Protein sequence comparison and Protein evolution

Tutorial - ISMB2000

William R. Pearson*

Department of Biochemistry and Molecular Genetics,
Jordan Hall, Box 800733
University of Virginia, Charlottesville, VA 22908, USA

October, 2001

Contents

*FAX: (804) 924-5069; email: wrp@virginia.EDU

1 Introduction

The concurrent development of molecular cloning techniques, DNA sequencing methods, rapid sequence comparison algorithms, and computer workstations has revolutionized the role of biological sequence comparison in molecular biology. As a result, the role of protein sequence data in molecular biology and biochemistry has dramatically changed. Twenty-five years ago, protein sequence determination was

usually one of the last steps in the characterization of a protein. Now the process is reversed, so that it is common to clone and sequence a gene of biological interest–e.g., one that is induced by serum stimulation, or a developmental change, or a chromosomal rearrangement associated with a disease. This is the fundamental premise of the human genome project—that one can first sequence all the genes in an organism and then infer their function by sequence analysis.

Today, the most powerful method for inferring the biological function of a gene (or the protein that it encodes) is by sequence 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. One of the more dramatic discoveries was the identification of a new tumor suppressor gene in humans that is related to yeast and *E. coli* DNA repair enzymes. This discovery, the result of a similarity search, both told the investigators that they had identified the appropriate gene and demonstrated clearly the nature of the oncogenic mutation. As entire genomes from bacteria, yeast, and simple eukaryotes become available, protein sequence comparison will become an even more powerful tool for understanding biological function.

Protein sequence comparison is our most powerful tool for characterizing protein sequences 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 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.

This tutorial examines how the information conserved during the evolution of a protein molecule can be used to infer reliably *homology*, and thus a shared protein fold and possibly a shared active site or function. We will start by reviewing a geological/evolutionary time scale. Many protein sequences can be used to infer reliably events that happened more than a billion years ago. Remarkably, some protein sequences change so slowly that they could be used to "date" events that took place more than 5 billion years ago, had the proteins existed. Next we will look at the evolution of several protein families. During the tutorial, these families will be used to demonstrate that homologous protein ancestry can be inferred with confidence. We will also examine different modes of protein evolution and consider some hypotheses that have been presented to explain the very earliest events in protein evolution.

The next part of the tutorial will examine the technical aspects of protein sequence comparison. Both optimal and heuristic algorithms and their associated parameters that are used to characterize protein sequence similarities are discussed. Perhaps more importantly, we will survey the statistics of local similarity scores, and how these statistics can both be used to improved the selectivity of a search and to evaluate the significance of a match.

We will then examine distantly related members of three protein families, the serine proteases, the glutathione transferases, and the G-protein-coupled receptors (GCRs). The serine proteases are used to emphasize that even when a highly conserved motif is found throughout a family, similarity extends over a much longer region. The glutathione transferases and GCRs are very diverse families whose members frequently do not share significant pair-wise similarity. The relative strengths of strategies to characterize

such relationships will be examined.

Finally, we will discuss how sequence similarity can be used to examine internal repeated or mosaic structures in proteins. Such repeated structures can arise from either divergence—calmodulin EF-hand repeats and EGF-domains—or convergence—tropomyosin and transcription factor coiled-coil.

This tutorial is directed towards examining protein evolution. Most of the algorithms and methods that are applied to protein evolution can be used with DNA sequences as well. However, in general, DNA sequence comparisons are far far less informative than protein sequence comparisons (see Fig. 8). DNA sequences that do not encode proteins or structural RNAs (e.g. ribosomal RNAs) diverge very rapidly, so that it is usually difficult to detect reliably non-coding DNA sequence homologies for sequences that diverged more than 200 million years ago. In contrast, even the most rapidly changing protein sequences can detect sequences that are 200 million years old; typically protein sequence comparisons detect sequences that diverged 1 billion years ago. Thus, the most important lesson from this tutorial is, when searching sequence databases for homologous sequences, to use protein sequences whenever possible.

1.1 Evolutionary time scales

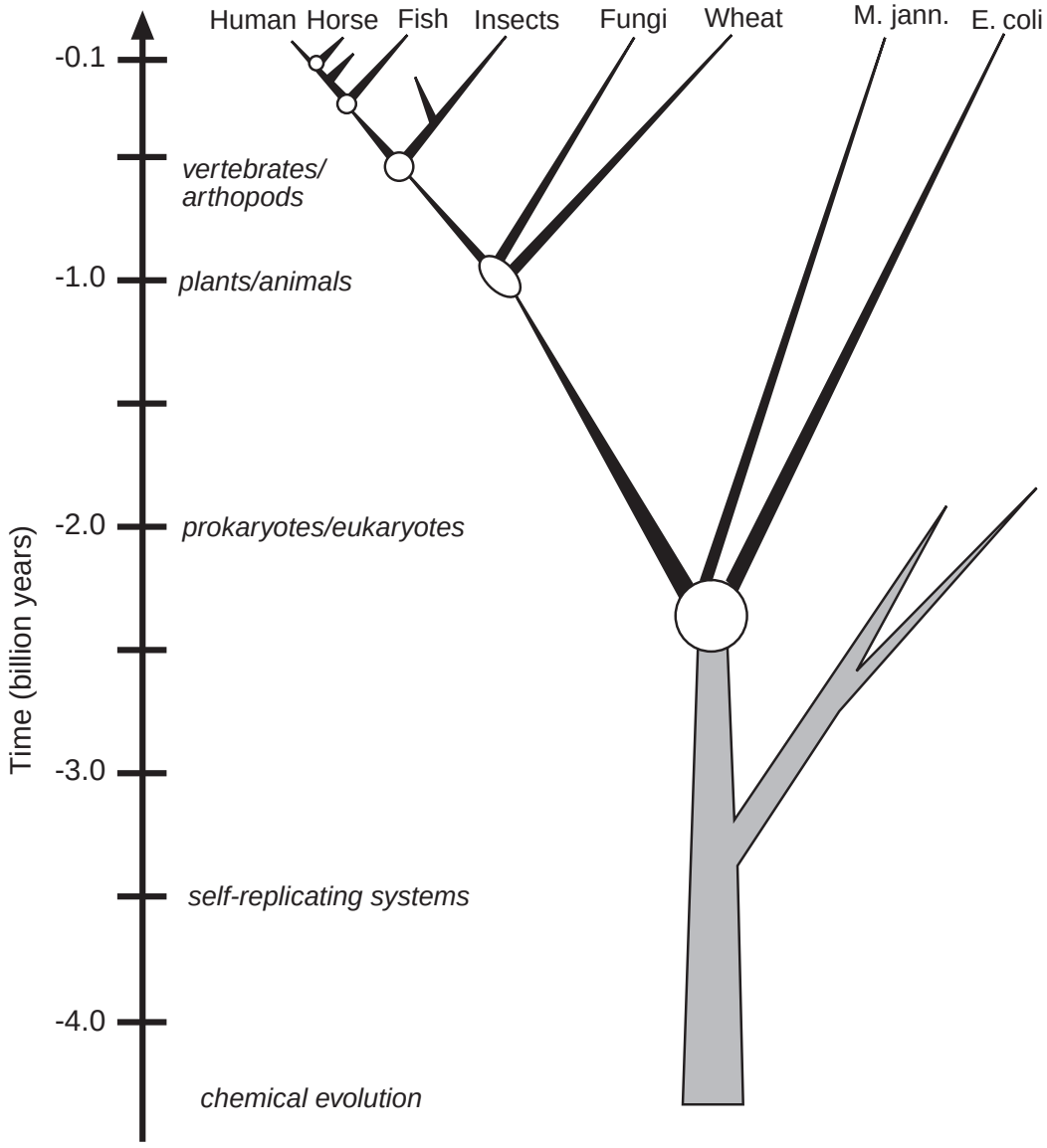
When we search for *homologous* proteins, we are trying to identify proteins that shared a common ancestor in the past. Fig. 1 shows a general evolutionary tree that reaches back to the beginning of the earth's history. The goal of protein sequence comparison is to take a protein sequence, for example from a human chromosome, and search a protein database to find *homologous* sequences, often from very divergent organisms. Thus, if the similarity search produces significant matches with a protein found in yeast, then an ancestral protein must have existed in an organism at least 1 billion years ago and that the descendants of that organism preserved the sequence in modern day humans and yeast. Likewise, if a yeast protein is homologous to one found in *E. coli*, that sequence must have existed in 2 billion years ago in the primordial organism that gave rise to bacteria and fungi.

When we examine protein or DNA sequences, we are almost always studying modern (present day) sequences. Thus, it does not make any sense to say that a yeast or bacterial sequence is more primitive than a mammalian sequence; all sequences are contemporary. As we will see later, however, there are examples of sequences that are found only in vertebrates, or only in animals or plants but not both. Such sequences are less ancient than those found both in mammals and bacteria.

For organisms that diverged within the past 600 My (million years), inferences about divergence times for modern organisms are taken from geological data; more ancient divergence times are inferred from extrapolations of evolutionary "clocks." Evolutionary clocks are based both on slowly changing protein sequences and on ribosomal RNA sequences; such divergence time estimates require a rate of change that is constant on average. The oldest fossils are of prokaryotes in rocks about 2.5 billion years old; this geological age is consistent with that inferred from evolutionary divergence rates.

Table 1 summarizes some important milestones in evolutionary time, and, when considered with Table 2, gives a better perspective on the evolutionary horizons provided by different protein families. The theoretical lookback times in Table 2 are based on the assumption that one can identify proteins that share about 20% sequence identity throughout their entire length. It will be clear from later examples that if two protein sequences share 25% identity across their lengths, they are homologous, and that in

Figure 1: The tree of life



Adapted from Dayhoff *et al.*, 1978.

Table 1: Some Important dates in history

Origin of the universe	-12[a]	±2
Formation of the solar system	-4.6	±0.4
First self-replicating system	-3.5	±0.5
Prokaryotic-eukaryotic divergence	-1.8	±0.3
Plant-animal divergence	-1.0	
Invertebrate-vertebrate divergence	-0.5	
Mammalian radiation beginning	-0.1	

[a]Billions of years. From Doolittle *et al.*, 1986.

some cases, convincing evidence of common ancestry can be deduced from similarities as low as 20%. These look-back times can be confirmed in practice; for example, with sensitive sequence comparison algorithms, significant similarity between plant and animal globins can be found.

1.2 Similarity, Ancestry and Structure

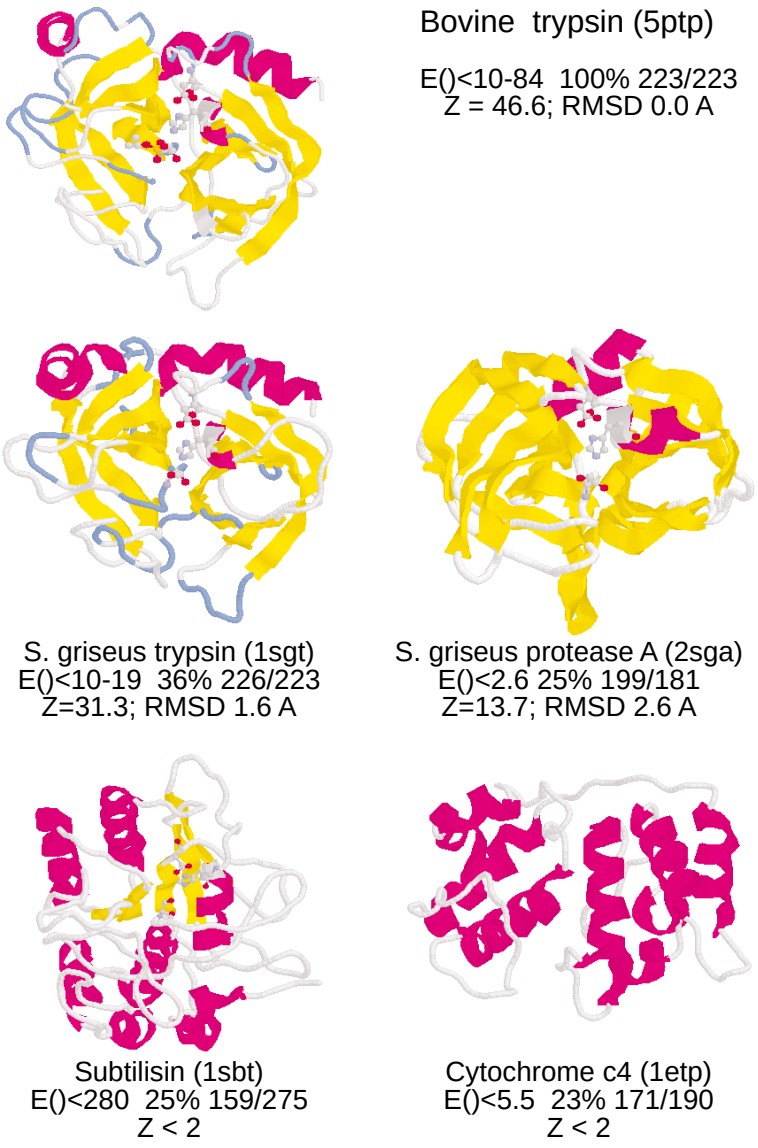
The inference of homology — common ancestry — is the most powerful conclusion that one can draw from a similarity search because homologous proteins share similar three-dimensional structures. This can be seen in Fig. 2, where the structures of three members of the serine protease superfamily are shown.

Table 2: Evolutionary Horizons

Protein	PAMs[a]/100 residues /10^8 years	Theoretical Lookback time[b]	Horizon
Pseudogenes	400	45[c]	Primates, Rodents
Fibrinopeptides	90	200	Mammalian Radiation
Lactalbumins	27	670	Vertebrates
Ribonucleases	21	850	Animals
Hemoglobins	12	1.5[d]	Plants/Animals
Acid Proteases	8	2.3	Prokayrotic/Eukarotic
Triosphosphate isomerase	3	6	Archaen
Glutamate dehydrogenase	1	18	

[a]PAMs, point accepted mutations. [b]Useful lookback time, 360 PAMs, 15% identity.
[c]Millions of years. [d]Billions of years. Adapted from Doolittle *et al.*, 1986

Figure 2: Structural similarity in related proteins – serine proteases



Expectation values (E()), percent identity, the length of the alignment are shown with respect to bovine trypsin. The last two numbers report the length of the alignment and the length of the library sequence whose structure is shown.

Two of these proteins, bovine chymotrypsin and *S. griseus* trypsin, share strong sequence similarity while the third related sequence, *S. griseus* protease A, does not share significant similarity (E()< 66) yet the protein has a very similar structure. Thus, as will be seen throughout this chapter, homologous proteins need not share statistically significant, or even detectable, sequence similarity.

Cytochrome c4 is an example of a very high-scoring, but unrelated protein whose structure is known.

This high scoring unrelated sequence does not share any structural similarity with trypsin or other serine proteases. If two proteins are not homologous, one cannot draw any conclusion about their structural similarity, even though they may have high similarity scores.

