CABIOS INVITED REVIEW

Vol. 13 no. 4 1997
Pages 325–332

Identifying distantly related protein sequences

William R.Pearson

Introduction

The most powerful method available today for inferring the biological function of a gene (or the protein that it encodes) from its sequence is similarity searching on protein and DNA sequence databases. With the development of rapid methods for sequence comparison, both with heuristic algorithms and powerful parallel computers, discoveries based solely on sequence homology have become routine. Indeed, the vast majority of the gene identifications in the recent descriptions of the *Haemophilus influenzae* (Fleischmann *et al.*, 1995), *Mycoplasma genitalium* (Fraser *et al.*, 1995), yeast (Dujon, 1996) and *Methanococcus janesscii* (Bult *et al.*, 1996) genomes are based only on protein sequence similarity. As more complete genomes become available, protein sequence comparison will become an even more powerful tool for understanding biological function.

Protein sequence comparison is a powerful tool because of the enormous amount of information that is preserved throughout the evolutionary process. For many protein sequences, an evolutionary history can be traced back 1–2.5 billion years. Proteins that share a common ancestor are called homologous. Sequence comparison is most informative when it detects homologous proteins. Homologous proteins always share a common three-dimensional folding structure and they often share common active sites or binding domains. Frequently, homologous proteins share common functions, but sometimes they do not. Our ability to characterize the biological properties of a protein based on sequence data alone stems almost exclusively from properties conserved through evolutionary time. Predictions of common properties for non-homologous proteins—similarities that have arisen by convergence—are much less reliable.

While sequence similarity searching is a routine method for characterizing newly determined DNA and protein sequences, researchers sometimes fail to exploit fully the information that is available from similarity searches of protein sequence databases. This review examines two strategies for using similarity search information more effectively: (i) looking for alignments that span an entire folding domain, rather than a short sequence motif, and (ii) re-examining sequences with high, but not statistically significant, similarity scores. For a broader perspective on sequence comparison and identification of homologous proteins, see Altschul *et al.* (1994) and Pearson (1996).

Members of the trypsin-like serine protease superfamily ('trypsin-like' distinguishes these serine proteases from other serine protease families—notably the subtilisins—that use serine in the active site but have very different structures and thus are not homologous) provide a classic example of a family of proteins with a highly conserved active site. While highly conserved motifs from this site are informative, serine proteases share similarity throughout the length of the protease domain, not just around the active site residues.

The trypsin-like serine protease family is quite diverse, with a number of very distantly related homologues. Thus, it can be difficult to demonstrate that *Streptomyces griseus* protease A and protease B are homologous based on sequence similarity alone. The second part of this review shows that by carefully re-examining sequences with high-scoring, but not statistically significant, similarity scores, it is possible to identify several proteins that share significant similarity with both the mammalian trypsin-like serine proteases and their distant prokaryotic homologues.

Motifs, homology, and the serine proteases

A common misconception in protein sequence comparison is that homologous proteins share sequence similarity mostly (or only) near the active site regions or other functional domains in a protein. This partly accounts for the popularity of databases of sequence motifs, such as PROSITE (Bairoch, 1991), which tabulate amino acid patterns that can be used to identify most of the members of a protein family. For features that result from convergence to a common property, such as glycosylation and phosphorylation sites, sequence motifs are uniquely informative. However, for features that result from divergence from a common ancestor, such as the serine protease active site residues, sequence motifs provide only a highly abstracted summary of the sequence conservation in a family. Because they share a common three-dimensional structure, homologous proteins share sequence similarity over large regions–typically the entire protein fold.

The trypsin-like serine protease superfamily is a classic example of a protein family whose members share several simple motifs that are diagnostic for the family (Figure 1).

