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 3.14. 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 3.16, *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 3.16. 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

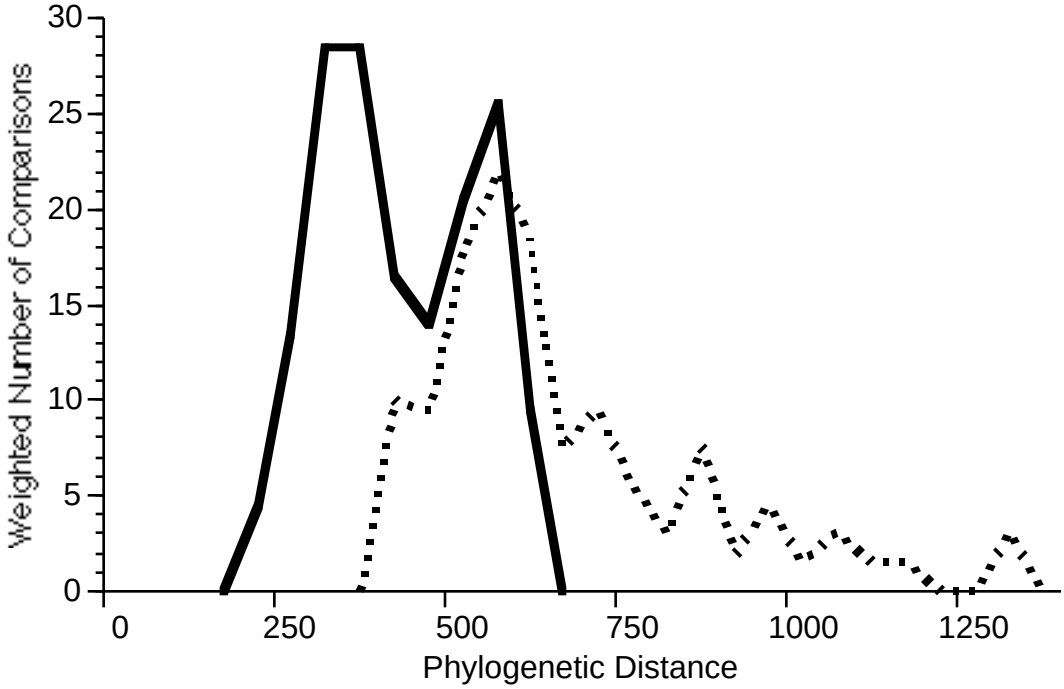

Figure 3.17. Weighted statistics of thirty sequences demonstrating distance differences between within-class comparisons (blue) and between-class comparisons (green). See text for a more complete description of methodology.

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 3.17). 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.[1] This finding is discussed in Section 3.19.

[1] 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.

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. 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 it 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 3.17, suggesting that this type of sampling bias is not present in the dataset of known trypsinogens (Figure 3.18). 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, 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

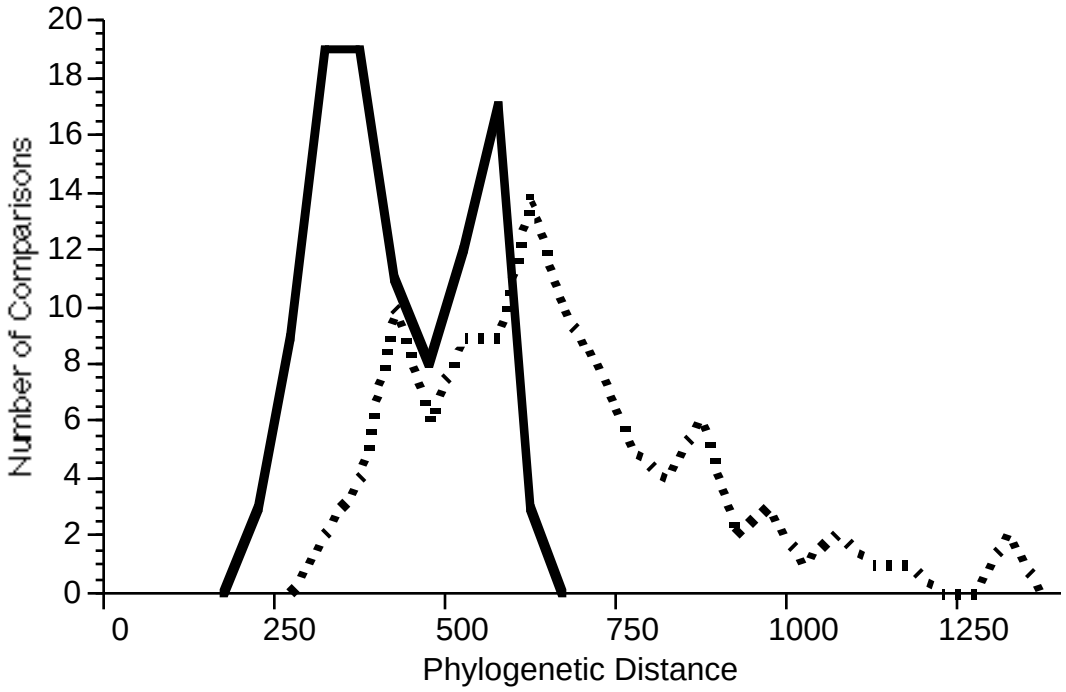

Figure 3.18. Unweighted statistics of thirty sequences demonstrating distance differences between within-class comparisons (blue) and between-class comparisons (green). See text for a more complete description of methodology.