1.3 Modes of Evolution

Figure 3: Orthologous sequences — The cytochrome ‘c’ family

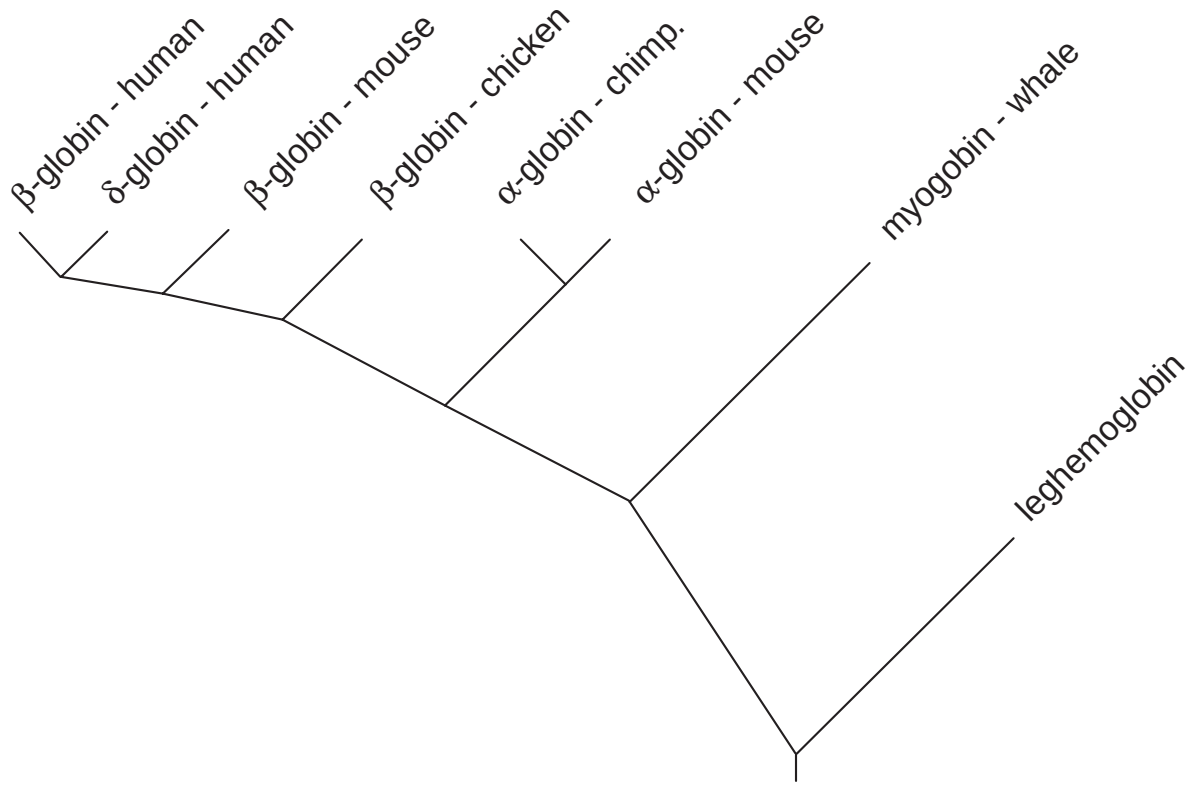
Cytochrome ‘c’s comprise a family of orthologous proteins that are found in all organisms. The sequences on this tree are *orthologous* — two cytochrome ‘c’s are different because they are in different species.

1.3.1 Conventional divergence from a common ancestor

Homologous sequences can be divided into two groups: (1) *orthologous* sequences — sequences that differ because they are found in different species; and (2) *paralogous* sequences — sequences that differ because of a gene duplication event. Fig. 3 shows an evolutionary tree for *orthologous* cytochrome ‘c’

sequences. The branching pattern, which reflects the differences between cytochrome ‘c’ sequences, matches the evolutionary relationships of the species that express the proteins.

Figure 4: Orthology and paralogy — The globin family



Members of the globin oxygen binding protein family have evolved through a series of gene duplications and speciation events. The human α and δ genes duplicated less than 50 Mya (δ chains are found in primates, but not in other mammals).

In general, the organismal tree and the sequence tree will not match if the sequences are *paralogous*. Members of the globin oxygen binding protein family are both *orthologous* — they differ because of speciation — and *paralogous* p — they differ because of gene duplications. Thus, human α-globin, mouse α-globin, and chicken α-globin are all orthologs, they differ because of the speciation events that gave rise to humans, rodents, and birds. Mouse β globin and human α globin are paralogous; they differ because of a gene duplication that created the α and β subunits some 600 Mya (million years ago). An evolutionary tree based on human α, chicken α, and mouse β would imply that humans are more closely related to chickens than to mice. While such a mistake is unlikely in a well-studied family like the globins, it can be quite common in large, diverse, and poorly characterized families like the G-protein-coupled receptors (Fig. 21).

1.3.2 Sequence similarity and homology, the H^+ ATPase

Our first example of the significant sequence similarity shared by homologous proteins will use one of the chains of the H^+-ATPase, or proton-pump, used to convert energy to ATP in the mitochrondria and chloroplasts of aerobic organisms.

Figure 5: The PAM250 matrix

Cys	12																			
Ser	0	2																		
Thr	-2	1	3																	
Pro	-1	1	0	6																
Ala	-2	1	1	1	2															
Gly	-3	1	0	-1	1	5														
Asn	-4	1	0	-1	0	0	2													
Asp	-5	0	0	-1	0	1	2	4												
Glu	-5	0	0	-1	0	0	1	3	4											
Gln	-5	-1	-1	0	0	-1	1	2	2	4										
His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

The similarity scores in Figs. 6–8 were calculated using the Smith-Waterman algorithm (Smith & Waterman, 1981,Sec. 2.2), a method that guarantees to calculate the best (optimal) score between any two protein or DNA sequences, given a scoring matrix and gap penalties. Fig. 5 shows the PAM250 matrix, which was developed almost 20 years ago by Dayhoff and her colleagues (Dayhoff *et al.*, 1978). The PAM250 matrix, or modern versions such as the BLOSUM50 matrix used here, incorporates information about the likelihood that one amino-acid will be mutated into another over evolutionary time. Thus, changes that are very unlikely to occur in evolution, for example the substitution of the very small glycine residue for the very large tryptophan residue, are given large negative scores (-7 in Fig. 5), while conservative changes, such as the substitution of lysine by arginine (both have basic side chains), are given positive scores ($+3$). The scores for identical matches also vary in the PAM250 matrix, depending on whether the amino-acids are common (e.g. serine and methionine), and thus likely to be aligned by chance, or rare (e.g. cysteine and tryptophan). There is a well-developed statistical theory for substitution matrices (Altschul, 1991), which will be discussed in section 3.2.

Table 3 reports similarity scores and their statistical significance from a search of the PIR annotated protein sequence database (PIR1, release 44, March, 1995) using the human H^+-ATPase as a query sequence. There is excellent agreement between the expected and actual distributions of similarity scores. In this search, all of the library sequences related (homologous) to the query sequence obtained scores higher than any of the unrelated sequences. However, a number of unrelated sequences obtained very

Figure 6: Searching with human ATP-ase, similarity scores

```
       opt      E()
< 20    17     0:=           one = represents 22 library sequences
  22     0     0:
  24     0     0:
  26     2     0:=
  28     7     3:*
  30     7    18:*
  32    45    68:===*
  34   166   184:========*
  36   337   379:================ *
  38   581   626:========================== *
  40   869   873:=======================================*
  42  1009  1067:================================================  *
  44  1276  1177:============================================================*====
  46  1253  1198:============================================================*==
  48  1199  1147:=========================================================*==
  50  1032  1047:================================================*
  52   949   920:=========================================*==
  54   838   786:=====================================*===
  56   578   657:==========================  *
  58   467   539:======================  *
  60   393   437:================== *
  62   339   350:===============*
  64   276   278:=============*
  66   214   220:==========*
  68   188   173:========*=
  70   140   136:=======*
  72   131   106:====*=
  74    88    83:====*
  76    71    64:==*=
  78    48    50:==*
  80    43    39:=*
  82    38    30:=*
  84    27    24:=*
  86    21    18:*
  88    15    14:*
  90    17    11:*
  92     7     8:*         :========*    = represents 1 library sequence
  94    22     7:*         :=======*================
  96     3     5:*         :=== *
  98     8     4:*         :===*=====
 100     6     3:*         :==*===
 102     5     2:*         :=*===
 104     9     2:*         :=*========
 106     4     1:*         :*===
 108     5     1:*         :*=====
 110     4     1:*         :*===
 112     4     1:*         :*===
 114     4     1:*         :*===
 116     6     0:=         *======
 118     1     0:=         *=
>120    32     0:==        *================================
```

high scores; 10 of the 32 sequences with z-scores > 120 (7 standard deviations above the mean [1]) are not members of the H+-ATPase family.

Fig. 6 shows the distribution of similarity scores between human H^+-ATPase (PIR entry PWHU6) and each protein sequence in the PIR1 (rel. 44) database. The '=' symbols in the histogram show the distribution of normalized similarity scores calculated during the search, thus, 393 sequences in the PIR1 library had scores of 60 or 61. The '$*$' symbols report the expected number of sequences with the indicated range of scores for a search of a database of this size, based on random chance. The basis for the statistical estimates will be discussed in section 3.

While Table 3 shows that all of the members of this family have siginificant similarity with the human enzyme, Fig. 7 gives a more realistic perspective of the family's evolutionary history by considering every possible pairwise alignment. When the *E. coli* enzyme is used to search the database for related H^+-ATPases, the ranking of the different sequences changes, but sequences distant from the *E. coli* sequence have more significant similarities than those distant from the human sequence.

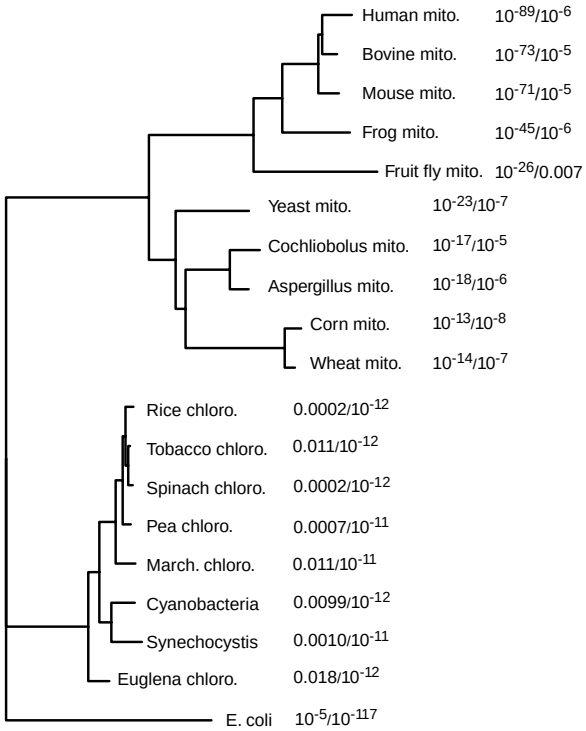
[1] The z-scores plotted have a mean of 50 and a standard deviation of 10.

Table 3: Searching with human ATP-ase, high-scoring sequences

The best scores are:		s-w	bits	E(14548)	%	len
PWHU6	H+-trans. ATP synth.—human mito.	1400	327	10^{-89}	100.0	226
PWBO6	H+-trans. ATP synth.—bovine mito.	1157	271	10^{-73}	77.9	226
PWMS6	H+-trans. ATP synth.—mouse mito.	1118	262	10^{-71}	75.7	226
PWXL6	H+-trans. ATP synth.—frog mito.	745	177	10^{-45}	53.3	226
PWFF6	H+-trans. ATP synth.—fruit fly mito.	471	115	10^{-26}	37.5	224
PWBY3	H+-trans. ATP synth.—yeast mito.	438	107	10^{-23}	36.2	232
PWAS6N	H+-trans. ATP synth.—aspergillus mito.	365	91	10^{-18}	30.4	230
PWKQ6	H+-trans. ATP synth.—Cochliobolus mito.	353	88	10^{-17}	31.3	214
PWWT6	H+-trans. ATP synth.—wheat mito.	309	78	10^{-14}	28.9	235
PWNT6M	H+-trans. ATP synth.—tobacco mito.	309	78	10^{-14}	28.3	233
PWZM6M	H+-trans. ATP synth.—corn mito.	283	72	10^{-13}	31.1	291
LWEC6	H+-trans. ATP synth.—E. coli	178	48	10^{-5}	23.3	236
LWRZ6	H+-trans. ATP synth.—rice chloro.	144	40	0.00063	24.2	231
PWPMA6	H+-trans. ATP synth.—pea chloro.	143	40	0.00074	25.0	232
PWYBAA	H+-trans. ATP synth.—Synechocystis	142	40	0.00099	26.5	170
PWSPA6	H+-trans. ATP synth.—spinach chloro.	138	39	0.00016	24.2	231
PWYCA6	H+-trans. ATP synth.—cyanobacteria	127	36	0.0099	26.3	167
LWNT6	H+-trans. ATP synth.—tobacco chloro.	126	36	0.011	22.1	231
LWLV6	H+-trans. ATP synth.—Marchiantia chloro.	126	36	0.011	24.0	167
PWEGAC	H+-trans. ATP synth.—Euglena chloro.	123	35	0.018	25.7	214
JQ0026	ATP/ADP translocase tlc1—Rickettsia	122	35	0.045	24.7	154
S17420	ubiquinol-cytochrome-c reductase	113	33	0.14	23.4	158
S17418	ubiquinol-cytochrome-c reductase	108	32	0.30	24.5	208
QXBO2M	NADH dehydrogenase (ubiquinone)	107	32	0.32	26.1	211
S17415	ubiquinol-cytochrome-c reductase	105	31	0.49	27.7	137
DNHUN2	NADH dehydrogenase (ubiquinone)	103	31	0.61	20.1	149
CBHU	ubiquinol-cytochrome-c reductase	102	31	0.79	26.8	205
QRECAA	amino acid trans. protein—E. Coli	104	31	0.82	23.4	111
S17419	ubiquinol-cytochrome-c reductase	101	30	0.92	23.4	158
S17407	ubiquinol-cytochrome-c reductase	99	30	1.3	23.6	140
QQBEN5	integral membrane protein—saimiriine herp	98	30	1.4	20.8	202

The horizontal line indicates the separation been the lowest scoring related sequences and the highest scoring unrelated sequence.

Figure 7: Phylogeny of H^+-ATPases



An evolutionary tree of H^+-ATPases (subunit 6). Sequences were aligned using the GCG PILEUP program, distances calculated using the GCG DISTANCES program, and the tree constructed using the Neighbor-Joining algorithm (GCG GROWTREE). Expectation values from a search with the human H^+-ATPase (PWHU6, Table 3) and a search with the *E. coli* sequence are shown.

Figure 8: Searching with human ATPase, high-scoring alignments

```
LWEC6 H+-transporting ATP synthase (EC 3.6.1.34) prot   (272 aa)
 s-w opt: 178  Z-score: 218.7  bits: 48.0 E(): 3.3e-06
Smith-Waterman score: 178;  23.729% identity (28.141% ungapped) in 236 aa overlap

                     10        20        30        40        50       60        70        80
PWHU6        MNENLFASFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTKGRTWSLMLVSLIIFIA
                    :..  ..:::   ....::  .   ...  . ... :. . .: ... ..  :.. :.. . . .......
LWEC6  QNPPATFWTINIDSMFFSVVLGL---LFLVLFRSVAKKATSG-VPGKFQTAIELVIGFVNGSVKDMYHGKSKLIAPLALTIFVWVF
              40        50           60        70         80        90      100       110

                        90              100       110       120      130            140
PWHU6  TTNLLGLLP---------HSF-------TPTTQLSMNLAMAIPLWAGTVIMGFRSKIKNALAHFLPQGTPTPL-----IPMLVIIE
        ::. :::          : .       .:.......:.::. ..   ... : :  .... :  . :  :.     ::. .:.:
LWEC6  LMNLMDLLPIDLLPYIAEHVLGLPALRVVPSADVNVTLSMALGVF---ILILFYSIKMKGIGGFTKELTLQPFNHWAFIPVNLILE
         120       130       140       150          160       170       180         190

        150       160       170       180       190        210       220
PWHU6  TISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYLHDNT
       .::: .:..:..:: .:. ::.:.. ::..     :   :      :: :::.          .::..: .:. .::
LWEC6  GVSLLSKPVSLGLRLFGNMYAGELIFILIAGLLPWWSQWILNVPWAIFHILIIT---------LQAFIFMVLTIVYLSMASEEH
        200       210       220       230       240       250                260       270

================

PWEGAC H+-transporting ATP synthase (EC 3.6.1.34) cha   (252 aa)
 s-w opt: 123  Z-score: 151.6  bits: 35.4 E(): 0.018
Smith-Waterman score: 123;  25.701% identity (30.220% ungapped) in 214 aa overlap

                       10        20        30        40        50        60            70
PWHU6            MNENLFASFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTK-GRT----WSLM
                                  .::: :  : : :.: :  . . ...: .:.:... .  . .:. :.     :  .
PWEGAC IANVEVGQHFYWSILGFQIHGQVLINSWIVILIIGF--LSIYTTKNL--TLVPANKQIFIELVTEFITDISKTQIGEKEYSKWVPY
      20        30        40        50          60          70        80        90      100

           80         90              100       110       120        130      140       150
PWHU6  LVSLIIFIATTNLLG-LLPHSFT--PTTQL---SMNLAMAIPLWAGTVIMGFRSKI-KNALAHFLPQGTPTPLIPMLVIIETISLL
       . ...:: ..:  : :.: ..   :. .:   .  ..  .  :   : .  : . . :..:..:     :::.. . :.: ..
PWEGAC IGTMFLFIFVSNWSGALIPWKIIELPNGELGAPTNDINTTAGLAILTSLAYFYAGLNKKGLTYFKKYVQPTPILLPINILEDFT--
          110       120       130       140       150       160       170      180

            160       170       180       190       200       210       220
PWHU6  IQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYLHDNT
        .:..:. :: .:: : .:.. .. :         .:  ::. . ::.: ..   .. ::: .:. : . :.
PWEGAC -KPLSLSFRLFGNILADELVVAVLVSL--------VP--LIVPVPLIFLGLF---TSGIQALIFATLSGSYIGEAMEGHH
       190        200       210                 220          230       240       250
```

Alignments of human H^+-ATPase with the *E. coli* homologue and a plant chloroplast homologue. Despite the considerable evolutionary distance (both sequences diverged at least 2 Bya), the pairs of sequence share more than 20% identity across almost their entire lengths. ‘:’ symbols denote identities; ‘.’ denote conservative substitutions. Searches were performed with the BLOSUM50 matrix and gap penalties of -12/-2.

