STATISTICAL SIGNIFICANCE IN BIOLOGICAL SEQUENCE COMPARISON[1]

William R. Pearson and Todd C. Wood

Department of Biochemistry and Molecular Genetics,
University of Virginia, Charlottesville, VA 22908

April 11, 2000 - *final*

[1] The corresponding author is William R. Pearson; Phone: (804) 924-2818; FAX: (804) 924-5069; email: `wrp@virginia.EDU`

The availability of comprehensive sequence databases, rapid sequence comparison methods, and accurate statistical estimates for sequence similarity has fundamentally changed the practice of biochemistry and molecular biology. With the possible exceptions of *E. coli* and *Saccharomyces*, the vast majority of the genes in newly sequenced genomes are characterized by sequence similarity searching. `blast`, `fasta`, and Smith-Waterman similarity searches provide the most informative and reliable method for inferring the biological function of an anonymous gene (or the protein that it encodes). Typically, 60–80% of eubacterial (and yeast) genes share statistically significant sequence similarity with sequences from another organism. Significant sequence similarity can be used to infer common ancestors and similar three-dimensional structures, and is routinely used to assign functions in metabolic pathways. Even for the first archaebacterial genome sequenced (*M. jannaschii*; Bult *et al.*, 1996), similarity-based functional gene assignments could be made for about 50% of the genes (Andrade *et al.*, 1997) and subsequent sequence analyses (Koonin, 1997) suggested functions for another 20% of the genes.

Unfortunately, some investigators are uncomfortable inferring the relationship between two sequences from a probability or expectation value; they prefer to think in terms of percent identity (sometimes mis-stated as percent homology). When current versions of the `blast` and `fasta` similarity searching programs are used, this concern is rarely justified. It is very unusual for a statistically significant sequence similarity not to reflect common ancestry, and thus common structure, for the two sequences.

This chapter will provide an overview of the role of statistical significance estimates in biological sequence comparison, focusing on local similarity searches. We will begin by discussing the relationship between "statistical significance" and "biological significance," addressing the question: "What biological inferences can be drawn from *statistically significant* sequence similarity?" Next, we will survey strategies that have been used to estimate the significance of local sequence similarity scores. Finally, we will discuss the reliability of statistical estimates for local sequence similarity scores.

Statistical Significance and Biological Significance

`Blast`, `fasta`, and other sequence similarity searching programs are designed to identify distantly related—homologous—sequences based on sequence similarity. When we say that two sequences are homologous, we are stating our belief that the two sequences diverged from a common ancestor in the past. A remarkable result of microbial genome sequencing projects has been that a large fraction of proteins, typically 50–80% of each newly sequenced genome, share statistically significant similarity with proteins in other organisms that diverged hundreds to thousands of millions of years in the past. Thus, it is common to observe very strong sequence similarity between prokaryotic and eukaryotic proteins that diverged more than two billion years ago.

The inference of homology, at least as the term is commonly used in sequence analysis, implies that the homologous proteins have similar structures. Indeed, structural similarity is the gold standard for homology. Almost without exception, if two sequences share statistically significant similarity, they will share significant structural similarity. However, the converse is not true; there are many examples of similar structures that do not share significant similarity (though perhaps not as many examples as are presented in the literature).

The concept of homology was given wide exposure and common usage by Richard Owen, the first curator of the British Museum. Owen defined a homolog as simply "the same organ in different animals" (Owen, 1843). He further divided homology into two types: special and serial. Special homology is essentially the definition of homology we use today, "the same organ in different animals." In contrast, serial homology specifically refers to similarity between structures in different body segments, such as the legs of a millipede. Darwin's theory of evolution by natural selection conferred upon homology the specialized meaning of structures or organs that share a common ancestor. Although he regarded Darwin's theory as little more than speculation, Owen did admit that special homology was the result of common ancestry (Owen, 1866).

Implicit in all definitions of anatomical homology is some kind of recognizable similarity, e.g. similarity of form or ontogeny. The classic example of anatomical homology is the similarity of forelimbs in the higher vertebrates. Whether adapted for grasping, running, swimming, or flying, the same basic skeletal pattern can be readily observed. Although forelimbs are detectably similar in their adult forms, some homologous structures are only similar in embryonic stages.