Department of Biochemistry, Jordan Hall #440, University of Virginia, Charlottesville, VA 22908, USA

E-mail: wrp@virginia.EDU

```
ID   TRYPSIN_HIS; PATTERN.
AC   PS00134;
DE   Serine proteases, trypsin family, histidine active site.
PA   [LIVM]-[ST]-A-[STAG]-H-C.
NR   /TOTAL=158(158); /POSITIVE=154(154); /UNKNOWN=2(2); /FALSE_POS=2(2);
NR   /FALSE_NEG=11(11);
CC   /TAXO-RANGE=??EP?; /MAX-REPEAT=1;
CC   /SITE=5,active_site;

ID   TRYPSIN_SER; PATTERN.
AC   PS00135;
DE   Serine proteases, trypsin family, serine active site.
PA   G-D-S-G-G.
NR   /TOTAL=160(160); /POSITIVE=151(151); /UNKNOWN=1(1); /FALSE_POS=8(8);
NR   /FALSE_NEG=16(16);
CC   /TAXO-RANGE=??EP?; /MAX-REPEAT=1;
CC   /SITE=3,active_site;
```

Fig. 1. Patterns for serine proteases. Patterns from PROSITE that identify 152/163 TRYPSIN_HIS or 143/159 TRYPSIN_SER members of the trypsin-like serine protease protein family.

Serine proteases cleave peptide bonds using a 'catalytic triad' of histidine, serine and aspartic acid that are required for the protease function. Because these residues are so highly conserved, patterns that focus on two of the regions (Figure 1) can be used to identify every member of the serine protease family. (The subtilisin-like serine proteases use exactly the same catalytic triad, but the families are non-homologous with very different three-dimensional structures.)

Most members of the trypsin-like serine protease superfamily are readily identified by sequence similarity searching. The results from a typical protein database search using the Smith–Waterman algorithm (Smith and Waterman, 1981) are shown in Figure 2. All of the eukaryotic trypsin-like serine proteases share statistically significant similarity with the bovine trypsin query sequence. However, as is often the case for divergent protein families, some prokaryotic members of the family do not share statistically significant similarity with bovine trypsin. These sequences are italicized in Figure 2; their membership in the serine protease family is usually inferred from their common three-dimensional structures (Figure 5).

The absolute conservation of residues in the 'catalytic triad' might suggest that sequence similarities shared by members of this family are limited to those regions. Indeed, two of the four 'High-Scoring segment Pairs' (Altschul *et al.*, 1994) reported by BLASTP correspond to TRYP_HIS and TRYP_SER regions (Figure 3). However, similarity in the serine proteases extends from one end of the protein to the other, with conservation throughout the sequence. Indeed, many parts of protein are conserved more strongly than the region around the aspartic acid in the catalytic triad (Figure 3). Thus, while the residues in the catalytic triad are an essential feature for a functional serine protease, it is the serine protease fold (two domains containing anti-parallel β barrels; Figure 5) that is required to bring these residues together. The evolutionary pressure to conserve the trypsin-like serine protease fold ensures that the folding domains share similar amino acids.

The requirement for a common folded structure in homologous proteins usually causes similarities to extend from one end of the protein to the other. With the exception of mosaic proteins that are the result of recent exon shuffling (Doolittle, 1995), optimal local sequence similarity is rarely confined only to a portion of two homologous sequences. (In mosaic proteins, the similarity extends throughout the exon-shuffled domain.) In general, it is incorrect to speak of homology at the N terminus or C terminus, even though only a portion of the protein may be aligned in 'High Scoring segment Pairs' by BLASTP. Indeed, the length of the locally similar region can sometimes be used to distinguish low-scoring related sequences from high-scoring unrelated sequences. Thus, all but two of the library sequences (including four with expectation values >0.02) that align over >80% of the length of the TRYP_BOVIN query sequence are members of the trypsin-like serine protease family. Figure 4 displays the locally similar regions for the related and unrelated sequences in Figure 2; the highest scoring unrelated sequences tend to have relatively short (<100 residue) regions of higher similarity (~30% identical), while related sequences have longer (140–300 residue) aligned regions, sometimes with lower (25%) sequence identity. In general, alignments with longer, lower identity are more significant than those with shorter, higher identity.

