Brachyury expression in the adult ctenophore, *Pleurobrachia bachei*.

Isaac Kareiva[1], Leonid Moroz[1,2], Billie Swalla[1,3], Andrea Kohn[1,2]

Marine Genomics Research Apprenticeship 2012
Spring 2012

[1]Friday Harbor Laboratories, University of Washington, Friday Harbor, WA 98250
[2]Department of Neuroscience, University of Florida, Gainesville, FL 32611
[3]Department of Biology, University of Washington, Seattle, WA 98195

Contact Information:
Isaac Kareiva
isaackareiva@yahoo.com

Keywords: T-box, *Pleurobrachia bachei*, ctenophore

Abstract:

Brachyury is a transcription factor important for mesoderm formation in chordates. Ctenophores are basal metazoans with little known about their germ layer organization. For this reason, *brachyury* was used as a possible mesodermal marker in the ctenophore *Pleurobrachia bachei* . The gene was cloned and then used for *in situ* hybridizations in adults. This revealed distinctive expression patterns in the tentacles and combs of the animal, with banding patterns indicative of possible muscle cells. The Pb *brachyury* sequence was also used to help validate extant gene models. It was found that these models did not correctly predict the *brachyury* gene structure. For this reason, the gene models were updated with new training data. The results of this were still pending at the time this paper was completed (June 1, 2012). Further research also needs to be done on *brachyury* expression in *Pleurobrachia bachei* embryos to completely understand its role in ctenophore development.

Introduction:

Brachyury is a founding member of the T-box family, a unique DNA binding domain found in transcription factors important to development. The members of the T-Box family can be identified by their binding patterns: all members of the family bind to the characteristic T-Box sequence TCACACCT (Edwards et al.).

Brachyury was first discovered in 1927 in mice, when it was observed that subjects homozygous for the mutated gene died after about 10 days due to improper mesoderm formation (Papaioannou and Lee M. Silver, 1998), whereas heterozygous mice lived but had shorter tails and an undifferentiated and incompletely developed notochord (Holland et al, 1995). This role of *brachyury* in mesoderm and notochord formation has been found to be remarkably well conserved in vertebrates (Scholz and Technau, 2003). It was further found that *brachyury* and its orthologues consistently induced mesoderm formation in animal cap assays (Marcellini et. al, 2003). However, in the invertebrate chordate ascidians, *brachyury* was found to be only expressed in the notochord and not the mesoderm (Yasuo and Satoh, 1993, 1994). Further research into other invertebrates lacking a notochord showed interesting patterns. For example, Yamada et. al found that in the ctenophore, *Mnemiopsis leidyi* (Ml), *brachyury* is required for the morphogenetic movements that form the blastopore and stomodaeum, but plays no role in

determining endomesodermal fates. Therefore, one current hypothesis is that *brachyury* evolved as a gene that regulated morphogenetic movement and only later became involved in mesoderm formation.

For both invertebrates and invertebrates, expression of *brachyury* is seen most often during the onset of gastrulation. However, while *brachyury* is seen around the blastopore, involuting mesoderm and notochord in most vertebrates, it is seen only in the blastopore and stomodaeum of invertebrates such as hemichordates (Tagawa 1998) and ctenophores (Martindale et al, 2010). Therefore, while *brachyury* is important in mesoderm formation in invertebrates and likely other, more derived phyla, its role in more basal metazoans remains unclear. For this reason, *brachyury* does not fit the role of an ideal mesodermal marker, but is nevertheless an interesting gene to investigate.

Materials and Methods

Animal collection

Adult specimens of *Pleurobrachia bachei* (Pb)were collected off the dock on Friday Harbor Laboratories. The animals were then sorted into two groups: those with copepods and those without. To avoid genetic contamination, only the animals without copepods were used in sequencing. These animals were then washed three times and used for RNA isolation.

RNA isolation and Library constructions

The RNAeasy kit was used to isolate RNA from the adult ctenophores. Primers were then designed by Andrea Kohn for PCR band isolation. These are shown below:

5’-CACAAGTACGAACCAAGACTC-3’

5’- GCTTCCGTGGGCTCTTCCT-3’

The isolated RNA was then ligated into the TOPO vector, grown on plates and sent off for sequencing. RNA from different stages of embryonic development was also used for PCR amplification and ran on a gel.

Sequence Analysis

The cloned Pb sequence was used to blast for other related sequences in NCBI. The top result for this was a full-length sequence for the closely related *Pleurobrachia pileus* (Pp). The *Brachyury* sequence for *Mnemiopsis Leidyi*(Ml) was also collected. These three sequences were then aligned to each other using muscle. The alignment is shown in figure 1.

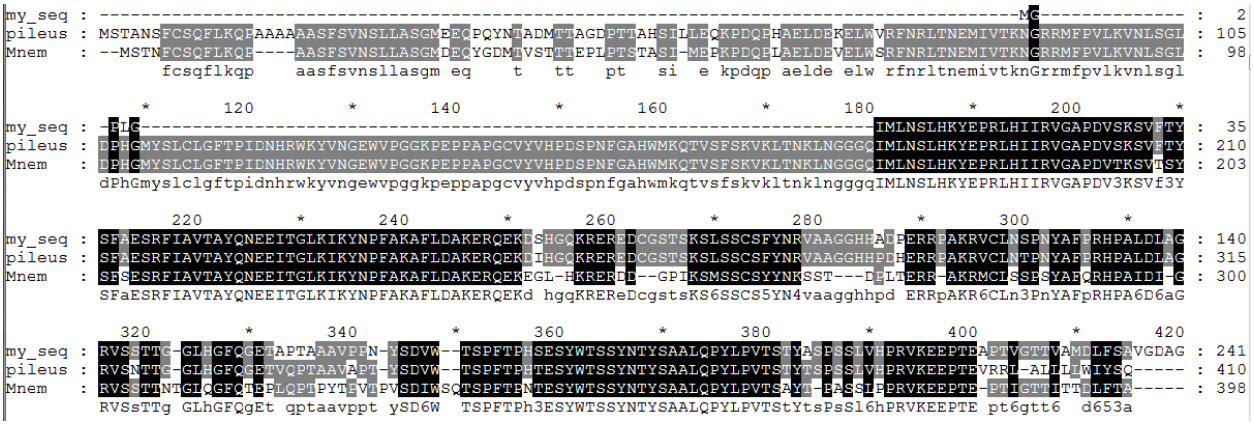


Figure 1: Sequence alignment of Ctenophore *Brachyury* sequences with Pb on top, Pp in the middle, and Ml on the bottom.

It can be seen that the downstream alignment between the three proteins is very good. However, there is a large chunk missing upstream in the Pb T-box sequence. This would indicate that the obtained Pb sequence was not full length. To try and obtain the sequence for this 5' half of the Pb gene, the Pp 5' half was blated in the UCSC genome browser. The blated sequence, corresponding to 166th position in the top row of the alignment, is shown below:

PHAELDEKELWVRFNRLTNEMIVTKNGRRMFPVLKVNLSGLDPHGMYSLCLGFTPIDNHRWKYVNGEWVPGGKPEP
PAPGCVYVHPDSPNFGAHWMKQTVSFSKVKLTNKLNGGGQ

The best hit for this had a low score of 54 and an identity of 73.7%. The genome alignment for this is shown below:

```
0214  G  K  P  E  P  P  A  P  G  C  V  Y  V  H  P  D  S  P  N  F  0273
<<<<  |  |  G  D  |  |  V  L  |  R  I  |  |  |  |  |  |  |  G  S  <<<<
3415 ggaaaaggagacccccctgtactggggaggatctacgtccacccagattcccctggttct 3356

0274  G  A  H  W  M  K  Q  T  V  S  F  S  K  V  K  L  T  N  0327
<<<<  |  |  |  |  G  |  |  |  |  |  |  D  |  |  |  |  |  |  <<<<
3355 ggggctcattggggtaaacagactgtcagttttgataaagtgaaactaacaaac 3302
```

This is a very poor alignment, and for this reason we conclude that the first half of the *brachyury* sequence is not in our assembled Pb genome (see discussion).

Alignments and Gene Tree Construction

Since both the Ml and Pp sequences were full length, they were used to blast for other *brachyury* sequences to be used in the gene tree. In order to obtain representative taxon sampling, other databases such as Neurobase, the Broad Institute and genome.jgi.doe were used as well. An attempt was made to pick only the full length sequences. However, due to possible annotation issues in NCBI, it was not always clear when this was the case. When the sequences did not align well (probably due to high levels of divergence) they were not used. Though it's important to know if sequences are divergent, including such sequences in a tree tends to significantly lower the bootstrap values even between closely related species.

Once the desired sequences were sorted out in this manner, the superphyla and some phyla (such as the chordates) were aligned with each other separately using muscle. These alignments were then combined into one file, and aligned manually (normally by shifting positions so conserved domains lined up). Then the alignment program, g-blocks, was used to filter out the conserved domains, with the parameters slightly tweaked (see supplement 1.6). It is this blocked alignment that was used for constructing a Maximum Likelihood tree (which used all sites, including gaps). The alignments for this

are shown in the supplement (section 1.6) and the obtained tree is shown in figure 8. All values below 50% were collapsed. Also note that our Pb *brachyury* sequence wasn't used in this tree due it not being full length.

Phylogenetic tree:

To further examine the expression of *brachyury* throughout the animal kingdom, the number of *brachyury* genes for different representative phyla in the phylogenetic tree was examined. To get these numbers, a search for *brachyury* with each species was done in PubMed to find relevant articles. Expectedly, not all species returned results. When this was the case, a blast search was done using a *brachyury* sequences for the closest related species to try and pick up any other matches. In this way it was found that while PubMed returned no results for *Lottia*[1] or *Capitella*, it did for the Lophotrochozoa *Platynereis dumerili.* This tree is shown in figure 9.

Domain analysis:

The Pb protein sequence was blasted in smart to obtain the domains. The only domain returned for the Pb sequence was the T-Box domain, with an E value of 9.44*10^-3. The domain position is from 1-74, which corresponds to the following sequence, where the blue represents an exon (given by the UCSC blat):

MGPLGIMLNSLHKYEPRLHIIRVGAPDVSKSVFTYSFAESRFIAVTAYQNEEITGLKIKYNPFAKAFLDAKERQ

The Pp sequence was also smarted for domain structure. The result was the same, except that the domain position extended from 64-249 (which is shown in red in the below sequence) and had an e-value of 2.57*10^-106. The underlined red corresponds to the overlapping alignments between the two

[1] Also note that while PubMed returned no results for *Lottia*, there was an annotated *Lottia brachyury* found in the uniprot database, which was used in the gene tree.

species (that is the blue part in the above sequence). The black is the flanking amino acid sequence in the Pp *brachyury* sequence.

MSTANSFCSQFLKQPAAAAAASFSVNSLLASGMEEQPQYNTADMTTAGDPTTAHSILLEQKPDQPHAELD
EKELWVRFNRLTNEMIVTKNGRRMFPVLKVNLSGLDPHGMYSLCLGFTPIDNHRWKYVNGEWVPGGKPEP
PAPGCVYVHPDSPNFGAHWMKQTVSFSKVKLTNKLNGGGQIMLNSLHKYEPRLHIIRVGAPDVSKSVFTY
SFAESRFIAVTAYQNEEITGLKIKYNPFAKAFLDAKERQEKDIHGQKREREDCGSTSKSLSSCSFYNRVA
AGGHHPDHERRPAKRVCLNTPNYAFPRHPALDLAGRVSNTTGGLHGFQGETVQPTAAVAPTYSDVWTSPF
TPHTESYWTSSYNTYSAALQPYLPVTSTYTSPSSLVHPRVKEEPTEVRRLALLLLWIYSQLLGTAPSTIP
VPICLGAQTSTADGQWWNLAHSSPDRMEQQCTIYESGMRCQITDYGH

Thus the red which is not underlined corresponds to the missing section of the T-box domain mentioned in the sequence analysis section. These findings are summarized in the below figure which compares the Pb sequence with that of the Pp *brachyury* gene.