Closely related concepts describe biological similarity that is not the result of a common evolutionary ancestry. Originally, Owen defined *analogy* as similarity of function, without regard to structure (Owen, 1843), and that definition was repeated by Neurath, Walsh, and Winter (Neurath *et al.*, 1967). The current definition of analogy adds the qualification that the similarity is not due to homology, that is, the similarity is primarily due to chance and is typically superficial (Kent, 1992). The horns of cows and rhinoceroses, and the limbs of insects and vertebrates are analogous.

Convergence is often invoked as a possible explanation of biological similarity, particularly when discussing protein sequence motifs. Properly understood, convergence refers to the process of evolution: two distantly related species developing a similar trait that was not present in their common ancestor. If convergence is observed over numerous stages of the evolution of two separate groups, it is termed *parallel evolution*. Examples of similarity from convergence include the body plans of sharks, dolphins, and ichthyosaurs. As each organism adapted to existence in the water, they developed a similar body plan by convergent evolution.

“Molecular” Homology

In the 1950’s and 1960’s, as protein sequences and three-dimensional structures were determined, researchers began to recognize surprising similarity between protein molecules. Though rigorous methods of understanding and detecting protein similarity were years away, the term “homology” was quickly applied to similarities observed among the trypsin-like proteases and the globins, implying that members of protein families shared their remarkable similarity because of divergence from a common ancestor.

Just as the term *homology* has been misused and misunderstood among anatomists, among biochemists the usage of homology as a synonym for similarity has unfortunately remained common. One often reads of “low homology” or even a quantified “percent homology” in papers reporting new sequences. Since homology is qualitative (having a common ancestor), it cannot be quantified as similarity can. Any two sequences have some measurable similarity, but a statement of homology implies that the similarity has some special meaning, specifically common ancestry.

Examples of Similarity in Proteins

Modern biochemical studies have revealed numerous examples of homology and analogy. In general, it is widely accepted that the three-dimensional structures of homologous proteins are more highly conserved than their sequences. Practically speaking, this means that homologous proteins with very low sequence similarity can and do have very similar structures. It is also believed that *orthologous* proteins—sequences that differ as a result of speciation events, in contrast to *paralogous* sequences, which result from gene duplication—share the same cellular function, and that new biological functions have arisen through the generation of paralogs by gene duplication.

Although mentioned frequently, convergent evolution as an explanation of protein similarity is not well defined. To avoid confusion, Doolittle (1994) proposed three categories of convergent evolution in proteins: mechanistic convergence, structural convergence, and functional convergence. The actual mechanisms that produce convergence are the subjects of ongoing research (Sanderson & Hufford, 1996). Here we will qualify convergence as similarity that arises by some kind of common selection. We will reserve the term analogy for similarity by chance, when no common selection is apparent.

Mechanistic convergence refers to similar active sites and residues in otherwise unrelated proteins. The classic example given is the mechanistic similarity between the trypsins and subtilisins. Although these proteins are entirely structurally dissimilar and thus almost certainly unrelated, they have geometrically and chemically equivalent catalytic triads. In mechanistic convergence, one can conclude that the need to accomplish a particular biochemical reaction is the selection producing the convergence. From principles of chemistry, it is reasonable to conclude that there are a limited number of enzymatic mechanisms available to accomplish particular reactions; thus, the occurrence of proteins with similar catalytic sites but distinct evolutionary histories is not surprising.