1.3.3 Protein families diverge at different rates

For many protein families with a variety of divergence rates, the rate of change over evolutionary time is relatively constant. These rates can be used to date the divergence events (e.g. plants and animals) that occurred more than 600 Mya and thus do not have a fossil record. However, different protein families diverge at different rates, so that, in general, the number of differences between a pair of sequences cannot be used to estimate the time the two sequences diverged. This is particularly true for paralogous sequences; once a sequence has duplicated, it may change very rapidly before selective pressure on its new function slows its rate of change. Thus, in Table 4 there are several members of growth hormone superfamily—growth hormone, sommatotropin, and prolactin—with different divergence rates.

Figure 9: The limits of sequence similarity

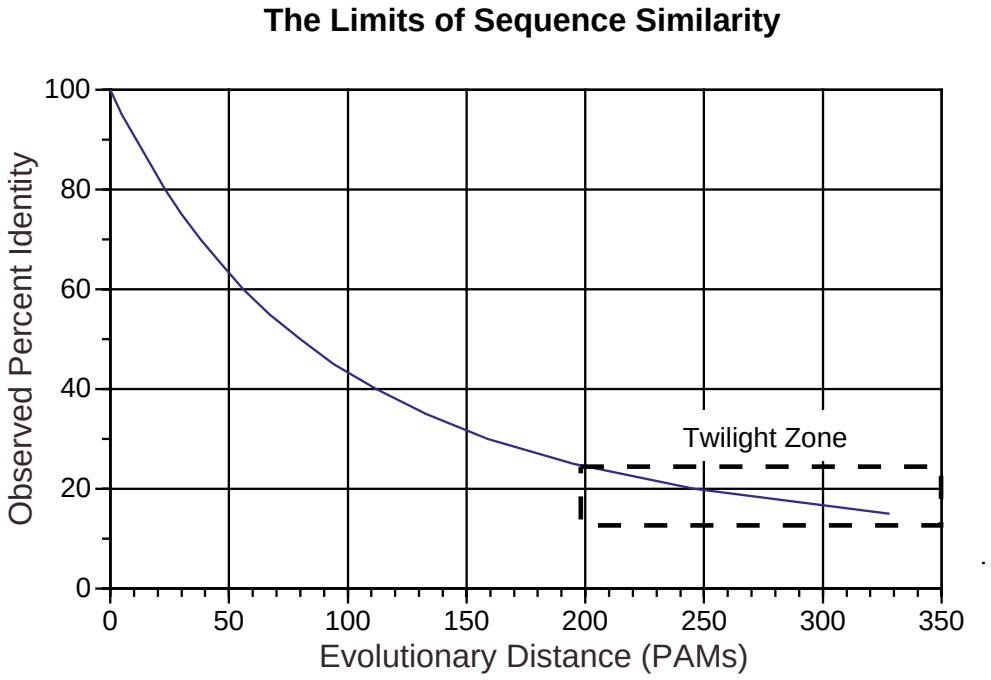


Table 4: Rates of change in protein families

Protein	*Rate*[a]	*Protein*	*Rate*
Fibrinopeptides	90	Thryrotropin beta chain	7.4
Growth hormone	37	Parathyrin	7.3
Ig kappa chain C region	37	Parvalbumin	7.0
Kappa casein	33	BPTI Protease inhibitors	6.2
Ig gamma chain C region	31	Trypsin	5.9
Lutropin beta chain	30	Melanotropin beta	5.6
Ig lambda chain C region	27	Alpha crystallin A chain	5.0
Complement C3a	27	Endorphin	4.8
Lactalbumin	27	Cytochrome b_5	4.5
Epidermal growth factor	26	Insulin	4.4
Somatotropin	25	Calcitonin	4.3
Pancreatic ribonuclease	21	Neurophysin 2	3.6
Lipotropin beta	21	Plastocyanin	3.5
Haptoglobin alpha chain	20	Lactate dehydrogenase	3.4
Serum albumin	19	Adenylate cyclase	3.2
Phospholipase A_2	19	Triosephosphate isomerase	2.8
Protease inhibitor PST1 type	18	Vasoactive intestinal peptide	2.6
Prolactin	17	Corticotropin	2.5
Pancreatic hormone	17	Glyceraldehyde 3-P DH	2.2
Carbonic anhydrase C	16	Cytochrome C	2.2
Lutropin alpha chain	16	Plant ferredoxin	1.9
Hemoglobin alpha chain	12	Collagen	1.7
Hemoglobin beta chain	12	Troponin C, skeletal muscle	1.5
Lipid-binding protein A-II	10	Alpha crystallin B-chain	1.5
Gastrin	9.8	Glucagon	1.2
Animal lysozyme	9.8	Glutamate DH	0.9
Myoglobin	8.9	Histone H2B	0.9
Amyloid A	8.7	Histone H2A	0.5
Nerve growth factor	8.5	Histone H3	0.14
Acid proteases	8.4	Ubiquitin	0.1
Myelin basic protein	7.4	Histone H4	0.1

[a]percent/100 My
From (Nei, 1987; Dayhoff *et al.*, 1978)

1.3.4 Mosaic proteins

"Conventional" protein families, e.g. the globins, cytochrome 'c's, H^+-ATPases, in which protein sequences have diverged from a common ancestor in a direct fashion, typically with only modest changes in the length of the sequence, have been known for more than 30 years. In the past 10 years, a more complex type of protein evolution has been observed—proteins that contain multiple domains from other proteins. These proteins have been called "mosaic" proteins; the domains are frequently inserted through a process called "exon shuffling." Table 7 lists a number of human proteins that are comprised of mosaic domains, but such proteins are not limited to mammals. Similar mosaic structures are common in DNA binding proteins, both in bacteria and eukaryotes.

Table 5: Classification of Protein Families

I. Ancient Proteins

   A. First editions. Direct-line descendacy to human and contemporary prokaryotes. Mostly mainstream metabolism enzymes. Example: triosphosphate isomerase (44.8% identical over 250 aa, E(59000) $< 10^{-36}$).

   B. Second edition. Homologous sequences in human and prokaryotic proteins, but apparently different functions. Example: human glutathione reductase and pseudomonas mercury reductase (31% identical over 438 aa, E(59000) $< 10^{-32}$).

II. Middle-age proteins. Proteins found in most eukaryotes but prokaryotic counterparts are unknown. Example: actin (human and yeast share 88% identical over 375 aa, E() $< 10^{-145}$, other yeast actin homologs share as little as little as 26.4 % over 489 aa, E() $< 10^{-14}$.

III. Modern proteins

   A. Recent vintage. Proteins found in animals or plants but not both. Not found in prokaryotes. Example: collagen.

   B. Very recent inventions. Proteins found in vertebrates but not elsewhere. Example: plasma albumin.

   C. Recent mosaics. Modern proteins clearly the result of exon shuffling. Example: LDL receptor.

From Doolittle *et al.*, 1986.

1.4 Introns Early/Late

The occurrence of mosaic proteins and the discovery of the "exon/intron" structure of genes in the late 1970's led several investigators to suggest that the exon structure of genes reflected the construction of proteins from modular domains (Gilbert & Glynias, 1993). While acceptance of this proposal is quite widespread, it is based on very little data. There is no question that many modern mosaic proteins are constructed by a process of "exon-shuffling" whereby exons from other genes have been combined to build new structures. In addition, for some proteins exons are associated with well defined structural

Table 6: Ancient human proteins

Human protein	Prokaryotic homologue	% identity	E(59,000)
A. First edition type			
Triosephosphate isomerase	*E. coli*	46	$< 10^{-36}$
Phosphoglyceraldehylde dehydrogenase	*B. stearothermophilus*	52	$< 10^{-78}$
Alkaline phosphatase	*E. coli*	28	$< 10^{-20}$
Dihydrofolate reductase	*E. coli*	28	$< 10^{-6}$
Superoxide dismutase (Cu-Zn)	*E. coli*	32	$< 10^{-7}$
Hypoxanthine-guanine phosphoribosyl transferase	*E. coli*	34	$< 10^{-17}$
B. Second edition type			
Glutathione reductase	Mercuric reductase, *Pseudomonas*	31	$< 10^{-32}$
Glutamate dehydrogenase (NAD)	Glutamate dehydrogenase,*E. coli*	29	$< 10^{-24}$
Ornithine transcarbamylase	Aspartate transcarbamylase, *E. coli*	26	$< 10^{-11}$

Adapted from Doolittle *et al.*, 1986

elements. The association of exons with structural elements may reflect and ancient construction of proteins from primordial exons. Alternatively, introns are also capable of invading genes; thus, the association of exons with structures may reflect modern invasions that are less disruptive when they occur between structural elements.

A recent test of the "introns" early hypothesis suggests there is little evidence to support the association of introns with structural boundaries (Stoltzfus *et al.*, 1994.

1.5 DNA vs Protein comparison

While all of the comparison methods described below work on either protein or DNA sequences, one's ability to identify distantly related sequences is reduced dramatically when DNA sequences are used. Fig. 8 compares the statistical significance of the best similarity scores obtained in a search of the GenBank DNA sequence database using a mouse glutathione transferase cDNA clone with the significance of the same alignment in a search of the GenPept protein sequence database (GenPept is derived from GenBank by translating DNA sequences into the encoded protein sequences). Many DNA sequences encoding clearly related proteins, e.g. RABGSTB have similarity scores that are expected to occur several times by chance in a DNA database search. DNA sequences are far less informative, both because they lack the inherent biochemical information that is retained in the PAM250 substitution matrix, and because many changes in DNA sequences (third-base changes) do not change the encoded protein.

Differences in the performance of sequence comparison algorithms are insignificant compared to the loss of information that occurs when one compares DNA sequences. If the biological sequence of interest

Table 7: Mosaic proteins

A. EGF-type	B. C9-type
Epidermal growth factor precursor	Complement C9
Tumor growth factors	LDL receptor
LDL receptor	Notch (*Drosophila*)
Factor IX	*lin-12* (*C. elegans*)
Protein C	
Tissue plasminogen activator	C. Fibronectin finger
Urokinase	Fibronectin
Complement C9	Tissue plasminogen activator
Notch protein (*Drosophila*)	
lin-12 (*C. elegans*)	D. Protease "Kringle"
	Plasminogen
	Tissue plasminogen activator
	Urokinase
	Prothrombin

From Doolittle *et al.*, 1986.

encodes a protein, protein sequence comparison is always the method of choice.

Table 8: DNA vs Protein Sequence Comparison

		score	E(DNA)	E(prot)	E(tx)
MUSGST	Mouse glutathione S-transferase class mu	5090	10^{-233}	10^{-90}	10^{-120}
MUSGSTA	Mouse, glutathione transferase GT9.3 mu	3693	10^{-167}	10^{-73}	10^{-120}
HUMGSTAB	Homo sapiens glutathione transferase	1930	10^{-84}	10^{-60}	10^{-80}
MAMGLUTRA	M.auratus mu class GST	399	10^{-11}	10^{-73}	10^{-11}
RATGSTYD	Rat glutathione S-transferase Yb subunit	399	10^{-11}	10^{-74}	10^{-10}
HSGSTM4	H.sapiens GSTM4 gene for GST	390	10^{-11}	10^{-69}	10^{-10}
RATGSTY	Rattus norvegicus GST	372	10^{-10}	10^{-71}	10^{-10}
HSGSTM1B	H.sapiens GSTM1b gene for GST	358	10^{-9}	10^{-63}	10^{-10}
HSGSTMU3	Human GSTmu3 gene for a GST	322	10^{-7}	10^{-25}	10^{-6}
BTGST	Bovine GST mRNA for GST	249	0.00013	10^{-16}	10^{-22}
HSGSTPI1	Human mRNA for anionic GST	237	0.00049	10^{-17}	10^{-21}
MUSGTF	Mus musculus GST mu	196	0.041	10^{-4}	10^{-6}
CRUGSTP	Chinese hamster GST	196	0.043	10^{-16}	10^{-21}
CRUGSTPIE	Cricetulus griseus GST pi	196	0.04	10^{-16}	10^{-21}
HAMGSTPIE	Mesocricetus auratus GST pi	191	0.13	10^{-16}	10^{-21}
BTRNAXOR	*B.taurus xanthine oxidoreductase*	184	*0.11*	> 10	> 5
HUMKAL2	*Human glandular kallikrein gene*	*170*	*0.59*	> 10	> 5
RNGSTYC2F	R.norvegicus GST Yc1	170	0.67	10^{-6}	> 5
MMGLUT	M.musculus mRNA for GST	168	1.0	10^{-7}	10^{-8}
MUSTHYGP	*Mouse Thy-1.2 glycoprotein*	*163*	*1.3*	> 10	> 5.0
HUMTROPI01	*Human troponin I, slow-twitch isoform*	*161*	*1.7*	> 10	> 5

Expectation values for searches against DNA (score, E(DNA)), protein (E(prot)), and translated DNA (E(tx) databases. A mouse glutathione transferase cDNA sequence (MUSGST) was used to search either the primate (GBPRI), rodent (GBROD), and mammalian (GBMAM) divisions of the GenBank DNA sequence database for the DNA sequence comparisons. Protein expectations (E(prot)) were calculated from a search the translated cDNA sequence against the GenPept sequence database, which includes all of translated GenBank. Unrelated sequences are *italicized*; E(prot) for unrelated sequences are $>> 100$.

2 Alignment methods

A variety of comparison algorithms and scoring parameters can be used to evaluate protein or DNA sequence similarity. In general, the choice the of "best" algorithm depends on the problem to be solved. Thus, algorithms that calculate a local comparison score—i.e., they find the strongest similarity between the two sequences, ignoring differences outside of the most similar region—are usually most appropriate for searching protein and DNA databases,[2] while global comparison algorithms are more appropriate when homology has been established, as when building evolutionary trees. Pattern-based, rather than similarity-based, comparison methods may be preferred when searching for functionally conserved non-homologous domains.

In searching protein sequence databases to identify distantly related homologous proteins, it is important to remember that avoiding high similarity scores with unrelated sequences can be more important as calculating high scores for related sequences. There are more than 50,000 protein sequences in comprehensive protein databases, while the typical family of proteins has fewer than 100 members. Thus, comparison algorithms, scoring matrices and gap penalties that produce the best alignments may not be the most effective for searching protein sequence databases (Pearson, 1995; Pearson, 1998).

2.1 Algorithms

Two general classes of comparison algorithms are used to calculate similarity scores to infer sequence homology: rigorous algorithms that are guaranteed to calculate an optimal similarity score, e.g. the NeedlemanWunsch (Needleman & Wunsch, 1970) and SmithWaterman (Smith & Waterman, 1981) algorithms, and rapid heuristic algorithms that do not guarantee to calculate an optimal score for every sequence in a library, e.g. FASTA (Pearson & Lipman, 1988) and BLAST(Altschul *et al.*, 1990). Table 2.1 summarizes widely used algorithms for biological sequence comparison.