The requirement for similarity over a large region is more evident when three-dimensional structures are examined. TRYP_BOVIN (structure not shown), TRYP_STRGR

LOCUS	Description	len	score	E(51,780)
TRYP_BOVIN	trypsinogen (EC 3.4.21.4).	229	1559	0
TRY2_HUMAN	trypsinogen II	247	1201	0
TRYP_PLEPL	trypsin	250	788	0
KLK2_HUMAN	glandular kallikrein 2	261	665	0
RVVA_VIPRU	vipera russelli proteinase	236	637	0
TRY1_ANOGA	trypsin 1	274	600	10^{-32}
TRYA_DROME	trypsin alpha	256	579	10^{-31}
FA9_RAT	coagulation factor IX	282	573	10^{-30}
PLMN_PIG	plasminogen	790	569	10^{-30}
TRY5_ANOGA	trypsin 5	274	550	10^{-29}
TRYP_FUSOX	trypsin	248	541	10^{-28}
FA7_RABIT	coagulation factor VII	443	519	10^{-27}
URTB_DESRO	salivary plasminogen activator β	431	508	10^{-26}
ACRO_PIG	acrosin	415	501	10^{-26}
PRTC_HUMAN	protein C	461	494	10^{-25}
TRYM_CANFA	mastocytoma protease	269	484	10^{-25}
TRYP_STRGR	trypsin	259	410	10^{-20}
HGF_HUMAN	hepatocyte growth factor prec.	728	397	10^{-18}
ACH1_LONAC	achelase I protease	213	352	10^{-16}
CERC_SCHMA	cercarial protease	264	203	10^{-6}
CO2_HUMAN	complement C2	752	198	10^{-5}
CFAB_MOUSE	complement factor B	761	170	0.00041
PRTZ_BOVIN	vitamin K-dependent protein Z	396	142	0.015
LORI_MOUSE	loricrin.	481	125	0.24
GSEP_BACLI	glutamyl endopeptidase	316	118	0.45
KRUC_SHEEP	keratin, ultra high-sulfur matrix	182	107	1.3
PRLA_LYSEN	alpha-lytic protease	397	107	3.1
AGI_URTDI	lectin/endochitinase precursor	372	105	3.9
KCR8_YEAST	prob. serine/threonine-protein kin.	603	107	4.7
G156_PARPR	156g surface protein precursor	715	117	5.0
YLK3_CAEEL	putative ser./thr.-protein kinase	895	114	5.4
AMY_CLOAB	putative alpha-amylase	469	104	5.7
AGI_HORVU	root-specific lectin precursor	212	98	6.2
YB9X_YEAST	hypothetical trp-asp repeats	878	105	9.5
PRTS_MOUSE	vitamin k-dependent protein S	675	103	9.8
DLK_HUMAN	delta-like protein	383	99	9.9
PRTB_STRGR	streptogrisin B (S. gris. prot. A)	299	94	16.
PRTA_STRGR	streptogrisin A (S. gris. prot. A)	297	85	64.

Fig. 2. Serine protease search—high-scoring sequences. High-scoring sequences from a search of SwissProt (Bairoch and Boechmann 1991; release 33, April 1996) with TRYP_BOVIN. Only 10% of the database sequences with E() < 10^{-6} are shown. Trypsin-like serine proteases with E() > 0.02 are in italics.

(Figure 5, 1sbt) and PRTA_STRGR (1sgc) share a very similar all-β fold with symmetrical β barrel structures and two short α helices. Very little of this structure is directly involved in forming the catalytic triad in the active site; yet the entire fold is conserved, thus requiring conservation of an amino acid sequence that adopts this fold.

Although almost all vertebrate trypsin-like serine proteases share significant sequence similarity with bovine trypsin, most bacterial serine proteases do not. For example, the similarity score for alignment of bovine trypsin with *S.griseus* protease A is not statistically significant (E() < 64), even though the structures of the two enzymes are very similar (Figure 5). Thus, while statistically significant similarity generally implies common ancestry, and thus common three-dimensional structure [the most common exceptions to this rule are regions with very low amino acid complexity, e.g.YSGGGGSSCGGGYSGGGGSSCGGGSSGGG from LORI_MOUSE (Altschul *et al.*, 1994)], lack of statistically significant similarity does not imply non-homology.

Figure 5 also shows the structures of two non-homologous proteins. Subtilisin (1sbt) is included because it is an example of 'convergent' evolution (Doolittle, 1994); subtilisin uses the same triad of catalytic residues (Asp, His and Ser) to cleave peptide bonds, but shares no structural similarity beyond the geometry of the active site of the enzyme. Subtilisin and subtilisin-like serine proteases are not homologous to the trypsin-like serine proteases. As expected, the different structures share no statistically significant sequence similarity (1500 random sequences from SwissProt would be expected to have a better similarity score than that obtained in the trypsin/subtilisin comparison).