Structural convergence refers to structural similarity that is not the result of common ancestry. The adaptive selection applied to the structure is not the protein's cellular or biochemical function as in mechanistic convergence, but rather the thermodynamic stability of the particular fold. Doolittle mentions the ubiquitous TIM barrels (named for their well known example, triosephosphate isomerase) as examples of structural convergence. The structural similarity in convergent TIM barrels is typically both topological and geometric; that is, both the ordering of the secondary structural elements in the peptide chain and the atomic positions in space are similar. A second type of structural convergence is restricted to geometry only, proteins that have a similar three-dimensional arrangement of secondary structural elements but a different ordering of those elements in the peptide. Examples of geometric structural convergence include the pleckstrin homology domain (PHd) and verotoxin (Orengo *et al.*, 1995), and the N-terminal β-barrels of *E. coli* transcription termination factor rho and the F1 ATPase subunits. In each case, the arrangement of atoms in space is very similar, but the tracing of the peptide chain through those atoms is different. In the case of the rho/F1 similarity, the rho barrel is actually traced in reverse order with respect to the F1 barrel (Allison *et al.*, 1998).

A third category of protein convergence defined by Doolittle is *functional convergence*. Multiple examples of independent origins of the same or similar enzymatic activities are known. For example, Rawlings and Barrett used a sequence analysis and manual structural evaluation to assign peptidases to 64 different "clans," each with an independent evolutionary origin (Rawlings & Barrett, 1993). Although Doolittle calls this similarity "functional convergence," no adaptive advantage or selection pressure is known or given for why so many different kinds of peptidases would exist. Analogy, or similarity by chance, seems a better description for this type of gross functional similarity.

Inferences from Protein Homology

The inference of protein homology from similarity is routinely used to assign biochemical and cellular functions of newly sequenced proteins when a protein of known function is available for comparison. This is of critical importance for initial analysis of genomic sequences. For example, the vast majority of function assignments of the open reading frames (ORFs) of the *Methanococcus jannaschii* genome were made based on protein homologues detected by sequence similarity (Bult *et al.*, 1996). By properly using computational tools for sequence comparison, inferring homology from sequence similarity is the single most powerful tool we have today for understanding the function and origin of a protein without actually performing biochemical experiments.

Since protein structure is conserved in divergent evolution, identifying homologous proteins of known structure can give both a general insight into the fold of the protein of interest as well as a detailed molecular model if the sequence similarity is high enough. Although a remarkable amount of information about the function of a protein or protein complex can be gained from traditional biochemical and genetic methods, nothing brings these data into such clear focus as an atomic-resolution protein structure. Unfortunately, solving a protein structure by NMR or crystallographic methods can be very time-consuming, much more so than determining the sequence. Deriving structural and mechanistic information from closely related proteins of known structure will remain an attractive means of understanding most proteins.

Estimating statistical significance for local similarity searches

The inference of homology from statistically significant sequence similarity is an application of Occam's razor; given two competing hypotheses: first that a particular sequence ordering arose twice independently by chance; or second, that the similarity reflects divergence from a common ancestor, it seems simpler to conclude that a particular structure arose only once in evolutionary history. Thus, in biological sequence analysis, we infer homology from statistically significant sequence similarity. The inference depends on two parts, (1) our ability to measure sequence similarity, and (2) accurate estimates for the statistical significance of the similarity measure to reduce the likelihood that the similarity could be expected by chance.

Measuring sequence similarity

Sequence comparison algorithms

Effective algorithms for comparing protein and DNA sequences have been available for more than thirty years, since the publication of a global sequence comparison algorithm by Needleman and Wunsch (1970). *Global* sequence comparison algorithms seek to align every residue in one sequence with every residue in a second, in contrast to the more commonly used *local* sequence alignment algorithms, which seek only the strongest region of similarity between two sequences. Global alignment algorithms are used for aligning families of sequences with similar lengths in preparation for phylogenetic analysis; global alignment scores can be transformed to the distance measures used for building evolutionary trees. Global similarity scores are rarely used to infer homology however, because the distribution of global similarity scores is not well understood and thus it is difficult to assign a statistical significance to a global similarity score. Moreover, many proteins are made up of domains that are homologous only over a portion of the protein sequence.