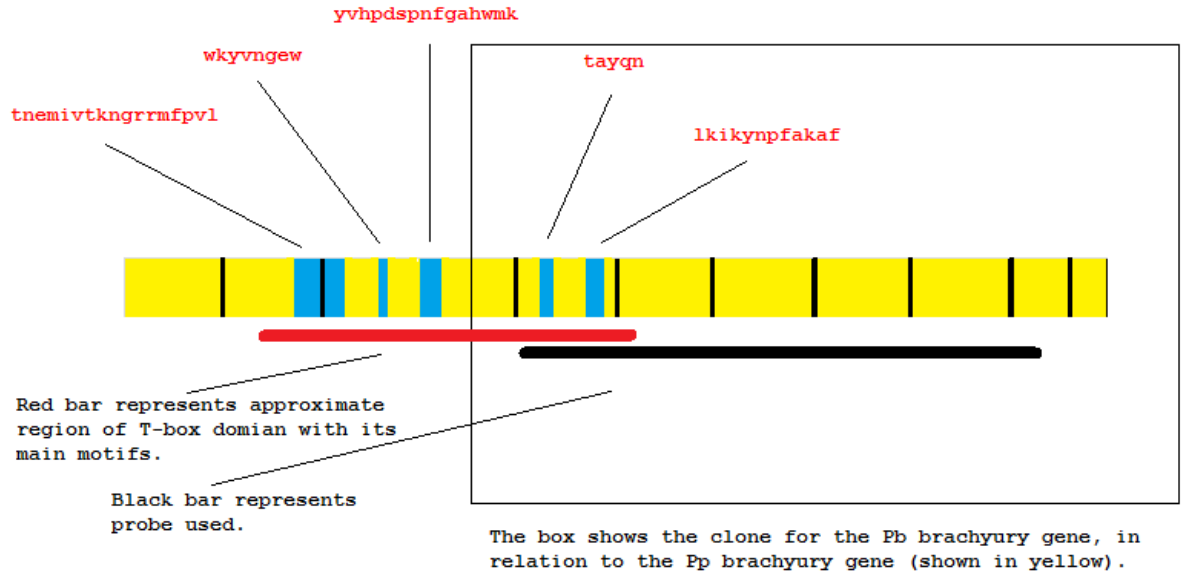


Figure 2:

The yellow block represents the Pp *brachyury* gene, with blue representing important, conserved motifs (the amino acid sequences of these are shown as well). These motifs were found by taking the black regions of the GeneDoc alignment greater than length 3 (see supplemental section 1.6). The black lines represent length demarcation of 50 bases each. The red bar represents the T-box domain, and the black line represents our probe. The box around the yellow Pp *brachyury* gene shows the approximate area of our Pb clone.

In-situ hybridization:

For the *in situ* hybridization the animals were first fixed in 4% paraformaldehyde on day 1, rinsed in PTW on day 2, and rehydrated and incubated overnight in hybridization buffer and probe overnight on day 3. The specimens were then blocked in 10% goat serum on day 4 and put in detection buffer on day 5 until expression became clear, at which point they were stopped in 4% paraformaldehyde and ethanol. For the full protocol see Supplement section 1.7. This protocol was repeated for 4 sets of animals with 3 on the first and second attempt, 2 on the 3rd, and 1 on the last. However, quality mounts were only obtained for the 3rd set of *in situs.*

Gene model validation

Our cloned Pb *brachyury* sequence was blated to the Pb genome using the UCSC genome browser. The intron and exon structure was noted for this, and then compared with our predicted gene structure from the Augustus gene models. To improve our predictions, a pipeline was set up to assemble a new training set for gene predictions. This training set was generated using PASA to align our Pb transcripts to the Pb genome. This is different from our original training set which used sponge data. Once the training set was assembled, it was used to train Augustus, which at the time this paper was written, was still building its model. For a more detailed than necessary but by no means comprehensive discussion on gene modeling, see supplement section 1.1-1.4.

Results:

Early development:

For early development a periodic pattern of *brachyury* expression was seen, where it peaked at the 8 cell stage, then disappeared, then returned during gastrulation.

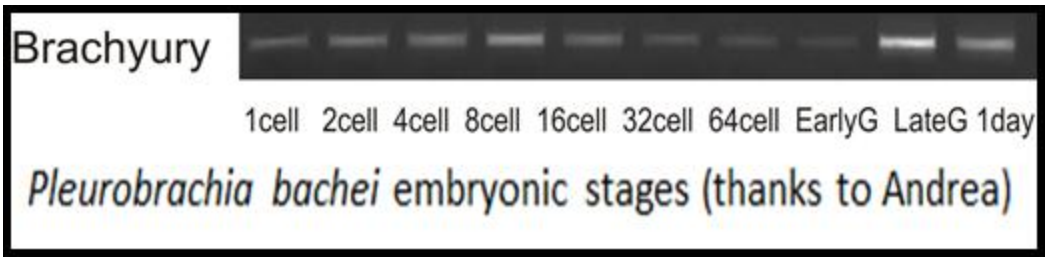


Figure 3: *brachyury* expression in Pb embryos.

Brachyury expression in adults

Of the 4 rounds of *in situs* performed, the only quality mounts obtained were from the 3rd round (this was due to bad mounts, not inconsistent staining). These are shown below. The expressions pattern seen in these are consistent with the other *in situs*, with some variability only on the expression levels. The first image shows the expression (shown in purple) in the tentacles of the adult ctenophore.

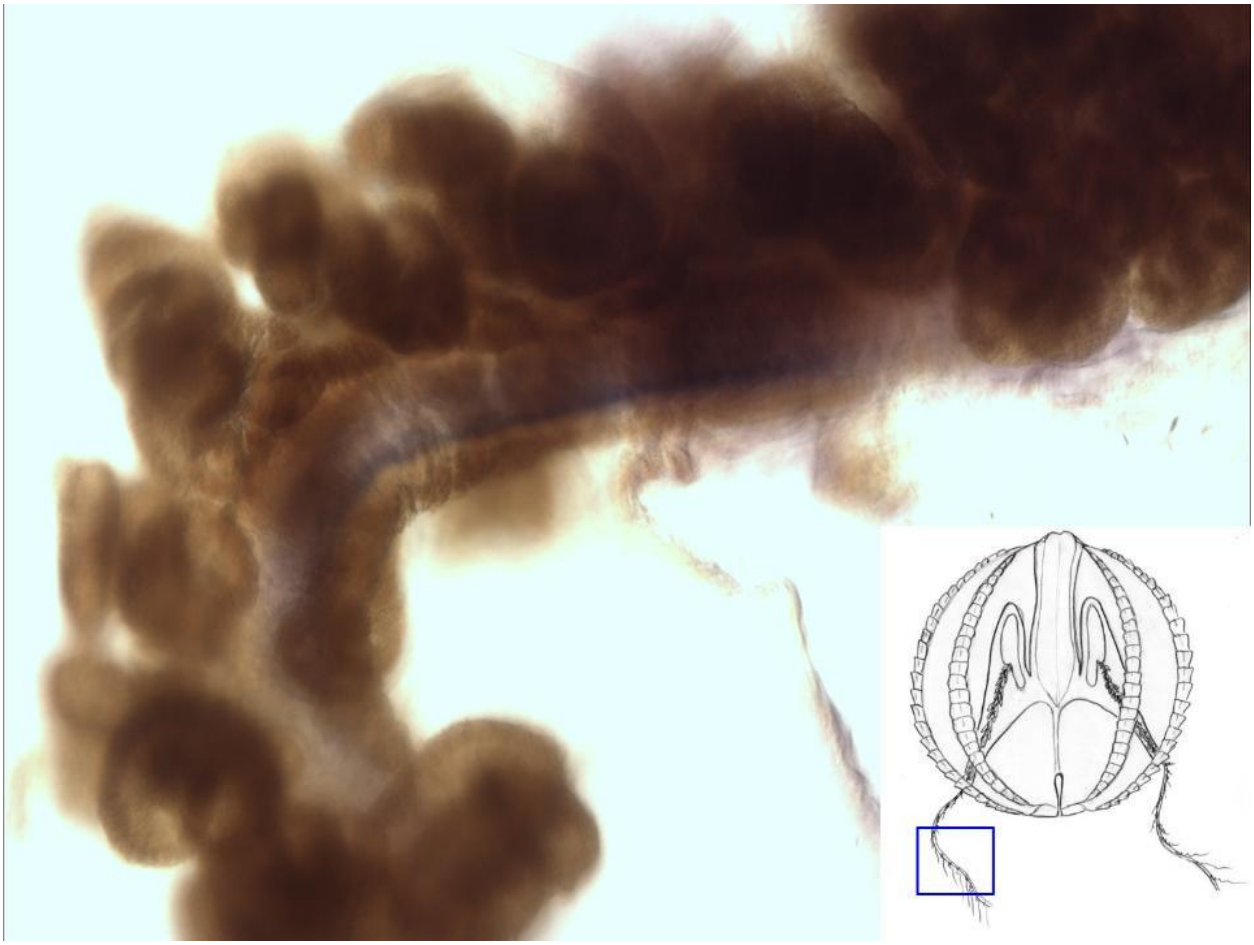


Figure 4. Picture taken at 20x of the tentacles of an adult Pb showing expression down along the middle.

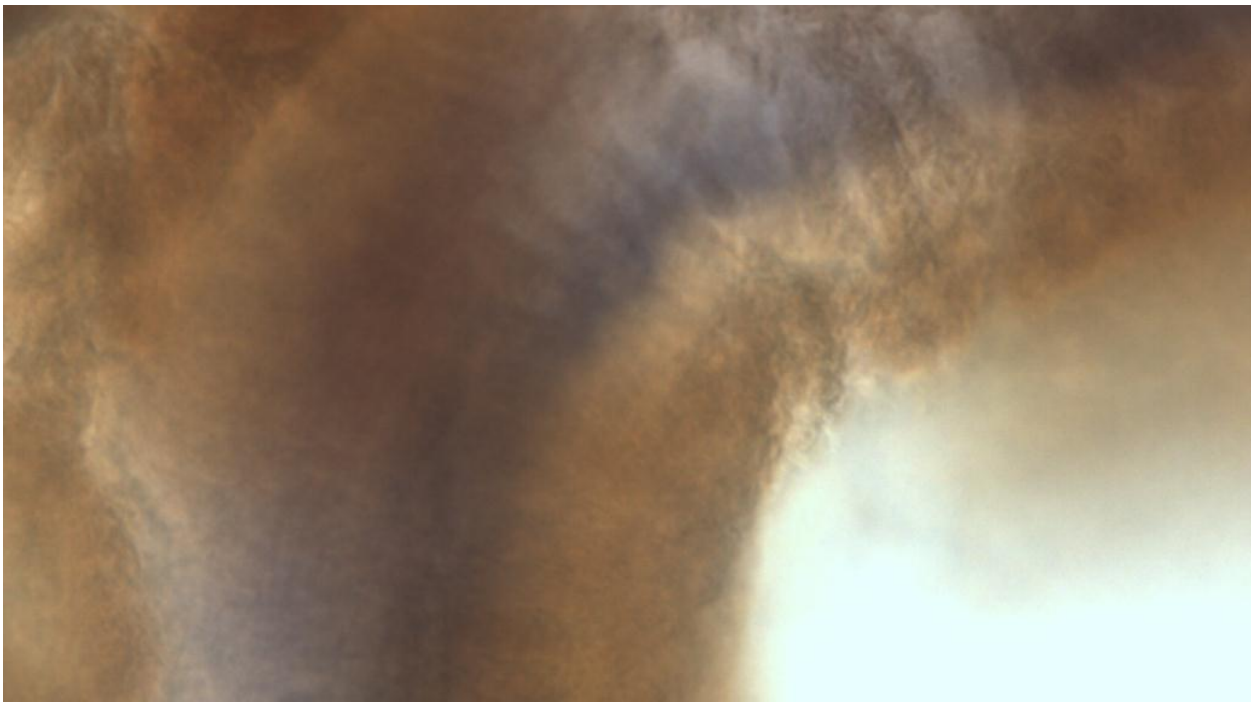


Figure 5. Enlarged picture of tentacles taken at 20x showing band like patterns.

