	S	G	I	Q	V	R	L	G	E	H	N	I	H	V	L	E	G	G	E	Q	F	I	D	A	S	K	I	R	H	P	K	Y																	
Bovine C	K	I	V	G	Y	T	C	G	A	N	T	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	N	D	Q	W	V	S	A	A	B	C	Y	K	-	-	S	G	I	Q	V	R	L	G	E	H	N	I	H	V	L	E	G	G	E	Q	F	I	D	A	S	K	I	R	H	P	K	Y																	
Pig	K	I	V	G	Y	T	C	A	A	N	S	I	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	N	S	Q	W	V	S	A	A	B	C	Y	K	-	-	S	G	I	Q	V	R	L	G	E	H	N	I	N	V	L	E	G	N	E	Q	F	I	S	A	S	K	S	I	V	H	P	S	Y																
Human T4	K	I	V	G	Y	N	C	E	E	N	S	V	P	Y	Q	V	S	L	-	S	G	Y	H	F	C	G	S	L	I	N	S	Q	W	V	S	A	A	B	C	Y	K	-	-	S	R	I	Q	V	R	L	G																																									

	22	8	10	120	(127)	128	130	146	(157)	150	(168)	160																																																												
Greyfish	DYDLLDND	I	S	L	L	K	S	G	S	L	T	F	N	N	N	V	A	I	A	L	P	A	Q	G	H	T	A	T	G	N	V	I	V	T	G	W	G	T	T	S	-	E	G	G	N	T	P	D	V	L	Q	K	V	T	P	L	V	S	D	A	E	C	R	D	D	Y	G	D	E	I		
B. villosa	DSATTDND	V	M	L	L	R	D	E	S	A	T	L	T	S	S	V	A	L	V	S	L	T	S	F	E	E	D	T	A	C	T	V	S	G	W	G	T	T	S	-	S	G	G	T	I	S	D	Y	L	M	K	V	E	N	V	D	Q	D	E	C	G	N	R	Y	G	S	L	T				
B. schlosseri	NSNTIQND	I	M	I	L	N	L	A	S	S	F	S	S	T	I	A	A	P	L	A	S	S	P	S	V	G	T	E	S	P	L	P	D	G	A	I	P	S	A	G	I	V	S	N	L	P	N	L	Q	Y	V	N	V	N	I	S	T	A	T	C	N	S	R	I	N	G	A	I				
Lamprey A1	NSWTIDND	I	M	L	I	K	L	S	S	P	A	T	L	N	Q	Y	A	Q	A	I	A	L	P	S	S	C	V	N	T	G	M	C	T	I	S	G	W	E	T	Q	-	T	S	V	G	S	P	D	V	L	M	C	V	A	P	V	L	S	D	T	S	C	R	N	S	Y	P	G	D	I		
Lamprey B1	SSATIDND	I	M	L	I	K	L	S	S	P	A	T	L	N	Q	Y	A	Q	A	V	P	L	P	S	S	C	V	G	T	G	M	C	T	I	S	G	W	E	T	Q	-	T	S	V	G	S	P	D	V	L	M	C	V	A	P	V	L	S	D	T	S	C	R	N	S	Y	P	G	D	I		
Dogfish	SGYDLND	I	M	L	I	K	L	S	K	P	A	L	N	R	N	V	D	L	I	S	L	P	T	G	C	A	Y	A	G	E	M	C	L	I	S	G	W	G	N	T	M	-	D	G	A	V	S	G	D	Q	L	Q	C	L	D	A	P	V	L	S	D	A	E	C	K	G	A	Y	P	G	M	I
Pufferfish	SSYNIDND	I	M	L	I	K	L	S	K	P	A	T	L	N	Q	Y	V	P	V	A	L	P	S	S	C	A	A	G	T	M	C	K	V	S	G	W	G	N	T	M	-	S	S	T	A	D	R	N	K	L	Q	C	L	N	I	P	L	S	D	R	D	C	E	N	S	Y	P	G	M	I		
Cod I	SSYNINND	I	M	L	I	K	L	T	K	P	A	T	L	N	Q	Y	V	H	A	V	A	L	P	T	E	C	A	A	D	A	T	M	C	T	V	S	G	W	G	N	T	M	-	S	S	V	A	D	G	D	K	L	Q	C	L	S	L	P	L	S	H	A	D	C	A	N	S	Y	P	G	M	I
Cod X	SSYNIDND	I	M	L	I	K	L	T	E	P	A	T	L	N	Q	Y	V	H	A	V	A	L	P	T	E	C	A	A	D	A	T	M	C	T	V	S	G	W	G	N	T	M	-	S	S	V	D	D	G	D	K	L	Q	C	L	N	L	P	L	S	H	A	D	C	A	N	S	Y	P	G	M	I
P. magellanicus	SSYNIDND	I	M	L	I	K	L	S	K	P	A	T	L	N	Q	Y	V	A	V	A	L	P	S	S	C	A	P	A	G	T	M	C	T	V	S	G	W	G	T	Q	-	S	S	A	D	G	N	K	L	Q	C	L	N	I	P	L	S	D	R	D	C	D	N	S	Y	P	G	M	I			
Salmon I	SSYNIDND	I	M	L	I	K	L	S	K	P	A	T	L	N	Q	Y	V	P	V	A	L	P	T	S	C	A	P	A	G	T	M	C	T	V	S	G	W	G	N	T	M	-	S	S	T	A	D	S	N	K	L	Q	C	L	N	I	P	L	S	Y	S	D	C	N	S	Y	P	G	M	I		
Salmon II	SSYNIDND	I	M	L	I	K	L	S	K	P	A	T	L	N	Q	Y	V	P	V	A	L	P	T	S	C	A	P	A	G	T	M	C	T	V	S	G	W	G	N	T	M	-	S	S	T	A	D	S	N	K	L	Q	C	L	N	I	P	L	S	Y	S	D	C	N	S	Y	P	G	M	I		
Salmon III	SSYNIDND	I	M	L	I	K	L	S	K	P	A	T	L	N	Q	Y	V	P	V	A	L	P	T	S	C	A	P	A	G	T	M	C	T	V	S	G	W	G	N	T	M	-	S	S	T	A	D	S	N	K	L	Q	C	L	N	I	P	L	S	Y	S	D	C	N	S	Y	P	G	M	I		
Chicken P1	NSRNLND	I	M	L	I	K	L	S	K	P	A	T	L	N	Q	Y	V	T	V	A	L	P	S	S	C	A	S	S	G	T	R	C	L	V	S	G	W	G	N	T	M	-	S	S	T	A	D	S	N	K	L	Q	C	L	N	I	P	L	S	Y	S	D	C	N	S	Y	P	G	M	I		
Chicken P29	NANTLNND	I	M	L	I	K	L	S	K	A	A	T	L	N	S	Y	V	N	T	V	P	L	P	T	S	C	V	A	G	T	T	C	L	I	S	G	W	G	N	T	M	-	S	S	L	Y	P	D	V	L	Q	C	L	N	A	P	V	L	S	S	Q	C	S	A	Y	P	G	R	I			
Chicken P38	SITLNND	I	M	L	I	K	L	S	A	V	E	S	A	D	I	Q	P	I	A	L	P	S	S	C	A	K	A	G	T	E	C	L	I	S	G	W	G	N	T	M	-	S	S	L	Y	P	D	V	L	Q	C	L	N	A	P	V	L	S	D	Q	E	C	A	Y	P	G	R	I				
Xenopus I	NANTLNND	I	M	L	I	K	L	S	K	A	A	T	L	N	S	Y	V	N	T	V	P	L	P	T	S	C	V	A	G	T	T	C	L	I	S	G	W	G	N	T	M	-	S	S	L	Y	P	D	V	L	Q	C	L	N	A	P	V	L	S	S	Q	C	S	A	Y	P	G	R	I			
Xenopus II	NSRNLND	I	M	L	I	K	L	S	T	T	A	R	L	S	A	N	I	Q	S	V	P	L	P	S	A	C	A	S	A	G	T	N	C	L	I	S	G	W	G	N	T	M	-	S	S	L	Y	P	D	V	L	Q	C	L	N	A	P	I	L	T	D	S	Q	C	N	S	Y	P	G	E	I	
Dog A	NSYTLND	I	M	L	I	K	L	S	S	P	A	S	L	N	A	A	V	N	T	V	P	L	P	S	G	C	S	A	G	T	S	C	L	I	S	G	W	G	N	T	M	-	S	S	G	N	Y	P	D	L	L	Q	C	L	N	A	P	I	L	T	N	A	Q	C	N	S	A	Y	P	G	E	I
Dog C	NSWILDND	I	M	L	I	K	L	S	S	P	A	V	L	N	A	R	V	A	T	I	S	L	P	R	A	C	A	A	G	T	Q	C	L	I	S	G	W	G	N	T	M	-	S	S	G	T	N	Y	P	D	L	L	Q	C	L	N	A	P	I	L	T	Q	A	C	E	A	S	Y	P	G	I	
Rat I	NANTIDND	I	M	L	I	K	L	S	S	P	A	V	L	N	R	V	A	I	A	L	P	K	S	C	P	A	A	G	T	Q	C	L	I	S	G	W	G	N	T	M	-	S	S	I	G	N	Y	P	D	V	L	Q	C	L	K	A	P	I	L	S	D	S	V	C	R	N	A	Y	P	G	I	
Rat II	SSWTLNND	I	M	L	I	K	L	S	S	P	V	K	L	N	R	V	A	P	V	A	L	P	S	A	C	A	P	A	G	T	Q	C	L	I	S	G	W	G	N	T	M	-	S	S	G	N	Y	P	D	L	L	Q	C	V	D	A	P	V	L	S	Q	A	D	C	E	A	Y	P	G	E	I	
Rat C	DRKTLNND	I	M	L	I	K	L	S	S	P	V	K	L	N	R	V	A	T	V	A	L	P	S	S	C	A	P	A	G	T	Q	C	L	I	S	G	W	G	N	T	M	-	S	S	G	V	N	E	P	D	L	L	Q	C	D	A	P	L	P	Q	A	D	C	E	A	Y	P	G	E	I		
Rat IV	NANTFDND	I	M	L	I	K	L	N	S	P	A	T	L	N	S	R	V	S	T	V	S	L	P	R	S	C	G	S	S	G	T	K	C	L	V	S	G	W	G	N	T	M	-	S	S	G	T	N	Y	P	S	L	L	Q	C	D	A	P	V	L	S	S	S	Q	C	S	Y	P	G	E	I	
Rat V	NKDTLDND	I	M	L	I	K	L	S	P	A	V	L	N	S	Q	V	S	T	V	S	L	P	R	S	C	A	S	T	D	A	Q	C	L	V	S	G	W	G	N	T	M	-	S	I	G	K	Y	P	A	L	L	Q	C	E	A	P	V	L	S	A	S	C	K	S	Y	P	G	E	I			
Bovine A	DKWTVND	I	M	L	I	K	L	S	P	A	T	L	N	S	K	V	S	T	I	P	L	P	Q	Y	C	P	T	A	G	T	E	C	L	V	S	G	W	G	-	V	L	K	F	G	F	E	S	P	V	L	Q	C	D	A	P	V	L	S	D	S	V	C	H	K	A	Y	P	R	Q	I		
Bovine C	SSWTLND	I	M	L	I	K	L	S	T	P	A	V	I	N	A	R	V	S	T	L	L	P	S	A	C	A	S	A	G	T	E	C	L	I	S	G	W	G	N	T	M	-	S	S	G	V	N	Y	P	D	L	L	Q	C	V	A	P	L	L	S	H	A	D	C	E	A	S	Y	P	G	E	I
Pig	NSWTLND	I	M	L	I	K	L	S	A	S	L	N	S	R	V	A	S	I	S	L	P	T	S	C	A	S	A	G	T	Q	C	L	I	S	G	W	G	N	T	M	-	S	S	G	T	S	P	D	V	L	K	C	L	A	P	I	L	S	D	S	S	C	K	A	Y	P	G	E	I			
Human T4	NGNTLDND	I	M	L	I	K	L	S	S	P	A	T	L	N	S	R	V	A	T	V	S	L	P	R	S	C	A	A	G	T	E	C	L	I	S	G	W	G	N	T	M	-	S	S	S	S	Y	P	S	L	L	Q	C	K	A	P	I	L	S	D	S	S	C	K	A	Y	P	G	E	I		
Human T6	DRKTLNND	I	M	L	I	K	L	S	S	R	A	V	I	N	R	V	S	T	I	S	L	P	T	A	P	P	A	G	T	K	C	L	I	S	G	W	G	N	T	M	-	S	S	A	D	Y	P	D	E	L	Q	C	D	A	P	V	L	S	Q	A	C	E	A	S	Y	P	G	E	I			
Human T8	NRITLNND	I	M	L	I	K	L	S	T	P	A	V	I	N	A	H	V	S	T	I	S	L	P	T	A	P	P	A	A	G	T	E	C	L	I	S	G	W	G	N	T	M	-	S	S	A	D	Y	P	D	E	L	Q	C	D	A	P	V	L	T	Q	A	C	K	A	S	Y	P	L	K	I	
Human T9	NSRTLND	I	M	L	I	K	L	S	S	P	A	V	I	N	R	V	S	A	I	S	L	P	T	A	P	P	A	A	G	T	E	S	L	I	S	G	W	G	N	T	M	-	S	S	A	D	Y	P	D	E	L	Q	C	D	A	P	V	L	S	Q	A	C	E	A	S	Y	P	G	E	I		
Mouse T20	NRDTLDND	I	M	L	I	K	L	S	S	P	A	V	I	N	R	V	S	T	I	S	L	P	T	A	P	P	A	A	G	T	E	C	L	I	S	G	W	G	N	T	M	-	S	S	A	D	Y	P	D																							

Grayfish	161	(182)	170	(189)	(191)	(192)	180	(193)	190	(201)	190	(217)	208	(220)	(227)	210	(232)	228								
<i>R. villosa</i>	FDSMICAGVPEGGK	DSCQ	GD	SG	GG	PLAA	TG	SL	AG	IV	SW	GG	CA	RP	GP	GV	Y	TE	VS	YH	VD	WI	K	--	NAV.	
<i>R. schlosseri</i>	GGMM-CL--	AS	GD	SG	GG	PA	VC	NG	VQ	YI	VS	WG	AG	CA	SV	LP	GV	Y	TR	VA	FT	WI	D	--	NMV.	
Lamprey A1	LSGMICGMNNGED	S	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	RG	CA	LP	NP	GP	Y	TK	VC	NY	NA	WI	A	Q	TIAAN.
Lamprey B1	TNNMICGLYLEGG	KDS	CQ	GD	SG	GP	VC	NG	Q	LQ	IV	SW	WG	RG	CA	LP	NP	GP	Y	TK	VC	NY	SW	IA	STMAAN.	
Dogfish	TNNMICGLYLEGG	KDS	CQ	GD	SG	GP	VC	NG	ML	QI	VS	WG	Y	GA	ER	HP	GP	Y	TR	VC	HY	SW	IH	ETIASV.		
Pufferfish	TNNMVCVGM	EGGKDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	LF	ND	WL	ETMASV.		
Cod I	TDAMFCAGYLEGG	KDS	CQ	GD	SG	GP	VC	NG	VL	Q	IV	SW	WG	Y	GA	ER	HP	GP	Y	AK	VC	LV	SG	WR	DTKANY.	
Cod X	TQSMFCAGYLEGG	KDS	CQ	GD	SG	GP	VC	NG	VL	Q	IV	SW	WG	Y	GA	ER	HP	GP	Y	AK	VC	LV	SG	WR	DTMASV.	
<i>P. magellanicus</i>	TQSMFCAGYLEGG	KDS	CQ	GD	SG	GP	VC	NG	VL	Q	IV	SW	WG	Y	GA	ER	HP	GP	Y	AK	VC	LV	SG	WR	DTMASV.	
Salmon I	TDAMFCAGYLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	LF	ND	WL	ETSMANY.		
Salmon II	TNAMFCAGYLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMASY.		
Salmon III	TNAMFCAGYLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMATY.		
Chicken P1	TSNMFCAGYLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Chicken P29	TSNMICGYLNGG	KDS	CQ	GD	SG	GP	VC	NG	Q	LQ	IV	SW	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.	
Chicken P38	TSNMICGYLNGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Xenopus I	TSNMICGYLNGG	KDS	CQ	GD	SG	GP	VC	NG	Q	LQ	IV	SW	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.	
Xenopus II	TKNMFCAFLAGG	KDS	CQ	GD	SG	GP	VC	NG	Q	LQ	IV	SW	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.	
Dog A	TANMICVGYMEGG	KDS	CQ	GD	SG	GP	VC	NG	Q	LQ	IV	SW	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.	
Dog C	TENMICAGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Rat I	SSMMCLGYMEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Rat II	TSSMICVGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Rat C	TDNMVCVGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Rat IV	TSNMFCLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Rat V	TSNMFCLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Bovine A	TNNMICAGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Bovine C	TNNMICAGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Pig	TNNMICAGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Human T4	TGNMICVGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Human T6	TNNMFCVGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Human T8	TSKMFCVGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Human T9	TNNMFCVGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
Mouse T20	TNSMFCVGFLEGG	KDS	CQ	RD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		
	TNNMICVGFLEGG	KDS	CQ	GD	SG	GP	VC	NG	EL	QI	VS	WG	Y	GA	ER	HP	GP	Y	AK	VC	IF	ND	WL	ETSMSSN.		

Figure 3.3. Trypsin multiple alignment (continued).

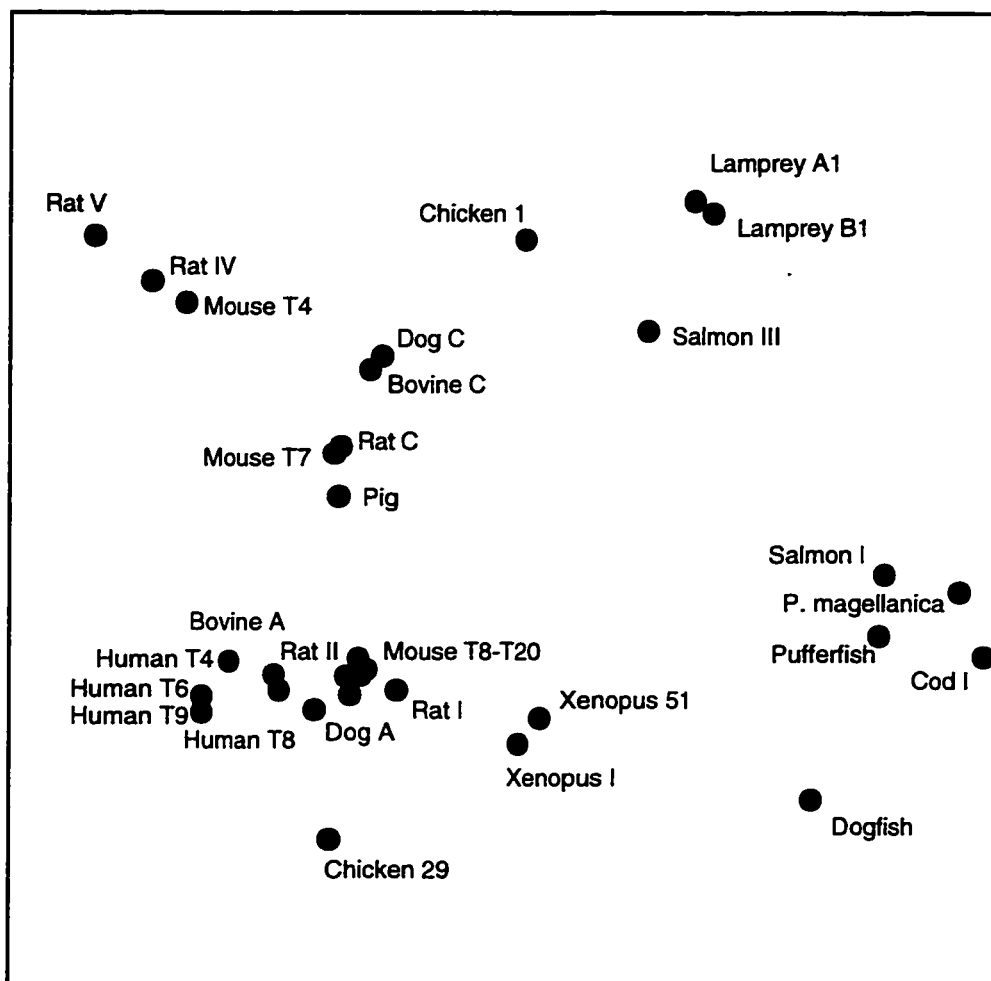


Figure 3.4. Multidimensionally scaled vertebrate trypsin sequence distances. Each point represents a trypsin sequence; The distance between two points corresponds to the calculated phylogenetic distance between the corresponding sequences. The program *SPSS*® 7.5 (SPSS, Inc.) was used to scale the trypsin pairwise distance matrix (from *Protdist*) as ratio data with a Euclidean distance model. Iterations continued until the S-stress altered by less than 0.0001 between iterations.

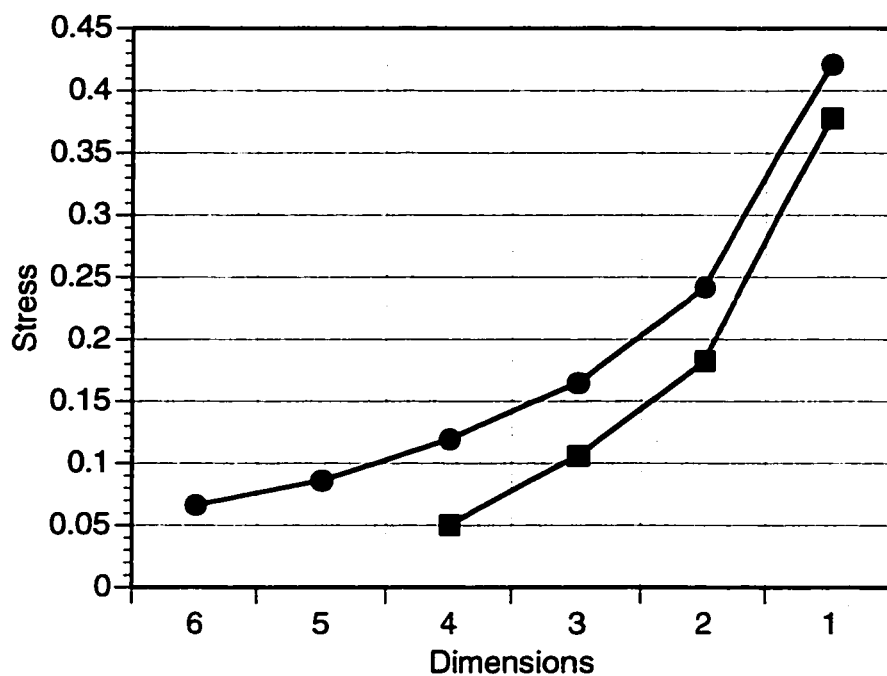


Figure 3.5. Stress vs. Dimensions. Data for Figure 3.4 (●) and Figure 3.12 (■) was sequentially multidimensionally scaled into successively fewer dimensions. The stress at each dimension is shown. There is a gradual rise in stress through two dimensions, with a larger rise occurring between two dimensions and one dimension, suggesting that two dimensions is sufficient to portray the data.

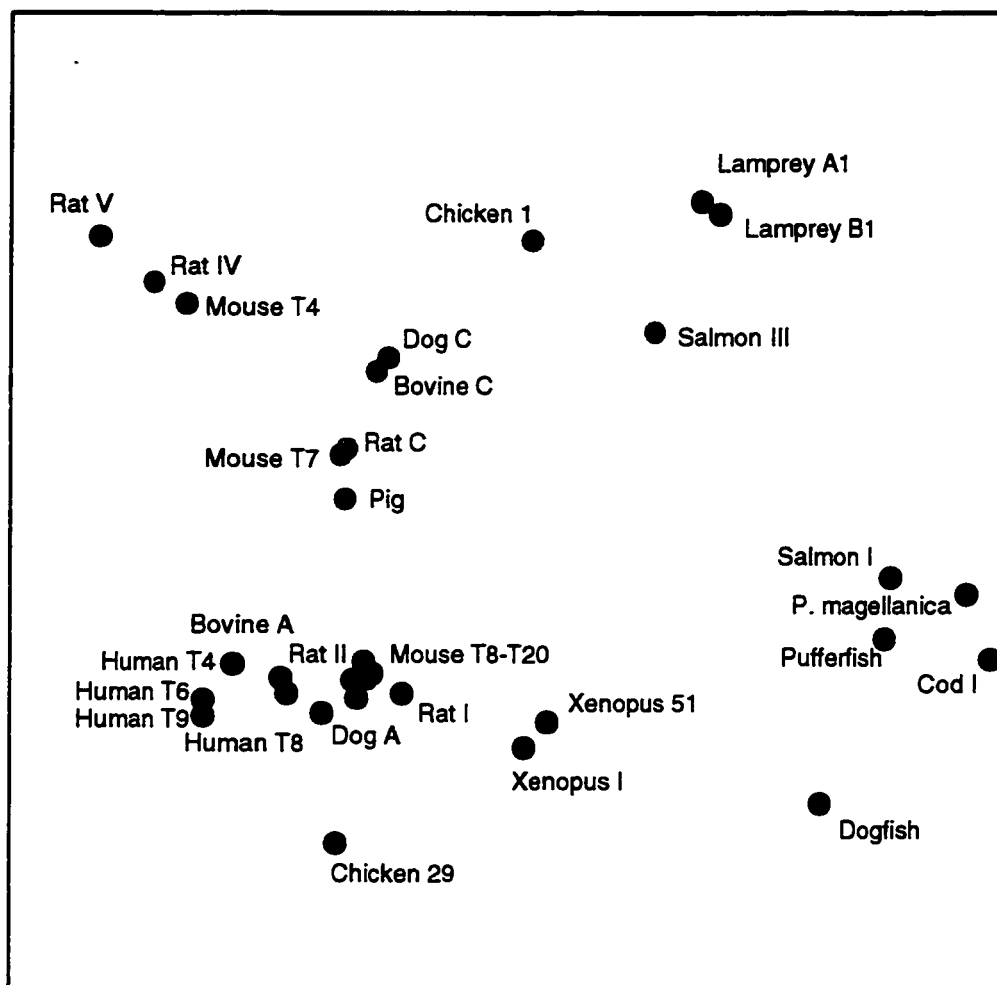


Figure 3.6. Group I vs. Group II. A multidimensionally scaled projection of the trypsin phylogenetic distances. Each point represents a trypsin sequence; The distance between two points corresponds to the calculated phylogenetic distance between the corresponding sequences. Group I trypsins are coded blue; Group II trypsins are coded red; the lamprey trypsins are green.

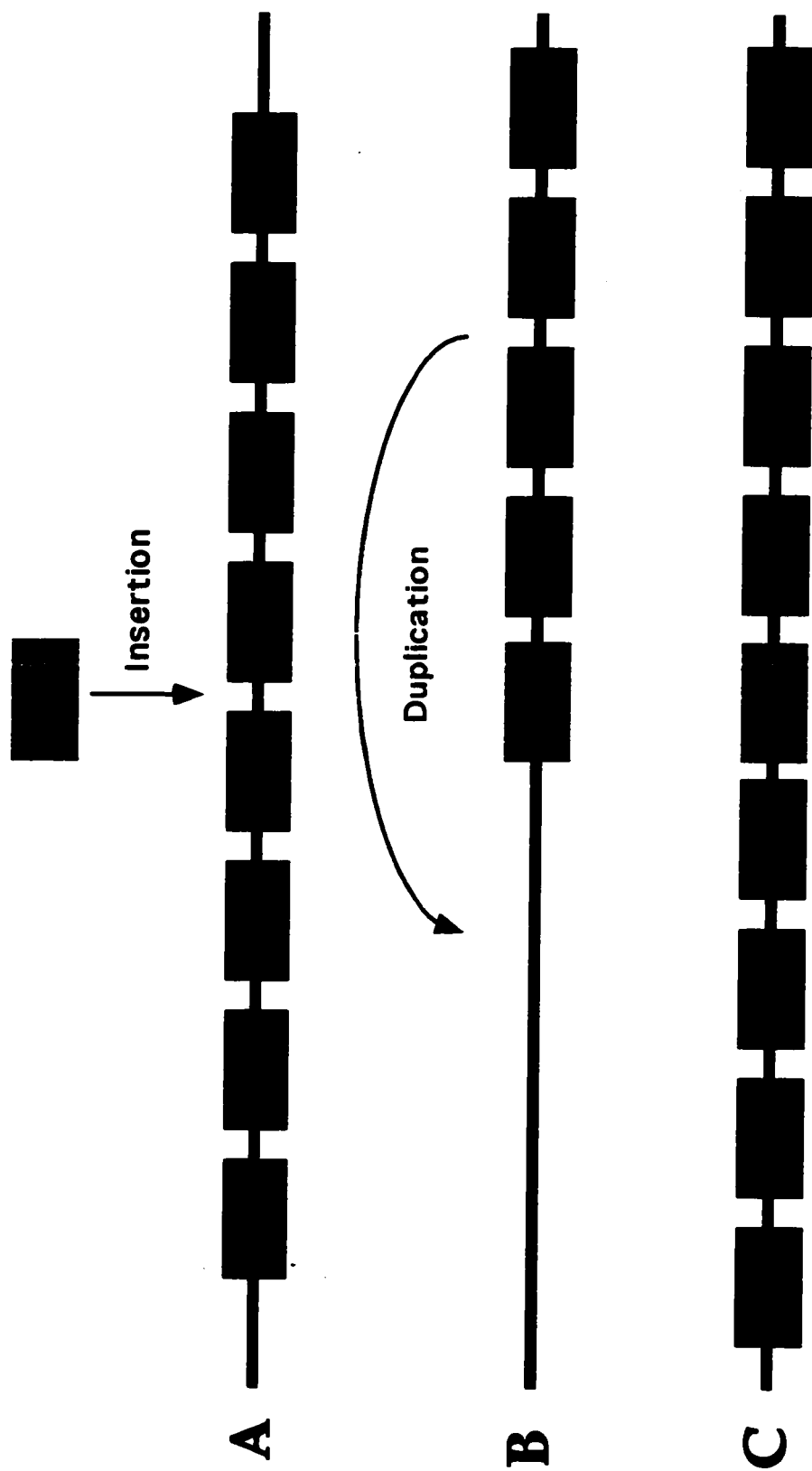


Figure 3.7. Allopatric division of the trypsinogen multigene family. Two general mechanisms can account for multigene allopatry. A, insertion. B, duplication; C, the resulting divided multigene family.

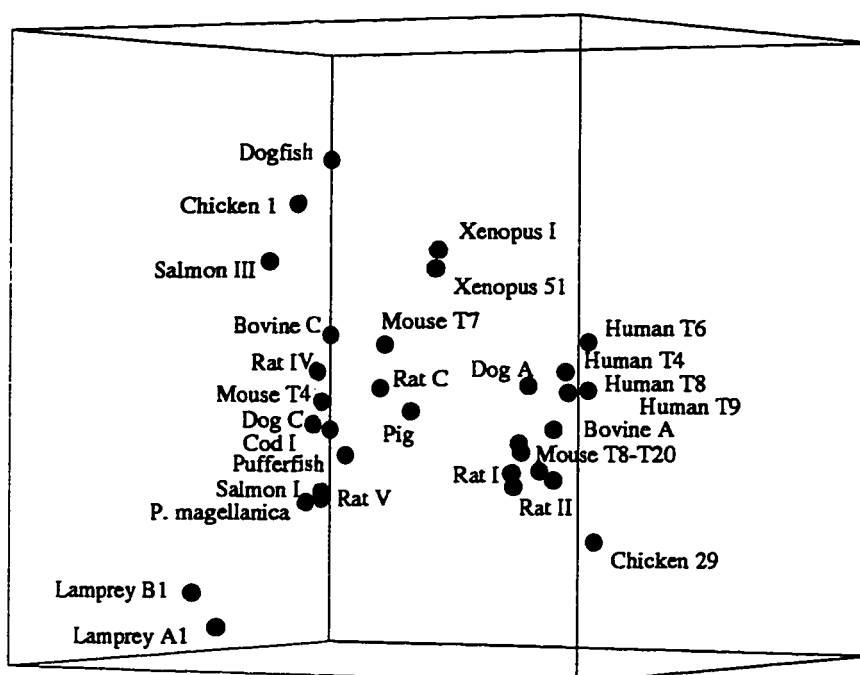


Figure 3.8. Group I vs. Group II. A shadow of a three-dimensional multidimensionally scaled projection of the trypsin phylogenetic distances. Group I trypsins are coded blue; Group II trypsins are coded red; the lamprey trypsins are green.

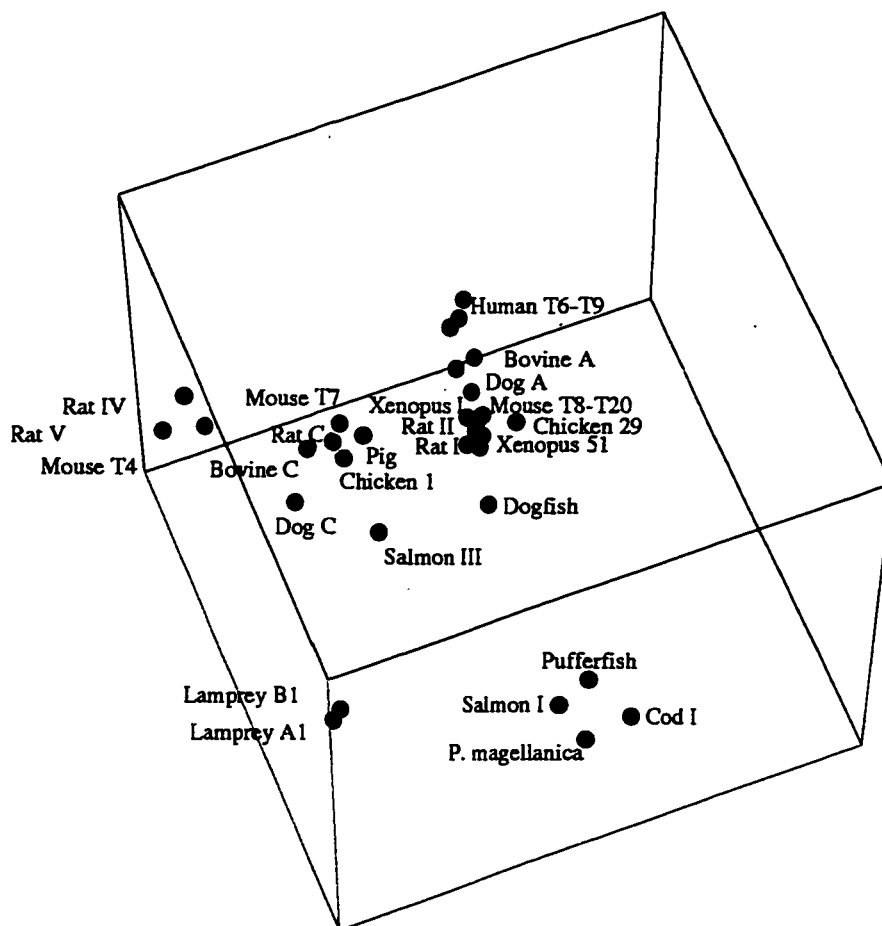


Figure 3.9. Group I vs. Group II. An alternative shadow of a three-dimensional multidimensionally scaled projection of the trypsin phylogenetic distances. Group I trypsins are coded blue; Group II trypsins are coded red; the lamprey trypsins are green.

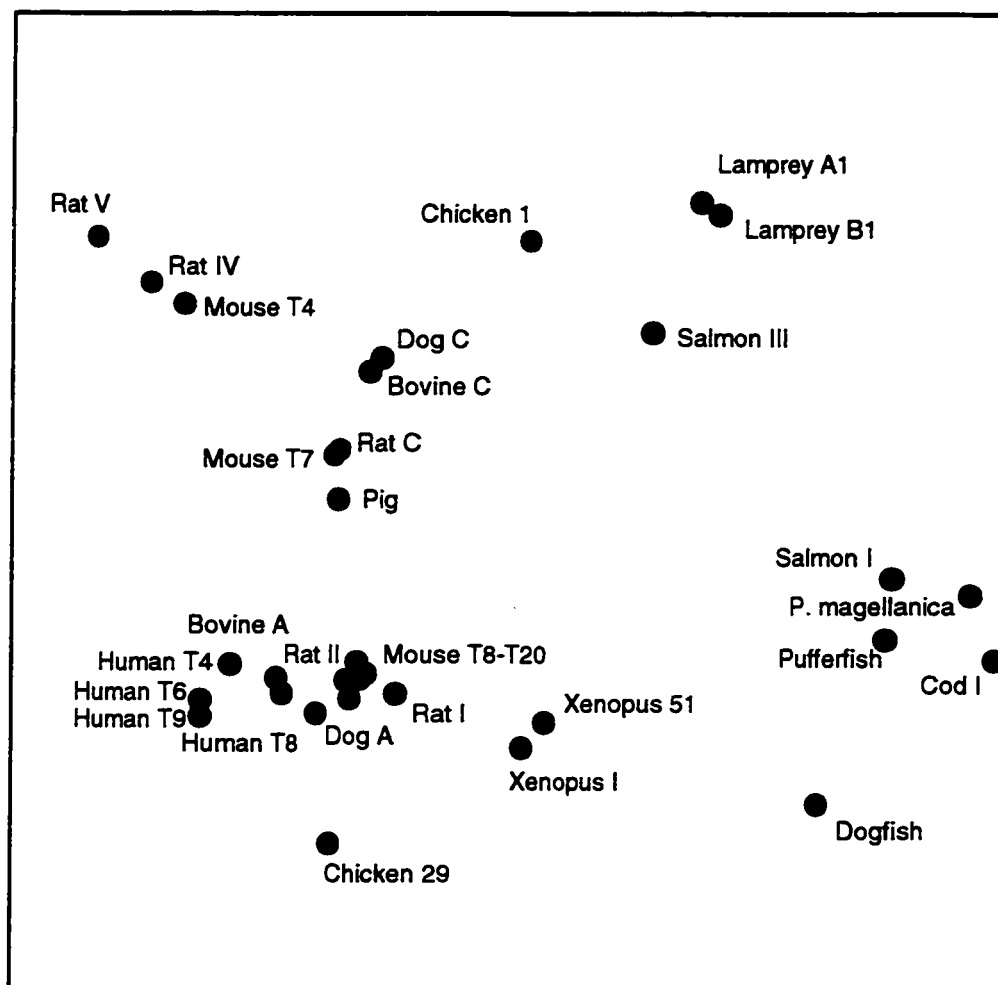


Figure 3.10. 5' vs. 3'. A multidimensionally scaled projection of the trypsin phylogenetic distances. Each point represents a trypsin sequence; The distance between two points corresponds to the calculated phylogenetic distance between the corresponding sequences. 3' trypsins are coded blue; 5' trypsins are coded red; sequences with unknown syntenic relationships are black.

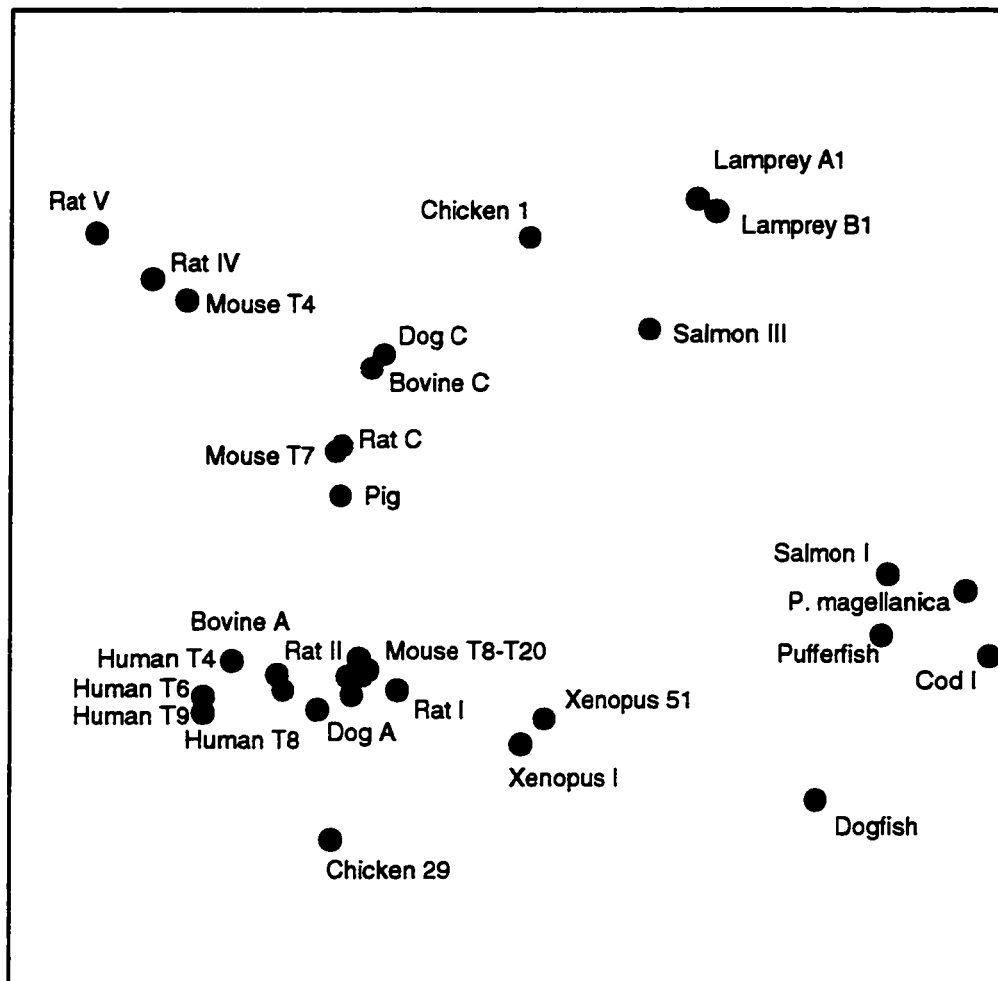


Figure 3.11. Anionic vs. Cationic. A multidimensionally scaled projection of the trypsin phylogenetic distances. Each point represents a trypsin sequence; The distance between two points corresponds to the calculated phylogenetic distance between the corresponding sequences. Anionic trypsins are coded blue; cationic trypsins are coded red.