The most widely used programs for searching protein and DNA sequence databases, including `blast`, `fasta`, and implementations of the Smith-Waterman algorithm, measure *local* sequence similarity. First described by Smith and Waterman (1981), local sequence alignment algorithms seek to align the most similar regions of two sequences. Local alignment algorithms have two dramatic advantages over global alignment methods when searching sequence databases for statistically significant matches: (1) the statistics of local similarity scores are well understood; and (2) local alignments allow one to identify conserved domains in proteins, which may not extend over the entire sequence. `blast` and `fasta` use heuristic methods that attempt to approximate the optimal local similarity shared by two sequences. `blast` is particularly efficient in identifying distantly related sequences because it spends very little time calculating similarity scores for sequences that are unlikely to share significant similarity. `fasta` is considerably slower than `blast`, because it calculates an approximate similarity score for every sequence in the database. `fasta` uses these approximate scores to estimate the parameters of the extreme-value distribution, λ and K, which describes the expected distribution of local similarity scores between random sequences.

Similarity scores for sequence comparison

All algorithms that calculate sequence similarity, global or local, optimal or heuristic, seek to maximize a measure of similarity. The earliest (and unfortunately most commonly cited even today) similarity measure was based on percent identical residues (Watson & Kendrew, 1961). Initially, the low percent identity of the myoglobin and hemoglobin sequences (typically less than 30%) was a surprising feature of two proteins with such similar structures. Later, researchers began to develop means to describe the similarity of different amino acid residues; the first such efforts were based on the redundancy of the genetic code, e.g. the minimal number of nucleotide substitutions required to convert one amino acid in the protein sequence to another (Fitch, 1966). In the 1970's, Margaret Dayhoff developed the notion of an *accepted point mutation* or PAM (Dayhoff *et al.*, 1978). The PAM concept centered around the natural selection against certain amino acid substitutions (thus an *accepted* point mutation) rather than simply the probability of mutations in the underlying DNA sequence. More recently the `BLOSUM` series of matrices, which tabulate the frequency with which different substitutions occur in conserved blocks of protein sequences, has been shown to be very effective in identifying distant relationships (Henikoff & Henikoff, 1992).

Dayhoff's PAM matrices are based on a well defined evolutionary model for protein sequences (Dayhoff *et al.*, 1978). Given an estimate for the probability that any amino-acid will change to each of the other amino-acids, or remain the same, after 1% change (1 accepted mutation per 100 residues), one can estimate the probability that any amino-acid will change into each of the others after $2\%, 10\%, \ldots, 40\%, \ldots 200\%$ change by multiplying the transition probability matrix by itself $2, 10, \ldots, 40, 200$ times. After incorporating the probability p_i of seeing a particular residue, the resulting matrix gives the probability $q_{i,j}$ of residue i aligning with residue j after a specified amount of evolutionary change. These probabilities are converted to log-odds scores by normalizing the alignment probabilities by the probability of seeing two residues align by chance, $p_i p_j$, yielding a scoring matrix $s_{i,j} = \log(\frac{q_{i,j}}{p_i p_j})$.

Fig. 1 shows parts of two `PAM` scoring matrices, `PAM40`, which incorporates transition probabilities between residues in sequences that have had 40 accepted mutations per 100 residues, and `PAM250`, which is "targeted" for sequences that have had 250 accepted mutations per 100 residues.[1] The `PAM40` and `PAM250` matrices differ dramatically in the relative scores of identities and substitutions; replacements that are considered unlikely at `PAM40`, e.g. '`R`' to '`N`' with $s_{N,R} = -7$ are

[1] Because different amino acids have different mutation probabilities, and an amino-acid can mutate to a different residue, which can then mutate again back to the original amino-acid, sequences that have changed by 250% are expected to remain about 20% identical, on average (Dayhoff *et al.*, 1978).

Figure 1: Similarity scoring matrices

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

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 $\log_2 \frac{q_{i,j}}{p_i p_j} = 2$, the entry in the matrix would be 6.

considered neutral $s_{N,R} = 0$ at PAM250. Likewise, replacements that are expected less frequently than chance ($s_{I,L} = -1$) after 40% change are more likely than chance substitutions ($s_{I,L} = 2$) after 250% change. Although the Dayhoff PAM matrices are based on the relatively small number of transitions available in 1978, a modern equivalent is available (Jones *et al.*, 1992), which performs well when appropriate gap penalties are used (Pearson, 1995).