Expression was also seen along the comb rows. Upon examining it under the microscope it became clear that the staining was actually under the combs. This was seen by adjusting the focus.

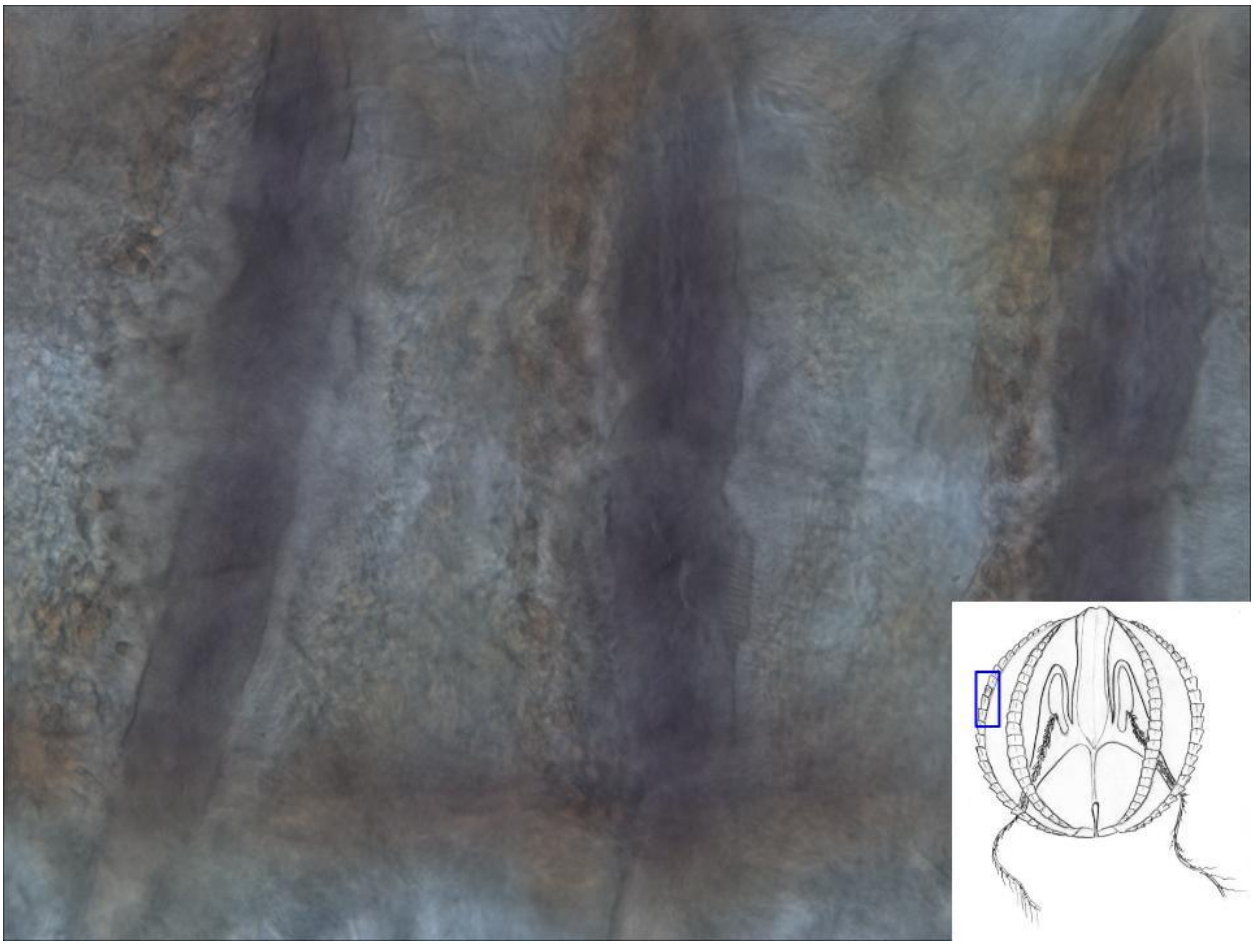


Figure 6. Expression of *brachyury* along comb rows.

Gene model comparisons:

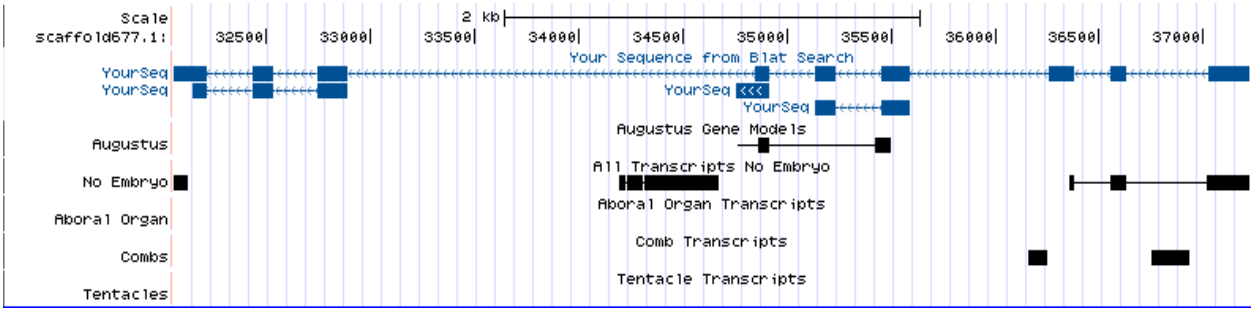


Figure 7. A snapshot of the cloned *brachyury* sequence compared to the predicted Augustus gene model. Transcripts from different regions of the Pb animal are shown below the Augustus model..

Discussion:

Incomplete clone:

The data collected thus far reveal a few interesting results. First is the absence of the 5' end of the T-box domain in the Pb genome. One possibility to explain this is that the genome assembly itself is not ideal. That is to say, Pb does have the full *brachyury* sequence, but it hasn't been correctly assembled into the genome. This could happen in several ways. First, the 5' end of the gene could have been missed in the initial sequencing step. This would be somewhat unusual considering the genome had 10000x base coverage (meaning that each base was sequenced on average 10,000 times). Of course, given that this is an average, the statistic leaves open the possibility of still missing entire stretches of DNA. Nevertheless, the second half of the gene clearly was present in the raw reads, and assembled into the genome. The other way that an incorrect assembly could occur would be if the 5' brachyury end was in the original raw reads, but that they were extremely fragmented. Assembly requires a certain threshold of overlap, and if the fragmentation was such that this threshold wasn't achieved, they wouldn't be assembled into the genome.

The other possibility is that the Pb genome truly does have an incomplete *brachyury* sequence. This is unlikely for the following reasons. First, it has been found in other ctenophores, namely Ml and Pp. Furthermore, *brachyury* plays an important functional role in determining morphogenetic movements in Ml development (Martindale et. al 2010): Ml with the morpholino oglionucleotide knockdown of *brachyury* were severely deformed. Therefore, it seems highly unlikely that a gene which plays such a critical role in one species plays no role in another closely related one. A similar argument pertains to the alignment between ctenophore amino acid sequences, which was extremely conserved (see figure 1). This conservation would suggest that the though the three species diverged at some point, selective pressure prevented divergence in the aligned segments of the protein. However, it makes little sense for selection to act on all but one half of a gene for one species, and then

for the whole length of the gene for the other two. In other words, why conserve only half a gene that in all likelihood needs its full length to be functional? Extending this argument, what was actually missing from the Pb sequence were distinctive conserved elements of the T-box domain. The T-box domain is found in a host of genes, not just *brachyury.* For that reason, proposing a loss in the first half of the *brachyury* sequence also implies a functional loss of all members of the T-box family. Considering that this domain has been found in every metazoan to date (Martinelli and Spring, 2005), this is very unlikely.

For the reasons outlined above, I assume that the true reason for the results is that an incomplete transcript was cloned. Further, due to an incomplete assembly, the putative other half of this gene was not found in the blated genome. This hypothesis can be easily tested by doing a 5' RACE. It would also be interesting to examine the original raw reads from the sequencing project to see if the 5' end is present at all. Proposing an assembly problem isn't nearly as interesting as suggesting a missing half of a highly conserved metazoan gene, but it's still an important problem that needs to be addressed.

Gene tree:

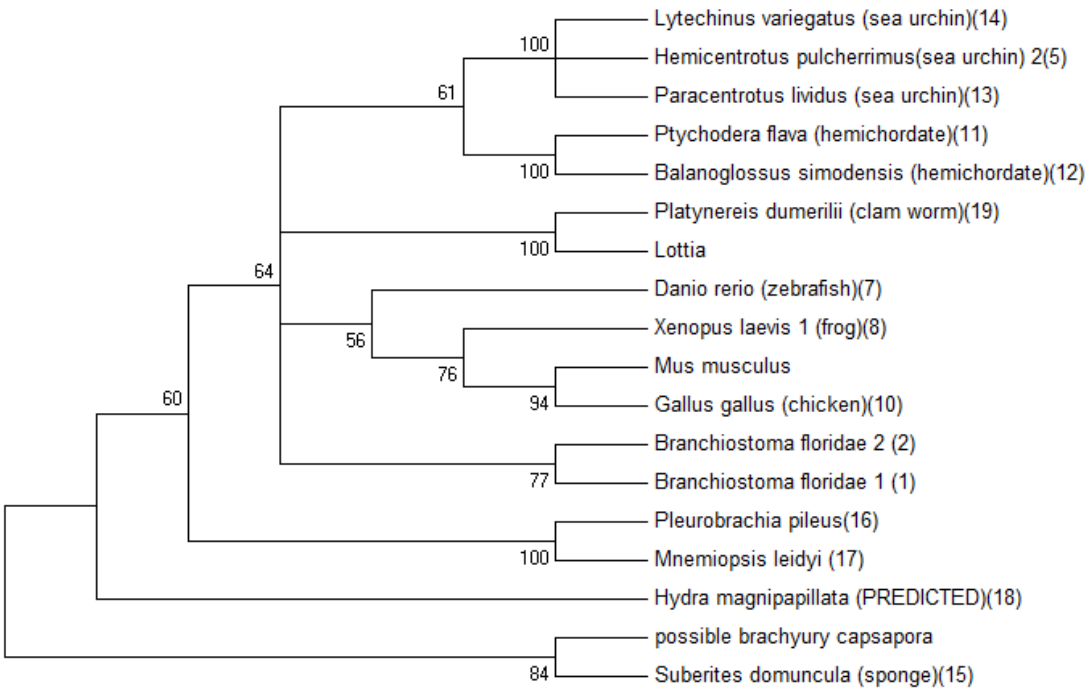


Figure 8: Gene tree for *brachyury* protein. The parenthesized numbers at the end of the species names are referencing a file used to organize the sequences, and not important to understanding the tree. All bootstrap values below 50% are collapsed.