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 3.16 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 3.16), 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 3.19). 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

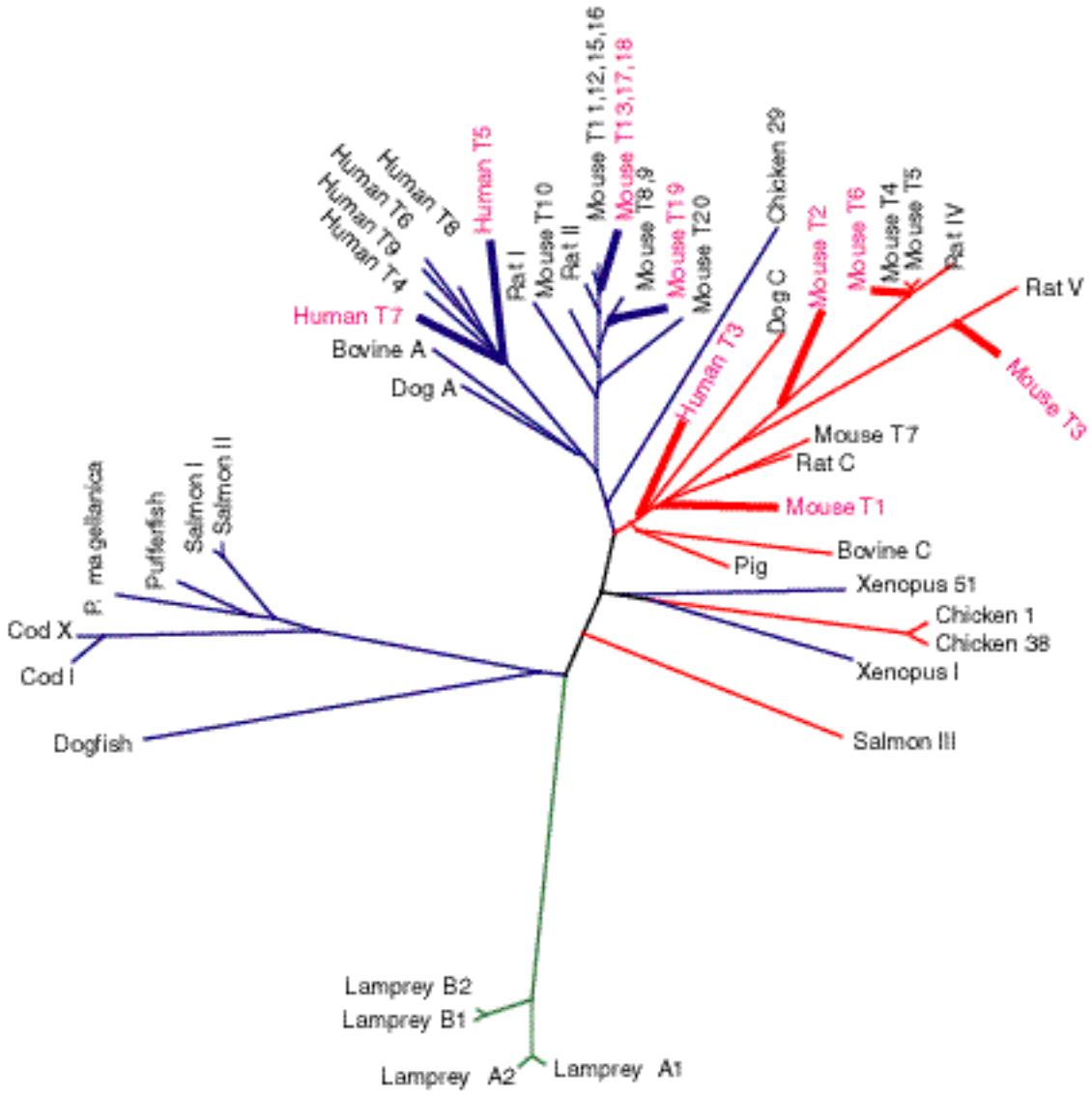

Figure 3.19. Pseudogenes added to the vertebrate trypsin phylogeny. Pseudogene names are in pink; pseudogene branches added to the phylogeny manually are bold. Group I branches are blue; group II branches are red. Branches in green diverged before group I and group II split. Black branches are indeterminate. The methodology of pseudogene addition is explained in the text. Pseudogene branch lengths are arbitrary; in this case, they have been selected for aesthetics.

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 3.19 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 3.17).

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 3.16, 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 are 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 3.20. 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 3.16 and Figure 3.20. This clade is found in 69% of the jackknifes executed with just the vertebrate data set, but only 43% of the jackknifes 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 platessa* sequence is not orthologous to group I or group II trypsins and therefore will not be subject to the same