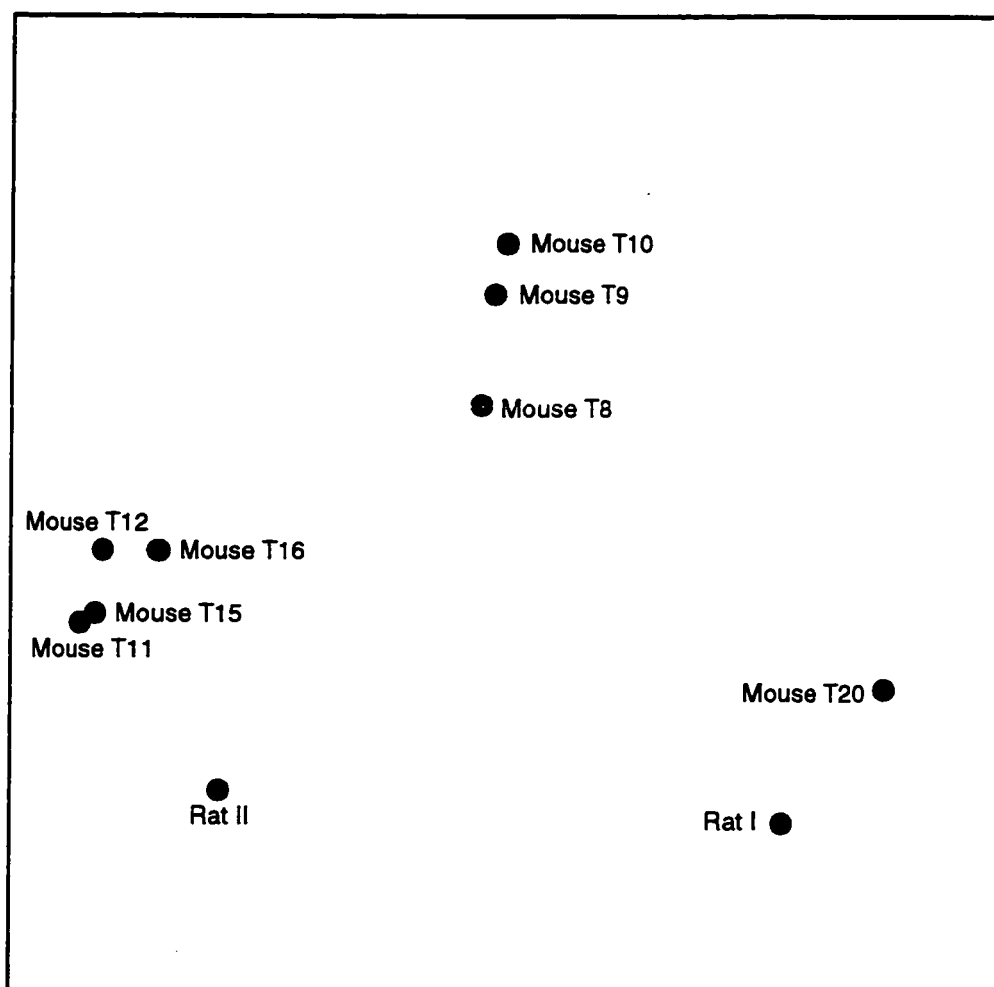


Figure 3.12. A multidimensionally scaled projection of the rodent group I trypsin phylogenetic distances. Each point represents a trypsin sequence; The distance between two points corresponds to the calculated phylogenetic distance between the corresponding sequences.

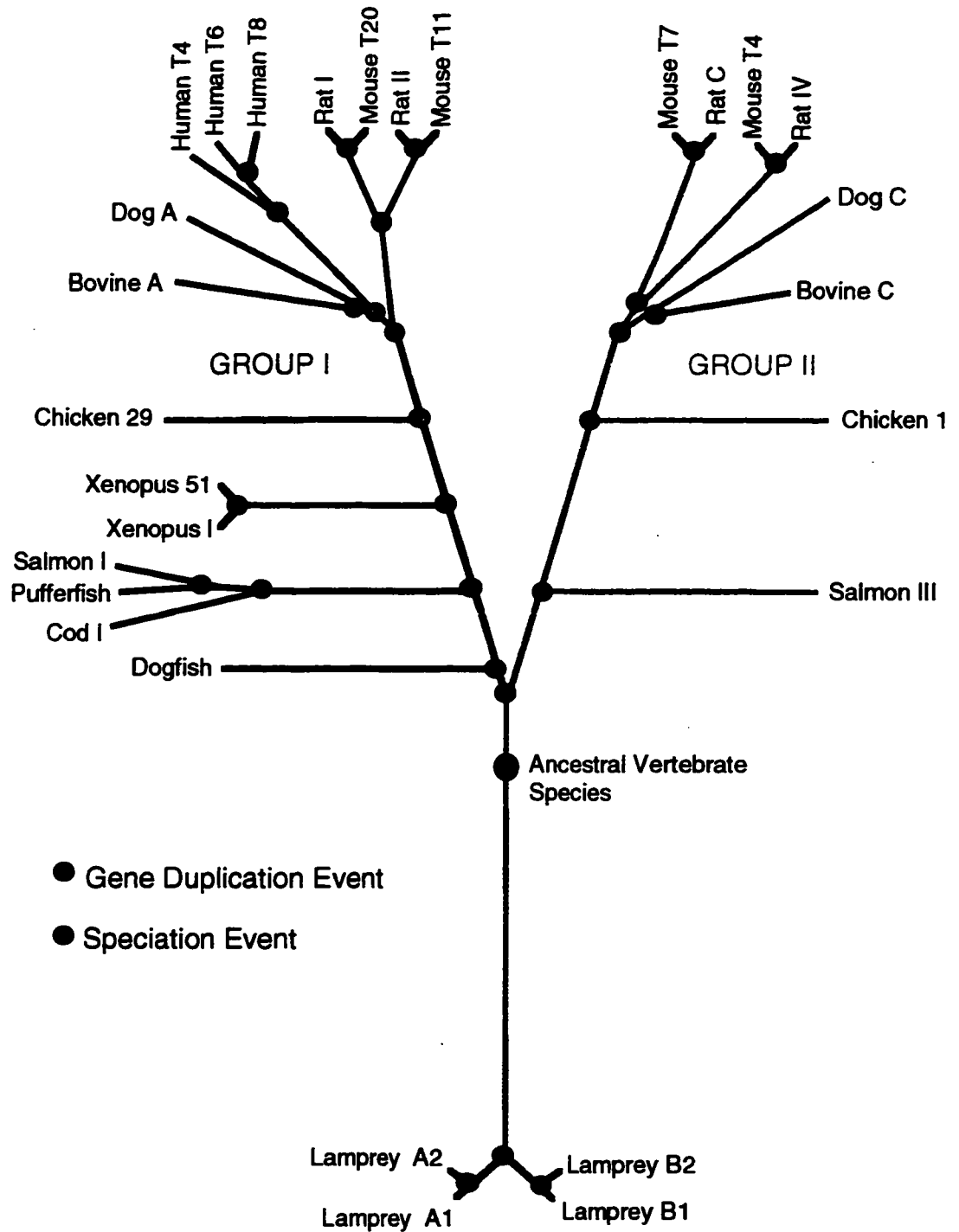


Figure 3.13. A hypothetical phylogeny of the vertebrate trypsinogens. Group I branches are blue; group II branches are red. Branches in green diverged before group I and group II split.

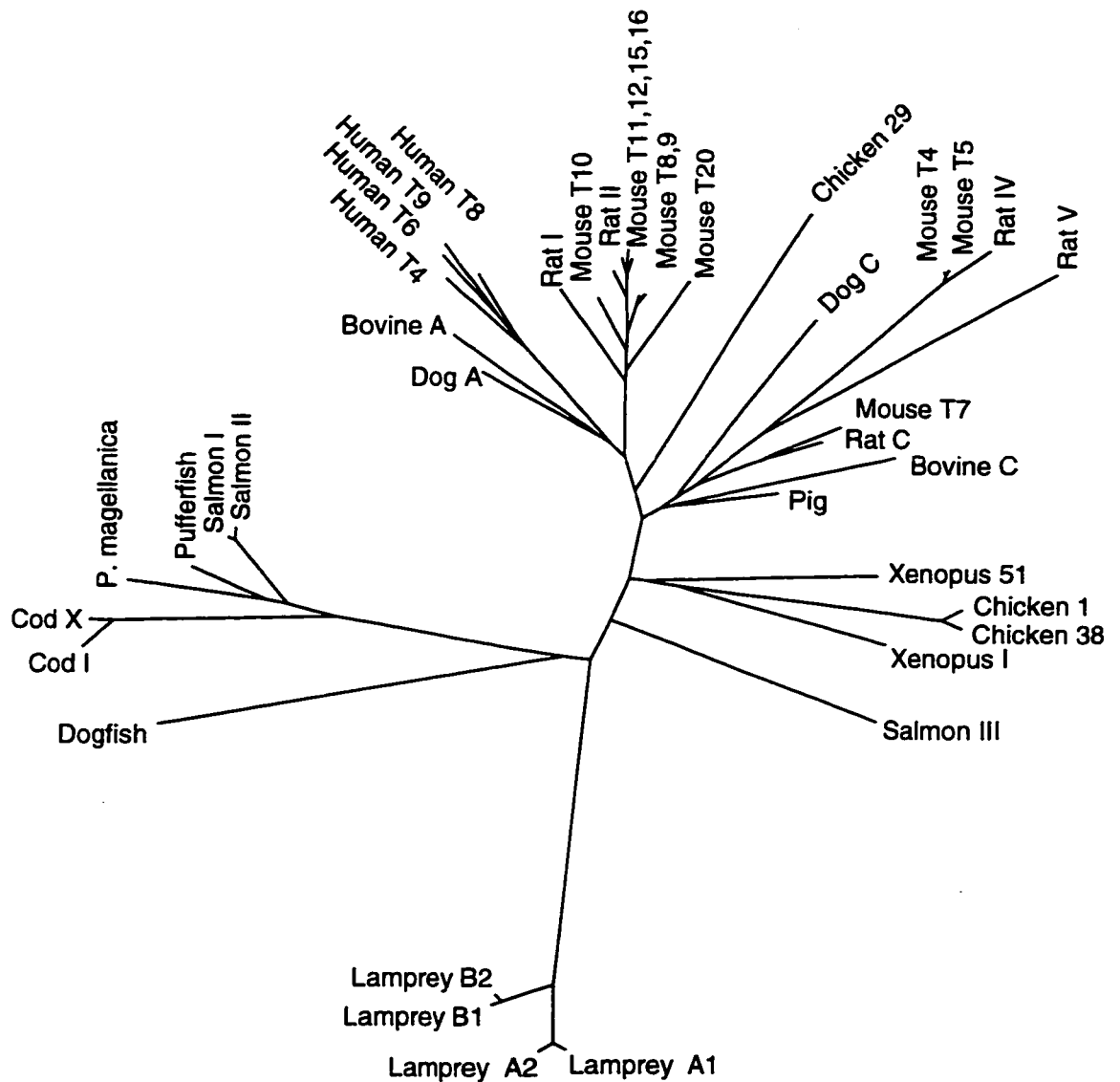


Figure 3.14. A Fitch-Margoliash phylogeny of forty-two vertebrate trypsinogens. Distances from the program *protdist*, with the Dayhoff matrix, were fed to the program *fit*ch, with global rearrangements and 20 random “jumbles” (Felsenstein, 1993).

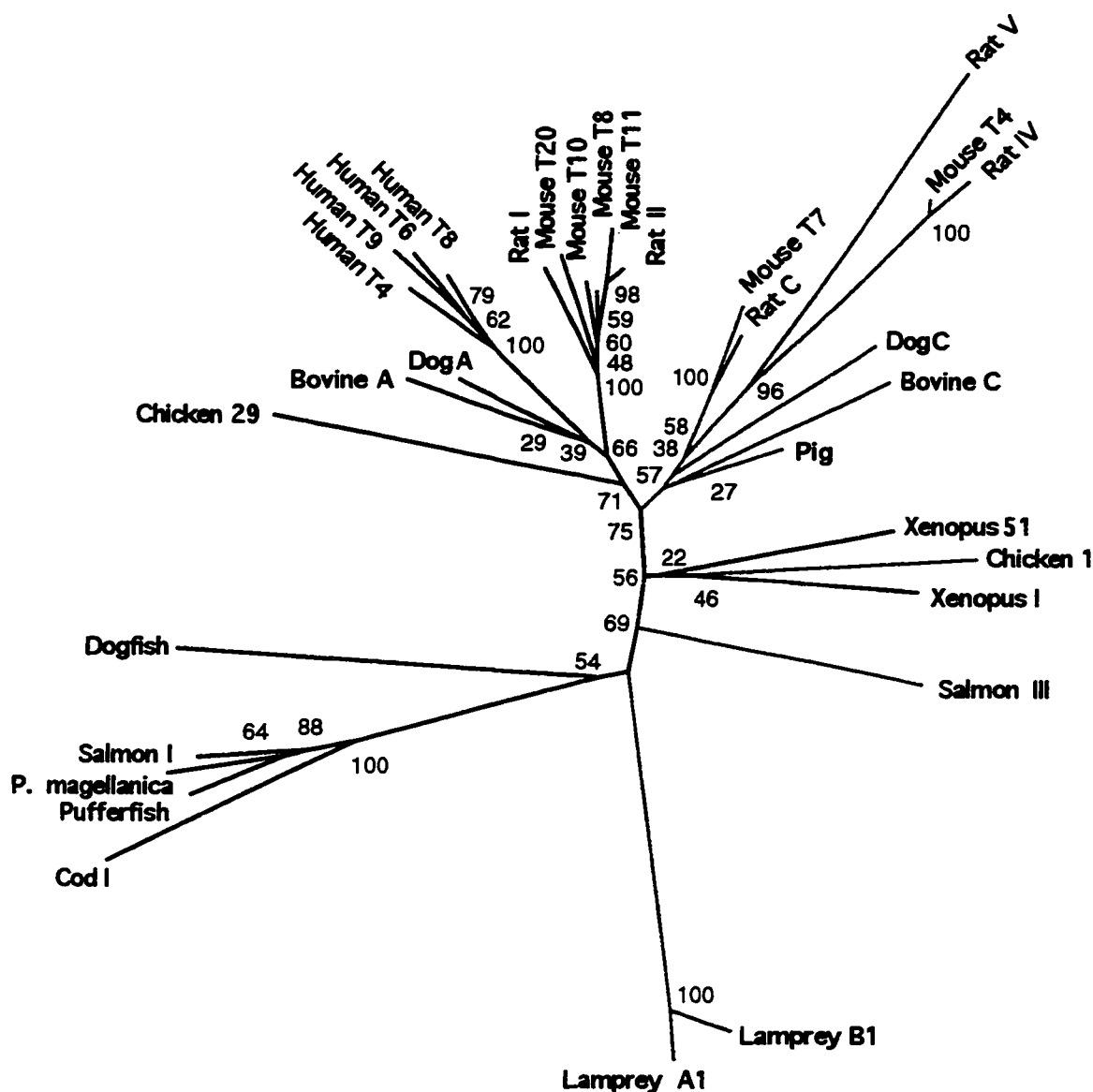


Figure 3.15. A Fitch-Margoliash phylogeny of thirty-two vertebrate trypsinogens. The graph is calculated as described for Figure 3.14. The numbers adjacent to nodes represent the number of times the clade distal to the unlabeled node was recovered during 100 delete-half-jackknives of the original data (but with only one random "jumble"). Group I branches are blue; group II branches are red. Branches in green diverged before group I and group II split. Black branches are indeterminate.

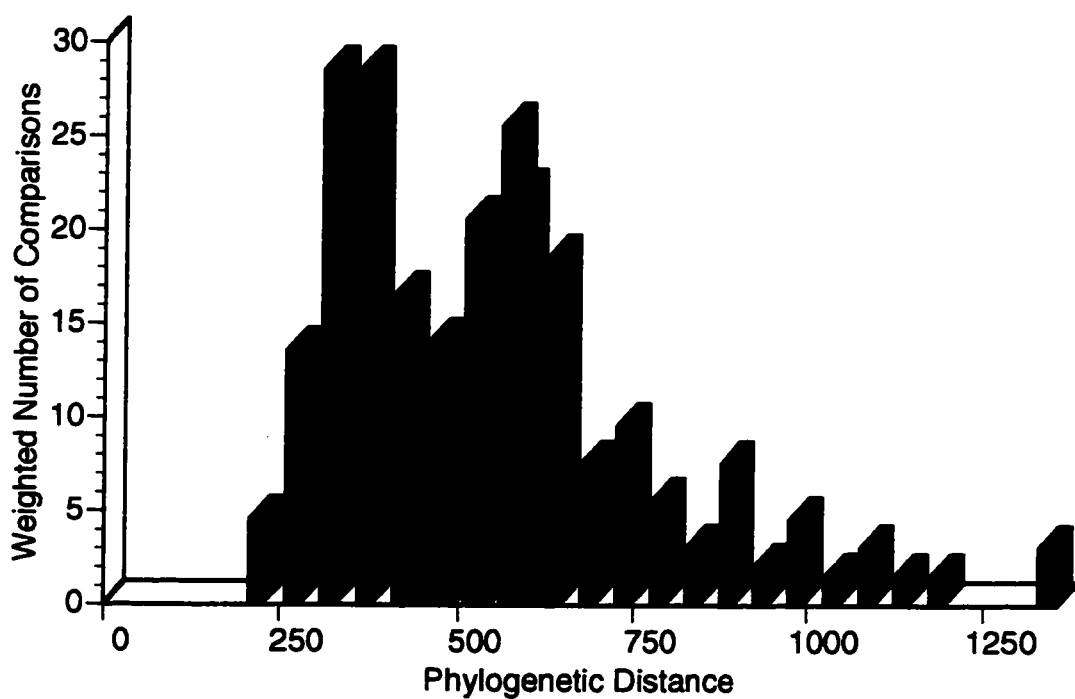


Figure 3.16. Weighted statistics of thirty sequences demonstrating distance differences between within-class comparisons (blue) and between-class comparisons (green). See text for a more complete description of methodology.

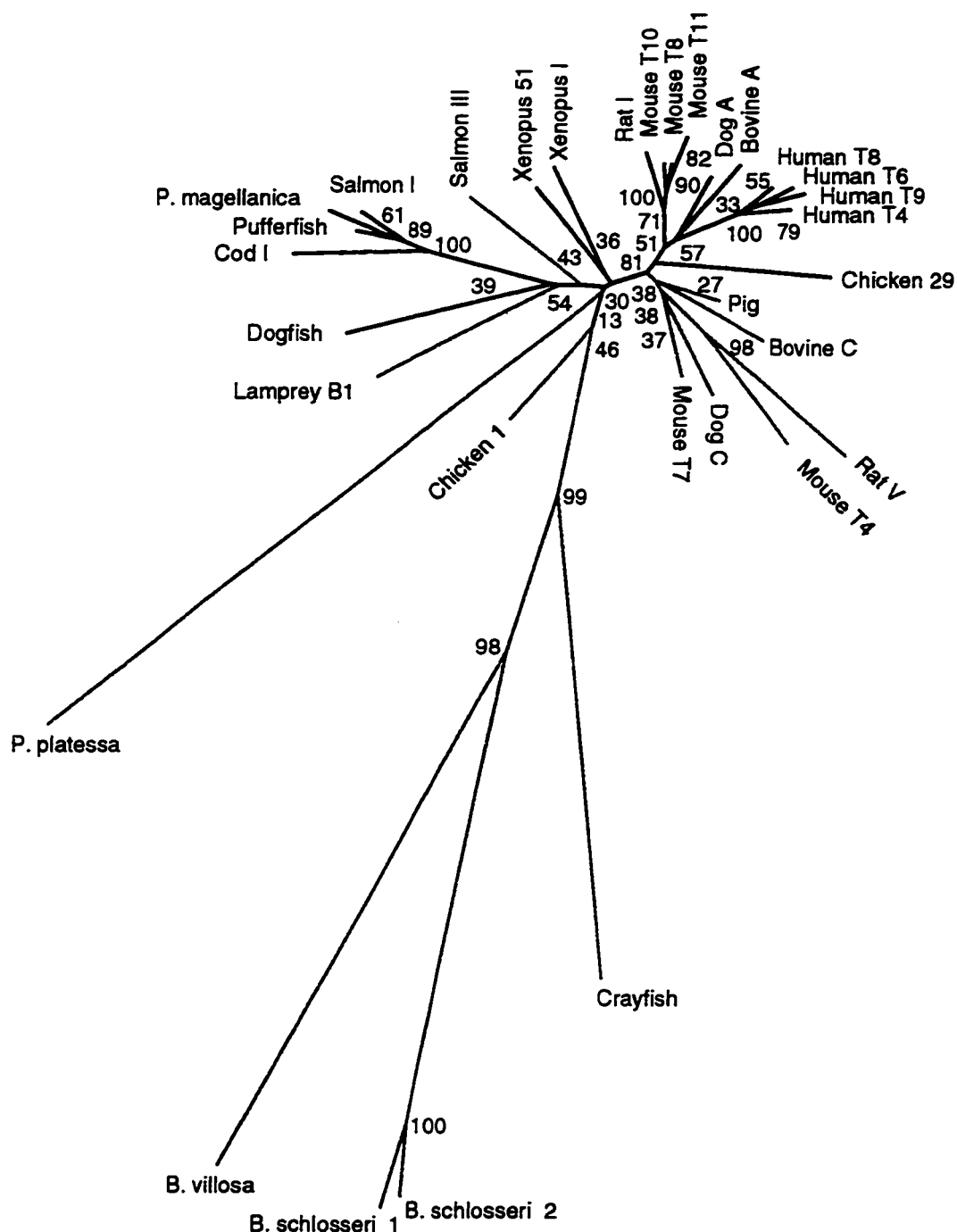


Figure 3.17. Pseudogenes added to the vertebrate trypsin phylogeny. Pseudogene names are in pink; pseudogene branches added to the phylogeny manually are bold. Group I branches are blue; group II branches are red. Branches in green diverged before group I and group II split. Black branches are indeterminate. The methodology of pseudogene addition is explained in the text. Pseudogene branch lengths are arbitrary; in this case, they have been selected for aesthetics.

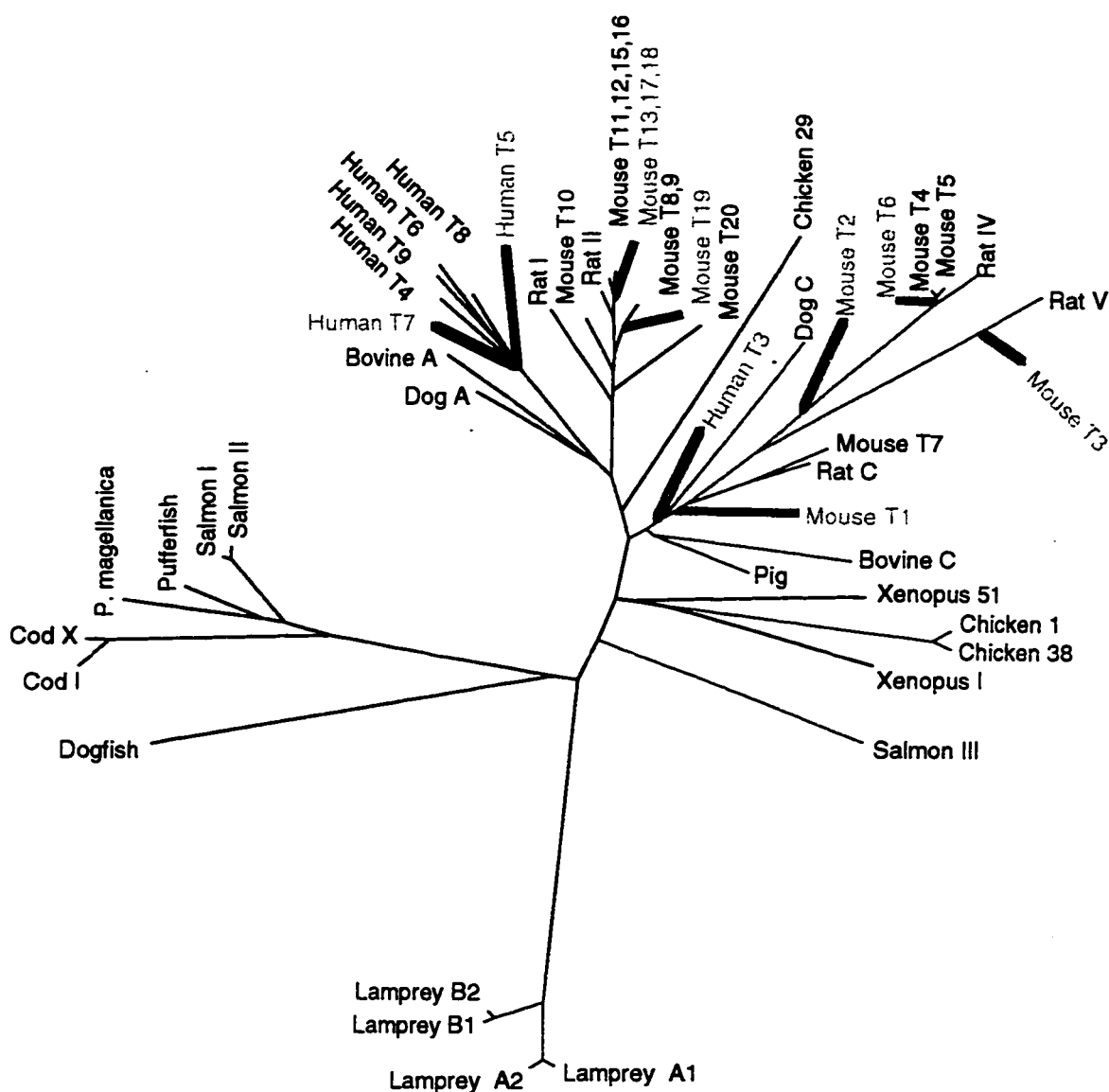


Figure 3.18. A Fitch-Margoliash phylogeny of thirty-two vertebrate trypsinogens, with the addition of five additional highly diverged sequences. The resulting phylogeny is highly skewed. The graph is calculated as described for Figure 3.14. The numbers adjacent to nodes represent the number of times the clade distal to the unlabeled node was recovered during 100 delete-half-jackknifes of the original data (but with only one random "jumble"). Group I branches are blue; group II branches are red. Branches in green diverged before group I and group II split. Black branches are indeterminate.

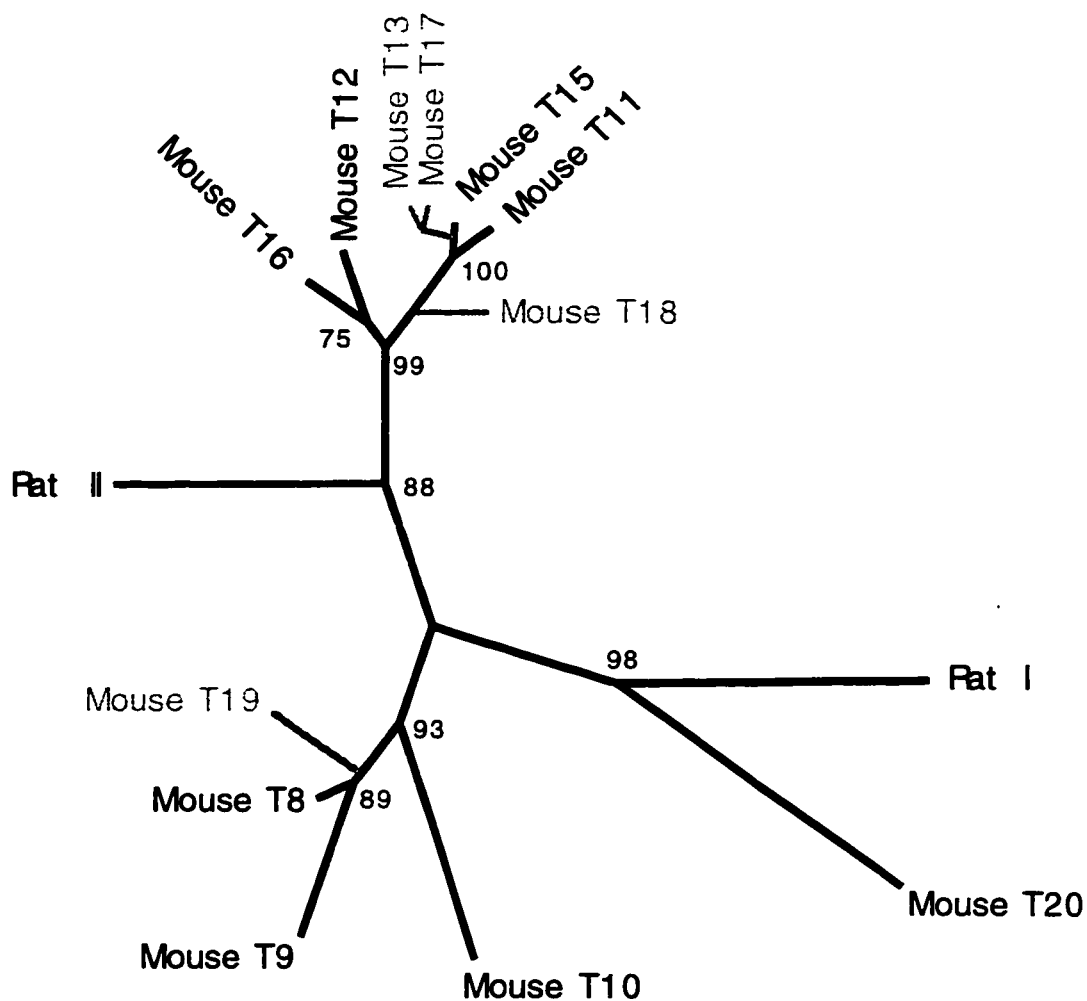


Figure 3.19. A phylogeny of the rodent group I trypsins. Constructed with the program *DNAML* with global rearrangements and 20 random “jumbles” (Felsenstein, 1993). Values at nodes are the result of 100 delete-half jackknifes of the original data (computed with global rearrangements but only one random “jumble”). Pseudogenes are in gray and were subsequently added under the protocol outlined for Figure 3.18.

Table 3.1. Literature references for the chordate trypsinogens.

SPECIES	ISOZYME	REFERENCE	ACCESSION NUMBER	COMMENTS
<i>Bolenia villosa</i>		Roach et al. (1997)	AF011897	cDNA
<i>Botryllus schlosseri</i>	1	Pancer et al. (1996)	X96387	cloned PCR product
<i>Botryllus schlosseri</i>	2	Pancer et al. (1996)	X96388	cloned PCR product
<i>Petromyzon marinus</i>	A1	Roach et al. (1997)	AF011352	cDNA
<i>Petromyzon marinus</i>	A2	Roach et al. (1997)	AF011898	cDNA
<i>Petromyzon marinus</i>	B1	Roach et al. (1997)	AF011899	cDNA
<i>Petromyzon marinus</i>	B2	Roach et al. (1997)	AF011900	cDNA
<i>Petromyzon marinus</i>	B3	Roach et al. (1997)	AF011901	cDNA
<i>Squalus acanthias</i>		Titani et al. (1975)	A00950	protein
<i>Gadus morhua</i>	I	Gudmundsdóttir et al. (1993)	X76886; X75998	cDNA
<i>Gadus morhua</i>	X	Gudmundsdóttir et al. (1993)	X76887; X75998	cDNA
<i>Gadus morhua</i>		direct submission	U47819	partial cDNA PCR product
<i>Salmo salar</i>	I	Male et al. (1995)	X70075	cDNA
<i>Salmo salar</i>	Ia	Male et al. (1995)	X70071	cDNA
<i>Salmo salar</i>	Ib	Male et al. (1995)	X70072	partial cDNA
<i>Salmo salar</i>	II	Male et al. (1995)	X70073	cDNA
<i>Salmo salar</i>	III	Male et al. (1995)	X70074	cDNA
<i>Takifugu rubripes</i>		Roach et al. (1997)	U25747	cDNA
<i>Paranotothenia magellanica</i>		Genicot et al. (1996)	X82223	cDNA
<i>Protopterus aethiopicus</i>		de Haën et al. (1977)	A61331; A27719	partial protein
<i>Xenopus laevis</i>	I	Shi and Brown (1990)	X53458	cDNA
<i>Xenopus laevis</i>	clone 51	Roach et al. (1997)	U72330	cDNA
<i>Gallus gallus</i>	P1	Wang et al. (1995)	GGU15155	cDNA and genomic
<i>Gallus gallus</i>	P29	Wang et al. (1995)	GGU15157	cDNA and genomic
<i>Gallus gallus</i>	P38	Wang et al. (1995)	GGU15156	cDNA and genomic
<i>Sus scrofa</i>		Hermanson et al. (1973)	A00947	protein
<i>Bos taurus</i>	Anionic	Le Huerou et al. (1990)	X54743	cDNA
<i>Bos taurus</i>	Cationic	Mikes et al. (1966)	P00760	protein
<i>Bos taurus</i>	Cationic	direct submission	D38507	cDNA

Table 3.1. Literature references for the chordate trypsinogens (continued).

SPECIES	ISOZYME	REFERENCE	ACCESSION NUMBER	COMMENTS
<i>Canis familiaris</i>	Anionic	Pinsky et al. (1995)	M11589	mRNA
<i>Canis familiaris</i>	Cationic	Pinsky et al. (1995)	M11590	mRNA
<i>Mus musculus</i>	1 (new gene)	direct submission	AE000663; AE000522	not really a trypsinogen
<i>Mus musculus</i>	2 (relic)	direct submission	AE000663; AE000522	genomic
<i>Mus musculus</i>	3 (pseudo)	direct submission	AE000663; AE000522	genomic
<i>Mus musculus</i>	4	direct submission	AE000663; AE000522	genomic
<i>Mus musculus</i>	5	direct submission	AE000663; AE000522	genomic
<i>Mus musculus</i>	6 (relic)	direct submission	AE000663; AE000522	genomic
<i>Mus musculus</i>	7	direct submission	AE000663; AE000522	genomic
<i>Mus musculus</i>	8	direct submission	AE000664; AE000522	genomic
<i>Mus musculus</i>	9	direct submission	AE000664; AE000522	genomic
<i>Mus musculus</i>	10	direct submission	AE000664; AE000522	genomic
<i>Mus musculus</i>	11	direct submission	AE000664; AE000522	genomic
<i>Mus musculus</i>	12	direct submission	AE000665; AE000522	genomic
<i>Mus musculus</i>	13 (relic)	direct submission	AE000665; AE000522	genomic
<i>Mus musculus</i>	14 (relic)	direct submission	AE000665; AE000522	genomic
<i>Mus musculus</i>	15	direct submission	AE000665; AE000522	genomic
<i>Mus musculus</i>	16	direct submission	AE000665; AE000522	genomic
<i>Mus musculus</i>	17 (relic)	direct submission	AE000665; AE000522	genomic
<i>Mus musculus</i>	18 (relic)	direct submission	AE000665; AE000522	genomic
<i>Mus musculus</i>	19 (relic)	direct submission	AE000665; AE000522	genomic
<i>Mus musculus</i>	20	Stevenson et al. (1986)	X04574	mRNA
<i>Mus musculus</i>	20	direct submission	AE000665; AE000522	genomic

Table 3.1. Literature references for the chordate trypsinogens (continued).

SPECIES	ISOZYME	REFERENCE	ACCESSION NUMBER	COMMENTS
<i>Rattus norvegicus</i>	I	MacDonald et al. (1982)	V01273	cDNA
<i>Rattus norvegicus</i>	II	MacDonald et al. (1982)	V01274	cDNA
<i>Rattus norvegicus</i>	Cationic (III)	Fletcher et al. (1987)	M16624	cDNA
<i>Rattus norvegicus</i>	IV	Lütke et al. (1989)	X15679	cDNA
<i>Rattus norvegicus</i>	Va	Kang et al. (1992)	X59012	cDNA
<i>Rattus norvegicus</i>	Vb	Kang et al. (1992)	X59013	cDNA
<i>Homo sapiens</i>	T1 (new gene)	Rowen et al. (1996)	U66059	not really a trypsinogen
<i>Homo sapiens</i>	T2 (relic)	Rowen et al. (1996)	U66059	genomic
<i>Homo sapiens</i>	T3 (pseudo)	Rowen et al. (1996)	U66059	genomic
<i>Homo sapiens</i>	T4 (I)	Emi et al. (1986)	M22612	cDNA
<i>Homo sapiens</i>	T4 (I)	Rowen et al. (1996)	U66061	genomic
<i>Homo sapiens</i>	T4 (I)	Whitcomb et al. (1996)	U70137	variant
<i>Homo sapiens</i>	T5 (pseudo)	Rowen et al. (1996)	AF009664; U66061	genomic
<i>Homo sapiens</i>	T6	Rowen et al. (1996)	U66061	genomic
<i>Homo sapiens</i>	T7 (pseudo)	Rowen et al. (1996)	U66061	genomic
<i>Homo sapiens</i>	T8 (II)	Emi et al. (1986)	M27602	cDNA
<i>Homo sapiens</i>	T8 (II)	Rowen et al. (1996)	AF009664; U66061	genomic
<i>Homo sapiens</i>	T9 (III)	Tani et al. (1990)	X15505	cDNA
<i>Homo sapiens</i>	T9 (III)	Wiegand et al. (1993)	X71345	cDNA
<i>Homo sapiens</i>	T9 (III)	Rowen et al. (1996)	AF029308	genomic

Table 3.2. Classification comments for the chordate trypsinogens.

SPECIES	ISOZYME	GENOMIC LOCALIZATION	ORIENTATION	GROUP	COMMON NAME
<i>Bolitena villosa</i>	I				tunicate (sea squirt)
<i>Botryllus schlosseri</i>	2				tunicate (sea squirt)
<i>Botryllus schlosseri</i>	A1				tunicate (sea squirt)
<i>Petromyzon marinus</i>	A2				sea lamprey
<i>Petromyzon marinus</i>	B1				sea lamprey
<i>Petromyzon marinus</i>	B2				sea lamprey
<i>Petromyzon marinus</i>	B3				sea lamprey
<i>Squalus acanthias</i>	I			I	sea lamprey
<i>Gadus morhua</i>	X			I	spiny dogfish
<i>Gadus morhua</i>	I			I	Atlantic cod
<i>Gadus morhua</i>	Ia			I	Atlantic cod
<i>Salmo salar</i>	Ib			I	Atlantic cod
<i>Salmo salar</i>	II			I	salmon
<i>Salmo salar</i>	III			II	salmon
<i>Takifugu rubripes</i>				I	pufferfish
<i>Paranotothenia magellanica</i>				I	Antarctic cod
<i>Protopterus aethiopicus</i>					marbled lungfish
<i>Xenopus laevis</i>	I			I	frog
<i>Xenopus laevis</i>	clone 51			I	frog
<i>Gallus gallus</i>	P1	5' end of TCR V β locus		II	chicken
<i>Gallus gallus</i>	P29	3' end of TCR V β locus		I	chicken
<i>Gallus gallus</i>	P38	5' end of TCR V β locus		II	chicken
<i>Sus scrofa</i>				III	pig
<i>Bos taurus</i>	Anionic			I	cow
<i>Bos taurus</i>	Cationic			II	cow
<i>Canis familiaris</i>	Anionic			I	dog
<i>Canis familiaris</i>	Cationic			II	dog
<i>Mus musculus</i>	1	5' to TCR V β on Chr. 6	←	II	mouse
<i>Mus musculus</i>	2	5' to TCR V β on Chr. 6	←	II	mouse
<i>Mus musculus</i>	3	5' to TCR V β on Chr. 6	←	II	mouse

Table 3.2. Classification comments for the chordate trypsinogens (continued).

SPECIES	ISOZYME	GENOMIC LOCALIZATION	ORIENTATION	GROUP	COMMON NAME
<i>Mus musculus</i>	4	5' to TCR V β on Chr. 6	←	II	mouse
<i>Mus musculus</i>	5	5' to TCR V β on Chr. 6	←	II	mouse
<i>Mus musculus</i>	6	5' to TCR V β on Chr. 6	←	II	mouse
<i>Mus musculus</i>	7	5' to TCR V β on Chr. 6	←	II	mouse
<i>Mus musculus</i>	8	3' to TCR V β on Chr. 6	→	I	mouse
<i>Mus musculus</i>	9	3' to TCR V β on Chr. 6	←	I	mouse
<i>Mus musculus</i>	10	3' to TCR V β on Chr. 6	→	I	mouse
<i>Mus musculus</i>	11	3' to TCR V β on Chr. 6	←	I	mouse
<i>Mus musculus</i>	12	3' to TCR V β on Chr. 6	→	I	mouse
<i>Mus musculus</i>	13	3' to TCR V β on Chr. 6	→	I	mouse
<i>Mus musculus</i>	14	3' to TCR V β on Chr. 6	→	I	mouse
<i>Mus musculus</i>	15	3' to TCR V β on Chr. 6	←	I	mouse
<i>Mus musculus</i>	16	3' to TCR V β on Chr. 6	→	I	mouse
<i>Mus musculus</i>	17	3' to TCR V β on Chr. 6	→	I	mouse
<i>Mus musculus</i>	18	3' to TCR V β on Chr. 6	→	I	mouse
<i>Mus musculus</i>	19	3' to TCR V β on Chr. 6	→	I	mouse
<i>Mus musculus</i>	20	3' to TCR V β on Chr. 6	←	I	mouse
<i>Rattus norvegicus</i>	I			I	rat
<i>Rattus norvegicus</i>	II			I	rat
<i>Rattus norvegicus</i>	III (Cationic)			II	rat
<i>Rattus norvegicus</i>	IV			II	rat
<i>Rattus norvegicus</i>	Va			II	rat
<i>Rattus norvegicus</i>	Vb			II	rat
<i>Homo sapiens</i>	T1	5' to TCR V β on Chr. 7	←	II	human
<i>Homo sapiens</i>	T2	5' to TCR V β on Chr. 7	←	II	human
<i>Homo sapiens</i>	T3	5' to TCR V β on Chr. 7	←	II	human
<i>Homo sapiens</i>	T4 (I)	3' to TCR V β on Chr. 7	→	I	human
<i>Homo sapiens</i>	T5	3' to TCR V β on Chr. 7	→	I	human
<i>Homo sapiens</i>	T6	3' to TCR V β on Chr. 7	→	I	human
<i>Homo sapiens</i>	T7	3' to TCR V β on Chr. 7	→	I	human
<i>Homo sapiens</i>	T8 (II)	3' to TCR V β on Chr. 7	→	I	human
<i>Homo sapiens</i>	T9 (III and IV)	3' to TCR V β on Chr. 9	→	I	human

Table 3.3. PCR primers used in the analysis of the chordate trypsinogens.

Degenerate Serine Protease Primers

H	CTSWCWGCWGCYCA YTG
S	YMSWGGKCCNCCRGARTC

Tunicate Trypsinogen Sequencing Primers

TUN2F1	TGGAACACGTGGAAAATAGTTCTC
TUN2R1	CGAGAACTATTTTCCACGTGTTCC
TUN2F2	CAAGCAGCGGAGGAACTATCTCCG
TUN2R2	TCCACTAACAGTACACGCGGTGTC
TUN19F1	GGTGTATACACCCGTGTTGCAGTG
TUN19R1	ACACTGCAACACGGGTGTATACAC
TUN2R3	TTTGGATGATTAAAGATTTTTTATTG

Trypsin specific primers

TRYA	TCCGGATCCTGATGACAAGATCGTTGGGGG
TRYB	TCCGGATCCTTCTGTGGAGGCTCCCTCAT
TRYC	TCCGGATCCATAGCCCCAGGAGAC
TRYD	TCCGGATCCTTGGTGTAGACACCAGG
TRYF	CTGGATCCGTGAGACTGGGAGAGCAC
TRYR	CTGGATCCGAATCCTTGCCTCCCTC

Lamprey Trypsinogen Sequencing Primers

LT2	AGCCAGTGGGTCCTGTCTG
LT3R	TCACGAAGATGTTGTGCTC
LT3	TCATGCTCATCAAGCTGTCCTC
LT4R	ACGCACATGAGGACGTCGGGAC
LT5R	AAGAGTAGTGTGTTAGATCCAC
XTA5	CCGGTGGCCCCGTGGTGTG

Table 3.4. Signal peptides and activation peptides of the chordate trypsinogens. The signal peptidase cleavage site is the predicted site (see text); only in the case of the canine (marked with *), have the sites been determined experimentally (Carne and Scheele, 1982). A dashed line (—) represents undetermined sequence. Intron/exon boundaries, where known, are indicated by bold lettering.