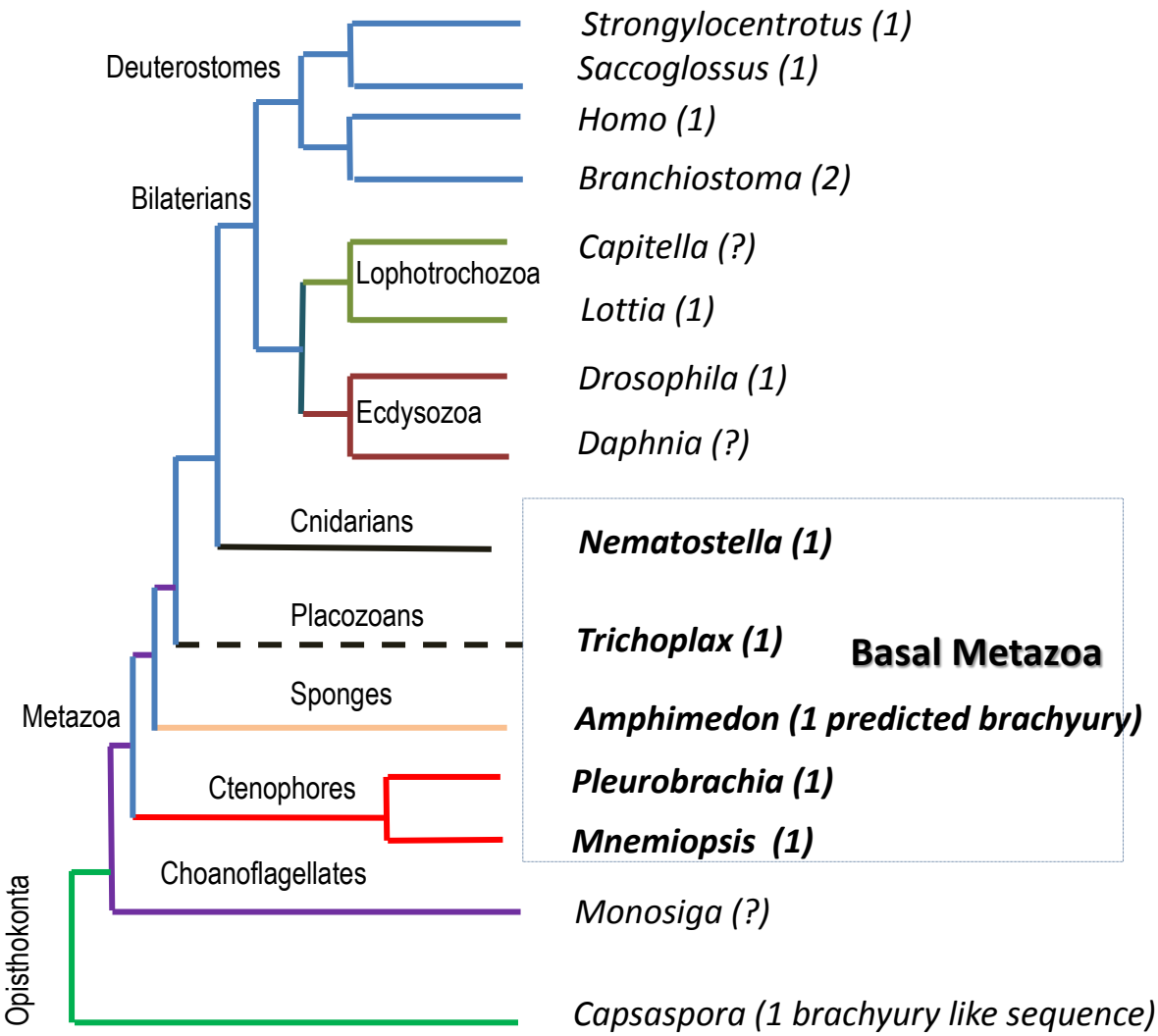


Figure 9:

This figure shows the number of *brachyury* genes throughout the phylogenetic tree. The numbers were obtained by searching PubMed databases for articles on the represented species. The numbers for the corresponding species were taken from the following articles: *Strongylocentrotus* (Peterson et. al, 1996); *Homo* (Mingullion and Logan, 2003); *Branchiostoma* (Holland et. al, 1995); *Drosophila* (kispert et. al, 1994) Nematostella (Scholz and Technau, 2003); Amphimedon (blast result); Tichoplax (Martinelli and Spring, 2003); *Pleurobrachia* (Martinelli and Spring, 2005); *Mnemiopsis* (Yamada et. al, 2010); Capsapora (Sebâe-Pedrâos et. al, 2011). For search results that did not return any results a question mark has been left. Also note that while no full length *brachyury* has been found in *Lottia* or *Capitella* (due to absence of research, not to negative results), *brachyury* has been found in the polychaete annelid *Platynereis dumerili*, a basal Lophotrochozoa (Technau, 2001). This suggest that in the absence of losses, *Lottia, Capitella* and other Lophotochozoas probably contain the transcription factor. Also note that while no PubMed articles could be found referencing *brachyury* in the sponge *Amphimendon queenslandica*, it did come up in a blast search as a predicted protein. Furthermore, *brachyury* has been isolated in the sponge *Suberites domuncula* (Adell & Muller, 2005).

The other interesting results thus far pertain to the gene tree (figure 8) and phylogenetic tree (figure 9). The gene tree groups the ctenophores together, but as more derived than sponges. Since there is some debate on which group is more basal, this is an interesting, but not particularly informative result. It is not particularly informative for the following reasons. First, the bootstrap values are low. And even where the values are not collapsed, the topology is not entirely correct (for instance, *Hydra* is placed more basal than ctenophores). Secondly, as a gene tree the information contained within it is necessarily limited: it is difficult to infer a phylogeny from a single gene. Finally, while representatives of the sponges (*Suberites Domuncula*) and non-metazoans (*Capsapora owczarzaki*) were included, these alignments were not very good, and kept more for the sake of comparison. These two species were probably placed as the most basal because they had the most divergent sequences, and thus had the greatest distance between the other species (which would put them as most related to each other and less to all other less divergent species). However, in reality, divergence doesn't necessarily imply a group is basal. Higher mutation rates in the sponges, could, for instance, account for greater divergence. Similarly, different selective pressures could affect the protein. In fact, Manuel et al. recently argued that while the sponge *brachyury* sequences are indeed divergent, this is due to autapomorphic changes within sponge lineages. That is to say, sponges are not different from other metazoans due to retention of a more basal state. And in fact, different species of sponges are very divergent even between each other (Manuel et. al 2004). For this reason, the constructed gene tree should be viewed as perhaps informative for grouping phyla, but not for inferring broader phylogenies.

The last result of interest is the notable lack of duplication of *brachyury* throughout the phylogenetic tree. Why this is remains to be seen. Interestingly, Yang et. al found that duplications of the *brachyury* gene in humans leads to significantly increased susceptibility to familial chordoma. Since untreated chordoma is generally fatal, it seems selective pressure would be great enough to prevent a duplication event in humans. Given this result, it would be interesting to see how over expressing

brachyury in other species would affect the organism. As it is, there are currently two confirmed cases of duplication: one in the *Hydra* (Bielen et. al 1995) and one in *Branchiostoma* (Holland et. al 1995). What's interesting is that in the *Hydra's* case, the two *brachyurys,-HyBra1* and *HyBra2*- are functionally unique: *Hybra1* was found mainly in the endoderm, whereas *Hydra2* was found mainly in the ectoderm (Bielen et. al, 2007). Furthermore, when the two genes were inserted into the vertebrate *Xenopus laevis*, they played functionally different roles as well. *HyBra1* acted most like the vertebrate *brachyury* by inducing mesoderm formation, whereas *HyBra2* was shown to induce neural activity (Bielen et. al 2007). Bielen et. al further showed that this was due to divergence in the sequence in the C-terminal domain. This suggests that even genes close enough to be annotated as *brachyury* in terms of sequence distance can play functionally different roles. This is interesting for two reasons. First, it is a possible explanation for why *brachyury* is so conserved: with just a small change in the amino acid sequence, the gene can perform differently. Perhaps if this change is in an important functional area, the affect could be negative (and *HyBra2* is a unique case where it is not). Extending this, the result is also interesting because it provides evidence for the co-option hypothesis, which states that *brachyury* has been co-opted for different uses in different lineages. For example, it is used in notochord formation in all chordates (Scholz and Technau, 2003), but not in basal groups like the ctenophores, where it is expressed mostly around the involuting blastopore (Martindale et. al 20010). Having conserved sites where small changes can cause broader developmental changes is one way for co-option to occur. Indeed, Marcellini et. al found that swapping the N-terminus and C-terminus domains of the *brachyury* orthologues and expressing them in the chick embryo provided different results. For instance, the *Xenopus* orthologues consistently induced mesoderm formation in the chick embryo, and the *Ascidian* orthologues consistently induced both endoderm and mesoderm formation. But swapping the N-terminus domains for these two orthologues and injecting them into embryos, the pattern was reversed. Therefore, it appears that different unique motifs in the brachyury sequence can lead to

different functional roles, once again supporting the co-option hypothesis. Further researcher on the ancestral function of *brachyury* still needs to be done, though. It will be particularly important to observe *brachyury* expression in Pb embryos. Yamada et. al has done so for the gastrulating stages, but little is known about the earlier stages. Ideally, development should be followed from the one cell stage to trace back the origins of cell fate determination.

Expression patterns

That *brachyury* is even expressed in adults is an interesting result in itself. To date, most research has focused on the developmental role of the gene, and little is known about what role it plays in adults of any species. If Yamada et. al is correct in concluding that *brachyury* is important in early morphogenetic movements, this is not *brachyury*'s only role, as these early cell movements should be complete in adults. One possibility is that *brachyury* is labeling mesodermal cells, specifically muscle cells. This is plausible for two reasons: where it's expressed and how it's expressed. The banding pattern in which it is expressed in the tentacles is indicative of possible striated muscles. This pattern is not present (or at least not distinctly so) in the combs, but this region could still contain possible smooth muscles. Further, the regions in which it is expressed-the combs and tentacles-make morphological sense: the tentacles contract and the combs move. Aside from this, the fact that *tropomyosin*, a typical mesodermal marker, also stained the tentacles and combs in a similar pattern (see Zander's report) supports this idea. The catch is that *tropomyosin* labeled other areas as well (mainly the blastopore) that *brachyury* did not. So *brachyury* and *tropomyosin* cannot simply just be labeling and mesoderm present. To investigate this relationship further, where the mesodermal and muscle cells are needs to be definitively determined. The best way to do this is to use electron microscopy. Phalloidin staining could be used as well.

Once the cell types have been determined, the relationship between gene expression patterns can be investigated further. Note, however, that suggesting *brachyury* is labeling mesodermal regions does not answer why it's expressed in adults. Even if *brachyury* induces mesoderm formation as it does in animal cap assays, the mesoderm should be completely formed in an adult ctenophore. If it is related to cell growth or cell repair, it's a surprisingly consistent pattern (that is, it's expressed compactly in every single comb and throughout every tentacle). So if it were related to cell repair, the tentacles would contain a long band of consistently damaged cells in need of repair. Further, cell repair is not the same as mesoderm determination. However, the expression could also possibly be indicative of muscle growth. But again, all this is needs much more research.