Likewise, high-scoring sequences that are not homologous to trypsin-like serine proteases rarely share structural similarity to the family, despite their 'strong' similarity. Wheat germ agglutinin (7wga) is the most similar non-serine protease sequence in the NRL_3D database of sequences whose structures are known, yet it does not contain a single β sheet. With the exception of membrane-spanning proteins, which frequently share hydrophobic regions with other unrelated membrane proteins, high sequence similarity—in

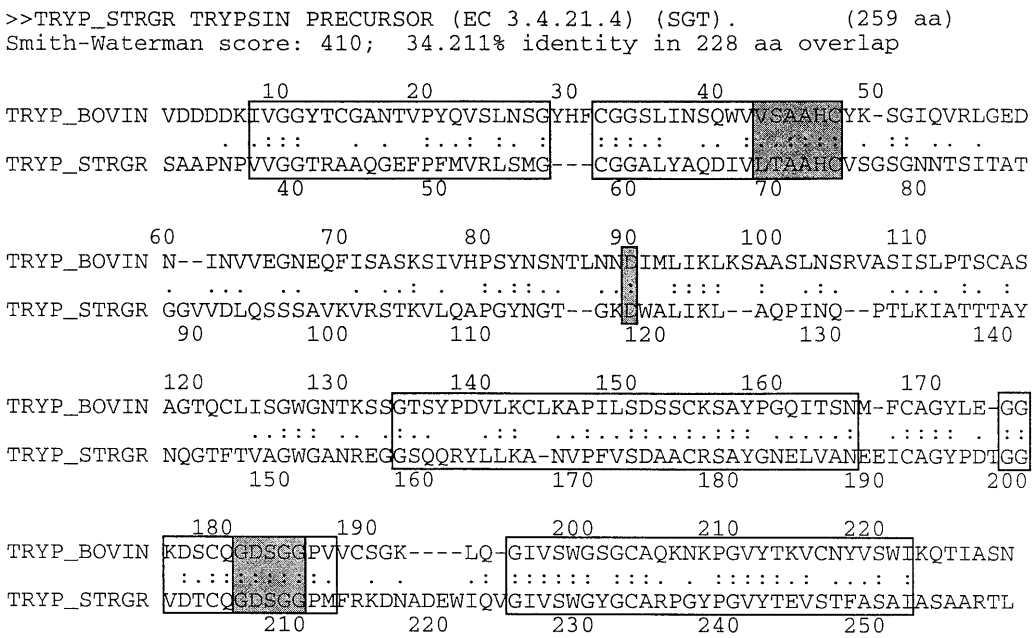

Fig. 3. Alignment of serine proteases. Alignment of bovine trypsinogen (TRYP_BOVIN) and *S.griseus* trypsin (TRYP_STRGR). Shaded boxes indicate the TRYP_HIS and TRYP_SER patterns shown in Figure 1 and the conserved 'D' that is the third component of the catalytic triad. Unshaded boxes indicate the consistent 'High Scoring segment Pairs' reported by BLASTP.

the absence of homology—provides no information about structural similarity.

Using statistical significance to explore distant relationships

A major advance in sequence identification by similarity searching has been the development of accurate statistical estimates for similarity scores (Altschul *et al.*, 1994). Since the similarity score from comparison of TRYP_BOVIN and TRYP_STRGR has an expectation value of E() < 10^{-20}, we conclude that these two sequences share similarity that would never be obtained by chance (or obtained once in 10^{20} searches of a database the size of SwissProt), and thus their similarity reflects a common ancestry for the two sequences. Current versions of the FASTA package of sequence comparison programs (version 2 and 3) include accurate statistical estimates for both FASTA and SSEARCH (Smith–Waterman) similarity scores (Pearson, 1996). Careful analysis of the high-scoring non-homologous sequences can be used both to confirm that the statistical estimates are reliable and to explore distantly related members of a protein family.

Identifying the highest-scoring non-homologous sequences in a database search may seem difficult if the protein family is very diverse. However, additional searches with high-scoring, but possibly unrelated sequences can be used to separate high-scoring unrelated sequences from distantly related sequences. Additional searches with high-scoring unrelated sequences will typically produce 'matches' with unrelated sequences, while additional searches with distantly-related sequences will produce 'matches' to protein family members. If the statistical estimates are accurate, high-scoring unrelated sequences will have E() values of ~1.0 , since one highest scoring sequence is expected in every search. If the E() value for the highest scoring unrelated sequences are unexpectedly low and the sequences do not contain low-complexity simple sequence repeats, additional searches can be carried out with higher gap penalties.