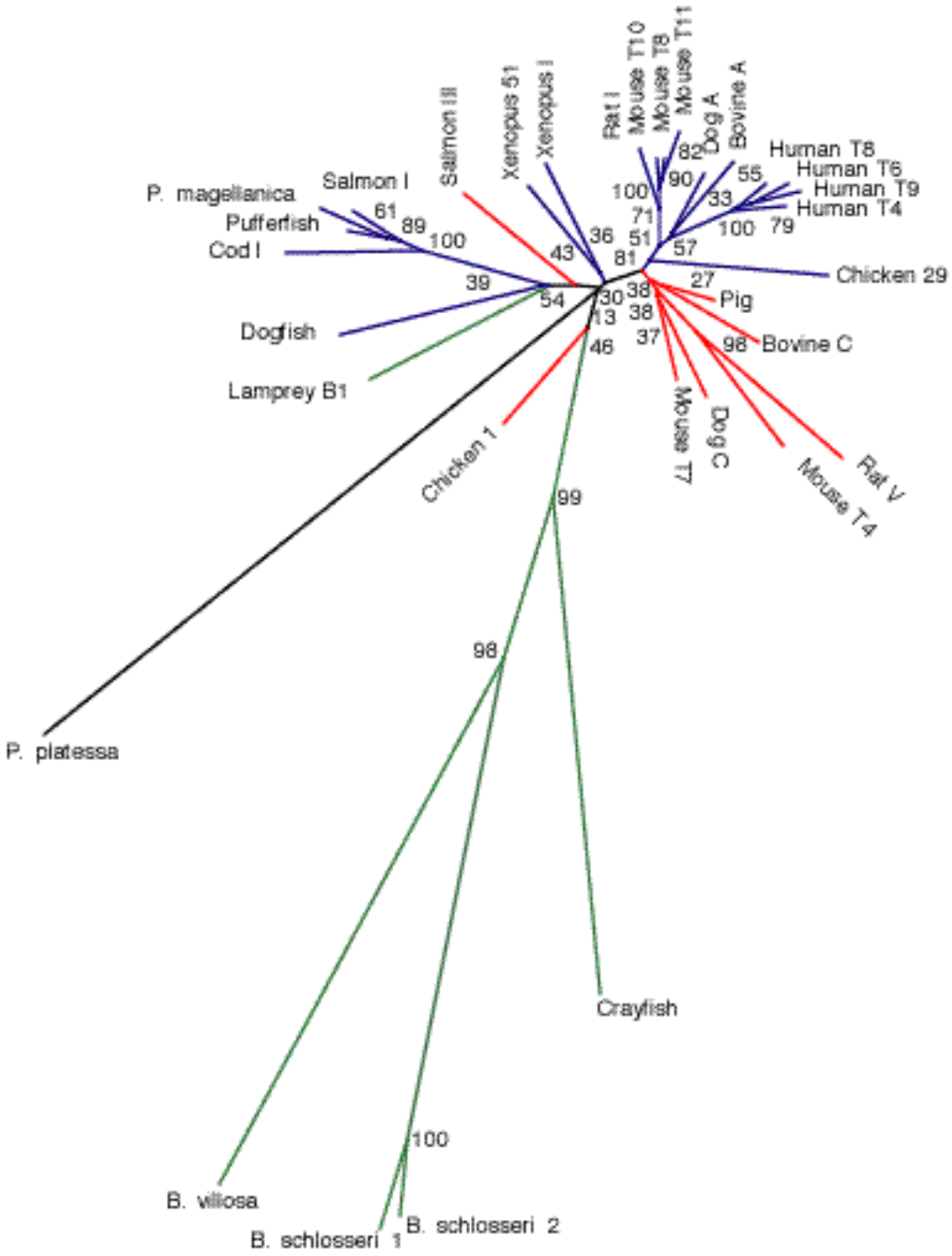

Figure 3.20. A Fitch-Margoliash phylogeny of thirty-two vertebrate trypsinogens, with the addition of five additional highly diverged sequences. The resulting phylogeny is highly skewed. The graph is calculated as described for Figure 3.15. The numbers adjacent to nodes represent the number of times 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. The *P. platessa* sequence is the originally submitted version, including several sequencing errors (see Section 3.3).

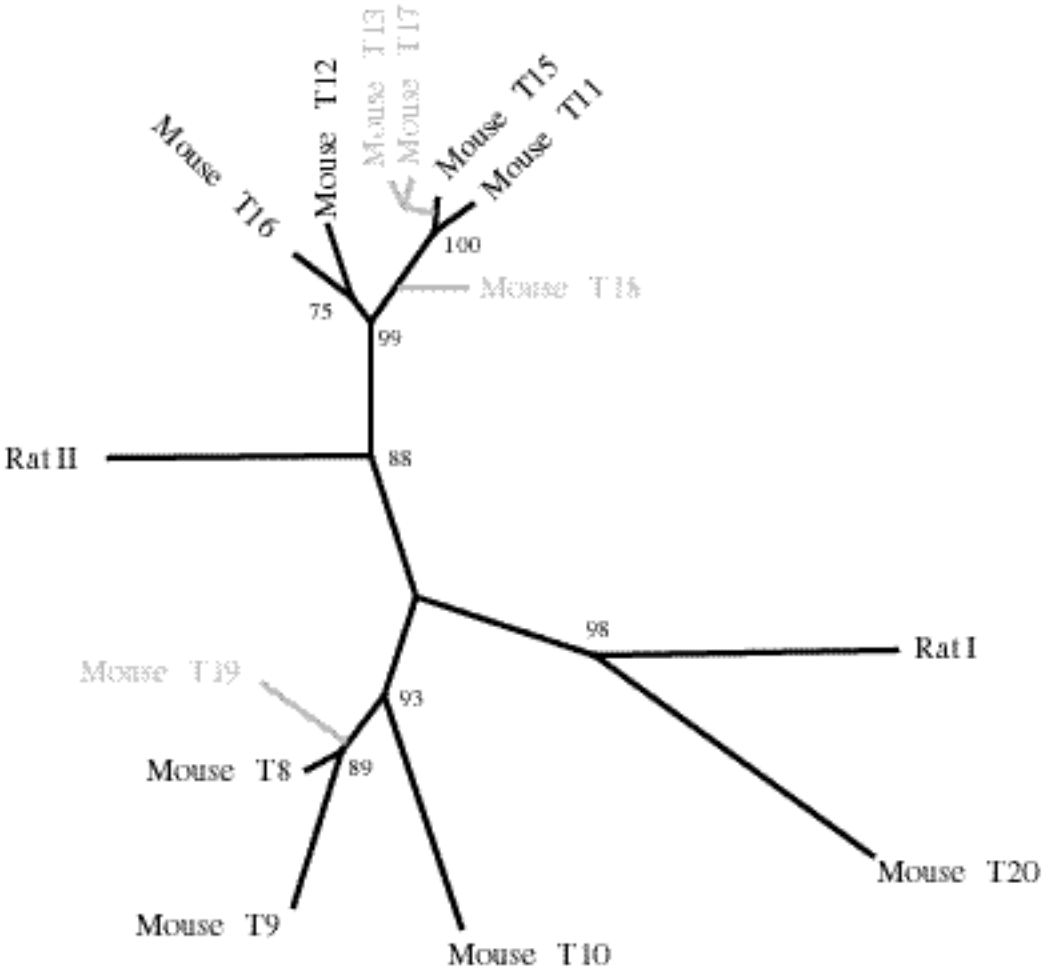

Figure 3.21. A phylogeny of the rodent group I trypsins. Constructed with the program DNAML with global rearrangements and 20 random "jumbles" (Felsenstein, 1993). Values at nodes are the result of 100 delete-half jackknifes of the original data (computed with global rearrangements but only one random "jumble"). Pseudogenes are in gray and were subsequently added under the protocol outlined for Figure 3.19.

selective pressures as the trypsins (see Subsection 3.18.2). Rypniewski et al. (1994) include the *Pleuronectes platessa* sequence in their phylogeny, quite possibly causing just such a distortion. Independently, Male et al. (1995) include the *Pleuronectes platessa* sequence in a phylogeny, again contributing to a skew. Both sequencing errors in the 1994 *Pleuronectes platessa* and its alterations in its function and therefore underlying selective pressure probably contributed to 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 3.16 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 Osteicthyes 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 3.21. 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.