Gene model comparisons:

The gene model comparison shows a poor match for the model and the actual gene structure. Of course, this is only one gene amongst many in the genome. Nevertheless, the gene models can undoubtedly be improved. It would also be useful to perform a summary of the predictions of the two models, the old one (a snap shot of which is shown in figure 7) and the new one currently still being built. One useful measure would be the number of genes predicted in each model, and the number of genes validated in each.

Conclusions:

The most that can be concluded from this research is limited to (1) *brachyury* is present in Pb and (2) it's expressed in defined regions. However, insight into its functional role in these regions hinges on more research, namely embryonic *in situs*. First and foremost, though, a full transcript needs to be obtained by 5'RACE.

Acknowledgements

I would like to thank the following people. First, the teachers for putting in the amazing amount of effort required for this course: Dr. Andrea Kohn for help in the lab and with sequencing, Dr. Billie Swalla for answering the many *brachyury* related questions and Dr. Leonid Moroz for the driving passion behind ctenophore research. I'd also like to thank all the TAs who helped, especially Mathew Citarella who put in countless hours in managing the bioinformatics end of the project and helped walk me through setting up the pipeline and the beginning of bioinformatics. Lastly, thanks to the University of Washington and to NSF for the funding required to make these apprenticeships possible.

Supplementary material

1.1 PASA pipeline outline:

To model gene structure, Augustus first needs some form of data telling it what introns, exons, starts, stops, and splice sites looks like. This data is called a training set. The training set for the first Augustus model used annotated genes from the sponge *Amphimedon queenslandica.* Since different species can have different DNA sequence patterns (or kmer frequencies), it was thought that deriving the training set from a sponge affected the original predictions. For that reason, the training was redone using *Pleurobrachia* data.

The pipeline, set up by Matthew Citarella, worked as follows: first, the transciptomes were assembled using inchworm. Second, PASA was used to align the transcripts to the genome and create an assembly of genes with gene structure. The three files used for this process were the genome file, the assembled transcriptomes (from inchworm) and our full length cDNAs. PASA generates a file which contains the best candidates for training data. This file, along with the genome and full length cDNAs was submitted to Augustus for training. When training is complete, the genome along with a file of hints (in the form of transcriptome files) can be used to re-predict the genes and their structure. Some (but not all) of the details of the PASA and Augustus aspects of the pipeline are explained below.

1.2 Hidden Markov Modeling: Augustus

The idea behind gene modeling is as follows. Given a random sequence of bases, use a model to predict where the introns, exons, splice sites, and etc. are. The way to do this is to construct a hidden markov model (HMM). For a good mathematical introduction to HMMs see Rabiner and Juang. The idea behind an HMM is this: First, you have a collection of *states*, and from each state you can *transition* to another state with some probability. Further, each state *emits a pattern* from a set of patterns with some probability. To give an example, imagine a casino where the dealer at a craps table has a loaded die and a normal die. If he is rolling the loaded die, there's a 75% chance he will switch back to the unloaded on the next turn. If, however, he's rolling the unloaded die, there's a 10% chance he'll switch to the loaded one. These rates are the transition probabilities. Further, each die emits a pattern: its number when rolled. For the unloaded die the probability of emitting pattern 1 (rolling a 1) is 1/6, the probability of emitting pattern 2, (rolling a 2) is 1/6 as well, and etc. For the loaded die, these probabilities are different (say 1/10, 1/10, 1/10, 1/10, 1/10, 5/10 for states 1, 2, 3, 4, 5 and 6 respectively). The point is this: given this set of die, a certain sequence of roles will have a corresponding probability. Further, each sequence of states will have a different probability of emitting the rolled sequence. So one could get the sequence '6,6,1' by rolling the loaded die 3 times, or by rolling the loaded die twice, then transitioning to the unloaded one. Or any number of ways. Some ways of rolling that sequence will be more likely than others, though. With this in mind we turn to DNA.

We said earlier that we wanted to predict the gene structure (exons, introns, splice sites, starts, stops) from the sequence. One way to do this is to break the structure into states. So call being an exon 'state exon', for example. Further, we could call the DNA sequence an emission. So the sate 'start' emits the sequence ATG 100% of the time (if the model is good and biologically accurate, at least). A schematic diagram of what a very simple model would like is shown below. Here, circles represent states, diamonds emission, and arrows different transitions to different states. Each arrow has a corresponding probability, and each emission has one as well. Notice, though, that not all the possible

arrows are drawn. For example, there's no arrow connecting a 'start' and a 'stop' because you don't see start and stop codons next to each other in a DNA sequence.

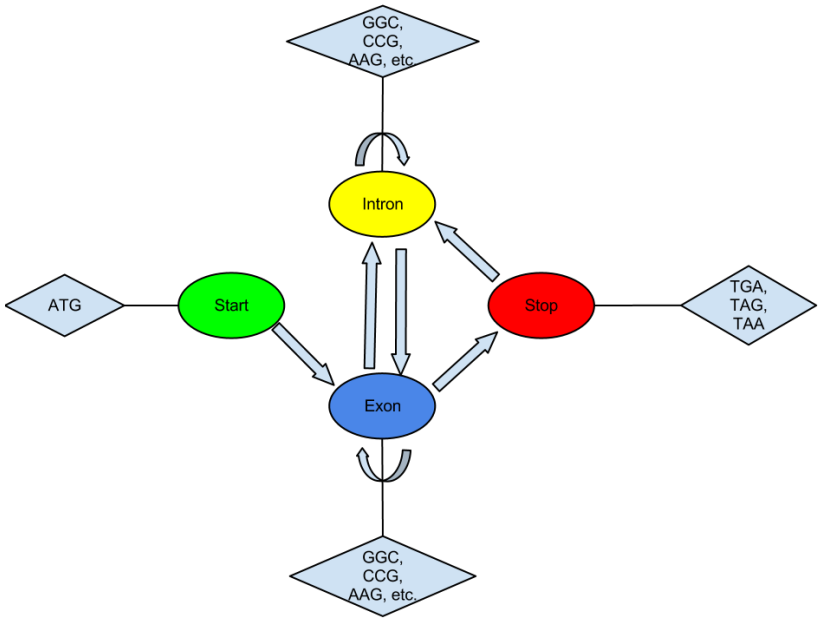


Figure 10. Schematic diagram of simple HMM

So if a computer were given some DNA sequence, say ATGCCGGACTGA, the goal would be to find the most likely path through the model, because this most likely path corresponds to the most likely gene structure. If the model were based in threes, it would look at the sequence as ATG GCC GAC TGA. So, one possible path would be Exon→Exon →Exon→Stop. This would correspond to starting in the EXON state, emitting an ATG, then looping back to the Exon state twice more and emitting a CCG then GAC, then transitioning to a Stop and emitting a TGA. Finding the likelihood of this path is as simple as multiplying the transition and emission probabilities. Whether or not this is a good gene structure depends on how it compares to other possible paths (maybe to the human eye it seems like a better guess would be to being in state 'Start', but such offhand knowledge is not possible to a computer. Further, for large data sets with thousands of possible paths and multiple possible reading frames, there is really no 'logical' first path). So to state the problem in formally, the question of finding the best gene

structure can be phrased as this: given a HMM with known states, transition probabilities and emission probabilities, and given an emission pattern (the DNA sequence) what is the shortest path through this HMM state machine?

In other words, if one has an emission from a HMM, can one recover the most likely state sequence? One simple way would be, as hinted at earlier, to test every possible path and see which one is the most likely. But this is computationally expensive. So instead, an algorithm is implemented, known as the Viterbi algorithm. For a good outline of the algorithm, see Forney. The idea is simply to calculate the shortest paths stepwise (from one transition to another) and then re-use these paths in latter calculations to cut down on the required computation. You can reuse the shortest paths up to a certain time step for the following reason. If one is trying to find the shortest path from A to Z, and has the shortest path from A to Y, to path AZ most go through this shorter path AY (assuming one must go through y, of course). Computational details aside, what's important to know is that this algorithm exists and it can be used to solve the gene model problem.

The above is a general introduction to HMMs in gene modeling. But it is the general idea that is important. In predicting our gene structure, we used a program called Augustus+ which employs the same general strategy of finding the shortest path through a given model, as outlined above. A diagram of the original Augustus model is shown below (note that diamonds don't mean emissions in this diagram (as they did in the example), but rather states which emit patterns of fixed length).

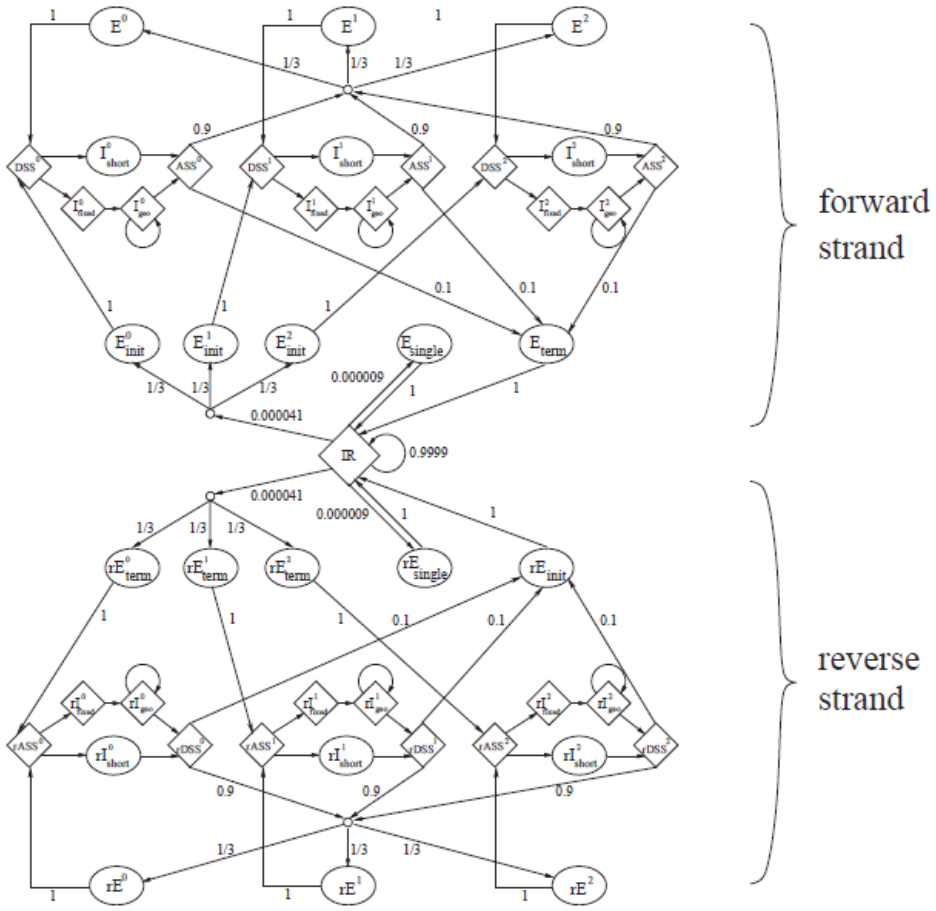


Figure 11. Diagram of model used in Augustus, taken from Stanke and Waack, 2003.

The model is necessarily more complex, but not fundamentally different from what has been described thus far. For the details of the Augustus model though see Stanke and Waack, 2003.