An alternative strategy for calculating scoring matrices was developed by Henikoff and Henikoff (1992). Rather than extrapolate transition probabilities for a very large amount of change from the frequencies obtained after a very small amount of change, they sought to measure transition probabilities directly, by building a very large set of conserved blocks of aligned amino acid residues and then tabulating the amino acid substitution frequencies by examining columns in the aligned blocks with different degrees of identity (Henikoff & Henikoff, 1992). These calculations were used to generate the BLOSUM series of scoring matrices; BLOSUM50, the default scoring matrix used by the fasta family of sequence comparison programs, reports substitution frequencies for residues in conserved blocks of sequences that show 50% identity or less; BLOSUM62, which is the default for the blast programs, is derived from blocks that are $\leq$ 62% identical, and BLOSUM80 reflects a very high degree of sequence conservation by including sequences up to 80% identical. The BLOSUM matrices are now more widely used than either the original or modern versions of the PAM matrices because they appear to perform better with many alignment algorithms (Henikoff & Henikoff, 1992) and over a broad range of gap penalties (Pearson, 1995).

Both the PAM and BLOSUM series of matrices provide similarity scores that are targeted for different levels of sequence identity (Altschul, 1991; Henikoff & Henikoff, 1992); PAM matrices range from low values PAM10–PAM40 for high identity to PAM200–PAM250 for low (25%–20% identity). BLOSUM matrices range from high (BLOSUM80) values for high identity to low values BLOSUM50–45 for distant relationships. However, despite this apparent similarity, the meaning of a "shallow" PAM20 matrix is quite different from that of very conservative BLOSUM80 substitution values. The PAM20 provides scores for sequences that have changed by only 20%; the amount expected for a comparison of mouse and human proteins, for example. In contrast, BLOSUM80 is targeted towards the most highly conserved regions in proteins, blocks that remain up to 80% identical within two sequences that may share less than 30% identity overall. Thus, low PAM matrices, but not BLOSUM80, are appropriate for short divergence times.

Although the PAM and BLOSUM matrices were built to target specific models of evolution and conservation, Altschul has shown (Altschul, 1991; States *et al.*, 1991) that any scoring matrix can be written in the form $s_{i,j} = \log(\frac{q_{i,j}}{p_i p_j})$ reflecting an implied target substitution frequency, which can be calculated using the formula $\lambda s_{i,j} = \log(\frac{q_{i,j}}{p_i p_j})$. In particular, the blastn2.0 program for DNA substitutions, which uses $+1$ for a match and -3 for a mismatch, has $\lambda = 1.374$. Rearranging the equation above, the target frequency for any nucleotide match, assuming $p_{A,C,G,T} = 0.25$, is $q_{A,A} = q_{C,C} = q_{G,G} = q_{T,T} = p_A p_A e^{\lambda(+1)} = 0.2469$ and the overall target identity is $\sum_{b=A,C,G,T} p_{b,b} = 0.988$. Thus, the blastn2.0 is optimally efficient at identifying homologous sequences that are 98.8% identical, and considerably less efficient at identifying sequences that share 80% identity or less. In contrast, the DNA match/mismatch values for blast1.4 and fasta are $+5/-4$, which, with $\lambda = 0.1915$, are targeted for alignments averaging 65% identity.

Statistical Significance of Local Similarity Scores

A major breakthrough in biological sequence comparison occurred in 1990, when Karlin and Altschul published their statistical analysis of local sequence similarity scores without gaps (Karlin & Altschul, 1990), and the `blast` program incorporated those statistics (Altschul *et al.*, 1990). Although a method for evaluating the statistical significance of sequence similarity scores, the `rdf` program, was included with the `fastp` program (Lipman & Pearson, 1985), along with the advice that sequence similarity scores that were 6 standard deviations above the mean of the distribution of shuffled sequence scores ($z > 6$) were "probably" significant, there was no statistical basis for this observation. Work from Waterman and Arratia (Arratia *et al.*, 1986) and Karlin and Altschul (Karlin & Altschul, 1990) demonstrated that local similarity scores, at least for alignments without gaps, were accurately described by the extreme-value distribution, which can be written as:

$$p(S \geq x) = 1 - \exp(-Kmn\, e^{-\lambda x}) \tag{1}$$

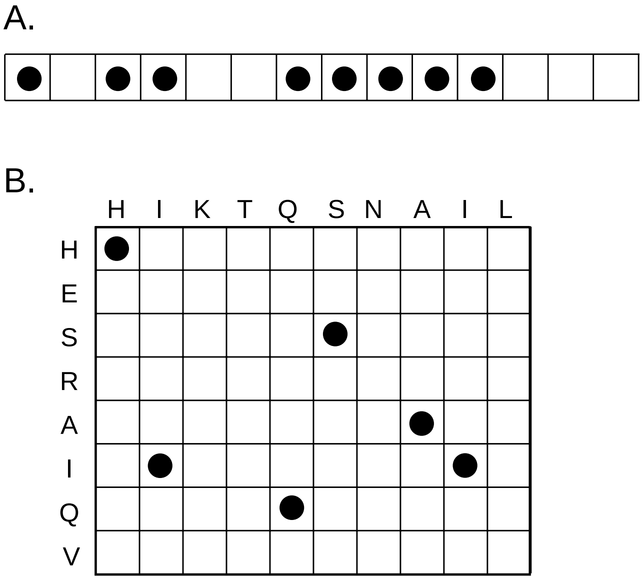
where λ and K can be calculated from the similarity scoring matrix $s_{i,j}$ and the amino acid compositions of the aligned sequences p_i, p_j, and m and n are the lengths of the two sequences.

Accurate similarity statistics allow us to discriminate reliably between statistically significant similarities, which reflect *homology*, and similarities that could have arisen by chance, *analogous* sequences. The availability of "Karlin-Altschul" statistics in the `blast` program (Altschul *et al.*, 1990) separated "first-generation" score-only programs from "second generation" methods. Without accurate statistics, it is impossible to do large scale sequence interpretation.

Statistics of Alignments Without Gaps

The first statistical models for local and global alignment scores applied to runs of similar amino acid or nucleotide residues, which are equivalent to alignments without gaps. Arratia, Gordan, and Waterman (Arratia *et al.*, 1986; Arratia *et al.*, 1990) and Karlin and Altschul (Karlin & Altschul, 1990; Karlin *et al.*, 1991) demonstrated that local similarity scores are expected to follow the extreme-value distribution. Waterman presents an intuitive argument in (Waterman, 1995), where, referring to Erdös and Renyí, he points out that the expected number of runs of heads of length l in n coin tosses is

Figure 2: Sequence comparison as coin tosses



(A) Results from tossing a coin 14 times; black circles indicate heads. The probability of 5 heads in a row is $p(5) = \frac{1}{2^5} = \frac{1}{32}$, but since there were 10 places that one could have obtained 5 heads in a row, the expected number of times that 5 heads occurs by chance is $E(5H) = 10 \times \frac{1}{32} = 0.31$. (B) Comparison of two protein sequences, with identities indicated as black circles. Assuming the residues were drawn from a population of 20, each with the same probability, the probability of an identical match is $p = 0.05$. In this example, there are $m = 10 \times n = 8$ boxes, so $E() = mnp = 80 \times 0.05 = 4$ matches are expected by chance. The probability of two successive matches is $p^2 = \frac{1}{20^2}$ so a run of two matches is expected about $nmp^2 = 8 \times 10 \times \frac{1}{20^2} = 0.2$ times by chance.

$E(l) \cong np^l$, where p is the probability of heads (Fig. 2). This relationship follows from the logic that the expected number of heads is the product of the probability of heads at each toss, times the number of tosses. If the longest run R_l is expected once, $1 = np^{R_l}$ and thus $R_l = \log_{1/p}(n)$. The longest run of heads coin toss example is equivalent to finding the highest scoring region (e.g. a hydrophobic patch) in a single protein sequence using a scoring matrix that assigns a positive value to some of the residues, and $-\infty$ to all of the others. The probability of a positive score, which corresponds to the probability of heads in the coin-toss example, is $\sum p_i$ for each of the residues p_i that obtain a positive score s_i.