	<i>signal</i>	<i>activation</i>	<i>mature</i>
<i>B. villosa</i>	MKIVILLLLGLAAVNA	DK	IVGG
<i>B. schlosseri</i> 1	MKVFAILLLLAFCGANA	DK	IIGG
<i>B. schlosseri</i> 2	MKVFAILLLLALYGANA	DK	IIGG
Lamprey A1	MHGLILALLVGVA ^{AA}	APYMYEDH	IVGG
Lamprey A2	MHGLILALLVGVA ^{AA}	APYMYEDH	IVGG
Lamprey B1 [†]	---LIFALLVGTA ^{AA}	APYMYEDH	IVGG
Lamprey B2 [†]	--GLIFALLVGTA ^{AA}	APYMYEDH	IVGG
Dogfish [†]	-----	APDDDDK	IVGG
Cod I	RKSLIFVLLLGAVFA	EEDK	IVGG
<i>P. magellanica</i>	MRSLVFVLLIGAFA	TEEDK	IVGG
Pufferfish [†]	-----LIAA ^{AA}	APIDEDDK	IVGG
Salmon I	MISLVFVLLIGAFA	TEDDK	IVGG
Salmon III [†]	-----FAVAFA	APIDEDDK	IVGG
<i>Xenopus</i> I	MKFLLLCVLLGAA ^{AA}	FDDDK	IIGG
<i>Xenopus</i> 51	MKFLVILVLLGA ^{AA}	FEDDDK	IVGG
Chicken 1	MLFLVLVAFVGVTV ^A	FPISDEDDDK	IVGG
Chicken 29	MLFLFLILSCLGA ^{AA}	FPGGADDDK	IVGG
Pig [†]	-----	FPTDDDDK	IVGG
Bovine A	MHPLLILAFVGA ^{AA}	FPSDDDDK	IVGG
Bovine C [†]	---FIFLALLGA ^{AA}	FPVDDDDK	IVGG
Dog A [*]	MNPLLILAF ^{LL} GA ^{AA}	TPTDDDDK	IVGG
Dog C [*]	MLTFIFLALLGAT ^V	FPIDDDDK	IVGG
Human T4 (I)	MNPLLILTFVAA ^{LLA}	APFDDDDK	IVGG
Human T6	MNPLLILAFVGA ^{AA}	VPFDDDDK	IVGG
Human T8 (II)	MNLLLILTFVAA ^{AA}	APFDDDDK	IVGG
Human T9 (III)	MNPFLILAFVGA ^{AA}	VPFDDDDK	IVGG
Human T3 (5' ϕ)	HEDLHLPALLGA ^{AA} T	FPTDDDDK	IVGG
Rat I	MSALLILALVGA ^{AA}	FPLEDDDK	IVGG
Rat II	MRALLILALVGA ^{AA}	FPVDDDDK	IVGG
Rat C	MLALIFLAF ^{LL} GA ^{AA}	LPLDDDDDK	IVGG
Rat IV	MLISIFFAF ^{LL} GA ^{AA}	LPVND ^{DD} K	IVGG
Rat V	MKICIFFTLLGT ^V AA	FPTEDNDDR	IVGG
Mouse T4	MKIITFFTF ^{LL} GA ^{AA}	LPANSDDK	IVGG
Mouse T5	MKIIFFFTF ^{LL} GA ^{AA}	LPANSDDK	IVGG
Mouse T7	MKTLIFLAF ^{LL} GA ^{AA}	LPLDDDDDK	IVGG
Mouse T8	MRALLFLALVGA ^{AA}	FPVDDDDK	IVGG
Mouse T9	MNSLLFLALVGA ^{AA}	FPVDDDDK	IVGG
Mouse T10	MSTLLFLALVGA ^{AA}	FPVDDDDK	IVGG
Mouse T11	MNALLILALVGA ^{AA}	FPVDDDDK	IVGG
Mouse T12	MSALLFLALVGA ^{AA}	FPVDDDK	IVGG
Mouse T15	MNAFLILALVGA ^{AA}	FPVDDDDK	IVGG
Mouse T16	MSALLFLALVGA ^{AA}	FPVDDDDK	IVGG
Mouse T20	MSALLILALVGA ^{AA}	FPVDDDDK	IVGG

Table 3.5. Cystine bridges of the chordate trypsinogens. A cross (X) indicates the predicted presence of a bridge between the residues designated at the top of each column. (*), experimentally determined (Kauffman, 1965; Jurasek et al., 1969; Jurásek and Smillie, 1973); (#), pseudogene missing one of two cysteine codons.

	C22-C157	C42-C58	C127-C232	C136-C201	C168-C182	C191-C220
Bacteria		X*			X*	X*
Crayfish		X			X	X
Tunicate		X		X	X	X
Lamprey		X	X	X	X	X
Non-Human Gnathostomata	X	X*	X*	X*	X*	X*
Group II Human (ϕ T3)	X#	X	X	X	X	X
Group I Human (except T8)	X	X		X	X	X
Group I Human (T8)	X	X			X	X

Table 3.6. Predicted isoelectric points and charges of the chordate trypsins (calculated with the program *DNA **[®], Madison, WI). Each Osteichthyes or tetrapod trypsin is assigned to either Group I or II based on phylogenetic considerations (see text). Experimental values (in parentheses) for isoelectric points are from Walsh (1970), Lütcke et al. (1989), and Asgeirsson et al. (1989). Note that measured isoelectric points depend not only on the net charge, but also on the distribution of the charge (Smalås et al. 1994). Charge predictions do not take this into account.

Isozyme	Group	Isoelectric Point	Charge at pH 7.0
Tunicate		3.9	-13.18
Lamprey A1		5.2	-5.62
Lamprey B1		5.8	-3.62
Dogfish	I	4.9	-10.11
Cod I	I	6.8 (6.6)	-0.79
Cod X	I	5.8 (5.5)	-5.95
<i>P. magellanica</i>	I	5.8	-5.28
Pufferfish	I	6.2	-2.61
Salmon I	I	5.9	-3.62
Salmon II	I	5.5	-4.62
Salmon III	II	8.1	4.21
Chicken P1	II	8.2	5.03
Chicken P29	I	4.6	-9.78
Xenopus	I	6.7	-0.79
Xenopus 51	I	7.7	2.21
Dog A	I	4.9	-5.94
Dog C	II	8.3	6.04
Pig	II	7.9 (10.8)	3.21
Bovine A	I	4.8	-7.62
Bovine C	II	8.3 (10.1)	6.03
Human T4 (I)	I	7.5	1.28
Human T6	I	6.9	-0.39
Human T8 (II)	I	5.0	-6.65
Human T9 (III)	I	6.8	-0.55
Mouse T4	II	6.8	-0.62
Mouse T5	II	6.5	-1.62
Mouse T7	II	8.3	6.20
Mouse T8	I	6.3	-1.79
Mouse T9	I	5.9	-2.79
Mouse T10	I	6.7	-0.79
Mouse T11	I	5.2	-4.78
Mouse T12	I	5.6	-3.78
Mouse T15	I	5.0	-5.78
Mouse T16	I	5.0	-5.78
Mouse T20	I	4.4	-9.78
Rat I	I	4.9 (4.4)	-6.62
Rat II	I	4.8 (4.3)	-6.78
Rat C	II	8.1 (8)	4.20
Rat IV	II	6.9 (6.2)	-0.29
Rat V	II	5.1	-9.11

Table 3.7. Genbank identification numbers (GIs) for the human trypsinogen ESTs, as of November 9, 1997. Isozyme identifications are in bold.

T4

1183500	1321005	1324454	1349808	1350157	1350390	1358723	1383458	1947217
1947222	1947267	1947302	1947441	1947671	1947685	1947767	1947841	1947876
1947953	1948021	1948066	1948120	2015973	2016030	2016134	2016136	2018354
2018413	2018415	2018424	2018431	2018496	2018504	2018544	2018665	2018691
2018790	2018855	2018892	2018989	2019014	2019039	2019046	2019072	2019152
2019195	2019233	2019234	2019267	2019367	2019440	2019467	2019475	2019554
2019557	2019623	2019627	2019634	2019697	2019740	2022066	391091	391188
704030								

T6

1947519	1947827	1948133	1965361	2015994	2019622
---------	---------	---------	---------	---------	---------

T8

1324185	1324290	1349770	1350016	1358098	1384253	1503188	1934275	1941124
1947213	1947436	1948039	1965364	2015956	2015978	2015987	2016017	2016029
2016033	2016110	2016148	2018350	2018364	2018381	2018394	2018405	2018421
2018423	2018527	2018547	2018574	2018584	2018592	2018599	2018633	2018646
2018654	2018662	2018696	2018720	2018805	2018808	2018812	2018893	2018907
2018913	2018923	2018954	2018955	2018957	2018978	2019049	2019102	2019132
2019162	2019167	2019243	2019284	2019318	2019338	2019352	2019363	2019366
2019371	2019398	2019410	2019427	2019430	2019442	2019447	2019449	2019460
2019476	2019510	2019513	2019563	2019576	2019588	2019618	2019619	2019653
2019718	2019720	2019725	2019734	2019752	2019772	2038313	391109	391414
391419	391425	475301	475318					

normally-spliced T9

1947892

alternatively-spliced T9

1968122	611445	1471327	1525435	1634309	1960755	1960950
---------	--------	---------	---------	---------	---------	---------

un-assigned T9

1183818	1329462	1423350	2015975	2018936	2397944
---------	---------	---------	---------	---------	---------

Table 3.8. Genbank accession numbers and tissues of origin for the mouse trypsinogen ESTs, as of November 24, 1997. Isozyme identifications are in bold.

T7

AA260562	liver
AA390094	lymph node
AA537998	diaphragm
AA570969	diaphragm
AA572665	diaphragm
AA615050	colon
AA638704	colon

T8

AA066788	diaphragm
AA066988	diaphragm
AA239834	liver
AA268196	liver
AA530444	diaphragm
AA537800	diaphragm
AA571068	diaphragm
AA571214	diaphragm
AA571693	diaphragm
AA572325	diaphragm
AA572330	diaphragm

T9

AA110460	testis
AA168368	spleen
AA512480	colon
AA537785	diaphragm
AA571280	diaphragm
AA607527	colon

AFTERWORD

It is my hope that interest in this dissertation may last some time into the future. Perhaps a few temporally distant readers may only know of microfilm as an antique curiosity. The dynamic nature of knowledge and publishing may encourage a future reader to search more versatile archives. To this end, such a reader may wish to start with my web page, currently located at weber.u.washington.edu/~roach. I will post relevant updates and related information to this web site.

BIBLIOGRAPHY

- Abdelhak, S., Kalatzis, V., Heilig, R., Compain, S., Samson, D., Vincent, C., Weil, D., Cruaud, C., Sahly, I., Leibovici, M., Bitner-Glindzicz, M., Francis, M., Lacombe, D., Vigneron, J., Charachon, R., Boven, K., Bedbeder, P., Van Regemorter, N., Weissenbach, J. and Petit, C. 1997. A human homologue of the *Drosophila* eyes absent gene underlies branchio-oto-renal (BOR) syndrome and identifies a novel gene family. *Nature Genetics* 15:157-164.
- Abita, J. P., Delaage, M. and Lazdunski, M. 1969. The mechanism of activation of trypsinogen. The role of the four N- terminal aspartyl residues. *European Journal of Biochemistry* 8:314-324.
- Allison, D. P., Kerper, P. S., Doktycz, M. J., Thundat, T., Modrich, P., Larimer, F. W., Johnson, D. K., Hoyt, P. R., Mucenski, M. L. and Warmack, R. J. 1997. Mapping individual cosmid DNAs by direct AFM imaging. *Genomics* 41:379-384.
- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410.
- Anantharaman, T. S., Mishra, B., and Schwartz, D. 1997. Genomics via optical mapping II: ordered restriction maps. *Journal of Computational Biology* 4(2):91-118.
- Ásgeirsson, B., Fox, J. W., and Bjarnason, J. B. 1989. Purification and characterization of trypsin from the poikilotherm *Gadus morhua*. *European Journal of Biochemistry* 180:85-94.
- Barrett, A. J., and Rawlings, N. D. 1993. The many evolutionary lines of peptidases. In "Innovations in Proteases and Their Inhibitors" (F. X. Avilés, Ed.), pp. 13-30, Walter De Gruyter, Berlin.
- Bartunik, H. D., Summers, L. J. and Bartsch, H. H. 1989. Crystal structure of bovine beta-trypsin at 1.5 Å resolution in a crystal form with low molecular packing density. Active site geometry, ion pairs and solvent structure. *Journal of Molecular Biology* 210:813-828.

- Baticle, E. 1935. Le problème de la répartition. *Comptes Rendus* 201:862-864.
- Bishop, D. T., Cannings, C., Skolnick, M., and Williamson, J. A. 1983. The number of polymorphic DNA clones required to map the human genome. In "Statistical Analysis of DNA Sequence Data" (Weir, B. S., Ed.), pp.181-200, Marcel Dekker, New York.
- Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453-1474.
- Bode, W. and Schwager, P. 1975. The single calcium-binding site of crystallin bovin beta-trypsin. *FEBS Letters* 56:139-143.
- Bodenteich, A., S. Chisoe, Y.-F. Wang, and B. A. Roe. 1994. Shotgun Cloning as the Method of Choice to Generate Templates for High-throughput Dideoxynucleotide Sequencing. In "Automated DNA Sequencing and Analysis" (M. Adams, C. Fields, and J. Venter, Eds.), pp. 42-50, Academic Press, New York.
- Bordo, D., Djinović, K. and Bolognesi, M. 1994. Conserved patterns in the Cu,Zn superoxide dismutase family. *Journal of Molecular Biology* 238:366-386.
- Borts, R. H. and Haber, J. E. 1987. Meiotic recombination in yeast: alteration by multiple heterozygosities. *Science* 237:1459-1465.
- Brenner, S. 1988. The molecular evolution of genes and proteins: a tale of two serines. *Nature* 334:528-530.
- Brown, D. D., Wensink, P. C. and Jordan, E. 1972. A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. *Journal of Molecular Biology* 63:57-73.
- Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S.

- M. and Venter, J. C. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273:1058-1073.
- Burland, V., Daniels, D. L., Plunkett, G., and Blattner, F. R. 1993. Genome sequencing on both strands: the Janus strategy. *Nucleic Acids Research* 21(15):3385-3390.
- Bystroff, C., and Baker, D. 1997. Improved local structure prediction for proteins using a library of sequence-structure motifs. Submitted.
- Carne, T., and Scheele, G. 1982. Amino acid sequences of transport peptides associated with canine exocrine pancreatic proteins. *Journal of Biological Chemistry* 257(8):4133-4140.
- Carreira, S., Fueri, C., Chaix, J. C. and Puigserver, A. 1996. Dietary modulation of the mRNA stability of trypsin isozymes and the two forms of secretory trypsin inhibitor in the rat pancreas. *European Journal of Biochemistry* 239:117-123.
- Chen, C. N., Su, Y., Baybayan, P., Siruno, A., Nagaraja, R., Mazzarella, R., Schlessinger, D. and Chen, E. 1996. Ordered shotgun sequencing of a 135 kb Xq25 YAC containing ANT2 and four possible genes, including three confirmed by EST matches. *Nucleic Acids Research* 24:4034-4041.
- Chen, E. Y., Schlessinger, D., and Kere, J. 1993. Ordered shotgun sequencing, a strategy for integrating mapping and sequencing of YAC clones. *Genomics* 17:651-656.
- Chen, L., DeVries, A. L. and Cheng, C. H. 1997. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences* 94:3811-3816.
- Cilia, V., Lafay, B., Christen, R. 1996. Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. *Molecular Biology and Evolution* 13(3):451-461.
- Clarke, L., and J. Carbon. 1976. A colony bank containing synthetic Col E1 hybrid plasmids representative of the entire *E. coli* genome. *Cell* 9:91-99.

- Corey, D. R., and Craik, C. S. 1993. Trypsin: a model enzyme for the introduction of novel properties into proteins. *In* "Innovations in Proteases and Their Inhibitors" (F. X. Avilés, Ed.), pp. 425-444, Walter De Gruyter, Berlin.
- Cox, T. F., and Cox, M. A. A. 1994. "Multidimensional Scaling," Chapman & Hall, London.
- Craik, C. S., Rutter, W. J. and Fletterick, R. 1983. Splice junctions: association with variation in protein structure. *Science* 220:1125-1129.
- Davis, C. A., Riddel, D. C., Higgins, M. J., Holden, J. A., and White, B. N. 1985. A gene family in *Drosophila melanogaster* for trypsin-like enzymes. *Nucleic Acids Research* 13(18):6605-6619.
- Davison, A. J. 1991. Experience in shotgun sequencing a 134 kilobase pair DNA molecule. *DNA Sequence* 1:389-394.
- Dawkins, R. 1990. "The Selfish Gene," Oxford University Press, Oxford.
- Dawkins, R. 1996. "The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design," W. W. Norton & Company, New York.
- Dayhoff, M. O. 1979. "Atlas of Protein Sequence and Structure, Volume 5, Supplement 3," National Biomedical Research Foundation, Washington, D.C.
- De Haën, C., Neurath, H., and Teller, D. C. 1975. The phylogeny of trypsin-related serine proteases and their zymogens. New methods for the investigation of distant evolutionary relationships. *Journal of Molecular Biology* 92:225-259.
- de Haën, C., Walsh, K. A., and Neurath, H. 1977. Isolation and amino-terminal sequence analysis of a new pancreatic trypsinogen of the African lungfish *Protopterus aethiopicus*. *Biochemistry* 16(20):4421-4425.
- Deininger, P. L. 1983. Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Analytical Biochemistry* 129:216-223.
- DeRisi, J. L., Iyer, V. R. and Brown, P. O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686.

- Dickerson, R. E. 1971. The structure of cytochrome c and the rates of molecular evolution. *Journal of Molecular Evolution* 1:26-45.
- Eddy, S. R., Mitchison, G., and Durbin, R. 1995. Maximum discrimination hidden Markov models of sequence consensus. *Journal of Computational Biology* 2(1):9-24.
- Edelman, G. M., and Gally, J. A. 1970. Arrangement and evolution of eukaryotic genes. In "The Neurosciences: Second Study Program" (Schmitt, F. O., Ed.), pp.962-972, Rockefeller University Press, New York.
- Edwards, A., and Caskey, T. 1991. Closure strategies for random DNA sequencing. *Methods: A Companion to Methods in Enzymology*. 3(1):41-47.
- Edwards, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C., Zimmerman, J., Erfle, H., Caskey, T., and Ansorge, W. 1990. Automated DNA sequencing of the human HPRT locus. *Genomics*. 6:593-608.
- Eigen, M., Winkler-Oswatitsch, R. and Dress, A. 1988. Statistical geometry in sequence space: a method of quantitative comparative sequence analysis. *Proceedings of the National Academy of Sciences* 85:5913-5917.
- Emi, M., Nakamura, Y., Ogawa, M., Yamamoto, T., Nishide, T., Mori, T., and Matsubara, K. 1986. Cloning, characterization and nucleotide sequences of two cDNAs encoding human pancreatic trypsinogens. *Gene* 41(2-3):305-310.
- Evans, G. A. 1991. Combinatoric strategies for genome mapping. *Bioessays*. 13(1):39-44.
- Everitt, B. S. 1993. "Cluster Analysis," Halsted Press, New York.
- Everitt, B. S., and Dunn, G. 1991. "Applied Multivariate Data Analysis," Edward Arnold, London.
- Fabret, C., Quentin, Y., Chapal, N., Guiseppi, A., Haiech, J. and Denizot, F. 1996. Integrated mapping and sequencing of a 115 kb DNA fragment from *Bacillus subtilis*: sequence analysis of a 21 kb segment containing the sigL locus. *Microbiology* 142:3089-3096.

- Farber, S., Shwachman, H., and Maddock, C. H. 1943. Pancreatic function and disease in early life. I. Pancreatic enzyme activity and the coeliac syndrome. *Journal of Clinical Investigation* 22:827-838.
- Feder, J. N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D. A., Basava, A., Dormishian, F., Domingo, R., Jr., Ellis, M. C., Fullan, A., Hinton, L. M., Jones, N. L., Kimmel, B. E., Kronmal, G. S., Lauer, P., Lee, V. K., Loeb, D. B., Mapa, F. A., McClelland, E., Meyer, N. C., Mintier, G. A., Moeller, N., Moore, T., Morikang, E., Prass, C. E., Quintana, L., Starnes, S. M., Schatzman, R. C., Brunke, K. J., Drayna, D. T., Risch, N. J., and Wolff, R. K. 1996. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genetics* 13:399-408.
- Felsenstein, J. 1983. Inferring evolutionary trees from DNA sequences. In "Statistical Analysis of DNA Sequence Data" (Weir, B. S., Ed.), pp.133-150, Marcel Dekker, New York.
- Felsenstein, J. 1993. *PHYLIP* (Phylogeny inference package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.
- Felsenstein, J., and Churchill, G. A. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* 13(1):93-104.
- Fields, C., Adams, M. D., White, O. and Venter, J. C. 1994. How many genes in the human genome? *Nature Genetics* 7:345-346.
- Fisher, R. A. 1940. On the similarity of the distributions found for the test of harmonic significance in harmonic analysis, and in Steven's problem in geometrical probability. *Annals of Eugenics* 10:14-17.
- Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Systematic Zoology* 19:99-113.
- Flatto, L., and A. G. Konheim. 1962. The Random division of an interval and the random covering of a circle. *SIAM Review* 43:211-222.

- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrman, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. and Venter, J. C. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496-512.
- Fletcher, T. S., Alhadeff, M., Craik, C. S., and Langman, C. 1987. Isolation and characterization of a cDNA encoding rat cationic trypsinogen. *Biochemistry* 26:3081-3086.
- Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A. and Venter, J. C. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270:397-403.
- Galper, A. R., and Brutlag, D. L. 1993. Computational simulations of biological systems. In "Biocomputing: Informatics and Genome Projects" (Smith, D. W., Ed.), pp.269-305, Academic Press, San Diego.
- Genicot, S., Rentier-Delrue, F., Edwards, D., vanBeeumen, J., and Gerday, C. 1996. Trypsin and trypsinogen from an Antarctic fish: molecular basis of cold adaptation. *Biochimica et Biophysica Acta* 1298(1):45-57.
- Gish, W. and States, D. J. 1993. Identification of protein coding regions by database similarity search. *Nature Genetics* 3:266-272.
- Gonick, L., and Wheelis, M. 1991. "The Cartoon Guide to Genetics," Harper Perennial, New York.

- Green, E. D., and Green P. 1991. Sequence tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods and Applications* 1:77-90.
- Gudmundsdóttir, Á., Gudmundsdóttir, E., Óskarsson, S., Bjarnason, J. B., Eakin, A. K., and Craik, C. S. 1993. Isolation and characterization of cDNAs from Atlantic cod encoding two different forms of trypsinogen. *European Journal of Biochemistry* 217:1091-1097.
- Gusfield, D. 1997. "Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology," Cambridge University Press, Cambridge.
- Haars, R., Kronenberg, M., Gallatin, W. M., Weissman, I. L., Owen, F. L. and Hood, L. 1986. Rearrangement and expression of T cell antigen receptor and gamma genes during thymic development. *Journal of Experimental Medicine* 164:1-24.
- Hagmann, M. 1997. RAGged repair: what's new in V(D)J recombination. *Biological Chemistry* 378:815-819.
- Hall, P. 1988. *Introduction to the Theory of Coverage Processes*. John Wiley & Sons, New York.
- Hartley, B. S. 1970. Homologies in serine proteinases. *Philosophical Transactions of the Royal Society of London. Series B. Sci* 257:77-87.
- Hartley, B. S. 1979. Evolution of enzyme structure. *Proceedings of the Royal Society of London. Series B.* 205:443-452.
- Haskins, C. H. 1923. "The Rise of Universities," Cornell University Press, London.
- Hedstrom, L., Farr-Jones, S., Kettner, C. A., and Rutter, W. J. 1994a. Converting trypsin to chymotrypsin: ground-state binding does not determine substrate specificity. *Biochemistry* 33:8764-8769.
- Hedstrom, L., Perona, J. J., and Rutter, W. J. 1994b. Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. *Biochemistry* 33:8757-8763.

- Hedstrom, L., Szilagyi, L., and Rutter, W. J. 1992. Converting trypsin to chymotrypsin: the role of surface loops. *Science* 255:1249-1253.
- Heiter, P. and Boguski, M. 1997. Functional genomics: it's all how you read it. *Science* 278:601-602.
- Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K. and Hood, L. 1997. Gene families: the taxonomy of protein paralogs and chimeras. *Science* 278:609-614.
- Hermanson, M. A., Ericsson, L. H., Neurath, H., and Walsh, K. A. 1973. Determination of the amino acid sequence of porcine trypsin by sequenator analysis. *Biochemistry* 12(17):3146-3153.
- Hewett-Emmett, D., Czelusniak, J. and Goodman, M. 1981. The evolutionary relationship of the enzymes involved in blood coagulation and hemostasis. *Annals of the New York Academy of Sciences* 370:511-527.
- Higashimoto, Y. and Liddle, R. A. 1993. Isolation and characterization of the gene encoding rat glucose-dependent insulinotropic peptide. *Biochemical and Biophysical Research Communications* 193:182-190.
- Hillis, D. M. 1996. Inferring complex phylogenies. *Nature* 383:130-131.
- Himmelreich, R., Hilbert, H., Plagens, H., Pirkel, E., Li, B. C. and Herrmann, R. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Research* 24:4420-4449.
- Holland, P. W. and Garcia-Fernández, J. 1996. Hox genes and chordate evolution. *Developmental Biology* 173:382-395.
- Hollingshead, S. K., Arnold, J., Readdy, T. L., Bessen, D. E. 1994. Molecular evolution of a multigene family in group A streptococci. *Molecular Biology and Evolution* 11(2):208-219.
- Hood, L., and Hunkapiller, T. 1991. Molecular evolution and the immunoglobulin gene superfamily. In "Evolution of Life" (Osawa, S., and Honjo, T., Eds.), pp.123-243, Springer-Verlag, Tokyo.

- Hood, L., Campbell, J. H., Elgin, S. C. 1975. The organization, expression, and evolution of antibody genes and other multigene families. *Annual Review of Genetics* 9:305-353.
- Huber, R. and Bode, W. 1978. Structural basis of the activation and action of trypsin. *Accounts of Chemical Research* 11:114-122.
- Hunkapiller, T. and Hood, L. 1989. Diversity of the immunoglobulin gene superfamily. *Advances in Immunology* 44:1-63.
- Hunkapiller, T., Gorman, J., Koop, B. F. and Hood, L. 1989. Implications of the diversity of the immunoglobulin gene superfamily. *Cold Spring Harbor Symposia on Quantitative Biology* 54(1):15-29.
- Iwabe, N., Kuma, K., and Miyata, T. Evolution of gene families and relationship with organismal evolution: rapid divergence of tissue-specific genes in the early evolution of chordates. *Molecular Biology and Evolution* 13(3):483-493.
- Jin, L., and Nei, M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution* 7(1):82-102.
- Jones, D. T., Taylor, W. R. and Thornton, J. M. 1994. A mutation data matrix for transmembrane proteins. *FEBS Letters* 339:269-275.
- Jones, D. T., Taylor, W. R., and Thornton, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* 8(3):275-282.
- Jukes, T. H., and Cantor, C. R. 1969. Evolution of protein molecules. pp. 21-132. In *Mammalian Protein Metabolism*, H. N. Munro (ed.), Academic Press, New York.
- Jurasek, L., Fackre, D. and Smillie, L. B. 1969. Remarkable homology about the disulfide bridges of a trypsin-like enzyme from *Streptomyces griseus*. *Biochemical and Biophysical Research Communications* 37:99-105.
- Jurášek, L. and Smillie, L. B. 1973. The amino acid sequence of *Streptomyces griseus* trypsin. I. The peptic peptides. *Canadian Journal of Biochemistry* 51:1077-1088.

- Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. and Tabata, S. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp.* strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Research* 3:109-136.
- Kang, J., Wiegand, U., and Müller-Hill, B. 1992. Identification of cDNAs encoding two novel rat pancreatic serine proteases. *Gene* 110:181-187.
- Kasai, H., Isono, S., Mineno, J., Akiyama, H., Kurnit, D.M., Berg, D.E., and Isono, K. 1992. Efficient large-scale sequencing of the *Escherichia coli* genome: implementation of a transposon- and PCR-based strategy for the analysis of ordered λ phage clones. *Nucleic Acids Research* 20(24):6509-6515.
- Kauffman, D. L. 1965. The disulphide bridges of trypsin. *Journal of Molecular Biology* 12:929-932.
- Kim, J. C., Cha, S. H., Jeong, S. T., Oh, S. K. and Byun, S. M. 1991. Molecular cloning and nucleotide sequence of *Streptomyces griseus* trypsin gene. *Biochemical and Biophysical Research Communications* 181:707-713.
- Kimura, M., and Ota, T. 1971. On the rate of molecular evolution. *Journal of Molecular Evolution* 1:1-17.
- Koonin, S. E. 1998. An independent perspective on the human genome project. *Science* 279:36-37.
- Koop, B. F., Rowen, L., Chen, W.-Q., Deshpande, P., Lee, H., and Hood, L. 1993. Sequence length and error analysis of Sequenase® and automated *Taq* cycle sequencing methods. *BioTechniques*. 14(3):442-447.
- Kossiakoff, A. A., Chambers, J. L., Kay, L. M., Stroud, R. M. 1977. Structure of bovine trypsinogen at 1.9 Å resolution. *Biochemistry* 16(4):654-664.

- Lander, E. S., and M. S. Waterman. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* 2:231-239.
- Lange, K., and Boehnke, M. 1982. How many polymorphic genes will it take to span the human genome? *American Journal of Human Genetics* 34(6):842-845.
- Lawrence, C. B., Honda, S., Parrott, N. W., Flood, T. C., Gu, L., Zhang, L., Jain, M., Larson, S. and Myers, E. W. 1994. The genome reconstruction manager: a software environment for supporting high-throughput DNA sequencing. *Genomics* 23:192-201.
- Le Huerou, I., Wicker, C., Guilloteau, P., Toullec, R., Puigserver, A. 1990. Isolation and nucleotide sequence of cDNA clone for bovine pancreatic anionic trypsinogen. *European Journal of Biochemistry* 193:767-773.
- Lévy, P. 1939. Sur la division d'un segment par des points choisis au hasard. *Comptes Rendus* 208:147-149.
- Li, C., and Tucker, P. W. 1993. Exoquence DNA sequencing. *Nucleic Acids Research* 21(5):1239-1244.
- Li, W.-H. 1997. "Molecular Evolution," Sinauer Associates, Sunderland, Massachusetts.
- Li, W.-H., and Graur, D. 1991. "Fundamentals of Molecular Evolution," Sinauer Associates, Sunderland, Massachusetts.
- Liu, G. L. 1968. "Introduction to Combinatorial Mathematics," McGraw Hill, New York.
- Logsdon, J. M., Jr. and Doolittle, W. F. 1997. Origin of antifreeze protein genes: a cool tale in molecular evolution. *Proceedings of the National Academy of Sciences* 94:3485-3487.
- Lütcke, H., Rausch, U., Vasiloudes, P., Scheele, G. A., and Kern, H. F. 1989. A fourth trypsinogen (P23) in the rat pancreas induced by CCK. *Nucleic Acids Research* 17(16):6736.
- MacDonald, R. J., Stary, S. J., Swift, G. H. 1982. Two similar but nonallelic rat pancreatic trypsinogens. Nucleotide sequences of the cloned cDNAs. *Journal of Biological Chemistry* 257(16):9724-9732.

- Male, R., Lorens, J. B., Smalås, A. O., and Torrissen, K. R. 1995. Molecular cloning and characterization of anionic and cationic variants of trypsin from Atlantic salmon. *European Journal of Biochemistry* 232:677-685.
- Maroux, S., Baratti, J. and Desnuelle, P. 1971. Purification and specificity of porcine enterokinase. *Journal of Biological Chemistry* 246:5031-5039.
- Martin-Gallardo, A., J. Lamardin, and A. Carrano. 1994. Shotgun sequencing. In "Automated DNA Sequencing and Analysis" (M. Adams, C. Fields, and J. Venter, Eds.), pp. 37-41, Academic Press, New York.
- Mayo, K. E., Cerelli, G. M., Rosenfeld, M. G. and Evans, R. M. 1985. Characterization of cDNA and genomic clones encoding the precursor to rat hypothalamic growth hormone-releasing factor. *Nature* 314:464-467.
- McKusick, V. 1997. Genomics: structural and functional studies of genomes. *Genomics* 45:244-249.
- Mikeš, O., Holeyšovský, V., Tomášek, V., and Šorm, F. 1966. Covalent structure of bovine trypsinogen. The position of the remaining amides. *Biochemical and Biophysical Research Communications* 24(3):346-352.
- Morgan, P. H., Robinson, N. C., Walsh, K. A. and Neurath, H. 1972. Inactivation of bovine trypsinogen and chymotrypsinogen by diisopropylphosphorofluoridate. *Proceedings of the National Academy of Sciences* 69:3312-3316.
- Mugasimangalam, R. C., Zevin-Sonkin, D., Shwartzburd, J., Rozovskaya, T. A., Sobolev, I. A., Chertkov, O., Ramanathan, V., Lvovsky, L., and Ulanovsky, L. E. 1997. DNA sequencing using differential extension with nucleotide subsets (DENS). *Nucleic Acids Research* 25(4):800-805.
- Müller, W. E. G., Pancer, Z., and Rinkevich, B. 1994. Molecular cloning and localization of a novel serine protease from the colonial tunicate *Botryllus schlosseri*. *Molecular Marine Biology and Biotechnology* 3(2):70-77.

- Murray-Rust, J., McDonald, N. Q., Blundell, T. L., Hosang, M., Oefner, C., Winkler, F. and Bradshaw, R. A. 1993. Topological similarities in TGF-beta 2, PDGF-BB and NGF define a superfamily of polypeptide growth factors. *Structure* 1:153-159.
- Nakai, K. and Kanehisa, M. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14:897-911.
- Neurath, H., Dixon, G. H. 1957. Structure and activation of trypsinogen and chymotrypsinogen. *Federation Proceedings* 16:791-801.
- Newberg, L. A. 1993. "Finding, Evaluating, and Counting DNA Physical Maps," Doctoral Thesis, University of California at Berkeley.
- Newberg, L. A. 1996. The number of clone orderings. *Discrete Applied Mathematics* 69(3):233-245.
- Northrup, J. H., Kunitz, M., and Herriott, R. 1948. "Crystalline Enzymes," Columbia University Press, New York.
- Ohno, S. 1978. The significance of gene duplication in immunoglobulin evolution (epimethean natural selection and promethean evolution). In "Immunoglobulins" (Litman, G. W., and Good, R. A., Eds.), pp.197-204, Plenum, New York.
- Olson, M. V., and Green, P. 1993. Criterion for the completeness of large-scale physical maps of DNA. *Cold Spring Harbor Symposia on Quantitative Biology*. 53:349-355.
- Ota, T., and Kimura, M. 1971. On the constancy of the evolutionary rate of cistrons. *Journal of Molecular Evolution* 1:18-25.
- Pancer, Z., Leuck, J., Rinkevich, B., Steffen, R., Müller, I., and Müller, W. E. G. 1996. Molecular cloning and sequencing analysis of two cDNAs coding for putative anionic trypsinogens from the colonial urochordate *Botryllus schlosseri* (Ascidacea). *Molecular Marine Biology and Biotechnology* 5(4):326-333.
- Pease, L. R., Horton, R. M., Pullen, J. K. and Yun, T. J. 1993. Unusual mutation clusters provide insight into class I gene conversion mechanisms. *Molecular and Cellular Biology* 13:4374-4381.

- Petes, T. D. and Hill, C. W. 1988. Recombination between repeated genes in microorganisms. *Annual Review of Genetics* 22:147-168.
- Petes, T. D., Malone, R. E., and Symington, L. S. 1991. Recombination in yeast. In "The Molecular and Cellular Biology of the Yeast *Saccharomyces*: Genome Dynamics, Protein Synthesis, and Energetics, Volume I" (J. Broach, E. Jones, and J. Pringle, Eds.), pp.407-521, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Pinsky, S. D., LaForge, K. S., and Scheele, G. 1985. Differential regulation of trypsinogen mRNA translation: full-length mRNA sequences encoding two oppositely charged trypsinogen isoenzymes in the dog pancreas. *Molecular and Cellular Biology* 5(10):2669-2676.
- Port, E., Sun, F., Martin, D., and Waterman, M. S. 1995. Genomic mapping by end-characterized random clones: a mathematical analysis. *Genomics* 26:84-100.
- Rapoport, T. A. 1990. Protein transport across the ER membrane. *Trends in Biochemical Sciences* 15:355-358.
- Raport, C. J., Schweickart, V. L., Chantry, D., Eddy, R. L., Jr., Shows, T. B., Godiska, R. and Gray, P. W. 1996. New members of the chemokine receptor gene family. *Journal of Leukocyte Biology* 59:18-23.
- Ravitch, M. M. 1973. Ectopic Pancreas – not so rare, not so innocent. *Medical Times* 101(6):57-59.
- Read, R. J. and James, M. N. 1988. Refined crystal structure of *Streptomyces griseus* trypsin at 1.7 Å resolution. *Journal of Molecular Biology* 200:523-551.
- Reynaud, C. A., Anquez, V., Dahan, A. and Weill, J. C. 1985. A single rearrangement event generates most of the chicken immunoglobulin light chain diversity. *Cell* 40:283-291.
- Richards, S., Muzny, D. M., Civitello, D. M., Lu, F., and Gibbs, R. A. 1994. Sequence map gaps and directed reverse sequencing for the completion of large sequencing

- projects. In "Automated DNA Sequencing and Analysis" (M. Adams, C. Fields, and J. Venter, Eds.), pp. 191-198, Academic Press, New York.
- Roach, J. C. 1995. Random subcloning. *Genome Research* 5:464-473.
- Roach, J. C., Boysen, C., Wang, K., and Hood, L. 1995. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* 26:345-353.
- Roach, J. C., Wang, K., Gan, L., Hood, L. 1997. The molecular evolution of the vertebrate trypsinogens. *Journal of Molecular Evolution* 45(6):640-652.
- Robbins, H. E. 1944. On the measure of a random set. *Annals of Mathematical Statistics* 15:70-74.
- Robbins, H. E. 1945. On the measure of a random Set. II. *Annals of Mathematical Statistics* 16:342-347.
- Rowen, L., and Koop, B. F. 1994. Zen and the art of large-scale genomic sequencing. In "Automated DNA Sequencing and Analysis" (M. Adams, C. Fields, and J. Venter, Eds.), pp. 167-174, Academic Press, New York.
- Rowen, L., Koop, B. F., and Hood, L. 1996. The complete 685-kilobase DNA sequence of the human β T cell receptor locus. *Science* 272:1755-1762.
- Rowen, L., Mahairas, G., and Hood, L. 1997. Sequencing the human genome. *Science* 278:605-607.
- Rypniewski, W. R, Perrakis, A., Vorgias, C. E., and Wilson, K. S. 1994. Evolutionary divergence and conservation of trypsin. *Protein Engineering* 7(1):57-64.
- Scheele, G., Bartelt, D. and Bieger, W. 1981. Characterization of human exocrine pancreatic proteins by two- dimensional isoelectric focusing/sodium dodecyl sulfate gel electrophoresis. *Gastroenterology* 80:461-473.
- Schick, J., Kern, H. and Scheele, G. 1984. Hormonal stimulation in the exocrine pancreas results in coordinate and anticonordinate regulation of protein synthesis. *Journal of Cell Biology* 99:1569-1574.

- Schneider, T. D. and Stephens, R. M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* 18:6097-6100.
- Seboun, E., Houghton, L., Hatem, C. J., Jr., Lincoln, R. and Hauser, S. L. 1993. Unusual organization of the human T-cell receptor beta-chain gene complex is linked to recombination hotspots. *Proceedings of the National Academy of Sciences* 90:5026-5029.
- Seboun, E., Joshi, N. and Hauser, S. L. 1992. Haplotypic origin of beta-chain genes expressed by human T-cell clones. *Immunogenetics* 36:363-368.
- Seboun, E., Robinson, M. A., Kindt, T. J. and Hauser, S. L. 1989. Insertion/deletion-related polymorphisms in the human T cell receptor beta gene complex. *Journal of Experimental Medicine* 170:1263-1270.
- Shi, Y.-B., and Brown, D. D. 1990. Developmental and thyroid hormone-dependent regulation of pancreatic genes in *Xenopus laevis*. *Genes and Development* 4(7):1107-1113.
- Siegel, A. F. 1979. Asymptotic coverage distributions on the circle. *Annals of Probability* 74:651-661.
- Siegel, A. F., and L. Holst. 1982. Covering the circle with random arcs of random sizes. *J Applied Probability* 19:373-381.
- Siegel, A. F., Roach, J. C., and van den Engh, G. 1998a. Expectation and variance of true and false fragment matches in DNA restriction mapping. *Journal of Computational Biology*. In Press.
- Siegel, A. F., Roach, J. C., Magness, C., Thayer, E., and van den Engh, G. 1998b. Optimization of restriction fragment DNA mapping. *Journal of Computational Biology*. In Press.
- Siegel, A. F., Trask, B., Roach, J. C., Mahairas, G. G., Hood, L., and van den Engh, G. Analysis of STC sequencing strategies. In Preparation.

- Siemieniak, D. L., Sieu, L. C., and Slightom, J. L. 1991. Strategy and methods for directly sequencing cosmid clones. *Analytical Biochemistry* 192:441-448.
- Smalas, A. O., Heimstad, E. S., Hordvik, A., Willassen, N. P. and Male, R. 1994. Cold adaptation of enzymes: structural comparison between salmon and bovine trypsins. *Proteins* 20:149-166.
- Smith, M. W., Holmsen, A. L., Wei, Y. H., Peterson, M., and Evans, G. A. 1994. Genomic sequence sampling: A strategy for high resolution sequence-based physical mapping of complex genomes. *Nature Genetics* 7(1):40-47.
- Smyth, M. J., O'Connor, M. D. and Trapani, J. A. 1996. Granzymes: a variety of serine protease specificities encoded by genetically distinct subfamilies. *Journal of Leukocyte Biology* 60:555-562.
- Solomon, H. 1978. "Geometric Probability," Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.
- Staden, R. 1980. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Research* 8:3673-3694.
- Steinhilber, W., Poensgen, J., Rausch, U., Kern, H. F. and Scheele, G. A. 1988. Translational control of anionic trypsinogen and amylase synthesis in rat pancreas in response to caerulein stimulation. *Proceedings of the National Academy of Sciences* 85:6597-6601.
- Stevens, W. L. 1939. Solution to a geometrical problem in probability. *Annals of Eugenics* 9:315-320.
- Stevenson, B. J., Hagenbüchle, O., and Wellauer, P. K. 1986. Sequence organisation and transcriptional regulation of the mouse elastase II and trypsin genes. *Nucleic Acids Research* 14(21):8307-8330.
- Stoltzfus, A., Logsdon, J. M., Jr., Palmer, J. D. and Doolittle, W. F. 1997. Intron "sliding" and the diversity of intron positions. *Proceedings of the National Academy of Sciences* 94:10739-10744.