However, Augustus+ is slightly different from the original Augustus. These differences are discussed in the following section.

1.3 Augustus+ and incorporating hints in gene prediction:

The idea of a hint is as follows: if one has a sequence of DNA, then that sequence can be 'blasted' to a database such as NCBI. If you're blasting a protein, these blast results can return homologous sequences. Further, if these homologous sequences are annotated, and an ATG on your

sequence lines up with an ATG on the known, homologous sequence, then this gives you pretty good confidence that you have a coding region with a start at that particular position. You can be more confident with that bit of information than without it. Thus, Augustus+ incorporates hints into its model which update the probability of the region in question. Conceptually, the model is very similar. All that's changed is the probability of certain events[2].

Augustus+ uses seven types of hints, each indicating a different structure. These types are [start, stop, donor splice site, acceptor splice site, exonpart, exon] and each is specific to a strand orientation[3].

An example of a hint is given below. In this example, one has taken a collection of sequences to be used as 'training' data (see other section). In this collection the true structure is presumed to be already known, and this collection of sequences has been blasted to the NCBI database. The notation q(start) means your model is proposing a 'start' site. The +/- refers to the compatibility of your model (we say model we mean the predicted protein structure). A '+' means that the start is compatible with the model and the sequence. A '-' means the hint is compatible with the sequence, but not the model. A hint compatible with a model means the structure agrees with the extrinsic information given by the hint. How this works will be clarified in the example.

(1) q+(start)=386/500=.77

For the training set there were 500 ATG start sites within the proteins. For 386 of these there came a hint (received from a protein database search) indicating a start site. Thus, if the sequence is in a given protein, there's a 77% chance of receiving a hint. It is a q+ because the hint is compatible with the sequence (a hint incompatible with a sequence would be anything proposing a start for a non ATG

[2] The alphabet of the model has to be slightly altered, so that instead of' a,c,g,t' there is a with 'a hint,' 'a without a hint', 'b with a hint', and etc. For details refer to Stanke et al.

[3] For details on how the hints are "oriented" on the strand, again, see Stanke et al.

sequence) and is compatible with the model because the hint is proposing a start site where there is actually one.

(2) q-(start)-47/145000)=3.2E-4

The ATG start motif also occurred 145000 times outside a protein but only received hints for 47 of these. This q- corresponds to the probability of a false negative. So the hint is compatible with the sequence, but not with the true gene model (since the ATG probably isn't actually the start of a protein)

In other words, given that your model says you have a protein (and that the model is correct) and you have an ATG sequence within this protein, you'll receive a hint that this particular codon is a start codon 77% of the time. However, you'll also receive a hint for a small portion of the time even when the model is incompatible.

What affect do hints have on the probability of a gene model?

As stated, a hint can either be in favor of the model or incompatible with it. There are two questions which will be addressed here. (1) How does one find probability that the model is correct given a certain hint? And (2) How does the presence of hints affect the probability of a particular model? For the sake of this discussion, we consider the model (M) and the hint (H) but not the sequence. What we are concerned with is P(M|H). This information can be used to find P(M,S,H), that is the joint probability of the hint, model, and sequence, which can ultimately be used to maximize P(M|H,S). The first question can be addressed simply by using Bayes Theorem:

$$P(M|H)=\frac{P(H|M)*P(M)}{P(H)} \qquad (1)$$

Which can trivially be written for the case that there is no hint (H=0) for the sake of manipulation:

$$P(M|H=0)=\frac{P(H=0|M)*P(M)}{P(H=0)} \qquad (2)$$

Setting each equation equal to P(M) and then setting the two equal to each other, we obtain:

$$\frac{P(M|H=0)*P(H=0)}{P(H=0|M)} = \frac{P(M|H)*P(H)}{P(H|M)} \quad (3)$$

This can be rewritten as:

$$P(M|H)=P(M|H=0)*\left[\frac{P(H=0)*P(H|M)}{P(H=0|M)*P(H)}\right] \quad (4)$$

So adding a hint modifies the probability that the model is correct by the factor in brackets. To take an example, suppose one has a series of possible start sites that have been found in a sequence of DNA. Some of those putative starts also have a hint. Using the values from before:

$P(H|M)=q+=.77$
$P(H=0)|M)=(1-P(H|M))=.23$
$P(H)=q-=3.2E-4$ (That is, for all ATGs in a sequence of DNA, there is a hint given about 3.2E-4 times)[4]
$P(H=0)=1-P(H) \approx 1.$

Plugging this into (4):

$$P(M|H)=P(M|H=0)*\left[\frac{1*.77}{.23*3.2E-4}\right] \quad (5)$$

Call the (1*.77)/(.23*3.2E-) the modifying factor. So P(M|H=0) gets modified by a factor of about 10462. This simply says models unsupported by hints are about 10462 times less likely than those with hints (for the particular case of a start codon, at least). Therefore, sequences which have hints attached to them are rewarded in the model. That unsupported models are less likely is intuitively obvious. The effect is nice, though. For instance, consider a sequence s for which 2 possible gene models, M1 and M2, are being compared. Both models are for a protein with some organization of introns and exons. Suppose the models are identical, except at one region, i-j. In this region M1 has no exon in its model and M2 has exactly one. We want to see how the models compare.

[4] P(H) is not actually the same as the q- value used here, since P(H) is the probability of receiving a hint in general, compatible or not. However, this is the value that the original paper used, though they derived their modifying value somewhat differently. For details, see Stanke et al.

Postition: I,… j

M1 ----------

M2----------\||||||||||||/------

Here the '--' indicate an intron position, the /, \ splice sites (donor and acceptor) and the '||'s exon parts. We see then that the probabilities of receiving a hint for any position in the two models are exactly the same except for at these splice sites and exon part regions. However, it is much more likely to receive a hint for any of these structures-that is a splice site or exon region-than just a length of non-coding intron[5] . Conversely, if we're more likely to receive a hint for these structures, we're less likely to not receive a hint. Using this, we want to see how P(M1|H=0) and P(M2|H=0) compare. However, since maximizing P(M,H) is same as maximizing P(M|H) (which is ultimately what we're trying to do with the state machine), we concern ourselves only with earlier[6]. For a joint probability, P(M,H)=P(M)*P(H|M). Therefore:

P(M1|H=0)=P(M1)*P(H1=0|M1)
P(M2|H=0)=P(M2)*P(H2=0|M2)

As stated earlier, P(H1=0|M2)<P(H1=0|M1). So if the models M1 and M2 have about the same likelihoods overall (since they have about the same structure except for the i-j region in question), but one proposes an exon at one position and the other does not, and neither model have hints, then the model with the proposed exon but no hint will be penalized. Therefore, models which propose certain structures should be backed up by hints, otherwise there will be a penalty. This can prevent a problem common in modeling where an excessive amount of splice sites are added. In other words, the absence of information can still be useful in determining which model is best.

[5] The reason for this is that exons are parts of proteins, and thus by definition functional. Therefore, they are also more often conserved. It's thus much more likely to find align a conserved region to another conserved region in a blast search than to align a non-conserved region to any other region.

[6] Take my word for it. Or refer to Stanke et al. if you're still actually reading this (and props if you are).

The math sketched out here is not a proof of the effectiveness and accuracy of the Augustus model. Rather, the point is to show as efficiently as possible the two main ideas behind hints. One, that attaching a hint to a structure can increase our confidence in our model and second, that the absence of hints is a valuable piece of information as well. In the more general sense, this hones the emission probabilities for the state machine, and thus changes likelihood of different paths through the model, with the overall result being a hopefully better estimation of the most likely gene model.

1.4 PASA and training sets:

Now that we have an understanding of how the Augustus model works to predict structures, the only question is how do we build it? That is, how do we find the transition and emission probabilities. Generally, finding these probabilities for a HMM is a very hard problem when one can only observe the model in action. But in our case, we want to build the model from scratch. The way to do this is to use known gene structures as a training set. The idea is that if we look at a large collection of known gene structures we can simply observe the transitions and emissions. For example, we did this in the above section with the hints example. We saw that known start sequences hit to a database search 77% of the time. Therefore the emission probability for an ATG with an attached hint from within a start site is 77%. Without going into details, finding the transitions and emissions for the other states uses the same idea of simply observing the frequency of events given the others. Augustus will do this for us if we just submit these gene structures to it as a training set.

So then all we needed to do was generate the training set. A training set is simply a set of DNA with known gene structures (exons, introns, splice sties, etc.) annotated. Fortunately, there is a program, PASA, which can be used to do this with given genome and transcript data. This program is outlined below.

PASA: alignment and assembly

First, assume one already has the assembled transcripts. The issue here, then, is to align these transcripts to the genome. Where these transcripts line up represent coding regions. If there are full length transcripts (that is, a transcript spanning from the start to the stop of gene), then one can get the intron and exon structure from these alignments (by looking at where the transcripts, which are necessarily exons, line up to the genome). So this is an alignment and assembly problem. A program which deals with this is PASA, and the general outline of its algorithm is described below, using a simplified example. For a detailed explanation see Haas et. al.

Suppose you have four transcripts, call them *a*, *b, c* and *d*, and each align to the genome at some point. Further suppose that from these transcripts you have inferred some exon boundaries (represented below by brackets)*: *a, b, c* and *d's* alignment to the genome is shown below:

```
Genome     : ------------------------------------------------------------------
Alignment a:      [III|||||||||]          [|||||||||||||||
Alignment b:           |||||]          [|||||||||||||||||||||]
Alignment c:         [|||||||]          [|||||||||||||||
Alignment d:             |||]           [||||||||
```

Note that while all align to the genome, and that while *a* 'agrees' with *b* and *b* 'agrees' with *c* (i.e., the inferred intron/exon boundaries match up in each) *a* does not agree with *c*. Therefore, agreement between alignments is not transitive. The issue then becomes, how does one create the largest possible 'agreeable' assembly? That is, we want all transcript alignments in the assembly to have the same intron/exon structure between each other, and we want the assembly as long as possible. The algorithm to do this is described below. First, defining terms:

L(x): longest assembly of transcripts that end in transcript x. So in the above picture, if it were an agreeable assembly, it would be Lb.

C(x,y): here x and y are two agreeing alignments. C(x,y) is the number of agreeing alignments within the span (that is from where a first aligns to the genome, to where the alignment ends, including regions in between transcripts) of *x* (including *x* itself), but not contained within the span of *y*. So in the figure above, C(c,b)=2 (Alignments *b, c* and *d* are within the span of *c,* but *d* is eliminated because it is also within the span of *b*). Note that to be *within* the span is not the same as being completely *contained* in the span.

C(x) will denote the number of x-agreeing alignments within the span of x. So in this case, C(c)=3

The maximizing equation is as follows :

Ly=max(x){Lx+Cx/y | x is compatible with y, x is strictly left of y, and x is not contained within y} (1)

The above equation is the equation used by PASA. A description of its implementation should elucidate how it works. First, align the transcripts. From here, Cx is easy to find and merging alignments can give Cx/y. Now start by scanning from left to right along the genome (plus all its aligned transcripts). The very first transcript hitting to the genome will be Lx. Count all alignments contained within Lx. Then slide over to where the next alignment, y, starts and you can begin using equation 1. Find all agreeing alignments (z,d,e,f, etc) contained within the linking piece, y. Then subtract out those found alignments that are also contained within x (to avoid counting duplicates). This is the equivalent to adding Cx/y to Lx. Note the requirements in equation (1), though. For instance, x must be strictly to the left of y and not contained within it. This avoids picking up non-agreeing alignments (as was seen in the example). However, every smaller alignment within and agreeing with a larger alignment, must agree with each other.