3.18.1 Recently Sequenced Trypsinogens[2]

Since the first printing of this thesis, several additional vertebrate trypsinogen sequences have become available. I have produced a hand-edited multiple alignment including these sequences (Figure 3.22).

---

[2] This section is new in the second printing, and modified in the third printing.

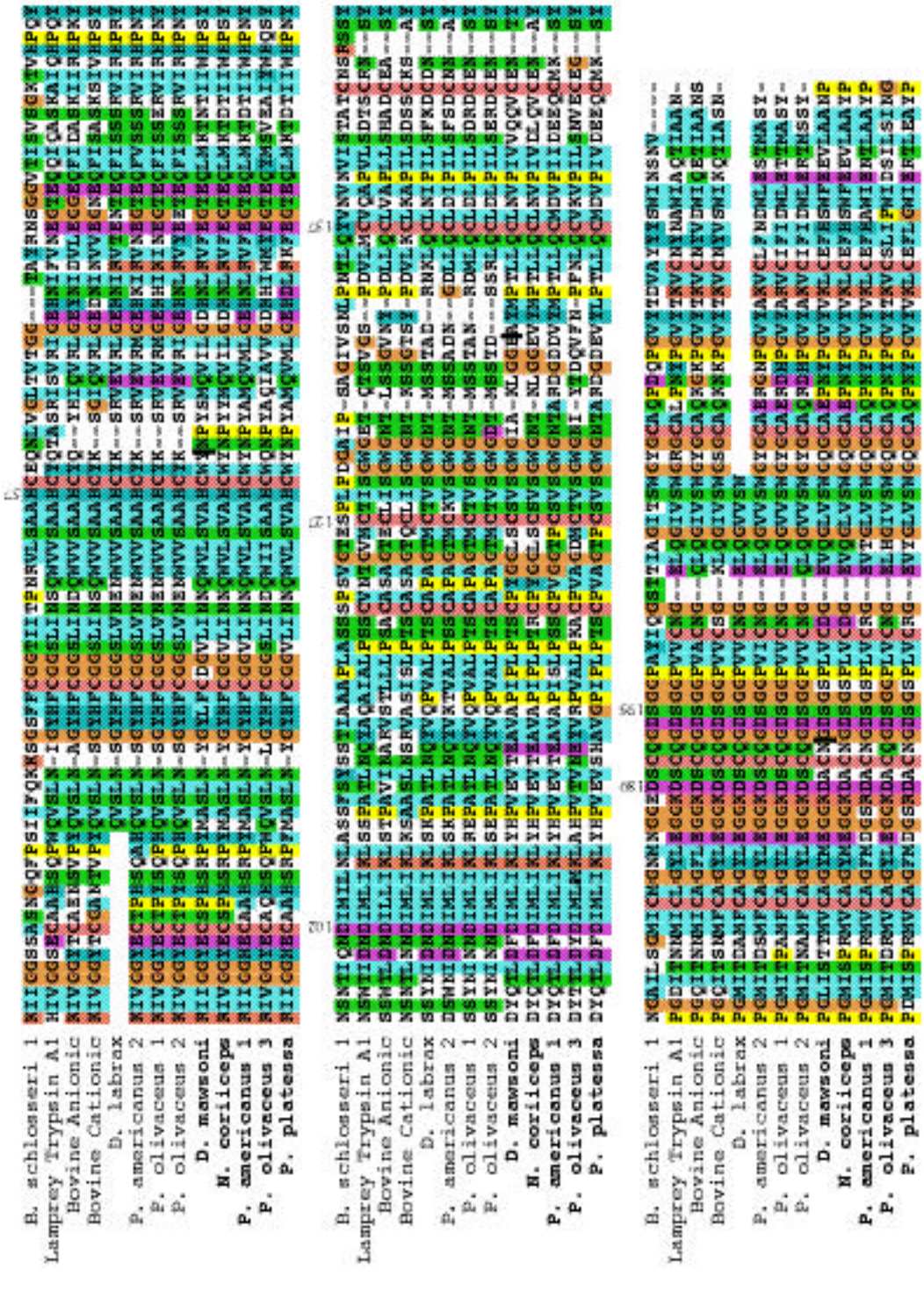

Figure 3.22. A hand alignment of recently identified trypsin sequences, together with several reference sequences. Psychrophilic trypsin names are in bold. The colors are the default biochemical scheme of *CLUSTAL W*. Introns/exon boundaries, where known, are indicated by a bar ().

The first, submitted August, 1998, is a cDNA from the winter flounder, *Pleuronectes americanus* (Douglas and Gallant, 1998). The authors named this cDNA "trypsinogen 2 precursor".[3] The authors have submitted two other sequences, named "trypsinogen 1 precursor" and "trypsinogen 3 precursor". which are not trypsinogens.[4] *Pleuronectes americanus* trypsinogen 1 is a cold-adapted trypsinogen (Section 3.18.2). *Pleuronectes americanus* trypsinogen 3 is not a serine protease, as it lacks key active site residues, and is therefore mis-named. The nomenclature of these two sequences was based in part on homology to the *Pleuronectes platessa* sequence.

The winter flounder trypsinogen 2 signal peptide is identical to that of *P. magellanica*. The activation peptide (LEDDK) differs by T**m**L and E**m**D substitutions. The mature trypsin has a predicted pI of 5.13 and charge at pH 7.0 of -6.98 (EMBL isoelectric point service). This is consistent with a group I classification. Southern hybridization reveals multiple genomic bands.

The second trypsinogen, a direct submission in October, 1998, is from the European sea bass, *Dicentrarchus labrax.*[5] It is a partial sequence.