Bovine trypsin (TRYP_BOVIN) shares statistically significant similarity with every full-length mammalian serine protease, but the bacterial alpha-lytic protease (PSLA_LYSEN) or *S.griseus* protease A or protease B do not share significant similarity with bovine trypsin. There is no question that these proteins are homologous to the mammalian trypsin-like enzymes because of their strong structural similarity (Figure 5). However, in the absence of high-resolution structural data, how can one decide whether a high-scoring, but not significantly similar, sequence is homologous?

Additional searches with the highest scoring, non-significant matches allow us to identify additional members of the family. A search with PRTZ_BOVIN, which has a marginally significant score, shows strong similarity (E() values < 10^{-10}) with a variety of other members of the family, thus confirming its homology. LORI_MOUSE gives a different result; while many serine proteases are highly ranked with

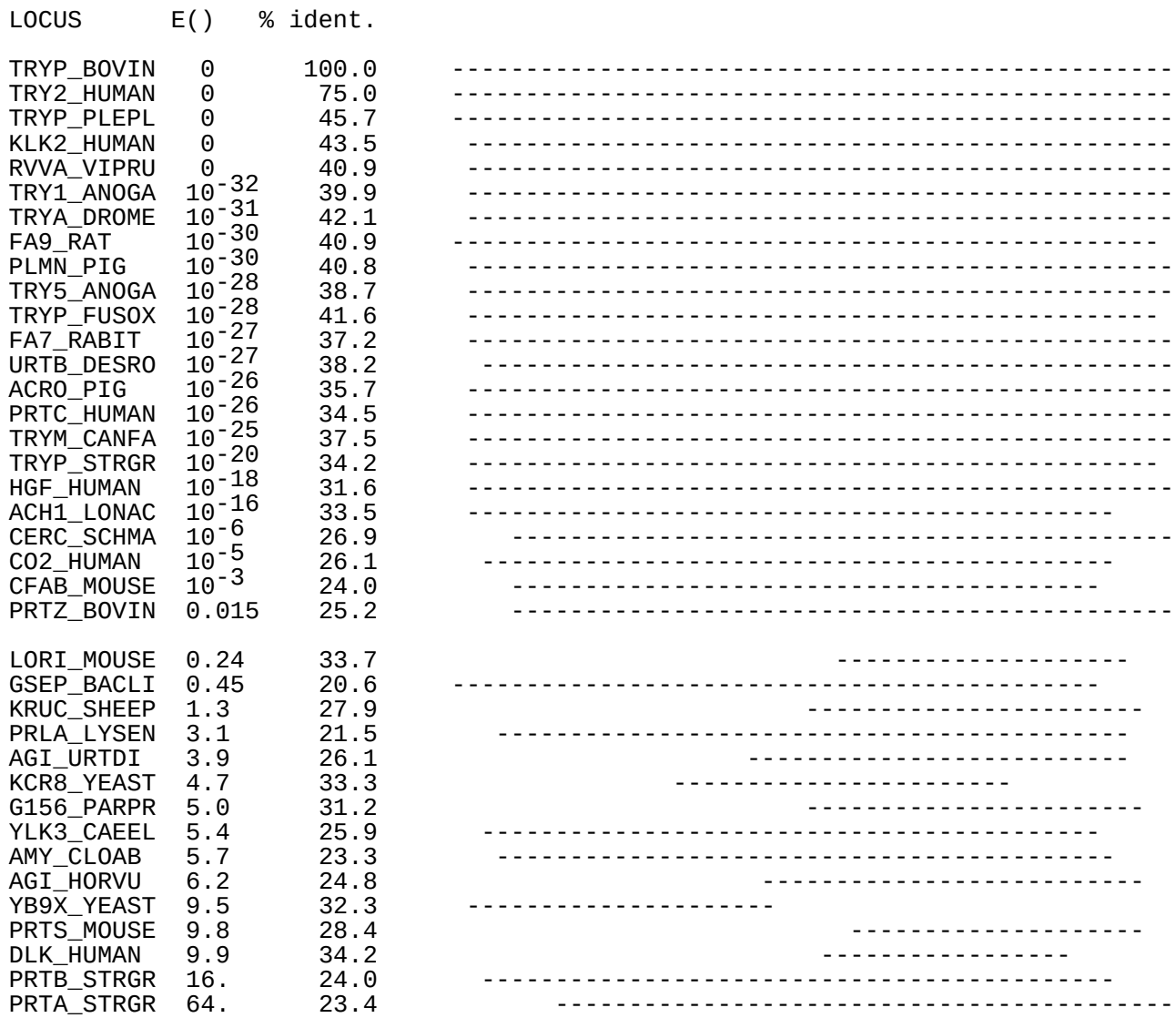

Fig. 4. Serine protease alignments The alignments of each of the high-scoring sequences reported in Figure 2 are indicated by mapping back to the TRYP_BOVIN query sequence. Thus, alignment of TRYP_BOVIN with itself extends from the beginning to the end of the query sequence; alignment of TRYP_BOVIN and TRYA_DROME extends over 85% of the TRYP_BOVIN query sequence. Members of the family with E() > 0.02 are italicized. The E() value and percent identity are also shown. The ssearch -m 4 option was used to produce this figure.

significant similarity, the sequence alignments contain a repeated glycine and serine motif. Thus, LORI_MOUSE is not homologous; it contains an unusual simple amino acid repeat sequence. On the other hand, GSEP_BACLI shares strong similarity with several bacterial serine proteases (E() < 10^{-9}) and weaker, but significant similarity with TRYP_SACER and TRYP_FUSOX, *Streptomyces* and yeast trypsins with very strong similarity to bovine trypsin. GSEP_BACLI is, therefore, a member of the trypsin-like serine protease family.