- Stubbs, L. 1992. Long-range walking techniques in positional cloning strategies. *Mammalian Genome* 3:127-142.
- Suyama, M., Matsuo, Y. and Nishikawa, K. 1997. Comparison of protein structures using 3D profile alignment. *Journal of Molecular Evolution* 44(Suppl. 1):S163-173.
- Sweet, R. M., Wright, H. T., Janin, J., Chothia, C. H., and Blow, D. M. 1974. Crystal structure of the complex of porcine trypsin with soybean trypsin inhibitor (Kunitz) at 2.6-Å resolution. *Biochemistry* 13(20):4212-4228.
- Szabo, C. I., Fransisco, L. V., Roach, J., Argonza, R., Wagner, L. A., Ostrander, E. A., King, M.-C. 1996. Human, canine, and murine BRCA1 genes: sequence comparison among species. *Human Molecular Genetics* 5(9):1289-1298.
- Tani, T., Kawashima, I., Mita, K., Takiguchi, Y. 1990. Nucleotide sequence of the human pancreatic trypsinogen III cDNA. *Nucleic Acids Research* 18(6):1631.
- Tatusov, R. L., Koonin, E. V. and Lipman, D. J. 1997. A genomic perspective on protein families. *Science* 278:631-637.
- Taylor, W. R., and Jones, D. T. 1993. Deriving an amino acid distance matrix. *Journal of Theoretical Biology* 164(1):65-83.
- The MITRE Corporation. 1997. "JASON Report JSR-97-315," The MITRE Corporation, McLean, Virginia.
- Thoeni, R. F., and Gedgaudas, R. K. 1980. Ectopic pancreas: usual and unusual features. *Gastrointestinal Radiology* 5(1):37-42.
- Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22:4673-4680.
- Thorne, J. L., Goldman, N., and Jones, D. T. 1996. Combining protein evolution and secondary structure. *Molecular Biology and Evolution* 13(5):666-673.

- Titani, K., Ericsson, L. H., Neurath, H., and Walsh, K. A. 1975. Amino acid sequence of dogfish trypsinogen. *Biochemistry* 14(7):1358-1366.
- Titani, K., Sasagawa, T., Woodbury, R. G., Ericsson, L. H., Dörsam, H., Kraemer, M., Neurath, H., and Zwilling, R. 1983. Amino acid sequence of crayfish (*Astacus fluviatilis*) trypsin I_f. *Biochemistry* 22:1459-1465.
- Townes, P. L. 1972. Trypsinogen deficiency and other proteolytic deficiency diseases. *Birth Defects: Original Article Series* 8(2):95-101.
- Venter, J. C., Smith, H. O., and Hood, L. 1996. A new strategy for genome sequencing. *Nature* 381(6581):364-366.
- Vingron, M., and von Haeseler, A. 1997. Towards integration of multiple alignment and phylogenetic tree construction. *Journal of Computational Biology* 4(1):23-34.
- von Heijne, G. 1985. Signal sequences. The limits of variation. *Journal of Molecular Biology* 184:99-105.
- von Heijne, G. 1986. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research* 14(11):4683-4690.
- Wachtel, S. S., and Tiersch, T. R. 1993. Variations in genome mass. *Comparative Biochemistry and Physiology* 104B(2):207-213.
- Walsh, K. A. 1970. Trypsinogens and trypsins of various species. *Methods in Enzymology* 19:41-63.
- Walsh, K. A., and Wilcox, P. E. 1970. Serine proteases. *Methods in Enzymology* 19:31-41.
- Wang, K., Gan, L., Lee, I., and Hood, L. 1995. Isolation and characterization of the chicken trypsinogen gene family. *Biochemical Journal* 307:471-479.
- Waterman, M. S. 1995. "Introduction to Computational Biology: Maps, sequences, and genomes," Chapman & Hall, London.

- Whitcomb, D. C., Gorry, M. C., Preston, R. A., Furrey, W., Sossenheimer, M. J., Ulrich, C. D., Martin, S. P., Gates, L. K., Amann, S. T., Toskes, P. P., Liddle, R., McGrath, K., Uomo, G., Post, J. C., and Ehrlich, G. D. 1996. Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. *Nature Genetics* 14:141-145.
- Whitworth, W. A. 1897a. "Choice and Chance," Cambridge University Press, Cambridge.
- Whitworth, W. A. 1897b. "Exercises in Choice and Chance," Cambridge University Press, Cambridge.
- Wiegand, U., Corbach, S., Minn, A., Kang, J., and Müller-Hill, B. 1993. Cloning of the cDNA encoding human brain trypsinogen and characterization of its product. *Gene* 136:167-175.
- Wiemann, S., Stegemann, J., Zimmermann, J., Voss, H., Benes, V., and Ansorge, W. 1996. "Doublex" fluorescent DNA sequencing: two independent sequences obtained simultaneously in one reaction with internal labeling and unlabeled primers. *Analytical Biochemistry* 234(2):166-174.
- Wills, C. 1991. "Exons, Introns, and Talking Genes: The Science Behind the Human Genome Project," pp. 92-95, Harper Collins, New York.
- Wu, C. F. J. 1986. Jackknife, bootstrap and other resampling plans in regression analysis. *Annals of Statistics* 14:1261-1295.
- Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* 10(6):1396-1401.
- Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39(3):306-314.
- Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139(2):993-1005.
- Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* 13:555-556.

- Yee, D. P. and Dill, K. A. 1993. Families and the structural relatedness among globular proteins. *Protein Science* 2:884-899.
- Zhang, Z., Pearson, W. R., and Miller, W. 1997. Aligning a DNA sequence with a protein sequence. *Journal of Computational Biology* 4(3):339-349.
- Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W. and Wilson, A. C. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proceedings of the National Academy of Sciences* 77:2158-2162.
- Zwilling, R., and Neurath, H. 1981. Invertebrate proteases. *Methods in Enzymology* 80:633-664.

APPENDIX A. DOUBLE-BARREL SHOTGUN ALGEBRA

I define the length of an insert to be 1, without loss of generality. Insert length is assumed to be constant. For this appendix, in a departure from the notation of the main text, I assume a constant sequence read length of length f , with $f \in [0, 0.5]$. If $f > 0.5$, the fragments overlap at the center of the insert. To maintain as much simplicity as possible, I will consider only the limiting case as $T \rightarrow 0$. I will explicitly calculate only the probability of obtaining a single scaffold (i.e., all clones form one scaffold). However, the equations developed along the way can be adapted with a healthy dose of algebra to give distributions for the lengths of scaffolds as well as other variables of interest. I assume stepwise addition of characterized inserts to a project, in order to aid my descriptions. I will refer to these inserts (i.e., “clones”) as C_1 , C_2 , or C_3 , with the subscript designating the order of addition to the project. Adding all the clones simultaneously does not alter the results. I assume a circular target with $G > n$. A less restrictive assumption on target size can be made with little loss of accuracy.¹

To aid my descriptions I will refer to each sequence read by whether or not it is the right or left read from an insert, with a subscript for which of the inserts the sequence read is obtained from. Thus the left read from C_1 is designated L_1 ; the right read from C_3 is R_3 .

I will use the symbol \cap to designate overlap. I will use S to indicate the state of a project with a single scaffold. Therefore $P(S|L_1 \cap L_2)$ means “the probability of a project having a single scaffold given that the left end of clone 1 overlaps the left end of clone 2.” $P(S|(L_1 \cap L_2) \wedge \sim (L_1 \cap R_3))$ means “the probability of a project having a single scaffold given that the left end of clone 1 overlaps the left end of clone 2 but not the right end of clone 3.”

¹ If G is smaller than this, it becomes possible for islands to wrap around on themselves. This alters the number and character of the possible topologies. It will tend to increase the probability of a single island, possibly quite significantly. If the target is linear, slight “edge effects” alter the calculations. These effects tend to be minor, particularly if $G \gg I$. Therefore, the calculations presented here as exact for really big circular targets are actually better approximations to real linear targets than they are to real circular targets.

To aid in the descriptions that follow, cartoon sketches of possible clone topologies are provided in Figure A.1 and Figure A.2.

A.1 ONE CLONE

One clone, by definition, will always form one scaffold.

A.2 TWO CLONES

There are two cases to consider.

A.2.1 CASE 1 $\frac{1}{2} \geq f \geq \frac{1}{3}$

The second clone will intersect the first clone with probability $2/G$. Note that the average number of clones in a scaffold will be:

$$\frac{1\left(\frac{2}{G}\right) + 2\left(1 - \frac{2}{G}\right)}{2\left(\frac{2}{G}\right) + 1\left(1 - \frac{2}{G}\right)} = \frac{2 - \frac{2}{G}}{1 + \frac{2}{G}}$$

A.2.2 CASE 2 $f < \frac{1}{3}$

The second clone will intersect the first clone with probability $6f/G$. Note that the average number of clones in a scaffold will be:

$$\frac{1\left(\frac{6f}{G}\right) + 2\left(1 - \frac{6f}{G}\right)}{2\left(\frac{6f}{G}\right) + 1\left(1 - \frac{6f}{G}\right)} = \frac{2 - \frac{6f}{G}}{1 + \frac{6f}{G}}$$

A.3 THREE CLONES

The cases with one and two clones were relatively simple. The case with three clones is also simple, in that no complex mathematics is needed, but some care must be

taken to ensure that all the topologies are properly accounted for. There are three cases to consider.

A.3.1 CASE 1 $\frac{1}{2} \geq f \geq \frac{1}{3}$

When $f \geq \frac{1}{3}$, a fragment cannot fall in the unsequenced gap between the two characterized ends of a clone without overlapping at least one of the ends. This limits the number of topologies that must be considered.

There are two possible configurations for the first two clones. Either they overlap, or they do not.

A.3.1.1 configuration 1 $C_2 \cap C_1$

This configuration will occur with probability $2/G$. All possible extents of overlap are equally likely. Fixing the leftmost end of C_1 at 1, then the leftmost end of C_2 will vary from 0 to 2. The expected probability of C_3 overlapping either or both of C_1 and C_2 can be calculated as

$$\frac{\frac{1}{G} \int_0^1 (3-x) dx + \frac{1}{G} \int_1^2 (1+x) dx}{2} = \frac{5}{2G}$$

For example, when the leftmost end of C_2 is at the origin, the leftmost end of C_3 can be anywhere from -1 to 2 and still overlap C_1 or C_2 , giving it a probability of $3/G$ of joining the scaffold. Due to symmetry, the two integrals contribute equally.

A.3.1.2 configuration 2 $\sim(C_2 \cap C_1)$

This configuration will occur with probability $1 - \frac{2}{G}$. All possible non-overlapping states are equally likely. Fixing the rightmost end of C_1 at 0, and letting x represent the rightmost end of C_2 , the probability of C_3 overlapping both of C_1 and C_2 is:

$$\frac{\frac{1}{G} \int_0^1 (1-x) dx + \frac{1}{G} \int_1^{G-3} 0 dx + \frac{1}{G} \int_{G-3}^{G-2} (x - (G-3)) dx}{G-2} = \frac{1}{G(G-2)}$$

Again, due to symmetry, the first and last integrals give the same result.

A.3.1.3 combining the probabilities for the two configurations

The probability of obtaining a single scaffold given $\frac{1}{2} \geq f \geq \frac{1}{3}$ is the sum of the probabilities from the last two sub-sections:

$$P(S) = \left(1 - \frac{2}{G}\right) \frac{1}{G(G-2)} + \frac{2}{G} \frac{5}{2G} = \frac{6}{G^2}$$

A.3.2 CASE 2 $\frac{1}{3} \geq f \geq \frac{1}{4}$

There are four possible topologies for the first two clones.

A.3.2.1 configuration 1 $L_1 \cap L_2$

This configuration will occur with probability $2f/G$. Explicitly writing out the probability of the third clone intersecting the scaffold formed by the first two clones:

$$P(S) =$$

$$P(S|R_3 \cap L_1)$$

$$+ P[S | (R_3 \cap R_1) \wedge \sim(R_3 \cap L_1)]$$

$$+ P[S | (R_3 \cap L_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1)]$$

$$+ P[S | (R_3 \cap R_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2)]$$

$$+ P[S | (L_3 \cap L_1) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2)]$$

$$+ P[S | (L_3 \cap R_1) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2) \wedge \sim(L_3 \cap L_1)]$$

$$+ P[S | (L_3 \cap L_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2) \wedge \sim(L_3 \cap L_1) \wedge \sim(L_3 \cap R_1)]$$

$$+P[SI(L_3 \cap R_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2) \wedge \sim(L_3 \cap L_1) \wedge \sim(L_3 \cap R_1) \wedge \sim(L_3 \cap L_2)] =$$

$$2f/G$$

$$+2f/G$$

$$+ \frac{-1 + 8f - 14f^2}{4fG}$$

$$+ \frac{-1 + 8f - 14f^2}{4fG}$$

$$+0$$

$$+ \frac{8f - 8f^2 - 1}{4fG}$$

$$+0$$

$$+ \frac{-1 + 8f - 14f^2}{4fG} =$$

$$\frac{8f - 8f^2 - 1}{4fG} + 3 \left(\frac{-1 + 8f - 14f^2}{4fG} \right) + \frac{16f^2}{4fG} =$$

$$\frac{-2 + 16f - 17f^2}{2fG}$$

The following identities help in the above calculation:

$$1. P(SI R_3 \cap L_1) = 2f/G$$

$$2. R_3 \text{ cannot hit } L_1 \text{ if } R_3 \text{ hits } R_1 \text{ so } P[SI(R_3 \cap R_1) \wedge \sim(R_3 \cap L_1)] = P(SI R_3 \cap L_1) = 2f/G$$

$$3. P[SI(R_3 \cap L_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1)] =$$

$$\begin{aligned}
& \frac{\frac{1}{2} \left(\int_0^f (f-x) dx + \int_f^{4f-1} (x-f) dx + \int_{4f-1}^{2f} (x-f) - [x - (1-2f)] dx \right)}{2f} = \\
& \frac{\frac{1}{2} \left(\int_0^f (f-x) dx + \int_f^{1-2f} (x-f) dx + \int_{1-2f}^{2f} (1-3f) dx \right)}{2f} = \\
& \frac{\frac{1}{2} \left(\left[fx - \frac{1}{2} x^2 \right]_0^f + \left[\frac{1}{2} x^2 - fx \right]_f^{1-2f} + (1-3f)(4f-1) \right)}{2f} = \\
& \frac{\frac{1}{2} \left(\frac{f^2}{2} + \frac{1-6f+9f^2}{2} + (-12f^2+7f-1) \right)}{2f} = \\
& \frac{\frac{1}{2} \left(\frac{-1+8f-14f^2}{2} \right)}{2f} = \\
& \frac{-1+8f-14f^2}{4fG}
\end{aligned}$$

4. If R_3 hits R_2 it cannot hit L_2 so:

$$\begin{aligned}
& P[SI(R_3 \cap R_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2)] = P[SI(R_3 \cap R_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1)] = \\
& \frac{-1+8f-14f^2}{4fG}
\end{aligned}$$

and this probability is in turn equal to $P[SI(R_3 \cap L_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1)]$.

5. If L_3 hits L_1 then R_3 hits R_1 , so:

$$P[SI(L_3 \cap L_1) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2)] = 0$$

6. If L_3 hits R_1 then L_3 cannot hit L_1 and R_3 cannot hit L_2 , L_1 , or R_1 so:

$$P[SI(L_3 \cap R_1) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2) \wedge \sim(L_3 \cap L_1)] =$$

$$P[SI(L_3 \cap R_1) \wedge \sim(R_3 \cap R_2)] =$$

$$\begin{aligned} & \frac{\frac{1}{G} \left(\int_0^{1-2f} 2f dx + \int_{1-2f}^{2f} (1-x) dx \right)}{2f} = \\ & \frac{\frac{1}{G} \left((2f - 4f^2) + (2f - \frac{1}{2}) \right)}{2f} = \\ & \frac{\frac{1}{G} \left(4f - 4f^2 - \frac{1}{2} \right)}{2f} = \\ & \frac{8f - 8f^2 - 1}{4fG} \end{aligned}$$

7. If L_3 hits L_2 then R_3 hits R_2 so:

$$P[SI(L_3 \cap L_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2) \wedge \sim(L_3 \cap L_1) \wedge \sim(L_3 \cap R_1)] = 0$$

8. If L_3 hits R_2 then: R_3 cannot hit L_1 , L_2 , or R_2 ; L_3 cannot hit L_2 ; L_3 hits L_1 if and only if R_3 hits R_1 . Therefore:

$$P[SI(L_3 \cap R_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2) \wedge \sim(L_3 \cap L_1) \wedge \sim(L_3 \cap R_1) \wedge \sim(L_3 \cap L_2)] =$$

$$P[SI(L_3 \cap R_2) \wedge \sim(R_3 \cap R_1) \wedge \sim(L_3 \cap R_1)] =$$

$$\frac{-1 + 8f - 14f^2}{4fG}$$

This last equality is symmetrical to the case in identity 4.

A.3.2.2 configuration 2 $\sim(C_1 \cap C_2)$

Since C_1 and C_2 do not overlap, in order for there to be one scaffold after the addition of C_3 , it must link C_1 and C_2 . I consider only the case where C_1 is to the left of C_2 . This case is symmetrical with the case where C_1 lies to the right of C_2 . The probability of one or the other of these cases occurring is $1-2/G$. However, with probability $1-4/G$, the

first two clones will be greater than a clone length apart, with zero chance that the third clone will link them into one scaffold. I will condition the following probabilities on (i) C_1 lies to the left of C_2 , and (ii) C_1 and C_2 are not separated by more than one clone length. With these conditions, the probability of C_3 hitting both C_1 and C_2 is:

$$P[SI(R_3 \cap R_1) \wedge (R_3 \cap L_2)] + P[SI(L_3 \cap R_1) \wedge (L_3 \cap L_2)] + P[SI(L_3 \cap R_1) \wedge (R_3 \cap L_2)] =$$

$$\frac{-1 + 8f - 6f^2}{2G}$$

The following identities help in the above calculation:

$$1. P[SI(R_3 \cap R_1) \wedge (R_3 \cap L_2)] = \int_0^f (f - x) dx + \int_f^1 0 = \frac{f^2}{2G}$$

$$2. P[SI(L_3 \cap R_1) \wedge (L_3 \cap L_2)] = \int_0^f (f - x) dx + \int_f^1 0 = \frac{f^2}{2G}$$

$$3. P[SI(L_3 \cap R_1) \wedge (R_3 \cap L_2)] = \int_0^{1-2f} (x + 4f - 1) dx + \int_{1-2f}^1 (1 - x) dx = \frac{-1 + 8f - 8f^2}{2G}$$

A.3.2.3 configuration 3 $R_1 \cap L_2$

This configuration will occur with probability $2f$. Note that this is symmetrical with the configuration $L_1 \cap R_2$, which will also occur with probability $2f$, so I will not explicitly address $L_1 \cap R_2$. Explicitly writing out the probability of the third clone intersecting the scaffold formed by the first two clones:

$$P(S) =$$

$$P(S|R_3 \cap L_1)$$

$$+ P(S|R_3 \cap R_1)$$

$$+P[SI(R_3 \cap L_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1)]$$

$$+P[SI(R_3 \cap R_2) \wedge \sim(R_3 \cap R_1)]$$

$$+P[SI(L_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2)]$$

$$+P[SI(L_3 \cap R_2) \wedge \sim(L_3 \cap R_1)] =$$

$$\frac{-1 + 8f - 7f^2}{fG}$$

The following identities help in the above calculation:

$$1. P(SI R_3 \cap L_1) = 2f/G$$

$$2. P(SI R_3 \cap R_1) = 2f/G$$

$$3. P[SI(R_3 \cap L_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1)] = \frac{-1 + 8f - 14f^2}{4fG}$$

$$4. P[SI(R_3 \cap R_2) \wedge \sim(R_3 \cap R_1)] = \frac{\int_0^{4f-1} (x+1-2f)dx + \int_{4f-1}^{2f} 2f dx}{2fG} = \frac{-1 + 8f - 8f^2}{4fG}$$

$$5. P[SI(L_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2)] = P[SI(L_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(L_3 \cap L_2)] =$$

$$\frac{\frac{1}{G} \left(\int_0^f (f-x)dx + \int_f^{1-2f} (x-f)dx + \int_{1-2f}^{2f} (1-3f)dx \right)}{2f} =$$

$$\frac{-1 + 8f - 14f^2}{4fG}$$

$$6. P[SI(L_3 \cap R_2) \wedge \sim(L_3 \cap R_1)] = \frac{\int_0^{4f-1} (x+1-2f)dx + \int_{4f-1}^{2f} 2f dx}{2fG} = \frac{-1+8f-8f^2}{4fG}$$

A.3.2.4 configuration 4 $(C_1 \cap C_2) \wedge \sim((R_1 \vee L_1) \cap (R_2 \vee L_2))$

This is the case where the clones overlap, but still form two scaffolds because none of their characterized ends overlap. The clones are interleaved. Again, this can occur in two symmetrical states, one with the left end of C_1 to the left of the left end of C_2 , and one with C_1 to the right. Each of these states will occur with probability $(1-3f)/G$. I will condition on the state with C_1 to the left in the following. Explicitly writing out the probability of the third clone intersecting the scaffold formed by the first two clones:

$$P(S) =$$

$$P[SI(R_3 \cap L_1) \wedge (R_3 \cap L_2)]$$

$$+ P[SI(R_3 \cap R_1) \wedge (R_3 \cap L_2)]$$

$$+ P[SI(R_3 \cap R_1) \wedge (R_3 \cap R_2)]$$

$$+ P[SI(L_3 \cap R_1) \wedge (L_3 \cap R_2)] =$$

$$\frac{10f-2}{G}$$

The following identities help in the above calculation:

$$1. P[SI(R_3 \cap L_1) \wedge R_3 \cap L_2] = \frac{\int_0^{1-3f} (f-x)dx}{(1-3f)G} = \frac{5f-1}{2G}$$

$$2. P[SI(R_3 \cap R_1) \wedge (R_3 \cap L_2)] = \frac{\int_0^{1-3f} (f-x) dx}{(1-3f)G} = \frac{5f-1}{2G}$$

$$3. P[SI(R_3 \cap R_1) \wedge (R_3 \cap R_2)] = \frac{\int_0^{1-3f} (f-x) dx}{(1-3f)G} = \frac{5f-1}{2G}$$

$$4. P[SI(L_3 \cap R_1) \wedge (L_3 \cap R_2)] = \frac{\int_0^{1-3f} (f-x) dx}{(1-3f)G} = \frac{5f-1}{2G}$$

A.3.2.5 combining the probabilities for the four configurations

With each possible configuration represented by a roman numeral, we have:

$$P(S) = P(i)P(Sli) + P(ii)P(Slii) + P(iii)P(Sliii) + P(iv)P(Sliv) =$$

$$\begin{aligned} & \left(\frac{2f}{G} \right) \left(\frac{-2+16f-17f^2}{2fG} \right) + 2 \left(\frac{1}{G} \right) \left(\frac{-1+8f-6f^2}{2G} \right) \\ & + 2 \left(\frac{2f}{G} \right) \left(\frac{-1+8f-7f^2}{fG} \right) + 2 \left(\frac{1-3f}{G} \right) \left(\frac{-2+10f}{G} \right) = \\ & \frac{-11+88f-111f^2}{G^2} \end{aligned}$$

A.3.3 CASE 3 $\frac{1}{4} \geq f \geq 0$

There are four possible topologies for the first two clones. These are the same topologies considered in Case 2. However, the probabilities must be computed differently, as some of the final topologies that were previously impossible are now possible. Each of the configurations is as described in Case 2, so I will omit explanations where they are identical to Case 2.

A.3.3.1 configuration 1 $L_1 \cap L_2$

$$P(S) =$$

$$P(S|R_3 \cap L_1)$$

$$+P(S|R_3 \cap R_1)$$

$$+P[S|(R_3 \cap L_2) \wedge \sim(R_3 \cap L_1)]$$

$$+P[S|(R_3 \cap R_2) \wedge \sim(R_3 \cap R_1)]$$

$$+P(S|L_3 \cap R_1)$$

$$+P[S|(L_3 \cap R_2) \wedge \sim(L_3 \cap R_1)] =$$

$$2f/G$$

$$+2f/G$$

$$+f/2G$$

$$+f/2G$$

$$+2f/G$$

$$+f/2G$$

$$=15f/2G$$

$$\text{A.3.3.2 configuration 2} \quad \sim(C_1 \cap C_2)$$

$$P(S) = P[S|(R_3 \cap R_1) \wedge (R_3 \cap L_2)] + P[S|(L_3 \cap R_1) \wedge (L_3 \cap L_2)] + P[S|(L_3 \cap R_1) \wedge (R_3 \cap L_2)] = \frac{5f^2}{G}$$

The following identities help in the above calculation:

$$1. \int_0^f (f-x)dx + \int_f^1 0 = \frac{f^2}{2G}$$

$$2. \int_0^f (f-x)dx + \int_f^1 0 = \frac{f^2}{2G}$$

$$3. \int_0^{1-4f} 0dx + \int_{1-4f}^{1-2f} (x+4f-1)dx + \int_{1-2f}^1 (1-x)dx = \frac{4f^2}{G}$$

A.3.3.3 *configuration 3* $R_1 \cap L_2$

$$P(S)=P(S|R_3 \cap L_1)$$

$$+P(S|R_3 \cap R_1)$$

$$+P[S|(R_3 \cap L_2) \wedge \sim (R_3 \cap R_1)]$$

$$+P(S|R_3 \cap R_2)$$

$$+P[S|(L_3 \cap R_1) \wedge \sim (L_3 \cap L_2)]$$

$$+P(S|L_3 \cap R_2)=$$

$$\frac{9f}{G}$$

A.3.3.4 *configuration 4* $(C_1 \cap C_2) \wedge \sim ((R_1 \vee L_1) \cap (R_2 \vee L_2))$

$$P(S)=$$

$$P[S|(R_3 \cap L_1) \wedge (R_3 \cap L_2)]$$

$$+P[S|(R_3 \cap R_1) \wedge (R_3 \cap L_2)]$$

$$+P[S|(R_3 \cap R_1) \wedge (R_3 \cap R_2)]$$

$$+P[SI(L_3 \cap R_1) \wedge (L_3 \cap R_2)] =$$

$$\frac{2f^2}{(1-3f)G}$$

The following identities help in the above calculation:

$$1. P[SI(R_3 \cap L_1) \wedge R_3 \cap L_2] = \frac{\int_0^f (f-x)dx + \int_f^{1-3f} 0}{(1-3f)G} = \frac{f^2}{2(1-3f)G}$$

$$2. P[SI(R_3 \cap R_1) \wedge (R_3 \cap L_2)] = \frac{\int_0^f (f-x)dx + \int_f^{1-3f} 0}{(1-3f)G} = \frac{f^2}{2(1-3f)G}$$

$$3. P[SI(R_3 \cap R_1) \wedge (R_3 \cap R_2)] = \frac{\int_0^f (f-x)dx + \int_f^{1-3f} 0}{(1-3f)G} = \frac{f^2}{2(1-3f)G}$$

$$4. P[SI(L_3 \cap R_1) \wedge (L_3 \cap R_2)] = \frac{\int_0^f (f-x)dx + \int_f^{1-3f} 0}{(1-3f)G} = \frac{f^2}{2(1-3f)G}$$

A.3.3.5 combining the probabilities for the four configurations

$$P(S) = P(i)P(Sli) + P(ii)P(Slii) + P(iii)P(Sliii) + P(iv)P(Sliv) =$$

$$\left(\frac{2f}{G}\right)\left(\frac{15f}{2G}\right) + 2\left(\frac{1}{G}\right)\left(\frac{5f^2}{G}\right) + 2\left(\frac{2f}{G}\right)\left(\frac{9f}{G}\right) + 2\left(\frac{1-3f}{G}\right)\left(\frac{2f^2}{(1-3f)G}\right) =$$

$$\frac{65f^2}{G^2}$$

A.4 MORE THAN THREE CLONES

The algebra quickly gets more difficult. However, a few simple statements can be made.

As f becomes a smaller fraction of the clone length, more topologies become possible. This is the same as holding f constant and increasing the clone length, which is what is done in actual practice when longer inserts are used to build clone libraries. This is a partial explanation for why long clones are better. The more topologies the clones have available to them in order to form a single scaffold, the more likely they are to form a single scaffold. The probability of clones forming a particular topology never drops as the ratio of clone length to f increases. This probability is illustrated in Figure A.3 for the case of three inserts.

It is interesting to note that this curve is not smooth at the ratio of three. I predict that this is the only point at which this curve will not be smooth, regardless of the number of inserts (note that the curve is not defined for ratios less than two). Below a ratio of three, the opposite ends of the insert are present in each other's potential region of overlap. This results in a concave curve. Above a ratio of three, I predict that the curve will be both smooth and convex.

The number of topologies obviously does not rise continuously as f diminishes. Rather the number of topologies takes discrete jumps at each $f \in \{1/x \mid x \text{ is an integer} \leq n\}$. The maximum number of topologies is reached when $f=1/n$, so there is no further gain in the number of topologies by increasing the clone length beyond nf . Since, in practice, n is universally greater than 100, and f for sequencing projects is at least 400 bp, this implies no topology gains for subclones longer than 40 kb. Since 10 kb is the maximum routine length for sequencing templates, this limit will not be reached in practice.

Evaluating the number of topologies is not a simple task. Nevertheless, it is undoubtedly feasible, given sufficient effort. Lee Newberg (1993 and 1996) has evaluated the number of possible topologies for traditional random subcloning maps.² By his counting, the number of topologies for $n=2$ is 2, for $n=3$ is 10, and for $n=4$ is 94. This number rises exponentially, with 1.3×10^{10} topologies at $n=10$ and 3.2×10^{27} topologies at $n=20$. The number of pairwise topologies will rise even faster. Furthermore, the smaller the ratio of f to clone length (up to the $1/n$ limit), the faster the rise.

² Newberg uses a slightly different definition of topology. He treats the clones as distinguishable. However, his results are easily adaptable.

The number of topologies affects primarily the probability of the clones forming a single scaffold. This will also tend to increase the length of the average scaffold. Another factor influencing the length of the average scaffold will be the clone length. With the number of topologies held constant (i.e., with $1/x-1 < f < 1/x$), the average scaffold length will rise proportionally with the clone length. This rise will continue even after the limit of $f < 1/n$ is reached.

Therefore there is always a monotonically increasing rise in average scaffold length as insert length is increased. I postulate that this curve will be convex, indicating that the most benefit from increasing insert length will occur with shorter inserts (with the exception of insert:fragment ratios below three). I have not proved this postulate.

Despite the increase in scaffold length with respect to insert length, this increase is not always useful. Consider the limiting case of a single clone ($n=1$). The single scaffold formed will grow linearly in length with respect to insert length. If the insert length equals the target length, this scaffold will be complete by definition. This increase in scaffold length is however of little use to the researcher: as the completeness (i.e., resolution) of the map decreases, the density of SMGs diminishes. Increases in scaffold length accompanied by increases in SMG density are what is needed at the lab bench. Such an increase cannot be gained by longer clones alone — the longer clones must be accompanied by an increase in available topologies. Thus, although there is a theoretical gain in scaffold length above a clone:read ratio of n , there is no further practical gain. Again, this last point can be considered trivial, as the clone:read ratio will never reach n in practice.³

Much work remains to be done on this difficult problem and it may be that approaches that remain hidden today will ultimately provide more insight for a mathematical model for pairwise end sequencing.

³ One might imagine that it could happen during a BAC end sequencing project, where the clone:read ratio is approximately $100000:400 = 250$. However, such sequences are poorer quality sequences that will not usually be used for constructing finished sequence. They would rarely be used as part of a project undertaken strictly as outlined in this dissertation. Additionally, although 250 clones might be used for analyzing a small target, 250 is significantly less than the actual n that would be used for a human-genome-sized project (see Venter et al., 1996; Siegel et al., in preparation).

ONECLONE



TWO CLONES

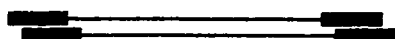


Figure A.1. Topologies for one- and two-clone double-barrel configurations. Single-scaffold topologies are highlighted in red. In some instances, sequence islands will span regions depicted here as SMGs.

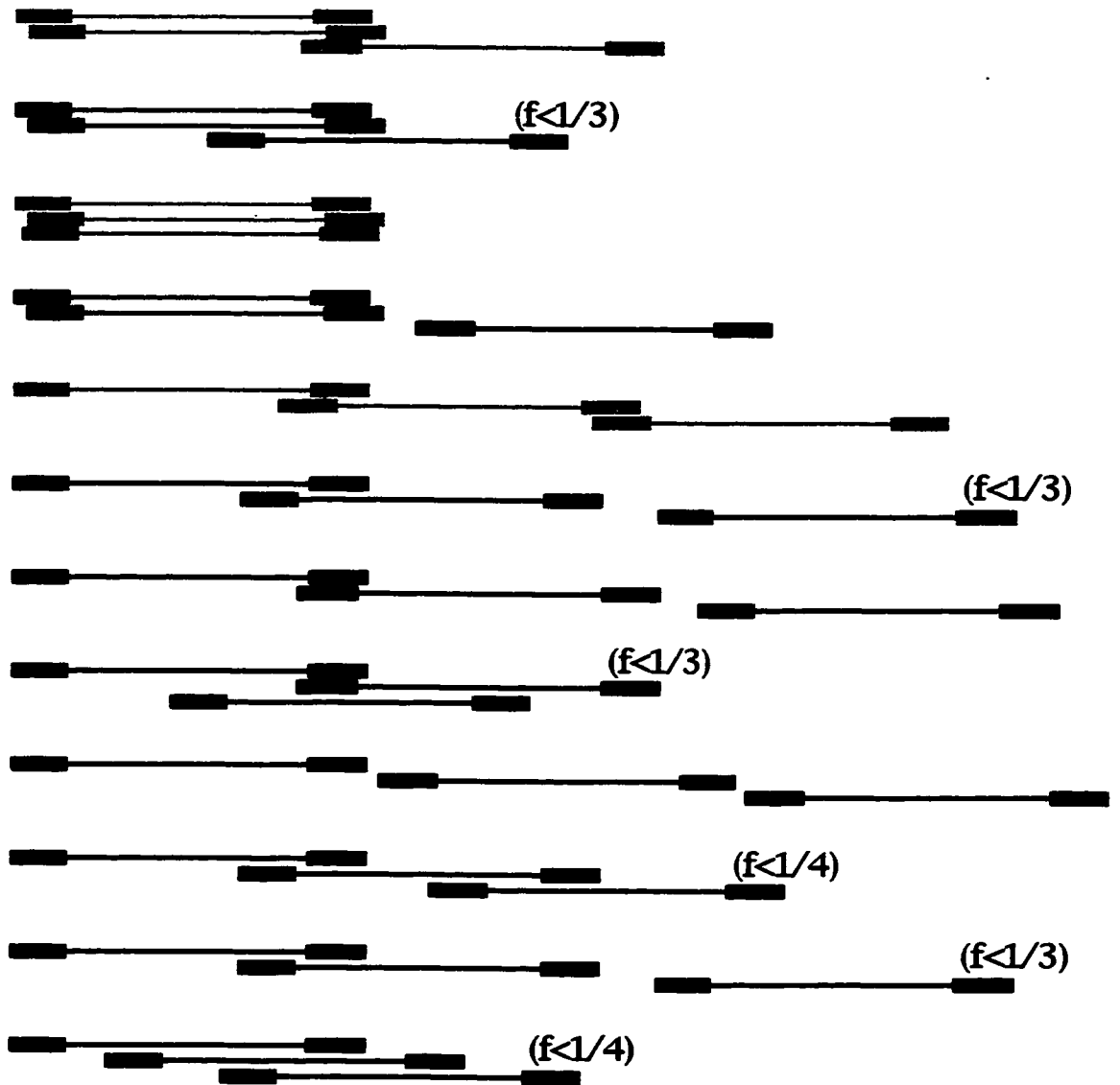


Figure A.2. Topologies for three-clone double-barrel configurations. The number of possible topologies rises rapidly with the number of clones. Single-scaffold topologies are highlighted in red. The first and third configurations shown here represent any topology in which their overlapping sequence ends form a sequence island, even if all three ends do not mutually intersect. In some instances, sequence islands will span regions depicted here as SMGs.

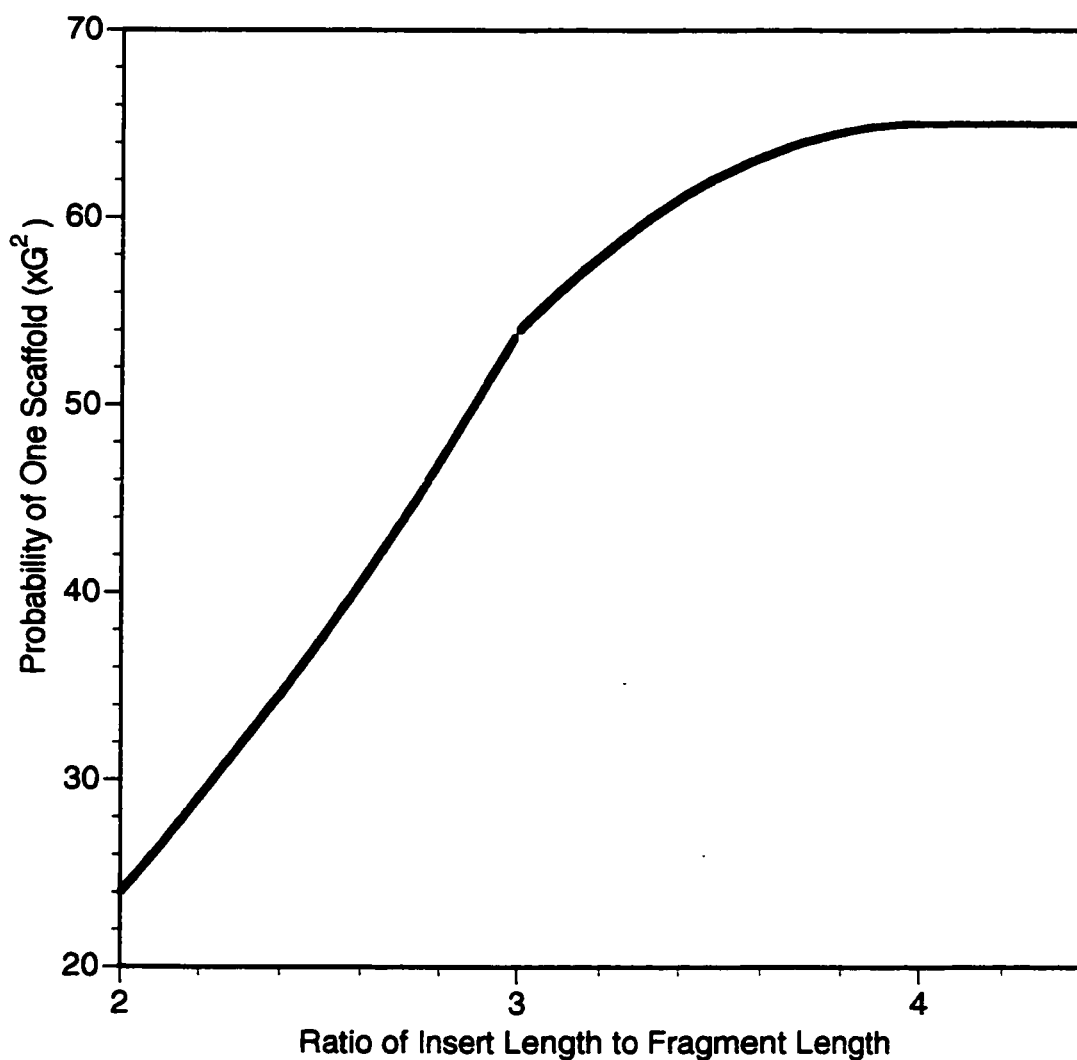


Figure A.3. For the case of three inserts, the probability of obtaining one scaffold versus the ratio of insert length to fragment length. The probability increases as the insert length is increased (and the fragment length remains constant). The character of the increase changes with each discrete increase in the ratio of insert length to read length, eventually reaching a plateau.

APPENDIX B. ALTERNATIVE SEQUENCE DISTANCE METRICS

There are limitations to all current distance metrics. The predominant distance metric used throughout Chapter 3 is based on the Dayhoff PAM matrix. This metric was used primarily for simplicity and speed of calculation, as it is an option of the *protdist* computer program, which is part of the *PHYLIP* package (Felsenstein, 1993). During the course of my efforts to grapple with multigene family protein evolution, I sought better methods to evaluate phylogenetic distances. Although it is not clear that I have succeeded, or even made a step in what might ultimately be the right direction, I have included this Appendix describing my preliminary efforts.

B.1 SEQUENCE DISTANCES

A simple way to compare two sequences is to count the number of residues they share in common. This determines their percent identity:

$$\text{Percent Identity} = \frac{\text{number of identical residues}}{\text{total residues}} \quad (\text{B.1})$$

Other methods compute values for similarity. In this case, a similarity value is assigned to each possible combination of two residues. Residues that are more likely to have replaced each other over the course of evolution have higher similarity values. For example, aspartate and glutamate, which are both acidic, are quite similar. There are many methods of computing sequence similarities and distances.

The distance between two sequences reflects the number of mutations that, on average, under conditions of natural selection, will convert one sequence into the other. If the mutation rate is constant and the conditions of natural selection are unchanging, then the number of mutations can be related to the amount of time separating the two sequences.

Mutations are typically assumed to operate as single base pair substitutions at the level of nucleotides. However, other types of mutations are possible. Some models allow for insertions and deletions of base pairs. A few models account for gene conversion events, but these suffer from lack of empirical data. Gene conversion is often considered to

be such a rare event that it can be ignored as a mutational mechanism, despite the large evolutionary impact a single gene conversion event can have. Multigene families are particularly susceptible to gene conversion, and so may provide a significant challenge to models that ignore gene conversion.

There are multiple requirements that must be met for sequence distances to be interpreted as a measure of the time separating two sequences. The hypothesis that sequence distance can be interpreted as time is called the *molecular clock hypothesis*. In some cases the molecular clock hypothesis holds, but in many cases it does not.¹ When a molecular clock can be used, molecular sequence data provides a powerful source of information on species phylogeny. It is important to distinguish between those cases in which the molecular clock is useful and those in which it is misleading. Making this distinction will be an important part of the analysis of the trypsinogens presented in this chapter.