The third and fourth new trypsinogens are from the bastard halibut, *Paralichthys olivaceus*, one of several commercially valuable Japanese flounders. These were submitted in July 1999. The two *Paralichthys olivaceus* trypsinogens are "trypsinogen 1" and "trypsinogen 2".[6] *Paralichthys olivaceus* "trypsinogen 3" is a cold-adapted trypsinogen (Section 3.18.2).[7] The initial sequence submitted to Genbank, AB029752.1, lacked a key cysteine for one of the cystines, so on my suggestion, the authors located errors in their raw sequence data that had produced a localized frameshift. This was corrected on 1/29/00 to Genbank entry AB029752.2. However, this sequence has a number of other residues not found in its clade-mates; the possibility of additional raw sequence errors remains.

A fifth new trypsinogen is the partial sequence from *Bothrops jararaca*, a South American pit viper.[8] It was submitted in October 1999. This sequence is not shown in Figure

---

[3] Genbank AAC32752.

[4] Genbank AF012462, AF012464.

[5] Genbank CAA07315.

[6] Genbank AB029750.1, AB029751.1.

[7] Genbank BAA82364.1.

3.21. However, a cursory analysis indicates that it is more similar to the group II trypsins than to the group I trypsins. If the full sequence fully supports this observation, then this fragment is the first sequenced example of an amphibian group II trypsin. Therefore, with this sequence, there are now examples of both group I and group II trypsins from all major vertebrate clades, with the notable exception of Agnatha.

A phylogeny created with *fitch* from Dayhoff-parameterized *protdist* is presented in Figure 3.23 (Felsenstein, 1993). This phylogeny includes the three new complete trypsinogens with a representative sample of the trypsinogens discussed in the main body of this Section, leaving out only single copies of nearly identical sequences. The cold-adapted trypsinogens are also included in this phylogeny (Subsection 3.17.2).

The three new full-length Osteicthyes trypsins are clearly members of group I and cluster with the other Osteicthyes trypsins. They offer no surprising new cladistic insights, but do confirm the distinct clustering of the group I Osteicthyes trypsins. The *Dicentrarchus labrax* trypsinogen fragment also clusters with the group I Osteicthyes trypsins (data not shown). The *Bothrops jararaca* trypsinogen fragment is too short for robust phylogenetic analysis.

3.18.2 Cold-Adapted Trypsinogens[9]

A *Pleuronectes platessa* "trypsinogen" sequence was a direct submission to Genbank in 1994 and has never been documented in any publication. It was named trypsinogen as it was more similar to the known trypsinogens than to any other sequence. I and others dismissed its importance, believing it to most likely be the result of sequencing and possibly PCR or cloning errors. In 1996, the sequence was updated. Also, additional sequences orthologous to the *Pleuronectes platessa* trypsinogen became available, thus ruling out the possibility of a cloning error. To date, four such orthologous sequences are known, for a total of five sequences forming a new clade.

Initially I hypothesized that these new sequences had an altered P1 substrate specificity due to changes in their active-site sequence, discussed below. This specificity pocket more resembles that of granzyme A than any other serine protease specificity pocket due to the Q192N and G197S differences. The P1 substrate specificity of granzyme A is for

---

[8] Genbank AF190273.1.

[9] This subsection is new in the third printing.