A search with alpha-lytic protease reveals a second group of closely related serine proteases, which includes *S.griseus* protease A and protease B. While none of the sequences in Figure 2 have significant similarity with PRLA_LYSEN, GLUP_STRGR, an *S.griseus* glutamyl endopeptidase, shares strong similarity with the *S.griseus* protease A and B, alpha-lytic protease, and weaker, but significant similarity with TRYA_DROME and several other *Drosophila* serine proteases (Figure 6). The insect sequences share strong similarity to mammalian trypsin-like serine proteases (Figure 2). Thus, by carefully exploring sequences with high, but not statistically significant, similarity scores, it is possible to construct statistically significant links between very distantly related serine proteases.

Distant sequence relationships can thus be established by moving from sequence **A** to significantly similar sequence **B**,

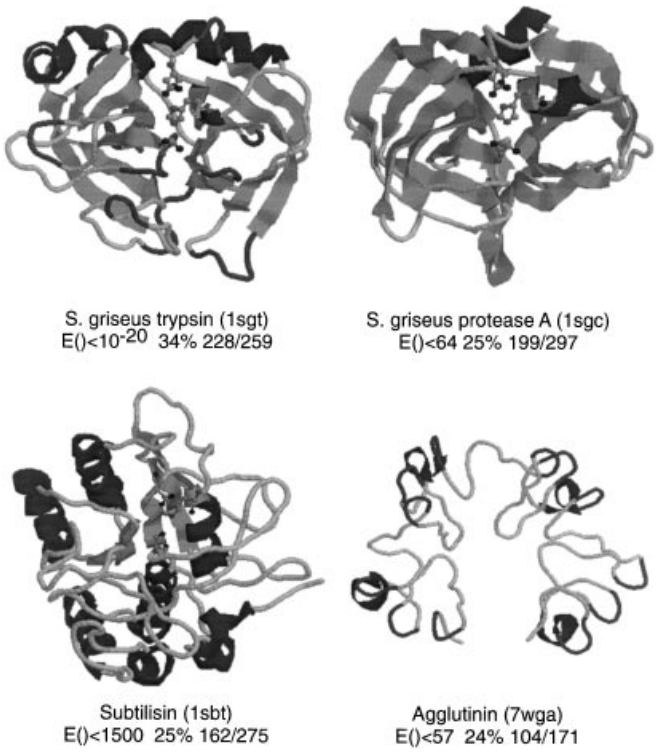

Fig. 5. Structures—homologous, convergent and unrelated. The structures of two members (1sgt, 1sgc) of the trypsin-like serine protease family are shown, along with subtilisin (1sbt)—a non-trypsin-like serine protease—and wheat germ agglutinin (7wga), one of the highest scoring non-serine proteases in the NRL_3D database (release 20) of sequences whose structures are known. Serine protease structures are aligned to present a similar view of the catalytic site. The expectation values shown are based on a comparison of bovine trypsin (TRYP_BOVIN) to the SwissProt (release 33) protein sequence database. Also shown are the percent identity and the length of the similar region with respect to the length of the sequence of the structure shown.