All molecular sequence data is contemporary.² Therefore, if two sequences are related, it is not because one evolved from the other, but because they both arose from a common ancestor. If two contemporary sequences shared a last common ancestral sequence 600 million years ago, then evolutionary processes have been operating to diverge those sequences for a net 1.2 billion years of evolution. To rephrase, 1.2 billion years of divergent evolution separate the two sequences.

It is generally assumed that evolution is symmetric. This implies that the evolutionary distance separating a modern sequence from its ancestor would be calculated identically if the sequences were switched. This assumption allows one to unambiguously calculate the distances between two sequences without knowing the position of the last common ancestral sequence in the phylogeny. Seldom does this assumption hold

¹ Several papers in the first issue of the *Journal of Molecular Evolution* address this topic, including Dickerson (1971), Kimura and Ota (1971) and Ota and Kimura (1971). Dickerson provides a chart showing the rates of evolution for cytochrome c, hemoglobin, and the fibrinopeptides. A molecular clock hypothesis holds for each of these proteins, according to the original data in these publications.

² Barring discovery of intact genetic material, it is not possible to obtain sequence from extinct species. In those rare cases where such sequence has been obtained, it is often incomplete. Furthermore, such sequences are seldom more than a few thousand years old, which is an almost insignificant period compared to evolutionary time scales.

absolutely. However, it almost universally holds well enough to be a useful tool. Without this assumption, molecular distance calculations become wickedly complicated.

Strictly speaking, it is not necessary to calculate pairwise distances between sequences in order to construct a phylogeny. In fact, in an ideal situation, it is not even desirable to do so. Information is lost when a data set of multiply aligned sequences is reduced to a diagonal matrix of pairwise distances. Some of this lost information is valuable not only in constructing the topology of a phylogeny but also in determining the evolutionary distances between sequences. Methodologies that employ maximum likelihood are most capable of properly utilizing all multiple alignment information. Such methodologies are computationally intensive. Additionally, no maximum likelihood computer programs are available capable of employing algorithms optimized for trypsinogen phylogenies. Even if such a program were to be written, it is likely that, today, the computer resources required to do a full trypsinogen analysis would exceed those available to trypsinogen researchers. These circumstances are likely to change in the future, both as novel programs are written and as computing power grows. In the meantime, to produce a trypsinogen phylogeny, it is necessary to use pairwise distances.

In order to produce as accurate a phylogeny as possible, it is important to choose the best method of calculating sequence distances. It is interesting to ponder what is meant by "best" method. A method should accurately produce a value for distance that reflects the amount of evolution that has occurred between two sequences. The "amount of evolution" is usually understood to mean the number of fixed point mutations, although it might conceivably mean other things, such as gene conversion events. Furthermore, if a molecular clock hypothesis holds, a calculated distance value should be proportional to the time of divergent evolution separating two sequences.

There is no perfect way to evaluate whether or not one particular distance metric is better than another. One can compare a phylogeny produced from calculated distances with a known phylogeny. However, there are few phylogenies known with certainty. Those that are known are usually sparsely populated, and only the topologies are absolutely certain, not the distances. Furthermore, one cannot be certain that a phylogeny made from sequence distances should necessarily reproduce a phylogeny produced by some other means, such as from paleontological and morphological considerations, or from molecular data derived from another gene locus. The phylogeny of one gene will not necessarily reflect the

phylogeny of another gene. Gene phylogenies will not necessarily reproduce species phylogenies.

Therefore one must rely on imperfect methods to evaluate a distance metric. Three are mentioned here. The first approach is to compare the metric with known phylogenies with the assumption that these phylogenies are correct and reflect the phylogeny of the gene in question. Secondly, one can simulate a phylogeny using an assumed model for evolution. One then determines whether the metric in question can reproduce the phylogeny. The problem with this second approach is that never is the process of evolution absolutely known, so this approach relies on the correctness of the hypothetical model for the evolutionary process.

Both of the first two evaluations methods are empirical. While employing such evaluations, one is not concerned with the details of how the distances are calculated. If a distance metric reproduces known phylogenies, then regardless of the details of the calculations, it is useful.

The third approach to evaluate distance metrics is not empirical. In this approach, the details of the evolutionary process are assumed, and a metric is derived from this evolutionary model. Such metrics are “perfect,” in the sense that they precisely measure distances (subject to random error) produced by an evolutionary process identical to that specified in the model. The obvious problem with such metrics is that actual evolution may not obey the assumptions of the model.

I will now consider the details of the original model for distance metrics, due to Jukes and Cantor (1969).

B.2 THE JUKES-CANTOR MODEL

We assume that in a given unit of evolutionary time, a base has a probability μ of mutating. We assume that if a base mutates, then it has an equal probability of being replaced by any of the four bases, including itself.³ It follows that any base that has

³ A commonly encountered alternative definition of μ specifies that when a base mutates, it has an equal probability of mutating to any of the *other* three bases. These definitions

mutated one or more times has an equal probability of currently being any of the four bases. The probability of a base remaining unchanged after τ units of time is

$$P(\text{base never mutated}) = (1 - \mu)^\tau \quad (\text{B.2})$$

If one is uncomfortable with designating τ as a unit of time, then one may refer to it as a unit of *pseudotime*. The probability of a base mutating at least once, but currently being in its original state, is one quarter of the probability that the base has mutated at least once, or

$$P(\text{base returned to original state}) = \frac{1}{4} [1 - (1 - \mu)^\tau] \quad (\text{B.3})$$

Therefore the probability of a base being in its original state after τ units of time is

$$\begin{aligned} P(\text{base in original state}) &= (1 - \mu)^\tau + \frac{1}{4} [1 - (1 - \mu)^\tau] \\ &= \frac{1}{4} [3(1 - \mu)^\tau + 1] \end{aligned} \quad (\text{B.4})$$

If τ is allowed to vary continuously, then equation (B.4) becomes

$$P(\text{base in original state}) = \frac{1}{4} [3e^{-\mu\tau} + 1] \quad (\text{B.5})$$

Equations (B.4) and (B.5) are equivalent for any practical purpose, as the unit of time will always be chosen small relative to the minimum distance between two sequences. I tend to employ discrete forms as it is slightly easier to program these into computer algorithms. To obtain evolutionary time, one can convert equation (B.4) into the following form:

produce equivalent models, but the value of μ in the alternative definition is $\frac{3}{4}$ of the value of μ employed here.

$$\tau = \frac{\ln\left(\frac{4P-1}{3}\right)}{\ln(1-\mu)} \quad (\text{B.6})$$

Now, one need only compare two sequences and make the maximum likelihood estimate for P as the percent identity of the sequences to obtain an estimate for τ as the divergence time. This model is very nice because it is simple.

Unfortunately, it is impossible to use the Jukes-Cantor model with amino acid sequence data.

B.3 A PEPTIDE VIEW OF THE JUKES-CANTOR MODEL

One can hypothesize an evolutionary mechanism similar to the Jukes-Cantor model, but that works at the level of residues instead of nucleotides. The mathematics of such a model would be nearly identical to the Jukes-Cantor model, except that there would be twenty states available for mutation in place of four. One quickly derives the following equation:

$$\begin{aligned} P(\text{residue in original state}) &= (1-\mu)^\tau + \frac{1}{20} \left[1 - (1-\mu)^\tau \right] \\ &= \frac{1}{20} \left[19(1-\mu)^\tau + 1 \right] \end{aligned} \quad (\text{B.7})$$

There is a major objection to this model. Mutations occur to nucleotides and not to residues. It is already a big assumption that all possible nucleotide substitutions are equally likely. It is yet another assumption to assume that all residues are equally likely mutations from a given codon. In fact, it is only possible for a single nucleotide substitution to change a given codon into 9 of the 63 other codons. At most, a codon might be able to mutate so as to code for 9 other residues, but due to the degeneracy of the genetic code, the widest repertoire of available residue mutations is seven; the narrowest is four (the average is 5.8).

For the model in equation (B.7), the concept of a “mutation” changes from that of equation (B.4). For (B.7), anything that changes an amino acid residue counts as a mutation. This could be one or more point mutations, a intron/exon boundary slide, or a gene conversion event. For (B.4), only point mutations are considered.

Recall that evolution works by natural selection as well as mutation. To a first approximation: mutations alter nucleotides; selection operates on residues. Therefore the use of a nucleotide-based model tends to emphasize the role of mutation in evolution. The use of a residue-based model emphasizes selection. For a non-coding region unaffected by selection it would be appropriate to use a nucleotide model and foolhardy to use a residue model. For a region strongly affected by selection, or by mutational events operating above the scale of a single nucleotide substitution, then it may make more sense to use a residue model.

If selection plays a significant role in the evolution of a protein, then there may be a significant probability that several silent mutations at the nucleotide level occur in one unit of evolutionary time. The rate of residue change is predicted to be much smaller than the rate of nucleotide change. Allowing for several silent mutations to occur between state changes at the site of a particular residue, then contemplation of the genetic code will indicate that widest number of available residue mutations is twelve and the narrowest five (with an average of 7.4). Non-silent mutations at a site are less likely to be fixed. Since selection operates on the site, it is likely that a non-silent mutation will decrease fitness and be selected against. Thus under conditions of selection, a residue model becomes more reasonable.⁴

I will primarily be employing a residue model to analyze the trypsinogen sequences. There are three reasons for this. First, my data is amino acid sequences, at least partially. Second, trypsin is a highly conserved enzyme with a critical digestive function. There are strong selective constraints governing the evolution of trypsin. Third, the trypsinogen genes are members of a multicopy gene family. They often reside in repeated elements in chromosomes. Within a genome, there is a large reservoir of trypsinogen gene material for gene conversion, unequal crossing over, and other methods of genetic exchange.

Gene conversion events can convert lengths of DNA ranging from a single nucleotide to several thousand (Li, 1997). A codon may undergo a change at all three of its

⁴ Aaron Halpern, currently at the University of New Mexico, has done some work building more complex residue models that incorporate aspects of the genetic code (personal communication). It may be that enough data will eventually be available to conduct empirical evaluations of different distance metrics. Currently, one can only note that for vertebrate trypsinogen data, residue-based and nucleotide-based metrics give roughly similar distances.

positions as a result of a gene conversion event. Furthermore, this change is less likely to be selected against, as the donor DNA may be a functional allele. Thus, it may be improper to consider nucleotide-based models for change when gene conversion is a prominent force for random variation. Mutation by gene conversion, effectively at the level of codons and residues, may have been common and dominant during the course of trypsinogen evolution.

The preceding arguments suggest that a residue model may be more appropriate for trypsinogen evolution than a nucleotide model. However, the initial model described by equation (B.7) may be too simplistic. There are several possible modifications to this basic model. I will describe two simple extensions.

A first possible extension to the basic residue model assumes that strong selection at a site permits only a few residues to be accepted as substitutions at that site. A change to another residue, other than these few, will result in a low fitness and effectively a zero probability that the mutation will be fixed. This, in essence, “slows down” mutation at that site. The modification to equation (B.7) is as follows:

$$P(\text{residue in original state}) = \frac{1}{N} \left[(N-1) \left(1 - \frac{N\mu}{20} \right)^r + 1 \right] \quad (\text{B.8})$$

Here N is the number of allowed residues at the site. Equation (B.7) can be obtained from equation (B.8) by setting $N=20$.

A second possible extension to the basic model is that there is no selection, but the number of possible mutations at any particular site is limited. This might occur if mutations occurred solely as a result of gene conversion from a limited number of alternative alleles. The modification to equation (B.7) would be as follows:

$$P(\text{residue in original state}) = \frac{1}{N} \left[(N-1)(1-\mu)^r + 1 \right] \quad (\text{B.9})$$

It is hard to make a case for practical use of the model implemented in equation (B.9), partly because one of the motivations for using a residue model is that natural selection plays a significant role in evolution, as the available pool of residues is likely to

change over time. Additionally, it seems unlikely that gene conversion from a limited pool of residues would be a dominant mode of evolution. Equation (B.9) is presented here mainly for reference. Therefore equation (B.8) will form one basis for my evaluation of the trypsinogens. Note that equation (B.9) can be transformed into equation (B.8) merely by employing a smaller probability of mutation per unit time ($\frac{N}{20}\mu$ in place of μ). This reflects the idea that selection “slows down the evolutionary rate.”

For the present paper, I arbitrarily set the rate μ to 0.01. Were a molecular clock hypothesis to hold, μ could be set empirically. Small alterations in μ affect the scale of a derived phylogenetic tree, but not its topology or proportions. To demonstrate this last point for the Jukes-Cantor model, I make the observation that $(1 - \mu)^\tau \approx 1 - \tau\mu$. Note that $\mu \ll 1$. Therefore from equation (B.4) we have

$$\tau \approx \frac{4(1 - (\text{percentage of bases in original state}))}{3\mu} \quad (\text{B.10})$$

Evolutionary time τ is inversely proportional to the mutation rate μ . This should not be surprising. If the mutation rate doubles, it should take half as long for a sequence to acquire the same number of changes. Because of back mutations and saturating effects, this proportionality is approximate.

Not all sites of a sequence will necessarily have the same number of permitted residues. Therefore one cannot use an equation similar to equation (B.6) or (B.10) when evaluating the most likely τ for equation (B.8). Rather one must determine the τ that results in the maximum likelihood of the observed identities and differences between two sequences. This can easily be done by computer, although multiple iterations may consume considerable processor time.

All of the models presented here assume that each site of a sequence evolves independently. This assumption clearly does not hold in reality, as gene conversion events may convert many adjacent residues simultaneously. However, any attempt to account for covariant effects is extremely unwieldy. Additionally, our knowledge of selective pressures is too limited to provide a model more accurate than what can be obtained with an assumption of independence.

In order to implement equation (B.8) in practice, one needs to determine the number of possible states possible at each site. For this, I assume that I have a large enough collection of sequences to have observed all possible permitted residues at each site. This assumption is most valid for sites with few observed residues. These sites are the most significant contributors to distance calculations, so this assumption is reasonable. A major incentive to use this model is to account for a slower rate of change at sites that, due to functional constraints, have few permitted residues. Note that if a site has only one observed state, I will assume that site to be invariant. Such sites contribute no information to evolutionary distance calculations and must be ignored.⁵ A plot of the variability of each trypsin site is shown in Figure B.1.

A bizarre but fascinating combination of Sesame Street and information theory has produced a method of graphing data from multiple sequence alignments. Such graphs are known as sequence logos (Schneider and Stephens, 1990). A sequence logo for pretrypsinogen is shown in Figure B.2. Because my aligned pretrypsinogens do not represent a uniform sampling of the vertebrate phylogeny, the relative residue frequency depicted by character height in the sequence logo may not be strongly correlated with biological significance. Nevertheless, the sequence logo strikingly depicts highly conserved regions. The information in the sequence logo complements the information in Figure B.1.

Some authors have attempted to incorporate site-to-site variability into phylogeny calculations (e.g., Yang, 1993, 1994, and 1995; Felsenstein and Churchill, 1996; Thorne et al., 1996). Jones et al. (1994) present a mutation data matrix for transmembrane proteins, which is a step in the right direction. However, this matrix cannot be used for trypsin, which lacks transmembrane motifs.

Hidden Markov models (HMM) are currently in vogue as a method for incorporating site-to-site variability into models. It is not at all clear that a HMM algorithm would be appropriate for trypsin data. Firstly, the “hidden” aspect forces the researcher to discard what is known about conserved residues and selective constraints. Secondly, effective use of an HMM requires that the persistence length of site conservation be at least moderately greater than a single residue. This constraint is violated repeatedly by trypsin,

⁵ Technically, using the model of equation (B.8), they do not have to be ignored, as the probability of the base resting in its original state is 1, so contributes neither negatively or positively to a maximum likelihood estimate for τ .

as a glance at the numerous narrow spikes and valleys of Figure B.1 shows.⁶ This also limits the effectiveness of Γ distribution models. Although the use of HMM algorithms may be a step in the direction of incorporating site specific information into protein phylogenies, it is my feeling that they currently offer no improvement on traditional methodologies.

Ideally, I would like to construct a separate distance matrix for each trypsin residue site. Most current distance metrics treat all sites equivalently (e.g., Jones et al., 1992; Taylor and Jones, 1993). Reliable construction of site specific distance matrices for trypsinogen (and most other proteins) would require more sequences and a better understanding of evolution than is currently available.

A preliminary, although simplified, approach to construction of a metric would be to tabulate all possible residues at a particular site that maintain function at the molecular level and fitness at the organismal level. This would require a statistically significant sampling of sequences spanning all clades of the phylogeny in question. Since known trypsinogen sequences from some vertebrate classes are either very sparse or completely missing, reliable estimates of the number of permitted residues at a given trypsinogen site cannot be made. In years to come, reliable estimates will be possible. At such a time, extensions to this approach for site specific distance estimation might include accommodations for the underlying mechanisms of mutation or for covariation of sites. Additionally, computational advances may permit maximum likelihood approaches to be combined with such models for evolution. For the present, I suspect that the simple “limited state” method of equation (B.8) provides a reasonable approximation to what might be obtained with more data.

A phylogeny analagous to that of Figure 3.14 is shown in Figure B.3. This phylogeny uses the distances calculated according to the methodology described above in place of the *protdist* distances used for the same purpose in Chapter 3. The topology of the phylogeny in Figure B.3 is essentially the same as the phylogeny in Figure 3.14. The distances of the phylogeny in Figure B.3 produce a phylogeny that has slightly more resemblance to a “star phylogeny” than does Figure 3.14. This results from smaller relative distances between pairs of sequences calculated with the methodology of this appendix.

⁶ Jin and Nei (1990) provide some further discussion on site specific rates of change.

An attempt to use multidimensional scaling on these distances produces plots with unacceptably high stress (approximately 0.35; data not shown). This is due to the moderately “spherical” character of this data in 31 dimensions, making it difficult to compact the data without skewing it. Despite this, as Figure 3.16 shows, these data still significantly support a hypothesis of coincidental evolution. If the true divergence time distances were indeed spherical, that would imply multiple early divisions followed by coincidental evolution that has created the appearance of a single early division with corresponding clustering of sequence distances. This possibility is discussed briefly in Section 3.16. The accumulation of more trypsinogen sequences and further refinement of sequence distance metrics will ultimately differentiate between the various hypotheses for the timing of the division(s) of the trypsinogen multigene family.

In order to develop data suitable for inclusion in site-specific matrices, a great deal of knowledge must be known about the selective pressures on the site in question. It is conceivable that at some point in the future that such knowledge may be gained from intense biochemical and genetic study of the gene product in question. At present, it is beginning to be possible to assign sites based on homology and known or predicted secondary and tertiary structure. Such predictive efforts work better on some proteins than others. There are currently several models for sequence based protein structure prediction. The simplest models employ only three structures: helix, sheet, and loop, so are unlikely to provide much extra power to phylogenetic analyses.

There are a few models with increased sophistication. One example is the model of I-sites by Bystroff and Baker (1997). I-site theory was originally and primarily designed to identify distant relationships between proteins based on similarities between protein folding initiation sites. However, the theory can be adapted to provide data for site-specific distance metrics. In short, if a site from a protein of interest can be assigned with confidence to a specific I-site position, then that position can be assumed to evolve according to the constraints on that I-site. Such constraints can be estimated from the entire body of I-site data, rather than limiting oneself to the possibly-skimpy data available from the set of proteins in question. This approach can form the basis for a more complex model. For example, there may be additional constraints on a residue beyond those imposed by its inclusion in an I-site.

I-site motifs can be assigned to 77 of 238 analyzed trypsin sites. Unfortunately, all but eight of these are assigned with a confidence statistic 0.80 or less, and all but 23 have a confidence statistic less than 0.50.⁷ Furthermore, 47 of the positions assigned to I-sites were either absolutely conserved or showed only two different residues in all vertebrate sequences. This high conservation is not unexpected, as the original intention of I-site prediction is to predict structures important for initiation of protein folding, which should also be conserved.⁸ However, since at these 47 conserved sites the trypsins show an even more restricted range of variation than is seen in I-site consensus sequences, it suggests that even stronger selective pressure operates on these positions in trypsin than merely enough to maintain an I-site consensus. Therefore it would be inappropriate to use evolutionary change matrices based on I-sites at these positions. These considerations lead me to conclude that an attempt to incorporate current data on protein motif consensuses would not have noticeably improved trypsin distance calculations. It may be that this will change as more knowledge accumulates on the relationships between protein structure and sequence. Also, I-site data may prove more useful in analyzing proteins other than trypsin, which might have fewer absolute constraints such as those dictated by the maintenance of a proteolytic active site.

⁷ A confidence statistic of 0.50 to 0.80 is judged to be "OK," while above 0.80 is "good." The confidence statistic is described by Bystroff and Baker (1997).

⁸ There are a total of 113 positions in the vertebrate trypsinogen alignment which permit either one or two observed residues. Of these, 35 are absolutely conserved. Approximately two thirds of these highly conserved positions are not assignable to a known protein structure motif (I-site predictions correlate well with other structure prediction algorithms). Many of these positions are involved in catalysis, substrate specificity, or cystine bridge formation.

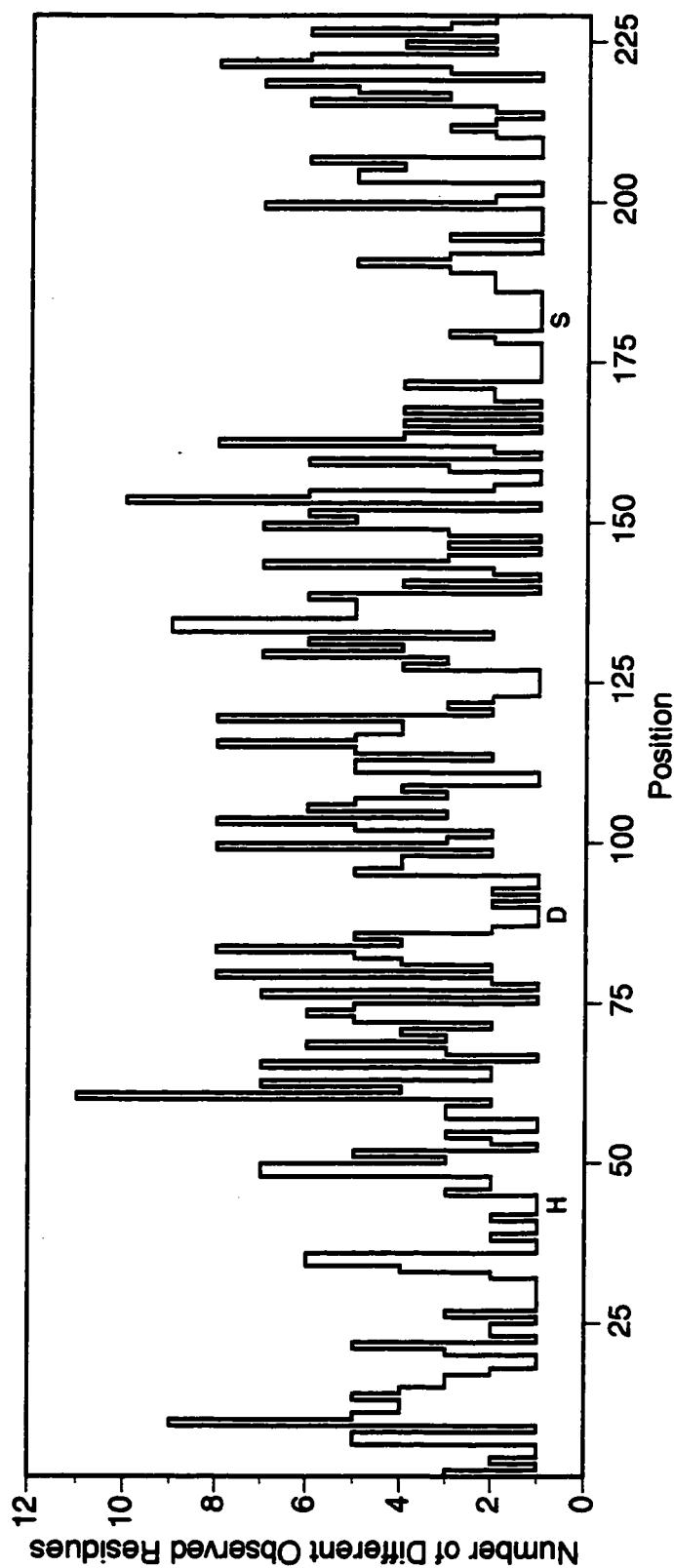


Figure B.1. Trypsin site variability. The letters H, D, and S identify the catalytic triad residues. This graph does not address the frequency of a particular residue at a site, as each different residue is counted once, whether it occurs in one sequence or many. The last site of the activation peptide is the first site shown in this graph and is one of three possible residues: lysine, arginine, or histidine. The last site in this graph represents either a serine or a stop codon. There are 228 sites graphed.

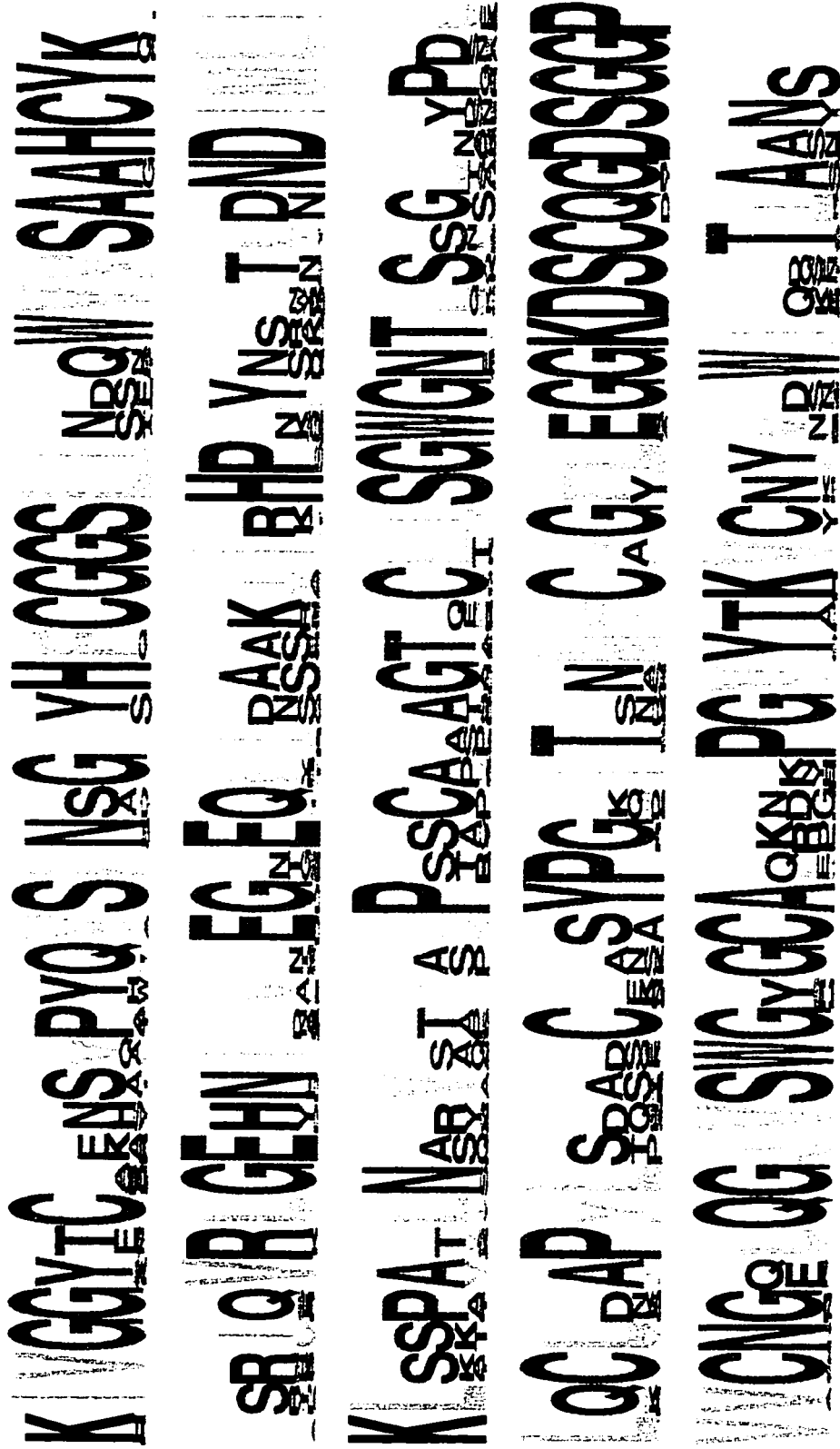


Figure B.2. Pretrypsinogen sequence logo. The total height of the characters at a given site indicates the information content of that site, while the height of each character reflects the relative frequency of the represented residue at that site. The color of a residue reflects its chemical structure. This sequence logo was generated by the engine at www.bio.cam.ac.uk/seqlogo/logo.cgi. Consult Schneider and Stephens (1990) for details of sequence logos.

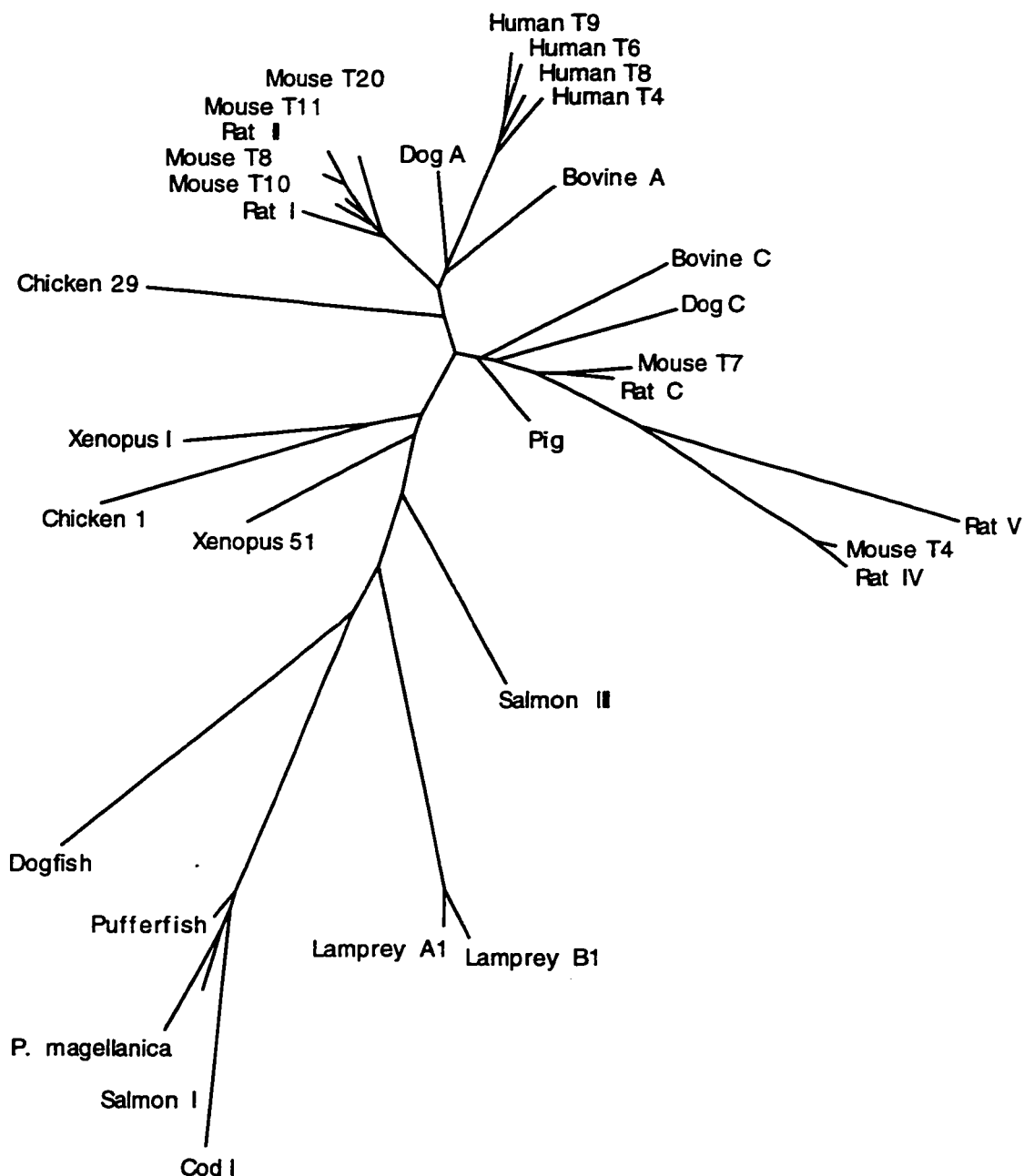


Figure B.3. A Fitch-Margoliash phylogeny of thirty-two vertebrate trypsinogens. Distances, calculated according to the methodology described in Appendix B, were fed to the program *fitch*, with global rearrangements and 40 random “jumbles” (Felsenstein, 1993).

APPENDIX C: PCR EVALUATION OF TRYPSINOGEN EXPRESSION

The polymerase chain reaction is a highly sensitive method for detecting the presence of a specific nucleic acid sequence. It can be used on RNA derived from a tissue to ascertain the presence of a sequence of interest, such as a trypsinogen mRNA.

PCR suffers from several drawbacks. It cannot accurately assay the quantity of a particular sequence. Its high sensitivity makes it vulnerable to false positives from contaminated samples. During the amplification process, highly identical sequences can recombine, producing artifacts. The error rate of PCR can be high, making it difficult to accurately assign the sequence of an amplified product to a particular allele. Therefore, PCR results must be interpreted with caution.

For this study, I have examined several tissues from the human and mouse to determine which trypsinogens were present.

C.1. METHODS AND RESULTS

Tissues were obtained as described in Sections C.2 and C.3. Total RNA was isolated with the guanidine thiocyanate protocol (Promega). Whole frozen tissues were ground under liquid nitrogen before RNA isolation. Frozen ground powder was added to denaturation solution and homogenized with a dounce. For peripheral blood cells, isolated cells were added directly to denaturation solution.

Purified total RNA samples were employed as templates for RT-PCR. First strand reverse synthesis was accomplished with a poly-T primer. The PCR primers used in this study are tabulated in Table 3.3. Nested PCR was required to obtain products from all tissues studied, except the pancreas. Initial amplification was performed with primers TRYA and TRYD, and subsequent amplification with either TRYB and TRYC or TRYF and TRYR. Each of these primers has a BamHI 5' leader. None of the predicted trypsinogen PCR products has an internal BamHI site. Twenty to thirty cycles were employed for each round of amplification, annealing at 55° for 30 seconds and extending at 72° for 50 to 80 seconds. PCR products were isolated, cleaved with BamHI, and cloned into the BamHI site of the m13mp9 vector. Individual clones were isolated and sequenced.

Human tissues were obtained from four sources. Pancreas and liver samples were obtained from a 45-year-old male who died of arrhythmia 18 hours before the tissue was quick frozen in an ethanol and dry-ice slurry. Fetal liver, spleen, and thymus samples were obtained from an 84-day embryo. Tissue was immediately quick frozen. Peripheral blood mononuclear cells were isolated from a 24-year-old volunteer and purified on a Ficoll gradient. In addition to these samples, pancreas-, thymus-, liver-, and spleen-specific cDNA samples were obtained commercially (Clontech). The results are summarized in Table C.1.

Mouse thymus, liver, and spleen tissues were obtained from a freshly killed Balb/C mouse. Tissues were immediately quick frozen. The results are summarized in Table C.2.

C.2. DISCUSSION

All human trypsinogens with hypothetically-functional genomic sequences are expressed, as seen by PCR. This observation is consistent with the presence of all of them in the EST database (Table 3.7). The pancreas shows the broadest range of expression of the various isozymes. Human trypsinogen T9 is found by PCR in all tissues examined. However, human T9 also appeared inconsistently in negative control PCR reactions, indicating the possibility of contamination accounting for this observation. Human T6 was only observed in the thymus by PCR, but has been found in pancreatic cDNA libraries (Table 3.7). This raises the possibility that human T6 is somewhat thymus-specific. These data suggest that there is differential expression of trypsinogen isozymes, although little can be said with respect to relative abundance in various tissues.

An inconsistent competitive advantage of human T6 during PCR could also account for this observation. Alternatively, the tissue sample employed for PCR may have come from an individual homozygous for a deletion of human T6.

Assuming that the PCR data from the mouse spleen represents pancreatic contamination, the mouse pancreas also demonstrates the broadest range of trypsinogen isozyme expression (Table C.2). Not all hypothetically-functional genomic sequences from the mouse are observed by PCR – only mouse T8, T9, T10, and T11. ESTs for mouse T7, T8, and T9 are present in the EST database (Table 3.8). Additionally, the sequence of the cloned mRNA for mouse T20 is present in Genbank. There is currently no concrete

evidence for the expression of the hypothetically-functional sequences mouse T4, T5, T12, T15, and T16. It may be that they are

expressed at very low levels, or not at all. Alternatively, any of the possible systematic errors of PCR or EST sequencing, discussed above, may account for the failure to detect their mRNAs.

ESTs for both the human and mouse “trypsinogen” T1 are present in the database, as discussed in Section 3.21.

Table C.1. Number of cDNAs sequenced by PCR from each of several human tissues.

	T8	T4	T9	T6	Total
Pancreas	6	8	5	Ø	19
Cadaver Pancreas	Ø	15	2	Ø	17
Thymus	Ø	4	11	17	32
Fetal Thymus	1	10	8	Ø	19
Liver	Ø	Ø	11	Ø	11
Cadaver Liver	Ø	Ø	32	Ø	32
Fetal Liver	Ø	1	17	Ø	18
Fetal Spleen	Ø	Ø	2	Ø	2
Spleen	Ø	Ø	4	Ø	4
PBMC	Ø	2	3	Ø	5

Table C.2. Number of cDNAs sequenced by PCR from each of several mouse tissues.

	T8	T9	T10	T11	Total
Thymus	10	11	Ø	Ø	21
Liver	14	13	Ø	Ø	27
Spleen	4	5	8	3	20

RESEARCH

Random Subcloning

Jared C. Roach

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195

Random subcloning strategies are commonly employed for analyzing pieces of DNA that are too large for direct analysis. Such strategies are applicable to gene finding, physical mapping, and DNA sequencing. Random subcloning refers to the generation of many small, directly analyzable fragments of DNA that represent random fragments of a larger whole, such as a genome. Following analysis of these fragments, a map or sequence of the original target may be reconstructed. Mathematical modeling is useful in planning such strategies and in providing a reference for their evaluation, both during execution and following completion. The statistical theory necessary for constructing these models has been developed independently over the last century. This paper brings this theory together into a statistical model for random subcloning strategies. This mathematical model retains its utility even at high subclone redundancies, which are necessary for project completion. The discussion here centers on shotgun sequencing, a random subcloning strategy envisioned as the method of choice for sequencing the human genome.

Random subcloning is a common tool for large-scale physical mapping and DNA sequencing. Large DNA targets are intractable to direct analysis and must be broken down into smaller fragments before techniques such as restriction mapping or sequencing can be employed. A map or sequence of the original target can be deduced following analysis of the derived fragments, termed subclones. Ideally, a direct strategy is pursued by analyzing a minimum number of fragments such that a shortest tiling path is followed. This requires prior knowledge of the relation of each fragment to the original target. However, such information is not necessarily or easily available. Thus, many strategies resort to picking and analyzing fragments at random.

In random subcloning strategies, fragments are generated from a vast number of identical target sequences, so the resulting library from which they are selected for further analysis is redundant. Therefore, individual fragments may overlap in the sense that they mutually possess some bit of target sequence. The presence of such overlaps allows retrospective determination of which fragments represent adjacent target sequences. When enough overlapping fragments have been analyzed, the original sequence or map may be deduced (Fig. 1).

From a theoretical standpoint, random mapping and sequencing strategies can be treated identically, albeit on a different scale. Most phys-

ical mapping projects employ targets on the megabase scale or larger. Such targets are randomly fragmented into yeast artificial chromosome (YAC), bacterial artificial chromosome (BAC), cosmid, or phage subclones ranging in size from tens of kilobases to several megabases. Analysis techniques include restriction mapping, sequence-tagged site (STS) content mapping, in situ hybridization, and many others. Sequencing projects employ both smaller targets and smaller subclones. In particular, the targets of sequencing projects are often the subclones of mapping projects. Fragments of these subclones (i.e., subclones of subclones) are small enough to be employed as sequencing templates for automated DNA sequencing machines. The effective fragment size for sequencing projects is the sequence read length, which is the amount of sequence that can be read from one fragment by a sequencing machine. This length currently ranges from several hundred to 1000 bases.

The appeal of random subcloning strategies lies in the absence of need for prior information about particular subclones. This allows projects to be undertaken with a great deal of automation and with a decreased need for highly trained human intervention. The drawback of such "shotgun" strategies is their dependence on overdetermination of information, with a need to generate several times as much raw data as an ideal directed strategy would. Accordingly, actual strategies may be a mix of both random and directed approaches, beginning with random and pro-

E-MAIL: roach@u.washington.edu; FAX (206)685-7301.

RANDOM SUBCLONING

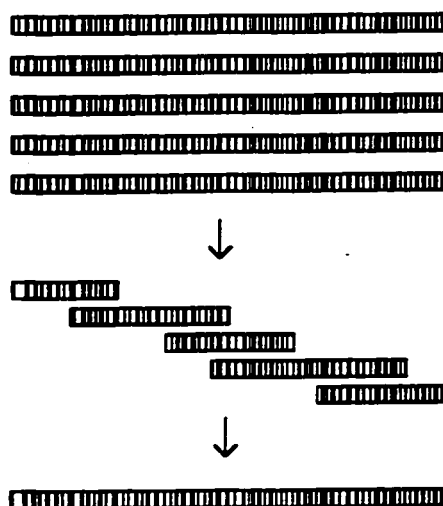


Figure 1 An example of a shotgun strategy, with a bar code length employed as an analogy to a DNA sequence. A large number of identical but unknown target sequences are randomly fragmented. These fragments are analyzed and aligned based on unique overlapping sequences. When enough fragments have been analyzed, the original target sequence may be deduced. The number of fragments is typically much larger than depicted here.

gressing to directed when the cost of choosing and sequencing directed subclones is judged to be less than the cost of continued shotgunning. Such decisions are predicated on the ability to determine such costs. Experience, simulations, and analytical models are the tools for this analysis.

A useful analytic model of the random shotgun strategy has been developed previously and is in common use (Lander and Waterman 1988). This model is accurate at low subclone redundancies. However, most sequencing projects and many mapping projects employ higher redundancies, which are necessary to approach complete coverage and closure of targets such as cosmids and BACs (for sequencing) or the human genome (for mapping), where closure is defined as the absence of gaps in the knowledge of the target sequence.

The present analysis initially specifies the distribution of the lengths of the spacings between fragment start sites. With this distribution, the number of gaps can be determined simply by

counting the number of spacings greater than the length of a fragment. All orderings of these spacings among spacings not forming gaps are equally likely, and because spacing lengths are independent of their ordering, this permits the determination of the distribution of the number of clones in an island as well as island length. Extensions permit additional probabilities of interest to be calculated.