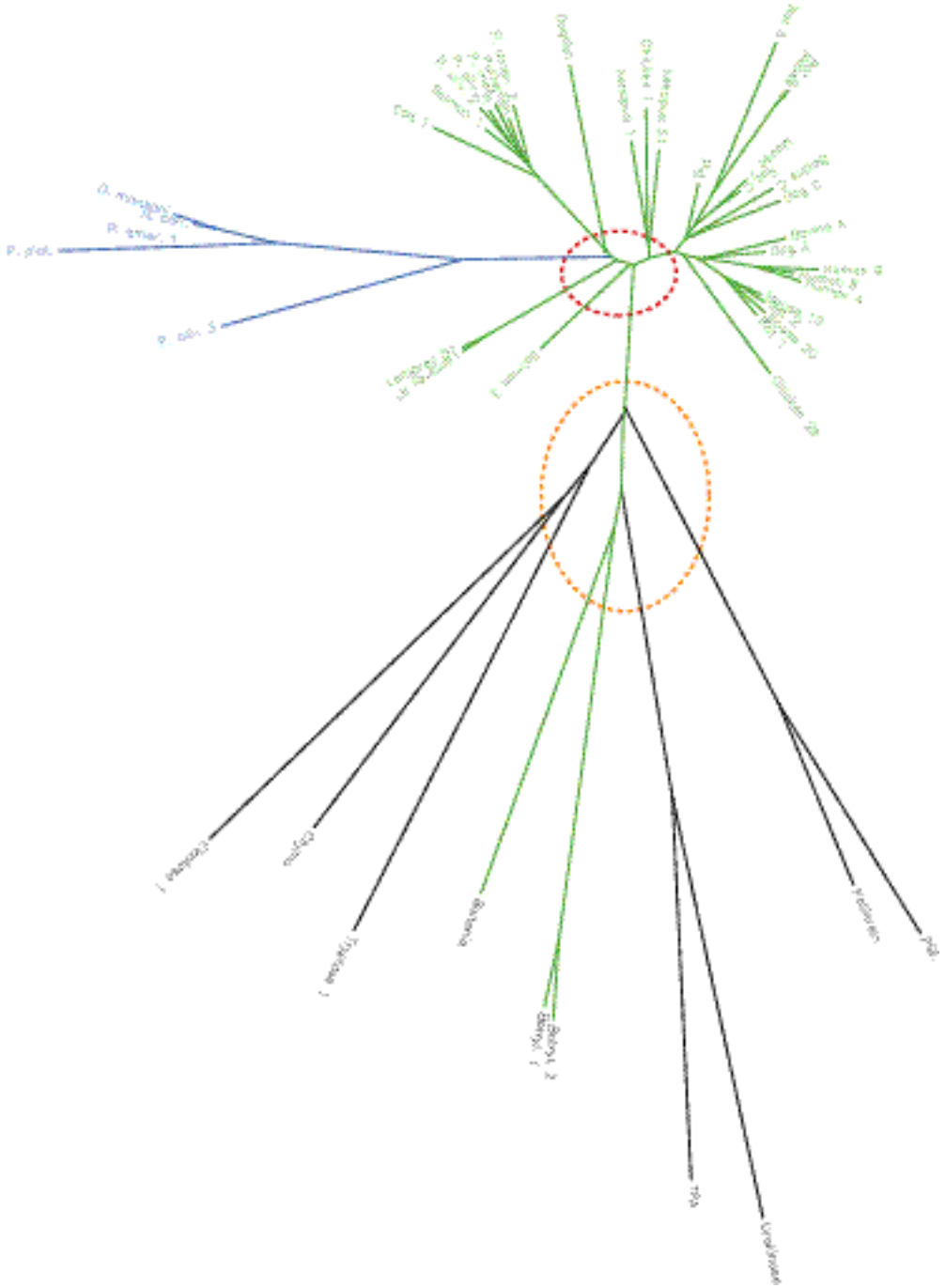

Figure 3.23. A Fitch-Margoliash phylogeny of thirty-five chordate trypsinogens, together with the five known "cold-adapted" trypsinogens, and seven representative human serine proteases. The arrow indicates a putative location for an ancestral root. The branches originating in the dashed region have weak statistical support for their topological ordering.

lysine and arginine, although phenylalanine can also be recognized (Enzyme Nomencalture database).[10] Therefore my hypothesized differences in substrate specificity may actually have been relatively minor.

In 1997, a sequence from the giant Antarctic toothfish *Dissosthicus mawsoni* was determined by PCR and identified as a trypsinogen based on its resemblance to the *Pleuronectes platessa* sequence (Chen et al., 1997). Subsequently, the mRNA and genomic sequences of this gene were determined (Cheng and Chen, 1999).

In August 1988 the orthologous mRNA from the winter flounder *Pleuronectes americanus* was sequenced and labeled “trypsinogen 1”.

In May 1999 the orthologous mRNA from the black rockcod *Notothenia coriiceps* was sequenced.

In July 1999 the orthologous mRNA from the bastard halibut *Paralichthys olivaceus* was sequenced and labeled “trypsinogen 3”.

Clearly, these five sequences together form a new clade of mutual orthology. These sequences are therefore paralogous to the other vertebrate trypsinogens. Following a conversation with Charles Craik, as well as several email exchanges with Chris Cheng-DeVries and Susan Douglas, it seems most likely that these are functionally active trypsins, and therefore are correctly classified with an enzyme classification number of 3.4.21.4. The most likely explanation for the strong selective pressure responsible for this new clade is that these enzymes are highly adapted to extremely cold temperatures. For this reason, I have named them, at least for my own purposes, “psychrophilic trypsins”. To support this claim, I have considered common changes in the primary sequence of the psychrophilic trypsins. Conversations with Charles Craik, Michael Levitt, and David Baker helped focus my thoughts.

Differences within the nearly absolutely conserved sequence near the active-site serine are particularly interesting, as they may indicate a difference in substrate specificity (Perona and Craik, 1995). A change not observed in any other naturally occurring trypsinogen occurs at position 190. However, this S190A difference probably does not substantially alter the function of the enzyme. Evnin et al. (1990) and Perona et al. (1993) have characterized a double D189E/S190A mutant. This mutant showed an increased preference for substrates

[10]http://www.chem.qmw.ac.uk/iubmb/enzyme

Table 3.10: Psychrophilic trypsin residue differences at highly conserved positions[a]

Position[b]	Mesophilic residue	Psychrophilic residue	Psychrophilic exceptions[c]	Mesophilic exceptions[c]
15	**K**	**R**		Rat 5
~30	**Q**	**M**	*P. olivaceus* 3	
	I,A,S	**Y**	*P. olivaceus* 3	
45	**S**	**V**	*P. olivaceus* 3	
55	**A**	**V**	*P. olivaceus* 3	
	Q,K	**Y**	*P. olivaceus* 3	
	S	**Y**		Bovine A
~59-63	**WYNPYAM**	**YK--SRI**		(lamprey)
	S	**Y**		
101	**N**	**F,Y**		
	S,K,R	**Y,A**		
~115	**N**	**T**		
	A	**P**		
	N,S,P	**Y,V**		
~136-150	**(indel)**	**(indel)**		

with a lysine at their P1 site. Wild type D189/S190 trypsin has a lysine:arginine preference of 4.3; the D189E/S190A mutant has a ratio of 1.3. The mutant also shows a decrease in k_{cat} and an increase in K_m. Evnin et al. (1990) and Perona et al. (1993) generally conclude that a negative charge at either the 189 or 190 position is critical for recognition of a positively charged lysine or arginine in the substrate P1 site. In psychrophilic trypsins this negative charge is maintained by the absolutely conserved glutamate at position 189. Although no trypsin mutant that has the S190A mutation without the D189E mutation is known, a mutant with a sole S190A mutation is likely to be qualitatively similar. The differences Q192N and G197S have not been observed in any other trypsins, but are minor biochemical changes and are again unlikely to result in major alterations in enzyme specificity.

Most of the other differences between mesophilic and psychrophilic trypsin are also predicted to result in minor biochemical changes (Table 3.10). However, there are a large

number of positions that have aromatic residues in the cold-adapted trypsins not present in the mesophilic trypsins, as well as an increase in the number of hydrophobic residues. This suggests a global adaptation to cold temperature, perhaps by increasing the overall number of hydrophobic contacts. There appear to be additional prolines and glycines, suggesting increased flexibility, which is a mechanism postulated to be important for enzyme adaptation to cold. I am currently pursuing efforts to model the various trypsin isozymes, both physically and virtually.

The data in Table 3.10 suggest that *P. olivaceus* 3 may be an intermediate form, as it shares specific residues in cases with the mesophilic trypsins, and in cases with the psychrophilic trypsins. The intermediacy of this trypsin may stem either from evolutionary history or from ecological niche and resulting temperature selection pressure.