and then from **B** to **C**, even though **A** does not share significant similarity with **C**. The strategy is effective because of the implicit evolutionary tree that connects all the members of a protein family. Thus, in Figure 7, a sequence on a relatively short branch, TRYA_DROME, can be used to establish significant relationships with very diverse members of the family. For large and diverse protein families, it is usually easy to identify a number of 'less-divergent' family members that can be used to link distant branches of the tree. Naturally, such inferences are more reliable if statistically significant similarity scores are produced with different sets of scoring matrices and gap penalties, and if they are established with several different linking sequences.

A phylogenetic tree was produced from selected vertebrate, invertebrate and prokaryotic trypsin-like serine proteases. Sequences were aligned using ClustalW (Thompson *et al.*, 1994) and protein distances estimated and distance trees built using the PHYLIP package (Felsenstein, 1989). The three numbers to the right of the sequence names report the statistical significance of the alignment score between the sequence and bovine trypsin (TRYP_BOVIN), *Drosophila* trypsin A (TRYA_DROME) and *S.griseus* glutamyl endopeptidase (GLUP_STRGR), respectively. MPR_BACSU is an example of another sequence that links eukaryotic and prokaryotic serine proteases, although it does not share statistically significant similarity with the three query sequences used for expectation values here.

Summary

Protein sequence comparison is the most powerful tool available today for inferring structure and function from sequences because of the constraints of protein evolution—a

LOCUS	Description	len	score	E(51,934)
GLUP_STRGR	glutamyl endopeptidase II	188	1223	0
SFA1_STRFR	serine protease 1	357	1019	0
PRTA_STRGR	streptogrisin A	297	681	0
PRTB_STRGR	streptogrisin B	299	624	10^{-30}
SFA2_STRFR	serine protease 2	174	583	10^{-28}
PRLA_LYSEN	alpha-lytic protease	397	349	10^{-14}
SP1_RARFA	serine protease I	525	297	10^{-10}
TRYA_DROME	trypsin alpha	256	160	0.0031
LORI_HUMAN	loricrin.	316	157	0.0057
LORI_MOUSE	loricrin.	481	160	0.0058
TRYB_DROME	trypsin beta	253	152	0.009
AIDA_ECOLI	adhesin AIDA-I	1286	155	0.032
TRYD_DROME	trypsin delta	253	139	0.054
GSEP_BACLI	glutamyl endopeptidase	316	140	0.059
TRYG_DROME	trypsin gamma	253	138	0.061
TRYP_FUSOX	trypsin	248	135	0.091
APMU_PIG	apomucin mucin core protein	1150	144	0.13
SLAP_CAUCR	S-layer paracrys. surf. prot	1025	142	0.15
TRY4_LUCCU	trypsin alpha-4	255	130	0.19

Fig. 6. From glutamyl endopeptidase to TRYA_DROME.

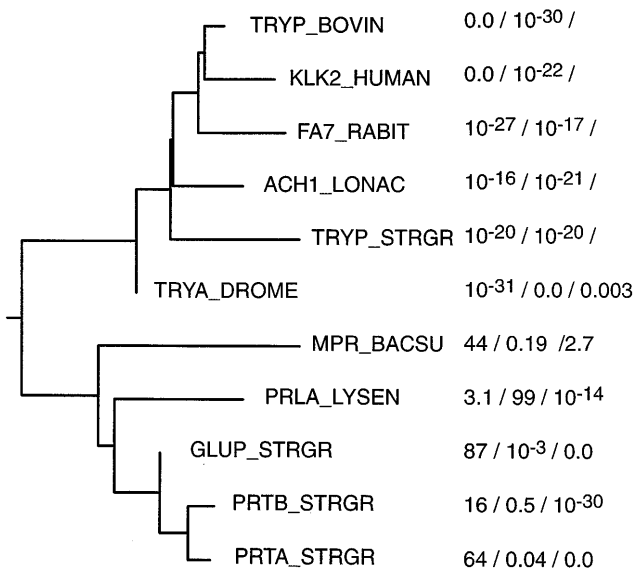

Fig. 7. Similarity and homology—a serine protease family tree.