Two optimal algorithms for calculating similarity scores have been described, the NeedlemanWunsch algorithm (Needleman & Wunsch, 1970), which calculates a "global" similarity score between two sequences, and the Smith-Waterman algorithm (Smith & Waterman, 1981), which calculates a "local" similarity score. Global scores require the alignment to begin at the beginning of each sequence and extend to the end of each sequence. Global alignments cannot be used to detect the relationship between DNA binding domains in homeobox proteins or the calcium binding domains shared between calmodulin and calpain. Likewise, global alignment algorithms often do not detect the relationships between mosaic proteins. Global similarity scores can be calculated with or without penalties for gaps at the ends of the sequences.

Local alignment algorithms identify the most similar region shared between two sequences. Thus, homologous calcium binding domains embedded in non-homologous proteins can be detected with local alignment methods. In addition, a local alignment algorithm can be used to find the exons in a genomic DNA sequence by aligning it with its encoded mRNA. Local alignment algorithms are required to identify homologies in mosaic proteins, and they can be used to detect internal domain duplications as well. Table 10 compares the scores of global, global without end-gap-penalties, and local similarity scores for

[2] For genomic DNA sequences, there is no logical alternative.

Table 9: Algorithms for comparing protein and DNA sequences

algorithm	value calculated	scoring matrix	gap penalty	time required	
Needleman-Wunsch	global similarity	arbitrary	penalty/gap q	$O(n^2)$	Needleman and Wunsch, 1970
Sellers	(global) distance	unity	penalty/residue rk	$O(n^2)$	Sellers, 1974
Smith-Waterman	local similarity	$\hat{S}_{ij} < 0.0$	affine $q+rk$	$O(n^2)$	Smith and Waterman, 1981 Gotoh, 1982
FASTA	approx. local similarity	$\hat{S}_{ij} < 0.0$	limited gap size $q+rk$	$O(n^2)/K$	Lipman and Pearson, 1985 Pearson and Lipman, 1988
BLASTP	maximum segment score	$\hat{S}_{ij} < 0.0$	multiple segments	$O(n^2)/K$	Altshul et al., 1990

a variety of related and unrelated proteins.

Rigorous sequence comparison algorithms, like the Smith-Waterman algorithm, require time proportional to $O(mN)$, where m is the length of the query sequence and N is the number of amino acids in the protein sequence library. Modern high-performance unix workstations can compare a 300 residue protein sequence (human opsin) to the $40,000$ entry, 15,000,000 amino acid Swiss-Prot 31 database in less than 10 minutes.

Although very rapid[3] algorithms are available for calculating optimal global similarity scores between two sequences, particularly with unit cost scores, such algorithms are rarely appropriate for biological sequence comparison. Unit cost algorithms must discard the substantial biochemical information encoded in the PAM250 matrix. Most rapid optimal algorithms calculate only global similarities; such comparisons are not useful for DNA sequence comparison because the "ends" required for a global sequence comparison are usually arbitrary.

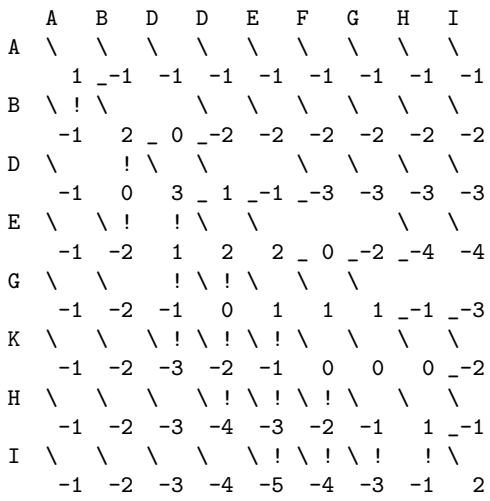
[2] $O(Nd)$, where N is the length of a sequence and d is the number of differences between the two sequences.

Table 10: Global and local sequence similarity scores

PIR Entry				Similarity Score			Distance
				Global		Local	
				End Penalty	No End Penalty		
HBHU	vs	HBHU	Hemoglobin beta-chain—human	725	725	725	0
		HAHU	Hemoglobin alpha-chain—human	314	320	322	152
		MYHU	Myoglobin—Human	121	164	166	212
		GPYL	Leghemoglobin—Yellow lupin	8	28	43	239
		LZCH	Lysozyme precursor—Chicken	−107	16	32	220
		NRBO	Pancreatic ribonuclease—Bovine	−124	16	31	280
		CCHU	Cytochrome c—Human	−160	10	26	321
MCHU	vs	MCHU	Calmodulin—Human	671	671	671	0
		TPHUCS	Troponin C, skeletal muscle	395	430	438	161
		PVPK2	Parvalbumin beta—Pike	−57	103	115	313
		CIHUH	Calpain heavy chain—Human	−2085	89	100	2463
		AQJFNV	Aequorin precursor—Jelly fish	−65	48	76	391
		KLSWM	Calcium binding protein—Scallop	−89	45	52	323
QRHULD	vs	EGMSMG	Epidermal growth factor precursor	−591	475	655	2549

Figure 10: Global and local alignment paths

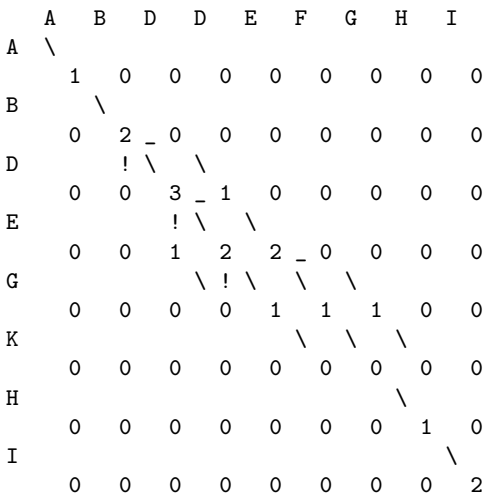
A. Global



Optimal global alignments (score 2):

```
    A B D D E F G H I  (top)
    A B D - E G K H I  (side)
or  A B - D E G K H I
```

B. Local



Optimal local alignment (score 3):

```
    A B D  (top)
    A B D  (side)
```

2.2 Dynamic Programming Algorithms

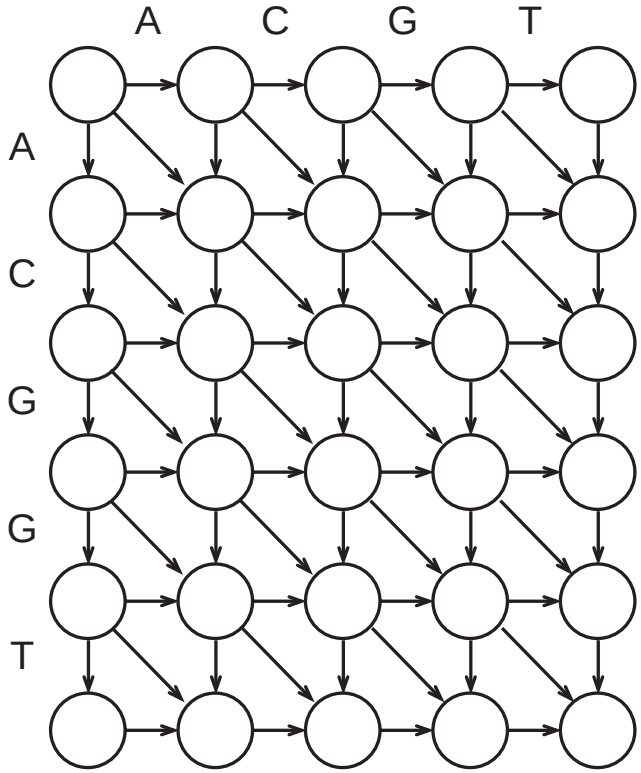
The algorithms used to calculate the maximum similarity scores between two sequences are most easily visualized with an alignment matrix or path graph. Figs. 10–11 demonstrate the correspondence between an alignment path graph and an actual alignment. The goal along the path is to maximize the similarity score for the alignment that ends at each potential vertex. For the figures, similarity scores are increased by $+1$ for diagonal edges if the two residues along the path are identical; if they are different, the diagonal edge cost is -1. The cost along either a horizontal or vertical edge, which corresponds to an insertion in the top sequence (horizonal edge) or an insertion in the left-side sequence (vertical edge) is -2. To produce a global alignment from a path graph, simply begin at the bottom-right corner of the graph and follow the "active" paths, noted by $\backslash$, _ or ! to the upper-left corner, aligning the two residues along the diagonal path, or aligning a residue with a gap if a horizontal or vertical path is taken.

For the global alignment in Fig. 10A, there are two alignments that produce the optimal score. Optimal comparison algorithms guarantee to produce the best score, given the match, mismatch, and gap costs, but frequently there are several optimal alignments for a single score. For the local alignment in Fig. 10B, there are several sub-optimal alignments with scores of 2. Note that the local alignment in Fig. 10B would extend from one end of each sequence to the other if the gap cost were reduced to -1.

Fig. 11 provides an exercise for the reader.

While there are an exponential number of potential alignments with gaps between two protein or

Figure 11: An alignment path matrix



DNA sequences, dynamic programming algorithms are available that can calculate the optimal score in $O(MN)$ steps. This efficiency is achieved by determining the optimal score for each prefix of each string, and then extending each prefix by considering the three paths that can be used to extend an alignment: (1) by extending the alignment by one residue in each sequence; (2) by extending the alignment by one residue in the first sequence and aligning it with a gap in the second; or (3) extending the alignment by one residue in the second sequence and aligning it with a gap in the first. This decision must be made for each of the MN prefixes of sequences of length M and N.

The first algorithm for comparing protein sequences (Needleman & Wunsch, 1970) calculates a "global" similarity score. A simplified global algorithm is shown in Fig. 12. Since a global algorithm requires that the alignment extend from the beginning to the end of the alignment, it is sufficient to report the score in the lower right ($S(M,N)$) of the scoring matrix.

Local alignment algorithms must consider alignments that begin and end at each of the MN positions in the alignment matrix. Despite this added complexity, they only add two additional steps to the global alignment algorithm. Since every possible starting position must be considered, similarity scores cannot fall below zero and a 0 term is added to the *max* comparison in Fig. 12. Since they can end at any position in the matrix, the *best* score must be saved at each step. In practice, global and local comparison algorithms require the same amount of computation.

Figure 12: Algorithms for Global and Local similarity scores

$S(0,0) \leftarrow 0$
for $j \leftarrow 1$ **to** N **do**
    $S(0,j) \leftarrow S(0,j-1) + \sigma(\ {}^{-}_{b_j}\ )$
for $i \leftarrow 1$ **to** M **do**
[$S(i,0) \leftarrow S(i-1,0) + \sigma(\ {}^{a_i}_{-}\ )$
    for $j \leftarrow 1$ **to** N **do**
        $S(i,j) \leftarrow max[S(i-1,j-1) + \sigma(\ {}^{a_i}_{b_j}\ ), S(i-1,j) + \sigma(\ {}^{a_i}_{-}\ ), S(i,j-1) + \sigma(\ {}^{-}_{b_j}\ )]$
]
write "Global similarity score is" $S(M,N)$

best $\leftarrow 0$
for $j \leftarrow 1$ **to** N **do**
    $S'(0,j) \leftarrow 0$
for $i \leftarrow 1$ **to** M **do**
[$S'(i,0) \leftarrow 0$
    for $j \leftarrow 1$ **to** N **do**
    [$S'(i,j) \leftarrow max[0, S'(i-1,j-1) + \sigma(\ {}^{a_i}_{b_j}\ ), S'(i-1,j) + \sigma(\ {}^{a_i}_{-}\ ), S'(i,j-1) + \sigma(\ {}^{-}_{b_j}\ )]$
        $best \leftarrow max(S'(i,j), best)$
]
write "Local similarity score is" *best*

2.3 Heuristic Algorithms

Two rapid heuristic algorithms are frequently used for searching protein and DNA sequence databases, FASTA (Pearson & Lipman, 1988) and BLASTP (Altschul *et al.*, 1990). These methods are 5–50 times faster than the rigorous Smith-Waterman algorithm, and can produce results of similar quality in many cases.

Fig. 13 summarizes the difference between the FASTA, BLASTP, and Smith-Waterman algorithms. BLASTP and FASTA are faster than Smith-Waterman because they examine only a portion of the potential alignments between two sequences. FASTA focuses on regions where there are either pairs (*ktup=2*) or single aligned *ktup=1* identities; BLASTP examines regions that include triples of conserved amino acids.

Figure 13: Heuristic strategies for sequence comparison

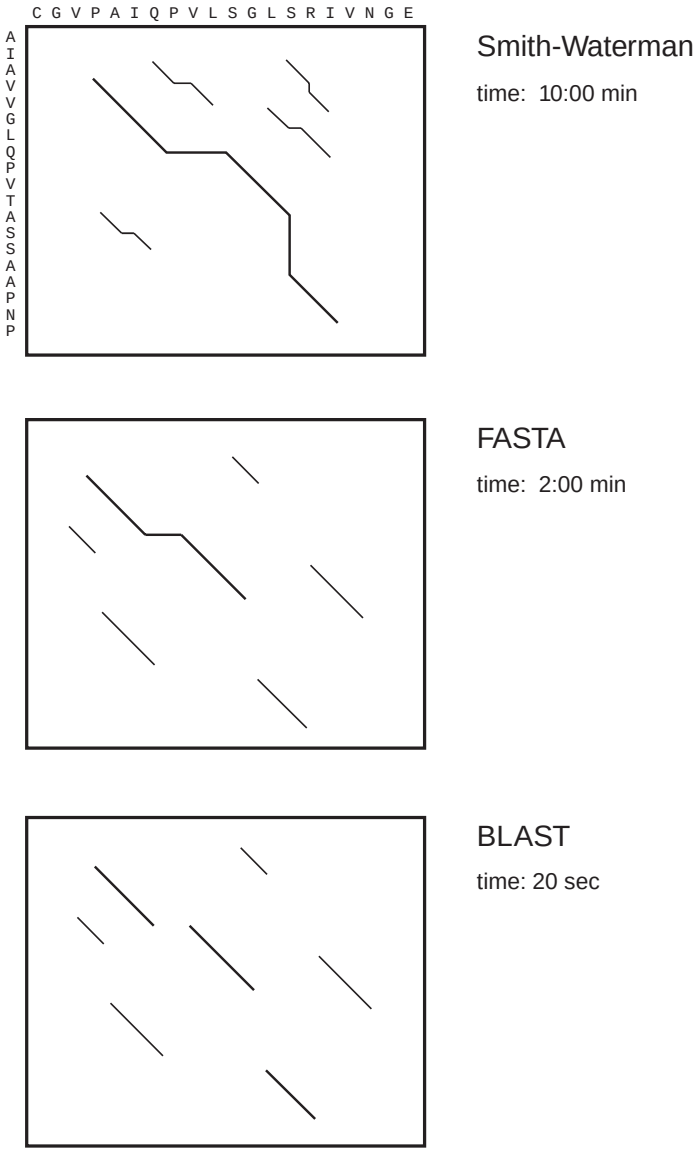


Table 11: Sequence similarity with BLASTP

Step 1 For each three amino acids in the query sequence, identify all of the substitutions of each word that have a similarity score greater than a threshold score $T =$. In practice, word-matches with scores $\geq T$ are seen on average 150 times per library sequence.

Step 2 Build a discrete finite automaton (DFA) to recognize the list of identical and substituted three letter words.

Step 3 Use the DFA to identify all of the matching words in sequences in the database. If two matches are found, extend each match both forwards and backwards using the BLOSUM62 matrix, allowing gaps, to produce a score that is higher than a threshold score. Save all of the high scoring regions shared by the query sequence and each library sequence.

Step 4 Report all of the significant alignments. Frequently, a query and library sequence will contain several MSPs because of the requirement that they do not contain gaps.