A phylogeny created with neighbor joining from *neighbor* is presented in Figure 3.24 (Felsenstein, 1993). An original idea in creating this phylogeny was to explore the idea that the psychrophilic trypsins might have novel specified, and in such, might be more closely related to another serine protease than to the trypsins. However, the psychrophilic trypsins clearly form a clade with the other trypsins. They also are clearly monophyletic. One is thus led to speculate on when the psychrophilic trypsins diverged from the other trypsins.

Initially, based on sequence similarity, the divergence of the psychrophilic and mesophilic trypsins would seem to be quite old. If the molecular clock were to hold universally for all chordate trypsins, then the psychrophilic trypsins would have diverged at the dawn of vertebrate evolution. However, we have observed in other cases, that the molecular trypsin clock is not universal (e.g., Rat IV and Rat V). Furthermore, we believe there to have been strong selection pressure on the psychrophilic trypsins. Given a high selection pressure, the cold-adapted trypsins might very well have diverged quite recently from the other trypsins. For example, certain antifreeze proteins are known to have evolved within the last few tens of millions of years in the Antarctic notothenioids. Psychrophilic trypsins might well have done the same.

However, the five known psychrophilic trypsins come from a wide variety of species, not just the notothenioids. *N. coriiceps* and *D. mawsoni* are both notothenioids, and diverged about 25 million years ago (Gon and Heemstra, 1990). Both of these fish seldom visit waters warmer than freezing. *P. platessa* and *P. americanus* are Arctic-boreal fishes, and travel between sub-zero Arctic waters and warmer waters (Demel and Rutkowicz, 1958). It is interesting to note that, considering the known trypsins, the notothenioids possess only a

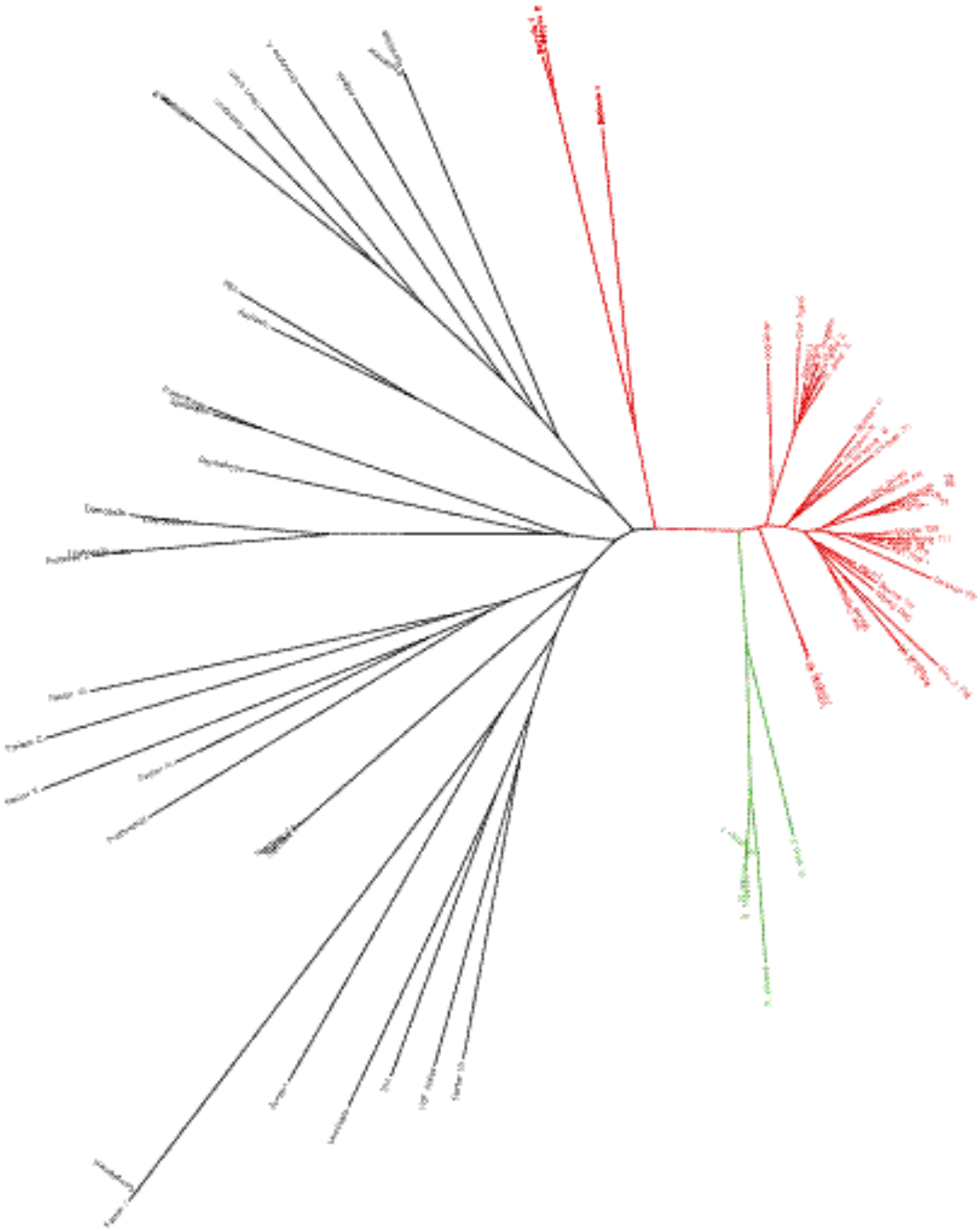

Figure 3.24. A neighbor-joining phylogeny of the trypsinogens and human serine proteases.

psychrophilic trypsin, and the Arctic-boreal fish appear to have both a mesophilic and psychrophilic trypsin. This suggests that the migratory fish may adapt to cold digestion by regulating expression of the different trypsins. This must remain a hypothesis until more complete data is obtained on the presence or absence of each trypsinogen gene in more species. The divergence of modern Percomporph families probably occurred in the late Cretaceous, about 98-65 million years ago (Eastman, 1991). It is therefore probable that psychrophilic trypsins have been around for at least this long, if not longer.