Formulation

A linear discrete target of length G , such as a genome, is assumed. For a given project, n fragments of constant length L are generated from the target and analyzed in a manner in which overlaps between fragments are detectable. All fragments are generated from distinct identical copies of G . No fragments may start within $L-1$ bases of the last, right-most base of G , as such fragments would not be contained entirely within G . Thus, the effective length G_e available for fragment start sites is $G - L + 1$. The starting, or left-most, base pair of each fragment is designated S_k , such that S_1 is the start site of the left-most fragment, with $S_k \in [1, G_e]$. S_n begins the right-most, or last fragment of G . The start site may be either the 5' or 3' base pair of the Crick strand of the fragment it begins, depending on fragment orientation relative to the target. In this model the S_k are an ordered sample of n independently, identically, and uniformly distributed observations on the interval $(0, G_e)$. The formulation is drawn schematically in Figure 2. Let $D_k = S_{k+1} - S_k$ represent the distance between start sites, for $k = 1, 2, \dots, n-1$. D_0 represents the length of the uncovered target region before the first fragment, and equals S_1 . D_n is the distance $G_e - S_n$. An assumption is made that an overlap of length of at least T is necessary and sufficient to detect adjacency of two fragments. Redundancy, R , is defined as nL/G . For notational ease, the effective fractional coverage f_e of the target provided by one fragment is defined as $(L - T)/G_e$, and the effective redundancy R_e is defined as nf_e . The notation and symbols used here are summarized in Table 1.

By genomic conventions, an island is a maximal set of fragments, each of which is connected to all other island members by one or more paths of fragments overlapping by T . A contig is an island consisting of at least two fragments (Sta-

ROACH

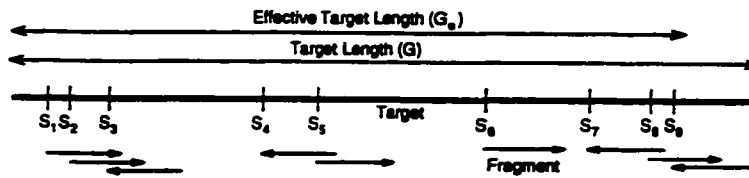


Figure 2 Schema of notation.

spacing lengths is ignored. However, because n is universally large for genome projects, this correlation approaches zero. With these considerations, the density function for the lengths of spacings between start sites is

$$f_{D_k}(x) = n(1 - \frac{x}{G_e})^{n-1} \quad (1)$$

$$\text{and } E(D_k) = \frac{G_e}{n+1}.$$

den 1980). In general, a target region not covered in any fragment is a gap. Adjacent islands are thus separated by gaps. The length of a gap is $D_k - L$, for $D_k > L - T$. A negative gap length indicates that an overlap is present, but not detected, so this is still considered a gap.

A simple geometric observation underlies many of the equations presented here: The domain space of the spacings D_k is the surface of the simplex $D_0 + D_1 + D_2 + \dots + D_n = G_e$, and their joint probability density is constant. This observation permits many probabilities of interest to be calculated by geometric considerations (Levy 1939). Here, G_e and D_k will usually be treated as continuous rather than discrete. This approximation is quite minor, given the scope of a genome, or even a cosmid, compared with the unit of divisibility, a base pair. Thus, the normalized length of a single spacing is described by a beta distribution, which is defined on the interval $[0,1]$ and is related here to the effective target length by the factor G_e . By employing the beta distribution in this case, the correlation between

Distribution of the Number of Gaps and Islands in a Project

A gap will occur following a fragment starting at S_k if and only if $D_k > L - T$. A spacing less than this will result in a detectable overlap between the two fragments; a greater spacing will result in no overlap, or an undetectable one. Thus, the probability of a gap following a given fragment is equivalent to the probability that $D_k > L - T$ (note the change of variable $y = x/G_e$):

$$P_{gap} = \int_{L-T}^{G_e} f_{D_k}(x) dx = \int_{L-T}^{G_e} n \left(1 - \frac{x}{G_e}\right)^{n-1} dx \quad (2)$$

$$= n \int_{f_G}^1 (1-y)^{n-1} dy = (1 - f_G)^n$$

Table 1. Notation and Symbols

G , target length
$G_e = G - L + 1$, effective target length
L , fragment length
n , number of fragments in a project
T , bases necessary to determine overlap
$R = \frac{nL}{G}$, redundancy
$R_e = \frac{n(L-T)}{G_e}$, effective redundancy
$f_G = \frac{L-T}{G_e}$, effective fractional target coverage per fragment
S_k , starting base of the k th leftmost fragment
$D_k = S_{k+1} - S_k$, spacing between adjacent fragment start sites
z_m , number of fragments contained within the m th leftmost island
l_m , length of the m th leftmost island

RANDOM SUBCLONING

Employing again the assumption that the lengths of the spacings between each S_k are independent, the distribution for the total number of gaps in a project is binomial. Again, this assumption is reasonable when there are a large number of spacings, the usual case for genome projects. In a given project there are $n - 1$ opportunities for a gap to occur, one between each pair of adjacent fragment start sites. Thus,

$$P(N_{\text{gaps}} = x) = \binom{n-1}{x} p_{\text{gap}}^x (1 - p_{\text{gap}})^{n-1-x} \quad (3)$$

$$= \binom{n-1}{x} (1 - f_G)^{nx} [1 - (1 - f_G)^n]^{n-1-x}$$

One immediately has the probability of project closure as

$$P(N_{\text{gaps}} = 0) = (1 - p_{\text{gap}})^{n-1} = [1 - (1 - f_G)^n]^{n-1} \quad (4)$$

Also, the expected number of gaps in a project is

$$E(N_{\text{gaps}}) = (n-1)(1 - f_G)^n \quad (5)$$

As noted above, the distribution for the number of gaps can be made exact by summing appropriate areas of $n + 1$ dimensional simplex. This somewhat awkward but nevertheless elegant distribution is provided by Stevens (1939) and in slightly different form by Flatto and Konheim (1962). Stevens' distribution is approximated by equation 3. Siegel (1979) provides an alternate derivation of equation 4 in slightly different form.

Not counting the two ends of the target as gaps, the number of islands will be one greater than the number of gaps, as each gap separates two adjacent islands. So by definition,

$$N_{\text{islands}} = N_{\text{gaps}} + 1, \text{ and} \quad (6)$$

$$E(N_{\text{islands}}) = 1 + (n-1)(1 - f_G)^n$$

Distribution of the Number of Clones in an Island

The total number of fragments in a project is n , and the total number of islands is N_{islands} , so the expected number of fragments z_m in an arbitrary island is clearly $E(z_m | N_{\text{islands}}) = n/N_{\text{islands}}$. To obtain the probability distribution of z_m one may divide the spacings $\{D_k | k = 1, 2, \dots, n-1\}$ into two subsets: those spacings $D_k > L - T$ that are

long and contain a gap, and those $D_k \leq L - T$ that are short and do not. The number of spacings in the first subset is N_{gaps} . By elimination, the number in the second subset is $n - 1 - N_{\text{gaps}}$. The number of fragments in an island is equal to one plus the number of short spacings between its two bounding long spacings. The last island might end with a short spacing, which would reduce its number of fragments by one. This minor effect is ignored here, but it may be accounted for, if desired, at the cost of a little algebra. Now all orderings of long and short spacings are equally likely, as the D_k are exchangeable. The probability distribution for z_m can be analyzed combinatorically (see approaches to similar problems by Whitworth 1897a; also see Batcille 1935), but to maintain simplicity a continuous approximation analogous to that employed previously to model spacing length may be employed (equation 1). In short, this determines the probability that no long spacings occur before the z_m th spacing. If there are enough short spacings, z_m may be treated as a continuous variable. This approximation of continuity is valid when n is large, the usual case for genome projects. Employing a uniform distribution of gaps over the continuous interval $[0, n - 1 - N_{\text{gaps}}]$ and scaling by a factor of $n - 1 - N_{\text{gaps}}$, a beta distribution is obtained as before. The conditional probability density for z_m is therefore

$$f_z[x | N_{\text{gaps}}] = N_{\text{gaps}} \left(1 - \frac{x-1}{n-1-N_{\text{gaps}}} \right)^{N_{\text{gaps}}-1} \quad (7)$$

Note that the number of short spacings in an island is beta distributed. An island will contain one additional spacing (and a fragment to go with it) because of its terminating long spacing. So with z_m the number of fragments in an island, $z_m - 1$ is the number of short spacings in that island. The expected number of clones in an arbitrary island (conditioned on the number of gaps) is, as expected,

$$E(z_m | N_{\text{gaps}}) = \frac{n-1-N_{\text{gaps}}}{N_{\text{gaps}}+1} + 1 = \frac{n}{N_{\text{islands}}} \quad (8)$$

The fraction of singleton islands expected in a project can be obtained by integrating the probability density in equation 7 over the range $x \in [1, 2]$; the remaining islands will be contigs. Also if desired, the necessity for conditioning on N_{gaps} can be dropped by performing a weighted summation over all possible values of N_{gaps} .

The distribution of the number of clones in

ROACH

an island enables the determination of the distribution of the length of that island. Some motivation also exists to predict the length of the longest island resulting from a project, as it is a readily identifiable feature of a work in progress. In particular, a failure to achieve islands of predicted length is often an indication of a technical inability to detect overlaps, and thus points to a problem that needs to be addressed. Whitworth (1897b) shows that for a given project, if the island are ordered by increasing number of fragments, the expected number of fragments in the x th island is

$E(\text{no. of fragments in } x\text{th smallest island} | N_{\text{gaps}}) =$

$$1 + \frac{n-1-N_{\text{gaps}}}{N_{\text{gaps}}} \sum_{i=1}^x \frac{1}{N_{\text{gaps}} - i + 1} \quad (9)$$

This expected value may be substituted in equation 11 below, and enables the prediction of the longest expected island for a project.

Expected Island Length

Each island is the union of one or more fragments starting at base pairs $S_k, S_{k+1}, S_{k+2}, \dots$, and S_{K+z_m-1} . The total length l_m of an island with S_k beginning its first fragment is the sum of the spacings between its fragment start sites plus the entire length of the last fragment in the island (Fig. 2):

$$l_m = \begin{cases} L + \sum_{k=S_k}^{K+z_m-2} D_k & \text{if } z_m > 1 \\ L & \text{if } z_m = 1 \end{cases} \quad (10)$$

Spacings are exchangeable in that the joint distribution of all D_k is unchanged under any permutation of subscripts. Or rephrased, the lengths of the spacings are independent of their order. Expected island length conditioned on z_m is therefore

$$E(l_m | z_m) = L + E(D_k)(z_m - 1) \quad (11)$$

And so one may approximate expected island length as

$$E(l_m) = L + E(D_k)(E(z_m) - 1) \quad (12)$$

$$= L + G_e \left(\frac{1}{n+1} \right) \left(\frac{n}{1 + (n-1)(1-f_G)^n} - 1 \right) \quad (13)$$

This approximation is most valid when the relative variance of z_m is small, the usual case for genome projects. Note that the use of $E(D_k)$ as calculated above constitutes an additional approximation, as not all spacings can be included in islands. To account for this, a modification to $E(D_k)$ must be made (Appendix A). Also, the accuracy of this approximation can be improved with the aid of a computer by summing equation 11 over all possible values of z_m , rather than employing the expected value of z_m . Based on evidence from computer simulations, however, equation 13 appears to offer enough accuracy for most purposes.

The expected fraction, f , of the target covered by fragments can be calculated directly: $f = E(l)E(N_{\text{islands}})$. This fraction can also be calculated independently (Appendix B).

Under low redundancy conditions, where $n \gg e^R$, $E(z) = e^R$. With this approximation $E(l) = [L(e^R - 1)]/R$ is obtained, identical to that of Lander and Waterman (1988) for the limit as T approaches zero. This defines an upper bound as a function of n (or R) for the accuracy of the Lander and Waterman equations. This bound occurs at a redundancy of approximately threefold. In particular, above this limit, a geometric distribution ceases to be a good model for the expected number of clones in an island. At the redundancies of six to eightfold used in common practice, geometric approximations also have the nonsensical disadvantage of predicting a fractional number of islands in conjunction with an average island length longer than the target. This arises when a significant fraction of the clones in a project are contained within a single island and the stopping probability of a geometric approximation increases. The result is that the "lack-of-memory" necessary for a Poisson analysis fails. Thus, such equations work best when a large number of islands is expected and an approximation of island length is independent.

Probability of an Island Greater Than a Critical Length

Care must be used when considering more than one island from a given project, as their lengths are not independent. For example, if there are two islands in a project, it might be that either one is greater than half the length of the target, but it is certain that both of them are not (barring undetected overlap). Now, the probability that a project contains at least i islands of length

RANDOM SUBCLONING

greater than a certain critical length, C , may be considered. Such a probability finds importance in evaluating performance of fragment assembly algorithms or in planning projects with limited goals.

Let N_{long} be the number of long spacings $D_k > L - T$. Let g_m be the number of short spacings contained between adjacent long spacings. Now $l_m \approx L + E(D)g_m$. Thus,

$$P(l_m > C) = P\left[g_m > \frac{C-L}{E(D)}\right].$$

The number of short spacings preceding the first long spacing is g_0 ; the number following the last long spacing is $g_{N_{long}}$. Because either D_0 or D_n may be long, N_{long} may be up to two greater than N_{sup} . N_{gaps} is nevertheless a good approximation to N_{long} . This is true at low redundancies where $N_{sup} \gg 2$. At high redundancies $n \gg N_{long}$ so it is unlikely that either D_0 or D_n is long. If greater accuracy is desired, an appropriate summation can be made. Because the distribution of long spacings is uniform among the set of all spacings, the distribution of g is defined on the simplex

$$\frac{g_0 + g_1 + g_2 + \dots + g_{N_{long}}}{(n+1 - N_{long})} = 1$$

with all points of the simplex equiprobable. Let R be the number of islands exceeding length C , let

$$c = \frac{C-L}{E(D)(n+1 - N_{long})}$$

and let k be the greatest integer less than $1/c$. Then, directly from Stevens (1939) one has

$$P(R > i | N_{sup}) = \sum_{j=1}^k (-1)^{j-1} \frac{(N_{sup} + 1)!}{(N_{sup} + 1 - j)! (j - i)! (i - 1)!} \frac{(1 - jc)^{N_{sup}}}{j} \quad (14)$$

In particular, the probability of having a contig in a project greater than half the length of the target can be approximated (i.e., $c = 1/2$) as follows:

$$\begin{aligned} P\left(\text{one contig} > \frac{G}{2}\right) &= \\ \sum_{v=1}^n P(N_{sup} = v) P\left(\text{one contig} > \frac{G}{2} | N_{sup} = v\right) &= \quad (15) \\ = \sum_{v=1}^n \binom{n-1}{v-1} (1 - f_G)^v [1 - (1 - f_G)^v]^{n-v} \frac{v+1}{2^v} \end{aligned}$$

At moderate to high redundancies, only the first few terms of this last sum are necessary.

Circular Targets

In the foregoing, considerations of a linear target forced somewhat awkward equations or mild approximations to obtain simpler equations. With circular targets, many of these considerations are unnecessary. In particular, $G_e = G$. Modifications to equations 1 and 3 result in

$$f_{D_i}(x) = (n-1) \left(1 - \frac{x}{G}\right)^{n-2} \quad (17)$$

$$\begin{aligned} P(N_{gaps} = x) &= P(N_{islands} = x) \\ &= \left(\frac{n}{x}\right) p_{sup}^x (1 - p_{sup})^{n-x} \end{aligned} \quad (3')$$

Other results follow appropriately. Note that these equations are not significantly different from those for a linear target. The literature provides an exact formula for the expected number of fragments needed for closure of a circular target (Flatto and Konheim, 1962). Begging "edge effects," this equation can also be applied to the line. It is, as T approaches zero, and where B is the greatest integer smaller than G/L .

$E(n \text{ needed for closure}) =$

$$1 - \sum_{k=1}^B (-1)^k \frac{\left(1 - k \frac{L}{G}\right)^{k-1}}{\left(k \frac{L}{G}\right)^{k-1}} \quad (16)$$

Also, some results are available for the case of a varying parameter $L - T$, where its distribution is known (Siegel and Holst 1982). These results include the distribution for the number of gaps and its corollary, the probability of project closure. These results may be applied also as approximations to the linear case. Conversely, the linear results may also be applied as approximations to the circular case.

Simulations and Experimental Data

A large number of Monte Carlo simulations of projects can be generated quickly with a computer. They provide a useful comparison with the mathematical models (Fig. 3) and demonstrate the accuracy of the present model. Agreement with experimental data, where available, is good also. There are few compilations of robust statis-

ROACH

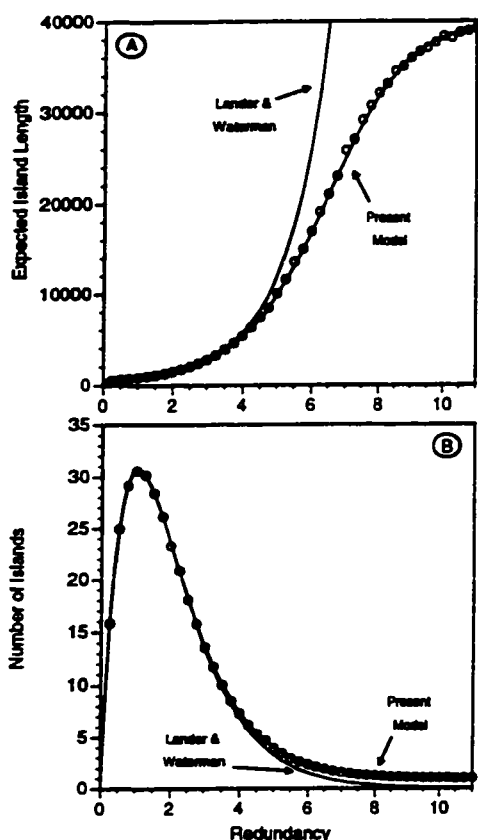


Figure 3 (A) Expected island length modeled as a function of redundancy for a typical cosmid sequencing project. Project parameters are G , 40,000; L , 500; T , 20. Data points are the average of 1000 independent Monte Carlo simulations (\circ). The Lander and Waterman model (1988) is shown as a reference. (B) The expected number of islands modeled with the same parameters.

tical data taken during intermediate project assembly points, particularly because compiling a statistically significant set of such data is burdensome. Nevertheless, the experience in our laboratory is that cosmid sequencing projects require redundancies around sevenfold for closure (Rowen and Koop 1994). Results from other laboratories support this view (Davison 1991; Bodenteich et al. 1994; Martin-Gallardo et al. 1994).

DISCUSSION

The major utility of the mathematical approach presented here stems from an initial determina-

tion of the distribution of the spacings between adjacent fragment start sites and the subsequent use of geometric probability. This general approach to modeling finite genomes may be useful for other mapping and sequencing strategies, such as those based on random transposon insertion. Such strategies represent a genomics implementation of the well-established "coverage process" theory. Hall provides a nice entry into some of the relevant mathematics literature (1988).

The equations derived here have many applications. To begin with, a strategist is interested in the amount of work necessary to complete a project. This can be expressed as the probability of project completion at a given redundancy, where project completion is defined as having closed all gaps. This is given by equation 4 and is graphed in Figure 4. These results are consistent with the expected redundancy needed for closure (equation 16), which is also indicated in Figure 4. Not surprisingly, longer targets have a higher cost in redundancy to close. Also, in general, fewer longer sequences are more desirable than a proportionally greater number of shorter sequences (or their mapping equivalent).

The cost of directed sequencing is roughly constant per gap, no matter how long the gap is. However, there is an exponentially increasing cost in redundancy to close gaps as shotgun projects proceed. Therefore, choosing whether and at which point to stop shotgunning and begin directed sequencing is a fundamental economic question. For this purpose it is useful to calculate the incremental redundancy cost of shotgun projects per gap expected to be closed (Fig. 5). This cost can be compared with the cost of directed sequencing to determine if and when directed methodology is appropriate for a project.

Several simplifying assumptions were employed in the present work. In particular, uniform distribution of fragment start sites may not be the case for some shotgun projects (see Deininger 1983). However, when variations in uniformity are local, the uniform distribution is an excellent approximation. This is valid for most projects, particularly with modern DNA shearing techniques. Larger variations in uniformity tend to be target idiosyncratic and difficult to model. Such variation may result from target regions that are genetically unstable and thus absent from the subclone library. This points out one particular utility of mathematical models: Unexpected deviations from predicted values serve as

RANDOM SUBCLONING

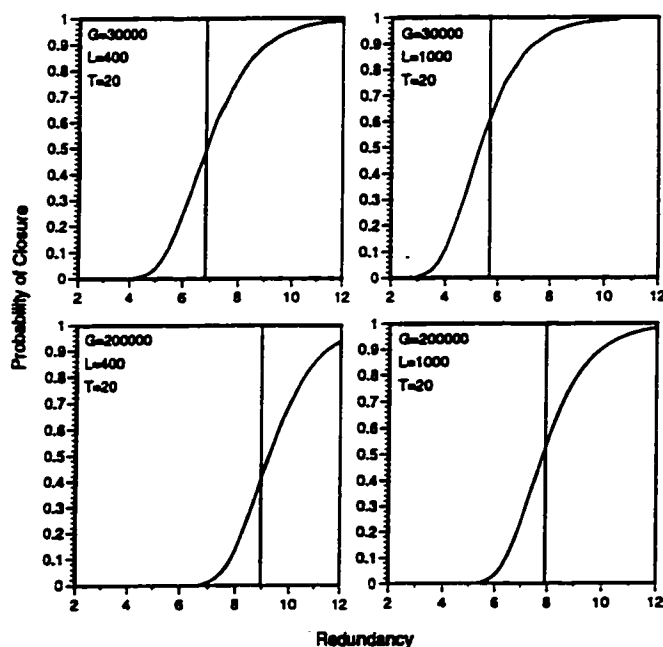


Figure 4 The probability of project completion graphed vs. redundancy, calculated using the exact equation of Stevens (1939). This equation is approximated in the text by equation 4. Four parameterizations are shown. The vertical lines intersect the expected redundancy necessary for closure, calculated using the exact equation of Flatto and Konheim (1962), with $\alpha = (L - T)/G$.

an indication of subcloning problems. Then, such problems can be addressed and corrected.

Although held constant in the present model, both fragment length (L) and necessary overlap (T) may vary in practice. Additionally, overlap may be expressed as a probability, not a certainty, and this "probability of overlap" is affected further when more than two fragments overlap at the same position. Repeated sequence elements in the target tend to decrease the probability of certain overlaps. Such considerations can be incorporated at the price of complexity into the present model by employing distributions for the values L and T , which can be combined into a distribution for L' , where $L' = L - T$. However, variations in fragment length should not cause much concern for the average genomicist. Siegel and Holst (1982) provide a proof that the expected number of gaps is dependent only on the expected fragment length, conditioning on the number of fragments. This proof is pro-

vided for coverage of a circle and also applies to the infinite line. Variation of the expected number of gaps on a finite line caused by variation of fragment length thus is expected only because of edge effects and is predicted to be small. Computer simulations confirm this prediction (data not shown). Although the expected number of gaps remains constant with varying fragment lengths, the distribution of island sizes will change. The probability of project closure will also be affected. In practice, these effects are likely to be small because L does not vary greatly, particularly in sequencing projects. Additionally, in most cases $T \ll L$ and can be approximated as 0.

Computer simulations were employed in the present work to verify that the approximations reported in this paper were useful for modeling genomic projects. In general, because of the complexities involved with modeling even the simplest random sequencing strategies, computer simulations should be consid-

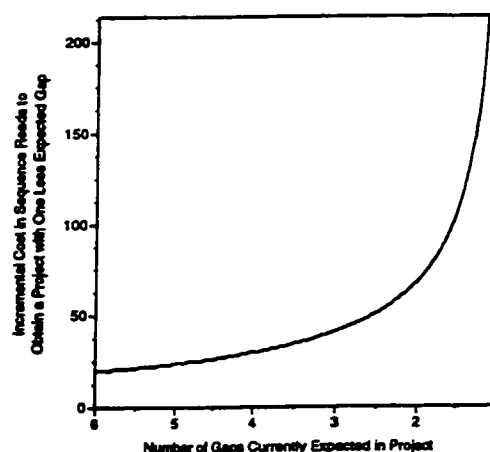


Figure 5 The incremental cost of closing one gap. This is calculated from the number of expected gaps in a project with no knowledge of a prior state of that project (see equation 5). Note that it is impossible to plan a project with zero expected gaps, because gaps always remain a small but finite possibility. G , 40,000; L , 500; T , 20.

ROACH

ered as an essential adjuvant to planning any large-scale project. Directed aspects of a project, variations in parameters, and boundary conditions are easily included in such simulations. Simplifying assumptions can be avoided. The value of simulations cannot be overemphasized. Mathematical models such as the one presented here are useful in conjunction with such simulations.

In summary, the model presented here should have utility for planning both sequencing and mapping projects, as well as for choosing realistic endpoints for such projects. In particular, the model may provide useful benchmarks for evaluating the progress of large-scale genomic projects, such as those contemplated for the human genome.

ACKNOWLEDGMENTS

I thank Alan Blanchard for inspiration, Nicholas Lovejoy for stimulating and enabling contributions, and Andrew Siegel for kind and critical commentary on the manuscript. I owe a tremendous debt to the reviewers for their many valuable comments, and to Leroy Hood, who has provided constant support. I am also grateful for a grant from the Life & Health Insurance Medical Research Fund.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- Battelle, E. 1935. Le problème de la répartition. *Comptes Rendus* **201**: 862–864.
- Bodenteich, A., S. Chisoe, Y.-F. Wang, and B.A. Roe. 1994. Shotgun cloning as the method of choice to generate templates for high-throughput dideoxynucleotide sequencing. In *Automated DNA sequencing and analysis* (ed. M. Adams, C. Fields, and J. Venter), pp. 42–50. Academic Press, New York, NY.
- Clarke, L. and J. Carbon. 1976. A colony bank containing synthetic Col E1 hybrid plasmids representative of the entire *E. coli* genome. *Cell* **9**: 91–99.
- Davison, A.J. 1991. Experience in shotgun sequencing a 134 kilobase pair DNA molecule. *DNA Sequence* **1**: 389–394.
- Deininger, P.L. 1983. Random subcloning of sonicated DNA: Application to shotgun DNA sequence analysis. *Anal. Biochem.* **129**: 216–223.
- Flatto, L. and A.G. Konheim. 1962. The random division of an interval and the random covering of a circle. *SIAM Rev.* **43**: 211–222.
- Hall, P. 1988. *Introduction to the theory of coverage processes*. John Wiley & Sons, New York, NY.
- Lander, E.S. and M.S. Waterman. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2**: 231–239.
- Lévy, P. 1939. Sur la division d'un segment par des points choisis au hasard. *Comptes Rendus* **208**: 147–149.
- Martin-Gallardo, A., J. Lamardin, and A. Carrano. 1994. Shotgun sequencing. In *Automated DNA sequencing and analysis* (ed. M. Adams, C. Fields, and J. Venter), pp. 37–41. Academic Press, New York, NY.
- Roach, J.C., C. Boysen, K. Wang, and L. Hood. 1995. Pairwise end sequencing: A unified approach to genomic mapping and sequencing. *Genomics* **26**: 345–353.
- Robbins, H.E. 1944. On the measure of a random set. *Ann. Math. Stat.* **15**: 70–74.
- Rowen, L. and B.F. Koop. 1994. Zen and the art of large-scale genomic sequencing. In *Automated DNA sequencing and analysis* (ed. M. Adams, C. Fields, and J. Venter), pp. 167–174. Academic Press, New York, NY.
- Siegel, A.F. 1979. Asymptotic coverage distributions on the circle. *Ann. Probab.* **74**: 651–661.
- Siegel, A.F. and L. Holst. 1982. Covering the circle with random arcs of random sizes. *J. Appl. Probab.* **19**: 373–381.
- Staden, R. 1980. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res.* **8**: 3673–3694.
- Stevens, W.L. 1939. Solution to a geometrical problem in probability. *Ann. Eugen.* **9**: 315–320.
- Whitworth, W.A. 1897a. *DCC exercises in choice and chance*. Cambridge University Press, Cambridge, UK.
- . 1897b. *Choice and chance*. Cambridge University Press, Cambridge, UK.

Received October 11, 1995; accepted in revised form November 21, 1995.

APPENDIX A: TRUNCATED SPACING DISTRIBUTION

Spacings greater than $L - T$ form gaps, and so are not included in the subset of spacings that may be included in the length of an island. To proceed, these spacings must be eliminated from the distribution of D_k (equation 1) by truncating and normalizing. The expected value of this truncated distribution is (note the change of variable $y = x/G_c$):

$$E(D_k | D_k \leq L - T) = \frac{\int_0^{L-T} xn \left(1 - \frac{x}{G_c}\right)^{n-1} dx}{\int_0^{L-T} n \left(1 - \frac{x}{G_c}\right)^{n-1} dx} = \quad (17)$$

$$G_c \frac{\int_0^{f_c} y(1-y)^{n-1} dy}{\int_0^{f_c} (1-y)^{n-1} dy} = G_c \left[\frac{1 - (1 + nf_c)(1 - f_c)^n}{(n+1)[1 - (1 - f_c)^n]} \right]$$

This value should be used in place of $E(D_k)$ in equation 12.

APPENDIX B: TARGET COVERAGE

The nondegenerate fraction of the target f represented in fragments is

$$f = 1 - e^{-R} \quad (18)$$

This is a characteristic equation for the coverage derived from a random cloning strategy, and in genomics literature is attributed to Clarke and Carbon (1976). A similar derivation is presented below, with comments to extend its applicability, particularly to pairwise projects where both ends of a fragment, but not the center, are characterized (see Roach et al. 1995).

Let a shotgun strategy be executed such that p_x is the probability of coverage of base pair x by any given fragment. The probability that the base pair is covered by at least one fragment is thus $1 - (1 - p_x)^G$. The expected value and higher moments of f follow from the results of Robbins (1944), with

$$E(f) = \frac{1}{G} \sum_{x=1}^G [1 - (1 - p_x)^G] \quad (19)$$

When $p_x = L/G$ for all x , this expected value is approximated by equation 18. Note that for pairwise projects L is replaced by the characterized length of each fragment. For linear targets p_x is not constant, and falls off near the edges. Despite this, unless L is a significant fraction of G equation 18 remains an adequate approximation to equation 19.

Using the approach of the present paper, f may be obtained also by subtracting the sum of the gap lengths from the total target length. This may be employed as an alternative approach to calculating expected island length. Note that an excess of negative gap lengths will result in clonal coverage in apparent excess of the total target length. This is most apparent when T is large. If the actual coverage is desired, the length of a gap should be calculated as $D - L$ for $D > L$.

Pairwise End Sequencing: A Unified Approach to Genomic Mapping and Sequencing

JARED C. ROACH,^{*}¹ CECILIE BOYSEN,[†] KAI WANG,^{*} AND LEROY HOOD^{*}

^{*}Department of Molecular Biotechnology, University of Washington, Seattle, Washington, 98195; and
[†]Division of Biology, California Institute of Technology, Pasadena, California 91125

Received September 19, 1994; accepted November 30, 1994

Strategies for large-scale genomic DNA sequencing currently require physical mapping, followed by detailed mapping, and finally sequencing. The level of mapping detail determines the amount of effort, or sequence redundancy, required to finish a project. Current strategies attempt to find a balance between mapping and sequencing efforts. One such approach is to employ strategies that use sequence data to build physical maps. Such maps alleviate the need for prior mapping and reduce the final required sequence redundancy. To this end, the utility of correlating pairs of sequence data derived from both ends of subcloned templates is well recognized. However, optimal strategies employing such pairwise data have not been established. In the present work, we simulate and analyze the parameters of pairwise sequencing projects including template length, sequence read length, and total sequence redundancy. One pairwise strategy based on sequencing both ends of plasmid subclones is recommended and illustrated with raw data simulations. We find that pairwise strategies are effective with both small (cosmid) and large (megaYAC) targets and produce ordered sequence data with a high level of mapping completeness. They are ideal for fine-scale mapping and gene finding and as initial steps for either a high- or a low-redundancy sequencing effort. Such strategies are highly automatable. © 1995

Academic Press, Inc.

INTRODUCTION

The maturing science of genomics is developing an armamentarium of techniques and strategies both to map and to sequence genomes (Evans, 1991; Burland *et al.*, 1993; Li and Tucker, 1993; Kasai *et al.*, 1992; Siemieniak *et al.*, 1991). Large-scale genomic sequencing projects are typically divided into two phases: mapping followed by sequencing. Some strategies iterate this process with interwoven mapping and sequencing

phases. A common strategy is to produce a rough map of approximately 40-kb completeness (terminology of Olson and Green, 1993), which is the level of cosmids. Cosmids provided by such a mapping effort can then be employed in sequencing strategies, often using a random shotgun approach, which employs random start locations for sequence reads. Alternatively, a directed strategy may be used to provide a map of extremely fine detail—with markers spaced less than a sequence read length apart—followed by “one-pass” sequencing.

To a large extent, directed strategies exist to overcome the major drawback of the shotgun strategy, the need for overdetermination of target sequence to maximize gap closure and minimize errors. As shotgun projects progress, due to their random nature, it becomes exponentially more difficult to generate novel data and to close gaps in the target sequence. Directed strategies bridge this problem by deriving fine-scale maps before sequencing, but at a cost of time and effort. Most sequencing projects seek a balance between directed and random approaches. A review of this dichotomy is provided by Chen *et al.* (1993).

Here we seek a complete integration of mapping and sequencing, with a fine-scale map arising automatically from sequence data as a project proceeds. The strategy that we describe retains the simplicity of random shotgun approaches, but, due to the fine-scale map produced, eliminates the need for more than minimal overdetermination of target sequence. Its primary process of “scaffold-building,” described below, is highly automatable and requires neither iterative steps nor intervention from highly trained individuals. We note that this strategy requires low sequence redundancy to achieve target-spanning maps. However, since sequence accuracy is largely a function of its redundancy, we also discuss methods for improving local coverage and thus global accuracy. Nevertheless, we emphasize the utility of a “low-pass” approach.

FORMULATION

A sequencing project begins with a target, denoted *G*, often cosmid-sized, but conceivably much larger. The

¹ To whom correspondence should be addressed. Telephone: (206) 685-7367. Fax: (206) 685-7301. E-mail: roach@u.washington.edu.

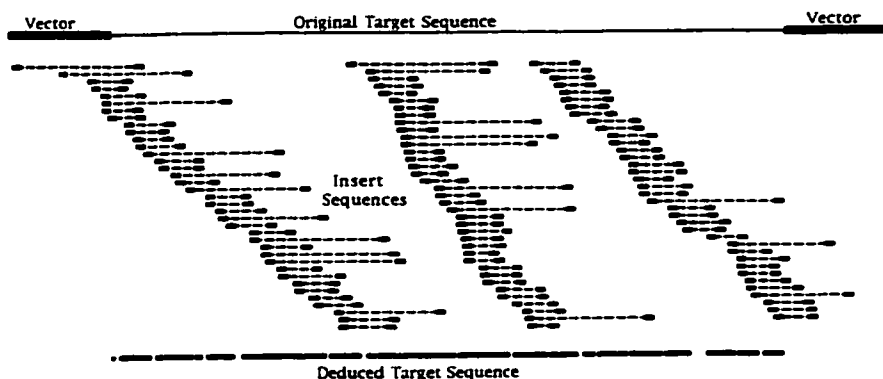


FIG. 1. A model "double-barrel shotgun" assembly. A 2.25 sequence redundancy produces 18 contigs that span 90% of an original target cosmid at 99.9% accuracy. Contig orientation and order are determined as shown. All but one gap are less than 400 bp; the remaining is 751 bp. More statistics are presented in Table 1.

strategy presented here envisions fragmenting multiple independent copies of this target into many small subcloned inserts, sequencing both ends of these inserts, and building "scaffolds" from analyzed sequence data (Fig. 1). In this paper we simulate insert sizes ranging from 1 to 10 kb. Our computer simulations, except where indicated, assume a constant insert length, L . All possible subcloned fragments are considered equiprobable.

Sequence read length, denoted L , can vary from 350 bp to over 800 bp, depending on sequencing protocols and instrumentation. For computer simulations we assume L to be a constant 400 bp. The number of inserts successfully sequenced in a project is denoted n . Since inserts are sequenced at both ends, the total number of sequences will be $2n$, and the total amount of sequence determined will be $2nL$. The redundancy of sequence data, denoted R_s , is defined to be $2nL/G$. Most of our results depend primarily on redundancy and only secondarily on sequence read length or quantity. Thus, many short sequence reads are roughly equivalent to proportionally fewer long reads, and our choice of 400 bp for L is not critical. We also note that in mapping projects redundancy is usually defined as the total length of all subcloned inserts analyzed. In the formulation presented here this quantity is denoted R_m and defined to be nL/G . The use of R_m permits comparison of our mapping results with other mapping techniques, such as restriction mapping.

After all insert end sequences have been determined, data can be analyzed and sequences can be assembled into islands and contigs (terminology of Staden, 1980). An island is a set of overlapping sequences such that a path can be traced between any two members of the set; a contig is an island consisting of at least two sequences. With a pairwise sequencing strategy, assembly of contigs is facilitated by knowledge of the pairwise orientation of sequences derived from the same insert. Such knowledge was first used extensively during the sequencing of the HPRT locus (Edwards *et al.*, 1990).

Pairwise knowledge also permits discrete islands to be ordered and oriented with respect to each other. This ordering and orienting creates a map with sequence islands as landmarks and permits mapping to be integrated into the sequencing phase of a project. Such landmarks have been referred to as "mapped and sequenced tags," or MASTs (Smith *et al.*, 1994). The size of the gaps between islands determines the completeness of the map (Olson and Green, 1993).

At low redundancies, it will not necessarily be possible to determine a single nondegenerate map for a project, as there may be sequence islands for which order or orientation is not determined. For a map to be finished, there must exist a path of bridging inserts between any two sequence islands, either directly or indirectly through other islands. Until enough redundancy is present to overcome this potential problem, there may be multiple coexisting and possibly overlapping maps. To address and discuss this issue, we define an ordered and oriented list of sequence islands to be a "scaffold." Since a scaffold consists of one or more overlapping subcloned inserts, it could also be legitimately called an island, and would be if our discussion centered solely on mapping issues. Here, we reserve the term "island" to denote a set of overlapping sequences.

Beyond a certain point, as redundancy increases, the number of both islands and scaffolds will decrease, ultimately resulting in a single scaffold. Such scaffolds usually contain the entire target and as such are termed "complete." A complete scaffold usually contains vector sequence as well, but for statistical purposes is considered to be equal to the length of the target. The longest scaffold resulting from a project is termed the "maximum" scaffold. Gaps in sequence data internal to a scaffold have previously been termed "sequence-mapped gaps," or SMGs² (Edwards and Caskey, 1991).

² Abbreviations used: YAC, yeast artificial chromosome; BAC, bacterial artificial chromosome; PCR, polymerase chain reaction; SMG, sequence-mapped gap; STS, sequence-tagged site.

For our computer simulations, we assume that a mutual overlap of length T is necessary and sufficient to detect overlap between two sequence reads. This overlap T was set at 30 bp, but we note that the effects of varying T are slight, particularly because $T \ll L$. In particular, assigning T any value between 1 and 50 does not noticeably alter our results (data not shown).

For most projects, a target sequence will have been fragmented along with its vector (i.e., YAC, BAC, cosmid, phage). To minimize the sequencing of vector, usually one employs either target sequence as a probe to pick positive inserts or vector sequence as a probe to screen out vector. We present here only simulations of the first strategy, which we also find to be representative of other strategies (data not shown). To this end, we assume that any insert that contains at least 40 bp of target sequence is a candidate for inclusion in a project. One advantage of this "positive screening" approach is that a few inserts will overlap vector sequence and can be used to anchor the ends of some scaffolds to the vector. However, this effect is slight, especially with longer target lengths.

Our analysis centers on two target lengths: 35 and 200 kb. We chose 35 kb as a representative length for cosmids, which currently form the majority of our targets. We chose 200 kb as a representative length for a BAC or YAC, to demonstrate the feasibility of using pairwise data to facilitate shotgun sequencing targets of that size or larger. All computer simulation data represent the average of 100 determinations.

COMPUTER SIMULATIONS

We find that complete scaffolds are an ideal project endpoint and thus have sought to determine optimal methods for their derivation. We have also characterized expected values for certain parameters, including average SMG size and total scaffold length. To these ends, we employed computer simulations that we in turn supplemented with a raw data simulation based on a highly redundant random shotgun project.

The results of our computer simulations are presented for cosmid- (Fig. 2) and BAC-sized (Fig. 3) targets. In general, the number of scaffolds rises sharply at low redundancies and then declines at higher redundancies. The sharp rise occurs since each insert sequence data pair added to a project at low redundancy has a high probability of forming a new scaffold. At higher redundancies, inserts begin to merge scaffolds and their number drops. For most projects, single scaffolds formed at a sequence redundancy between twofold and threefold. Slightly greater sequence redundancies were necessary to achieve single scaffolds from 200-kb targets than from 35-kb targets. Nonetheless, when 10-kb inserts were used, a single scaffold was always obtained at a redundancy less than twofold. In general, fewer scaffolds resulted when longer insert lengths were used. This is a result of longer inserts having a higher probability of spanning greater distances be-

tween sequence islands and emphasizes the value of using as long an insert length as possible, which maximizes R_m .

At high redundancies complete scaffolds are always obtained, as seen from our graphs of average maximum scaffold length (Figs. 2 and 3). For example, when 1.2-kb inserts are used for a 200-kb target, complete scaffolds are obtained around sevenfold redundancy. However, to obtain an improvement over traditional random shotgun sequencing strategies, complete scaffolds should be obtained at lower redundancies. This was clearly possible when longer insert lengths were employed. For example, redundancies of twofold were sufficient to ensure complete scaffolds when 10-kb inserts were simulated. When a project resulted in a single scaffold, this scaffold was also complete, or nearly so (data not shown).

We did not notice significant differences in redundancies necessary to achieve analogous results for 35-kb (Fig. 2), 200-kb (Fig. 3), or even 1-Mb targets (data not shown). This suggests that sequencing effort scales roughly linearly to results, and not exponentially, even with relatively large targets. This rough linearity stems from the use of pairwise data and indirectly from the high mapping redundancy R_m .

The number of SMGs in maximum scaffolds increases, then decreases, as sequence redundancy increases. The initial increase is due both to the increasing length of the maximum scaffold, enabling it to contain more gaps, and to the division of large gaps into smaller gaps as sequence islands bisect them. The subsequent decrease in SMGs is due to additional sequence data closing gaps. Roughly speaking, the largest number of SMGs tends to occur in complete scaffolds that have been obtained with a minimum of sequence redundancy. Thus, at redundancies between twofold and fourfold, which we envision as reasonable for pairwise projects, a significant number of SMGs are likely to result.

For many projects a complete target sequence is desired, with no gaps fragmenting continuity. For other projects, such as gene finding, complete sequence is not a priority, but gap characterization may be of interest. In general, project design should aim for gaps no longer than a single sequence read, or at most two reads. Our simulations (Figs. 2 and 3) demonstrate that large gaps occur as expected at very low redundancies, but at redundancies above 1.5 average gap length tends to be less than a single sequence read length. Also, for all projects with sequence redundancies above twofold, the maximum observed gap length tended to be less than 800 bp, requiring at most two sequence read lengths to close. Occasionally longer gaps occur. For example, at a redundancy of 2.5 with a 35-kb target, 100 simulations of a project employing 2-kb inserts contained one gap greater than 800 bp in 17 cases, and two such gaps in a single case. Above twofold redundancies, there were no significant differences in SMG length resulting from alternative choices of insert size.

