VERTEBRATE TRYPSINOGEN EVOLUTION

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

Henry James

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

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

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

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

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

Trypsin is important commercially. Currently, the use of cold-adapted proteases is anticipated to have utility in food processing. Understanding the differences in function of the trypsin isozymes may aid fish farming techniques as well as genetic engineering of fish in farms. Additionally, antifreeze proteins, some of which evolved from trypsinogen, will increasingly play a role in genetically engineered foods (Macouzet et al., 1999).

3.1 COINCIDENTAL EVOLUTION

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

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

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

Dover et al. (1982) have additionally complicated the semantics of the field with the term “molecular drive”. This term is usually employed to describe the phenomenon of a single mutation spreading throughout a multigene family (“homogenization”) within a species (“fixation”). Thus, it is precisely analogous to the population genetics terminology of “fixation of gene in a population”. Dover emphasizes that molecular drive is a “process”, but if so, then he has needlessly coined another term synonymous to horizontal evolution, coincidental evolution, and concerted evolution. None of these terms are referenced in his 1982 monograph, so it is quite possible that he was unaware of their existence at that time. It is clear from his definition that he excludes mutational bias and selective pressure as causes of the phenomenon of “molecular drive”. Furthermore, he makes clear that molecular drive can be due to either neutral or selected (“stochastic or biased”) processes (Dover et al, 1993). If neutral processes are permitted, then clearly the word “drive” has unfavorable connotations. Use of the term “molecular drive” is therefore unfortunate.

Several mechanisms can account for coincidental evolution. Horizontal transfer of genetic information has a sudden and dramatic effect on the similarity of two genes. Horizontal transfer can be accomplished by unequal crossing over, gene conversion, or possibly other mechanisms. Common selective pressures on function can produce a slow but inexorable convergence of similarity (see, for example, Stewart and Wilson, 1987). A similar

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

effect can be produced by a common bias in the mechanism of random variation, such as a nucleotide usage bias, or exposure to an environmental mutagen. However, it is unlikely that biases in random variation play a detectable role in vertebrate coincidental evolution. An overview of the mechanisms of coincidental evolution is provided by Li (1997).

3.2 HISTORICAL PERSPECTIVE ON TRYPSINOGEN

Trypsin from the bovine pancreas was one of the first proteases isolated with sufficient purity and enough quantity for precise biochemical studies (Northrup et al., 1948). This bovine trypsin was one of the first proteins to be sequenced (Walsh and Wilcox, 1970). Three-dimensional trypsin and trypsinogen protein structures were early conquests of X-ray crystallography (Sweet et al., 1974; Kossiakoff et al., 1977). Thus, over a period of several decades, the details of the sequence, structure, and mechanism of action of trypsin were worked out (reviewed by Male et al., 1995). The early availability of data on trypsin and other serine proteases helped fuel the birth and development of the field of molecular evolution.

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

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

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

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

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

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

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

As a rule of thumb, group II trypsinogens appear 5' to TCR Vβ gene segments, while group I trypsinogens appear 3' to TCR Vβ gene segments. In the human there are no functional group II trypsinogens, but there are two trypsinogen relics (T1 and T2) and a pseudogene (T3), all of which are immediately 5' to the entire TCR β locus.[3] These are all derived from group II trypsinogens. The interval between Vβ29S1 and Dβ1 in humans contains 3 functional trypsinogen genes (T4, T6, and T8) and two trypsinogen pseudogenes (T5 and T7). All of these trypsinogens are group I. The 3' end of the human TCR β locus is duplicated on chromosome 9, where a fourth functional group I trypsinogen (T9) is found 3' to the orphon Vβ29S2 gene segment.

A similar organizational grouping of the trypsinogens is found in the mouse. However, the mouse has several functional group II trypsinogens. Two pseudogenes, two relics, and three functional group II genes lie 5' to the mouse TCR β locus (T1-T7).[4] Five relics and eight functional group I trypsinogens lie between mouse Vβ18S1 and Dβ1 (T8-T20).

The organization of the trypsinogens in the chicken differs somewhat. There are two families of V gene segments in the chicken TCR β locus: Vβ1 and Vβ2. All the Vβ1 segments appear 5' to the Vβ2 segments. Each V segment is tandemly linked to a trypsinogen gene. There are three known chicken group I trypsinogens, in tandem with the three known Vβ2 segments, with the last trypsinogen located between the most 3' Vβ2 segment and Dβ1. There are approximately 6 group II trypsinogens and Vβ1 segments, tandemly linked with opposite orientations. Characterization of the chicken TCR β locus is not complete (Kai Wang, personal communication).

Linkage of the TCR β locus to trypsinogen genes has not been established in other organisms, but I postulate that the two will always occur together. The synteny of trypsinogen and TCR β may be due to an important functional synergism of the two loci or merely due to random association. If a random rearrangement was originally responsible for the synteny of the two loci, then there may be a mechanism that prevents successful chromosomal rearrangements that split this synteny. With trypsinogen genes internal to the TCR β locus, it may be that splitting rearrangements destroy the TCR β locus, and significantly decrease the fitness of the resulting mutant. Alternatively, there may be no such constraining selective force. In this case, the two loci remain syntenic merely because a random event has not separated them. I will return to this discussion in Section 3.19.

The tandemly repeated nature of trypsinogens has been observed in other organisms. Three repeated trypsinogens are present in the pufferfish genome, two in tandem and one with opposite orientation (Kai Wang, personal communication). Two or more repeated

[3] Human T1 is not a true relic; see Section 3.22.

[4] Mouse T1 is not a true relic; see Section 3.22.

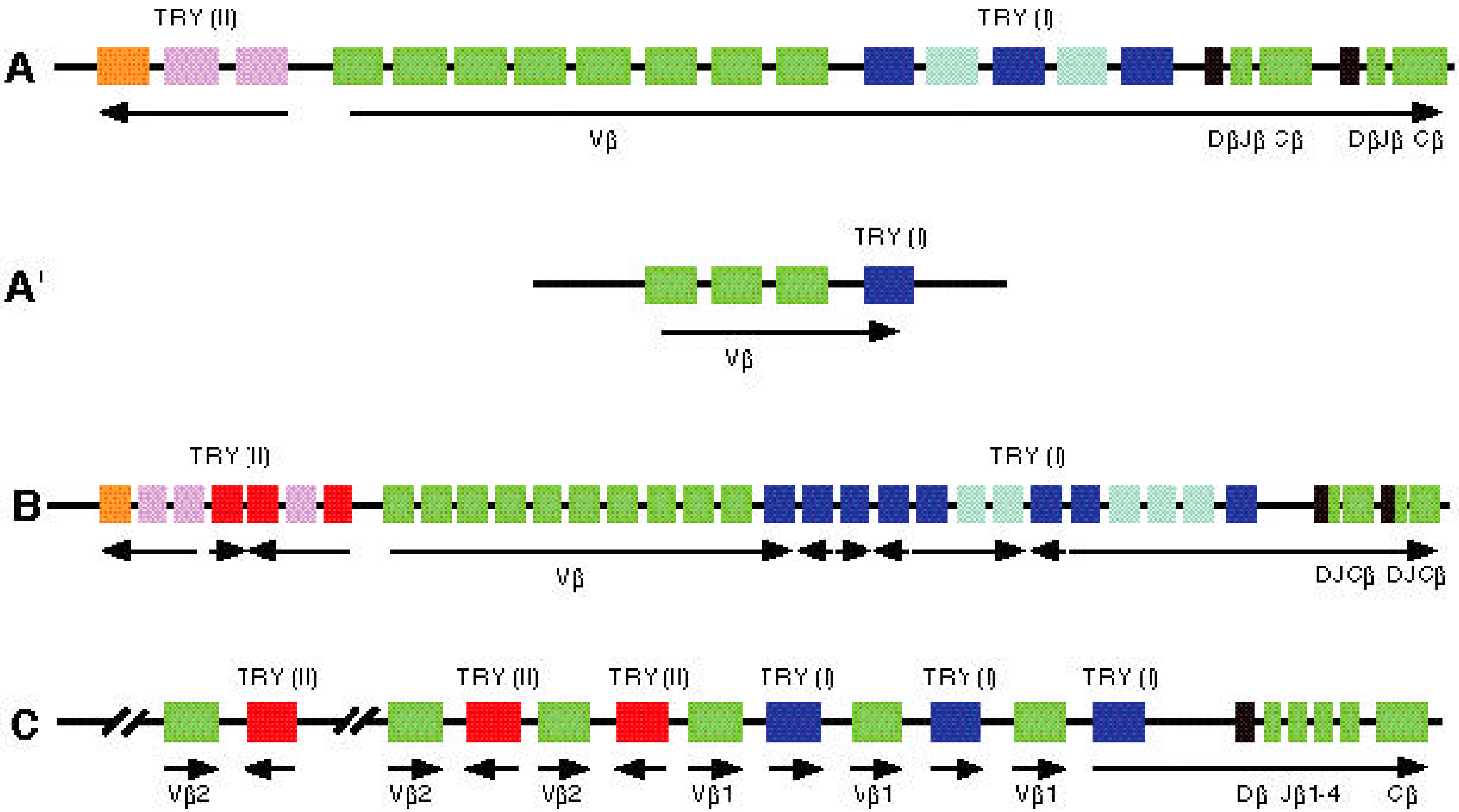


Figure 3.1. Cartoons of the topology of genomic trypsinogen organization. A, human TCR/TRY locus; A', partial translocation of human TCR/TRY locus to chromosome 9; B, murine TCR/TRY locus; C, chicken TCR/TRY locus. Blue, group I trypsinogen; Light blue, group I pseudotrypsinogen; Red, group II trypsinogen;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

trypsinogens are linked in tandem in the lamprey genome. The *Drosophila* genome has four tandem trypsinogen genes in alternating orientations, but presumably not linked to a TCR locus (Davis, 1985), as the immune receptor loci are hypothesized to have first evolved in the chordate lineage.

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

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

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

3.3 THE COMPILATION OF TRYPSIN AND TRYPSINOGEN SEQUENCES

Previously published trypsin and trypsinogen sequences were culled from Genbank and SwissProt using a variety of text and homology based searches (Gish and States, 1993;

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

Altschul et al., 1990). The homology searches were used to rule out the possibility that a trypsin or trypsinogen sequence might be present in a database under a different name. However, no such “mislabeled” vertebrate sequences were identified.

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

Two vertebrate trypsinogen-like cDNA sequences present in Genbank in early 1997 were not included in any of my analyses: those of the plaice *Pleuronectes platessa* and the notothenioid *Dissosthicus mawsoni*.[6] These are discussed in Subsection 3.18.2, where I argue that they are not orthologous to either group I or group II trypsins. In addition to the lack of orthology, the original *Pleuronectes platessa* sequence submission to Genbank contained several sequencing errors including a pair of frameshifts that altered a 36 bp subsequence. This sequence was revised in 1996 from accession number X56744.0 to X56744.1. For historical reasons, the X56744.0 sequence was used for one of the analyses of this thesis (Figure 3.20). Trypsinogens submitted to Genbank after mid-1997 were also not included. These are discussed in Subsection 3.18.1.

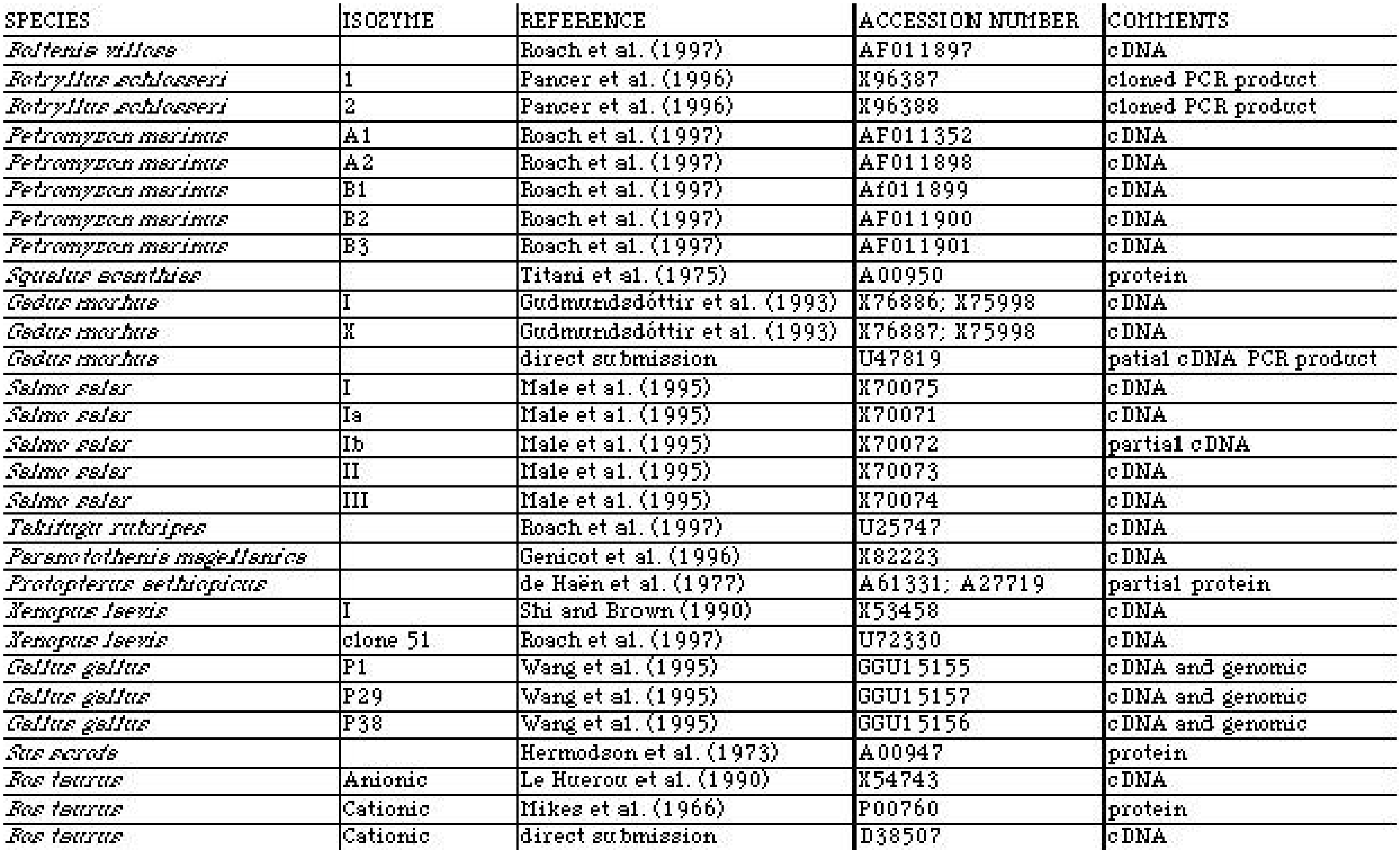
3.4 CLONING AND SEQUENCING *PETROMYZON MARINUS* TRYPSINOGEN

Poly(A)-mRNA was prepared from the dissected gut of a *Petromyzon marinus* ammocoete (a gift of James Seeley, Hammond Bay Biological Station, MI). The mRNA was reverse transcribed and cloned as cDNA into the λ-ZAP directional-cloning vector (Stratagene). Additionally, RT-PCR was performed on the poly(A)-mRNA with the trypsin specific primers TRYF and TRYR. The PCR primers used in this study are tabulated in Table 3.3. The sequence of the 387 bp TRYF-TRYR PCR product was consistent with trypsin and was used to probe the *Petromyzon* gut cDNA library. Thirty-one positive plaques were sequenced at their 5' ends with the m13 reverse primer. Several of these plaques hybridized weakly to the probe, and were picked due to the possibility that the probe might have hybridized to something unexpected but interesting.

Of these 31 plaques picked for sequencing, 17 were trypsinogen, 4 were

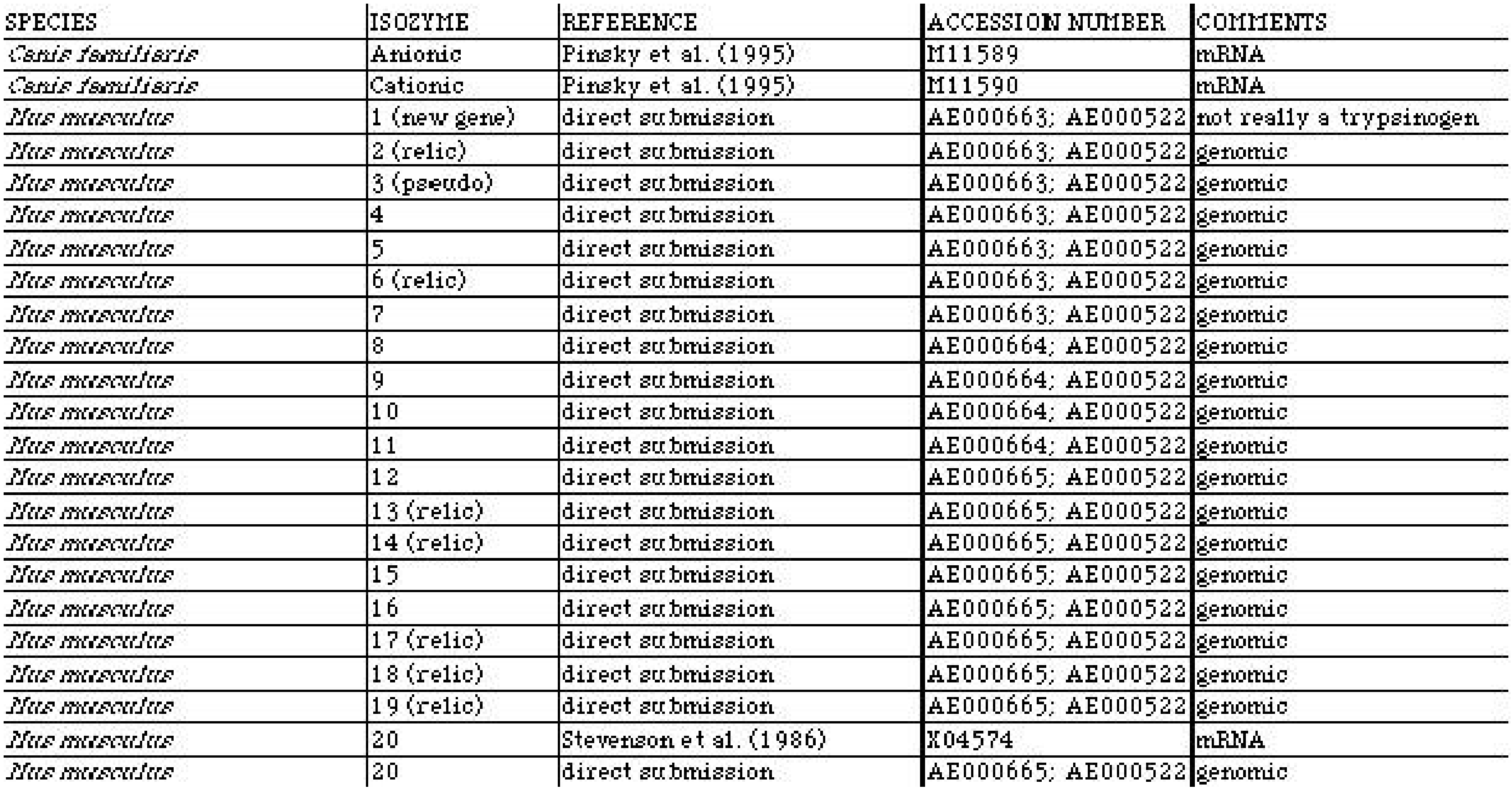
[6] Genbank X56744 and U58835.

Table 3.1. Literature references for the chordate trypsinogens.