2.3.1 BLAST

Advances in the statistical theory of sequence alignments without gaps (Karlin & Altschul, 1990) provided the theoretical basis for the BLASTP program (Altschul *et al.*, 1990). BLASTP is now the most widely used program for rapid sequence comparison, in large part because of its accurate estimates for the statistical significance of similarity scores (see 3. BLASTP, like FASTA, uses a word-based scanning procedure to identify regions of local similarity (Table 11) with out gaps. BLASTP is effective because it combines high sensitivity with excellent selectivity. BLASTP combines good sensitivity with exceptional selectivity. Except when the query sequence contains a low complexity region, BLASTP rarely calculates scores for unrelated sequences.

2.3.2 FASTA

The current version of FASTA provides several significant improvements over earlier versions. FASTA now calculates optimized scores (step 4 in Table 12)) for most of the sequences in the database and provides accurate estimates for statistical significance (3). Calculation of optimized scores improves substantially the performance of FASTA. Without the calculation, FASTA performs significantly worse than BLASTP; however, with the calculation of optimized scores (and normalization of the scores based on library sequence length), FASTA performs significantly better than BLASTPv1.4 and almost as well as the Smith-Waterman algorithm (Pearson, 1995). In addition, FASTA now uses the Smith-Waterman algorithm to produce final alignments; previous versions limited the size of gaps, which sometimes led to incomplete alignments.

Every database search for members of a diverse protein family involves a tradeoff between sensitivity—the ability to identify distantly related members of the family—and selectivity—the ability to avoid high similarity scores for unrelated sequences. Table 3.3 compares how effectively the three algorithms maintain this balance for a large protein family—the G-protein-coupled receptors. Thus, BLASTP calculates a very highly significant score for the closely related opsin and dopamine D2 receptors, and a significant score for the more distantly related thromboxane A_2 receptor, but it does not detect the similarity between

Table 12: Sequence similarity with FASTAv33

Step 1 Identify regions shared by the two sequences with the highest density of identities (*ktup=1*) or pairs of identities (*ktup=2*). If the `-S` option is used, low-complexity (lower-case) letters are ignored in steps 1–5.

Step 2 Rescan the ten regions with the highest density of identities using the BLOSUM50 matrix. Trim the ends of the region to include only those residues contributing to the highest score. Each region is a partial alignment without gaps.

Step 3 If there are several initial regions with scores greater than the CUTOFF value, check to see whether the trimmed initial regions can be joined to form an approximate alignment with gaps. Calculate a similarity score that is the sum of the joined initial regions minus a penalty (usually 20) for each gap (*initn*). The score of the single best initial region found in Step 2 is also reported (*init1*).

Step 4 For sequences with scores greater than a threshold, construct an optimal local alignment of the query sequence and the library sequence, considering only those residues that lie in a band centered on the best initial region found in Step 2. For protein searches with *ktup=2* a 16 residue band is used by default. A 32 residue band is used with *ktup=1*. This is the optimized (*opt*) score.

Step 5 (Statistical estimates) After the first 60,000 scores have been calculated, normalize the raw similarity scores using estimates for the statistical parameters of the extreme value distribution. The default strategy regresses the similarity score against ln(library-sequence length) and calculating the average variance in similarity scores. *Z-values* (normalized scores with mean 0 and variance 1) are calculated, and the calculation is repeated with library sequences with *z-values* greater than 5.0 and less than -5.0 removed. These z-values are used to rank the library sequences, but other estimation strategies are available.

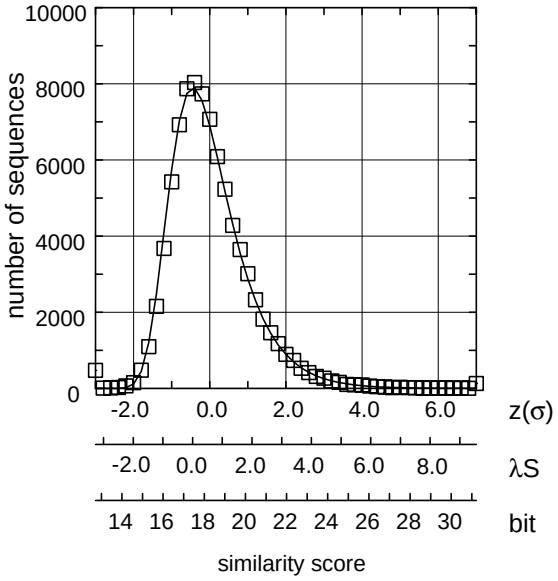
Step 6 The Smith-Waterman algorithm (without limitation on gap size) is used to display alignments. Final alignments score low-complexity regions.

opsin and the very distantly related *Dictyostelium* cAMP (CAR1) receptor. In addition, BLASTP would never suggest a relationship between opsin and cytochrome oxidase. FASTA (*ktup=2* does a better job at recognizing the relationship between opsin and thromboxane A2, fails to detect the cAMP-1 receptor, and is more ambiguous about a possible relationship with cytochrome oxidase. FASTA with *ktup=1* and Smith-Waterman calculate statistically significant relationships between opsin and cAMP-1, but also good (but not significant) scores for opsin and cytochrome oxidase.

3 The statistics of sequence similarity scores

The development of accurate statistical estimates for local sequence similarity scores (Karlin & Altschul, 1990; Mott, 1992) has allowed dramatic improvement in our ability to reliably recognize distantly related proteins. The statistical estimates calculated by BLASTP are used widely in large scale sequence comparison, e.g. to characterize all of the genes on a yeast chromosome or all of the genes in a bacterial genome. The incorporation of statistical estimates into FASTA and SSEARCH (a Smith-Waterman implementation) have significantly improved the performance of these programs as well.

Figure 14: The extreme value distribution



3.1 Sequence alignments without gaps

The statistics of local similarity scores for alignments without gaps but with an arbitrary substitution matrix have been described by Karlin & Altschul, 1990. Local similarity scores are described by the *extreme value* distribution. Using the parameters λ and K, which can be derived from the scoring matrix and the amino acid composition of the query sequence, the probability that a normalized similarity score:

$$S' = \lambda S - \ln Kmn \tag{1}$$

(Karlin & Altschul, 1990; Altschul *et al.*, 1994) where m is the length of the query sequence and n is the length of the library sequence can be calculated as:

$$P(S' \geq x) = 1 - exp(-e^{-x}) \tag{2}$$

or

$$P(S \geq x) = 1 - exp(-Kmne^{-\lambda x}) \tag{3}$$

Since a typical database search typically involves thousands of pairwise comparisons, the expectation of finding a score $S' \geq X$ for a search of D sequences is: $E(S' \geq X) = PD$. (Thus, searches of highly redundant databases may be less informative, because D is larger but the number of different sequences is not.)

3.2 Scoring matrices

The scoring matrices used for protein sequence comparison are much more sophisticated than $+1$ for a match and -1 for a mismatch. The most effective matrices are based on the actual frequency of

substitutions that occur between related proteins. Similarity scoring matrices can differ in three ways: (1) the method by which they are constructed; (2) their information content, which is related to the number of residues that must be aligned to produce a statistically significant score, and (3) their scale - the amount of information provided per unit score.

Two general approaches have been used to produce scoring matrices. The original PAM250 matrix (Fig. 5) was produced by examining several hundred alignments between very closely related proteins, and then calculating the frequency with which each amino-acid residue changed into each of the others at a very short evolutionary distance—one where only 1% of the residues had kchanged (Dayhoff *et al.*, 1978). This replacement frequency, when corrected for the amino-acid abundance, can be used to calculate the PAM1 scoring matrix (PAM is "Point Accepted Mutation"). If the matrix is multiplied against itself 250 times, a PAM250 matrix, which reflects the frequency of change for proteins that have diverged 250%. If two protein sequences have diverged by 250%, it is expected that they will share about 20% sequence identity (Fig. 9). Since 20% identity is at the edge of where significant similarity can be detected, the PAM250 matrix has been widely used. The PAM250 matrix is based on small number of amino acid substitutions; modern extrapolated matrices based both on sequence alignments (Jones *et al.*, 1992) and structural alignments (Johnson & Overington, 1993) are available.

Alternatively, substitution matrices can calculated directly by examining "blocks" of aligned sequences that differ by no more than X% (Henikoff & Henikoff, 1992). Thus, the BLOSUM62 matrix, which is used by the BLASTP rapid comparison program, is derived from substitution data for blocks of aligned sequences that are no more than 62% identical. BLOSUM62 performs substantially better than extrapolated matrices with BLASTP and FASTA (Henikoff & Henikoff, 1993), but both BLOSUM and extrapolated matrices can perform well when used with optimal gap penalties (Pearson, 1995).

PAM-style matrices and BLOSUM matrices differ not only in the way they they are built, but also in the way they are used. "Shallow" PAM matrices (`PAM20`, `PAM20`, `PAM40`) have low numbers, and indicate that very little evolutionary change has taken place. "Conservative" BLOSUM matrices (e.g. `BLOSUM80`) have high numbers, and indicate a high degree of sequence conservation, in contrast to a small amount of evolutionary change.

Altschul (1991) has provided a information-theory based perspective for evaluating scoring matrices in general for alignments without gaps (Altschul, 1991). With this interpretation, one can think of each entry s_{ij} in a scoring matrix as a "log-odds" value $s_{ij} = \log(\frac{q_{ij}}{p_i p_j})$, where q_{ij} is the "target" residue substitution frequency—the frequency with which residue `A` is replaced by `M`, or vice-versa, after a certain amount of evolutionary change—and p_i, p_j are the probabilities that the residues would align by chance, based solely on their frequency in a sequence. For a given pairwise sequence comparison, p_i, p_j will remain the same for scoring matrices at different evolutionary distances, but q_{ij} will vary, depending on how much change is expected. Thus, in Fig. 15, the `PAM40` matrix gives much larger positive scores for identical matches, lower negative scores for non-conservative replacements, and gives some replacements negative scores that are scored as conservative replacements when `PAM250` is used. In addition, one can calculate the the average score, information content, or relative entropy H, that is expected per aligned residue: $H = \lambda \sum_{ij} q_{ij} s_{ij} = \sum_{ij} q_{ij} \log_2 \frac{q_{ij}}{p_i p_j}$ if λs_{ij} is scaled in bits. For the `PAM40` and `PAM250` matrices in Fig 15, $H = 2.26$ and $H = 0.354$ bits/residue. If 40 bits of information are required to obtain an expectation value $E(80,000) < 0.001$, then an alignment of only $40/2.26 = 18$ residues would be required on average for `PAM40`, while 113 residues would be required with `PAM250`. However, if the two sequences last shared a common ancestor 600 Mya, and thus the sequences are separated by 1200 My,

Figure 15: Where similarity scoring matrices come from

A. PAM40

```
   A   R   N   D   E   I   L
A  8
R -9  12
N -4  -7  11
D -4 -13   3  11
E -3 -11  -2   4  11
I -6  -7  -7 -10  -7  12
L -8 -11  -9 -16 -12  -1  10
```

B. PAM250

```
    A   R   N   D   E   I   L
A   2
R  -2   6
N   0   0   2
D   0  -1   2   4
E   0  -1   1   3   4
I  -1  -2  -2  -2  -2   5
L  -2  -3  -3  -4  -3   2   6
```

q_{ij} : replacement frequency at PAM40, 250

$$
\begin{array}{rcl}
q_{R:N}(40) = 0.000435 & & p_R = 0.051 \\
q_{R:N}(250) = 0.002193 & & p_N = 0.043 \\
\lambda_2 S_{ij} = \log_2(q_{ij}/p_i p_j) & \lambda_e S_{ij} = \ln(q_{ij}/p_i p_j) & p_R p_N = 0.002193 \\
\lambda_2 S_{R:N}(40) = & \log_2(0.000435/0.00219) = & -2.333 \\
\lambda_{b/3} = 1/3 & S_{R:N}(40) = -2.333/\lambda_{b/3} = & -7 \\
\lambda S_{R:N}(250) = & \log_2(0.002193/0.002193) = & 0
\end{array}
$$

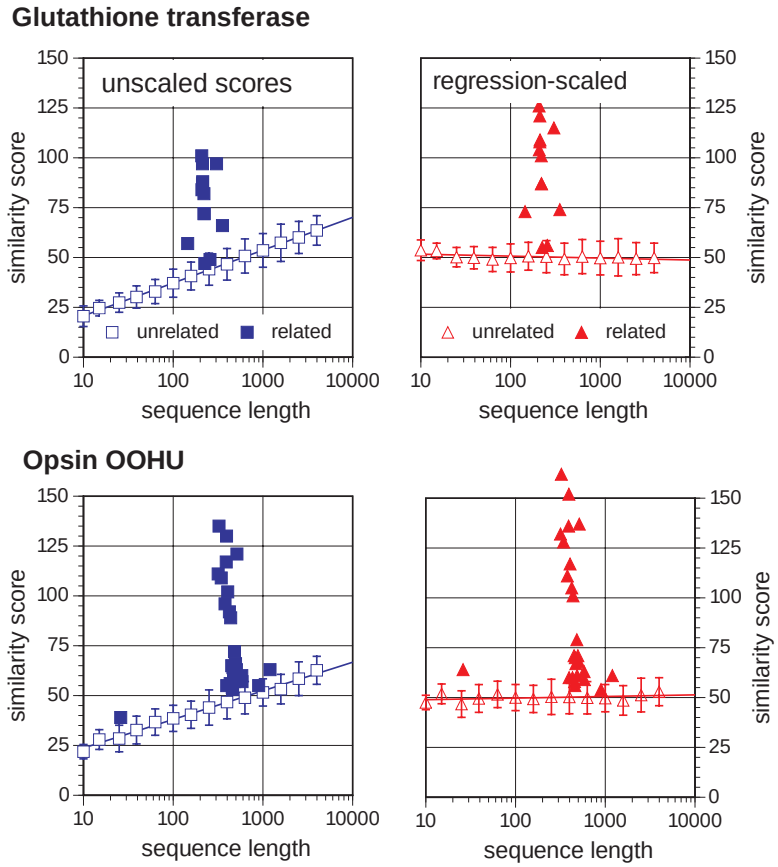
The PAM40 and PAM250 similarity scoring matrices are shown for 6 amino-acid residues. The substitution matrices are symmetric. Diagonal elements are the scores given to amino-acid identities; off-diagonal elements are the scores used for amino-acid substitutions. Both the PAM40 and PAM250 matrices are scaled to 0.33 bits per unit raw score. Thus, if $\log_2 \frac{q_{ij}}{p_i p_j} = 2$, the entry in the matrix would be 6.

and changed at an "average" rate of 15 PAMs/100 My, their significant similarity would not be detectable with PAM40, but could easily be seen, if the similar domain were longer than 120 residues, with PAM250 (the alignment would be even more significant, however, with PAM180, which should be optimal for the evolutionary distance).

If scoring matrix entries s_{ij} are thought of as "log-odds" ratios, we still do not know what the "odds" are unless we know the base for the logarithm. Thus, Dayhoff (1978) used $\log_{10}$, then multiplied each of those values by 10 and rounded to integers, with the result that $10 \log_{10} \frac{q_{ij}}{p_i p_j} = \log_2 \frac{q_{ij}}{p_i p_j}/3$; thus PAM250 values are sometimes described as scaled to $1/3$ bit units.[3] Indeed, scaling s_{ij} by 10, or 20, or even 100 will not change the length of the alignment, its relative ranking in a database search, or its statistical significance as long as all scores are calculated with the same s_{ij} (and the gap penalties are scaled appropriately as well). Thus, a raw similarity score $S = 100$ tells us nothing about the significance of a match;

[3] Unfortunately, BLOSUM62 s_{ij} are scaled in $1/2$-bit units, but BLOSUM50 is scaled in $1/3$-bit units. As a result, a gap-penalty of -12 for the first residue in a gap and -2 for each additional residue, which is commonly used for $1/3$-bit matrices, is too high for $frac$12-bit matrices; which should use $-8,-1$ or $-8,-2$.