protein must fold into a functional structure—which are reflected in its sequence. Protein sequence similarity can routinely be used to infer relationships between proteins that last shared a common ancestor 1–2.5 billion years ago. Our ability to identify distantly related proteins has improved over the past 5 years with the use of optimized scoring parameters (Pearson, 1995) and the development of accurate statistical estimates. In using sequence similarity to infer homology, one should remember the following.

1. Always compare protein sequences if the genes encode proteins. Protein sequence comparison will typically double the look-back time over DNA sequence comparison.
2. Homologous sequences are usually similar over an entire sequence or domain. Matches that are > 50% identical in a 20–40 amino acid region frequently occur by chance.
3. While most sequences that share statistically significant similarity (E() < 0.02) are homologous, many distantly related homologous sequences do not share significant homology. (Significant similarity in low-complexity regions does not imply homology.)
4. By focusing on the statistical significance of a similarity and identifying the highest scoring unrelated sequence in a database search, you can both confirm that the statistical estimates are accurate and potentially identify distantly related family members.
5. Homologous sequences share a common ancestor, and thus a common protein structure. Depending on the evolutionary distance and divergence path, two or more homologous sequences may have very few absolutely conserved residues. However, if homology has been inferred between **A** and **B**, between **B** and **C**, and between **C** and **D**, **A** and **D** must be homologous, even if they share no significant similarity when compared directly. In evaluating the results of a similarity search, remember that there is an evolutionary tree that connects the family members.

Motifs revisited

This review argues that sequence similarity searching, rather than motif identification, is the most reliable method for identifying distantly related protein sequences. However, motif searches are frequently used to characterize a newly determined sequence. While motifs can be very valuable for identifying functional sites in a protein, one must be very careful in basing sequence identifications on motif patterns alone. Thus, if a newly determined protein sequence contains the G-D-S-G-G motif, but does not share strong similarity (E() < 20) with any of the hundreds of trypsin-like serine proteases in the protein databases, is it likely to be homologous to trypsin and share the same protein fold? It seems unlikely, since so many very distantly related members of the family are known. However, if a protein sequence shares high, but not significant (0.02 < E() < 20) sequence similarity with several distantly related members of the family, the presence of the two motifs in Figure 1 would provide strong supporting evidence that a new branch in the serine protease family had been found.

Alternatively, if a sequence shares significant similarity with proteins from several branches of the serine protease family tree, but does not contain the G-D-S-G-G motif, it is very likely that it adopts the serine protease protein fold, although it may not function as a protease. Thus, when enzymatic mechanisms are known, motifs can be used to confirm functional aspects of homologous proteins. However, in the absence of strong similarity to any member of a large protein family, motifs are unreliable for inferring protein homology.

References

Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) Issues in searching molecular sequence databases. *Nature Genet.*, **6**, 119–129.

Bairoch,A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.*, **19(Suppl.)**, 2241–2245.

Bairoch,A. and Boechmann,B. (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, **19(Suppl.)**, 2247–2249.

Bult,C.J. *et al.* (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058–1073.

Doolittle,R.F. (1994) Convergent evolution: the need to be explicit. *Trends Biochem. Sci.*, **19**, 15–18.

Doolittle,R.F. (1995) The multiplicity of domains in proteins. *Annu. Rev. Biochem.*, **64**, 287–314.

Dujon,B. (1996) The Yeast Genome Project, what did we learn? *Trends Genet.*, **12**, 263–270.

Felsenstein,J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

Fleischmann,R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Fraser,C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.

Pearson,W.R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.*, **4**, 1145–1160.

Pearson,W.R. (1996) Effective protein sequence comparison. *Methods Enzymol.*, **266**, 227–258.
Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) ClustalW: Improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Received on November 21, 1996; revised on January 10, 1997; accepted on January 28, 1997