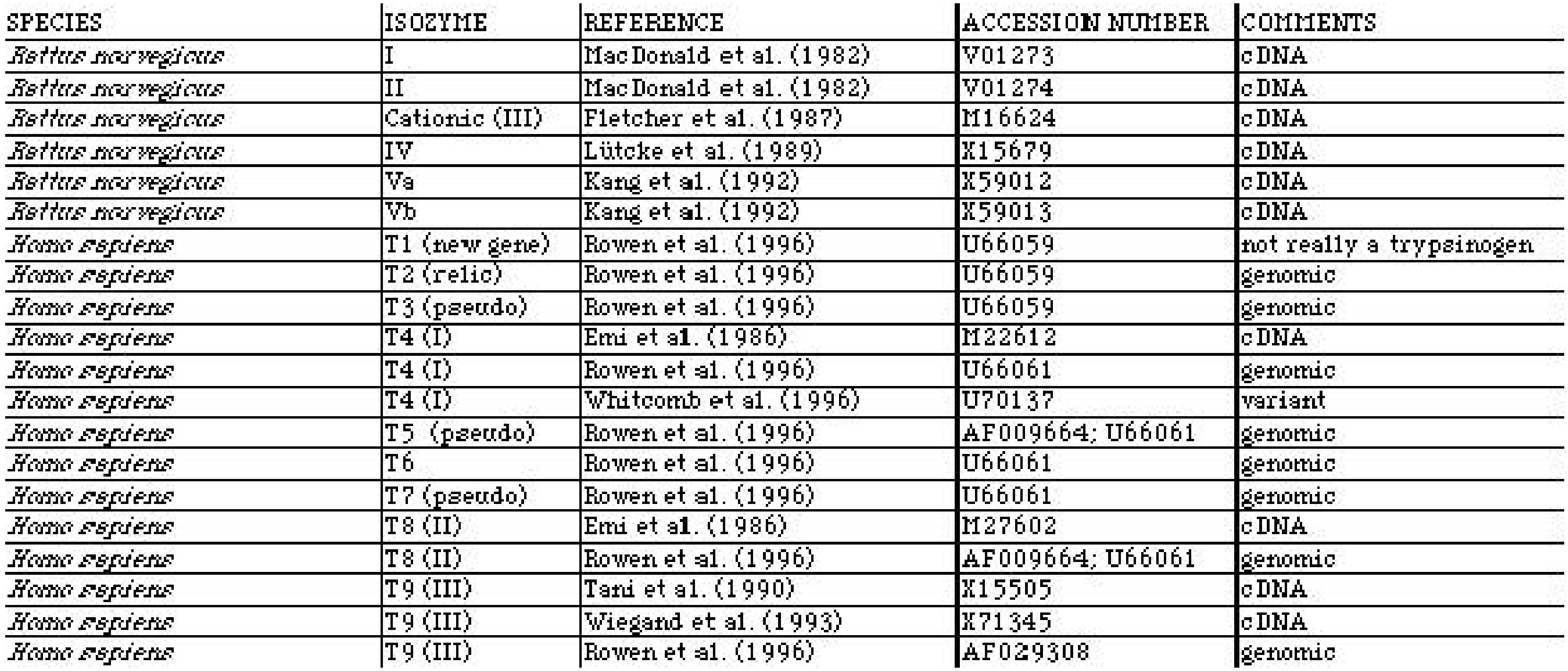
SPECIES	ISOZYME	REFERENCE	ACCESSION NUMBER	COMMENTS
Boltenia villosa		Roach et al. (1997)	AF011897	cDNA
Botryllus schlosseri	1	Pancer et al. (1996)	X96387	cloned PCR product
Botryllus schlosseri	2	Pancer et al. (1996)	X96388	cloned PCR product
Petromyzon marinus	A1	Roach et al. (1997)	AF011352	cDNA
Petromyzon marinus	A2	Roach et al. (1997)	AF011898	cDNA
Petromyzon marinus	B1	Roach et al. (1997)	Af011899	cDNA
Petromyzon marinus	B2	Roach et al. (1997)	AF011900	cDNA
Petromyzon marinus	B3	Roach et al. (1997)	AF011901	cDNA
Squalus acanthias		Titani et al. (1975)	A00950	protein
Gadus morhua	I	Gudmundsdóttir et al. (1993)	X76886; X75998	cDNA
Gadus morhua	X	Gudmundsdóttir et al. (1993)	X76887; X75998	cDNA
Gadus morhua		direct submission	U47819	patial cDNA PCR product
Salmo salar	I	Male et al. (1995)	X70075	cDNA
Salmo salar	Ia	Male et al. (1995)	X70071	cDNA
Salmo salar	Ib	Male et al. (1995)	X70072	partial cDNA
Salmo salar	II	Male et al. (1995)	X70073	cDNA
Salmo salar	III	Male et al. (1995)	X70074	cDNA
Takifugu rubripes		Roach et al. (1997)	U25747	cDNA
Paranotothenia magellanica		Genicot et al. (1996)	X82223	cDNA
Protopterus aethiopicus		de Haën et al. (1977)	A61331; A27719	partial protein
Xenopus laevis	I	Shi and Brown (1990)	X53458	cDNA
Xenopus laevis	clone 51	Roach et al. (1997)	U72330	cDNA
Gallus gallus	P1	Wang et al. (1995)	GGU15155	cDNA and genomic
Gallus gallus	P29	Wang et al. (1995)	GGU15157	cDNA and genomic
Gallus gallus	P38	Wang et al. (1995)	GGU15156	cDNA and genomic
Sus scrofa		Hermodson et al. (1973)	A00947	protein
Bos taurus	Anionic	Le Huerou et al. (1990)	X54743	cDNA
Bos taurus	Cationic	Mikes et al. (1966)	P00760	protein
Bos taurus	Cationic	direct submission	D38507	cDNA

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



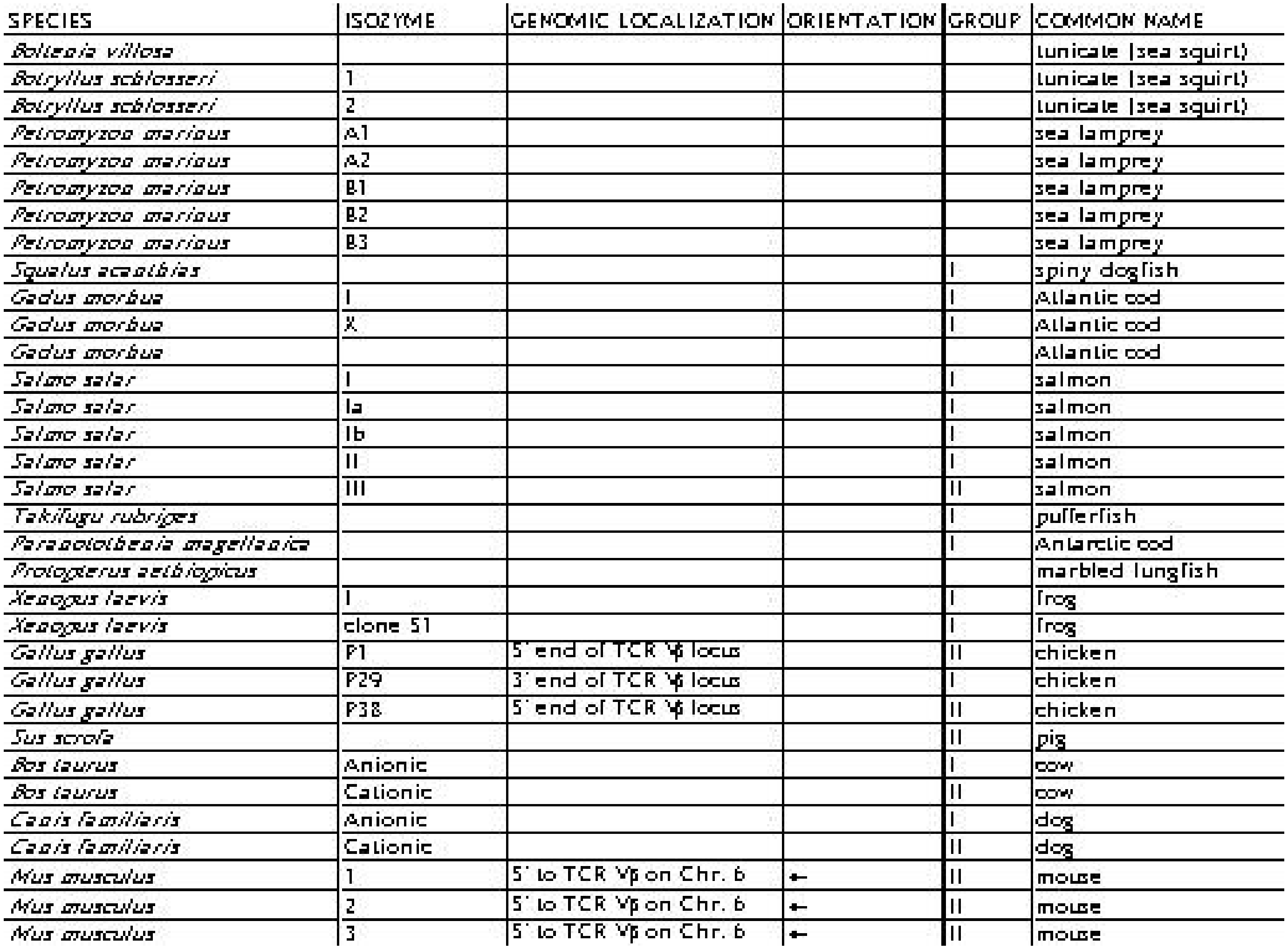
SPECIES	ISOZYME	REFERENCE	ACCESSION NUMBER	COMMENTS
Canis familiaris	Anionic	Pinsky et al. (1995)	M11589	mRNA
Canis familiaris	Cationic	Pinsky et al. (1995)	M11590	mRNA
Mus musculus	1 (new gene)	direct submission	AE000663; AE000522	not really a trypsinogen
Mus musculus	2 (relic)	direct submission	AE000663;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

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



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

Table 3.2. Classification comments for the chordate trypsinogens.



SPECIES	ISOZYME	GENOMIC LOCALIZATION	ORIENTATION	GROUP	COMMON NAME
Boltenia villosa					tunicate (sea squirt)
Botryllus schlosseri	1				tunicate (sea squirt)
Botryllus schlosseri	2				tunicate (sea squirt)
Petromyzon marinus	A1				sea lamprey
Petromyzon marinus	A2				sea lamprey
Petromyzon marinus	B1				sea lamprey
Petromyzon marinus	B2				sea lamprey
Petromyzon marinus	B3				sea lamprey
Squalus acanthias				I	spiny dogfish
Gadus morhua	I			I	Atlantic cod
Gadus morhua	X			I	Atlantic cod
Gadus morhua					Atlantic cod
Salmo salar	I			I	salmon
Salmo salar	Ia			I	salmon
Salmo salar	Ib			I	salmon
Salmo salar	II			I	salmon
Salmo salar	III			II	salmon
Takifugu rubripes				I	pufferfish
Paranotothenia magellanica				I	Antarctic cod
Protopterus aethiopicus					marbled lungfish
Xenopus laevis	I			I	frog
Xenopus laevis	clone 51			I	frog
Gallus gallus	P1	5' end of TCR $V\beta$ locus		II	chicken
Gallus gallus	P29	3' end of TCR $V\beta$ locus		I	chicken
Gallus gallus	P38	5' end of TCR $V\beta$ locus		II	chicken
Sus scrofa				II	pig
Bos taurus	Anionic			I	cow
Bos taurus	Cationic			II	cow
Canis familiaris	Anionic			I	dog
Canis familiaris	Cationic			II	dog
Mus musculus	1	5' to TCR $V\beta$ on Chr. 6	←	II	mouse
Mus musculus	2	5' to TCR $V\beta$ on Chr. 6	←	II	mouse
Mus musculus	3	5' to TCR $V\beta$ on Chr. 6	←	II	mouse

Table 3.2. Classification comments for the chordate trypsinogens (continued).

SPECIES	ISOZYME	GENOMIC LOCALIZATION	ORIENTATION	GROUP	COMMON NAME
Mus musculus	4	5' to TCR Vβ on Chr. 6	←	II	mouse
Mus musculus	5	5' to TCR Vβ on Chr. 6	←	II	mouse
Mus musculus	6	5' to TCR Vβ on Chr. 6	←	II	mouse
Mus musculus	7	5' to TCR Vβ on Chr. 6	←	II	mouse
Mus musculus	8	3' to TCR Vβ on Chr. 6	→	I	mouse
Mus musculus	9	3' to TCR Vβ on Chr. 6	←	I	mouse
Mus musculus	10	3' to TCR Vβ on Chr. 6	→	I	mouse
Mus musculus	11	3' to TCR Vβ on Chr. 6	←	I	mouse
Mus musculus	12	3' to TCR Vβ on Chr. 6	→	I	mouse
Mus musculus	13	3' to TCR Vβ on Chr. 6	→	I	mouse
Mus musculus	14	3' to TCR Vβ on Chr. 6	→	I	mouse
Mus musculus	15	3' to TCR Vβ on Chr. 6	←	I	mouse
Mus musculus	16	3' to TCR Vβ on Chr. 6	→	I	mouse
Mus musculus	17	3' to TCR Vβ on Chr. 6	→	I	mouse
Mus musculus	18	3' to TCR Vβ on Chr. 6	→	I	mouse
Mus musculus	19	3' to TCR Vβ on Chr. 6	→	I	mouse
Mus musculus	20	3' to TCR Vβ on Chr. 6	←	I	mouse
Rattus norvegicus	I			I	rat
Rattus norvegicus	II			I	rat
Rattus norvegicus	III (Cationic)			II	rat
Rattus norvegicus	IV			II	rat
Rattus norvegicus	Va			II	rat
Rattus norvegicus	Vb			II	rat
Homo sapiens	T1	5' to TCR Vβ on Chr. 7	←	II	human
Homo sapiens	T2	5' to TCR Vβ on Chr. 7	←	II	human
Homo sapiens	T3	5' to TCR Vβ on Chr. 7	←	II	human
Homo sapiens	T4 (I)	3' to TCR Vβ on Chr. 7	→	I	human
Homo sapiens	T5	3' to TCR Vβ on Chr. 7	→	I	human
Homo sapiens	T6	3' to TCR Vβ on Chr. 7	→	I	human
Homo sapiens	T7	3' to TCR Vβ on Chr. 7	→	I	human
Homo sapiens	T8 (II)	3' to TCR Vβ on Chr. 7	→	I	human
Homo sapiens	T9 (III and IV)	3' to TCR Vβ on Chr. 9	→	I	human

Table 3.3. PCR primers used in the analysis of the chordate trypsinogens.

Degenerate Serine Protease Primers	
H	CTSWCWGCWGCYCAYTG
S	YMSWGGKCCNCCRGARTC
Tunicate Trypsinogen Sequencing Primers	
TUN2F1	TGGAACACGTGGAAAATAGTTCTC
TUN2R1	CGAGAACTATTTTCCACGTGTTCC
TUN2F2	CAAGCAGCGGAGGAACTATCTCCG
TUN2R2	TCCACTAACAGTACACGCGGTGTC
TUN19F1	GGTGTATACACCCGTGTTGCAGTG
TUN19R1	ACACTGCAACACGGGTGTATACAC
TUN2R3	TTTGGATGATTAAGGATTTTTATTG
Trypsin specific primers	
TRYA	TCCGGATCCTGATGACAAGATCGTTGGGGG
TRYB	TCCGGATCCTTCTGTGGAGGCTCCCTCAT
TRYC	TCCGGATCCATAGCCCCAGGAGAC
TRYD	TCCGGATCCTTGGTGTAGACACCAGG
TRYF	CTGGATCCGTGAGACTGGGAGAGCAC
TRYR	CTGGATCCGAATCCTTGCCTCCCTC
Lamprey Trypsinogen Sequencing Primers	
LT2	AGCCAGTGGGTCCTGTCTG
LT3R	TCACGAAGATGTTGTGCTC
LT3	TCATGCTCATCAAGCTGTCCTC
LT4R	ACGCACATGAGGACGTCGGGAC
LT5R	AAGAGTAGTGTGTTAGATCCAC
XTA5	CCGGTGGCCCCGTGGTGTG

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

3.5 LAMPREY TRYPSINOGENS

The lamprey trypsinogen cDNA sequences contain all of the important sequence features expected of a trypsinogen. They possess overwhelming similarity to the known trypsinogens. In particular, they possess the three absolutely conserved cystine bridges present in all serine proteases, the four key trypsin “pocket specificity” residues, the three residues of the catalytic triad, a signal peptide (described in Section 3.8), an activation peptide (described in Section 3.9), a stabilizing trypsin amino-terminus, and a conserved calcium binding site. The conserved residues that are characteristic of the vertebrate trypsinogens are detailed Section 3.15. The overwhelming similarity of the lamprey trypsinogens to all other vertebrate trypsinogens, coupled with the origin of the cDNA library from the gut, permitted the lamprey sequences to be classified as trypsinogen with certainty.

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

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

[7] Genbank AF011352 and AF011898-AF011901.

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

The lampreys diverged from the other members of the vertebrate subphylum early in vertebrate evolutionary history, so the presence of tandemly repeated trypsinogens in lampreys is strongly suggestive that this general organization of the trypsinogens has been maintained throughout vertebrate history. The presence of highly similar repeats is known to facilitate mechanisms of horizontal transfer of genetic information, such as unequal crossing over and gene conversion (Li, 1997). Therefore the vertebrate trypsinogens have been prime candidates for horizontal gene transfer throughout their history. Such horizontal transfer would result in coincidental evolution (see Section 3.1 and Section 3.18).

3.6 CLONING AND SEQUENCING *BOLTENIA VILLOSA* TRYPSINOGEN

Poly(A)-mRNA was prepared from the dissected gut of a specimen of *Boltenia villosa* (a gift of William Moody, University of Washington, WA). The mRNA was reverse transcribed and cloned as cDNA into the λ-ZAP directional-cloning vector (Stratagene). Additionally, RT-PCR was performed on the poly(A)-mRNA with degenerate primers designed to amplify serine proteases: H and S. The PCR primers used in this study are tabulated in Table 3.3. The H and S primers correspond to conserved sequences of the serine protease active site. Similar primers are described by Kang et al. (1992) and Wiegand et al. (1993). A resulting H-S PCR product of approximately 350 bp was agarose-gel isolated and used as a probe to screen the cDNA library. This band failed to sequence due to polyclonality and was judged to be a diverse mixture of serine-protease-derived products. Twenty-three positive plaques were picked and sequenced. Seven cDNAs identified by 5' end sequencing as trypsinogen were completely sequenced by primer walking on both strands with the following primers: TUN2F1, TUN2R1, TUN2F2, TUN2R2, TUN19F1, TUN19R1, and TUN2R3. All seven of the tunicate trypsinogen cDNAs I sequenced appeared to represent the same allele, as I could distinguish no sequence variation between them.[8] Of the other sequences, five were chymotrypsinogen, two were ribosomal proteins, one was actin, one was glutathione S-

[8] Genbank AF011897.

transferase, one was from the mitochondrial 16S RNA, and six were not positively identified. The chymotrypsinogen sequences were identified based on the presence of a methionine at the position that is 192 according to the bovine chymotrypsinogen numbering system, as well as overall similarity.

3.7 TUNICATE TRYPSINOGEN

The *Boltenia* trypsinogen sequence was identified based on its presence in a gut derived cDNA library and its sequence similarity to the other chordate trypsinogens. The *Boltenia* trypsinogen cDNA sequence contains all but one of the important sequence features expected of a trypsinogen. These were mentioned in Section 3.5 for the lamprey trypsinogens, and are further detailed in Section 3.15.

The *Boltenia* trypsin appears to lack the residues forming the calcium-binding site found in all vertebrate trypsins. The two known trypsinogens from the tunicate *Botryllus schlosseri* also appear to lack these residues. Therefore, the tunicate trypsinogens may bind calcium at an alternative site, much as the *Streptomyces griseus* trypsin does (Read and James, 1988). It has been suggested that calcium binding confers stability against thermal or chemical denaturation (Martin, 1984). Additionally, a requirement for calcium may ensure that trypsin is only active extracellularly, since intracellular calcium concentrations are extremely low (Kretsinger, 1976). There is still uncertainty as to the exact location, function, and importance of calcium binding sites in active trypsin (Read and James, 1988; Smalås et al., 1994).

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

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

Ascidiacea, but to different families within this order. It is unclear when the two families diverged.

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

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

3.8 SIGNAL SEQUENCES

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

An alignment of chordate pretrypsinogen signal sequences and activation peptides is shown in Table 3.4. In the case of the canine, the sites have been determined experimentally by comparison of pretrypsinogen with trypsinogen (Carne and Scheele, 1982). All of the other signal peptidase cleavage sites in Table 3.4 were predicted with the program *PSORT* (Nakai and Kanehisa, 1992), which implements the algorithm of von Heijne (1986). These sites are consistent with the vast body of literature on the sequencing of trypsinogen activation peptides (e.g., Bricteux-Grégoire et al., 1972).

In many cases, the sequences of the activation peptides, and particularly their amino-

	signal	activation	mature
B. villosa	MKIVILLLLGLAAVNA	DK	IVGG
B. schlosseri 1	MKVFAILLLAFCGANA	DK	IIGG
B. schlosseri 2	MKVFAILLLALYGANA	DK	IIGG
Lamprey A1	MHGLILALLVGVAAA	APYMYEDH	IVGG
Lamprey A2	MHGLILALLVGVAAA	APYMYEDH	IVGG
Lamprey B1‡	---LIFALLVG**T**AAA	APYMYEDH	IVGG
Lamprey B2‡	--GLIFALLVG**T**AAA	APYMYEDH	IVGG
Dogfish‡	---------------	APDDDDK	IVGG
Cod I	RKSLIFVLLLGAVFA	EEDK	IVGG
P. magellanica	MRSLVFVLLIGAAFA	TEEDK	IVGG
Pufferfish‡	--------LI**A**AAYA	APIDEDDK	IVGG
Salmon I	MISLVFVLLIGAAFA	TEDDK	IVGG
Salmon III‡	---------FAVAFA	APIDDEDDK	IVGG
Xenopus I	MKFLLLCVLLGAAAA	FDDDK	IIGG
Xenopus 51	MKFLVILVLLGAAVA	FEDDDK	IVGG
Chicken 1	MLFLVLVAFVGV**T**VA	FPISDEDDDK	IVGG
Chicken 29	MLFLFLILSCLGA**A**VA	FPGGADDDK	IVGG
Pig‡	---------------	FPTDDDDK	IVGG
Bovine A	MHPLLILAFVGAAVA	FPSDDDDK	IVGG
Bovine C‡	---FIFLALLGAAVA	FPVDDDDK	IVGG
Dog A	MNPLLILAFLGAAVA	TPTDDDDK	IVGG
Dog C	MLTFIFLALLGATVA	FPIDDDDK	IVGG
Human T4 (I)	MNPLLILTFVAA**A**LA	APFDDDDK	IVGG
Human T6	MNPLLILAFVGA**A**VA	VPFDDDDK	IVGG
Human T8 (II)	MNLLLILTFVAA**A**VA	APFDDDDK	IVGG
Human T9 (III)	MNPFLILAFVGA**A**VA	VPFDDDDK	IVGG
Human T3 (5' ϕ)	HEDLHLPALLGA**A**AT	FPTDDDDK	IVGG
Rat I	MSALLILALVGAAVA	FPLEDDDK	IVGG
Rat II	MRALLILALVGAAVA	FPVDDDDK	IVGG
Rat C	MLALIFLAFLGAAVA	LPLDDDDDK	IVGG
Rat IV	MLISIFFAFLGAAVA	LPVNDDDK	IVGG
Rat V	MKICIFFTLLGTVAA	FPTEDNDDR	IVGG
Mouse T4	MKIITFFTFLGAAVA	LPANSDDK	IVGG
Mouse T5	MKIIFFFTFLGAAVA	LPANSDDK	IVGG
Mouse T7	MKTLIFLAFLGAAVA	LPLDDDDDK	IVGG
Mouse T8	MRALLFLALVGAAVA	FPVDDDDK	IVGG
Mouse T9	MNSLLFLALVGAAVA	FPVDDDDK	IVGG
Mouse T10	MSTLLFLALVGAAVA	FPVDDDDK	IVGG
Mouse T11	MNALLILALVGAAVA	FPVDDDDK	IVGG
Mouse T12	MSALLFLALVGAAVA	FPVDDDK	IVGG
Mouse T15	MNAFLILALVGAAVA	FPVDDDDK	IVGG
Mouse T16	MSALLFLALVGAAVA	FPVDDDDK	IVGG
Mouse T20	MSALLILALVGAAVA	FPVDDDDK	IVGG

Table 3.4. Signal peptides and activation peptides of the chordate trypsinogens. The signal peptidase cleavage site is the predicted site (*PSORT*); In the case of the canine, marked with an asterix (*), the sites have been determined experimentally (Carne and Scheele, 1982). A dashed line (--) represents undetermined sequence; The diesis (‡) indicates that available sequence is too short for *PSORT* to make a prediction. Intron/exon boundaries, where known, are indicated by bold lettering.

terminal residues, have been determined. De Haën et al. (1977) provide a nice review, including sequences from dogfish, pig, man, and cow. These sequences are all consistent with the predictions of Table 3.4. Curiously, in the 1960's and 1970's, the activation peptides alone were used as the basis for phylogenetic studies. Bricteux-Grégoire et al. (1972) provide an entry into this literature. In retrospect, statistical support for phylogenies based on these octapeptides is lacking.