Figure 16: Similarity scores and library sequence length



The distribution of Smith-Waterman similarity scores is plotted as a function of $\log(n)$, n is the length of the library sequence. Filled symbols indicate individual related sequences (only the most distant related sequences are shown); open symbols show the average and std. error of similarity scores for unrelated sequences.

one must also know the scoring matrix and its "scale", or, more specifically, its λ, or describe the score in terms of the number of standard deviations above the mean for sequences of that size. Current versions of the `blast2` and `fasta3` programs report scores in terms of `bits`, where$S_{bit} = (\lambda S_{raw} - \ln K)/\ln 2$. Thus, substituting in equation 3:

$$E(S_{bit}) = mn2^{-S_{bit}} = \frac{mn}{2^{S_{bit}}} \tag{4}$$

For ungapped alignments, λ and K can be calculated analytically (Karlin & Altschul, 1990). For alignments with gaps, they must be estimated.

Table 13: Search Algorithms and Statistical Significance

algorithm	closely related	related	distantly related	unrelated
	dopamine D2[a]	thromboxane A2[b]	cAMP-1[c]	cytochrome oxidase[d]
Smith-Waterman	4×10^{-8}	0.00016	0.058	3.8
PRSS[e]	2.6×10^{-8}	0.0012	0.046	28
PRSS(window=20)[e]	6.0×10^{-6}	0.91	10	330
fasta, *ktup=1, opt*	2.6×10^{-7}	0.00026	0.067	3.3
fasta, *ktup=2, opt*	0.00012	0.00026	> 10	2.3
BLASTP2.0	3×10^{-10}	0.19	> 10.0	> 10.0

[a]D2DR_HUMAN, [b]TA2R_HUMAN, [c]CAR4_DICDI, [d]APPC_ECOLI

Expected number of times that a similarity score as high or higher than that obtained by the indicated library sequence would be obtained by chance in a search of Swiss-Prot ($\approx 100,000$ entries) with the OPSD_HUMAN (human opsin) query sequence. [e]Expected times this score would be obtained after $100,000$ shuffles of the indicated library sequence with either global (prss) or local (window=20) amino acid exchanges.

3.3 Empirical statistics for alignments with gaps

The normalization in equation 1 shows that scores for alignments without gaps between random sequences increase as $\ln Kmn$, or since K and m are fixed for a given search, $\ln n$, the length of the library sequence. This is seen empirically with scores for alignments that contain gaps (Collins *et al.*, 1988; Mott, 1992) and is shown in Fig. 16. For local similarities, the variance of the score should be independent of library sequence length. Thus, normalization of similarity scores by fitting a line to the relationship of similarity score to $\ln n$ will reduce the scores of long, unrelated sequences, and make it possible to detect more distant relationships (Pearson, 1995).

Accurate statistical estimates for alignments with gaps can can be calculated by normalizing similarity scores to remove the $\ln n$ dependence for similarity scores. This can be seen in Fig. 6, where the '$*$'s show the fit of an extreme value distribution to the observed data ('=='). FASTA and SSEARCH estimate statistical significance by fitting a line to S vs $\ln n$ and calculating the average variance for the scores. The regression line and variance are used to calculate

$$Z - score = (S - (a + b \ln n))/\sqrt{var} \tag{5}$$

The distribution of $Z - score$'s should follow the extreme value distribution, so that:

$$P(Z > x) = 1 - exp(-e^{-1.282Z - 0.5772}) \tag{6}$$

and, as before, $E(Z > x) = PD$.

3.4 Statistical significance by random shuffling

Statistical estimates derived from database searches measure the difference between an observed similarity score and that expected for a sequence with the amino acid composition of the database. Such tests may overestimate significance in cases where the query sequence's amino acid composition differs from that of the database. Thus, membrane proteins with their hydrophobic transmembrane domains may have statistically significant scores with non-homologous membrane proteins. A more challenging test compares the similarity score between a query and library sequence with the distribution of scores obtained by comparing the query sequence to random sequences with the same length and amino acid composition as the library sequence. Such sequences are easily generated by randomly shuffling the library sequence, either globally, by exchanging randomly each amino acid with any other position in the sequence, or locally, by performing the exchanges within a window of 10–20 residues. Because this Monte Carlo test measures the significance of the order of the two amino acid sequences, rather than the difference between the highest scoring sequences and the rest of the database, it tends to be more demanding.

As before, similarity scores for random sequences should follow the extreme value distribution, and a fit of the distribution of scores can be used to estimate the significance of an unshuffled score. However, to extrapolate an expectation value from shuffled sequences to that for a library search, the "E()-value" must be multiplied by the ratio of the number of sequences in the library to the number of shuffled sequences. Thus, in the example below, an E()-value from 500 shuffles must be multiplied by 80 to be comparable to an E()-value from the $40,000$ entry Swiss-Prot. As expected, the E()-value from the actual search—2×10^{-4}—is slightly more significant than that from the shuffled distribution—3×10^{-3}.

```
 Comparison of OOHU (human opsin) with TA2R_HUMAN (thromboxane A2 receptor)
 (shuffled) MLE statistics: Lambda= 0.1476;  K=0.01206
 Smith-Waterman (3.39 May 2001) function [BL50 matrix (15:-5)], open/ext: -10/-2
 PRSS34 - 1000 shuffles;  uniform shuffle
 unshuffled s-w score: 173;  bits(s=173|n_l=369): 43.2 p(173) < 1.25774e-08
For 1000 sequences, a score >= 173 is expected 1.258e-05 times
```

Although accurate statistical estimates can be very valuable in interpreting the results of similarity searches, they must be evaluated with caution. Distantly related homologous sequences often do not share statistically significant similarity. Thus, overreliance on statistical estimates, particularly after a single search, can miss genuine homologies. Conversely, sequences with low-complexity regions often share significant similarity but are not homologous. Finally, some structures, such as the coiled-coil structure in tropomyosin, share statistical significance because of a common repeated structure, because of convergence (analogy), rather than homology.

4 Identifying distantly related protein sequences

In this section, we will examine similarity searches in three diverse families of protein sequences, serine proteases, glutathione S-transferases, and the G-protein-coupled receptors. The serine proteases are considered because they provide a classic example of a family of proteins with a highly conserved active

Figure 17: Patterns for serine proteases

```
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;
```

Patterns from PROSITE that identify 152/163 (`TRYPSIN_HIS` or 143/159 `TRYPSIN_SER` members of the serine protease protein family.

site; the glutathione transferases are a very diverse family where many members do not share significant similarity with all other members, while the G-protein-coupled receptors are a very large and diverse family of membrane proteins.

4.1 Serine proteases

Serine proteases cleave peptide bonds using a "catalytic triad" of histidine, serine, and aspartic acid; these residues are underlined in Fig. 19. Because these residues are so highly conserved, patterns that focus on two of the regions (Fig. 17) can be used to identify every member of the serine protease family. Fig. 18 shows the highest scoring unnormalized similarity scores. As is often the case for divergent protein families, several members of the family do not share statistically significant similarity with bovine trypsin. These sequences are italicized in Fig. 18; their membership in the serine protease family is based on common three-dimensional structures. As expected from the discussion in section 16, several of the highest scoring unrelated sequences are substantially longer than genuine serine proteases. These scores have much higher (less significant) expectation values when the $\ln n$ correction is used.

The absolute conservation of residues in the "catalytic triad" might suggest that similarities between members of this family are limited to those regions. This is not the case, as can be seen in Fig. 19. Similarity in the serine proteases typically extends from one end of the protein to the other, with strong conservation throughout the sequence. Indeed, the region around one of the residues in the catalytic triad—the apartic acid—is not well conserved. While the residues in the catalytic triad is an essential feature of serine proteases, the serine protease fold (two domains containing anti-parallel β-barrels) are required to bring these residues together.

Figure 18: Serine protease search - high scoring sequences

LOCUS	Description	len	score	E(12,000)
TRBOTR	trypsin precursor - bovine	229	1559	10^{-97}
TRRT2	trypsin II precursor - rat	246	1240	10^{-76}
KQHU	tissue kallikrein precursor -	262	669	10^{-37}
NGMSG	7S NGF gamma chain I	237	645	10^{-36}
KQRTTN	tonin - rat	235	623	10^{-34}
KYBOA	chymotrypsin A precursor - bovine	245	609	10^{-34}
PLHU	plasmin precursor - human	790	580	10^{-31}
TRFF	trypsin-like proteinase	256	579	10^{-31}
KFHU	coagulation factor IXa	461	578	10^{-31}
ELRT2	pancreatic elastase II	271	559	10^{-30}
KYBOB	chymotrypsin B precursor - bovine	245	556	10^{-30}
KFHU1	coagulation factor XIa	625	547	10^{-29}
WMMS28	complement factor D homolog	259	541	10^{-29}
EXBO	coagulation factor Xa	492	518	10^{-27}
DBHU	complement factor D	246	517	10^{-27}
KXBO	protein C (activated)	456	515	10^{-27}
UKHU	u-plasminogen activator precu	431	507	10^{-26}
TBHU	thrombin precursor - human (fr	615	472	10^{-24}
TRSMG	trypsin - Streptomyces griseus	221	409	10^{-20}
C1HURB	complement subcomponent C1r p	705	356	10^{-16}
HPHU1	haptoglobin-1 precursor - human	347	335	10^{-15}
TRPGAZ	azurocidin - pig	219	316	10^{-14}
HPRT	haptoglobin - rat (fragments)	297	289	10^{-12}
C2HU	complement C2 - human	752	198	10^{-6}
BBHU	complement factor B - human	739	169	0.00014
KXBOZ	protein Z - bovine	396	142	0.0041
TRYXB4	alpha-lytic proteinase	396	107	0.83
OKBY8W	probable protein kinase YCR008W	603	107	1.3
RRIHM2	RNA-directed RNA polymerase	4488	99	37
IJFFTM	cadherin-related tumor suppressor	5147	99	42
GNNYE7	genome polyprot. - enterovirus 70	2194	98	20
VGIHHC	E2 glycoprotein - coronavirus	1173	96	14
QRRBVD	VLDL receptor - rabbit	873	96	10
PRSMBG*	proteinase B - S. griseus	185	96	1.9
MMMSB2	laminin chain B2 precursor - mouse	1607	95	23
RERTK	renin precursor - rat	402	94	6.0
MMMSA	laminin chain A - mouse	3084	93	61
LNRZ	lectin precursor - rice	227	90	6.0
PRSMAG*	proteinase A - S. griseus	182	89	5.5

Figure 19: Alignment of serine proteases

```
TRSMG trypsin (EC 3.4.21.4) precursor - Streptomyces griseus (259 aa)
Smith-Waterman score: 385;    33.6% identity in 247 aa overlap

                                    10          20        30        40
KYBOA                      CGVPAIQPVLSGLSR--IVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINE
                             : ..::: .. .   .:.: .:. : .:..: :.   :   :::.:  .
TRSMG  MKHFLRALKRCSVAVATVAIAVVGLQPVTASAAPNPVVGGTRAAQGEFPFMVRLS--MG---CGGALYAQ
               10        20        30        40        50             60

      50            60        70        80        90       100       110
KYBOA  NWVVTAAHC----GVTTSDVVVAGEFDQGSSSEKIQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFS
       . :.::::::    : .:: ....:  :  :::    :.. .::..   ::.   ..: .:.::  :  ..
TRSMG  DIVLTAAHCVSGSGNNTSITATGGVVDLQSSSA--VKVRSTKVLQAPGYNGT--GKDWALIKL--AQPIN
          70        80        90         100       110         120

         120       130       140       150       160       170       180
KYBOA  QTVSAVCLPSASDDFAAGTTCVTTGWGLTRYTNANTPDRLQQASLPLLSNTNCKKYWGTK-IKDAMICAG
       : .  .   .:   .  ::  :. ::: .:  ...    : .:..:..:.. :.. .:.. . .  ::::
TRSMG  QPTLKIATTTA---YNQGTFTVA-GWGANR-EGGSQQRYLLKANVPFVSDAACRSAYGNELVANEEICAG
     130       140           150        160       170       180       190

             190       200        210       220       230       240
KYBOA  ---ASGVSSCMGDSGGPLVCKKNG-AWTLVGIVSWGSSTCSTSTPGVYARVTALVNWVQQTLAAN
          ..::..:.:::::::.  : :.  :  ::::::: .    . ::::..:..... .
TRSMG  YPDTGGVDTCQGDSGGPMFRKDNADEWIQVGIVSWGYGCARPGYPGVYTEVSTFASAIASAARTL
          200       210       220       230       240       250
```

The requirement for a common folded structure in homologous proteins usually causes similarities to extend from one end of the protein to the other, or for mosaic proteins, from one end of a domain to the other. Fig. 20 displays the locally similar regions for the related and unrelated in Table 18; the highest scoring unrelated sequences tend to have relatively short (< 100 residue) regions of higher similarity ($\approx$ 30% identical) while related sequences have longer (140–400), though sometimes lower (25%) similarity. In general, shorter, higher similarities are less significant than longer, lower similarities.