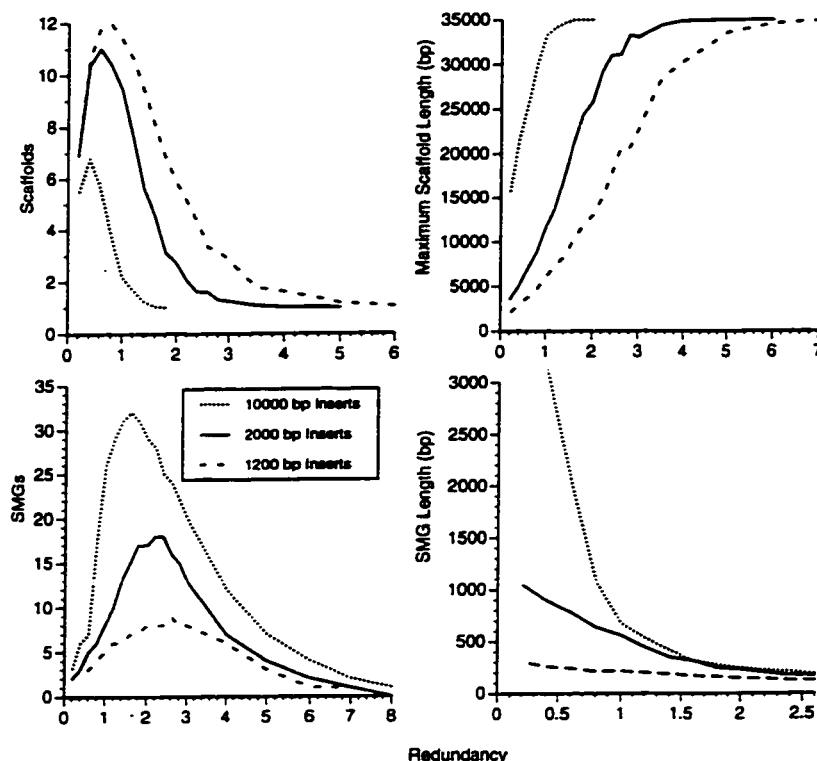


FIG. 2. Parameters from a 35-kb pairwise project evaluated as a function of sequence redundancy.

Long insert lengths are not always convenient sequencing templates. For this reason, we sought a strategy that minimized the need for longer inserts and explored strategies that employed mixtures of insert sizes. In general, we found that benefits derived from large inserts could be obtained even when they represented a small fraction of the total number of inserts sequenced. In particular, we simulated strategies that employed a mixture of 2000- and 10,000-bp inserts (Fig. 4). For these simulations we held redundancy constant at 2.25 and assumed a 200-kb target. We found no significant differences between projects utilizing entirely 10,000-bp inserts and those that used only 15% 10,000-bp inserts.

We also envisioned strategies that mix pairwise data with data derived from a single strand only, such as might be obtained with M13 templates. A relatively small fraction of pairwise data suffices for the formation of complete scaffolds that are composed largely of single-strand data (Fig. 5). With a mixture of 60% single-strand data, 30% 2000-bp pairwise insert data, and 10% 10,000-bp pairwise insert data, a maximum scaffold was reached before threefold redundancy for a 35-kb target. This simulation addresses a practical question, for sequencing reactions will occasionally fail, which implies that most pairwise projects will be supplemented with a cohort of widowed sequences.

For random subcloning projects, the fraction of the target present in subclones is approximated by the equation $1 - e^{-R}$ (Roach, submitted). Our simulations met this prediction (data not shown). At any given redundancy R , target coverage will be the same for either a traditional shotgun or a pairwise sequencing strategy. We emphasize that increased target coverage is not an advantage of pairwise strategies. At the redundancies of about 2.5 necessary to build complete scaffolds, target coverage will be about 92%.

RAW DATA SIMULATION

We wished to verify that results from our computer simulations accurately modeled real projects. Such projects utilize raw sequence data and might employ templates with significant repeat elements. In addition, we were interested in determining the ease of assembling scaffolds by hand. To our knowledge, no computer programs are yet available for this purpose. To this end, we designed a simulation built around a cosmid from the human T-cell receptor β locus sequenced in our lab.

The cosmid A1-4 has been sequenced using a random shotgun strategy to a final redundancy of 8.4 (Koop *et al.*, 1993). This cosmid consists of a 35,343-bp target

PAIRWISE END SEQUENCING

349

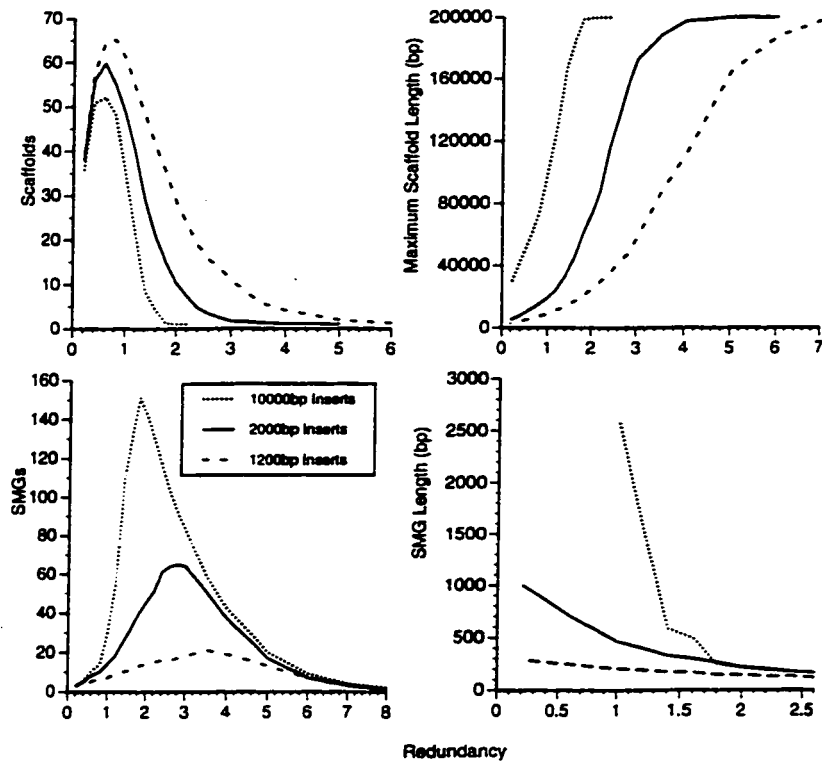


FIG. 3. Parameters from a 200-kb pairwise project evaluated as a function of sequence redundancy.

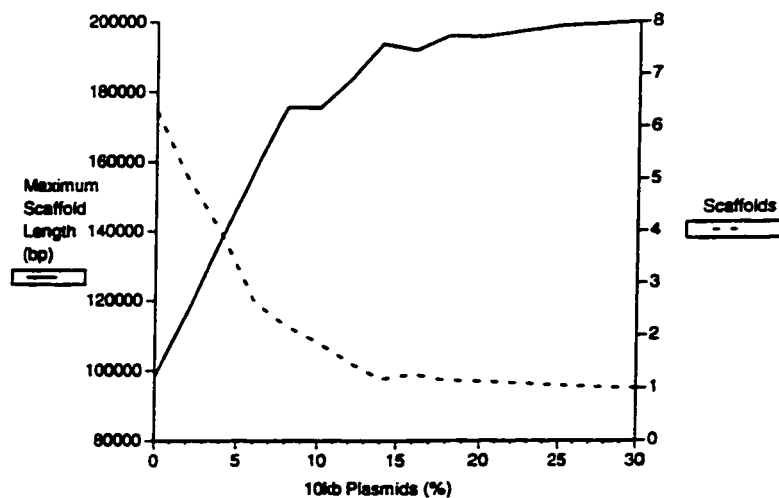


FIG. 4. Pairwise strategies employing a mix of insert sizes were simulated. Here, a mix of 2000- and 10,000-bp inserts was simulated at a constant sequence redundancy (2.25). As seen, a small proportion of larger inserts produces results comparable to those achieved when only large inserts are used. At 2.25 redundancy, complete scaffolds can be obtained with only a 15% mix of longer insert lengths. A 200-kb target was assumed.

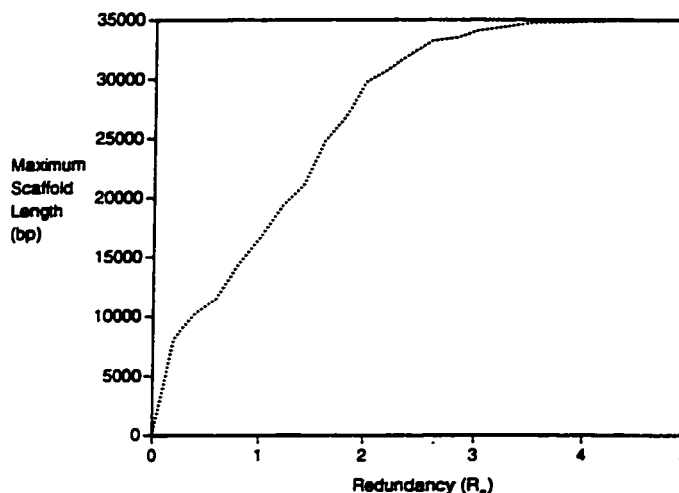


FIG. 5. A hybrid strategy employing a combination of single-strand and pairwise data was simulated. A mix of 60% single strand, 30% 2000 bp, and 10% 10,000 bp data was employed. With this approach a complete scaffold can be obtained at less than threefold redundancy. A 35-kb target was assumed.

cloned into an 8213-bp vector. The target is notable in that it contains several repeats, including two 8.4-kb homologous elements. Their similarity ranges from 85% to over 99% when 400-bp sliding windows are used for analysis. For this reason, the cosmid A1-4 was judged to represent a significant challenge for assembly (Lee Rowen, Seattle, pers. comm., 1994). The sequences used for the original assembly of A1-4 were derived primarily from single-stranded M13 templates and were sequenced with either Sequenase or *Taq* cycle sequencing protocols.

For our pairwise assembly simulation we chose a subset of these 678 sequences that might represent typical data from a pairwise project. To this end we planned for a 2.25 final redundancy R_p . We wished to pursue a strategy that employed a mix of long and short templates, so we simulated 88 2500-bp fragments and 22 7000-bp fragments. We determined the start locations of these fragments with a random number generator. The length of the fragments was modified randomly with a squarewave to simulate uncertainty in fragment length, as might occur if such fragments were size selected by banding on an agarose gel. Sequence reads of the proper orientation were then chosen from the A1-4 data set to represent the pairwise end sequences of our hypothetical fragments. The closest raw sequences to our randomly generated fragment endpoints were used, although no sequence was used twice. The final range of short fragment lengths was 1738–3418 bp (2375 ± 287 SD), while the range of long fragments was 5312–8245 bp (6819 ± 781 SD). For our initial assembly, all sequences longer than 400 bp were clipped to 400 bp to demonstrate that long sequence reads are not necessary for the success of pairwise assemblies. In addition, a few sequences were shorter,

although no sequence was less than 250 bp. Our final redundancy R_p was thus slightly less than 2.25. We judged our protocol for sequence selection to be a reasonable approximation of what might be likely to result from an actual pairwise project.

One of us (J.C.R.) then assembled these pairwise sequences into a single scaffold. Before and during this assembly he was blind to the nature of the repeats in A1-4 other than that it was a "difficult" cosmid. Additionally he was blind to the exact length of the fragments, other than that they were either "long" or "short." Sequence contigs were assembled with the software package DNA* (Madison, WI). Scaffolds were assembled by sliding pieces of paper on a large table and were ultimately merged into a single scaffold (Fig. 1). This assembly took about a day, illustrating that it would be a task suitable for software. Following the generation of this scaffold, each sequence contig was edited by hand for maximum accuracy. At this point, for editing purposes, the ends of sequences extending beyond 400 bp were used. Some of these sequences, although often low accuracy, served to verify the ordering and orienting of the contigs within the scaffold. Additionally, they helped improve the overall accuracy of the sequence data.

The results of this raw data simulation compared favorably with the averages predicted by our computer simulations (Table 1). Eighty-nine percent of the target sequence was represented in this scaffold. The remaining unknown sequence was contained in 17 SMGs. Sequence accuracy was 99.9%. All but one of the 44 errors were present in regions covered by only a single strand. This suggests that double strand coverage is capable of obtaining extremely high accuracy, which could be obtained for these regions by sequencing oppo-

TABLE 1

Results from a Raw Data Simulation of a Pairwise Strategy Compared with Values from a Computer Simulation

	Computer simulation	Raw data simulation
Number of scaffolds	1.02	1
Scaffold length (bp)	35,267	35,343
Number of SMGs	21	17
Average SMG length (bp)	169	223
% target covered	90.1	89.2

Note. Results from a raw data simulation of a pairwise strategy employed on a 35,343-bp cosmid that had previously been shotgun sequenced to ninefold redundancy. For this simulation, 2.25 sequence redundancy was derived from the end sequences of a mixture of hypothetical subclone inserts. 20% approximately 7000 bp in length and 80% approximately 2500 bp in length. These results agree with results from our computer simulations and suggest the practicality of obtaining complete scaffolds in the range of twofold redundancy. The computer simulation column depicts average values from 100 independent determinations.

site strands. The exact lengths of the 17 SMGs were unknown, but could be estimated. Ten of these SMGs were spanned by the low-quality ends of sequence reads present in our data set. These data were insufficient for base calling, but allowed the estimation of gap lengths to within a few basepairs. The lengths of the remaining SMGs could be estimated based on the lengths of the fragments that spanned them. As subsequently verified, all 17 SMGs were less than 800 bp, and all but two were less than 300 bp.

DISCUSSION

We have simulated a variety of approaches to sequencing large targets. Our simulations revolve around the utility of pairs of sequence data derived from opposite ends of subclones. We refer to strategies employing such data as pairwise end-sequencing, or more colloquially, "double-barrel shotgun" strategies. We were primarily interested in determining the minimum amount of sequence redundancy necessary to reach satisfactory project endpoints. We were also interested in determining the optimum fragment size to sequence.

We defined a "scaffold" to be an ordered and oriented list of sequence islands. SMGs between such islands are determined in such a manner that their length is approximately known. We feel that production of a scaffold that equals or exceeds target length is an ideal endpoint for a random strategy. Strategies based on our simulations achieve such scaffolds at sequence redundancies around twofold.

A key factor in producing scaffolds at twofold redundancy is the choice of insert lengths. We found that the longer an insert is, the more useful it is. This is in considerable contrast with a common misconception that the ideal insert length is three times the sequence read length. Nevertheless, there is a practical upper

limit to useful insert size. This limit depends on three factors. First, it is difficult to clone large fragments routinely. Second, longer inserts have correspondingly more sequence complexity, which tends to degrade the quality of the raw data. Third, assembly becomes more difficult with longer fragments, as the absolute uncertainty of the length between pairwise ends tends to increase. These limitations vary in stringency depending on available technology and resources. Thus, the optimal choice of fragment size may vary from one laboratory to another. However, given the option, fragment sizes should be chosen as large as possible. It should be noted that in addition to their advantages in scaffold-building, large fragments are also extremely useful in detecting and resolving repetitive elements in target sequence.

The use of a mixture of small and large inserts gains most of the advantages that would occur with the sole use of large inserts (Figs. 4 and 5). This is true even when the large inserts represent a relatively small fraction of the total. Generally speaking, the total length of all of the inserts should be chosen to maximize the mapping redundancy R_m . If for technical reasons an insert library is constructed of a single intermediate size, a slightly higher sequence redundancy can be used to ensure completeness (Figs. 2 and 3). The exact balance between redundancy and insert lengths will depend on the laboratory and should be determined on a case-by-case basis with the aid of computer simulations.

We find that "double-barrel" shotgun sequencing has many advantages over traditional "single-barrel" shotgun sequencing. Notably the mapping redundancy R_m for single-barrel sequencing is $2nL = R_s$. The mapping redundancy for double-barrel sequencing is nL , which should be several times greater than R_s . This creates a high-redundancy mapping situation, which permits efficient low-redundancy sequencing. This sequencing builds complete scaffolds for which the location and lengths of all gaps are determined. Pairwise strategies are not confined to low-pass sequencing and are equally valuable at high redundancies, particularly for sequence assembly. For these reasons, we feel that all random strategies should employ pairwise data, at least with the goal of generating complete scaffolds as a basis for further sequencing. Such sequencing can either continue to be random or switch to directed approaches.

We expect that most projects will move to directed sequencing after a complete scaffold is obtained. This "gap closure" will entail obtaining sequence for SMGs as well as reverse sequence from regions of single-strand coverage. The templates localized during scaffold construction are ideal substrates for such directed sequencing. One gap closure methodology is to sequence a PCR product spanning the gap, which should be capable of closing any of the gaps present in complete scaffolds. If the entire target sequence is not needed, only gaps of interest need be filled. This might

be the case for a gene finding effort, or across an element such as an *Alu* repeat.

For any reasonably large project, computational tools will be necessary to assemble and analyze scaffolds. We expect that such software will evolve in the near future. A simple assembly algorithm is to first assemble individual sequences into islands, blind to their pairwise nature. Second, order the resulting sequence islands by linking together sequences with their mates from opposite ends of the inserts. Third, check for inconsistencies, remove suspect pairs of sequences, and iterate the process. Finally, make rough estimates of gap distances based on insert lengths and on low-quality sequence read end data. This algorithm was successfully employed to assemble the cosmid A1-4, which we believe to have been a robust test of its efficiency.

Algorithms employing pairwise assembly data are more robust and accurate than traditional assembly algorithms. Each sequence offers a positional check on its mate, allowing a majority of misplaced sequences to be located immediately following an assembly. Without this check, we would indeed have misplaced several sequences during our raw data simulation of the A1-4 assembly, particularly within one of the 8.4-kb homologous repeats. In general, repeats pose little problem for these algorithms. The key limit for repeat detection with any sequencing strategy is the length of the longest effective subclone. For pairwise sequencing this limit is the insert length, which can be an order of magnitude greater than that of traditional shotgun sequencing.

One aspect of potential concern to pairwise sequencing is uncertainty in the exact lengths of pairwise inserts. Such uncertainty will arise if fragments are band purified on an agarose gel and not subsequently characterized. In extreme cases, particularly at low redundancies, such uncertainty might result in indeterminate island order within a scaffold. However, in our raw data simulation of cosmid A1-4, we found that a redundancy R , of 2.25 was more than enough to avoid any such problems. Thus, a knowledge of exact insert lengths would have contributed little to our project.

The computer simulations presented in this paper hold both sequence read length L and insert lengths I constant. In actual projects, such as that represented by our raw data simulation, these parameters are expected to vary. We have incorporated these considerations into several additional computer simulations (data not shown), particularly by allowing I to vary as a squarewave centered on a target value. No significant differences in predicted results were noticed when I was allowed to vary. Variations in L also have no significant effect, as long as redundancy remains constant (data not shown).

Another assumption of our computer simulations was that all target fragments are equiprobable. The accuracy of this approximation is dependent on the fragmentation method (see, for example, Deininger, 1983). For most cases this approximation is quite valid,

because the regions of fluctuation in fragmentation probability tend to be smaller than the length of the inserts. Such deviations are easily modeled in computer simulations, but due to their idiosyncratic nature are not explicitly presented here.

We believe that there are many reasonable implementations of the pairwise strategy, which was originally proposed by Edwards and Caskey (1991). One example known as "ordered shotgun sequencing," or OSS, has been suggested by Chen *et al.* (1993). OSS is characterized by a low-redundancy pairwise approach that produces multiple unlinked scaffolds that form the basis for further directed sequencing. Further simulations and a review of pairwise strategies have been provided by Richards *et al.* (1994). In the present paper we present pairwise strategies in their most general form, accompanied by computer simulations that provide previously unavailable information, allowing the pursuit of optimal strategies. Obviously, such strategies should be adapted to individual projects, techniques, and laboratories. However, projects should be designed in accord with the general principles elucidated in this paper. In particular, an effort should be made to maximize insert lengths. Also, we see no advantage in sequencing one insert end initially and subsequently a subset of the pairwise ends. We believe it is more efficient to sequence all ends simultaneously. Additionally, we see no particular advantage in halting pairwise projects before complete scaffolds are achieved. The extra sequence redundancy necessary to achieve complete scaffolds is relatively small compared with the labor that otherwise would be necessary to assemble several unlinked scaffolds into a map. An extreme example of this last option is presented by Smith *et al.* (1994), in which cosmid clones are entirely mapped before their ends are sequenced.

In particular, we find a PCR-based implementation of the double-barrel strategy attractive (K. Wang, L. Gan, C. Boysen, and L. Hood, submitted). In this implementation we use colony PCR to recover plasmid inserts. The PCR products are sequenced with forward and reverse primers. A 96-well format is used from start to finish, and ABI 373 gels are used to generate sequence data. A complete scaffold is generated quickly and efficiently. Low-redundancy pairwise strategies such as this are particularly useful for gene finding, as they provide most of the sequence data from a target region, which can then be utilized in similarity or feature identification searches. Regions of interest can be singled out for special attention facilitated by the structured nature of the scaffold.

We believe that pairwise strategies can effectively handle megabase targets. Our simulations demonstrate that sequence redundancies between two- and threefold are more than adequate to span such targets with complete scaffolds (data not shown). By permitting direct shotgun sequencing, double-barrel strategies eliminate the need to use intermediate subclones of large mapping vectors such as BACs or YACs. This

elimination of cosmid subcloning and mapping can represent a significant increase in the efficiency of genomic sequencing efforts. We would particularly like to recommend double-barrel shotgun sequencing for small bacterial and viral genomes.

In summary, pairwise end sequencing can be characterized as mapping at high redundancy, but sequencing at low redundancy. It generates complete scaffolds more economically and more quickly than traditional shotgun sequencing. The advantages of this strategy include its simplicity and the absence of any need for clone mapping other than that which results as an incidental by-product of sequencing. It is capable of handling relatively large repeats or complex templates. Its utility includes STS generation, gene finding, low- and high-pass sequencing, and ultra-fine-scale mapping.

ACKNOWLEDGMENT

J.C.R. is supported by a grant from the Life and Health Insurance Medical Research Fund.

REFERENCES

- Burland, V., Daniels, D. L., Plunkett, G., and Blattner, F. R. (1993). Genome sequencing on both strands: The Janus strategy. *Nucleic Acids Res.* 21(15): 3385-3390.
- Chen, E. Y., Schlessinger, D., and Kere, J. (1993). Ordered shotgun sequencing, a strategy for integrating mapping and sequencing of YAC clones. *Genomics* 17: 651-656.
- Deininger, P. L. (1983). Random subcloning of sonicated DNA: Application to shotgun DNA sequence analysis. *Anal. Biochem.* 129: 216-223.
- Edwards, A., and Caskey, T. (1991). Closure strategies for random DNA sequencing. *Methods: Companion Methods Enzymol.* 3(1): 41-47.
- Edwards, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C., Zimmermann, J., Erfle, H., Caskey, C. T., and Ansorge, W. (1990). Automated DNA sequencing of the human HPRT locus. *Genomics* 6: 593-608.
- Evans, G. A. (1991). Combinatoric strategies for genome mapping. *Bioessays* 13(1): 39-44.
- Green, E. D., and Green, P. (1991). Sequence tagged site (STS) content mapping of human chromosomes: Theoretical considerations and early experiences. *PCR Methods Appl.* 1: 77-90.
- Kasai, H., Isono, S., Mineno, J., Akiyama, H., Kurnit, D. M., Berg, D. E., and Isono, K. (1992). Efficient large-scale sequencing of the *Escherichia coli* genome: Implementation of a transposon- and PCR-based strategy for the analysis of ordered λ phage clones. *Nucleic Acids Res.* 20(24): 6509-6515.
- Koop, B. F., Rowen, L., Chen, W.-Q., Deshpande, P., Lee, H., and Hood, L. (1993). Sequence length and error analysis of Sequenase[®] and automated Taq cycle sequencing methods. *BioTechniques* 14(3): 442-447.
- Li, C., and Tucker, P. W. (1993). Exoquence DNA sequencing. *Nucleic Acids Res.* 21(5): 1239-1244.
- Olson, M. V., and Green, P. (1993). Criterion for the completeness of large-scale physical maps of DNA. *Cold Spring Harbor Symp. Quant. Biol.* 53: 349-355.
- Richards, S., Muzny, D. M., Civitello, D. M., Lu, F., and Gibbs, R. A. (1994). Sequence map gaps and directed reverse sequencing for the completion of large sequencing projects. In "Automated DNA Sequencing and Analysis" (M. D. Adams, C. Fields, and J. C. Venter, Eds.), pp. 191-198, Academic Press, New York.
- Siemieniak, D. L., Sieu, L. C., and Slightom, J. L. (1991). Strategy and methods for directly sequencing cosmid clones. *Anal. Biochem.* 192: 441-448.
- Smith, M. W., Holmsen, A. L., Wei, Y. H., Peterson, M., and Evans, G. A. (1994). Genomic sequence sampling: A strategy for high resolution sequence-based physical mapping of complex genomes. *Nature Genet.* 7(1): 40-47.
- Staden, R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res.* 8: 3673-3694.

The Molecular Evolution of the Vertebrate Trypsinogens

Jared C. Roach, Kai Wang,* Lu Gan,* Leroy Hood

Department of Molecular Biotechnology, University of Washington, Seattle, Box 357730, Washington, 98195, USA

Received: 12 July 1997

Abstract. We expand the already large number of known trypsinogen nucleotide and amino acid sequences by presenting additional trypsinogen sequences from the tunicate (*Boltenia villosa*), the lamprey (*Petromyzon marinus*), the pufferfish (*Fugu rubripes*), and the frog (*Xenopus laevis*). The current array of known trypsinogen sequences now spans the entire vertebrate phylogeny. Phylogenetic analysis is made difficult by the presence of multiple isozymes within species and rates of evolution that vary highly between both species and isozymes. We nevertheless present a Fitch-Margoliash phylogeny constructed from pairwise distances. We employ this phylogeny as a vehicle for speculation on the evolution of the trypsinogen gene family as well as the general modes of evolution of multigene families. Unique attributes of the lamprey and tunicate trypsinogens are noted.

Key words: Trypsinogens — Vertebrate — Molecular evolution

Introduction

Trypsin is one of the most well studied enzymes. Its ready availability in robust quantities from the lumen of the vertebrate gut makes it particularly amenable to laboratory study (Walsh and Wilcox 1970). The trypsin and trypsinogen structures were early conquests of X-ray

crystallography (Sweet et al. 1974; Kossiakoff et al. 1977). Trypsinogen amino acid sequences were early targets of protein sequencing. For these reasons, the birth and development of the field of molecular evolution was fueled, in part, by trypsinogen sequence data.

Trypsinogen genes have received renewed attention recently, following the serendipitous discovery of trypsinogen genes within the human T-cell receptor (TCR) β locus. This genomic localization of trypsinogen genes has also been observed in mice and chickens (Lee Rowen, personal communication; Kai Wang, unpublished observation). Intrigued by these observations, we sought to expand our knowledge of trypsinogen gene sequences and their modes of evolution. Over the past few decades, a large number of protein and gene sequences have been determined for either trypsin or trypsinogen from a variety of vertebrate species. In an effort to acquire molecular data from all vertebrate classes and to increase representation from within some classes, we obtained additional cDNA and/or genomic trypsinogen sequences from the lamprey *Petromyzon marinus*, the pufferfish *Fugu rubripes*, and the frog *Xenopus laevis*. In addition, in order to examine a phylogenetic outgroup, we sequenced a trypsinogen cDNA from a urochordate, the tunicate *Boltenia villosa*.

Materials and Methods

Boltenia villosa. Poly(A) mRNA was prepared from the dissected gut of a specimen of *Boltenia villosa* (a gift of William Moody, University of Washington, WA). mRNA was reversed transcribed and cloned as cDNA into the λ -ZAP directional cloning vector (Stratagene). Additionally, RT-PCR was performed on the poly(A) mRNA with degenerate primers designed to amplify serine proteases: H 5'-CTSWCWGCWGCYCA YTG and S 5'-YMSWGGKCCNC-

* Present address: Darwin Molecular Corporation, 1631 220th St. SE, Bothell, Washington, 98021, USA

Correspondence to: J.C. Roach; e-mail roach@u.washington.edu

CRGARTC. These two primers correspond to conserved sequences of the serine protease active site. Similar primers are described by Kang et al. (1992) and Wiegand et al. (1993). A resulting PCR product of approximately 350 bp was agarose gel isolated and used as a probe to screen the cDNA library. This band failed to sequence due to polyclonality and was judged to be a diverse mixture of serine protease-derived products. Twenty-three positive plaques were picked and sequenced. The cDNAs identified by 5' end sequencing as trypsinogen were completely sequenced by primer walking on both strands with the following primers: TUN2F1 5'-TGGAACACGTGGAAAA-TAGTTCTC. TUN2R1 5'-CGAGAACTATTTCCACGTGTTCC. TUN2F2 5'-CAAGCAGCGGAGGAAGTATCTCCG. TUN2R2 5'-TCCACTAACAGTACACGCGGTGTC. TUN19F1 5'-GGTGTATACACCGTGTTCAGTG. TUN19R1 5'-ACACTGCAACACGCGGTATATACAC. and TUN2R3 5'-TTTGGATGATTAAG-GATTTTTATTG.

Petromyzon marinus. Poly(A) mRNA was prepared from the dissected gut of a *Petromyzon marinus* ammocoete (a gift of James Seeley, Hammond Bay Biological Station, MI). mRNA was reverse transcribed and cloned as cDNA into the λ -ZAP directional cloning vector (Stratagene). Additionally, RT-PCR was performed on the poly(A) mRNA with the trypsin-specific primers TRYF 5'-CTGGATCCGTGAGAC-TGGGAGAGCAC and TRYR 5'-CTGGATCCGAATCCTTGCCCTCCTC. The sequence of the 387-bp product was consistent with trypsin and was used to probe the cDNA library. Thirty-one positive plaques were sequenced at their 5' ends with the m13 reverse primer. Seven cDNAs identified by 5' end sequencing as trypsinogen were completely sequenced by primer walking on both strands with the following primers: LT2 5'-AGCCAGTGGGTCTCTGTCTG. LT3R 5'-TCACGAAGATGTTGTGCTC. LT3 5'-TCATGCTCATCAAGC-TGTCTC. LT4R 5'-ACGCACATGAGGACGTCGGGAC. LT5R 5'-AAGAGTAGTGTGTTAGATCCAC. XTA5 5'-CCGGTGCC-CCGTGGTGTG. and TRYR.

Sequence Analysis. Previously published trypsin and trypsinogen sequences were culled from Genbank using a variety of text- and homology-based searches.¹ A table of previously published sequences with original references can be found in Rypniewski et al. (1994). Additional published sequences not otherwise cited in the text are found in Titani et al. (1975), Gudmundsdottir et al. (1993), Genicot et al. (1996), and Pancer et al. (1996). Unknown sequences were analyzed with BLASTN and BLASTX searches against Genbank release 96.0 (Gish and States 1993; Altschul et al. 1990). Additionally, BEAUTY was employed as a search tool (Worley et al. 1993). Multiple sequence alignments were performed manually. The *Takifugu* (TRU25747) and *Xenopus* (XLU27330) sequences were isolated and sequenced as described previously (Wang et al. 1995).

Phylogenetic Analysis. Phylogenetic trees were estimated with FITCH, an element of the PHYLIP package (Felsenstein 1993). Protein distances were calculated with a special metric to take advantage of the large number of known biological constraints affecting trypsin residues. Most current distance metrics treat all sites equivalently. Ideally, we would construct a separate distance matrix for each trypsin residue site. Reliable construction of such a matrix would require more sequences than we have at present, but may eventually be possible. We thus employed a simple method that we suspect provides a reasonable approximation to what might be obtained with more data. We anticipate future computational advances that will permit a combination of site-

specific data and maximum likelihood methodology to surpass our current approach.

We assume that we have a large enough collection of sequences to have observed all possible permitted residues at each site. This assumption is more valid for sites with few observed residues. These sites are the most significant contributors to distance calculations, making this assumption reasonable. Let the number of permitted residues at site i be N_i . Ignore sites where $N_i = 1$. Calculate the probability of a residue remaining unchanged at site i after τ time periods as:

$$P = \frac{N_i - 1}{N_i} \left(1 - \frac{Nm}{R} \right)^\tau + \frac{1}{N_i}$$

where R is the number of possible residues ($R = 20$), and m is the probability of a residue mutating per time period. This equation represents a Jukes-Cantor model with R states (Jukes and Cantor 1969). However, it assumes that mutations to a nonpermitted residue are selected against and thus are observed as no mutational event. This model treats residues, rather than nucleotides, as mutable elements. The main incentive to use this model is to account for a slower rate of change at sites that, due to functional constraints, have few permitted residues. Our distances are calculated as the value of τ that gives the maximum likelihood of the observed differences between two sequences. For the present paper, the rate m is arbitrarily set to 0.01. Were a molecular clock hypothesis to hold, m could be set empirically. Small alterations in m affect the scale of a derived phylogenetic tree, but not its topology or proportions.

Results

Tunicate Trypsinogen

We screened a tunicate (*Boltonia villosa*) intestine cDNA library with a degenerate serine protease PCR probe. We sequenced 23 positive plaques, including several false positives that were not serine proteases. Of these 23 sequences, seven were the trypsinogen sequence reported here (Fig. 1). Five were chymotrypsinogen, two were ribosomal proteins, one was actin, one was glutathione S-transferase, one was from the mitochondrial 16S RNA, and six were not positively identified. The chymotrypsinogen sequences were identified based on the presence of a methionine at the position that is 192 according to the bovine chymotrypsinogen numbering system (Zwilling and Neurath 1981), as well as overall similarity.

The tunicate trypsinogen cDNA we present here was identified based on its sequence and its presence in a gut-derived cDNA library. All seven of the cDNAs we sequenced appeared to represent the same allele, as we could distinguish no sequence variation between them. The tunicate trypsinogen cDNA contains all of the important sequence features of a trypsinogen and possesses overwhelming similarity to the known trypsinogens. It contains the six absolutely conserved cysteine residues necessary to build the three cystine bridges observed in all vertebrate trypsins (see below). It contains the four key pocket specificity residues: aspartate, glutamine, glycine, and glycine at chymotrypsinogen positions 189, 192, 217, and 227. It contains the three key catalytic residues: histidine, aspartate, and serine, all in the correct sequence contexts and positions. It contains signal and activation peptides (discussed further below). The predicted active trypsin sequence begins with the isoleu-

¹ Nucleic acid sequences described for the first time in this report have accession numbers TRU25747, XLU27330, AF011352, AF011897-901, AF028829, and AA618632-58.

67

	(22)	10	20	(42)	30	40	(58)	50	60	70	80																															
Crayfish	X	I	V	G	G	T	D	A	V	L	G	E	F	P	Y	Q	L	S	F	Q	E	F	S	F	H	C	G	A	S	I	N	E	N	A	I	T	A	G				
Turkate	K	I	V	G	G	E	A	G	A	I	P	Y	Q	A	R	L	Q	S	A	G	S	I	F	Q	K	S	G	S	F	C	G	G	T	I	T	P	N	R	V	L	C	A
Botryllus I	K	I	G	G	S	S	A	N	G	Q	F	S	I	I	F	Q	K	S	G	S	F	C	G	G	T	I	T	P	N	R	V	L	S	A								
Botryllus II	K	I	G	G	S	A	A	N	G	Q	F	S	I	I	F	Q	E	K	S	G	S	F	C	G	G	T	I	I	S	A	N	R	V	L	S	A						
Lamprey A1	H	I	V	G	S	E	C	A	A	H	S	Q	P	W	Q	V	S	L	N	-	I	G	Y	H	F	C	G	G	S	L	I	N	S	Q	W	V	S	A				
Lamprey A2	H	I	V	G	S	E	C	A	A	H	S	Q	P	W	Q	V	S	L	N	-	I	G	Y	H	F	C	G	G	S	L	I	N	S	Q	W	V	S	A				
Lamprey B1	H	I	V	G	G	Y	E	C	A	A	H	S	Q	P	W	Q	V	S	L	N	-	I	G	Y	H	F	C	G	G	S	L	I	S	S	E	W	V	S	A			
Lamprey B2	H	I	V	G	G	Y	E	C	A	A	H	S	Q	P	W	Q	V	S	L	N	-	I	G	Y	H	F	C	G	G	S	L	I	S	S	E	W	V	S	A			
Dogfish	K	I	V	G	G	Y	E	C	P	K	H	A	A	P	T	V	S	L	N	-	V	G	Y	H	F	C	G	G	S	L	I	A	P	G	W	V	S	A				
Pufferfish	K	I	V	G	G	Y	E	C	R	K	N	S	V	A	Y	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	N	E	N	W	V	S	A			
Cod I	K	I	V	G	G	Y	E	C	T	K	H	S	Q	A	H	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	V	S	K	D	W	V	S	A		
Cod X	K	I	V	G	G	Y	E	C	T	R	H	S	Q	A	H	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	V	S	K	D	W	V	S	A		
P. magellanicus	K	I	V	G	G	K	E	C	S	P	S	Q	P	H	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	N	E	N	W	V	S	A				
Salmon I	K	I	V	G	G	Y	E	C	K	A	S	Q	T	H	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	N	E	N	W	V	S	A				
Salmon II	K	I	V	G	G	Y	E	C	K	A	S	Q	T	H	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	N	E	N	W	V	S	A				
Salmon III	K	I	V	G	G	Y	E	C	R	K	N	S	A	S	Y	Q	A	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	S	T	W	V	S	A				
Chicken P1	K	I	V	G	G	Y	E	C	A	R	S	A	A	P	Y	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	S	S	Q	W	V	S	A			
Chicken P29	K	I	V	G	G	Y	T	C	P	E	H	S	V	P	Y	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	N	S	Q	W	V	S	A			
Chicken P38	K	I	V	G	G	Y	S	C	A	R	S	A	A	P	Y	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	S	S	Q	W	V	S	A			
Xenopus I	K	I	V	G	G	T	C	A	K	N	A	V	P	Y	Q	V	S	L	N	-	A	G	Y	H	F	C	G	G	S	L	I	N	S	Q	W	V	S	A				
Xenopus II	K	I	V	G	G	T	C	A	K	S	S	V	P	Y	I	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	T	N	Q	W	V	S	A				
Dog A	K	I	V	G	G	T	C	E	N	S	V	P	Y	Q	V	S	L	N	-	A	G	Y	H	F	C	G	G	S	L	I	S	D	Q	W	V	S	A					
Dog C	K	I	V	G	G	T	C	S	R	N	S	V	P	Y	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	N	S	Q	W	V	S	A				
Rat I	K	I	V	G	G	T	C	P	E	H	S	V	P	Y	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	N	D	Q	W	V	S	A				
Rat II	K	I	V	G	G	T	C	Q	E	N	S	V	P	Y	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	N	D	Q	W	V	S	A				
Rat C	K	I	V	G	G	T	C	Q	K	N	S	L	P	Y	Q	V	S	L	N	-	A	G	Y	H	F	C	G	G	S	L	I	N	S	Q	W	V	S	A				
Rat IV	K	I	V	G	G	T	C	P	K	H	L	V	P	Y	Q	V	S	L	N	-	I	S	H	Q	C	G	G	S	L	I	S	D	Q	W	V	S	A					
Rat V	R	I	V	G	G	T	C	Q	E	H	S	V	P	Y	Q	V	S	L	N	-	A	G	S	H	I	C	G	G	S	L	I	T	D	Q	W	V	S	A				
Bovine A	K	I	V	G	G	T	C	G	A	N	T	V	P	Y	Q	V	S	L	N	-	A	G	Y	H	F	C	G	G	S	L	I	N	S	Q	W	V	S	A				
Bovine C	K	I	V	G	G	T	C	A	N	S	I	P	Y	Q	V	S	L	N	-	S	G	S	H	F	C	G	G	S	L	I	N	S	Q	W	V	S	A					
Pig	K	I	V	G	G	Y	N	C	E	N	S	V	P	Y	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	N	E	Q	W	V	S	A				
Human I (T4)	K	I	V	G	G	Y	N	C	E	N	S	V	P	Y	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	N	E	Q	W	V	S	A				
Human T6	K	I	V	G	G	Y	T	C	E	N	S	V	P	Y	Q	V	S	L	N	-	S	G	S	H	F	C	G	G	S	L	I	S	E	Q	W	V	S	A				
Human II (T8)	K	I	V	G	G	Y	T	C	E	N	S	V	P	Y	Q	V	S	L	N	-	S	G	Y	H	F	C	G	G	S	L	I	S	E	Q	W	V	S	A				
Human III (T9)	K	I	V	G	G	Y	T	C	E	N	S	L	P	Y	Q	V	S	L	N	-	S	G	S	H	F	C	G	G	S	L	I	S	E	Q	W	V	S	A				
Mouse I	K	I	V	G	G	Y	T	C	R	E	S	S	V	P	Y	Q	V	S	L	N	-	A	G	Y	H	F	C	G	G	S	L	I	N	D	Q	W	V	S	A			
	C	V	Y	G	D	S	G	L	Q	I	V	A	G	E	L	D	M	S	V	N	E	G	S	E	Q	T	I	T	V	S	K	I	L	H	E	N	F					
	C	Q	S	A	M	K	I	V	L	G	L	Y	Q	A	S	N	A	D	N	E	A	G	V	T	F	N	V	N	A	Q	T	P	N	S	D							
	C	E	Q	-	-	-	N	L	V	G	L	T	V	T	G	T	A	Y	R	N	S	G	G	T	I	S	V	S	G	K	T	V	H	P	Q							
	C	E	Q	-	-	-	N	L	V	G	L	T	V	T	G	T	A	S	R	S	G	G	V	T	I	S	V	T	G	K	T	V	H	P	Q							
	C	Y	Q	T	A	S	R	I	S	V	R	I	G	E	H	N	I	F	V	N	E	G	T	Q	I	Q	A	S	K	A	I	Q	H	P	Q							
	C	Y	Q	T	A	S	R	I	S	V	R	I	G	E	H	N	I	F	V	N	E	G	T	Q	I	Q	A	S	K	A	I	Q	H	P	Q							
	C	Y	Q	T	A	S	R	I	S	V	R	I	G	E	H	N	I	F	V	T	E	G	T	Q	R	I	Q	A	S	K	A	I	R	H	P	Q						
	C	Y	Q	T	A	S	R	I	S	V	R	I	G	E	H	N	I	F	V	T	E	G	T	Q	R	I	Q	A	S	K	A	I	R	H	P	Q						
	C	Y	Q	-	-	-	R	I	Q	V	R	L	G	E	H	N	I	S	A	N	E	G	D	E	T	I	D	S	S	M	V	I	R	H	P	Q						
	C	Y	K	-	-	-	S	R	V	V	R	L	G	E	H	N	I	R	A	N	E	G	T	Q	F	I	S	S	S	R	V	I	R	H	P	Q						
	C	Y	K	-	-	-	S	V	L	R	V	R	L	G	E	H	N	I	R	N	E	G	T	Q	F	I	S	S	S	S	V	I	R	H	P	Q						
	C	Y	K	-	-	-	S	V	L	R	V	R	L	G	E	H	N	I	R	N	E	G	T	Q	F	I	S	S	S	S	V	I	R	H	P	Q						
	C	Y	K	-	-	-	S	R	V	E	R	M	G	E	H	N	I	R	V	T	E	G	K	Q	F	I	S	S	S	R	V	I	R	H	P	Q						
	C	Y	K	-	-	-	S	R	V	E	R	L	G	E	H	N	I	K	V	T	E	G	S	E	Q	F	I	S	S	S	R	V	I	R	H	P	Q					
	C	Y	Q	-	-	-	S	R	V	E	R	L	G	E	H	N	I	Q	V	T	E	G	S	E	Q	F	I	S	S	S	R	V	I	R	H	P	Q					
	C	Y	K	-	-	-	S	R	I	Q	V	R	L	G	E	H	N	I	A	N	E	G	T	Q	F	I	D	S	V	K	V	I	M	H	P	S						
	C	Y	K	-	-	-	S	S	I	Q	V	R	L	G	E	N	I	A	Q	D	G	S	E	Q	T	I	S	S	S	K	V	I	R	H	S	G						
	C	Y	K	-	-	-	S	S	I	Q	V	R	L	G	E	N	I	D	Q	E	D	S	E	V	R	S	S	S	V	I	I	R	H	P	K							
	C	Y	K	-	-	-	S	S	I	Q	V	R	L	G	E	N	I	A	Q	D	G	S	E	Q	T	I	S	S	S	K	V	I	R	H	S	G						
	C	Y	K	-	-	-	S	R	I	Q	V	R	L	G	E	N	I	A	N	E	G	T	Q	F	I	D	S	V	K	V	I	K	H	P	N							
	C	Y	K	-	-	-	A	S	I	Q	V	R	L	G	E	N	I	A	L	S	E	G	T	Q	F	I	S	S	S	K	V	I	R	H	S	G						
	C	Y	K	-	-	-	S	R	I	Q	V	R	L	G	E	N	I	D	V	L	E	G	N	E																		