Intriguingly, there have been several studies on the adaptation to cold of trypsin isozymes. These have focused on trypsins from the Atlantic salmon, Greenland cod, and *P. magellanica*, all on isozymes within the clade that I refer to as "mesophilic" (Smalås et al., 1994; Male et al., 1995; Genicot et al., 1988; Genicot et al., 1996; Simpson and Haard, 1984a; Simpson and Haard, 1984b). There have been a number of studies on biochemically isolated trypsins as well, but without sequence data (e.g., Hofer et al., 1975). Some moderate differences between these trypsins and mammalian trypsins were identified in these studies, with the implication that these differences were important for cold adaptation. This is likely to be true. However, if my hypothesis is correct, then the trypsins considered in these studies were mesophilic for the organisms a in question, and only psychrophilic by comparison to mammalian trypsins. These trypsins may have been the first to be identified if they were isolated from fish most recently living in non-freezing water, as might be the case if they were caught in warmer waters, or if they were kept in a tank before sacrifice. My prediction is that the true psychrophilic trypsins will have more dramatic structural changes for adaptation to extreme cold.

There are several proposed mechanisms by which enzymes might adapt to cold (Low et al., 1972; Feller and Gerday, 1997). These include adjustment in enzyme concentration, change of expressed isozyme, decrease in activation energy ΔG^*, increase in specific activity or turnover number k_{cat}, and increased K_m. Enzymes not limited by substrate concentration, such as trypsin amidst a food-filled gut, might have a particularly high K_m, as binding substrate would not be limiting for reaction rate. In particular, one might imagine that substrate binding might be much weaker for a psychrophilic trypsin. The changes observed at the P1 site in the psychrophilic trypsins are consistent with this prediction. Psychrophilic enzymes are also predicted to be more flexible than their mesophilic counterparts. An increase in prolines and glycines in the psychrophilic trypsins may produce just this effect. Hydrophobic contacts are weakened at lower temperatures. An increase in large hydrophobic residues such as tyrosine may be compensation. Michael Levitt first pointed out to me the large number of tyrosines in the psychrophilic clade of trypsins. The psychrophilic trypsins

may also have several other adaptations such as in the number, strength, and location of salt bridges.

3.19 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.[11] 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 families. However, statistically significant exchange did occur, as documented in Figure 3.17.

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 3.15 & Figure 3.16). 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. Selective pressure due to body temperature may be especially important, and may in particular explain coincidental evolution within ectothermic and endothermic groups. Clearly, adaptation to cold played a major role in the evolution of the psychrophilic trypsins (Section 3.18.2).

The molecular traces of 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

[11]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 3.17) would have to be explained by coincidental evolution.

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.[12] 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

[12]As an addition to pun-derived terminology, I propose the word "trypsinogene" to designate these trypsinogen "alleles."

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 event. They may also provide a genetic reservoir of material for recombination events with functional trypsinogens. It would seem that if this were a major, or even an important minor, mode of evolution that sooner or later the active site serine would mutate from TCN to AGY. Since this has not happened, one must assume that psuedogene gain-of-function mutations must be rare.

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 3.16 and Section 3.18.

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.

3.20 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 3.11). 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.[13] All 64 human T4 and all 94 human T8 cDNAs are pancreatic in origin.[14] Of the 14

---

[13]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 sequenced 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.

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.[15] 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 consistent with an immunological

[14]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.

[15]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.

Table 3.11. 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

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 3.12).[16] 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 3.11 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.

3.21 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 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

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

Table 3.12. Genbank accession numbers and tissues of origin for the mouse trypsinogen ESTs, as of October 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

has no role in such cells, for a percentage of peripheral T-cells maintain an unrecombined TCR β locus.[17] 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.

3.22 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.

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

[17]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.

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, the cold-adapted trypsinogens discussed in Subsection 3.18.2 are almost certainly 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 the giant Antarctic toothfish, *Dissosthicus mawsoni* evolved by replacing the center region of a psychrophilic trypsinogen gene with a repeated nine-nucleotide element encoding hydrophobic residues (Cheng and Chen, 1999; Chen et al., 1997b). It retains the trypsinogen signal sequence and 3' untranslated region. The resulting gene bears little resemblance to a serine protease, but clearly evolved from a cold-adapted trypsinogen gene. The psychrophilic trypsin, in turn, is likely to have a mesophilic trypsinogen gene as its ancestor, thus participating as both a child and a parent in a multi-generational spectrum of trypsinogen-derived genes. Antifreeze proteins were probably not necessary in the warm waters of the late Cretaceous and early Tertiary periods about 100 million years ago. Therefore, antifreeze proteins most likely evolved between 5 and 65 million years ago (Gon and Heemstra, 1990).

At least some Arctic and Antarctic antifreeze proteins evolved independently (Duman and DeVries, 1974; Chen et al., 1997a). Logsdon and Doolittle (1997) provide a nice commentary on the antifreeze proteins.

It might initially appear that there is no particular reason for an antifreeze gene to have evolved from a trypsin gene. There may be some justification for such an event, however. First, since trypsins exist as a multigene family, a spare gene is free to mutate without abolishing function of the other isozymes, limiting a change in fitness of the organism. Second, since the antifreeze protein emerged from a psychrophilic trypsin, it would already have a regulatory mechanism for expression in cold environments. Thus only function would need evolve, not regulation. This might facilitate evolution in short geological time. Chen et al. (1997b) have previously noted that the signal sequence of trypsin permits secretion of the antifreeze protein and that it may have evolved first to prevent freezing of the contents of the digestive tract.

Evidence suggests that there may have been a major phase of gene duplication at the dawn of vertebrate evolution (Holland and Garcia-Fernàndez, 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 cold-adapted trypsinogen. If this were the case, one would expect all vertebrates to have essentially the same complement of genes. Specialty genes, like the cold-adapted trypsinogens, might form an exception to this rule. 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.