Figure 20: Serine protease alignments

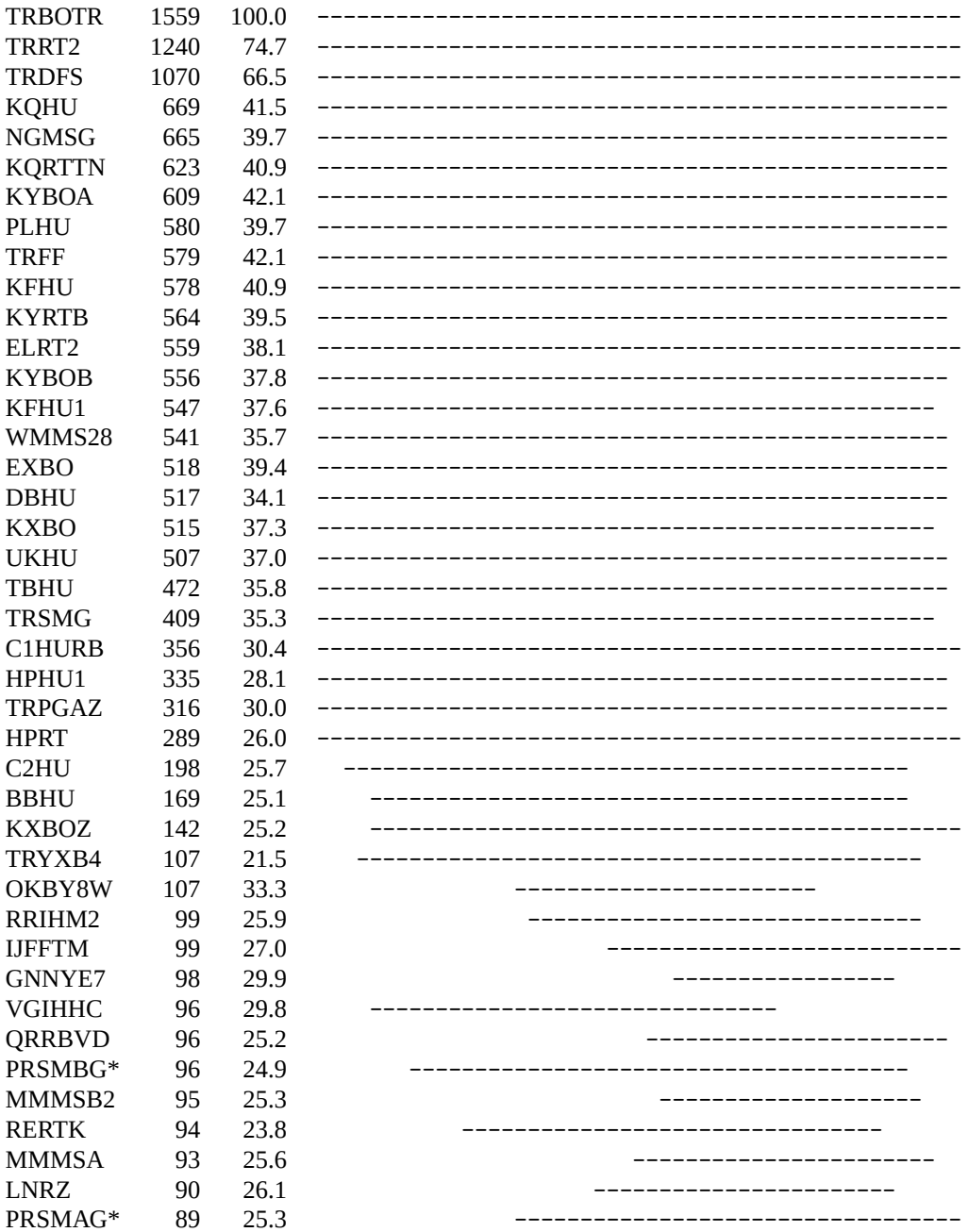


Table 14: Glutathione S-transferases

The best scores are:		s-w	bits	E(96,308)
GTM1_MOUSE	Glutathione S-transferase GT8.7	1490	366	10^{-101}
GTM1_RAT	Glutathione S-transferase YB1	1406	346	10^{-94}
GTM1_HUMAN	Glutathione S-transferase	1235	305	10^{-82}
GTM2_CHICK	Glutathione S-transferase 2	954	237	10^{-62}
GTP1_MOUSE	Glutathione S-transferase P	361	93	10^{-18}
GTA1_MOUSE	Glutathione S-transferase GT41A	218	61	10^{-9}
SC2_OCTDO	S-crystallin 2 (OL2).	224	61	10^{-9}
GTA2_MOUSE	Glutathione S-transferase Ya	229	59	10^{-8}
GTC_MOUSE	Glutathione S-transferase Yc	215	58	10^{-8}
GTA1_HUMAN	Glutathione S-transferase A1-1	206	56	10^{-7}
GT28_SCHHA	Glutathione S-transferase 28 kd	203	55	10^{-7}
GTA4_MOUSE	Glutathione S-transferase GST 5.7	183	50	10^{-5}
GT28_SCHJA	Glutathione S-transferase 28 kd	169	47	10^{-4}
GTS2_DROME	Glutathione S-transferase 2	164	46	10^{-4}
SC1_OCTVU	S-crystallin 1.	159	45	10^{-4}
GTA2_CHICK	Glutathione S-transferase, CL-3.	144	43	0.00054
SC18_OMMSL	S-crystallin SL18.	131	39	0.016
GTT1_MUSDO	Glutathione S-transferase 1	122		0.055
GTH1_MAIZE	Glutathione S-transferase I	120		0.056
GTXA_TOBAC	Auxin-regulated protein	117	34	0.220
GT32_MAIZE	Glutathione S-transferase III	115	34	0.31
GTT1_DROME	Glutathione S-transferase 1-1	100	30	3.5
GTH1_WHEAT	Glutathione S-transferase 1	98	30	5.4
GT_PROMI	Glutathione S-transferase GST-6.0	97	30	5.5
VP2_AHSV3	Outer capsid protein (1057)	108	32	6.8
DCMA_METSP	Dichloromethane dehalogenase	98	30	7.2
SLT_HAEIN	putative transglycosylase (593)	103	31	7.6
GTY2_ISSOR	Glutathione S-transferase Y-2	94	29	8.4
MOD5_YEAST	tRNA isopentenyltransferase	100	30	8.9
GTX2_TOBAC	Auxin-induced PGNT35/PCNT111.	93	29	12
GTT1_RAT	Glutathione S-transferase 5	93	29	13
SPCB_HUMAN	Spectrin beta chain, erythrocyt (2137)	108	31	16
DAPF_YERPE	Diaminopimelate epimerase	90	28	17
LIGE_PSEPA	β-etherase	91	28	22
EF1G_HUMAN	Elongation factor 1γ	94	29	23

All of the unitalicized sequences are known to be members of the glutathione transferase family.

4.2 Glutathione S-transferases

The glutathione transferase family of enzymes is a very diverse family of proteins found, in various forms, in animals, plants, and prokaryotes. Fortunately, many of the members of this family have a common enzyme activity so that they can be recognized by name. Table 14 shows that for this family, there are many homologues that do not show significant similarity when the database is searched with a single query sequence.

Frequently, clear identification of a distant homology will require several database searches, with either different algorithms or additional query sequences. For example, in Table 14, one might wish to test the possibility that glutathione S-transferases share homology with elongation factors, which are among the high scoring sequences. The result of a search using `EF1G_HUMAN` is shown in Table 15. Here, there is a clear relationship between this elongation factor and the class-theta glutathione transferases. An additional search with a class-theta sequence reveals the most distant relationships in this family more clearly.

Table 15: Distant glutathione transferase homologs

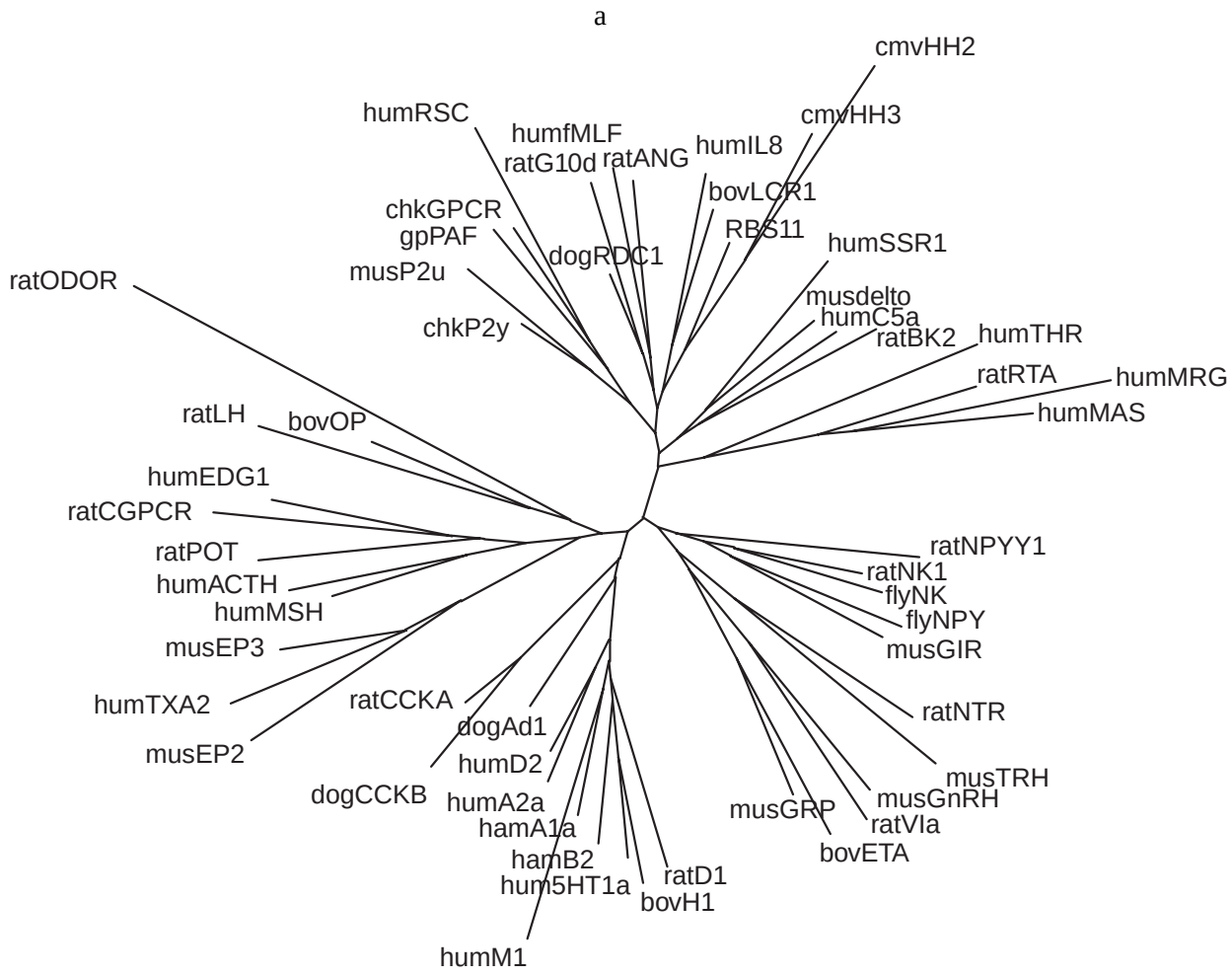
Re-search with LIGE_PSEPA

The best scores are:		s-w	bits	E(100,225)
LIGE_PSEPA	β-etherase	1993	477	10^{-134}
GTZ1_DIACA	Glutathione S-transferase 1	170	46	10^{-4}
GTX6_SOYBN	Probable glutathione S-transferase	168	46	10^{-4}
GTX3_TOBAC	Probable glutathione S-transferase	165	45	10^{-4}
GTXA_ARATH	Glutathione S-transferase	161	44	0.00056
GTX2_TOBAC	Probable glutathione S-transferase	157	43	0.00031
GTX1_SOLTU	Probable glutathione S-transferase	149	41	0.002
GTX1_TOBAC	Probable glutathione S-transferase	147	41	0.0028

Re-search with EF1G_HUMAN

The best scores are:		s-w	bits	E(100,225)
EF1G_HUMAN	Elongation factor 1γ (EF-1γ)	2977	634	10^{-181}
EF1G_XENLA	Elongation factor 1γ (EF-1γ)	2370	506	10^{-143}
EF1H_YEAST	Elongation factor 1γ 2 (EF-1γ)	769	169	10^{-41}
EF1G_TRYCR	Elongation factor 1γ (EF-1γ)	715	158	10^{-37}
SYV_HUMAN	valyl-tRNA synthetase	440	101	10^{-20}
GTH1_MAIZE	Glutathione S-transferase I	222	56	10^{-7}
GTH3_MAIZE	Glutathione S-transferase III	193	49	10^{-5}
GTH1_WHEAT	Glutathione S-transferase 1	186	47	10^{-4}
GTH1_TOBAC	Glutathione S-transferase	184	46	10^{-4}
GTY2_ISSOR	Glutathione S-transferase Y-2	175	45	0.00021
GTH2_WHEAT	Glutathione S-transferase 2	175	44	0.00062
GTX6_SOYBN	Probable glutathione S-trans.	171	43	0.00082
GTH2_TOBAC	Glutathione S-transferase	169	43	0.001
GTT1_DROME	Glutathione S-transferase 1-1	162	42	0.0028

Figure 21: G-protein-coupled receptors



4.3 G-protein-coupled receptors

The G-protein-coupled receptors (GCRs) are one of the largest known gene families; members of the family transduce signals from light, peptides, cationic amines, lipid mediators, odors, and many more small molecules. An evolutionary tree that summarizes the diversity of this family is shown in Fig. 21. Based on hydrophobicity plots and the structure of bacteriorhodopsin (a protein that does not share significant similarity with members of this family), the GCRs are thought to contain seven transmembrane domains, so that the N-terminus of the proteins is extracellular, while the C-terminus is intracellular.

Because GCRs have transmembrane domains, the highest scoring unrelated sequences are frequently other membrane proteins. Table 16 lists sequences from Swiss-Prot that have marginally significant matches with a human opsin sequence (there are more than 500 related sequences with expectations ranging from 0–0.01 that are not shown). As with most divergent families, the question becomes, “how do I know that XXX is/is not a GCR?” This is more difficult with the GCRs, because they have long

Table 16: GCRs distant from human opsin

The best scores are:		s-w	bits	E(100,225)
MC3R_RAT	melanocortin-3 receptor	140	38	0.034
MC3R_MOUSE	melanocortin-3 receptor	139	38	0.039
OLF6_CHICK	olfactory receptor-like protein	139	38	0.044
ML1A_XENLA	melatonin receptor type 1A	133	37	0.048
GU27_RAT	gustatory receptor GUST27	137	38	0.051
AG2T_RAT	type-1C angiotensin II receptor	132	36	0.064
OLF2_RAT	olfactory receptor-like protein F12	135	37	0.07
MAS_MOUSE	MAS proto-oncogene	133	37	0.097
PAFR_MACMU	platelet activating factor receptor	130	36	0.1
MAS_RAT	MAS proto-oncogene.	131	36	0.13
OLF2_CHICK	olfactory receptor-like protein C	129	36	0.17
CAR1_DICDI	cyclic AMP receptor 1	130	36	0.18
YS96_CAEEL	hypothetical 110.4 KD protein	133	36	0.26
5H2A_CAVPO	5-hydroxytryptamine 2A receptor (	121	34	0.46
PE24_RAT	prostaglandin E2 receptor EP4	124	35	0.55
CAR3_DICDI	cyclic AMP receptor 3	124	35	0.56
OLF4_CHICK	olfactory receptor-like protein c	121	34	0.58
ML1B_RAT	melatonin receptor type 1B	115	33	0.59
UL33_HSV7J	G-protein coupled receptor homolog U12	121	34	0.63
OLF5_CHICK	olfactory receptor-like protein C	120	34	0.67
MAS_HUMAN	MAS proto-oncogene.	120	34	0.7
NU2M_CHOCR	*NADH-ubiquinone oxidoreductase chain 2*	122	34	*0.76*
PE24_HUMAN	prostaglandin E2 receptor EP4	120	34	1
OLF1_CHICK	olfactory receptor-like protein C	117	33	1.1

variable length loops in both their extracellular and intracellular domains.

As before, two strategies can be used to evaluate the hypothesis of homology: re-searching the library and statistical significance from shuffling. A search of the Swiss-Prot database reveals that `MAS_HUMAN` shares significant similarity ($E(58,500) < 0.01$) with 205 GCRs; 100 additional scores with less statistical significance also belong to the GCR family before the first non-GCR is encountered. In contrast, the highest ranking scores from the `NU2M_CHOCR` are (more than 100 NADH oxidoreductase sequences are not shown):

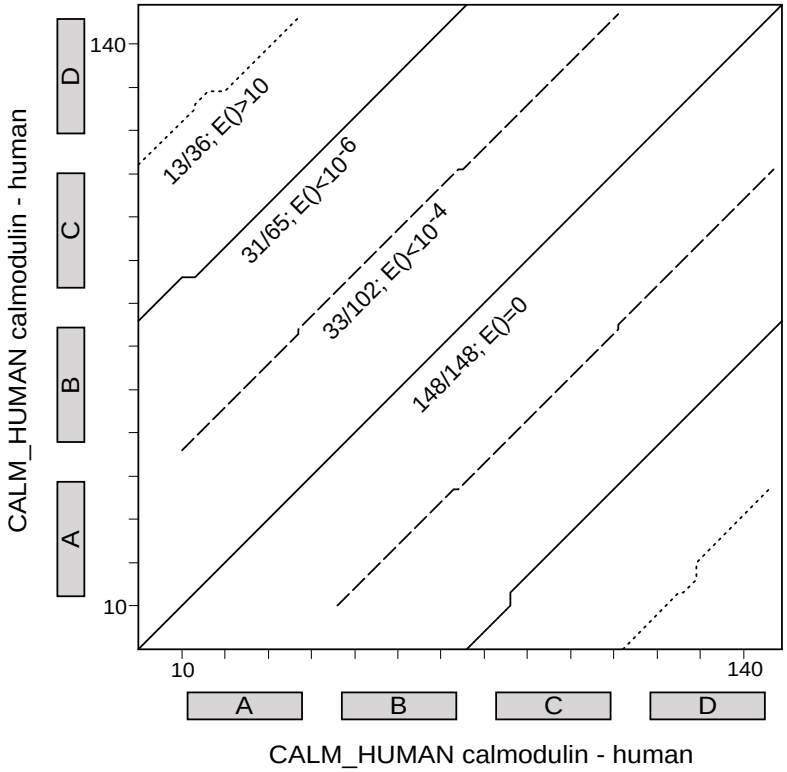
```
The best scores are:                                          s-w bits E(100,225)
NU2M_CHOCR NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 2  ( 497) 3181  531 3.3e-150
NUON_RHOCA NADH DEHYDROGENASE I CHAIN N (EC 1.6.5. ( 478)  928  162 3.1e-39
NU2C_MARPO NADH-PLASTOQUINONE OXIDOREDUCTASE CHAIN ( 501)  827  146   3e-34
NU2M_PODAN NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 2  ( 556)  788  140 2.8e-32
NU2M_ANOGA NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 2  ( 341)  460   86 2.7e-16
NU2M_RAT NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 2 (E ( 345)  393   75 5.4e-13
NU2M_CROLA NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 2  ( 348)  312   61 5.3e-09
NU5M_XENLA NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 5  ( 604)  230   48 9.1e-05
NDHF_BACSU NADH DEHYDROGENASE SUBUNIT 5 (EC 1.6.5. ( 505)  190   42  0.0074
COX1_PECMA CYTOCHROME C OXIDASE POLYPEPTIDE I      ( 549)  172   38   0.054
COX1_LEITA CYTOCHROME C OXIDASE POLYPEPTIDE I      ( 549)  154   36    0.47
CCMF_RHIME CYTOCHROME C-TYPE BIOGENESIS PROTEIN CY ( 676)  153   36    0.62
Y825_HAEIN HYPOTHETICAL PROTEIN HI0825.            ( 244)  145   34    0.67
RFBX_SALTY RFBX PROTEIN.                           ( 430)  148   35    0.76
ATP6_OENBE ATP SYNTHASE A CHAIN (EC 3.6.1.34)      ( 281)  141   33     1.2
YM04_PARTE HYPOTHETICAL 18.8 KD PROTEIN (ORF4).    ( 156)  135   32     1.5
YC43_ODOSI HYPOTHETICAL 30.1 KD PROTEIN YCF43 (ORF ( 263)  138   33     1.6
NU4M_ANOAR NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 4  ( 266)  135   32     1.9
COP_CLOPE COPY NUMBER PROTEIN (ORF4).              ( 198)  134   32       2
YJG2_YEAST HYPOTHETICAL 94.9 KD PROTEIN IN MRPL8-N ( 830)  143   34     2.3
CAPE_STAAU CAPE PROTEIN.                           ( 440)  138   33     2.4
OPSD_MOUSE RHODOPSIN.                              ( 348)  134   32     3.1
```