This process will potentially leave out many disagreeing alignments. An attempt is therefore made to pick up these disagreeing alignments and create a maximum assembly from them. Suppose a' is an alignment which was left out in the first maximal assembly attempt. A search to the left of it can

pick up La', the nearest agreeing assembly to the left of a. To extend this, a reverse scan is performed from right to left, giving Ra'. The two La' Ra' are merged and the overlap is subtracted out. In this way, each alignment is guaranteed to be included in a maximal assembly Lx.

Chimeric Assemblies and Alternate Splicing:

When the orientation of an alignment is not known (which is often the case for short sequences like ESTs), potential problems arise. For instance, if the ambiguous piece has sticky ends for two transcripts in opposite directions along the genome, these two transcripts can be incorrectly merged. To deal with this, PASA runs the algorithm by setting the orientations in one direction for one computation, and then doing a second computation for the other direction. The maximum assembly of all these computations is then taken to be the best one[7].

Where maximal assemblies of the same size contain different intron/exon structures, alternative splice can be inferred. For example, if in one assembly exon 1 ends at position 5, whereas in another maximal alignment exon 1 ends at position 10, it can be inferred that there exists an alternate splicing site which can either splice along position 5 or continue coding.

Therefore, it can be seen how a gene structure can be obtained from transcript to genome alignments. PASA has the convenient feature of also generating file which includes the best canidates for a training set. Generally, these best candidates are the largest assemblies. For details on how this file is generated, refer to the PASA homepage: <http://pasa.sourceforge.net/>

[7] This is generally a fair assumption because if an ambiguous alignment is in the incorrect orientation, it is less likely to agree with the alignment it is contained within (that is, it might align to the genome, but its intro/exon boundaries are less likely to match). In general, the very maximal assembly will include all consistent alignments.

1.5 Alignments:

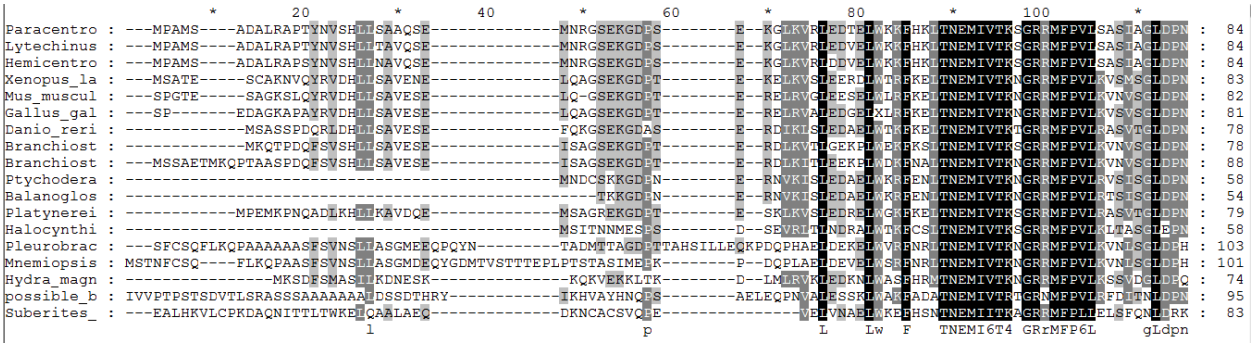
Our cloned Pb *brachyury* amino acid sequence is shown below:

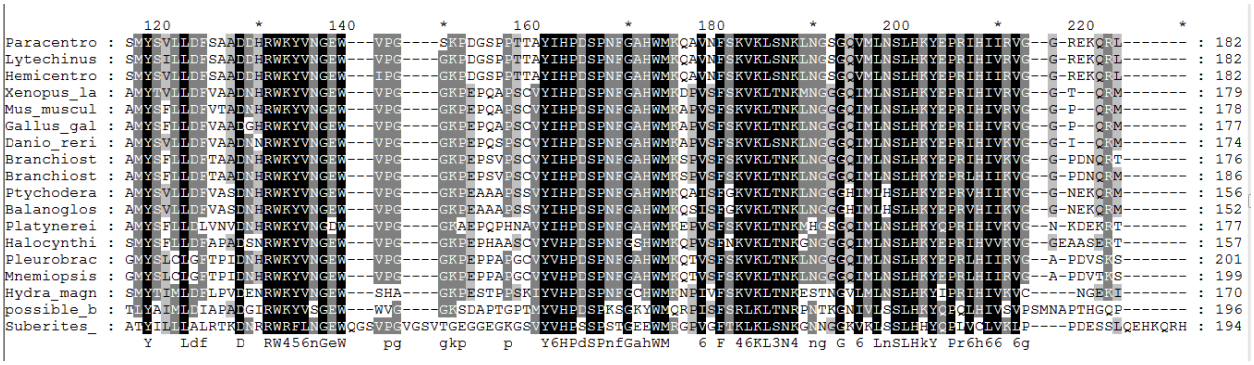
>Pleurobrachia Brachrury sb|11393705| scaffold677.1|size41701 #FGENESH: 4 10 exon (s) 32052 - 37351 370 aa, chain -
MGPLGIMLNSLHKYEPRLHIIRVGAPDVSKSVFTYSFAESRFIAVTAYQNEEITGLKIKYNPFAKAFLDAKERQEKDSHGQ
KREREDCGSTSKSLSSCSFYNRVAAGGHHADPERRPAKRVCLNSPNYAFPRHPALDLAGRVSSTTGGLHGFQGETAPTA
AAVPPNYSDVWTSPFTPHSESYWTSSYNTYSAALQPYLPVTSTYASPSSLVHPRVKEEPTEAPTVGTTVAMDLFSAVGD
AGRVSSTTGGLHGFQGETAPTAAAVAPNYSDVWTSPFTPHSESYWTSSYNTYSAALQPYLPVASTYTSPSSLVHPRVKE
EPTEAPTVGTTVAMDLFSAVGDGTEHDPSSYLPWSANFDSRRSVVEPSPQFP

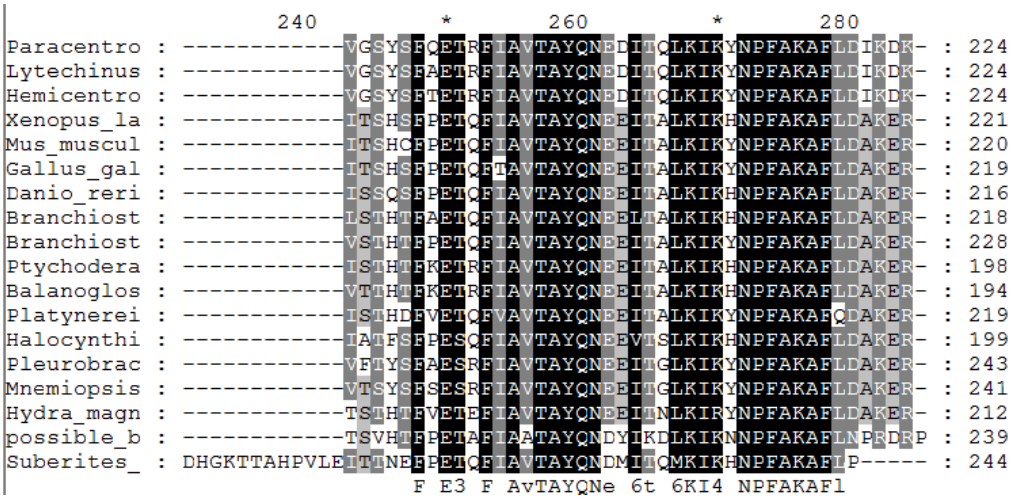
And the DNA sequence:

>sb|11408135| scaffold677.1|size41701
ATGGGTCCTCTAGGGATAATGCTAAATTCACTGCACAAGTACGAACCAAGACTCCACATAATTCGAGTTGGTGCCC
TGATGTGTCCAAGAGTGTTTTCACCTACTCGTTCGCTGAGAGCCGCTTTATAGCCGTGACTGCCTACCAAAACGAA
GAAATCACCGGTCTAAAGATCAAATATAACCCATTCGCTAAGGCTTTCCTCGATGCTAAAGAAAGGCAAGAGAAA
GATAGTCATGGACAGAAACGGGAGAGAGAGGATTGTGGGTCGACCAGTAAATCTCTTTCTTCGTGCAGTTTCTAC
AATCGTGTCGCAGCTGGAGGCCATCACGCCGATCCGGAGCGGAGACCTGCTAAACGAGTGTGTCTGAATAGCCCC
AATTATGCTTTCCCTCGCCATCCCGCACTTGACCTTGCTGGTAGAGTGAGCAGTACGACAGGAGGTCTACACGGGT
TTCAGGGAGAAACAGCACCAACTGCAGCAGCTGTACCACCTAACTACTCTGATGTTTGGACGTCCCCCTTTACACCT
CACTCAGAATCCTATTGGACTTCATCCTACAATACATACAGTGCAGCACTACAGCCGTACCTCCCGGTGACCTCAAC
CTATGCCTCACCATCCTCCCTTGTACATCCCAGAGTAAAGGAAGAGCCCACGGAAGCGCCAACGGTTGGTACCACA
GTTGCCATGGACTTGTTCTCAGCAGTTGGAGACGCTGGTAGAGTGAGCAGTACGACAGGAGGTCTACATGGATTT
CAAGGTGAAACGGCACCAACTGCAGCTGCTGTAGCACCCAACTACTCTGATGTGTGGACGTCCCCCTTTACACCTC
ACTCAGAATCCTATTGGACTTCATCCTACAATACATACAGTGCAGCACTACAGCCGTACCTCCCTGTGGCCTCAACC
TATACCTCACCATCCTCCCTTGTACATCCCAGAGTGAAGGAAGAGCCCACGGAAGCGCCAACGGTTGGTACCACAG
TTGCCATGGACTTGTTCTCAGCAGTTGGAGACGGTACCGAGCACGACCCCAGTTCCTATCTACCATGGAGCGCCAA
TTTTGACAGTAACGATCTGTGGTAGAACCTAGTCC CCAGTTTCCCTGA

The alignments used in making the gene tree are shown in GeneDoc format, with black representing the most conserved regions, and gray lesser conserved regions.







Figures 12-14. Showing alignments used in constructing gene tree. Pb sequence was not included in these alignments.