In some studies, only an octapeptide is detected as the activation peptide. For example, Davie and Neurath (1955) determine the activation peptide of bovine cationic trypsinogen to be VDDDDK. By contrast, Louvard and Puigserver (1974) determine the activation peptide of bovine anionic trypsinogen to be FPSDDDDK. It is remotely possible that this difference is due to a variation in the signal peptidase cleavage site, but it seems likely that degradation during isolation was the cause, given the high identity of the primary sequences of the pretrypsinogens.

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

The short length of these sequences limits the statistical significance of any conclusions to be drawn from the alignment. Wang et al. (1995) have suggested that slight differences in signal sequences between “anionic” and “cationic” pretrypsinogens might bias targeting towards different cellular locales or influence the rate of secretion. The leucine-isoleucine-leucine sequence present at positions 5-8 of several anionic pretrypsinogens is suggested to fill this role. I feel that it is more likely that this similarity stems from neutral mutations occurring during the common ancestry that these group I pretrypsinogens share. There is no need to invoke differential selection to explain trends observed from the alignment of the pretrypsinogen signal sequences.

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

3.9 ACTIVATION PEPTIDES

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

Following activation, the newly freed amino-terminus of the trypsin enzyme tucks itself into an internal pocket of the globular protein. This “isolecuine-valine-glycine-glycine” sequence stabilizes the conformation of the enzyme, raising its catalytic rate constant by several orders of magnitude (Huber and Bode, 1978; Morgan et al., 1972).[9] This sequence is shown in orange in Figure 3.2. The amino-terminus sequence of *Xenopus* I and the two *Botryllus* trypsins is isolecuine-isolecuine-glycine-glycine, but this difference seems too minor to alter its function.

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

Strikingly, none of the activation peptides for the lamprey trypsinogens end in a lysine or arginine residue. All lamprey trypsinogen activation peptides end in a histidine. Thus it

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

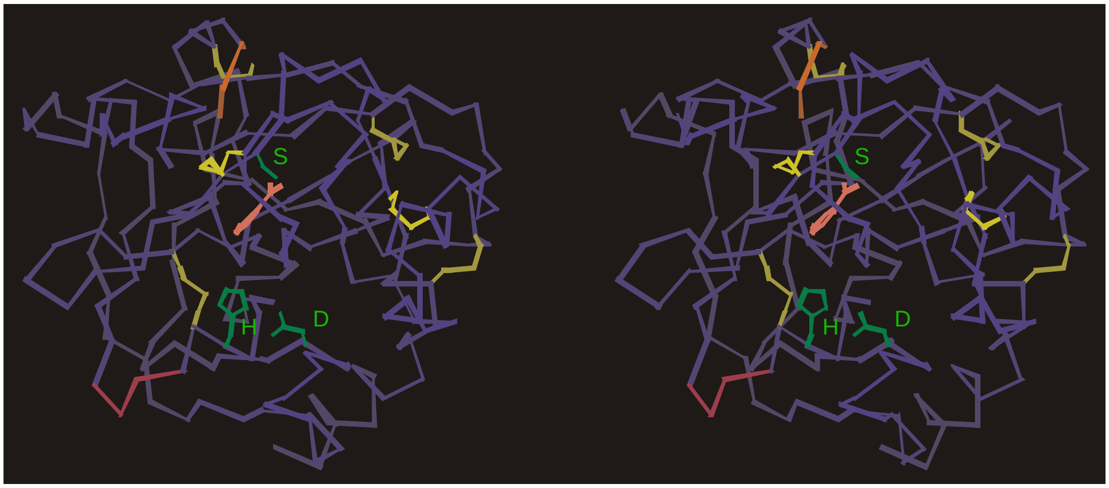


Figure 3.2. A stereoscopic view of the trypsin backbone. The *MAGE* program (David Richardson, author) was employed to visualize rat trypsin II (PDB designation 1ANE). The "IVGG" amino-terminus is tucked into the interior at the top of this view (residues 16-19 of 1ANE; orange). The three catalytic residues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

seems unlikely that lamprey trypsin is capable of autocatalyzing its own activation, as trypsin is not capable of cleaving after a histidine residue. This suggests that lampreys rely exclusively on enterokinase for trypsinogen activation. The life cycle of the lamprey may explain selective pressure for greater control of digestive enzyme activation. For example, adult lampreys will go months to years without eating. The lamprey chymotrypsinogen activation peptide ends in an arginine and so could be activated by trypsin.[10] This would allow lamprey enterokinase to function as a master control switch for digestion, allowing for little or no basal digestive enzyme activation.

3.10 CYSTINE BRIDGES

Acquisition and loss of cystine bridges is a rare evolutionary event and thus a useful phylogenetic marker. It is, however, unclear exactly how useful they are. Each cysteine residue alone is highly conserved, and the added knowledge that a bridge links two does not

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

necessarily provide much additional information to phylogeny inference. Some bridges will be more conserved than other bridges, and the rare gain or loss of such a bridge will be particularly informative, much as would be a change in an active site residue. These issues aside, consideration of cystine bridges will strengthen the already strong case that the human T3 trypsinogen pseudogene is descended from a group II trypsinogen, in contrast to all of the functional human trypsinogens.

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

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

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

	C22-C157	C42-C58	C127-C232	C136-C201	C168-C182	C191-C220
Bacteria		X*			X*	X*
Crayfish		X			X	X
Tunicate		X		X	X	X
Lamprey	X	X	X	X	X	X
Non-Human Gnathostomata	X*	X*	X*	X*	X*	X*
Group II Human (ψT3)	X#	X	X	X	X	X
Group I Human (except T8)	X	X		X	X	X
Group I Human (T8)	X	X			X	X

Table 3.5. Cystine bridges of the chordate trypsinogens. Across (X) indicates the predicted presence of a bridge between the residues designated at the top of each column. (*), experimentally determined (Kauffman, 1965; Jurasek et al., 1969; Jurášek and Smillie, 1973); (#), pseudogene missing one of two cysteine codons.

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

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

3.11 INSERTIONS AND DELETIONS

Evolutionarily conserved insertions and deletions are expected to be rare events, and thus serve as good markers for tracking gene family phylogenies over large time scales. Figure 3.3 and Figure 3.22 display slightly different multiple alignments, both nearly optimal, which help illustrate possibilities for the orthology of the trypsinogen indels.

The tunicates share a single residue insertion at position 21 in common with rat trypsin IV. This event is most likely a coincidence, as it is found in no other vertebrates. A second tunicate insertion occurs near residues 45-51. This coincides precisely with the boundary between exons two and three. The crayfish shares this insertion (Titani et al., 1983). The lampreys also have an insertion at this site, but the lamprey insertion consists of only two residues. This is consistent with a progressive loss of residues at this site during early vertebrate evolution, with five lost prior to the Agnathan divergence, and another two lost prior to the elasmobranch divergence. These residues appear to be part of a surface loop (Figure 3.2). They are thus unlikely to have much functional significance, other than possibly a role in determining substrate specificity. The cold-adapted trypsinogens also have an insertion at this site (see Section 3.18.2). The variability in sequence length near position 45 is most likely due to junctional sliding and so independent indel events may be more likely here than elsewhere (see Section 3.12).

In the tunicate *Boltenia villosa*, an insertion of five residues occurs near positions 115-119. *Boltenia villosa* also has a deletion around position 165. The tunicate *Botryllus schlosseri* has insertions near positions 155 and 190. All tunicate sequences have a somewhat truncated carboxy-terminus, probably representing a deletion near position 223. None of these events is observed in any other vertebrate trypsin, consistent with the 600-700 million-year period of independent evolution since the urochordate/vertebrate split. The tunicate *Botryllus schlosseri* also has an insertion near positions 138, which, along with the indel polymorphism at position

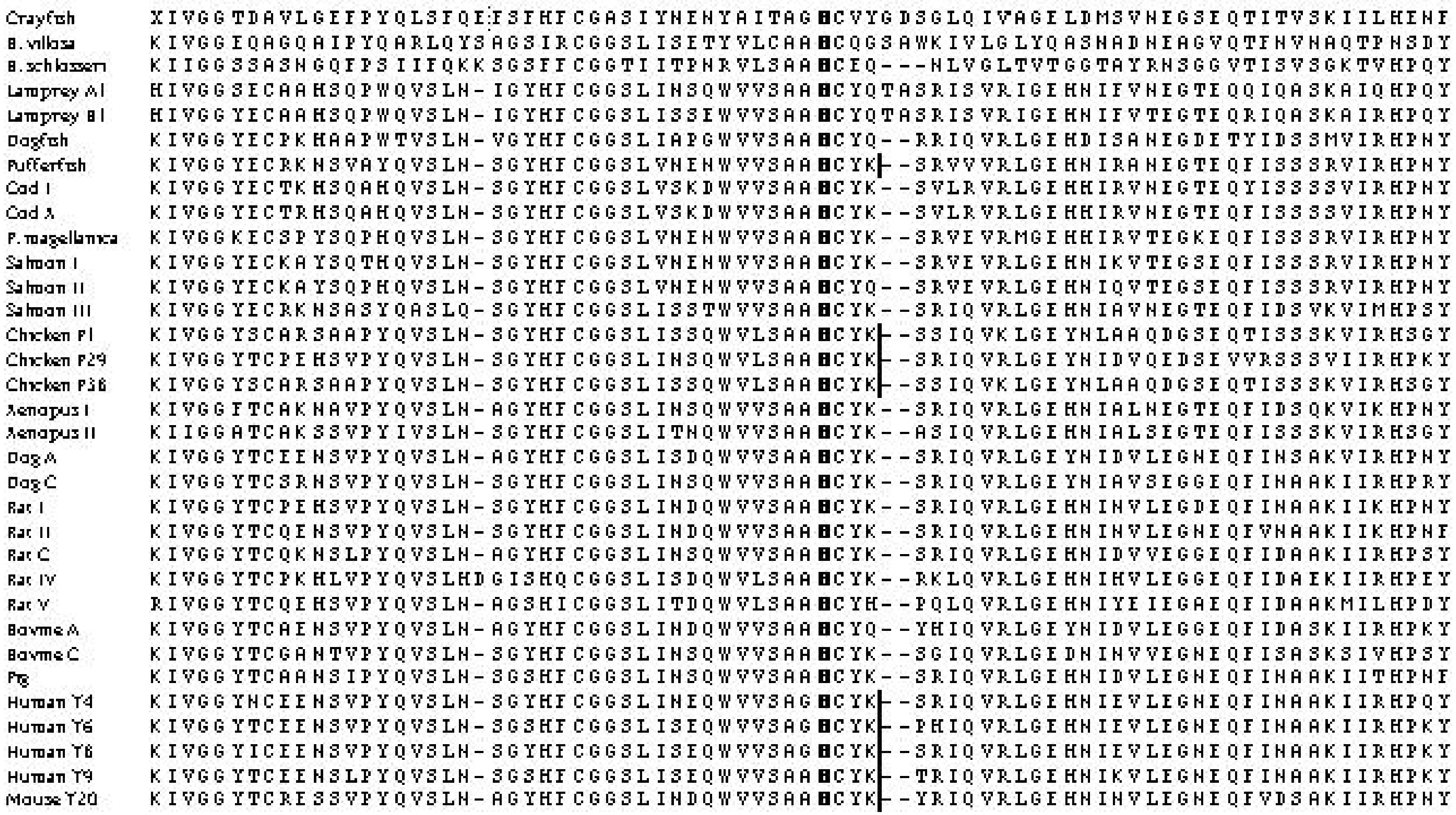
130 discussed in the next paragraph, is likely due to junctional sliding (see Section 3.12).

The tunicate *Botenia villosa*, lamprey, dogfish, and all but one of the Osteicthyes trypsins lack a residue at position 130 that is found in all other vertebrate trypsinogens as well as the tunicate *Botryllus schlosseri*. A residue is present at this position in salmon trypsin III. Therefore, most likely, there were at least two trypsinogen isozymes present in the common Osteicthyes/tetrapod ancestor, one of which gained a residue at position 130. Both variations were maintained by the Osteicthyes, but the insertion became the exclusive variant for the tetrapods, perhaps due to coincidental evolution and/or gene copy number contraction and expansion (Hood et al., 1975). Note that rat trypsin V lacks a residue near position 130. This most likely represents an independent deletion event, especially considering that this gene appears to have undergone rapid evolution in recent times. This recent “burst” of evolution is discussed further in Section 3.19. An alternative multiple alignment with only a minor loss in alignment score permits the rat V deletion to align precisely with the deletion noted in Osteicthyes. It is conceivable that these deletions descend from a common ancestor. However, in order for this hypothesis to be supported, the deletion would have to be present in tetrapod trypsins yet to be sequenced.

3.12 INTRON/EXON BOUNDARIES

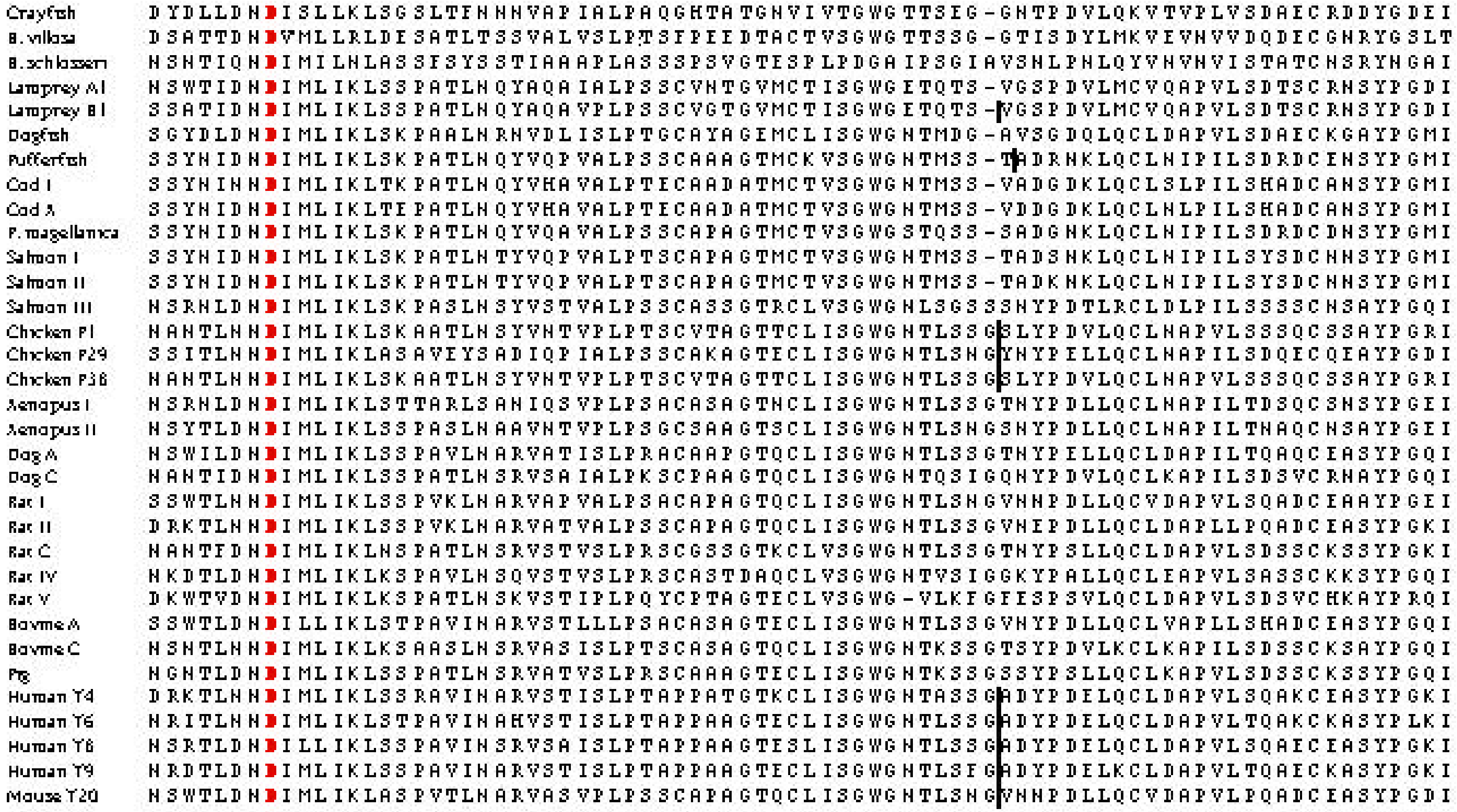
Point mutations that alter codons are not the only means of introducing variability into genes. Nevertheless, as discussed in Appendix B, most statistical models for evolution focus exclusively on point mutations. One potentially important mechanism for variability is junctional sliding. Junctional sliding refers to the reassignment of a splice acceptor or donor site for an intron. Junctional sliding has been referred to by other names, such as “intron shifting” or “intron sliding,” but this has resulted in confusion with a mechanism that reassigns both acceptor and donor sites simultaneously. The frequency, if any, of intron sliding is under debate (Stoltzfus et al., 1997). Junctional sliding is well documented (e.g., Mayo et al., 1985; Higashimoto and Liddle, 1993). Junctional sliding and intron positioning have been postulated to play a role in the development of diversity within the serine protease family (Craik et al., 1982; Craik et al., 1983; Rogers, 1983).

Intron/exon boundaries for the vertebrate trypsinogens, where known, are indicated in multiple alignment displays (Figure 3.3, Figure 3.22, and Table 3.4). These are deduced for trypsinogens with known genomic sequences (human, mouse, chicken, giant Antarctic toothfish, and lamprey). Of note is the absolutely conserved location of the boundary between exons four and five which occurs near the active site serine. The 1/2 exon boundary is also