The results from the MAS_HUMAN and NU2M_CHOCR, which show that MAS_HUMAN is clearly a member of the GCR family, contrast with the statistical significance calculated with the PRSS program. Comparing the OOHU with RTA_RAT score with the distribution of scores calculated after shuffling RTA_RAT 1000 times with a local window of 20 suggests that the unshuffled score (109) is expected 6 times in 1000 shuffles. In contrast, the NU2M_CHOCR score is expected only 1.7 times in 1000 shuffles. From this perspective, the NU2M_CHOCR score is somewhat more significant, but, in fact, neither similarity score is statistically significant. It is not until MAS_HUMAN is compared with other members of the family, e.g. the angiotensin, fMet-Leu-Phe, thrombin, or substance-P receptors with E-values from 10^{-12}–10^{-6}, that the relationship is apparent.

Table 3.3 compares the statistical significance inferred from database searches with those determined by Monte-Carlo shuffling. As expected, the significance of the scores when compared with locally (window) shuffled sequences is 10-fold lower than the comparison with globally shuffled scores. It is unclear how to compare the expectation from shuffles with the expectation from a search. In the table, the expectation from a search of a $43,000$ entry library is compared to the expectation from $1,000$ shuffles. For global shuffles, the expectations are quite comparable while local shuffles are more conservative, yet all but one of the similarity scores judged significant from the database search are still significant when compared with the local-shuffle distribution.

Nevertheless, these examples show both that current statistical models for the similarity scores of unrelated sequences are quite accurate, but also that homologous sequences frequently do not share significant pair-wise similarity scores. Thus, a lack of statistical significance cannot be used to infer non-homology, but strong statistical significance is a good indicator of common ancestry.

Figure 22: Internal duplications in calmodulin



Comparison of human calmodulin with itself. Each diagonal line represents a potential local alignment of calmodulin with itself. Values below the diagonal lines show the number of identities and length of the aligned region (e.g. 33/102) and the expectation value for the similarity score of the alignment.

5 Repeated structures in proteins

So far, we have focussed on the identification and statistics of the single most significant similarity score shared by two sequences. As can be seen in Fig. 10B, however, there are frequently several non-overlapping local alignments with optimal similarity scores. In addition, there can be non-overlapping sub-optimal alignments with significant scores that can be used to infer the duplication events that gave rise to the protein sequence. An algorithm for the best *N* non-overlapping local alignments was described by (Waterman & Eggert, 1987).

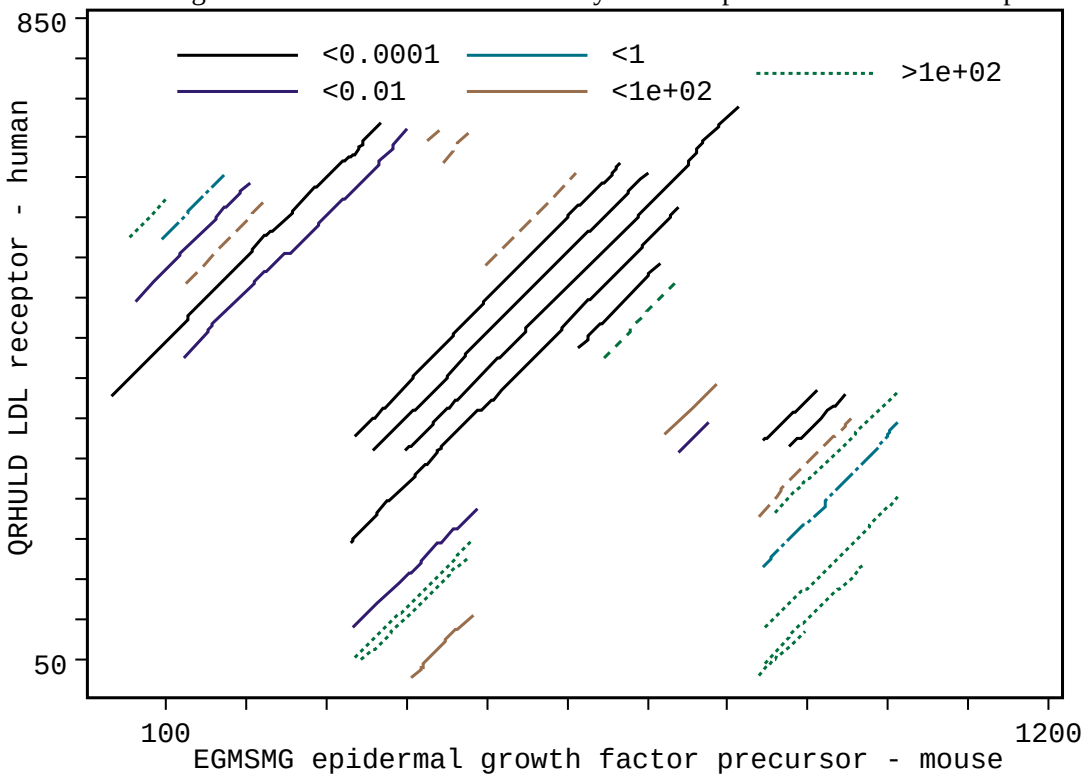
Figs. 22 and 23 show a graphical plot of the local similarities within the calmodulin calcium binding protein. Calmodulin contains four EF-hand Ca^{+}-binding domains that are well conserved. The highest scoring alignment in Fig. 23 aligns domains `A-B` with `C-D`; the second highest aligns `A-B-C` with `B-C-D`; the third aligns `A` with `D`.

A similar pattern of local similarity can be seen in Fig. 24, which shows the mosaic relationship between the EGF-precursor and the LDL-receptor.

Figure 23: Calmodulin internal alignments

```
 Comparison of:
(A) MCHU - Calmodulin - Human, rabbit, bovine, rat,   - 148 aa
(B) MCHU - Calmodulin - Human, rabbit, bovine, rat,   - 148 aa
 using matrix file: BLOSUM50, gap penalties: -14/-4

  47.7% identity in 65 aa overlap; score:  214 E(10,000): 3.4e-13

               20        30        40        50        60        70
MCHU   EFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARK
       :..::: .::::::.: :.. ::  :: .::.. :. :...:: :.: ::.: ... ::. ::. :
MCHU   EIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK
            90       100       110       120       130       140
----------
  32.4% identity in 102 aa overlap; score:  177 E(10,000):  1e-09

      10        20        30         40        50        60        70
MCHU   AEFKEAFSLFDKDGDGTITTKELGTVM-RSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMKD
       ::... ..  : ::.:::   :. :.: :.. .. .: :... .   : :::: :.  :.  .:. .. .
MCHU   AELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMT-NLGE
          50        60        70        80        90       100       110

       80        90       100       110
MCHU   TDSEEEIREAFRVFDKDGNGYISAAELRHVMT
         ..::. : .:  : ::.: ..  :. ..::
MCHU   KLTDEEVDEMIREADIDGDGQVNYEEFVQMMT
          120       130       140
----------
  34.2% identity in 38 aa overlap; score:   58 E(10,000):     39

              10         20        30
MCHU   MADQLTEEQIAEF-KEAFSLFDKDGDGTITTKELGTVM
       ...::.:.. :. .::    : :::: .. .:.  .:
MCHU   LGEKLTDEEVDEMIREA----DIDGDGQVNYEEFVQMM
            120           130       140

----------
  40.0% identity in 20 aa overlap; score:   53 E(10,000): 1.1e+02

       70        80
MCHU   LTMMARKMKDTDSEEEIREA
       .: ...:. : . .: ::::
MCHU   MTNLGEKLTDEEVDEMIREA
      110       120
----------
```

Figure 24: Mosaic domains shared by the EGF-precursor and LDL-receptor



Some non-homologous structures, particularly proteins containing the coiled-coil structure, have a periodic structure that is easily seen in local similarity plots. Fig. 25 shows local similarities in tropomyosin. All the alignments shown have local similarity scores greater than 120, and each would be significant in a conventional database search.

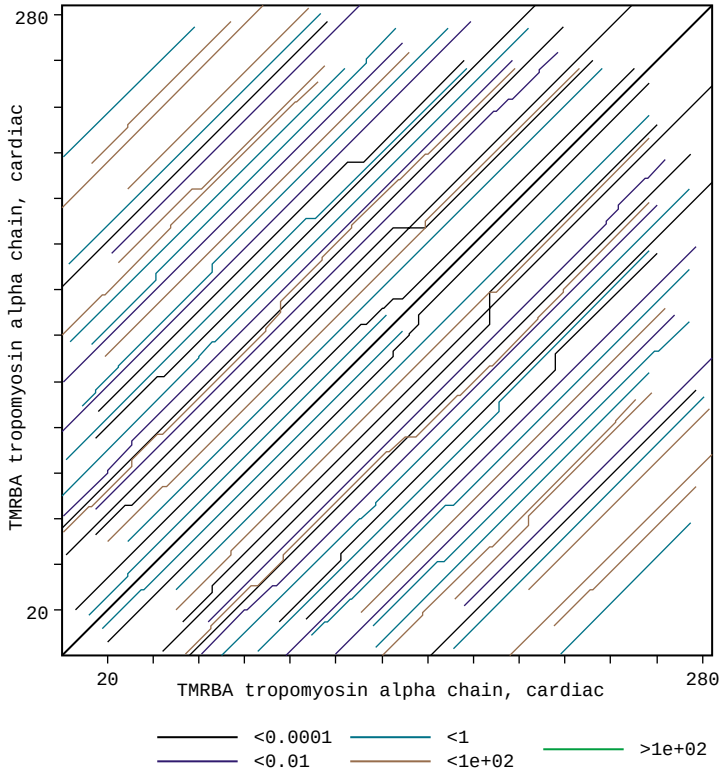
6 Summary

Protein sequence comparison is the most powerful tool available today for inferring structure and function from sequence because of the constraints of protein evolution—a protein fold into a functional structure. 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 five years with the development of accurate statistical estimates, which have provided better normalization methods, and with the use of optimized scoring parameters. In using sequence similarity to infer homology, one should remember:

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. While most sequences that share statistically significant similarity are homologous, many distantly related homologous sequences do not share significant homology. (Low complexity regions can

Figure 25: Coiled-coil structures share local similarity



display significant similarity in the absence of homology). Homologous sequences are usually similar over an entire sequence or domain. Matches that are more than 50% identical in a 20–40 amino acid region occur frequently by chance.

3. Homologous sequences share a common ancestor, and thus a common protein fold. 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.

4. Similarity searching techniques can be improved either by increasing the ability of a method to recognize distantly related sequences—increased sensitivity—or by lowering scores for unrelated sequences—increased selectivity. Since there are generally 1000-times more unrelated than related sequences in a sequence database, improvements that reduce the scores of unrelated sequences can have dramatic effects. The most dramatic improvements in comparison methods recently have used this approach.

References

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219**, 555–565.

Altschul, S. F. (1993). A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.* **36**, 290–300.

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

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Collins, J. F., Coulson, A. F. W. & Lyall, A. (1988). The significance of protein sequence similarities. *Comp. Appl. Biosci.* **4**, 67–71.

Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, (Dayhoff, M., ed.), vol. 5, supplement 3, pp. 345–352. National Biomedical Research Foundation Silver Spring, MD.

Doolittle, R. F., Feng, D. F., Johnson, M. S. & McClure, M. A. (1986). Relationships of human protein sequences to those of other organisms. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 447–455.

Gilbert, W. & Glynias, M. (1993). On the ancient nature of introns. *Gene,* **135**, 137–144.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci. USA,* **89**, 10915–10919.

Henikoff, S. & Henikoff, J. G. (1993). Performance evalutation of amino-acid substitution matrices. *Proteins,* **17**, 49–61.

Johnson, M. S. & Overington, J. P. (1993). A structural basis for sequence comparisons. an evaluation of scoring methodologies. *J. Mol. Biol.* **233**, 716–738.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.* **8**, 275–282.

Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA,* **87**, 2264–2268.

Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. U S A,* **90**, 5873–7.

Mott, R. (1992). Maximum-likelihood estimation of the statistical distribution of smith-waterman local sequence similarity scores. *Bull. Math. Biol.* **54**, 59–75.

Needleman, S. & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48**, 444–453.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia Univ. Press, New York, NY.

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

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71–84.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA,* **85**, 2444–2448.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. & Doolittle, W. F. (1994). Testing the intron theory of genes: the evidence from protein structure. *Science,* **265**, 202–207.

Waterman, M. S. & Eggert, M. (1987). A new algorithm for best subsequences alignment with application to tRNA-rRNA comparisons. *J. Mol. Biol.* **197**, 723–728.

7 Suggested Reading

7.1 General Protein evolution

R. F. Doolittle, D. F. Feng, M. S. Johnson, and M. A. McClure. Relationships of human protein sequences to those of other organisms. *Cold Spring Harb. Symp. Quant. Biol.*, 51:447–455, 1986.

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

7.1.1 Introns Early/Late

W. Gilbert and M. Glynias. On the ancient nature of introns. *Gene*, 135:137–144, 1993.

A. Stoltzfus, D. F. Spencer, M. Zuker, J. M. Logsdon, and W. F. Doolittle. Testing the intron theory of genes: the evidence from protein structure. *Science*, 265:202–207, 1994.

7.2 Alignment methods

7.2.1 Algorithms

S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*, 48:444–453, 1970.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

W. R. Pearson and W. Miller. Dynamic programming algorithms for biological sequence comparison. In L. Brand and M. L. Johnson, editors, *Meth. Enz.*, volume 210, pages 575–601. Academic Press, San Diego, 1992.

D. J. Lipman and W. R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.

W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.

W. R. Pearson. Rapid and sensitive sequence comparison with FASTP and FASTA. In R. F. Doolittle, editor, *Meth. Enz.*, volume 183, pages 63–98. Academic Press, San Diego, 1990.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs *Nucleic Acids Res.*, 25:3389–3402, 1997.

W. R. Pearson. Empirical statistical estimates for sequence similarity searches *J. Mol. Biol.* 276:71–84, 1998.

7.2.2 Scoring methods

M. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, supplement 3, pages 345–352. National Biomedical Research Foundation, Silver Spring, MD, 1978.

S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, 219:555–565, 1991.

D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.*, 8:275–282, 1992.

S. Henikoff and J. G. Henikoff. Performance evalutation of amino-acid substitution matrices. *Proteins*, 17:49–61, 1993.

7.3 Evaluating matches - statistics of similarity scores

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

S. F. Altschul, and W. Gish. Local Alignment Statistics *Meth. Enzymol.*, 266:460–480, 1996.