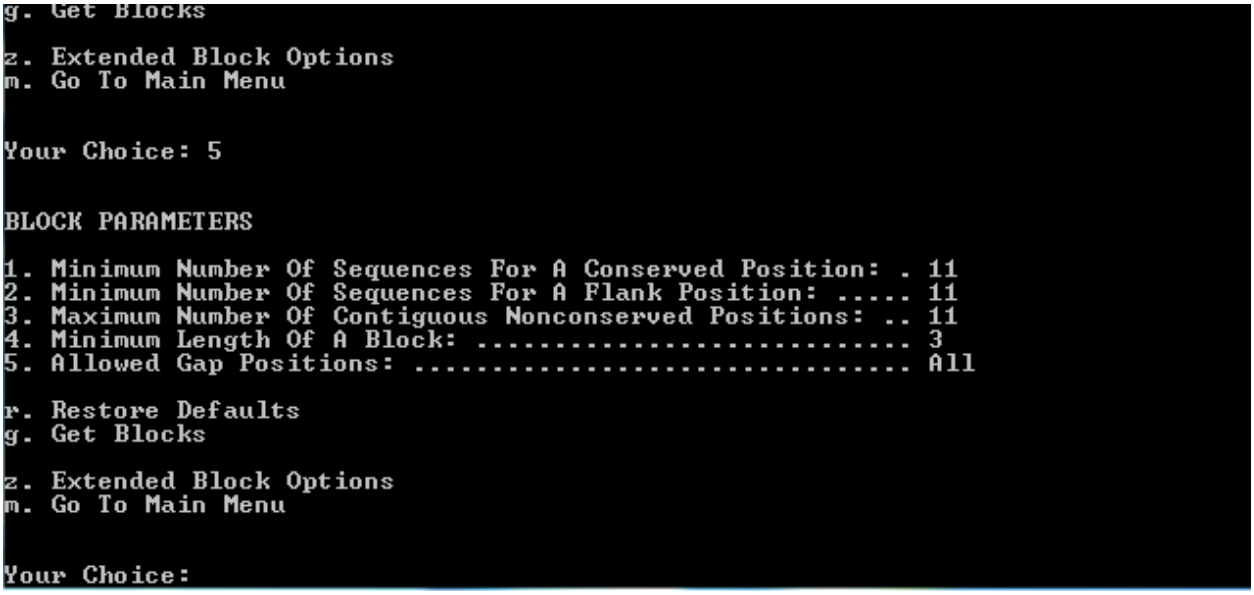


Figure 15. Screen shot of Gblocks showing blocking parameters used to obtain alignments.

1.7 Full in situ protocol

Day 1

Fix whole specimen in 4% paraformaldehyde in Filtered Sea Water (FSW) overnight (O/N) at 4^0C

Place no more than 10 animals in a 50 ml conical tube. To mix, hold on side and rotate gently.

4% Formaldehyde in Filtered Sea Water (5.4mL Formaldehyde in 44.6 ml FSW)

Fresh sea water is obtained from our tanks then filtered on 0.2 um pore membrane filter unit.

Day2

Rinse 3 x for 10 min in PTW (PBST) at Room Temperature

To mix, hold on side and rotate gently. Dispose of all solutions in hazardous waste.

PTW (PBST) 50ul Tween 20 in 50mL 1 x PBS (1 x PBS is 5 ml of 10x PBS in 45ml MQ H_2O)

Wash in 1:1 Methanol (MeOH)/PTW (to equilibrate to MeOH) 10 min at Room Temperature

1:1 MeOH/PTW (25 ml Methanol and 25 ml PTW)

Store in 100% MeOH at -20C for 2 hours up to a week

Place on side in freezer to allow animals to be separated

Day 3

Rehydrate specimen for 10 minutes in MeOH/PTW 3:1, 1:1, 1:3, 0:1 at Room Temperature

30mL/20mL/10mL MeOH fill to 40 ml with MQ H_2O

Wash in 1:1 solution of hybridization buffer (HB) and PTW for 15 minutes at Room Temperature

25ml PTW in 25 ml HB buffer

Incubate (prehybridize) in pre-HB buffer for 1 hours at 60^0C

*Hybridization buffer (HB) (50% formamide, 5mM EDTA, 5X SSC, 1X Denhardt solution (in 1.5 ml tubes at –20^0C) (0.02% ficoll, 0.02% polyvynilpirrolidon, 0.02% BSA), 0.1% Tween 20, 0.5 mg / ml yeast RNA (in–20^0C (Invitrogen) = (**for 50 ml)** 25 ml formamide, 0.5 ml 0.5M EDTA, 12.5 ml 20X SSC, 50 ul Tween20, 1 ml 50X Denhardt s-n, 25 mg yeast RNA)*

NOTE: pre-HB buffer does not contain Denhart or tRNA

Incubate (hybridize) in HB with DIG-RNA probe O/N at 60^0C

Add Denhardt and yeast tRNA to pre-HB buffer to make HB buffer. Mix 1 ml of hybridization buffer (HB) with 2-10 ul(200-400 ng Qubit) of probe.Then remove prehybridization buffer from tube with animals and add mixed hybridization buffer. Very gently rock O/N.

Day 4

Wash in pre-HB for 30 min at 60^{0}C

Remove old HB buffer and replace with 1 ml fresh pre-HB in same well

Wash in 1:1 HB/PTW for 30 min at 60^{0}C

Remove old HB buffer and replace with 1 ml 1:1 pre-HB/PTW in same well

Wash in PTW for 30 min at Room Temperature

Remove old 1:1 pre-HB/PTW and replace with 1ml PTW in same well

Block in 10% Goat Serum (GS) for 60 min at Room Temperature

Remove old PTW and replace with 1ml 10% Goat Serum in same well

1mL GS (in 1.5 ml tubes at –20^{0}C) in 10mL PBT

Incubate in anti-DIG 1/2000 at 4^{0}C O/N

Remove 10% Goat Serum replace with 1 ml 1% GS + 1:2000 (0.5 ul / 1 ml 1% GS) of alkaline phosphatase -conjugated DIG-antibodies in same well

0.1mL GS (in 1.5 ml tubes at –20^{0}C) in 10mL PBT

1:2000 (0.5 ul alkaline phosphatase -conjugated DIG-Antibody/ 1 ml 1% GS)

Day 5

Wash 4 x 30 min in PBS Room Temperature

Prepare 24 well plates label wells with marker the names of probes.

Make detection buffer and aliquot 1mL into clean well for each sample. When ready to develop, add 20uL of NBT/BICP mix until dissolved. Should be yellow in color! **NOW** add samples. **Put on ICE and cover with tin foil.**

*Detection buffer (100 mM NaCl, 50mM MgCl2, 0.1% Tween 20, 1mM levamisole, 100mM Tris HCl = (**for 50 ml**) 1 ml of 5 M NaCl, 2.5ml of 1 M MgCl(10g/50mlH_2O), 0.012g levamisole, 50ul Tween 20, 5ml of 1M Tris-HCl pH=8; adjust PH to 9.5 with 10M NaOH, filter the resulting solution)*

Do not keep detection buffer longer than 2 weeks or if it becomes cloudy.

Watch for appropriate color development

If crystals begin pooling non-specifically in the mesoglea, cut animal in half and shake gently to evacuate interior

Stop in 4% paraformaldehyde in MeOH

Wash in 4% paraformaldehyde in MeOH 30 min Room Temperature

Timing of this step depends on the strength of the signal and the background. This time is for high signal with low background

Change solution couple times

5.4mL Formaldehyde in 44.6 ml Methanol

Wash 3x 10 min in Ethanol (EtOH) Room Temperature

Begin mounting immediately, or store in 100% EtOH for up to 4 days at $4^{o}C$

Specific staining begins to fade after 4 days

References

Adell, T., & Müller, W. E. (January 01, 2005). Expression pattern of the Brachyury and Tbx2 homologues from the sponge Suberites domuncula. *Biology of the Cell / Under the Auspices of the European Cell Biology Organization, 97,* 8, 641-50.

Bielen, H., Oberleitner, S., Marcellini, S., Gee, L., Lemaire, P., Bode, H. R., Rupp, R., ... Technau, U. (December 01, 2007). Divergent functions of two ancient Hydra Brachyury paralogues suggest specific roles for their C-terminal domains in tissue fate induction. *Development (09501991), 134,* 23.)

Edwards, Y. H., Putt, W., Lekoape, K. M., Stott, D., Fox, M., Hopkinson, D. A., & Sowden, J. (January 01, 1996). The human homolog T of the mouse T(Brachyury) gene; gene structure, cDNA sequence, and assignment to chromosome 6q27. *Genome Research, 6,* 3, 226-33.

Forney, G. D. (n.d.). The viterbi algorithm. *Proceedings of the Ieee, 61,* 3, 268-278.

Haas, B. J., Rusch, D. B., Mount, S. M., Delcher, A. L., Wortman, J. R., Smith, J. R. K., Hannick, L. I., ... White, O. (October 01, 2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research, 31,* 19, 5654-5666.

Holland, P. W., Koschorz, B., Holland, L. Z., & Herrmann, B. G. (January 01, 1995). Conservation of Brachyury (T) genes in amphioxus and vertebrates: developmental and evolutionary implications. *Development (cambridge, England), 121,* 12, 4283-91.

Kispert, A., Herrmann, B. G., Leptin, M., & Reuter, R. (January 01, 1994). Homologs of the mouse Brachyury gene are involved in the specification of posterior terminal structures in Drosophila, Tribolium, and Locusta. *Genes & Development, 8,* 18, 2137-50.

Manuel, M., Le, P. Y., & Borchiellini, C. (January 01, 2004). Comparative analysis of Brachyury T-domains, with the characterization of two new sponge sequences, from a hexactinellid and a calcisponge. *Gene, 340,* 2, 291-301

Marcellini, S., Technau, U., Smith, J. C., & Lemaire, P. (January 01, 2003). Evolution of Brachyury proteins: identification of a novel regulatory domain conserved within Bilateria. *Developmental Biology, 260,* 2, 352-61.

Martinelli, C., & Spring, J. (January 01, 2003). Distinct expression patterns of the two T-box homologues Brachyury and Tbx2/3 in the placozoan Trichoplax adhaerens. *Development Genes and Evolution, 213,* 10, 492-9.

Martinelli, C., & Spring, J. (September 12, 2005). T-box and homeobox genes from the ctenophore Pleurobrachia pileus: Comparison of Brachyury, Tbx2/3 and Tlx in basal metazoans and bilaterians. *Febs Letters, 579,* 22.)

Minguillon, C., & Logan, M. (January 01, 2003). The comparative genomics of T-box genes. *Briefings in Functional Genomics & Proteomics, 2,* 3, 224-33.

Peterson, K. J., Harada, Y., Cameron, R. A., & Davidson, E. H. (January 01, 1999). Expression pattern of Brachyury and Not in the sea urchin: comparative implications for the origins of mesoderm in the basal deuterostomes. *Developmental Biology, 207,* 2, 419-31.

Rabiner, L., & Juang, B. (n.d.). An introduction to hidden Markov models. *Ieee Assp Magazine, 3,* 1, 4-16.

Scholz, C. B., & Technau, U. (January 01, 2003). The ancestral role of Brachyury: expression of NemBra1 in the basal cnidarian Nematostella vectensis (Anthozoa). *Development Genes and Evolution, 212,* 12, 563-70.

Sebâe-Pedrâos, A., de, M. A., Lang, B. F., Degnan, B. M., & Ruiz-Trillo, I. (January 01, 2011). Unexpected Repertoire of Metazoan Transcription Factors in the Unicellular Holozoan Capsaspora owczarzaki. *Molecular Biology and Evolution, 28,* 3, 1241-1254.

Stanke, M., & Waack, S. (January 01, 2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics (oxford, England), 19,* 215-25.

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (January 01, 2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *Bmc Bioinformatics, 7.*

Technau, U. (January 01, 2001). Brachyury, the blastopore and the evolution of the mesoderm. *Bioessays : News and Reviews in Molecular, Cellular and Developmental Biology, 23,* 9, 788-94.

Yamada, A., Martindale, M. Q., Fukui, A., & Tochinai, S. (March 01, 2010). Highly conserved functions of the Brachyury gene on morphogenetic movements: Insight from the early-diverging phylum Ctenophora. *Developmental Biology, 339,* 1, 212-222.

Yang, X. R., Ng, D., Goldstein, A. M., Parry, D. M., Alcorta, D. A., Li, S., Kelley, M. J., ... Sheridan, E. (November 01, 2009). T (brachyury) gene duplication confers major susceptibility to familial chordoma. *Nature Genetics, 41,* 11, 1176-1178.