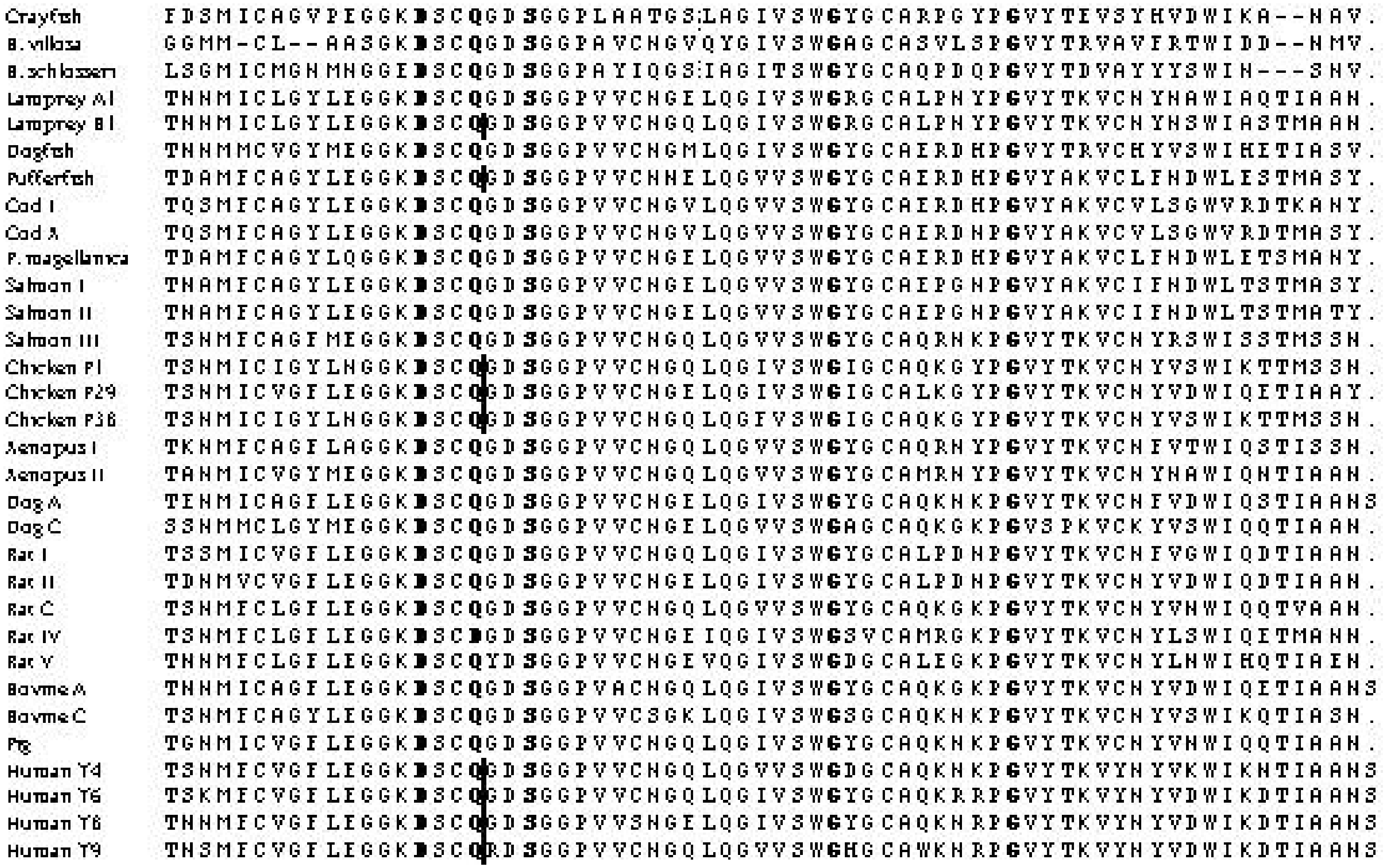
```
Crayfish        X I V G G T D A V L G E F P Y Q L S F Q E : F S F H F C G A S I Y N E N Y A I T A G ■ C V Y G D S G L Q I V A G E L D M S V H E G S E Q T I T V S K I I L H E N F
B. villosa      K I V G G E Q A G Q A I P Y Q A R L Q Y S A G S I R C G G S L I S E T Y V L C A A ■ C Q G S A W K I V L G L Y Q A S N A D N E A G V Q T F N V N A Q T P N S D Y
B. schlosseri   K I I G G S S A S N G Q F P S I I F Q K K S G S F F C G G T I I T P N R V L S A A ■ C E Q - - - N L V G L T V T G G T A Y R N S G G V T I S V S G K T V H P Q Y
Lamprey A1      H I V G G S E C A A H S Q P W Q V S L N - I G Y H F C G G S L I N S Q W V V S A A ■ C Y Q T A S R I S V R I G E H N I F V N E G T E Q Q I Q A S K A I Q H P Q Y
Lamprey B1      H I V G G Y E C A A H S Q P W Q V S L N - I G Y H F C G G S L I S S E W V V S A A ■ C Y Q T A S R I S V R I G E H N I F V T E G T E Q R I Q A S K A I R H P Q Y
Dogfish         K I V G G Y E C P K H A A P W T V S L N - V G Y H F C G G S L I A P G W V V S A A ■ C Y Q - - R R I Q V R L G E H D I S A N E G D E T Y I D S S M V I R H P N Y
Pufferfish      K I V G G Y E C R K N S V A Y Q V S L N - S G Y H F C G G S L V N E N W V V S A A ■ C Y K|- - S R V V V R L G E H N I R A N E G T E Q F I S S S R V I R H P N Y
Cod I           K I V G G Y E C T K H S Q A H Q V S L N - S G Y H F C G G S L V S K D W V V S A A ■ C Y K - - S V L R V R L G E H H I R V N E G T E Q Y I S S S S V I R H P N Y
Cod X           K I V G G Y E C T R H S Q A H Q V S L N - S G Y H F C G G S L V S K D W V V S A A ■ C Y K - - S V L R V R L G E H H I R V N E G T E Q F I S S S S V I R H P N Y
P. magellanica  K I V G G K E C S P Y S Q P H Q V S L N - S G Y H F C G G S L V N E N W V V S A A ■ C Y K - - S R V E V R M G E H H I R V T E G K E Q F I S S S R V I R H P N Y
Salmon I        K I V G G Y E C K A Y S Q T H Q V S L N - S G Y H F C G G S L V N E N W V V S A A ■ C Y K - - S R V E V R L G E H N I K V T E G S E Q F I S S S R V I R H P N Y
Salmon II       K I V G G Y E C K A Y S Q P H Q V S L N - S G Y H F C G G S L V N E N W V V S A A ■ C Y Q - - S R V E V R L G E H N I Q V T E G S E Q F I S S S R V I R H P N Y
Salmon III      K I V G G Y E C R K N S A S Y Q A S L Q - S G Y H F C G G S L I S S T W V V S A A ■ C Y K - - S R I Q V R L G E H N I A V N E G T E Q F I D S V K V I M H P S Y
Chicken P1      K I V G G Y S C A R S A A P Y Q V S L N - S G Y H F C G G S L I S S Q W V L S A A ■ C Y K|- - S S I Q V K L G E Y N L A A Q D G S E Q T I S S S K V I R H S G Y
Chicken P29     K I V G G Y T C P E H S V P Y Q V S L N - S G Y H F C G G S L I N S Q W V L S A A ■ C Y K|- - S R I Q V R L G E Y N I D V Q E D S E V V R S S S V I I R H P K Y
Chicken P38     K I V G G Y S C A R S A A P Y Q V S L N - S G Y H F C G G S L I S S Q W V L S A A ■ C Y K|- - S S I Q V K L G E Y N L A A Q D G S E Q T I S S S K V I R H S G Y
Xenopus I       K I V G G F T C A K N A V P Y Q V S L N - A G Y H F C G G S L I N S Q W V V S A A ■ C Y K - - S R I Q V R L G E H N I A L N E G T E Q F I D S Q K V I K H P N Y
Xenopus II      K I I G G A T C A K S S V P Y I V S L N - S G Y H F C G G S L I T N Q W V V S A A ■ C Y K - - A S I Q V R L G E H N I A L S E G T E Q F I S S S K V I R H S G Y
Dog A           K I V G G Y T C E E N S V P Y Q V S L N - A G Y H F C G G S L I S D Q W V V S A A ■ C Y K - - S R I Q V R L G E Y N I D V L E G N E Q F I N S A K V I R H P N Y
Dog C           K I V G G Y T C S R N S V P Y Q V S L N - S G Y H F C G G S L I N S Q W V V S A A ■ C Y K - - S R I Q V R L G E Y N I A V S E G G E Q F I N A A K I I R H P R Y
Rat I           K I V G G Y T C P E H S V P Y Q V S L N - S G Y H F C G G S L I N D Q W V V S A A ■ C Y K - - S R I Q V R L G E H N I N V L E G D E Q F I N A A K I I K H P N Y
Rat II          K I V G G Y T C Q E N S V P Y Q V S L N - S G Y H F C G G S L I N D Q W V V S A A ■ C Y K - - S R I Q V R L G E H N I N V L E G N E Q F V N A A K I I K H P N F
Rat C           K I V G G Y T C Q K N S L P Y Q V S L N - A G Y H F C G G S L I N S Q W V V S A A ■ C Y K - - S R I Q V R L G E H N I D V V E G G E Q F I D A A K I I R H P S Y
Rat IV          K I V G G Y T C P K H L V P Y Q V S L H D G I S H Q C G G S L I S D Q W V L S A A ■ C Y K - - R K L Q V R L G E H N I H V L E G G E Q F I D A E K I I R H P E Y
Rat V           R I V G G Y T C Q E H S V P Y Q V S L N - A G S H I C G G S L I T D Q W V L S A A ■ C Y H - - P Q L Q V R L G E H N I Y E I E G A E Q F I D A A K M I L H P D Y
Bovine A        K I V G G Y T C A E N S V P Y Q V S L N - A G Y H F C G G S L I N D Q W V V S A A ■ C Y Q - - Y H I Q V R L G E Y N I D V L E G G E Q F I D A S K I I R H P K Y
Bovine C        K I V G G Y T C G A N T V P Y Q V S L N - S G Y H F C G G S L I N S Q W V V S A A ■ C Y K - - S G I Q V R L G E D N I N V V E G N E Q F I S A S K S I V H P S Y
Pig             K I V G G Y T C A A N S I P Y Q V S L N - S G S H F C G G S L I N S Q W V V S A A ■ C Y K - - S R I Q V R L G E H N I D V L E G N E Q F I N A A K I I T H P N F
Human T4        K I V G G Y N C E E N S V P Y Q V S L N - S G Y H F C G G S L I N E Q W V V S A G ■ C Y K|- - S R I Q V R L G E H N I E V L E G N E Q F I N A A K I I R H P Q Y
Human T6        K I V G G Y T C E E N S V P Y Q V S L N - S G S H F C G G S L I S E Q W V V S A G ■ C Y K|- - P H I Q V R L G E H N I E V L E G N E Q F I N A A K I I R H P K Y
Human T8        K I V G G Y I C E E N S V P Y Q V S L N - S G Y H F C G G S L I S E Q W V V S A G ■ C Y K|- - S R I Q V R L G E H N I E V L E G N E Q F I N A A K I I R H P K Y
Human T9        K I V G G Y T C E E N S L P Y Q V S L N - S G S H F C G G S L I S E Q W V V S A A ■ C Y K|- - T R I Q V R L G E H N I K V L E G N E Q F I N A A K I I R H P K Y
Mouse T20       K I V G G Y T C R E S S V P Y Q V S L N - A G Y H F C G G S L I N D Q W V V S A A ■ C Y K|- - Y R I Q V R L G E H N I N V L E G N E Q F V D S A K I I R H P N Y
```

Figure 3.3. A multiple alignment of the chordate trypsin protein sequences. Intron/exon boundaries, where known, are indicated by a bar (|). The "catalytic triad" residues are red, and the four "pocket specificity" residues are blue. Residues in the invertebrate sequences that are not found in any vertebrate sequences are not shown. Numbers in parentheses correspond to the chymotrypsinogen numbering system. Dotted boundaries in the non-vertebrate sequences indicate undisplayed



```
Crayfish        DYDLLDNDISLLKLSGSLTFNNNVAPIALPAQGHTATGNVIVTGWGTTSEG-GNTPDVLQKVTVPLVSDAECRDDYGDEI
B. villosa      DSATTDNDVMLLRLDESATLTSSVALVSLPTSFPEEDTACTVSGWGTTSSG-GTISDYLMKVEVNVVDQDECGNRYGSLT
B. schlosseri   NSNTIQNDIMILNLASSFSYSSTIAAAPLASSSPSVGTESPLPDGAIPSGIAVSNLPNLQYVNVNVISTATCNSRYNGAI
Lamprey A1      NSWTIDNDIMLIKLSSPATLNQYAQAIALPSSCVNTGVMCTISGWGETQTS-VGSPDVLMCVQAPVLSDTSCRNSYPGDI
Lamprey B1      SSATIDNDIMLIKLSSPATLNQYAQAVPLPSSCVGTGVMCTISGWGETQTS-VGSPDVLMCVQAPVLSDTSCRNSYPGDI
Dogfish         SGYDLDNDIMLIKLSKPAALNRNVDLISLPTGCAYAGEMCLISGWGNTMDG-AVSGDQLQCLDAPVLSDAECKGAYPGMI
Pufferfish      SSYNIDNDIMLIKLSKPATLNQYVQPVALPSSCAAAGTMCKVSGWGNTMSS-TADRNKLQCLNIPILSDRDCENSYPGMI
Cod I           SSYNINNDIMLIKLTKPATLNQYVHAVALPTECAADATMCTVSGWGNTMSS-VADGDKLQCLSLPILSHADCANSYPGMI
Cod X           SSYNIDNDIMLIKLTEPATLNQYVHAVALPTECAADATMCTVSGWGNTMSS-VDDGDKLQCLNLPILSHADCANSYPGMI
P. magellanica  SSYNIDNDIMLIKLSKPATLNQYVQAVALPSSCAPAGTMCTVSGWGSTQSS-SADGNKLQCLNIPILSDRDCDNSYPGMI
Salmon I        SSYNIDNDIMLIKLSKPATLNTYVQPVALPTSCAPAGTMCTVSGWGNTMSS-TADSNKLQCLNIPILSYSDCNNSYPGMI
Salmon II       SSYNIDNDIMLIKLSKPATLNTYVQPVALPTSCAPAGTMCTVSGWGNTMSS-TADKNKLQCLNIPILSYSDCNNSYPGMI
Salmon III      NSRNLDNDIMLIKLSKPASLNSYVSTVALPSSCASSGTRCLVSGWGNLSGSSSNYPDTLRCLDLPILSSSSCNSAYPGQI
Chicken P1      NANTLNNDIMLIKLSKAATLNSYVNTVPLPTSCVTAGTTCLISGWGNTLSSGSLYPDVLQCLNAPVLSSSQCSSAYPGRI
Chicken P29     SSITLNNDIMLIKLASAVEYSADIQPIALPSSCAKAGTECLISGWGNTLSNGYNYPELLQCLNAPILSDQECQEAYPGDI
Chicken P38     NANTLNNDIMLIKLSKAATLNSYVNTVPLPTSCVTAGTTCLISGWGNTLSSGSLYPDVLQCLNAPVLSSSQCSSAYPGRI
Xenopus I       NSRNLDNDIMLIKLSTTARLSANIQSVPLPSACASAGTNCLISGWGNTLSSGTNYPDLLQCLNAPILTDSQCSNSYPGEI
Xenopus II      NSYTLDNDIMLIKLSSPASLNAAVNTVPLPSGCSAAGTSCLISGWGNTLSNGSNYPDLLQCLNAPILTNAQCNSAYPGEI
Dog A           NSWILDNDIMLIKLSSPAVLNARVATISLPRACAAPGTQCLISGWGNTLSSGTNYPELLQCLDAPILTQAQCEASYPGQI
Dog C           NANTIDNDIMLIKLSSPATLNSRVSAIALPKSCPAAGTQCLISGWGNTQSIGQNYPDVLQCLKAPILSDSVCRNAYPGQI
Rat I           SSWTLNNDIMLIKLSSPVKLNARVAPVALPSACAPAGTQCLISGWGNTLSNGVNNPDLLQCVDAPVLSQADCEAAYPGEI
Rat II          DRKTLNNDIMLIKLSSPVKLNARVATVALPSSCAPAGTQCLISGWGNTLSSGVNEPDLLQCLDAPLLPQADCEASYPGKI
Rat C           NANTFDNDIMLIKLNSPATLNSRVSTVSLPRSCGSSGTKCLVSGWGNTLSSGTNYPSLLQCLDAPVLSDSSCKSSYPGKI
Rat IV          NKDTLDNDIMLIKLKSPAVLNSQVSTVSLPRSCASTDAQCLVSGWGNTVSIGGKYPALLQCLEAPVLSASSCKKSYPGQI
Rat V           DKWTVDNDIMLIKLKSPATLNSKVSTIPLPQYCPTAGTECLVSGWG-VLKFGFESPSVLQCLDAPVLSDSVCHKAYPRQI
Bovine A        SSWTLDNDILLIKLSTPAVINARVSTLLLPSACASAGTECLISGWGNTLSSGVNYPDLLQCLVAPLLSHADCEASYPGQI
Bovine C        NSNTLNNDIMLIKLKSAASLNSRVASISLPTSCASAGTQCLISGWGNTKSSGTSYPDVLKCLKAPILSDSSCKSAYPGQI
Pig             NGNTLDNDIMLIKLSSPATLNSRVATVSLPRSCAAAGTECLISGWGNTKSSGSSYPSLLQCLKAPVLSDSSCKSSYPGQI
Human T4        DRKTLNNDIMLIKLSSRAVINARVSTISLPTAPPATGTKCLISGWGNTASSGADYPDELQCLDAPVLSQAKCEASYPGKI
Human T6        NRITLNNDIMLIKLSTPAVINAHVSTISLPTAPPAAGTECLISGWGNTLSSGADYPDELQCLDAPVLTQAKCKASYPLKI
Human T8        NSRTLDNDILLIKLSSPAVINSRVSAISLPTAPPAAGTESLISGWGNTLSSGADYPDELQCLDAPVLSQAECEASYPGKI
Human T9        NRDTLDNDIMLIKLSSPAVINARVSTISLPTAPPAAGTECLISGWGNTLSFGADYPDELKCLDAPVLTQAECKASYPGKI
Mouse T20       NSWTLDNDIMLIKLASPVTLNARVASVPLPSSCAPAGTQCLISGWGNTLSNGVNNPDLLQCVDAPVLPQADCEASYPGDI
```

Figure 3.3. Trypsin multiple alignment (continued).



```
Crayfish        FDSMICAGVPEGGKDSCQGDSGGPLAATGS:LAGIVSWGYGCARPGYPGVYTEVSYHVDWIKA--NAV.
B.villosa       GGMM-CL--AASGKDSCQGDSGGPAVCNGVQYGIVSWGAGCASVLSPGVYTRVAVFRTWIDD--NMV.
B.schlosseri    LSGMICMGNMNGGEDSCQGDSGGPAYIQGS:IAGITSWGYGCAQPDQPGVYTDVAYYYSWIN---SNV.
Lamprey A1      TNNMICLGYLEGGKDSCQGDSGGPVVCNGELQGIVSWGRGCALPNYPGVYTKVCNYNAWIAQTIAAN.
Lamprey B1      TNNMICLGYLEGGKDSCQGDSGGPVVCNGQLQGIVSWGRGCALPNYPGVYTKVCNYNSWIASTMAAN.
Dogfish         TNNMMCVGYMEGGKDSCQGDSGGPVVCNGMLQGIVSWGYGCAERDHPGVYTRVCHYVSWIHETIASV.
Pufferfish      TDAMFCAGYLEGGKDSCQGDSGGPVVCNNELQGVVSWGYGCAERDHPGVYAKVCLFNDWLESTMASY.
Cod I           TQSMFCAGYLEGGKDSCQGDSGGPVVCNGVLQGVVSWGYGCAERDHPGVYAKVCVLSGWVRDTKANY.
Cod A           TQSMFCAGYLEGGKDSCQGDSGGPVVCNGVLQGVVSWGYGCAERDNPGVYAKVCVLSGWVRDTMASY.
P. magellanica  TDAMFCAGYLQGGKDSCQGDSGGPVVCNGELQGVVSWGYGCAERDHPGVYAKVCLFNDWLETSMANY.
Salmon I        TNAMFCAGYLEGGKDSCQGDSGGPVVCNGELQGVVSWGYGCAEPGNPGVYAKVCIFNDWLTSTMASY.
Salmon II       TNAMFCAGYLEGGKDSCQGDSGGPVVCNGELQGVVSWGYGCAEPGNPGVYAKVCIFNDWLTSTMATY.
Salmon III      TSNMFCAGFMEGGKDSCQGDSGGPVVCNGQLQGVVSWGYGCAQRNKPGVYTKVCNYRSWISSTMSSN.
Chicken P1      TSNMICIGYLNGGKDSCQGDSGGPVVCNGQLQGIVSWGIGCAQKGYPGVYTKVCNYVSWIKTTMSSN.
Chicken P29     TSNMICVGFLEGGKDSCQGDSGGPVVCNGELQGIVSWGIGCALKGYPGVYTKVCNYVDWIQETIAAY.
Chicken P38     TSNMICIGYLNGGKDSCQGDSGGPVVCNGQLQGFVSWGIGCAQKGYPGVYTKVCNYVSWIKTTMSSN.
Xenopus I       TKNMFCAGFLAGGKDSCQGDSGGPVVCNGQLQGVVSWGYGCAQRNYPGVYTKVCNFVTWIQSTISSN.
Xenopus II      TANMICVGYMEGGKDSCQGDSGGPVVCNGQLQGVVSWGYGCAMRNYPGVYTKVCNYNAWIQNTIAAN.
Dog A           TENMICAGFLEGGKDSCQGDSGGPVVCNGELQGIVSWGYGCAQKNKPGVYTKVCNFVDWIQSTIAANS
Dog C           SSNMMCLGYMEGGKDSCQGDSGGPVVCNGELQGVVSWGAGCAQKGKPGVSPKVCKYVSWIQQTIAAN.
Rat I           TSSMICVGFLEGGKDSCQGDSGGPVVCNGQLQGIVSWGYGCALPDNPGVYTKVCNFVGWIQDTIAAN.
Rat II          TDNMVCVGFLEGGKDSCQGDSGGPVVCNGELQGIVSWGYGCALPDNPGVYTKVCNYVDWIQDTIAAN.
Rat C           TSNMFCLGFLEGGKDSCQGDSGGPVVCNGQLQGVVSWGYGCAQKGKPGVYTKVCNYVNWIQQTVAAN.
Rat IV          TSNMFCLGFLEGGKDSCDGDSGGPVVCNGEIQGIVSWGSVCAMRGKPGVYTKVCNYLSWIQETMANN.
Rat V           TNNMFCLGFLEGGKDSCQYDSGGPVVCNGEVQGIVSWGDGCALEGKPGVYTKVCNYLNWIHQTIAEN.
Bovine A        TNNMICAGFLEGGKDSCQGDSGGPVACNGQLQGIVSWGYGCAQKGKPGVYTKVCNYVDWIQETIAANS
Bovine C        TSNMFCAGYLEGGKDSCQGDSGGPVVCSGKLQGIVSWGSGCAQKNKPGVYTKVCNYVSWIKQTIASN.
Pig             TGNMICVGFLEGGKDSCQGDSGGPVVCNGQLQGIVSWGYGCAQKNKPGVYTKVCNYVNWIQQTIAAN.
Human T4        TSNMFCVGFLEGGKDSCQGDSGGPVVCNGQLQGVVSWGDGCAQKNKPGVYTKVYNYVKWIKNTIAANS
Human T6        TSKMFCVGFLEGGKDSCQGDSGGPVVCNGQLQGIVSWGYGCAQKRRPGVYTKVYNYVDWIKDTIAANS
Human T8        TNNMFCVGFLEGGKDSCQGDSGGPVVSNGELQGIVSWGYGCAQKNRPGVYTKVYNYVDWIKDTIAANS
Human T9        TNSMFCVGFLEGGKDSCQRDSGGPVVCNGQLQGVVSWGHGCAWKNRPGVYTKVYNYVDWIKDTIAANS
```

Figure 3.3. Trypsin multiple alignment (continued).

highly conserved.

The 2/3 and 3/4 exon boundaries both display evidence of junctional sliding. Most indels in trypsinogen evolution have occurred near these boundaries (Figure 3.3 and Figure 3.22). Junctional sliding is most likely responsible for the majority of indels during vertebrate trypsinogen evolution. Junctional sliding in highly conserved regions is not observed due to strong selection against such changes. The creation of indels by junctional sliding may be much more frequent than the creation of indels by other processes. Therefore such indels may be less reliable as “unique evolutionary events” for use in establishing evolutionary topology by parsimony (Section 3.11).

3.13 CATIONIC AND ANIONIC TRYPSINS

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

Steiner et al. (1997) have suggested that the cat expresses only a single trypsinogen isoform. This is based on extensive biochemical efforts that, they argue, would have identified a second trypsinogen, if present. This study was done without the aid of nucleic acid expression or sequencing data. Additionally, it could not rule out the possibility of multiple isoforms with similar biochemical properties. Furthermore, it would have missed trypsins expressed at different developmental or environmental timepoints, or in feline individuals not studied. Nevertheless, this work strongly supports the hypothesis that cats express only a single trypsin. This trypsin is cationic with a pI greater than 10.0, and therefore almost certainly group II. Hofer et al. (1975) biochemically identified multiple isozymes from three frog species, the trout *Salmo trutta*, the tench *Tinca tinca*, and the lizard *Lacerta muralis*, but only one isozyme from a fourth frog species, Rana temporaria. They only examined the tadpole of this last species.