Crayfish	DYDLN	ISLLKLSGSLTFNNNVAPIALPAQGHATGNIIVTGWGTTT - EGGNTPDVLQKVTVPLVSDAECDYDGEI	150
Turkate	DSATTDN	VMLRLDESATLTSSVALSLPTSPFEEDTACTVSGWGTTS - SGOTISDYLKMKVEVNVVVDQDECGRNRYGSLT	151
Botryllus I	NSNTIQN	IMILNLASSPSYSTIAAAPLASSPSVGTESPLPDCAIPASGIVSNLQYVNVVISTDCNSTRYNGAI	152
Botryllus II	NSNTIQN	IMILNLGSSFSLGSTIAAAPLASSPSVGTESPLPDCAIPASGIVSNLQYVNVVISTDCNSTRYNGAV	153
Lamprey A1	NSWTIDN	IMLIKSSPATLNQYAAIAIPSSCVNTGVMCTISGWGETQ - TSVGSPDVLQVAPVLSDTSCRNSTYPGDI	154
Lamprey A2	NSWTIDN	IMLIKSSPATLNQYAAIAIPSSCVNTGVMCTISGWGETQ - TSVGSPDVLQVAPVLSDTSCRNSTYPGDI	155
Lamprey B1	SSATIDN	IMLIKSSPATLNQYAAIAIPSSCVNTGVMCTISGWGETQ - TSVGSPDVLQVAPVLSDTSCRNSTYPGDI	156
Lamprey B2	NSATIDN	IMLIKSSPATLNQYAAIAIPSSCVNTGVMCTISGWGETQ - TSVGSPDVLQVAPVLSDTSCRNSTYPGDI	157
Dogfish	SGYDLN	IMLIKSKPAALNRNVDLISLPTGCAYAGEMCLISGWGNTM - DGAVSGDQLQCLDAPVLSDAECKGAYPGMI	158
Pufferfish	SSYNIDN	IMLIKSKPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSTADRNKLQCLNIPILSDRDCENSYPGMI	159
Cod I	SSYNINN	IMLIKLTKPATLNQYVHVALPTECAADATMCTVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	160
Cod X	SSYNIDN	IMLIKLTKPATLNQYVHVALPTECAADATMCTVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	161
<i>P. magellanicus</i>	SSYNIDN	IMLIKSKPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	162
Salmon I	SSYNIDN	IMLIKSKPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	163
Salmon II	SSYNIDN	IMLIKSKPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	164
Salmon III	NSRNLN	IMLIKSKPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	165
Chicken P1	NANTLNN	IMLIKSKAATLNSYVNTVPLPTSCVTAAGTCLISGWGNTLSSGSLYPDVLQCLNAPVLSQSSAYPGRI	166
Chicken P29	SSITLNN	IMLIKSKAATLNSYVNTVPLPTSCVTAAGTCLISGWGNTLSSGSLYPDVLQCLNAPVLSQSSAYPGRI	167
Chicken P38	NANTLNN	IMLIKSKAATLNSYVNTVPLPTSCVTAAGTCLISGWGNTLSSGSLYPDVLQCLNAPVLSQSSAYPGRI	168
Xenopus I	NSRNLN	IMLIKSTTARLSANIQSVPLPSACASAGTCLISGWGNTLSSGSLYPDVLQCLNAPVLSQSSAYPGRI	169
Xenopus II	NSYTLN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	170
Dog A	NSWIDLN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	171
Dog C	NANTIDN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	172
Rat I	SSWTLNN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	173
Rat II	DRKTLNN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	174
Rat C	NANTFDN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	175
Rat IV	NKDTLDN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	176
Rat V	DKWTVDN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	177
Bovine A	SSWTLNN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	178
Bovine C	SSWTLNN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	179
Pig	NGNTLDN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	180
Human I (T4)	DRKTLNN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	181
Human T6	NRITLNN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	182
Human II (T8)	NRITLNN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	183
Human III (T9)	NRITLNN	IMLIKSSPATLNQYVQVVALPSSCAAAGTMCVSGWNTM - SSVADGDKLQCLSLPILSHADCANSTYPGMI	184
Mouse I	NSWTLN	IMLIKSPVTLNARVASVPLPSSCAPAGTQCLISGWGNTLSSGSLYPDVLQCLNAPVLSQSSAYPGRI	185

Fig. 1. Continued

SEGTEQWQASKAIRHPGYVSSSTIDNDIMLVKLAKPATLN
 AYAQPVALPTACATVGMCTISGWGETETPIGSFDVLMCV
 EAPVLSVADCESGYPGDITDNMVCLGYMEGG

Fig. 2. Translation of a trypsin-like PCR product from the lamprey *Petromyzon marinus*.

levels. It is likely that this probe, as an uncloned PCR product, was heterogeneous, but with one species dominant enough to produce a clear signal when sequenced. This suggests that extreme care must be taken when evaluating PCR sequences derived from multigene families, particularly for phylogenetic purposes (see also Cilia et al. 1996).

We sequenced 31 cDNA plaques positive for this probe, a few of which were false positives. Of these 31 plaques, 17 were trypsinogen, four were chymotrypsinogen, one was similar to the chitotriosidase precursor cDNA, one was similar to the oligosaccharyl transferase STT3 subunit, and the other eight cDNAs were not positively identified. Of the 17 trypsinogen cDNAs, seven were completely sequenced by primer walking. Based on contig assemblies, the 17 cDNAs fell into at least five clusters, indicating the presence of at least five different expressed lamprey trypsinogen isozyme genes (or alleles). These sequences were identified as trypsinogen based on the same criteria as the tunicate trypsinogen (see above).

We designated the five lamprey trypsinogen clusters: A1, with nine cDNAs (two completely sequenced); A2, with three cDNAs (one completely sequenced); A3, with one completely sequenced cDNA; B1, with three cDNAs (two completely sequenced); and B2, with one completely sequenced cDNA. The untranslated 3' tails of the three A cluster trypsinogens are 92.0–96.0% identical. The untranslated 3' tails of B1 and B2 are 85.9% identical. The A and B cluster tails could not be aligned with each other. The coding regions of the A cluster sequences are 98.9–99.7% identical. The coding regions of B1 and B2 are 97.7% identical. The identity between the coding sequences of the A and B clusters is 92.5–93.2%. Because of the possibility of confounding sequencing errors with allelic or isozymic variation, we cannot rule out the possibility of more than five clusters. We are confident of these five clusters however, as we estimate our error rate to be less than 1 in 5,000. Additionally, we specifically confirmed discrepancies between clusters.

Different lamprey trypsinogen genes (or alleles) can clearly be over 99% identical similar across regions longer than a single sequence read. This high similarity of multiple trypsinogen genes is observed in other species (Wang et al. 1995). The lamprey trypsinogen genes are probably coded for by highly similar tandem repeats, as is the case in humans, mice, and chickens. Two of the lamprey B cluster trypsinogens have been observed to be linked in tandem following double barrel shotgun sequencing of a genomic cosmid clone (data not shown).

Double barrel shotgun sequencing is a low-redundancy methodology for establishing ordered and oriented sequence contigs across a large target sequence (Roach et al. 1995). Genomic Southern blots suggest a large number of lamprey trypsinogen genes (data not shown).

Insertions and Deletions

The serine protease gene family possesses a number of features that make it particularly amenable to multiple sequence alignment. These include a cleavage site following an activation peptide, six cysteine residues necessary to form the three absolutely conserved cystine bridges, four active site pocket specificity residues embedded in conserved sequences, three catalytic residues embedded in conserved sequences, and several other highly conserved sequences. These conserved sequences are spread throughout the length of the protein, allowing members of the gene family to be easily aligned, as regions of low similarity are inevitably flanked by conserved residues. Variations in sequence length between two conserved residues can be recognized as insertions or deletions. Evolutionarily conserved insertions and deletions are expected to be rare events and thus serve as good markers for tracking gene family phylogenies over large time scales.

The tunicate shares a single residue insertion at position 21 in common with rat trypsin IV (X15679). This event is most likely a coincidence, as it is found in no other vertebrates. A second tunicate insertion occurs near residues 45–51. This coincides precisely with the boundary between exons 2 and 3. The crayfish shares this insertion (Titani et al. 1983). Of the vertebrates, only the lampreys have an insertion at this site, but the lamprey insertion consists of only two residues. This is consistent with a progressive loss of residues at this site during early vertebrate evolution, with five lost prior to the Agnathan divergence and another two lost prior to the elasmobranch divergence. Molecular models predict that these residues are extensions to a surface loop (Michael Levitt, personal communication). They are thus unlikely to have much functional significance, other than possibly a role in determining substrate specificity. A third tunicate insertion of five residues occurs near positions 115–119. The tunicate also has a deletion of three residues around position 223. Neither of these events is observed in any other trypsin, consistent with the 600–700-million-year period of independent evolution since the urochordate/vertebrate split.

The tunicate, lamprey, dogfish, and all but one of the Osteichthyes trypsins lack a residue at position 130 that is found in all other vertebrate trypsinogens. A residue is present at this position in salmon trypsin III. Therefore, most likely, there were at least two trypsinogen isozymes present in the common Osteichthyes/tetrapod ancestor, one of which gained a residue at position 130. Both

Table 1. Cystine bridges of the chordate trypsinogens: identification of the bridges uses the bovine chymotrypsinogen numbering system (Zwillig and Neurath 1981)

	Cystine bridges ^a					
	22-157	42-58	127-232	136-201	168-182	191-220
Bacteria		X ^b			X ^b	X ^b
Crayfish		X			X	X
Tunicate		X		X	X	X
Lamprey	X	X	X	X	X	X
Nonhuman Gnathostomata	X ^b	X ^b	X ^b	X ^b	X ^b	X ^b
"Cationic" Human (cT3)	X ^c	X	X	X	X	X
"Anionic" Human (except T8)	X	X		X	X	X
"Anionic" Human (T8)	X	X			X	X

^a A cross (X) indicates the predicted presence of a bridge^b Experimentally determined^c Pseudogene missing one of two cysteine codons

variations were maintained by the Osteichthyes, but the insertion became the exclusive variant for the tetrapods, perhaps due to coincidental evolution and/or gene copy number contraction and expansion (Hood et al. 1975). Note that the rat trypsin V lacks a residue near position 130. This most likely represents an independent deletion event, especially considering that this gene appears to have undergone rapid evolution in recent times (see below). The codon for this residue lies directly on the intron/exon boundary between exons 2 and 3, so insertions and deletions here are likely to be much more frequent than elsewhere.

Cystine Bridges

Acquisition and loss of cystine bridges is a rare evolutionary event and thus a useful phylogenetic marker. A backbone phylogeny of the serine proteases can be developed by considering the parsimonious addition and loss of cystine bridges (De Haën et al. 1975). The assignment of cystine bridges can be made from considerations of the homology of cysteine residues. There are six cystine bridges in most vertebrate trypsinogens. Of these, three are absolutely conserved in all serine proteases. The bacterial and crayfish trypsins lack all three of the "optional" vertebrate bridges (Titani et al., 1983; Kim et al. 1991). The tunicate trypsin gains one bridge (between residues 136 and 201; Table 1). The lamprey trypsin gains another two bridges (22/157 and 127/232), to reach the vertebrate standard. Curiously, all human trypsins have lost the 127/232 bridge. Furthermore, human trypsin II has also lost the 136/201 bridge. Thus, a progressive increase of cystine bridges is seen during the course of vertebrate trypsin evolution. The human lineage shows a subsequent decrease.

The human T3 trypsinogen pseudogene 5' to the TCR β locus possesses 11 of the 12 cysteine residues necessary to make the six cystine bridges labeled in Table 1. Thus, the functional precursor to this pseudogene had all six bridges, in contrast to all functional

human trypsins. The loss of the cystine bridges from the human trypsins is therefore recent, as it occurred after the mammalian divergence, and in only one of the two major branches of the trypsinogen phylogeny.

Intron/Exon Boundaries

Intron/exon boundaries, where known, are shown in the multiple alignments as vertical lines (Fig. 1 and Table 2). Some genomic sequences for the trypsinogens are available from human, mouse, chicken, and lamprey. Additionally, intron/exon boundaries are available for some other serine proteases. Of note is the absolutely conserved location of the boundary between exons 4 and 5 which occurs near the active site serine. The 1/2 and 3/4 exon boundaries are also highly conserved. Shifting of intron/exon boundaries does not appear to have been a major mode of evolution for the trypsinogens, or even serine proteases in general. The possible exception is the 2/3 exon boundary, which occurs immediately adjacent to a position of inserted residues in lampreys and tunicates (see above).

Signal Sequences

An alignment of chordate trypsinogen signal sequences and activation peptides is shown in Table 2. These sequences are highly conserved and conform to the general rules for eukaryotic signal sequences (von Heijne 1986). Most vertebrate trypsinogen signal sequences are 15 residues in length. Notably, the tunicate and some of the Osteichthyes signals are two to three residues shorter. One of the chicken signals is 16 residues long.

Activation Peptides

The activation peptides of the chordate trypsinogens are shown in Table 2. The key feature of trypsinogen activation peptides is a cluster of at least three anionic resi-

Table 2. Signal peptides and activation peptides of the chordate trypsinogens*

	Signal	Activation	Mature
Human I (T4)	MNPLLILTFVAAA ILA	APFDDDDK	IVGG
Human T6	MNPLLILAFVGAA IVA	VPFDDDDK	IVGG
Human II (T8)	MNLLLILTFVAAA IVA	APFDDDDK	IVGG
Human III (T9)	MNPFLILAFVGAA IVA	VPFDDDDK	IVGG
Human 5' α (T3)	HEDLHLPALLGAA IAT	FPTDDDDK	IVGG
Bovine A	MHPLLILAFVGAAAA	FPSDDDDK	IVGG
Bovine C	-----	--VDDDDK	IVGG
Dog A	MNPLLILAFVGAAVA *	TPTDDDDK	IVGG
Dog C	MLTFIFLALLGATVA *	FPIDDDDK	IVGG
Mouse I	MSALLILALVGAAVA	FPVDDDDK	IVGG
Pig	-----	FPTDDDDK	IVGG
Rat I	MSALLILALVGAAVA	FPLEDDDK	IVGG
Rat II	MRALLILALVGAAVA	FPVDDDDK	IVGG
Rat C	MLALIFLAFVGAAVA	LPLDDDDK	IVGG
Rat IV	MLISIFFAFLGAAVA	LPVNDKK	IVGG
Rat V	MKICIFFTLLGTVA	FPTEDNDDK	IVGG
Chicken P1	MLFLVLVAFVGV TIVA	FPISEDDDK	IVGG
Chicken P29	MLFLFLILSCLGAA IVA	FPGVDDDK	IVGG
Xenopus I	MLFLLLCVLLGAAVA	FDDDK	IVGG
Xenopus II	MKFLVILVLLGAAVA	FEDDDK	IVGG
Cod	MLS LIFVLLGAV	FAEEDK	IVGG
P. magellanicus	MRS L VFVLLIGAA	FATEEDK	IVGG
Pufferfish	-----LIAIAA	YAAPIDEDDK	IVGG
Salmon I	MISLVFVLLIGAA	FATEDDK	IVGG
Salmon III	-----FAVA	FAAPIDDEDK	IVGG
Dogfish	-----	APDDDDK	IVGG
Lamprey A1	MHGLILALLVGVA	APYMYEDH	IVGG
Lamprey A2	MHGLILALLVGVA	APYMYEDH	IVGG
Lamprey B1	---LIFALLVGTIAAA	APYMYEDH	IVGG
Lamprey B2	--GLIFALLVGTIAAA	APYMYEDH	IVGG
Tunicate	MKIVILLLLGLA	AVNADK	IVGG

* The signal peptidase cleavage site is the predicted site (von Heijne 1986); only in the case of the canine (marked with *), have the sites been determined experimentally (Carne and Scheele 1982). A dashed line (---) represents undetermined sequences. Intron/exon boundaries, where known, are indicated by a bar (|).

duces preceding a lysine or arginine. However, the lamprey activation peptide has only two penultimate anionic residues, while the tunicate has just one (but complemented by an asparagine preceding an alanine). Many of the Osteichthyes trypsinogens and one of the *Xenopus* trypsinogens have three anionic residues, while the higher vertebrates tend to have four or more such residues, suggesting a progressive increase in selective pressure for such residues during the course of vertebrate evolution.

Strikingly, none of the activation peptides for the lamprey trypsinogens end in a lysine or arginine residue. All lamprey trypsinogen activation peptides end in a histidine. Thus it seems unlikely that lamprey trypsin is capable of autocatalyzing its own activation, as trypsin is not capable of cleaving after a histidine residue. This suggests that lampreys rely exclusively on enterokinase for trypsinogen activation. The life cycle of the lamprey may explain selective pressure for greater control of digestive enzyme activation. For example, adult lampreys will go months to years without eating. The lamprey chymotrypsinogen activation peptide ends in an arginine and so could be activated by trypsin. This would allow

lamprey enterokinase to be a master control switch for digestion, allowing for little or no basal digestive enzyme activation.

Sequence Distances and Derived Phylogenies

We have generated phylogenies for trypsin using a number of different methodologies. All of these methodologies give results that are qualitatively and quantitatively very similar. Because we do not have complete trypsinogen sequences in all cases, we have limited our formal analysis to multiple alignments of portions of the sequences coding for the mature trypsin peptide. Adding the signal and activation peptide data to our phylogenies does not significantly affect the results (data not shown). We obtain similar phylogenies using either nucleotide or amino acid sequence data. Note that we do not have nucleotide sequence data for all of the trypsinogens, as several were determined by protein sequencing. We present a trypsin phylogeny in Fig. 3.

This phylogeny is striking in two major respects. First, it fails to support a molecular clock hypothesis.



Fig. 3. A phylogeny of the vertebrate trypsins derived from distance matrix data.

This is most striking for several of the rodent trypsins (Rat IV, Rat V, and Rat C). Mice possess nearly identical homologs to all the known rat trypsinogens, indicating that these rate variations occurred before the mouse/rat divergence (Lee Rowen, personal communication). Therefore, since the mammalian radiation, rates of evolution have differed by as much as an order of magnitude between different isozyme loci within a species. This is consistent with "bursts of sudden evolution" at particular loci, perhaps due to gene conversion events. Molecular traces of such events, if present, have been largely obliterated by subsequent mutation. The presence of both coincidental evolution and vastly differing evolutionary rates therefore precludes the construction of any trypsin phylogeny with high confidence.

Secondly, the phylogeny in Fig. 3 fails to reproduce the topology of trypsinogen evolution. Note that if only one representative of each vertebrate class is considered, the resulting phylogeny is more consistent with a star than a tree. A likely explanation for this is that these trypsins have reached an equilibrium distance from each other, where converging mutations are as common as diverging mutations. Chaotic fluctuations away from equilibrium would be expected in this case, producing deviations from a perfect star phylogeny, notably small topological deviations in early branchings. Within vertebrate classes, our phylogeny successfully reproduces known phylogenetic relationships, suggesting that mo-

lecular trypsin data may have some utility for analysis of recent evolutionary events. A further topological deviation results from the coincidental evolution of the two groups of trypsinogen genes within the class Mammalia (see Discussion). It is likely that similar deviations will also be noted in other classes as more sequences are obtained.

Because of the impossibility of building a trypsin phylogeny with high confidence, we sought another informative method to display trypsin sequence distance data. To this end, we represented the sequences as points on a two-dimensional plot, minimizing the least-squares error in the plotted distances with respect to the calculated distances (Michael Levitt, personal communication). This plot is free of assumptions about phylogenetic relationships (Fig. 4). For this plot only, we calculated distance as the number of amino acid residue differences between two sequences. This metric emphasizes structural distance over evolutionary distance.

We feel that this diagram provides a better picture of trypsinogen evolution than a phylogenetic tree does, largely because it frees the observer of dependency upon incorrect underlying assumptions. In particular, we noted that the anionic and cationic trypsins group together, suggesting that they fall naturally into structurally related clusters. Coincidental selective constraints such as gene conversion may tend to keep these functional forms clustered.

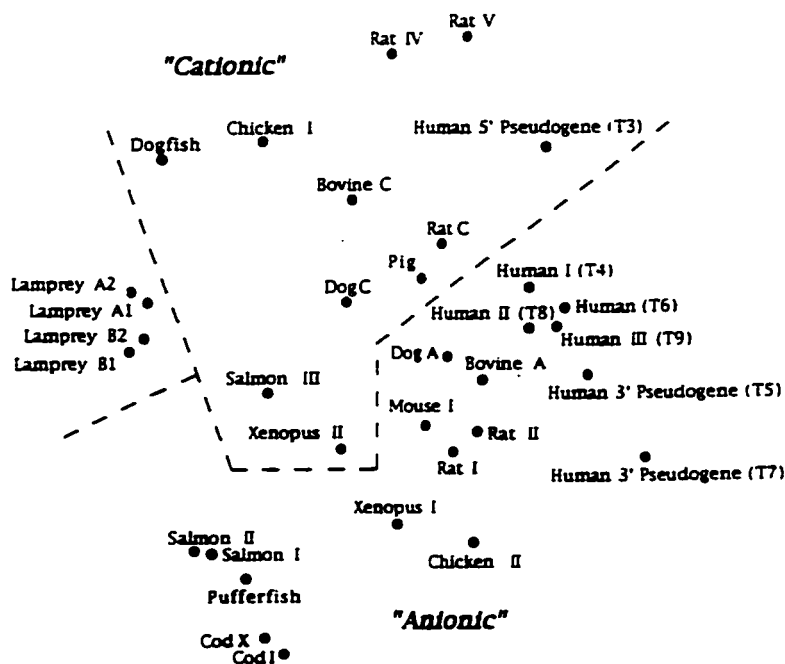


Fig. 4. A least-squares global minimization of protein-protein primary structural distances for the vertebrate trypsins. The arbitrary dashed line is intended to suggest a possible natural grouping of sequences.

Discussion

Cationic and Anionic Trypsins

It has been known for some time that vertebrate trypsinogens occur in at least two different isoforms, termed "cationic" and "anionic" (discussed by Le Huerou et al. 1990). Most species appear to possess one or more representatives of each of these isoforms. However, this is not universally the case. For example, humans possess no functional "cationic" trypsins. Additionally, it is unlikely that the lamprey or dogfish trypsins belong to either of these groups. No difference in function has been demonstrated between the "cationic" and "anionic" trypsinogens, although a possible difference in substrate specificity has been proposed (Fletcher et al. 1987).

A table of predicted and experimental isoelectric points for the chordate trypsins is presented in Table 3. Several discrepancies between this biochemical data and phylogenetic expectations can be noted. Rat trypsins IV and V, which phylogenetically should be "cationic," have predicted anionic charges. Human trypsin I, which phylogenetically should be "anionic," has a cationic charge. We are thus unsure of the utility of designating trypsins as "cationic" or "anionic" and suggest a re-evaluation of the nomenclature. It may be that both isoforms have a functional niche, and higher vertebrates require both. In this scenario, the biochemically cationic but phylogenetically "anionic" human trypsin I fills a crossover role by filling a niche vacated by the missing phylogenetically "cationic" trypsinogens. Also since, rat trypsin C is cationic, in this scenario there would be

little selective pressure on rat trypsins IV and V to remain cationic, and their charge could have "drifted."

However, we find the above scenario unlikely. As far as we were able to ascertain, lampreys and tunicates possess only biochemically anionic trypsins, suggesting that a vertebrate has no absolute need for trypsins of two different charges. Furthermore, there are no residues at specific sites that are characteristic of either the "cationic" or "anionic" trypsins. Rather, the net charge of trypsin is governed by highly variable surface residues. Consistent with this, the predicted isoelectric points of the vertebrate trypsins do not fall into two groups but span the pI spectrum continuously, from 4.4 to 8.3.

If there were two groups of trypsinogen isozymes at separate genomic locations, each coded for by tandem repeats, there might be significant genetic exchange or gene conversion within a group, but not between groups. Coupled with an underlying mutation rate providing diversity within members of a group, this could account for the maintenance of general biochemical features within a group in the absence of absolutely conserved characteristic residues. The evolutionary remnants of phylogenetically "cationic" trypsins are present in humans as repeated elements at a distinct location from the tandemly repeated "anionic" trypsinogens (Rowen et al. 1996). Chickens and mice also maintain a genomic separation of their "anionic" and "cationic" trypsinogens (Kai Wang, unpublished observation; Lee Rowen, personal communication), so this last scenario is possible. This mechanism can also account for similarities in signal sequences between members of a phylogenetic group (Table 2). Nevertheless, we cannot rule out the possibil-

Table 3. Predicted isoelectric points and charges of the chordate trypsins^a

	Group	Isoelectric point	Charge at pH 7.0
Tunicate		3.9	-13.18
Lamprey A I		5.2	-5.62
Lamprey B I		5.8	-3.62
Dogfish	II (?)	4.9	-10.11
Cod I	I	6.8 (6.6)	-0.79
Cod X	I	5.8 (5.5)	-5.95
<i>P. magellanica</i>	I	5.8	-5.28
Pufferfish	I	6.2	-2.61
Salmon I	I	5.9	-3.62
Salmon II	I	5.5	-4.62
Salmon III	II	8.1	-4.21
Chicken P1	II	8.2	5.03
Chicken P29	I	4.6	-9.78
Xenopus I	I	6.7	-0.79
Xenopus II	II	7.7	2.21
Dog A	I	4.9	-5.94
Dog C	II	8.3	6.04
Pig	II	7.9 (10.8)	3.21
Bovine A	I	4.8	-7.62
Bovine C	II	8.3 (10.1)	6.03
Human T4 (I)	I	7.5	1.28
Human T6	I	6.9	-0.39
Human T3 (II)	I	5.0	-6.65
Human T9 (III)	I	6.8	-0.55
Mouse I	I	4.4	-9.75
Rat I	I	4.9 (4.4)	-6.62
Rat II	I	4.8 (4.3)	-6.78
Rat C	II	8.1 (8)	4.20
Rat IV	II	6.9 (6.2)	-0.29
Rat V	II	5.1	-9.11

^a Each Osteichthyes or tetrapod trypsin is assigned to either group I or II based on phylogenetic considerations (see text). Experimental values (in parentheses) for isoelectric points are from Walsh (1970), Lütcke *et al.* (1989), and Aageirsson *et al.* (1989).

ity of functional differences between trypsinogen groups or of a small selective advantage of having multiple trypsins of different isoelectric points. Also, the role and importance of calcium binding in trypsin function remain unclear (Le Huerou *et al.* 1990).

Modes of Trypsinogen Evolution

It is clear that trypsinogen does not evolve as a classical single locus gene with a constant rate. Vertebrate trypsinogen evolution has been dynamic and multimodal. The most significant event of this evolution was the gross duplication of a whole or part of the trypsinogen locus shortly after the elasmobranch divergence, or close to that time. This gave rise to two groups of trypsinogen genes that were maintained in all Osteichthyes and tetrapods. We designate trypsinogens phylogenetically resembling classically "anionic" trypsinogens as group I and those resembling "cationic" trypsinogens as group II. Following the divergence of the two groups of trypsinogens, they tended not to exchange genetic informa-

tion with each other and acted largely as separately evolving gene families. Thus a particular group I trypsinogen is likely to be more similar to a group I trypsinogen from another species than to a group II trypsinogen from the same species.

The class Mammalia provides an exception to this general rule. All of the mammalian trypsinogens are more closely related to each other than to trypsinogens of other classes (Figs. 3 and 4). This is a clear result of coincidental evolution (Hood *et al.* 1975). Similar selective pressures on nucleotide and codon usage and on protein structure and function within the mammalian environment will cause sequence distances to converge. Additionally, transfer of genetic information by gene conversion or recombination may be an infrequent event that over the time scale of vertebrate class evolution may also tend to converge sequence distances. It is likely that these effects will be more readily observed in other vertebrate classes as more sequences are obtained.

Within a group, several modalities of evolution are likely to operate. Since trypsinogen gene loci tend to have repeat structures, one can expect both expansion and contraction of these repeats. The repeat structure of the loci provides for the maintenance of a large pseudogene reservoir. For example, there are five known human trypsinogen pseudogenes (Rowen *et al.* 1996). These pseudogenes can evolve rapidly, free of evolutionary constraint, as digestive function is provided by the many functional members of a locus. It is conceivable that rarely these pseudogenes will back mutate to functionality, perhaps jump started by a gene conversion event. Perhaps more importantly, they also provide a genetic reservoir of material for recombination events with functional trypsinogens. The proximity of trypsinogen to the TCR β locus in chickens has been noted (Wang *et al.* 1995). In chickens, gene conversion of a functional TCR β gene with a V pseudogene segment is common. One may thus hypothesize that periodically during vertebrate evolution, the immune receptor loci may have been particularly susceptible to recombinogenic events. This susceptibility might extend to the trypsinogen loci. For example, during T-cell development, TCR β locus episomes will contain several copies of group I trypsinogen genes. If such an episome somehow formed in or recombined with germline DNA, there could be a striking and sudden change in, transposition of, or even novel generation of a trypsinogen locus.

These kinds of gross and evolutionarily sudden events can cause the apparent rate of evolution of trypsinogen to vary highly between species and even between loci within a species. Such events are complemented by a background of intron/exon boundary shifting, as well as nucleotide transitions, transversions, insertions, and deletions. Taken together, these modes of evolution will significantly confound efforts to build reliable trypsinogen phylogenies over large time scales, such as those

separating classes. Phylogenies spanning more than one phylum are particularly problematic, and for these reasons we have excluded nonvertebrate trypsinogens from most of our analyses.

Determination of species relationships from phylogenies based on multigene family data is difficult (Cilia et al. 1996; Hollingshead et al. 1994). During the evolution of Animalia, trypsinogen genes are likely to have been consistently present in genomes as one or more multigene loci. For example, several insect species are known to have multiple trypsinogens (Davis et al. 1985). The principles of population genetics may well be better suited for studying such multigene families than traditional molecular evolution models (e.g., Dickerson 1971). We recommend great care in applying traditional analyses to the evolution of genes not definitely known to be single-copy.

Function of Trypsin

The only known function of trypsin involves the digestion of food, either directly, or by the activation of zymogens. The presence of multiple trypsin isozymes within an organism raises the possibility that they may perform different functions. The tight linkage of the trypsins to the TCR β locus suggests an immunological role. Their deletion from a functional TCR β locus seems inconsistent with a role for trypsinogen in mature T cells but does not rule out other immunological functions. Many serine proteases are known to play roles in immune defense (see, for example, Müller et al. 1994), so such a role is not out of the question for one of the trypsinogen isozymes. However, in the absence of concrete evidence to the contrary, we feel that the null hypothesis for trypsinogen is that it performs no function other than alimentary digestion. Large quantities of trypsin are needed on short notice for digestion, and one facet of gene regulation for such a protein could involve high gene dosage. An alimentary selective pressure for high gene dosage may be sufficient to explain the maintenance of multiple trypsinogen genes in a genome.

It remains possible that group II trypsins play a non-alimentary role in vertebrates other than humans. This putative role would either not be necessary in humans or, more likely, be subsumed by a novel serine protease.

To summarize, the vertebrate trypsinogens provide a fascinating example of a dynamic multigene family. The interest and value of studying multigene family evolution is likely to grow as more sequences become available. Studying such families will reveal much about the dynamics and mechanisms of evolution.

Acknowledgments. J.C.R. is supported by a grant from the Life & Health Insurance Medical Research Fund. Discussions with Chris Bystroff, Joe Felsenstein, Aaron Halpern, and Ed Thayer were instrumental in formulating approaches to, and recognizing the limitations of, the trypsinogen phylogeny. Michael Levitt provided novel insight into the functional and structural relationships of the trypsinogens.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
- Åsgerisson B, Fox JW, Bjarnason JB (1989) Purification and characterization of trypsin from the poikilotherm *Gadus morhua*. *Eur J Biochem* 180:85–94
- Came T, Scheele G (1982) Amino acid sequences of transport peptides associated with canine exocrine pancreatic proteins. *J Biol Chem* 257(8):4133–4140
- Cilia V, Lafay B, Christen R (1996) Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. *Mol Biol Evol* 13(3):451–461
- Davis CA, Riddell DC, Higgins MJ, Holden JA, White BN (1985) A gene family in *Drosophila melanogaster* for trypsin-like enzymes. *Nucleic Acids Res* 13(18):6605–6619
- De Haen C, Neurath H, Teller DC (1975) The phylogeny of trypsin-related serine proteases and their zymogens. New methods for the investigation of distant evolutionary relationships. *J Mol Biol* 92: 225–259
- Dickerson RE (1971) The structure of cytochrome c and the rates of molecular evolution. *J Mol Evol* 1:26–45
- Felsenstein J (1993) PHYLIP (phylogeny inference package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle
- Fletcher TS, Alhadeff M, Craik CS, Langman C (1987) Isolation and characterization of a cDNA encoding rat cationic trypsinogen. *Biochemistry* 26:3081–3086
- Genicot S, Renner-Delrue F, Edwards D, VanBeeumen J, Gerday C (1996) Trypsin and trypsinogen from an Antarctic fish: molecular basis of cold adaptation. *Biochimica et Biophysica Acta* 1298(1): 45–57
- Gish W, States DJ (1993) Identification of protein coding regions by database similarity search. *Nat Genet* 3:266–272
- Gudmundsdottir A, Gudmundsdottir E, Oskarsson S, Bjarnason JB, Eakun AK, Craik CS (1993) Isolation and characterization of cDNAs from Atlantic cod encoding two different forms of Trypsinogen. *Eur J Biochem* 217:1091–1097
- Hollingshead SK, Arnold J, Readdy TL, Bessen DE (1994) Molecular evolution of a multigene family in group A Streptococci. *Mol Biol Evol* 11(2):208–219
- Hood L, Campbell JH, Elgin SC (1975) The organization, expression, and evolution of antibody genes and other multigene families. *Annu Rev Genet* 9:305–353
- Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) *Mammalian protein metabolism*. Academic Press, New York, pp 21–132
- Kang J, Wiegand U, Mueller-Hill B (1992) Identification of cDNAs encoding two novel rat pancreatic serine proteases. *Gene* 110:181–187
- Kim JC, Cha SH, Jeong ST, Oh SK, Byun SM (1991) Molecular cloning and nucleotide sequence of *Sireptomycetes griseus* trypsin gene. *Biochem Biophys Res Commun* 181:707–713
- Kossiakoff AA, Chambers JL, Kay LM, Stroud RM (1977) Structure of bovine trypsinogen at 1.9 Å resolution. *Biochemistry* 16(4):654–664
- Le Huou L, Wicker C, Guilloteau P, Toullec R, Puigserver A (1990) Isolation and nucleotide sequence of cDNA clone for bovine pancreatic anionic trypsinogen. *Eur J Biochem* 193:767–773
- Lütcke H, Rausch U, Vasiloudes P, Scheele GA, Kern HF (1989) A fourth trypsinogen (P23) in the rat pancreas induced by CCK. *Nucleic Acids Res* 17(16):6736
- Müller WEG, Pancer Z, Rinkevich B (1994) Molecular cloning and localization of a novel serine protease from the colonial tunicate *Botryllus schlosseri*. *Mol Marine Biol Biotechnol* 3(2):70–77
- Pancer Z, Leuck J, Rinkerich B, Steffen R, Mueller I, Mueller WEG (1996) Molecular cloning and sequencing analysis of two cDNAs

- coding for putative anionic trypsinogens from the colonial urochordate *Botryllus Schlosseri* (Ascidacea). *Molecular Marine Biology and Biotechnology* 5(4):326-333
- Roach J, Boysen C, Wang K, Hood L (1995) Pairwise end-sequencing: a unified approach to mapping and sequencing. *Genomics* 26:345-356
- Rowen L, Koop BF, Hood L (1996) The complete 685-kilobase DNA sequence of the human β T cell receptor locus. *Science* 272:1755-1762
- Rypniewski WR, Perrakis A, Vorgias CE, Wilson KS (1994) Evolutionary divergence and conservation of trypsin. *Protein Eng* 7(1): 57-64
- Sweet RM, Wright HT, Janin J, Chothia CH, Blow DM (1974) Crystal structure of the complex of porcine trypsin with soybean trypsin inhibitor (Kunitz) at 2.6-Å resolution. *Biochemistry* 13(20):4212-4228
- Titani K, Ericsson LH, Neurath H, Walsh KA (1975) Amino acid sequence of dogfish trypsinogen. *Biochemistry* 14(7):1358-1366
- Titani K, Sasagawa T, Woodbury RG, Ericsson LH, Dörsam H, Kraemer M, Neurath H, Zwillig R (1983) Amino acid sequence of crayfish (*Astacus fluvianilis*) trypsin I. *Biochemistry* 22:1459-1465
- von Heijne G (1986) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res* 14(11):4683-4690
- Walsh KA (1970) Trypsinogens and trypsins of various species. *Methods Enzymol* 19:41-63
- Walsh KA, Wilcox PE (1970) Serine proteases. *Methods Enzymol* 19:31-41
- Wang K, Gan L, Lee I, Hood L (1995) Isolation and characterization of the chicken trypsinogen gene family. *Biochem J* 307:471-479
- Wiegand U, Corbach S, Minn A, Kang J, Muller-Hall B (1993) Cloning of the cDNA encoding brain trypsinogen and characterization of its product. *Gene* 136:167-175
- Worley KC, Wiese BA, Smith RF (1995) BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res* 5:173-184
- Zwillig R, Neurath H (1981) Invertebrate proteases. *Methods in Enzymology* 80:633-664

VITA

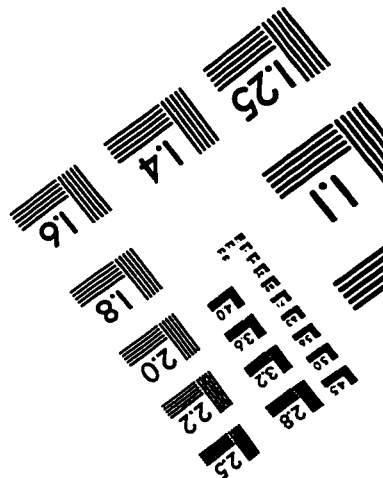
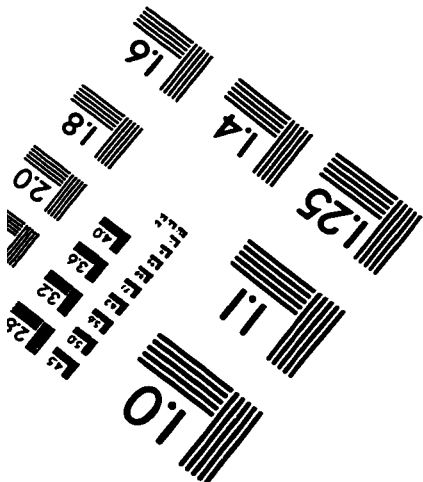
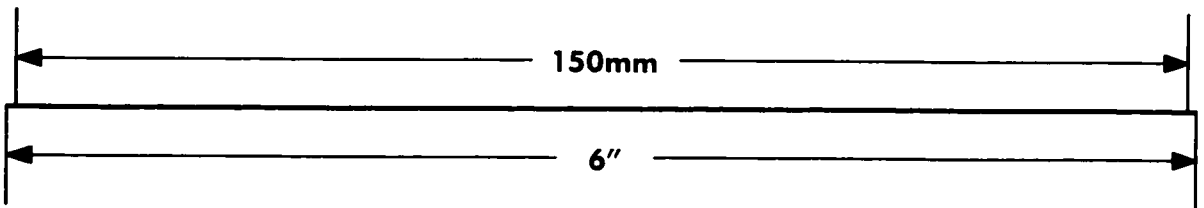
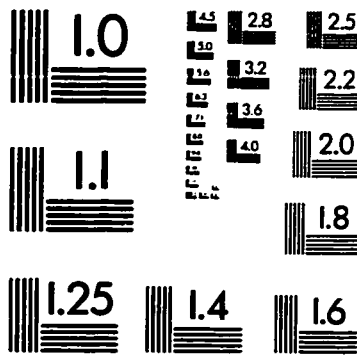
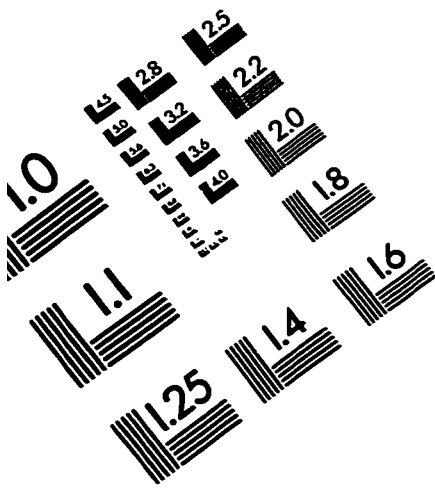
Jared C. Roach

University of Washington

1998

The author received a high school diploma from the North Carolina School of Science and Mathematics in 1985, the Baccalauréat C from the Lycée du Mont Blanc in 1986, and a Bachelor of Science in Biochemistry from Cornell University in 1990.

IMAGE EVALUATION TEST TARGET (QA-3)



APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved