APPENDIX B
ALTERNATIVE SEQUENCE DISTANCE METRICS

There are limitations to all current distance metrics. The predominant distance metric used throughout Chapter 3 is based on the Dayhoff PAM matrix. This metric was used primarily for simplicity and speed of calculation, as it is an option of the *protdist* program, which is part of the PHYLIP package (Felsenstein, 1993). During the course of my efforts to grapple with multigene family protein evolution, I sought better methods to evaluate phylogenetic distances. Although it is not clear that I have suceeeded, or even made a step in what might ultimately be the right direction, I have included this Appendix describing my preliminary efforts.

B.1 SEQUENCE DISTANCES

A simple way to compare two sequences is to count the number of residues they share in common. This determines their percent identity:

$$\text{Percent Identity} = \frac{\text{number of identical residues}}{\text{total residues}} \tag{B.1}$$

Other methods compute values for similarity. In this case, a similarity value is assigned to each possible combination of two residues. Residues that are more likely to have replaced each other over the course of evolution have higher similarity values. For example, aspartate and glutamate, which are both acidic, are quite similar. There are many methods of computing sequence similarities and distances.

The distance between two sequences reflects the number of mutations that, on average, under conditions of natural selection, will convert one sequence into the other. If the mutation rate is constant and the conditions of natural selection are unchanging, then the number of mutations can be related to the amount of time separating the two sequences.

Mutations are typically assumed to operate as single base pair substitutions at the level of nucleotides. However, other types of mutations are possible. Some models allow for insertions and deletions of base pairs. A few models account for gene conversion events, but these suffer from lack of empirical data. Gene conversion is often considered to be such a rare event that it can be ignored as a mutational mechanism, despite the large evolutionary impact a single gene conversion event can have. Multigene families are particularly susceptible to gene conversion, and so may provide a significant challenge to models that ignore gene conversion.

There are multiple requirements that must be met for sequence distances to be interpreted as a measure of the time separating two sequences. The hypothesis that sequence distance can be interpreted as time is called the "molecular clock hypothesis." In some cases the molecular clock hypothesis holds, but in many cases it does not.[1] When a molecular clock can be used, molecular sequence data provides a powerful source of information on species phylogeny. It is important to distinguish between those cases in which the molecular clock is useful and those in which it is misleading. Making this distinction will be an important part of the analysis of the trypsinogens presented in this chapter.

All molecular sequence data is contemporary.[2] Therefore, if two sequences are related, it is not because one evolved from the other, but because they both arose from a common ancestor. If two contemporary sequences shared a last common ancestral sequence 600 million years ago, then evolutionary processes have been operating to diverge those sequences for a net 1.2 billion years of evolution. To rephrase, 1.2 billion years of divergent evolution separate the two sequences.

It is generally assumed that evolution is symmetric. This implies that the evolutionary distance separating a modern sequence from its ancestor would be calculated identically if the sequences were switched. This assumption allows one to unambiguously calculate the distances between two sequences without knowing the position of the last common ancestral sequence in the phylogeny. Seldom does this assumption hold absolutely. However, it almost universally holds well enough to be a useful tool. Without this assumption, molecular distance calculations become wickedly complicated.

[1] Several papers in the first issue of the Journal of Molecular Evolution address this topic, including Dickerson (1971), Kimura and Ota (1971) and Ota and Kimura (1971). Dickerson provides a chart showing the rates of evolution for cytochrome c, hemoglobin, and the fibrinopeptides. A molecular clock hypothesis holds for each of these proteins, according to the original data in these publications.

[2] Barring discovery of intact genetic material, it is not possible to obtain sequence from extinct species. In those rare cases where such sequence has been obtained, it is often incomplete. Furthermore, such sequences are seldom more than a few thousand years old, which is an almost insignificant period compared to evolutionary time scales.

Strictly speaking, it is not necessary to calculate pairwise distances between sequences in order to construct a phylogeny. In fact, in an ideal situation, it is not even desirable to do so. Information is lost when a data set of multiply aligned sequences is reduced to a diagonal matrix of pairwise distances. Some of this lost information is valuable not only in constructing the topology of a phylogeny but also in determining the evolutionary distances between sequences. Methodologies that employ maximum likelihood are most capable of properly utilizing all multiple alignment information. Such methodologies are computationally intensive. Additionally, no maximum likelihood computer programs are available capable of employing algorithms optimized for trypsinogen phylogenies. Even if such a program were to be written, it is likely that, today, the computer resources required to do a full trypsinogen analysis would exceed those available to trypsinogen researchers. These circumstances are likely to change in the future, both as novel programs are written and as computing power grows. In the meantime, to produce a trypsinogen phylogeny, it is necessary to use pairwise distances.

In order to produce as accurate a phylogeny as possible, it is important to choose the best method of calculating sequence distances. It is interesting to ponder what is meant by “best” method. A method should accurately produce a value for distance that reflects the amount of evolution that has occurred between two sequences. The “amount of evolution” is usually understood to mean the number of fixed point mutations, although it might conceivably mean other things, such as gene conversion events. Furthermore, if a molecular clock hypothesis holds, a calculated distance value should be proportional to the time of divergent evolution separating two sequences.

There is no perfect way to evaluate whether or not one particular distance metric is better than another. One can compare a phylogeny produced from calculated distances with a known phylogeny. However, there are few phylogenies known with certainty. Those that are known are usually sparsely populated, and only the topologies are absolutely certain, not the distances. Furthermore, one cannot be certain that a phylogeny made from sequence distances should necessarily reproduce a phylogeny produced by some other means, such as from paleontological and morphological considerations, or from molecular data derived from another gene locus. The phylogeny of one gene will not necessarily reflect the phylogeny of another gene. Gene phylogenies will not necessarily reproduce species phylogenies.

Therefore one must rely on imperfect methods to evaluate a distance metric. Three are mentioned here. The first approach is to compare the metric with known phylogenies with the assumption that these phylogenies are correct and reflect the phylogeny of the gene in question. Secondly, one can simulate a phylogeny using an assumed model for evolution. One then determines whether the metric in question can reproduce the phylogeny. The problem with this second approach is that never is the process of evolution absolutely known, so this approach relies on the correctness of the hypothetical model for the evolutionary process.

Both of the first two evaluations methods are empirical. While employing such evaluations, one is not concerned with the details of how the distances are calculated. If a distance metric reproduces known phylogenies, then regardless of the details of the calculations, it is useful.

The third approach to evaluate distance metrics is not empirical. In this approach, the details of the evolutionary process are assumed, and a metric is derived from this evolutionary model. Such metrics are “perfect,” in the sense that they precisely measure distances (subject to random error) produced by an evolutionary process identical to that specified in the model. The obvious problem with such metrics is that actual evolution may not obey the assumptions of the model.

I will now consider the details of the original model for distance metrics, due to Jukes and Cantor (1969).

B.2 THE JUKES-CANTOR MODEL

We assume that in a given unit of evolutionary time, a base has a probability μ of mutating. We assume that if a base mutates, then it has an equal probability of being replaced by any of the four bases, including itself.[3] It follows that any base that has mutated one or more times has an equal probability of currently being any of the four bases. The probability of a base remaining unchanged after τ units of time is

$$P(\text{base never mutated}) = (1-\mu)^{\tau} \qquad \text{(B.2)}$$

If one is uncomfortable with designating τ as a unit of time, then one may refer to it as a unit of “pseudotime.” The probability of a base mutating at least once, but currently being in its

[3] A commonly encountered alternative definition of μ specifies that when a base mutates, it has an equal probability of mutating to any of the *other* three bases. These definitions produce equivalent models, but the value of μ in the alternative definition is $\frac{3}{4}$ of the value of μ employed here.

original state, is one quarter of the probability that the base has mutated at least once, or

$$P(\text{base returned to original state}) = \tfrac{1}{4}\left(1-(1-\mu)^{\tau}\right) \tag{B.3}$$

Therefore the probability of a base being in its original state after τ units of time is

$$\begin{aligned} P(\text{base in original state}) &= (1-\mu)^{\tau} + \tfrac{1}{4}\left(1-(1-\mu)^{\tau}\right) \\ &= \tfrac{1}{4}\left(3(1-\mu)^{\tau}+1\right) \end{aligned} \tag{B.4}$$

If τ is allowed to vary continuously, then equation (B.4) becomes

$$P(\text{base in original state}) = \tfrac{1}{4}\left(3e^{-\mu\tau}+1\right) \tag{B.5}$$

Equations (B.4) and (B.5) are equivalent for any practical purpose, as the unit of time will always be chosen small relative to the minimum distance between two sequences. I tend to employ discrete forms as it is slightly easier to program these into computer algorithms. To obtain evolutionary time, one can convert equation (B.4) into the following form:

$$\tau = \frac{\ln\left(\frac{4P-1}{3}\right)}{\ln(1-\mu)} \tag{B.6}$$

Now, one need only compare two sequences and make the maximum likelihood estimate for *P* as the percent identity of the sequences to obtain an estimate for τ as the divergence time. This model is very nice because it is simple.

Unfortunately, it is impossible to use the Jukes-Cantor model with amino acid sequence data.

B.3 A PEPTIDE VIEW OF THE JUKES-CANTOR MODEL

One can hypothesize an evolutionary mechanism similar to the Jukes-Cantor model, but that works at the level of residues instead of nucleotides. The mathematics of such a model would be nearly identical to the Jukes-Cantor model, except that there would be twenty states available for mutation in place of four. One quickly derives the following equation:

$$\begin{aligned} P(\text{residue in original state}) &= (1-\mu)^{\tau} + \tfrac{1}{20}\left(1-(1-\mu)^{\tau}\right) \\ &= \tfrac{1}{20}\left(19(1-\mu)^{\tau}+1\right) \end{aligned} \tag{B.7}$$

There is a major objection to this model. Mutations occur to nucleotides and not to residues. It is already a big assumption that all possible nucleotide substitutions are equally likely. It is yet another assumption to assume that all residues are equally likely mutations from a given codon. In fact, it is only possible for a single nucleotide substitution to change a given codon into 9 of the 63 other codons. At most, a codon might be able to mutate so as to code for 9 other residues, but due to the degeneracy of the genetic code, the widest repertoire of available residue mutations is seven; the narrowest is four (the average is 5.8).

For the model in equation (B.7), the concept of a “mutation” changes from that of equation (B.4). For (B.7), anything that changes an amino acid residue counts as a mutation. This could be one or more point mutations, a intron/exon boundary slide, or a gene conversion event. For (B.4), only point mutations are considered.

Recall that evolution works by natural selection as well as mutation. To a first approximation: mutations alter nucleotides; selection operates on residues. Therefore the use of a nucleotide-based model tends to emphasize the role of mutation in evolution. The use of a residue-based model emphasizes selection. For a non-coding region unaffected by selection it would be appropriate to use a nucleotide model and foolhardy to use a residue model. For a region strongly affected by selection, or by mutational events operating above the scale of a single nucleotide substitution, then it may make more sense to use a residue model.

If selection plays a significant role in the evolution of a protein, then there may be a significant probability that several silent mutations at the nucleotide level occur in one unit of evolutionary time. The rate of residue change is predicted to be much smaller than the rate of nucleotide change. Allowing for several silent mutations to occur between state changes at the site of a particular residue, then contemplation of the genetic code will indicate that widest number of available residue mutations is twelve and the narrowest five (with an average of 7.4). Non-silent mutations at a site are less likely to be fixed. Since selection operates on the site, it is likely that a non-silent mutation will decrease fitness and be selected against. Thus under conditions of selection, a residue model becomes more reasonable.[4]

I will primarily be employing a residue model to analyze the trypsinogen sequences. There are three reasons for this. First, my data is amino acid sequences, at least partially. Second, trypsin is a highly conserved enzyme with a critical digestive function. There are strong selective constraints governing the evolution of trypsin. Third, the trypsinogen genes

[4] Aaron Halpern, currently at the University of New Mexico, has done some work building more complex residue models that incorporate aspects of the genetic code (personal communication). It may be that enough data will eventually be available to conduct empirical evaluations of different distance metrics. Currently, one can only note that for vertebrate trypsinogen data, residue-based and nucleotide-based metrics give roughly similar distances.

are members of a multicopy gene family. They often reside in repeated elements in chromosomes. Within a genome, there is a large reservoir of trypsinogen gene material for gene conversion, unequal crossing over, and other methods of genetic exchange.

Gene conversion events can convert lengths of DNA ranging from a single nucleotide to several thousand (Li, 1997). A codon may undergo a change at all three of its positions as a result of a gene conversion event. Furthermore, this change is less likely to be selected against, as the donor DNA may be a functional allele. Thus, it may be improper to consider nucleotide-based models for change when gene conversion is a prominent force for random variation. Mutation by gene conversion, effectively at the level of codons and residues, may have been common and dominant during the course of trypsinogen evolution.

The preceding arguments suggest that a residue model may be more appropriate for trypsinogen evolution than a nucleotide model. However, the initial model described by equation (B.7) may be too simplistic. There are several possible modifications to this basic model. I will describe two simple extensions.

A first possible extension to the basic residue model assumes that strong selection at a site permits only a few residues to be accepted as substitutions at that site. A change to another residue, other than these few, will result in a low fitness and effectively a zero probability that the mutation will be fixed. This, in essence, "slows down" mutation at that site. The modification to equation (B.7) is as follows:

$$P(\text{residue in original state}) = \frac{1}{N}\left((N-1)\left(1-\frac{N\mu}{20}\right)^{\tau}+1\right) \tag{B.8}$$

Here N is the number of allowed residues at the site. Equation (B.7) can be obtained from equation (B.8) by setting N=20.

A second possible extension to the basic model is that there is no selection, but the number of possible mutations at any particular site is limited. This might occur if mutations occurred solely as a result of gene conversion from a limited number of alternative alleles. The modification to equation (B.7) would be as follows:

$$P(\text{residue in original state}) = \frac{1}{N}\left((N-1)(1-\mu)^{\tau}+1\right) \tag{B.9}$$

It is hard to make a case for practical use of the model implemented in equation (B.9), partly because one of the motivations for using a residue model is that natural selection plays a significant role in evolution, as the available pool of residues is likely to change over time. Additionally, it seems unlikely that gene conversion from a limited pool of residues would be a dominant mode of evolution. Equation (B.9) is presented here mainly for reference. Therefore equation (B.8) will form one basis for my evaluation of the trypsinogens. Note that equation (B.9) can be transformed into equation (B.8) merely by employing a smaller probability of mutation per unit time ($\frac{N}{20}\mu$ in place of μ). This reflects the idea that selection "slows down the evolutionary rate."

For the present paper, I arbitrarily set the rate μ to 0.01. Were a molecular clock hypothesis to hold, μ could be set empirically. Small alterations in μ affect the scale of a derived phylogenetic tree, but not its topology or proportions. To demonstrate this last point for the Jukes-Cantor model, I make the observation that $(1-\mu)^{\tau} \approx 1-\tau\mu$. Note that $\mu<<1$. Therefore from equation (B.4) we have

$$\tau \approx \frac{4\left(1-\left(\text{percentage of bases in orginal state}\right)\right)}{3\mu} \tag{B.10}$$

Evolutionary time τ is inversely proportional to the mutation rate μ. This should not be surprising. If the mutation rate doubles, it should take half as long for a sequence to acquire the same number of changes. Because of back mutations and saturating effects, this proportionality is approximate.

Not all sites of a sequence will necessarily have the same number of permitted residues. Therefore one cannot use an equation similar to equation (B.6) or (B.10) when evaluating the most likely τ for equation (B.8). Rather one must determine the τ that results in the maximum likelihood of the observed identities and differences between two sequences. This can easily be done by computer, although multiple iterations may consume considerable processor time.

All of the models presented here assume that each site of a sequence evolves independently. This assumption clearly does not hold in reality, as gene conversion events may convert many adjacent residues simultaneously. However, any attempt to account for covariant effects is extremely unwieldy. Additionally, our knowledge of selective pressures is too limited to provide a model more accurate than what can be obtained with an assumption of independence.

In order to implement equation (B.8) in practice, one needs to determine the number of possible states possible at each site. For this, I assume that I have a large enough collection of sequences to have observed all possible permitted residues at each site. This assumption is most valid for sites with few observed residues. These sites are the most significant contributors to distance calculations, so this assumption is reasonable. A major incentive to use this model is to account for a slower rate of change at sites that, due to functional constraints, have few permitted residues. Note that if a site has only one observed state, I will assume that site to be invariant. Such sites contribute no information to evolutionary distance calculations and must be ignored.[5] A plot of the variability of each trypsin site is shown in Figure B.1.

A bizarre but fascinating combination of Sesame Street and information theory has produced a method of graphing data from multiple sequence alignments. Such graphs are known as sequence logos (Schneider and Stephens, 1990). A sequence logo for pretrypsinogen is shown in Figure B.2. Because my aligned pretrypsinogens do not represent a uniform sampling of the vertebrate phylogeny, the relative residue frequency depicted by character height in the sequence logo may not be strongly correlated with biological significance. Nevertheless, the sequence logo strikingly depicts highly conserved regions. The information in the sequence logo complements the information in Figure B.1.

Some authors have attempted to incorporate site-to-site variability into phylogeny calculations (e.g., Yang, 1993, 1994, and 1995; Felsenstein and Churchill, 1996; Thorne et al., 1996). Jones et al. (1994) present a mutation data matrix for transmebrane prtoeins, which is a step in the right direction. However, this matrix cannot be used for trypsin, which lacks transmembrane motifs.

Hidden Markov models (HMM) are currently in vogue as a method for incorporating site-to-site variabilty into models. It is not at all clear that a HMM algorithm would be appropriate for trypsin data. Firstly, the “hidden” aspect forces the researcher to discard what is known about conserved residues and selective constraints. Secondly, effective use of an HMM requires that the persistence length of site conservation be at least moderately greater than a single residue. This constraint is violated repeatedly by trypsin, as a glance at the numerous narrow spikes and valleys of Figure B.1 shows.[6] This also limits the effectiveness of Γ distribution models. Although the use of HMM algorithms may be a step in the direction of incorporating site specific information into protein phylogenies, it is my feeling that they currently offer no improvement on traditional methodologies.

[5] Technically, using the model of equation (B.8), they do not have to be ignored, as the probability of the base resting in its original state is 1, so contributes neither negatively or positively to a maximum likelihood estimate for τ.

[6] Jin and Nei (1990) provide some further discussion on site specific rates of change.

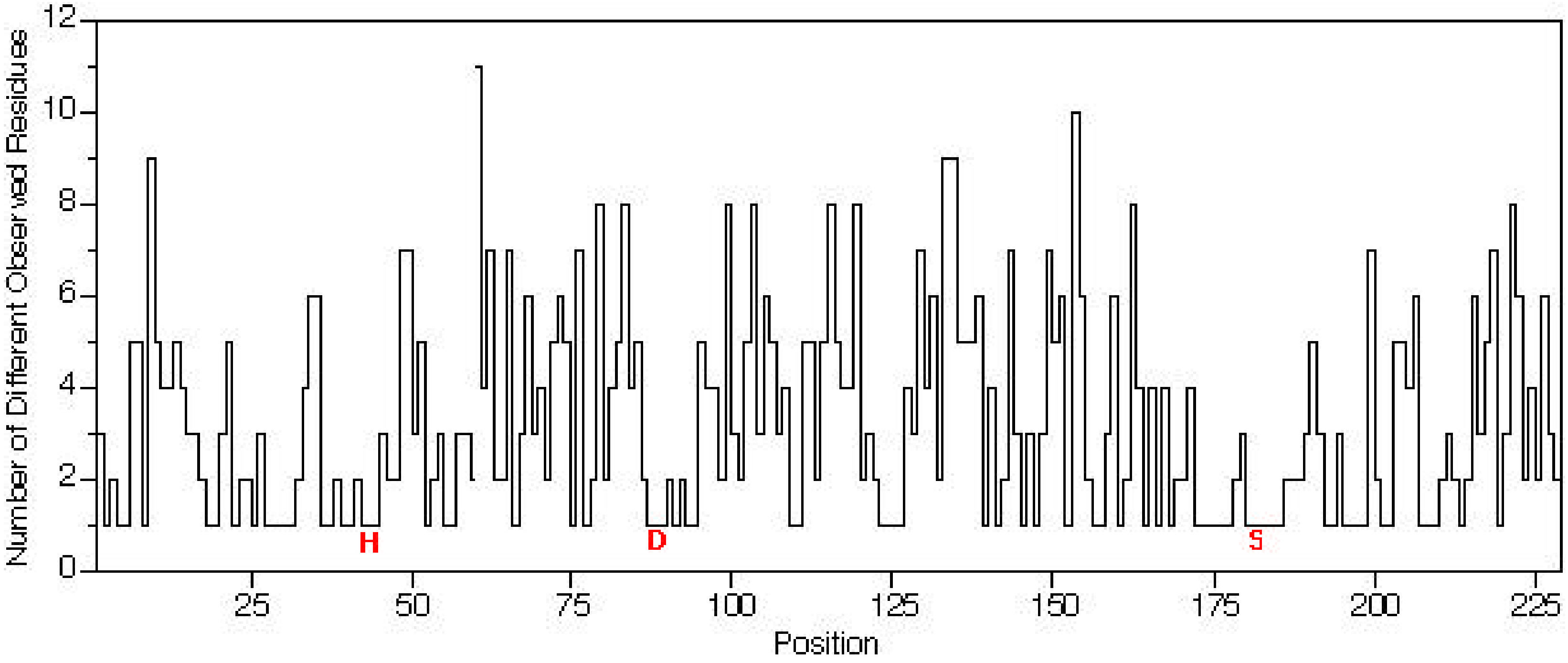

Figure B.1. Trypsin site variability. The letters (H, D, and S) are positioned in the vicinity of the corresponding catalytic triad residues. This graph does not address the frequency of a particular residue at a site, as each different residue is counted once, whether it occurs in one sequence or many. The last site of the activation peptide is the first site (1) in this graph: lysine, arginine, or histidine. The last site (228) represents either a serine or a stop codon.

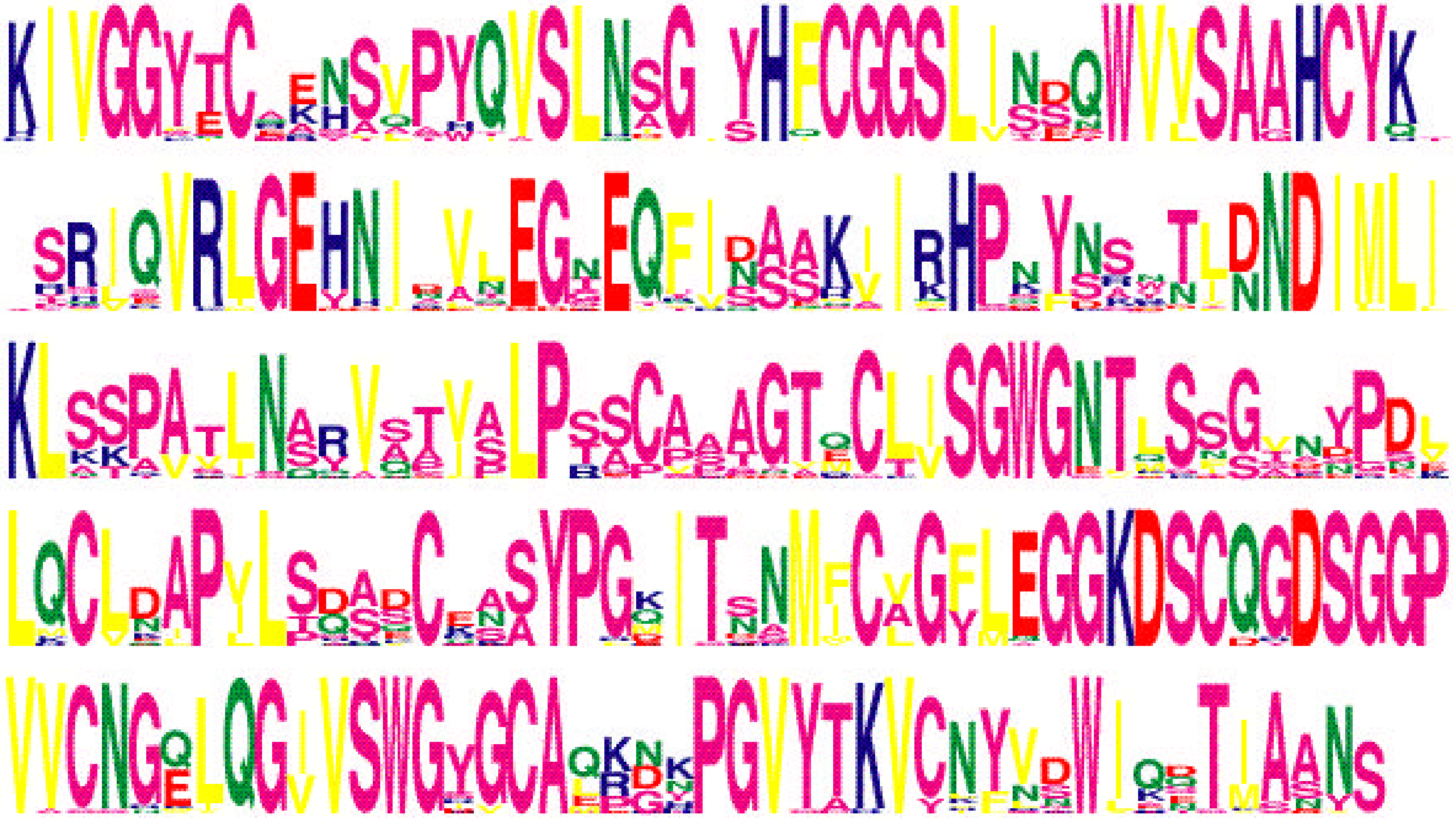

Figure B.2. Pretrypsinogen sequence logo. The total height of the characters at a given site indicates the information content of that site, while the height of each character reflects the relative frequency of the represented residue at that site. The color of a residue reflects its chemical structure. Consult Schneider and Stephens (1990) for details of sequence logos.

Ideally, I would like to construct a separate distance matrix for each trypsin residue site. Most current distance metrics treat all sites equivalently (e.g., Jones et al., 1992; Taylor and Jones, 1993). Reliable construction of site specific distance matrices for trypsinogen (and most other proteins) would require more sequences and a better understanding of evolution than is currently available.

A preliminary, although simplified, approach to construction of a metric would be to tabulate all possible residues at a particular site that maintain function at the molecular level and fitness at the organismal level. This would require a statistically significant sampling of sequences spanning all clades of the phylogeny in question. Since known trypsinogen sequences from some vertebrate classes are either very sparse or completely missing, reliable estimates of the number of permitted residues at a given trypsinogen site cannot be made. In years to come, reliable estimates will be possible. At such a time, extensions to this approach for site specific distance estimation might include accomodations for the underlying mechanisms of mutation or for covariation of sites. Additionally, computational advances may permit maximum likelihood approaches to be combined with such models for evolution. For the present, I suspect that the simple “limited state” method of equation (B.8) provides a reasonable approximation to what might be obtained with more data.

A phylogeny analagous to that of Figure 3.15 is shown in Figure B.3. This phylogeny uses the distances calculated according to the methodology described above in place of the *protdist* distances used for the same purpose in Chapter 3. The toplogy of the phylogeny in Figure B.3 is essentially the same as the phylogeny in Figure 3.15. The distances of the phylogeny in Figure B.3 produce a phylogeny that has slightly more resemblance to a “star phylogeny” than does Figure 3.15. This results from smaller relative distances between pairs of sequences calculated with the methodology of this appendix.

An attempt to use multidimensional scaling on these distances produces plots with unacceptably high stress (approximately 0.35; data not shown). This is due to the moderately “spherical” character of this data in 31 dimensions, making it difficult to compact the data without skewing it. Despite this, as Figure 3.17 shows, these data still significantly support a hypothesis of coincidental evolution. If the true divergence time distances were indeed spherical, that would imply multiple early divisions followed by coincidental evolution that has created the appearance of a single early division with corresponding clustering of sequence distances. This possibility is discussed briefly in Section 3.17. The accumulation of more trypsinogen sequences and further refinement of sequence distance metrics will

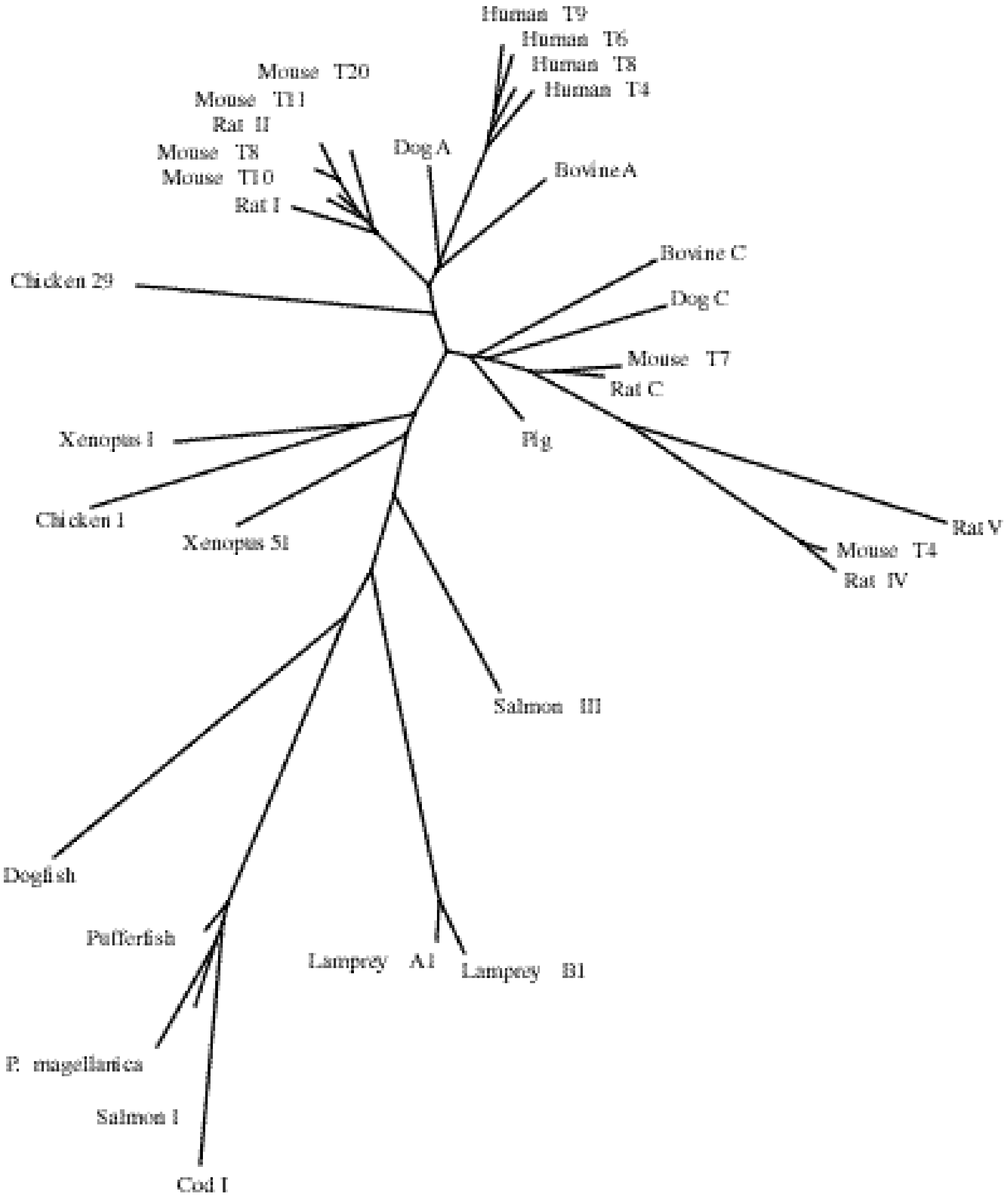

Figure B.3. A Fitch-Margoliash phylogeny of thirty-two vertebrate trypsinogens. Distances, calculated according to the methodology described in Appendix B, were fed to the program *fitch*, with global rearrangements and 40 random "jumbles" (Felsenstein, 1993).

ultimately differentiate between the various hypotheses for the timing of the division(s) of the trypsinogen multigene family.

In order to develop data suitable for inclusion in site-specific matrices, a great deal of knowledge must be known about the selective pressures on the site in question. It is conceivable that at some point in the future that such knowledge may be gained from intense biochemical and genetic study of the gene product in question. At present, it is beginning to be possible to assign sites based on homology and known or predicted secondary and tertiary structure. Such predictive efforts work better on some proteins than others. There are currently several models for sequence based protein structure prediction. The simplest models employ only three structures: helix, sheet, and loop, so are unlikely to provide much extra power to phylogenetic analyses.

There are a few models with increased sophistication. One example is the model of I-sites by Bystroff and Baker (1998). I-site theory was originally and primarily designed to identify distant relationships between proteins based on similarities between protein folding initiation sites. However, the theory can be adapted to provide data for site-specific distance metrics. In short, if a site from a protein of interest can be assigned with confidence to a specific I-site position, then that position can be assumed to evolve according to the constraints on that I-site. Such constraints can be estimated from the entire body of I-site data, rather than limiting oneself to the possibly-skimpy data available from the set of proteins in question. This approach can form the basis for a more complex model. For example, there may be additional constraints on a residue beyond those imposed by its inclusion in an I-site.

I-site motifs can be assigned to 77 of 238 analyzed trypsin sites. Unfortunately, all but eight of these are assigned with a confidence statistic 0.80 or less, and all but 23 have a confidence statistic less than 0.50.[7] Furthermore, 47 of the positions assigned to I-sites were either absolutely conserved or showed only two different residues in all vertebrate sequences. This high conservation is not unexpected, as the original intention of I-site prediction is to predict structures important for initiation of protein folding, which should also be conserved.[8] However, since at these 47 conserved sites the trypsins show an even more restricted range of variation than is seen in I-site consensus sequences, it suggests that even stronger selective pressure operates on these positions in trypsin than merely enough to maintain an I-site consensus. Therefore it would be inappropriate to use evolutionary change matrices based on

[7] A confidence statistic of 0.50 to 0.80 is judged to be “OK,” while above 0.80 is “good.” The confidence statistic is described by Bystroff and Baker (1998).

I-sites at these positions. These considerations lead me to conclude that an attempt to incorporate current data on protein motif consensuses would not have noticeably improved trypsin distance calculations. It may be that this will change as more knowledge accumulates on the relationships between protein structure and sequence. Also, I-site data may prove more useful in analyzing proteins other than trypsin, which might have fewer absolute constraints such as those dictated by the maintenance of a proteolytic active site.

[8] There are a total of 113 positions in the vertebrate trypsinogen alignment which permit either one or two observed residues. Of these, 35 are absolutely conserved. Approximately two thirds of these highly conserved positions are not assignable to a known protein structure motif (I-site predictions correlate well with other structure prediction algorithms). Many of these positions are involved in catalysis, substrate specificity, or cystine bridge formation.