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

Table 3.6. Predicted isoelectric points and charges of the chordate trypsins, calculated with the program *Protean* (DNA✻®, Madison, WI). Experimental values are in parentheses. Note that measured isoelectric points depend not only on the net charge, but also on the distribution of the charge (Smalås et al. 1994). Charge predictions do not take this into account.

Isozyme	Group	Isoelectric Point	Charge at pH 7.0
Tunicate		3.9	-13.18
Lamprey A1		5.2	-5.62
Lamprey B1		5.8	-3.62
Dogfish	I	4.9	-10.11
Cod I	I	6.8 (6.6)	-0.79
Cod X	I	5.8 (5.5)	-5.95
P. magellanica	I	5.8	-5.28
Pufferfish	I	6.2	-2.61
Salmon I	I	5.9	-3.62
Salmon II	I	5.5	-4.62
Salmon III	II	8.1	4.21
Chicken P1	II	8.2	5.03
Chicken P29	I	4.6	-9.78
Xenopus	I	6.7	-0.79
Xenopus 51	I	7.7	2.21
Dog A	I	4.9	-5.94
Dog C	II	8.3	6.04
Pig	II	7.9 (10.8)	3.21
Bovine A	I	4.8	-7.62
Bovine C	II	8.3 (10.1)	6.03
Human T4 (I)	I	7.5	1.28
Human T6	I	6.9	-0.39
Human T8 (II)	I	5.0	-6.65
Human T9 (III)	I	6.8	-0.55
Mouse T4	II	6.8	-0.62
Mouse T5	II	6.5	-1.62
Mouse T7	II	8.3	6.20
Mouse T8	I	6.3	-1.79
Mouse T9	I	5.9	-2.79
Mouse T10	I	6.7	-0.79
Mouse T11	I	5.2	-4.78
Mouse T12	I	5.6	-3.78
Mouse T15	I	5.0	-5.78
Mouse T16	I	5.0	-5.78
Mouse T20	I	4.4	-9.78
Rat I	I	4.9 (4.4)	-6.62
Rat II	I	4.8 (4.3)	-6.78
Rat C	II	8.1 (8)	4.20
Rat IV	II	6.9 (6.2)	-0.29
Rat V	II	5.1	-9.11

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

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

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

The relevance of the cationic and anionic groupings of the vertebrate trypsins to their phylogenetic origins is discussed in Section 3.18.

3.14 AMINO-ACID COMPOSITION[11]

Much of the early work on trypsin occurred before the advent of routine protein or DNA sequencing. However, the determination of amino-acid composition was relatively

[11] This section is new in the third printing.

straightforward. Therefore, the amino-acid compositions of many trypsinogens are known. Many of these trypsinogens have never been sequenced.

Before 1998, I briefly considered using amino-acid composition to assign unsequenced trypsins to clades. At that time, a cursory analysis indicated that composition alone would not be sufficient to determine clades with satisfactory confidence. I have subsequently re-examined this issue in light of additional sequence data. It now appears that some conclusions can be drawn from amino-acid composition data. I present them in this section.

For sequenced trypsins, amino-acid compositions are determined by enumerating hypothetically translated codon sequences, or in two cases, enumerating the residues of directly sequenced proteins. These, however, must be compared to enzymes with amino-acid compositions determined biochemically. Biochemical determinations historically neither distinguished between asparagine and aspartic acid nor between glutamine and glutamic acid. Thus, for the comparative analyses of this section, I have combined the counts for these two amino acid pairs. Additionally, some early biochemical determinations did not determine cysteine or tryptophan amounts. In these cases, I used a value equal to the average number of residues of cysteine or tryptophan for the sequenced trypsins. Studies of amino acid composition that lacked determinations of any other amino acids were not included. Thus, the very earliest studies of bovine-trypsin amino-acid composition were not included.

Compositions can be expressed as percentages by number of residues, or as absolute number of residues. The use of absolute number is often more suitable for phylogenetic cladistics, while the use of percentages is often more suitable for groupings based on biochemical properties. The analyses of this section are based on absolute counts. I did experiment with some analyses based on percentages, and observed no significantly different clusterings (data not shown).

Table 3.7 and Table 3.8 present the results of all published determinations of vertebrate trypsin amino acid compositions. References are provided in Table 3.9 for those compositions determined biochemically or by protein sequencing; other sequences were obtained from Genbank. For cases where the composition of trypsinogen was determined, the observed amino acid counts were adjusted for the sequence of the activation peptide. This is possible, as all publications that describe a composition of a trypsinogen also provide the sequence of the activation peptide of that trypsinogen. Additionally, Table 3.8 includes some composition data from enzymes described as “trypsin-like” by the authors of the relevant work. Table 3.7 also

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	B (D-N)	Z (E-Q)
Prostate Specific Antigen (Human)	11	10	11	10	3	21	10	10	9	24	5	6	17	12	10	21	12	20	8	7	17	22
Chymotrypsin B (Human)	21	9	15	6	7	21	2	11	15	17	2	9	11	9	5	21	17	22	8	2	24	15
Trypsin 2 (Human)	14	8	15	9	4	21	12	11	9	20	4	5	23	13	12	10	11	25	9	10	20	22
Shrimp (Penaeus vannamei)	19	9	20	12	6	28	5	17	5	12	2	10	11	12	3	22	11	18	5	10	30	24
Botryllus schlosseri 1	17	6	6	3	5	28	2	20	3	10	4	18	12	10	3	34	17	14	2	11	24	13
Botryllus schlosseri 2	24	6	7	6	5	31	2	15	4	12	5	16	8	8	3	27	19	16	2	9	23	14
Boltenia villosa 1	21	9	14	8	3	26	1	8	3	15	5	11	6	13	6	26	19	22	5	9	25	21
Lamprey A1	17	12	7	7	2	23	5	18	4	12	4	16	10	16	4	23	12	17	6	9	23	23
Lamprey B1	16	12	7	7	2	24	5	16	4	12	5	12	11	13	6	26	13	18	5	10	19	20
P. americanus 2	13	12	11	10	5	24	6	10	8	14	7	15	9	7	6	24	8	18	5	10	26	17
Dogfish	16	12	14	9	1	28	8	14	5	14	9	10	10	6	7	17	7	18	5	12	24	15
Pufferfish	15	12	10	10	4	21	5	10	8	14	6	17	8	7	9	24	7	19	4	12	27	17
Cod I	16	12	10	7	2	23	10	9	9	15	5	11	7	8	6	24	10	22	4	12	21	15
Cod X	15	12	12	8	3	23	9	9	6	15	6	11	7	8	7	24	10	22	4	11	23	16
P. magellanica	13	12	11	9	4	23	7	10	8	13	6	13	10	10	7	26	7	18	4	11	24	19
Salmon I	13	12	7	9	4	23	5	11	8	13	6	17	10	7	4	26	12	18	4	13	24	16
Salmon III	11	12	7	5	4	23	4	11	8	15	5	15	8	9	8	39	8	16	4	11	22	14
Chicken I	14	12	5	2	1	25	3	14	9	18	3	15	8	11	3	34	13	16	4	13	20	13
Chicken 29	13	12	9	12	2	23	4	19	7	17	2	12	10	11	4	24	7	17	4	14	21	23
Xenopus 31	15	12	7	4	6	23	4	14	8	17	2	20	9	13	5	24	12	15	4	9	27	17
Xenopus I	19	12	5	5	2	25	4	16	6	16	4	20	9	9	3	26	11	15	4	12	25	14
Dog A	17	12	8	10	4	23	3	17	7	17	2	16	10	13	5	20	9	16	5	10	24	23
Dog C	17	12	6	5	2	25	3	17	10	13	4	15	11	14	7	25	6	17	4	10	21	19
Rat I	17	12	10	7	4	24	5	15	7	17	2	17	13	11	3	19	7	20	5	8	27	18
Rat 2	15	12	12	8	4	23	4	13	9	19	2	18	12	11	4	17	8	20	4	8	30	19
Rat C	11	12	8	4	5	26	4	11	11	17	2	16	9	12	5	27	10	19	4	10	24	16
Rat 4	11	12	10	9	3	24	7	14	13	20	4	10	10	11	5	23	8	18	4	8	20	20
Rat 5	12	12	11	11	5	22	8	14	10	19	3	10	12	12	2	16	10	18	5	10	21	23
Bovine A	18	12	10	8	3	25	5	16	7	20	1	13	9	12	3	20	9	16	5	12	23	20
Bovine C	14	12	6	4	3	25	3	15	14	14	2	16	8	10	2	34	10	17	4	10	22	14
Pig (1975)	15	12	6	5	4	25	4	16	10	16	2	18	9	12	4	27	10	16	4	8	24	17
Human 4	15	10	8	10	4	24	4	15	13	14	2	17	10	11	6	20	9	18	4	10	25	21
Human 6	15	10	8	10	4	23	6	17	12	16	2	13	12	10	5	19	11	17	4	10	21	20
Human 8	15	8	9	13	4	24	4	17	9	16	1	15	11	9	6	24	7	17	4	11	24	22
Human 9	16	10	10	10	5	22	5	15	12	16	2	14	11	8	7	19	11	17	5	9	24	18
Mouse 4	11	12	11	8	3	24	5	15	12	18	4	12	10	11	6	24	8	18	4	8	23	19
Mouse 7	15	12	7	4	4	25	4	12	9	18	2	16	9	12	8	23	12	16	4	11	23	16
Mouse 8	15	12	9	6	5	23	4	14	8	19	2	22	11	11	5	17	9	19	4	8	31	17
Mouse 10	16	12	11	6	4	23	4	15	12	18	2	18	10	12	4	17	9	18	4	8	29	18
Mouse 11	16	12	10	8	5	23	4	12	8	20	2	18	12	11	5	16	10	19	4	8	28	19
Mouse 20	15	12	13	7	3	23	4	14	5	16	2	19	13	11	5	17	8	21	5	10	32	18
P. olivaceus 1	12	12	9	13	4	22	7	12	4	13	8	13	12	8	9	20	11	17	4	12	22	21
P. olivaceus 2	10	12	9	12	4	21	6	11	4	14	6	12	11	10	9	25	10	18	4	13	21	22
D. mawsoni	11	12	8	13	5	23	6	12	5	21	7	11	16	9	2	15	10	20	6	15	19	22
P. glauca	12	12	12	12	6	22	8	11	5	18	10	8	17	8	7	13	9	18	6	13	20	20
P. americanus 1	14	12	12	11	5	22	8	12	4	16	11	10	16	8	6	14	10	18	6	13	22	19
N. coriiceps	10	12	9	13	5	24	7	10	4	20	8	11	17	8	4	13	8	21	6	17	20	21
P. olivaceus 3	13	12	11	7	4	24	8	15	6	13	7	11	14	13	2	13	9	20	9	15	22	20

Table 3.7. Amino acid compositions of the trypsins as determined from sequence data. See text for details of the tabulation.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	B (D+N)	Z (E+Q)
Capelin I (Mallotus villosus)	7.82	5.71	ND	ND	2.3	12.83	2.98	3.64	4.25	6.22	2.28	ND	4.7	ND	2.34	10.61	4.02	6.63	ND	3.71	11.44	8.44
Capelin II (Mallotus villosus)	7.87	5.39	ND	ND	1.88	12.7	2.83	3.39	3.98	5.92	1.98	ND	4.13	ND	2.51	11	3.97	7.04	ND	3.37	11.37	10.3
Carp (Cyprinus carpio)	14	12	ND	ND	4	28	6	10	10	12	5	ND	9	ND	7	22	10	18	8	8	27	26
Catfish	13.3	12.2	ND	ND	5.6	24.4	7.4	13.2	11	13	6.1	ND	12.6	ND	5.2	23.1	8.9	18.1	7.3	9.1	24.9	21.9
Chum salmon (Oncorhynchus keta)	15	12	ND	ND	4	25	3	9	8	13	6	ND	12	ND	2	23	14	16	5	10	22	14
Cow (Bos taurus) (1952)	12.10	ND	ND	ND	2.76	20.26	2.86	11.98	11.14	12.02	1.83	ND	8.26	ND	2.09	26.16	8.14	13.37	5.23	8.14	18.77	12.10
Cow (Bos taurus) (1954)	13	ND	ND	ND	4	21	3	12	14	12	1	ND	7	ND	2	38	10	15	ND	9	24	10
Cow (Bos taurus) (1980)	14.2	11.5	ND	ND	3.02	25	3.1	12.3	13.7	14.1	1.2	ND	8.2	ND	2.05	32.4	9.6	11.8	ND	9	22	14.2
Cow (Bos taurus) (1984)	14	12	10	3	3	25	3	15	15	14	2	16	9	11	2	33	10	18	4	10	26	14
Cow Anionic (Bos taurus) (1974)	19	12	ND	ND	3	25	5	17	8	21	2	ND	9	ND	4	17	10	16	7.5	12	26	24
Dromedary (Camelus dromedarius)	14	8	ND	ND	3	20	4	12	14	14	2	ND	9	ND	5	26	10	16	3.5	8	22	14
Elephant seal (Mirounga leonina L.)	15	14	ND	ND	5	27.5	4	18	16	20	3	ND	12	ND	5	26	15	18	5	11	24	25
Goat (Capra hircus)	14.9	11.9	ND	ND	3.34	25.3	3.09	13.9	12.47	14.55	2	ND	9.8	ND	2.53	32.5	10.4	18.7	3	10.7	25.2	17.9
Greenland Cod (Gadus ogac)	16.07	8.05	ND	ND	3.39	27.34	7.32	7.87	5.91	13.89	3.06	ND	9.84	ND	4.92	31.7	9.37	16.04	1.33	6.78	22.85	18.67
Human	14	8	ND	ND	4	21	3	13	12	13	1	ND	10	ND	6	25	11	17	3	7	22	22
Human Anionic	15	8	ND	ND	6.8	22.6	5.2	12.5	12	17.9	2.3	ND	11.7	ND	6.7	19.8	10.6	14.9	3	7.6	24.6	24.2
Lungfish A (Protopterus aethiopicus)	16.4	11.7	ND	ND	3.1	23.7	4	17.9	5.5	15	2	ND	8.6	ND	5.5	25.6	7.6	13.9	5.1	11.9	27	18
Lungfish B (Protopterus aethiopicus)	11.7	11.3	ND	ND	2.55	24.5	6.17	14.6	6.57	13.8	2.87	ND	10.4	ND	5.42	30	8.9	18.6	5.3	11.8	23	16.6
P. magellanica 2	13	16	ND	ND	4	27	5	8	13	13	5	ND	10	ND	5	26	11	16	3	8	29	20
Pig (Sus scrofa) (1981)	15.57		ND	ND	4.206	25.21	3.811	15.05	9.756	16.17	2.018	ND	9.855	ND	4.11	25.21	10.05	15.77		8.021	25.92	16.86
Pig (Sus scrofa) (1982)	12.10	ND	ND	ND	3.94	16.90	3.10	9.96	8.38	12.00	1.27	ND	11.02	ND	3.60	13.82	9.12	11.89	5.16	4.83	18.10	13.82
Pig (Sus scrofa) (1985)	15.6	11.8	ND	ND	4	26.3	3.8	14.7	9.8	15.9	1.9	ND	9.7	ND	3.9	24.2	10.7	16.1	5.8	7.9	17.7	16.8
Pig Anionic (Sus scrofa) (1971)	16.6	8.3	ND	ND	4.1	20.7	4.2	10.3	8.3	20.7	2.1	ND	10.3	ND	4.1	16.6	20.8	12.4	8	8.3	24.8	16.6
Pig Anionic (Sus scrofa) (1974)	20	12	ND	ND	4	24	4	14	8	22	2	ND	12	ND	6	20	8	16	7.5	11	26	23
Pig Cationic (Sus scrofa) (1983)	15.55	12	ND	ND	4.2	24.7	3.85	15.15	9.75	16	2	ND	9.8	ND	4.05	25.2	10.05	15.8	4	7.85	24	16.9
Rat 1	16.5	9.8	ND	ND	4.8	22.4	5	13.4	9.8	17.4	2	ND	15.2	ND	3.2	15.1	8.4	20.2	3.9	6.8	28.7	19.6
Rat 2	14.1	11.3	ND	ND	5.2	22.8	5.3	10.3	12.2	17.5	1.7	ND	11.6	ND	4.1	22	10.8	20.3	4.8	8	25.9	16.7
Red Deer (Cervus elaphus)	17	10	ND	ND	4	25	3	14	12	14	2	ND	10	ND	2	29	11	19	ND	10.5	26	17
Roe Deer 1 (Capreolus capreolus)	15	9	ND	ND	3	25	3	13	10	14	2	ND	10	ND	3	29	11	18	ND	10.5	24	18
Roe Deer 2 (Capreolus capreolus)	15	10	ND	ND	3	25	3	13.5	10.5	14	2	ND	9	ND	3	29	11	18	ND	10	23	17
Rorqual (Balaenoptera acutorostrata)	16.7	11	ND	ND	4.7	22.9	4.7	12.9	11.1	19	3.7	ND	9.5	ND	5.9	18.7	9.6	15.3	4.2	7.6	19.3	17
Sardine I (Sardinops melanosticta)	14	6.9	ND	ND	4.1	21.1	3.2	9.7	10.2	12.8	0.7	ND	9.9	ND	5.6	12.9	17.5	12.2	4.3	6.9	19.1	20.4
Sardine II (Sardinops melanosticta)	15	12.6	ND	ND	3.5	29.7	2.8	7.7	18.6	10.1	0.9	ND	13	ND	6	16.2	17.1	21.1	4.9	14	23	18.6
Sardine III (Sardinops melanosticta)	14.8	7.8	ND	ND	4.3	27.4	3.2	10.7	17.4	9.5	0.8	ND	9.5	ND	6	16.9	18.6	17.9	4.4	13	30.6	16.4
Sheep (Ovis aries) (1988)	17	12	ND	ND	5	19	3	10	12	14	2	ND	9	ND	4	26	15	17	ND	6	20	14
Sheep (Ovis aries) (1989)	15.8	12	ND	ND	3.34	25.9	3.13	14.2	12.6	14.5	2	ND	9.1	ND	3.16	32.2	10.4	17.3	4	10.6	24.2	13.83
Sheep (Ovis aries) (1972)	12.29	ND	ND	ND	2.39	19.22	2.88	10.75	9.55	12.05	1.85	ND	8.28	ND	2.55	23.21	8.18	13.39	6.19	8.74	18.79	11.95
Spiny dogfish (Squalus acanthias)	16	11.9	ND	ND	1.2	27.8	7.6	13.5	5.3	14.6	8.1	ND	10.3	ND	6.5	16.5	7.3	17.4	5.4	11.5	23.9	15.5
Trout (Salmo gairdneri)	19	11	ND	ND	6	26	4	11	12	19	3	ND	11	ND	3	19	12	16	2	7	27	23
Turkey	11	7.7	ND	ND	1	19.3	2.8	9.7	6.7	13.4	1.4	ND	7.7	ND	2.4	22.8	10.7	14.4	5	10.2	15.5	16.1

Table 3.8. Amino acid compositions of the trypsins as determined biochemically. See text for details of the tabulation.

provide some compositions of select invertebrate trypsins and three other human serine proteases for reference.

Amino-acid composition data lends itself well to principal-component analysis. Principal-component analysis is a statistical method for determining combinations of variables that account for the principal components of variation in a data set. This permits visualization of eighteen-dimensional data in two dimensions, in a manner similar to multidimensional scaling.

I performed a principal-component analysis for trypsins with known sequences. I excluded biochemically determined compositions from this initial analysis. Biochemical data contains a large degree of experimental variation. My objective was to quantify evolutionary variation within sequence composition space, so it made sense initially to exclude data with experimental variation from this preliminary analysis. Subsequent to this preliminary analysis, I transformed the rest of the data set of Table 3.7 into the same space as the reference trypsins using the previously determined transformation matrix. The combined result of these transformations is shown in Figure 3.4.

Overall, Figure 3.4 supports the hypothesis that all trypsins have more or less the same composition. However, some conclusions can be drawn.

The biochemical determinations of bovine trypsin composition of 1954, 1960, 1962, and 1964 all orbit about the composition of cationic bovine trypsin (CowC in Figure 3.4). These determinations were most likely all performed on the same enzyme, although it remains possible that they were performed on slightly different isozymes or mixtures of isozymes. Thus the orbit of these data points provides a measure of the range of experimental error for biochemical determinations.

Similarly, the 1974 determination of bovine trypsin composition most likely corresponds to bovine anionic trypsin, a group I trypsin. This is consistent with its anionic nature (Louvard and Puigserver, 1974).

The determinations of pig trypsin composition of 1961, 1963, and 1965 correspond to the pig trypsin sequenced at the protein level in 1973, a group II trypsin. The determinations of pig trypsin composition of 1962, 1971, and 1974 correspond to one or more group I trypsin isozymes that have never been sequenced. This provides additional evidence that the vast majority of vertebrates possess at least one group I and one group II trypsin isozyme. In fact, this represents much of the original evidence of the existence of two groups of trypsins.

Table 3.9. References for trypsin amino acid composition data. The first two sequences are references to amino acid sequences. The capelin and sardine data are putative trypsins based on biochemical properties.

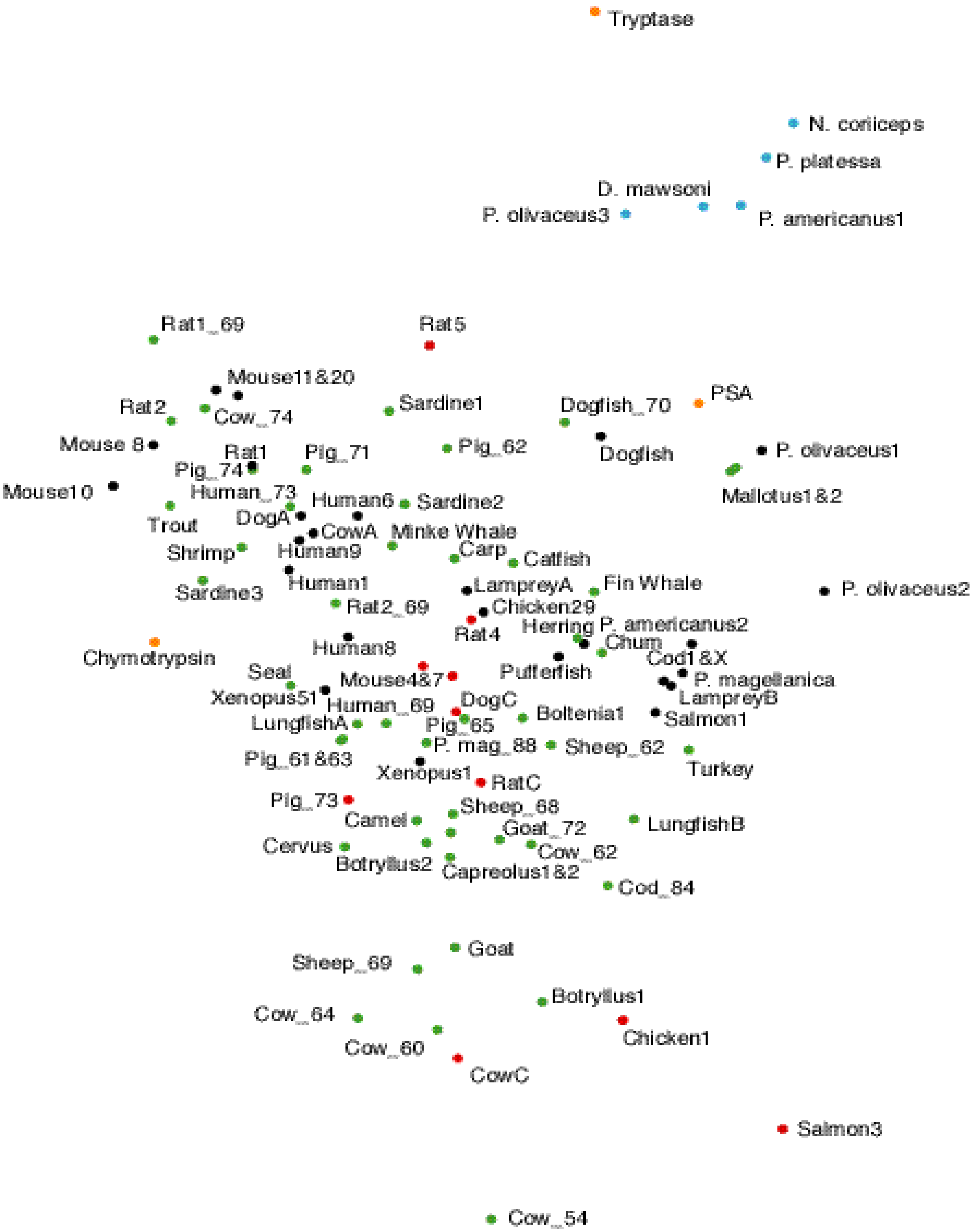


Figure 3.4. Principal-component analysis of trypsin amino-acid compositions. The first principal component is the vertical axis, the second is the horizontal axis. Black, group I trypsins; Red, group II trypsins; Blue, psychrophilic trypsins; Orange, reference sequences; Green, enzymes with biochemically determined compositions.

Voytek and Gjessing (1971) provide an early discussion of the elucidation of the differences between cationic and anionic trypsins.

The two human trypsin compositions of 1969 and 1973 and the two rat trypsin compositions of 1969 cannot be unambiguously assigned.

It is unclear if the determinations of sheep trypsin composition of 1962, 1968, and 1969 all correspond to the same isozyme, but it is possible, assuming wide experimental variation.

In general, it is not possible to unambiguously assign a trypsin to a group based solely on its amino-acid composition. However, the cow, chicken, and salmon group II sequences exclusively cluster towards the bottom of Figure 3.4. The pig and rat cationic trypsins also possess a lower first principal component than any group I trypsin, although not to the same degree as the cow, chicken, or salmon group II sequences. Therefore, it is reasonable to conclude that sequences with a very low first principal component are group II trypsins. Additionally, group II trypsins tend to have second principal components that are close to zero. With these considerations, it seems reasonable to conclude that the red deer, roe deer, camel, sheep, and goat isozymes in Figure 3.4 are all group II trypsins. This is consistent with the cationic nature of these trypsins. The cod isozyme composition determined in 1984 is also likely to represent a group II trypsin.

Lungfish trypsin A and lungfish trypsin B appear to occupy significantly different regions of trypsin composition space and therefore probably represent group I and group II, although assignments as to which is which cannot be made on the basis of these data.

The region of composition space occupied by the trypsins is not exclusive to trypsin isozymes. Therefore one cannot determine whether or not an enzyme is a trypsin solely on the basis of its composition. This point is illustrated by the presence of both prostate specific antigen and chymotrypsin within the sphere of composition space occupied by the trypsins. Tryptase lies well outside this sphere, however, so composition does provide some clue to clade and function. Thus, although one cannot be certain that the capelin and sardine sequences are trypsins based solely on their amino-acid compositions, these compositions are at least consistent with those of the known trypsins. Thus the original hypotheses that these isozymes are trypsins are supported by these composition data (Murakami and Noda, 1981; Hjeleland and Raa, 1982).

Notably, the five trypsins that I have called “psychrophilic trypsins” all cluster together unambiguously based solely on composition data. From a phylogenetic perspective (Section 3.18), they are less distinct from the majority of the vertebrate trypsins than are the invertebrate trypsins such as the shrimp and the three tunicate trypsins that have been provided as reference in Figure 3.4. Thus it is quite surprising that the psychrophilic trypsins cluster so distinctly. It should therefore be possible to distinguish a psychrophilic trypsin from the other trypsins based solely on amino-acid composition. This observation was a major motivation for my tabulation of composition data, as I sought to identify additional psychrophilic trypsins that might have associated biochemical kinetic measurements. However, it appears that none of the biochemically studied trypsins belong to this novel psychrophilic group.

It seems likely that selective pressure accounts for the amino-acid composition difference of the psychrophilic trypsins from the other chordate trypsins. Given the close clustering of invertebrate trypsins and even chymotrypsin with the majority of the vertebrate trypsins, it would be hard to explain the divergence of the psychrophilic trypsin compositions as due to neutral drift. The most likely selective pressure is adaptation to cold. One may therefore hypothesize that the amino acids weighted heavily in the first component are important in cold adaptation.

The first principal component accounts for 39.3% of the variation in the reference set of trypsin compositions. The second principal component accounts for 18.9% of that variation. The residues that contribute most strongly to the first principal component are, in order of eigenvalue magnitude: -0.84S +0.31P +0.24Z -0.19K +0.16M +0.13H +0.12L +0.12V. The residues that contribute most strongly to the second principal component are, in order of eigenvalue magnitude: 0.48B -0.45M -0.34Y +0.31I +0.31K +0.27A -0.23S -0.22H +0.18L. From my point of view, these eigenvector coefficients do not immediately suggest a mechanism for cold adaptation.

One may hypothesize that GC content of the underlying DNA sequence might play a role in determining amino-acid content. To this end one notes that residues F, L, Y, I, M, B, and K all have adenine or thymine in the first two positions of their codons, while P, R, A, and G all have guanine or cytidine in the first two positions of their codons. Neither of the first two eigenvalues seems to be dominated by considerations of underlying GC content. Note that the coefficient of methionine in the second principal component is negative, the opposite of what would be expected in a strongly GC-content determined eigenvector.

If one desired a more robust statistical test for the inclusion of an enzyme in the clade of psychrophilic trypsins based on amino-acid composition, discriminant analysis would be superior to principal-component analysis. My principal-component-analysis separation of the mesophilic and psychrophilic trypsins is expected to be conservative. I have yet to perform discriminant analysis.

3.15 MULTIPLE SEQUENCE ALIGNMENTS

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

The potential difficulty in sequence alignment is uncertainty in which positions of the sequences correspond to each other. The sequences may start or end at different positions. They may be different lengths. One sequence may have insertions or deletions with respect to the other. The similarity between the sequences may not be sufficient to recognize across the whole length of the sequences, but may be confined to one or a few small regions. Thus multiple alignment is often extremely difficult. To address this difficulty, a number of algorithms have been developed (e.g., *HMMER*, *CLUSTAL W*, and several others).[12] However, none of these algorithms are necessary for the trypsinogens. Vertebrate trypsin sequences can be aligned manually.

The trypsinogen multigene family possesses a number of features that make it particularly amenable to multiple sequence alignment. These include a cleavage site following an activation peptide, six cysteine residues necessary to form the three absolutely conserved cystine bridges, four active-site pocket-specificity residues embedded in conserved sequences, three catalytic residues embedded in conserved sequences, and several other highly conserved sequences. These conserved sequences are spread throughout the length of the protein, allowing members of the multigene family to be easily aligned, as regions of low similarity are inevitably flanked by conserved residues.[13] Variations in sequence length between two conserved residues can be recognized as insertions or deletions.

There is considerable variation in overall sequence length between the various subfamilies of the serine protease gene family, including some of the invertebrate trypsins, as

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

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

discussed in Section 3.7. However, there is very little variation in overall length within the vertebrate trypsinogens. Insertions and deletions have been rare during vertebrate trypsinogen evolution. Those that do exist are either one or two residues long. Both the amino- and carboxy-termini of trypsin are conserved, allowing for no uncertainty in aligning the protein ends.

The multiple alignment used for most of my analyses was performed at the amino acid level. This was necessary, as there are two vertebrate trypsinogens for which only amino acid sequence exists: the dogfish and pig sequences.[14] These sequences represent key nodes in the vertebrate trypsinogen phylogeny so it would be limiting to restrict an analysis solely to known nucleic acid sequences. The nucleic acid sequences were incorporated into these alignments as hypothetical translations.

In some cases complete pretrypsinogen sequences are not available. For example, the protein sequences of the pig and dogfish trypsinogens do not include signal sequences. Therefore, I have limited my formal analysis to multiple alignments of the portions of the sequences coding for the mature trypsin peptide. Also, activation peptides vary in length, which would lead to ambiguity in precise alignment and distance calculations for complete trypsinogen or pretrypsinogen sequences (see Section 3.9). However, the last residue of the activation peptide, which is always a lysine, arginine, or histidine, can be unambiguously included in multiple alignments. The last residue of the activation peptide has therefore been included in my alignments, as it provides a small amount of additional phylogenetic information.

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

All trypsins contain six absolutely conserved cysteine residues, which are necessary to

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

build the three cystine bridges observed in all serine proteases (see Section 3.10). These cysteines are at positions C42, C58, C168, C182, C191, and C220.

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

All trypsins contain four key “pocket specificity” residues: aspartate, glutamine, glycine, and glycine at positions C189, C192, C217, and C227 (Ken Walsh, personal communication). These four residues distinguish the trypsins from all other serine proteases. Hedstrom et al. (1994) provide a more extensive discussion of the sequence characteristics that determine the catalytic specificity of trypsin. Perona and Craik (1995 and 1997) also provide excellent reviews.

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

All vertebrate trypsins possess a calcium binding site on a “calcium loop,” characterized by the residues glutamate, asparagine, valine, glutamate, and glutamate at positions 56, 58, 61, 63, 66 (Bode and Schwager, 1975). The negatively charged acidic residues chelate the positively charged calcium ion. The role of calcium binding in trypsin was discussed in Section 3.7.

3.16 SEQUENCE DISTANCES

Once sequences have been aligned, evolutionary distances between them can be determined.[15] There are many methods for calculating evolutionary distance. Most of these are algorithms that operate on a pair of sequences. Some, based on maximum likelihood, operate on an entire data set. Maximum-likelihood methods are favored, but are computationally intensive, and so may not be possible with large data sets (Felsenstein, 1983). Maximum-likelihood algorithms operating on nucleic acid sequences are more advanced than those that operate on protein sequences.[16] Both because of lack of computational resources

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

and lack of implemented protein algorithms, most of my trypsin analyses were done with pairwise distance methods.

For multidimensional scaling, described in Section 3.17, and phylogeny construction and jackknifing, described in Section 3.18, I used distances derived from the program *protdist*, part of the *PHYLIP* package (Felsenstein, 1993). *Protdist* was executed with the “Dayhoff” algorithm, which utilizes Dayhoff’s PAM 001 matrix (Dayhoff, 1979). I chose this simple algorithm for this purpose for its ease of use and speed of execution. The distances used for the phylogeny in Appendix B (Figure B.3) and the statistics of coincidental evolution (Figure 3.17) were calculated with the algorithm described in Appendix B. In no case did I observe a qualitative difference in results when different distance algorithms were employed.

Distances calculated between pseudogenes were not generally considered. The model for distance calculation assumes that all genes are evolving under the same selective constraints. Pseudogenes evolve with a near lack of selective pressure, so it would be inappropriate to include them in distance calculations with functional genes. It is true that the assumption of uniform selective pressure is violated even by the functional trypsinogens. However, differences in selective pressure between functional genes will be negligible compared to a complete absence of selection. After construction of a phylogeny, pseudogenes can be assigned topological locations based on similarity or identity comparisons.

The differences in selective pressure operating on the functional trypsinogens may prevent accurate reconstruction of a phylogeny. This would occur if the rates of evolution in different branches of the phylogeny were skewed to such an extent as to warp the tree topology. This may indeed have happened, as discussed in Section 3.18.

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

The aligned vertebrate trypsin amino acid sequences contain 228 sites (Section 3.15). As discussed in Appendix B, 115 of these sites are highly or absolutely conserved. Most of the variation from sequence to sequence occurs at the other 113 sites. Thus, almost exactly fifty

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

percent of the vertebrate trypsin sequence is highly conserved. This suggests that two infinitely diverged vertebrate trypsin sequences will still share high identity.

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

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

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

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

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

One additional problem impedes distance calculations between trypsins. As mentioned above, the aligned vertebrate trypsin amino acid sequences are only 228 residues long. The corresponding nucleic acid sequences are three times that long, with 684 nucleotides. However, accurate reconstruction of the topologies of complex phylogenies is hypothesized to require about 2,000 sites, even in “easy” cases (Hillis, 1996). Thus one expects some uncertainty in a distance-based topology generated for the trypsin phylogeny. Such uncertainties can be evaluated with bootstrapping or jackknifing, as discussed in Section 3.18. Recent advances in tree-construction algorithms may ease the recovery of correct tree topologies for short sequences (Tandy Warnow, personal communication).

3.17 MULTIDIMENSIONAL SCALING

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

An introduction to the subject of clustering is provided by Everitt (1993). In many cases it is not possible to provide rigorous statistical support for or against alternative clusterings of data. Therefore, one of the main goals of cluster analysis is to provide hypotheses that must be confirmed with external data. There may be many alternative hypotheses. For example, there are 2.2x10^{12} ways to cluster the 42 known vertebrate trypsin sequences into 2 groups. This can be calculated as follows (Liu, 1968):

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

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

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

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

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

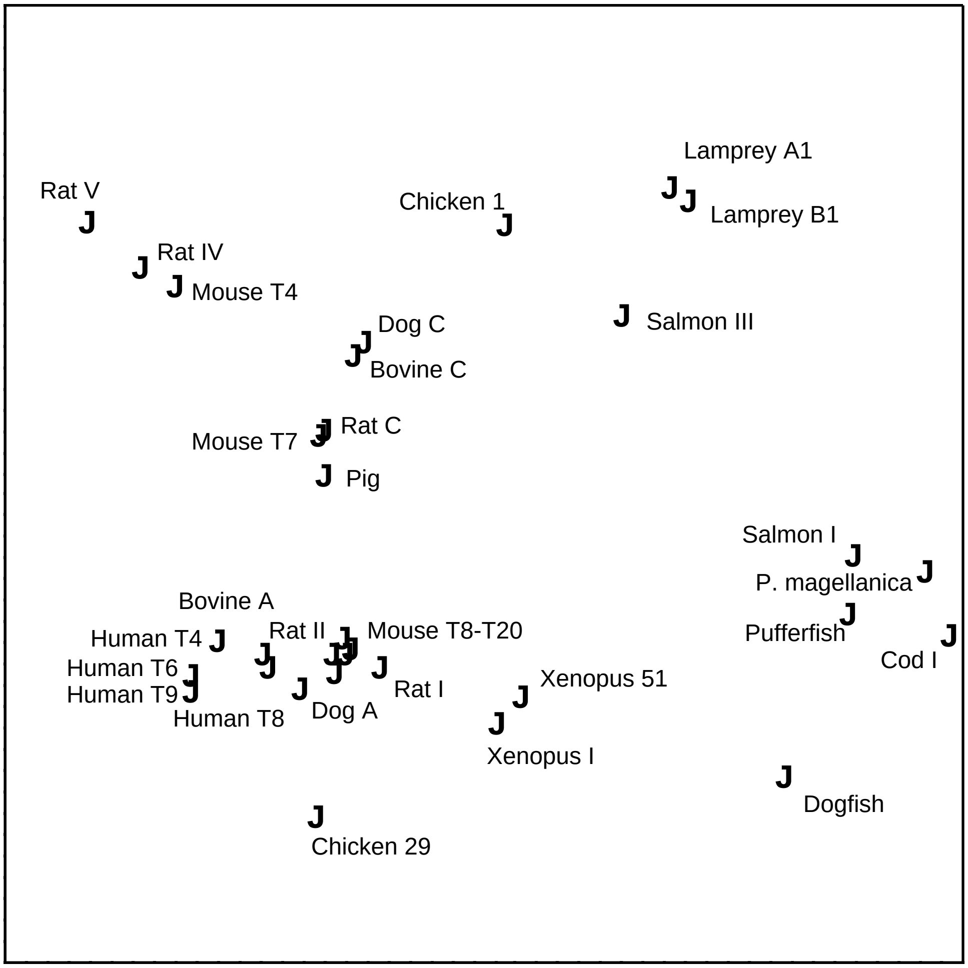


Figure 3.5. Multidimensionally scaled vertebrate trypsin sequence distances. Each point represents a trypsin sequence; The distance between two points corresponds to the calculated phylogenetic distance between the corresponding sequences. The program *SPSS*® *7.5* (SPSS, Inc.) was used to scale the trypsin pairwise distance matrix (from *protdist*) as ratio data with a Euclidean distance model. Iterations continued until the S-stress altered by less than 0.0001 between iterations.

in the process, provide evidence for the occurrence of coincidental evolution.

Multidimensional scaling of the vertebrate trypsin distances is shown in Figure 3.5. Only 32 of the 42 known sequences are utilized; each of the remaining 10 sequences is nearly identical to one of the included sequences. Note that multidimensional scaling is invariant to orthogonal transformations, which include rotations and magnifications. Therefore the rotational positioning of the axes is arbitrary, as are the units of distance. The informational content of the plot lies in the relative distances between points. If a “molecular clock” hypothesis held, as discussed in Appendix B, then these distances could be interpreted as units of time.

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

Figure 3.5 represents a global minimization but gives little insight into alternative possible local minima. Given the high variance of distance data from short sequences, and the large number of points, alternate minima may represent equally valid depictions of the data. However, I am unaware of programs that explore alternative local minima. Therefore, I would welcome the addition of such options in relevant computer programs, perhaps implemented with a simulated-annealing algorithm. I currently have no good method for evaluating alternative minima, so will not discuss it further.

The simplified subset of trypsin data is inherently 31-dimensional, one dimension less than the number of plotted sequences. Therefore, informational content is lost as the data is “compressed” into a lower-dimensional space. This loss in informational content can be characterized by the “stress” statistic, which is based on the squared differences between the original and the scaled distances (Cox and Cox, 1994). The stress for the multidimensionally scaled plot in Figure 3.5 is graphed in Figure 3.6. The stresses of the two-dimensional plots in Figure 3.5 and Figure 3.13 are below 25%, and thus low enough to indicate that these plots adequately portray the underlying structure of the data (Everitt and Dunn, 1991). The utility of this plot is supported by the gradual rise in stress through two dimensions, with a large increase occurring only between two dimensions and one dimension. This suggests that although one dimension is not enough to portray the data, two dimensions is sufficient. It is nevertheless interesting to examine the data in three dimensions. Views of a three-dimensional scaling of the data in Figure 3.5 are provided in Figure 3.9 and Figure 3.10.

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

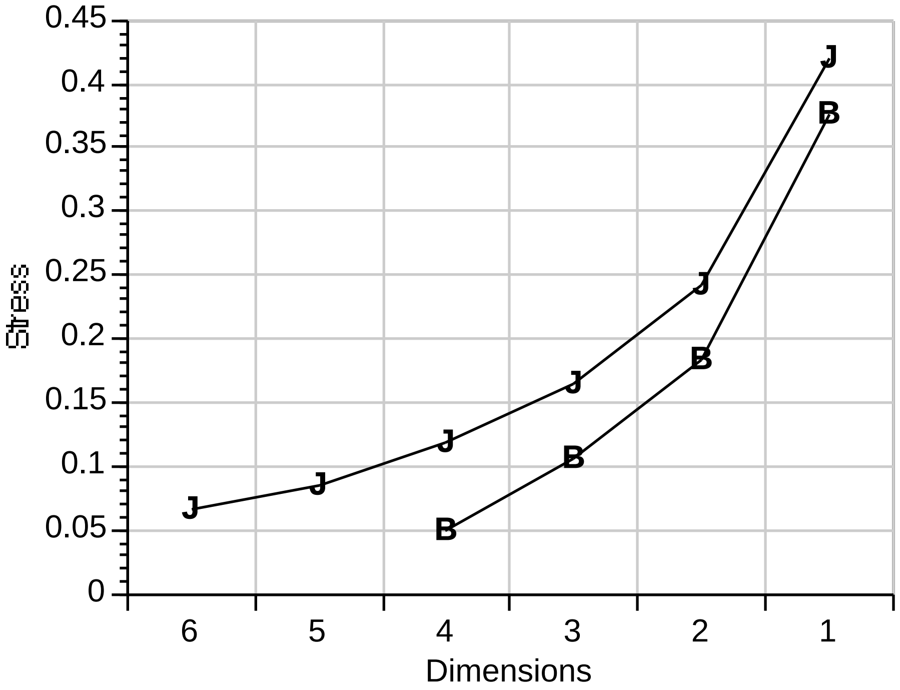


Figure 3.6. Stress vs. Dimensions. Data for Figure 3.5 (**J**) and Figure 3.13 (**B**) was sequentially multidimensionally scaled into successively fewer dimensions. The stress at each dimension is shown. There is a gradual rise in stress through two dimensions, with a larger rise occurring between two dimensions and one dimension, suggesting that two dimensions is sufficient to portray the data.

original phylogeny (data not shown). Phylogenies are discussed in greater detail in Section 3.18.

Having considered these cautionary notes, one can generate several possible hypotheses for clustering the trypsins are suggested by the multidimensionally scaled plot in Figure 3.5. In particular, there seems to be a natural division of the trypsins into two groups, one at the top of the plot, and one at the bottom of the plot. This forms the basis of a hypothesis that the vertebrate trypsins cluster naturally into two groups. I propose that these two groups correspond to the two groups of trypsinogens defined in Section 3.2. I propose that the origin of the cluster is a schism of the ancestral vertebrate trypsinogen multigene locus into two separate multigene loci that were maintained by all descendant species of the ancestral species. I also propose that this schism occurred after the Agnathans diverged from the ancestral vertebrate lineage, so that the Agnathan trypsins belong to neither of the two groups of trypsin. This hypothesis is displayed in color in Figure 3.7, which, other than the added colors, is a reproduction of Figure 3.5. In this and in all subsequent Figures, blue

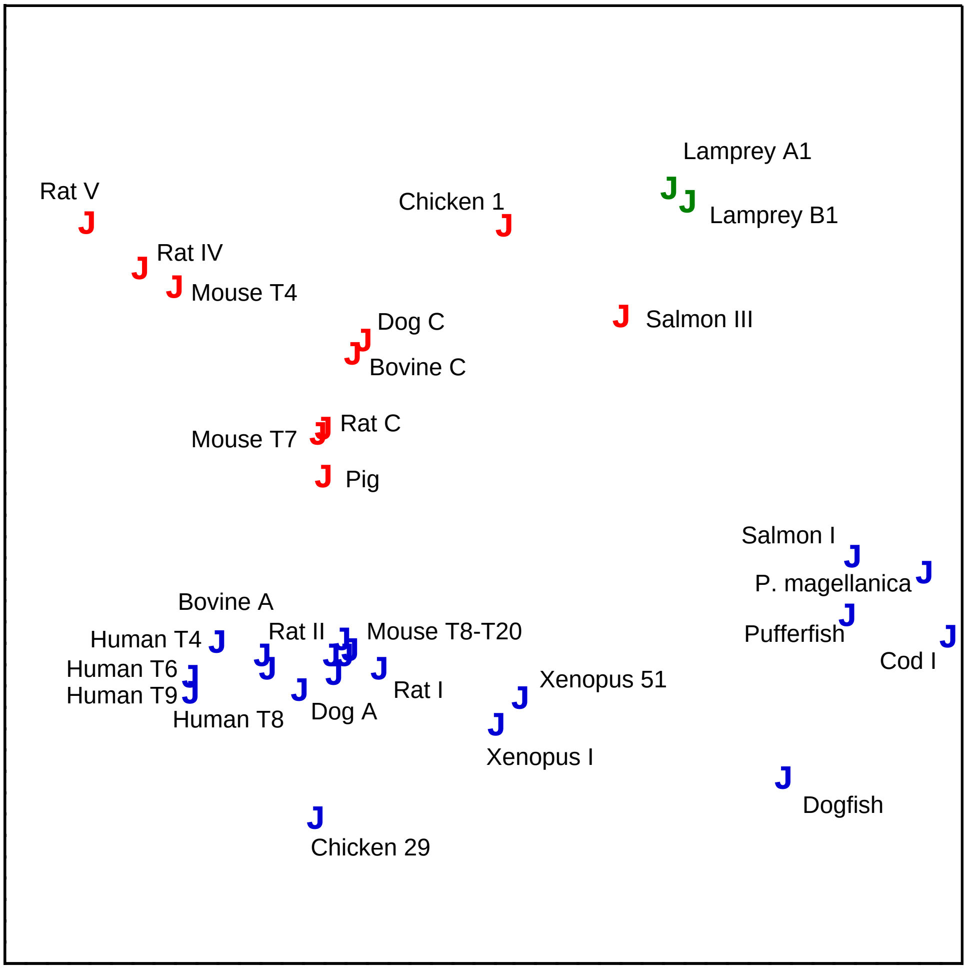


Figure 3.7. Group I vs. Group II. A multidimensionally scaled projection of the trypsin phylogenetic distances. Each point represents a trypsin sequence; The distance between two points corresponds to the calculated phylogenetic distance between the corresponding sequences. Group I trypsins are coded blue; Group II trypsins are coded red; the lamprey trypsins are green.

The two possible general mechanisms for the schism of an ancestral trypsinogen multigene family are shown in Figure 3.8. The first mechanism is the insertion of a foreign gene into the locus, dividing it. The second mechanism is the duplication of the locus to the opposite side of a foreign gene. There are a number of specific molecular mechanisms that could account for either general mechanism (see, for example, Li, 1997). For the remainder of this chapter, I will refer to both of these mechanisms as the “division” of the trypsinogen

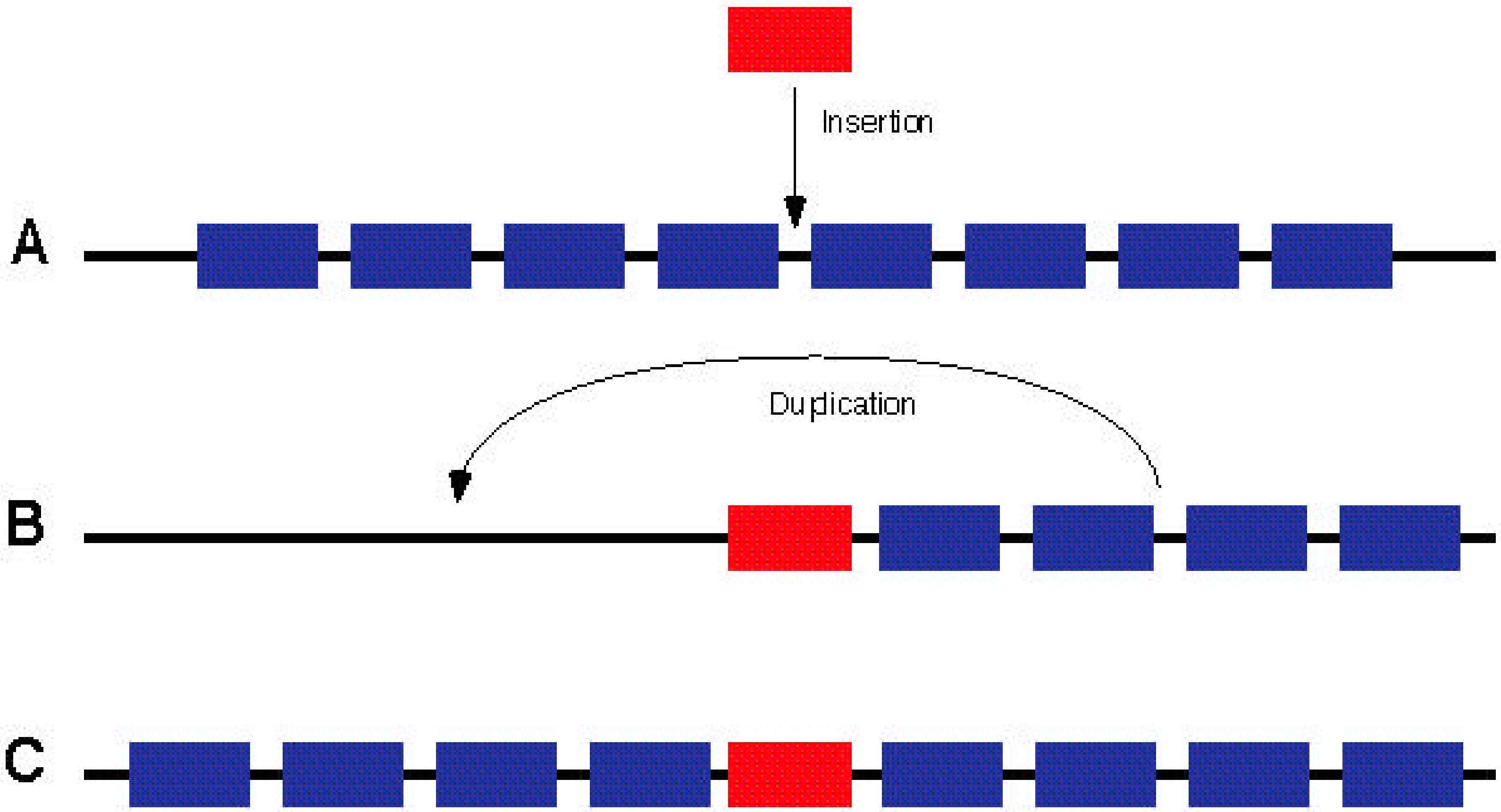


Figure 3.8. Allopatric division of the trypsinogen multigene family. Two general mechanisms can account for multigene allopatry. A, insertion; B, duplication; C, the resulting divided multigene family.

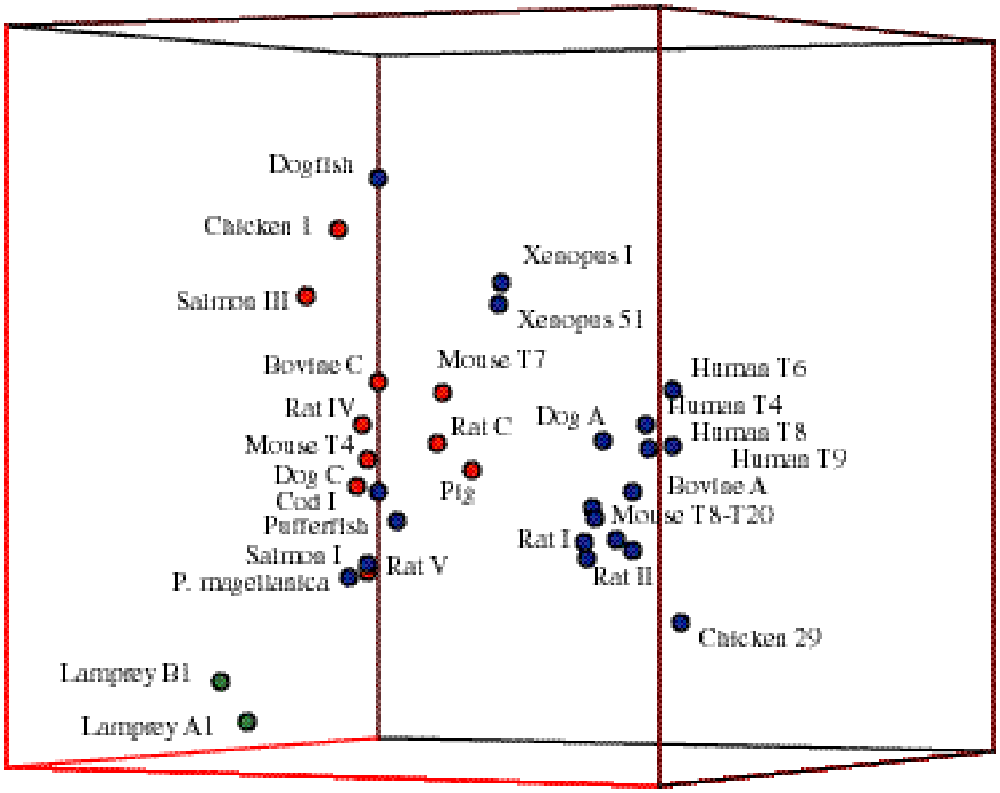


Figure 3.9. Group I vs. Group II. A shadow of a three-dimensional multidimensionally scaled projection of the trypsin phylogenetic distances. Group I trypsins are coded blue; Group II trypsins are coded red; the lamprey trypsins are green.

multigene family, incorporating the possibility of division by duplication into this meaning. This division is analogous to the “allopatric” division of a species. Allopatry is a term used in population genetics to refer to the division of a species by a geographic barrier, such as a mountain range. Such a barrier can result in speciation, which is the division of a single ancestral species into two descendant species. By analogy, I will refer to the division of the trypsinogen multigene family as an allopatric division.

Three-dimensional multidimensional scaling of the trypsin data supports the hypothesis of a single major division of the trypsinogen multigene family (Figure 3.9 and Figure 3.10). The three-dimensional views, in particular, support the hypothesis that the lamprey sequences belong to neither group I nor group II. The view in Figure 3.10 strikingly illustrates the division of the two groups along a central plane.

The syntenic positions of several trypsinogen genes with respect to the TCR β locus

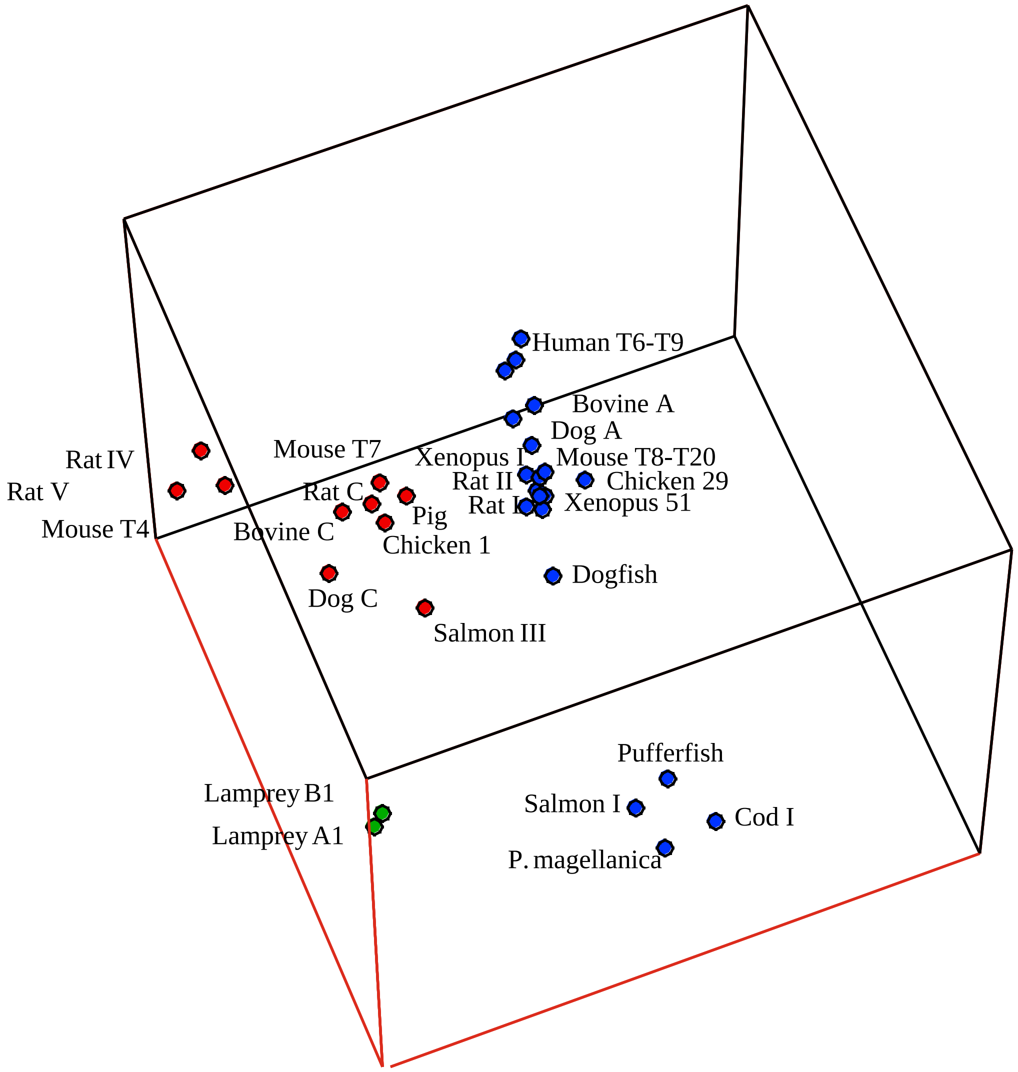


Figure 3.10. Group I vs. Group II. An alternative shadow of a three-dimensional multidimensionally scaled projection of the trypsin phylogenetic distances. Group I trypsins are coded blue; Group II trypsins are coded red; the lamprey trypsins are green.

are known. This allows these sequences to be assigned to either group I or group II with certainty. Additionally, each of the rat sequences can be assigned to a group with near certainty, due to the extreme similarity they share with orthologous mouse genes.[17] The

[17] It is difficult to prove orthology. Additionally, the term is somewhat meaningless for members of a multigene family. My use of the word "orthology" in this instance is meant to imply recent divergence from an ancestral gene with the same syntenic relationship to the TCR β locus. The orthology of the rodent trypsinogens is discussed further in Section 3.18.

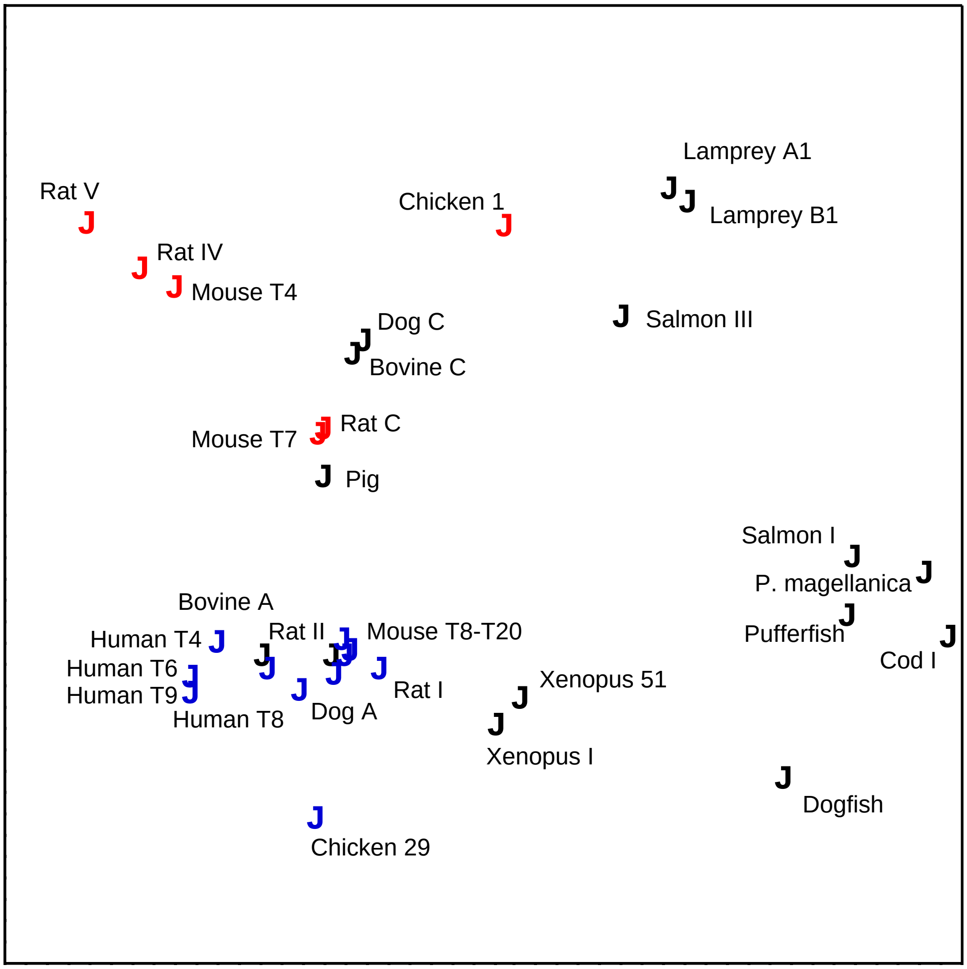


Figure 3.11. 5' vs. 3'. A multidimensionally scaled projection of the trypsin phylogenetic distances. Each point represents a trypsin sequence; The distance between two points corresponds to the calculated phylogenetic distance between the corresponding sequences. 3' trypsins are coded blue; 5' trypsins are coded red; sequences with unknown syntenic relationships are black.

trypsinogens with known synteny are colored in the plot in Figure 3.11. This data is absolutely consistent with the hypothesis suggested in the preceding paragraphs, but leaves a few sequences unassigned, other than based on the clustering suggested by multidimensional scaling (Figure 3.7).

The grouping of the remaining sequences is supported by a consideration of their isoelectric points. The isoelectric points of the trypsins are indicated on the plot in Figure 3.12; the actual values for the isoelectric points are tabulated in Table 3.6. It is immediately apparent that the group I trypsins are mostly anionic, while the group II trypsins are mostly

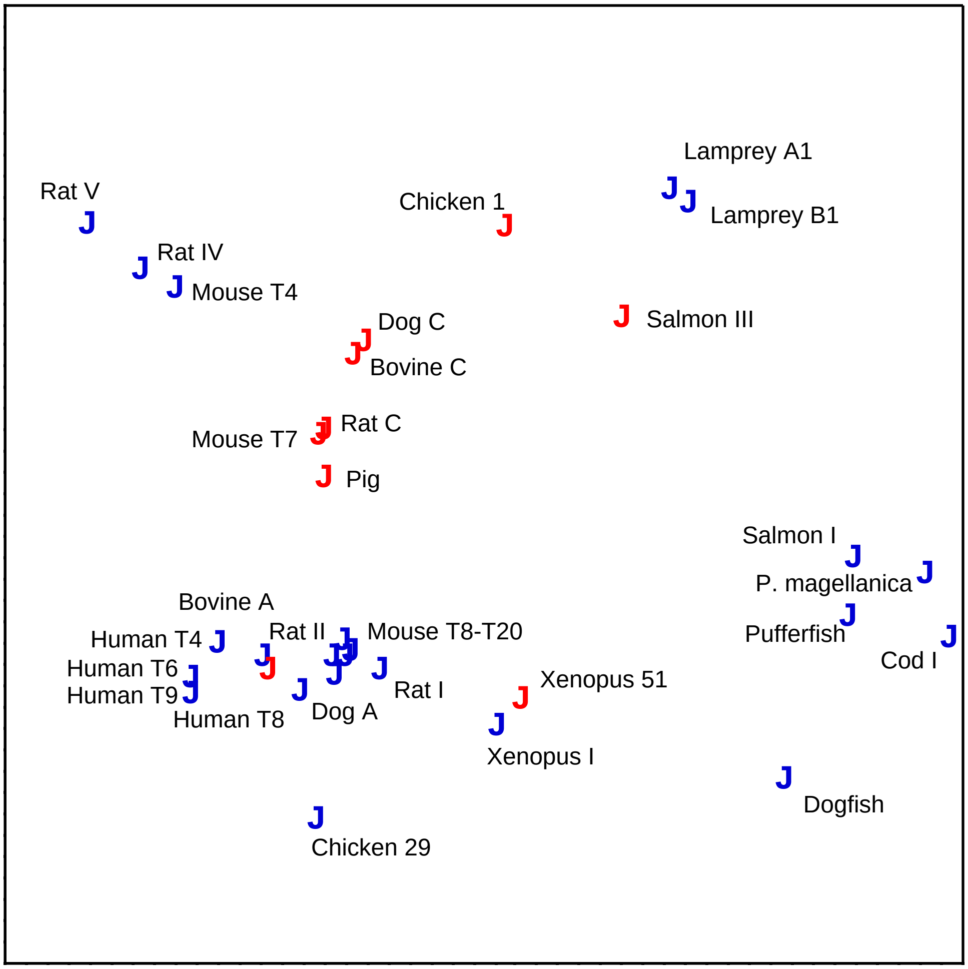


Figure 3.12. Anionic vs. Cationic. A multidimensionally scaled projection of the trypsin phylogenetic distances. Each point represents a trypsin sequence; The distance between two points corresponds to the calculated phylogenetic distance between the corresponding sequences. Anionic trypsins are coded blue; cationic trypsins are coded red.

cationic. Thus, there is a concordance of three different data types with independent components that supports the hypothesis of exactly two natural groupings of the vertebrate trypsins, with the lamprey sequences not belonging to either. These are: one, syntenic evidence; two, multidimensionally scaled distance data; and three, isoelectric point evidence.

Consideration of isoelectric points is particularly valuable for grouping the dogfish and Osteicthyes trypsins, for which no syntenic data is available. Without this data, an alternative grouping of the trypsinogens might gain consideration. From the multidimensional scaling plots, it is possible visualize the trypsins grouped into about four clusters (Figure 3.5). In addition to division of the plotted points by a horizontal line, one can also imagine division by a vertical line separating the fish sequences from the remaining sequences. If one divided the trypsins in this manner, that might lend credence to an alternative hypothesis that there were two or more independent major duplications or divisions of the trypsinogen multigene family. However, the assignment of isoelectric points for the fish trypsins is absolutely consistent with their division in to exactly two groups. This supports my hypothesis of a single major division of the trypsinogen multigene family.

Significantly, the lamprey trypsins are strongly anionic. This suggests that they do not belong to group II. Additionally, as noted above, they do not cluster with either group I or Group II in three-dimensional multidimensionally scaled plots (Figure 3.9 and Figure 3.10). Together, these data support the hypothesis that the division of the trypsinogens occurred after the divergence of the Agnathans.

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

The lack of absolute correlation of the isoelectric points with phylogenetic group is not a major obstacle to the “two group” hypothesis. First of all, all three discordant rodent sequences — mouse T4, rat IV and rat V — are already known to be group II based on synteny, as discussed above. The same is true of the discordant human T8 trypsinogen. This leaves only the discordant *Xenopus* 51 trypsinogen to be explained, but its similarity with *Xenopus* I coupled with its unambiguous position in the multidimensionally scaled plots leave little doubt as to its group I assignment. It remains possible that *Xenopus* 51 is a group II trypsin that has undergone extensive coincidental evolution, but if this were the case, it would have little impact on the conclusions of the “two group” hypothesis. It would, in fact, support the contention of Section 3.19 that coincidental evolution has played a major role in vertebrate trypsinogen evolution.

There is a possible explanation for the discordance of the human T8 isoelectric point. Recall from Section 3.2 that humans possess no functional group II trypsins. However, as mentioned in Section 3.13, it may be that both anionic and cationic isoforms have a functional niche, with the jawed vertebrates requiring both. In this scenario, human trypsin I, which is biochemically cationic but phylogenetically group I, would fill a crossover role by filling a

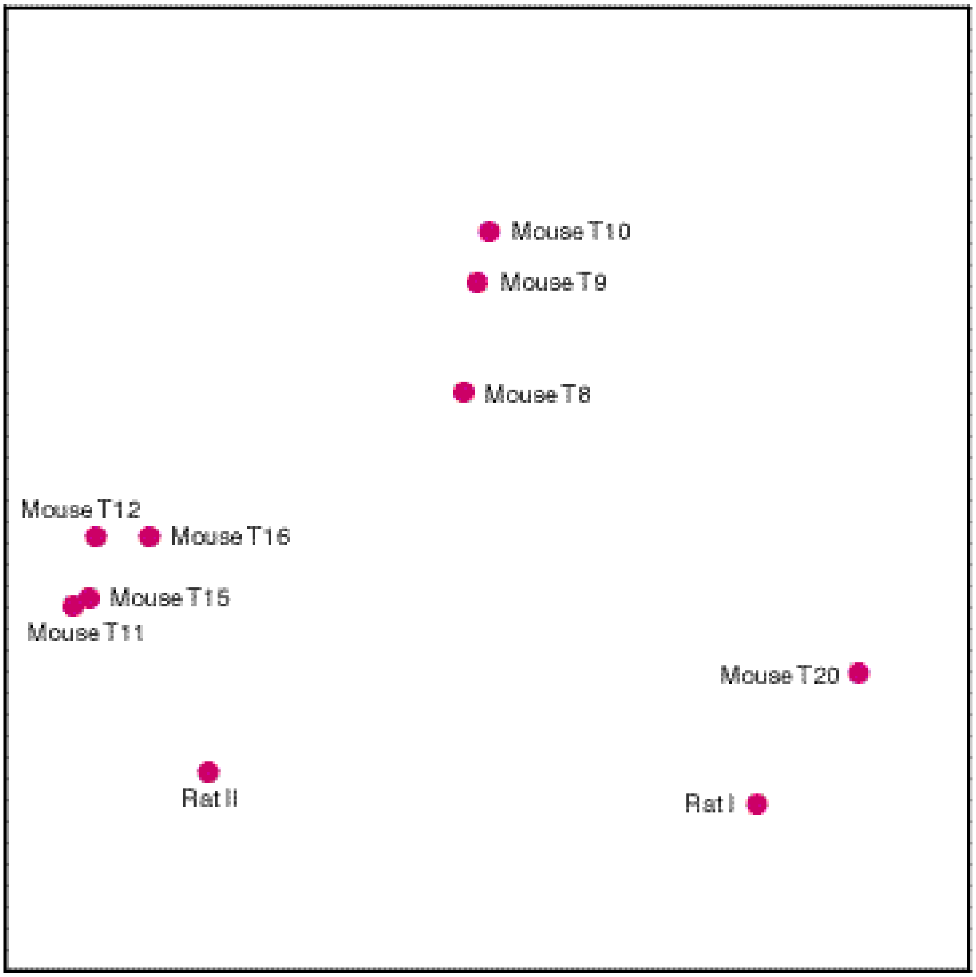


Figure 3.13. A multidimensionally scaled projection of the rodent group I trypsin phylogenetic distances. Each point represents a trypsin sequence; The distance between two points corresponds to the calculated phylogenetic distance between the corresponding sequences.

niche vacated by the missing group II trypsinogens. Also since rat C and mouse T7 are cationic, in this scenario there might be little selective pressure on mouse T4, rat IV, and rat V to remain cationic, so their charge could have “drifted.”

I did not incorporate pseudogenes into my multidimensionally scaled plots. The use of pseudogenes might skew the structure of the data. Such skewing effects can be severe for phylogenies, as shown in Section 3.18. However, skewing due to distorted distances for multidimensional scaling can be surprisingly mild (Everitt and Dunn, 1991). This may be the case for vertebrate trypsin data. Several pseudogenes are incorporated into the data for Figure 4 of Roach (1997). As expected, the group II pseudogenes cluster with the functional group II genes, and the group I pseudogenes cluster with the functional group I genes.

In an effort to further explore the utility of multidimensional scaling, I multidimensionally scaled the distances for the rodent group I trypsinogens, including several sequences that were not included in the simplified vertebrate data set (Figure 3.13). In this case, multidimensional scaling provides little insight beyond what can be obtained from a traditional phylogenetic analysis (see Section 3.18). A difference between the rodent group I data set and the vertebrate data set is that all of the group I rodent trypsins are very closely related. Multidimensional scaling may be most useful for examining distant relationships. The stress for the multidimensionally scaled plot in Figure 3.13 is graphed in Figure 3.6.

3.18 PHYLOGENIES

A consideration of the data presented in the previous section leads one to conclude that the trypsinogen multigene family underwent an allopatric division about 500 million years ago, during the Ordovician or Silurian Periods. One thus predicts the existence of two groups of trypsinogen in all of the jawed vertebrates. All of the members of each group should be more related to each other than any are to members of the other group, as each group shared a more recent ancestor. This assumes that no coincidental evolution has occurred (see Section 3.1). With these assumptions, a hypothetical phylogeny can be sketched (Figure 3.14). This Figure also assumes a roughly constant rate of evolution. This Figure can be compared to the computed phylogenies discussed below. Such comparisons highlight the impact of coincidental evolution on the vertebrate trypsinogens.

A phylogeny of the vertebrate trypsinogens can be computed from a pairwise distance matrix, calculated as discussed in Section 3.16. Such a phylogeny, for forty-two sequences, is shown in Figure 3.15. This phylogeny can be recalculated with fewer sequences. This can be done with little loss of information, as there are several sequences that are nearly identical to each other: mouse T11, T2, T15, and T16; mouse T8 and T9; mouse T4 and T5; chicken 1 and 38; salmon I and II; cod I and X; lamprey B1 and B2; lamprey A1 and A2. Not only is the resulting phylogeny less cluttered, but it also demands fewer computational resources to calculate.

A thirty-two-sequence phylogeny is shown in Figure 3.16. The branches of the phylogeny are colored to correspond with the group to which the sequences at the terminus of the branch belong.

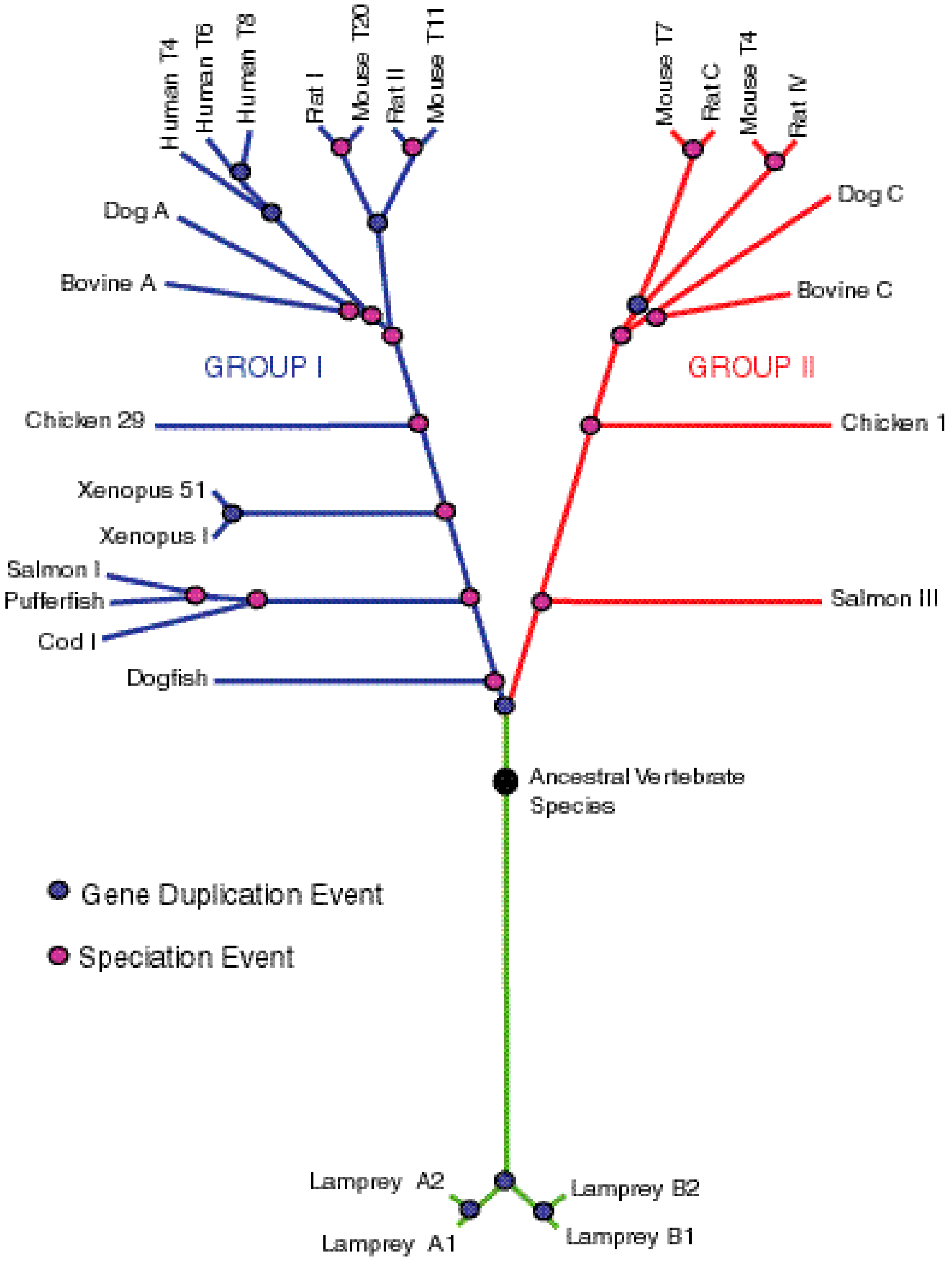


Figure 3.14. A hypothetical phylogeny of the vertebrate trypsinogens. Group I branches are blue; group II branches are red. Branches in green diverged before group I and group II split.

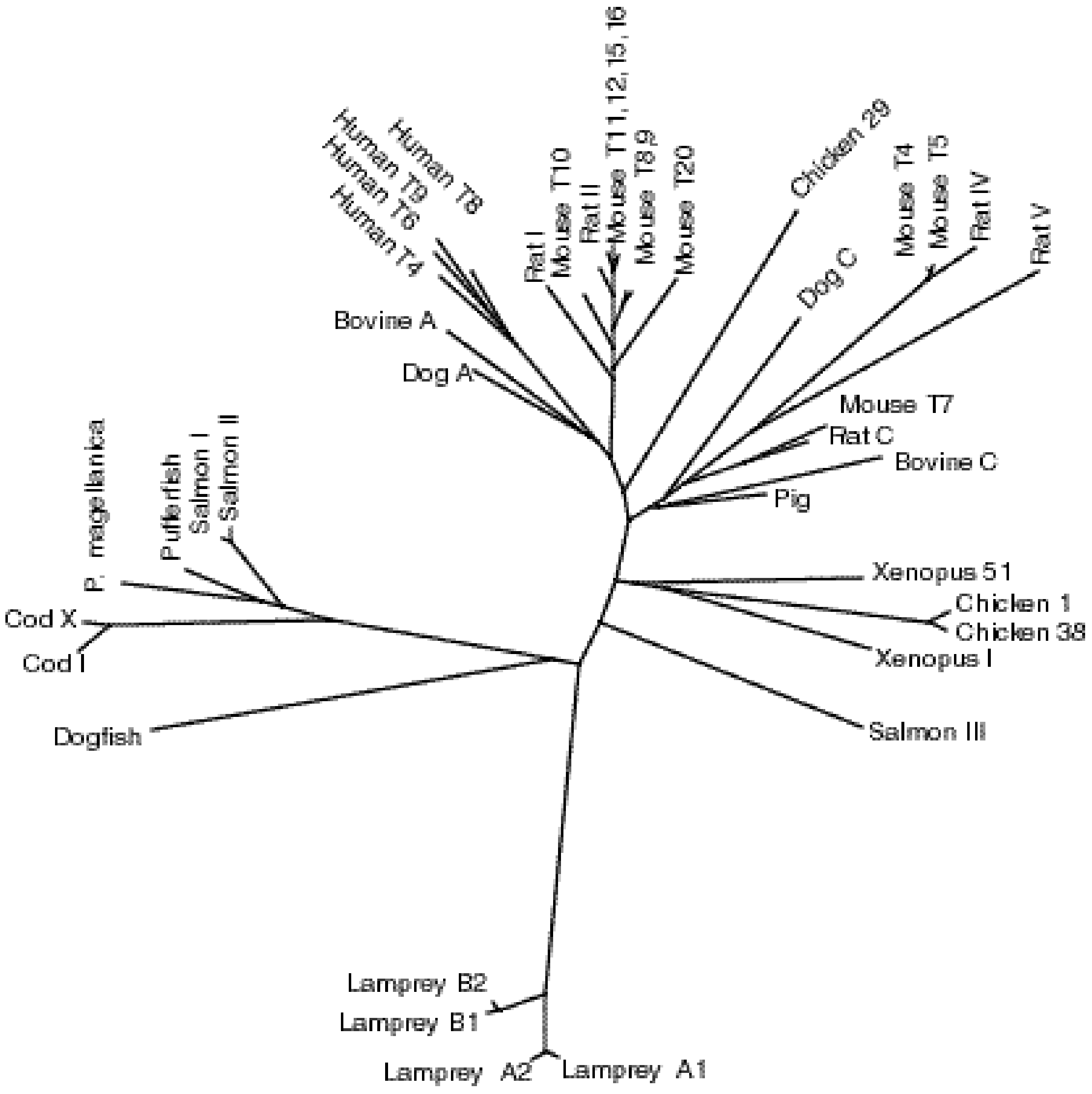


Figure 3.15. A Fitch-Margoliash phylogeny of forty-two vertebrate trypsinogens. Distances from the program *protdist*, with the Dayhoff matrix, were fed to the program *fitch*, with global rearrangements and 20 random "jumbles" (Felsenstein, 1993).

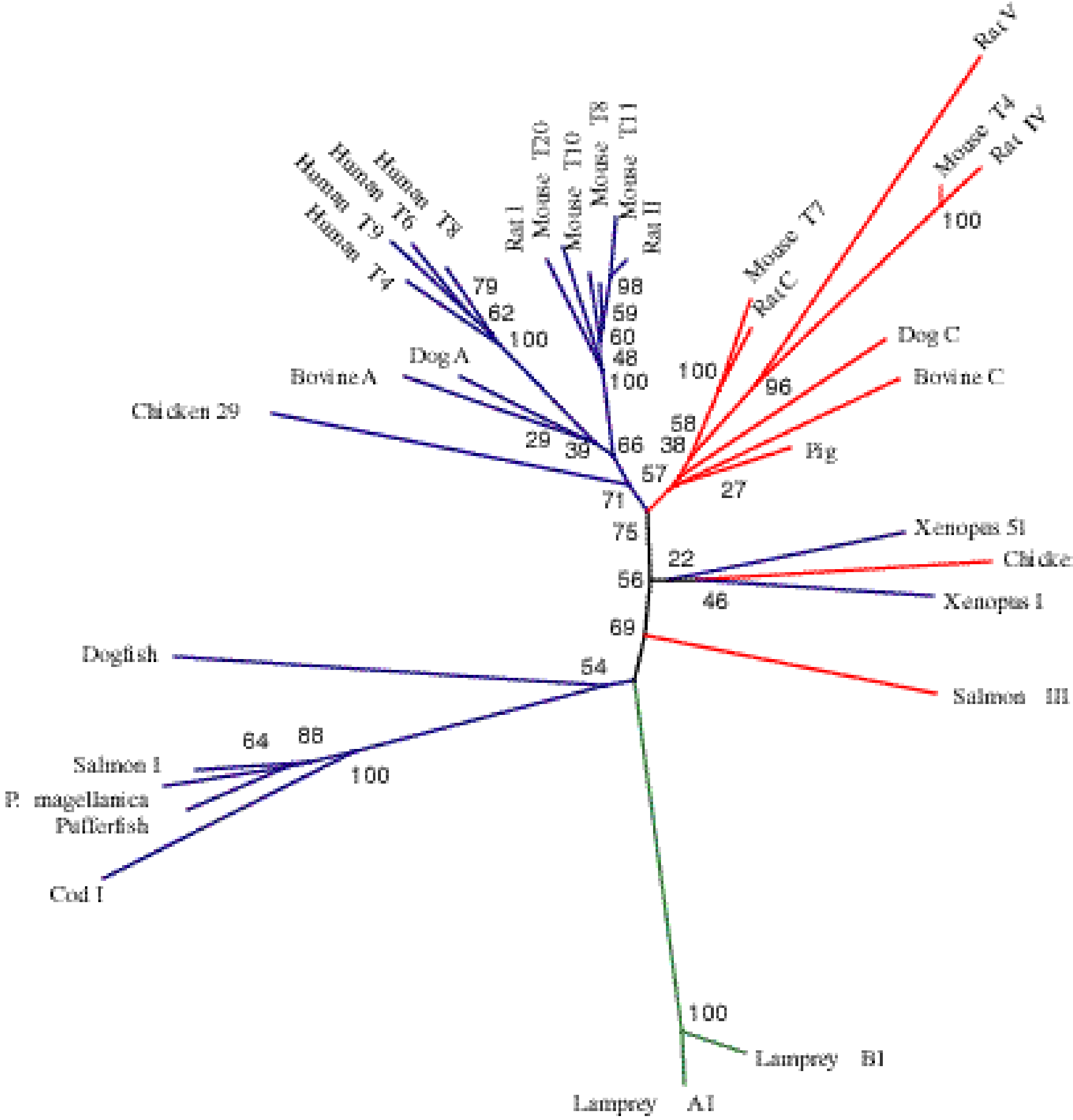


Figure 3.16. A Fitch-Margoliash phylogeny of thirty-two vertebrate trypsinogens. The graph is calculated as described for Figure 3.15. The numbers adjacent to nodes represent the number of times the clade distal to the unlabeled node was recovered during 100 delete-half-jackknifes of the original data (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.