The simple example of head-runs, or scores with $-\infty$ mismatch penalties, shows that *local* similarity scores for single sequences are expected to increase with the logarithm of the sequence length

n. In sequence comparison, we consider possible alignments of two sequences, $a_{1..m}$ and $b_{1..n}$, but the probability calculation is quite similar. Rather than calculate the probability of obtaining the k heads, where $p_k = pp_{k-1}$, we consider the case of matching at m positions, or equivalently giving a *head* score if $a_i = b_j$. If the sequences are placed as in Fig. 2B, the head-run problem corresponds to the longest run of matches along any of the diagonals. If the letters (residues) in the two sequences have equal probabilities p, then the probability of a match of residue a_i with b_j is p and the probability of a match of length l from a_i, b_j to a_{i+l-1}, b_{j+l-1} is again p^l. In this case, however, there are $m - l + 1 \times n - l + 1$ places where that match could start, so $E(l) \cong mnp^l$. Thus, the expected length of the longest match between two random sequences of length m and n when the match score is positive and the mismatch score is $-\infty$ is $M_{mn} = \log_{1/p}(mn)$ or $2\log_{1/p}(n)$ when $m = n$ (Waterman, 1995). The shift from $\log_{1/p}(n)$ for one sequence to $\log_{1/p}(n^2)$ for two sequences of length n, reflects the larger number of positions where a run of length M_l with probability $P(M_l) = p^{M_l}$ could start. As in the single sequence case, we can transform the problem from the probability of the longest match run to the probability of score $S_l \geq x$ by considering the probability $P(S \geq x)$ when a pair of residues $a_i b_j$ is matched with positive score $s_{i,j}$ and all negative scores are $-\infty$. For local pair-wise alignment scores with a mismatch score of $-\infty$ and no gaps, the expected number of runs of score $S \geq x$ has the general form: $E(S \geq x) \propto mnp^x$, or equivalently $E(S \geq x) \propto mn\, e^{x \ln p}$ or $mne^{-\lambda x}$ where $\lambda = -\ln p$.

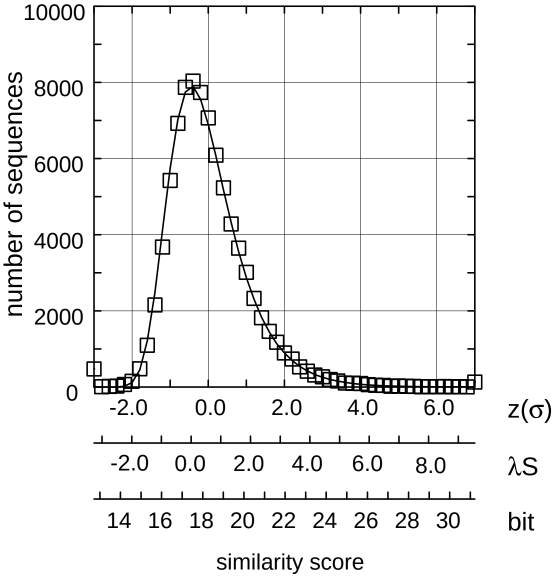
Karlin and Altschul provided a natural extension of the problem of head-runs, or match-runs, or positive similarity scores bounded by $-\infty$ mismatch scores, to the more general case of local sequence patches or local similarity scores for non-intersecting alignments without gaps. To ensure the scores are local, the requirement that $E(s_{i,j}) = \sum_{i,j} p_i p_j s_{i,j} < 0$ must first be met. If so, the expected number of alignments with score S is:

$$E(S \geq x) = Kmn\, e^{-\lambda x} \qquad (2)$$

Karlin and Altschul derived analytical expressions for K and λ (Karlin & Altschul, 1990). $K < 1$ is a proportionality constant that corrects the $m\,n$ “space factor” for the fact that there are not really $m\,n$ independent places that could have produced score $S \geq x$. Compared to λ, K has a modest effect on the statistical significance of a similarity score.

The λ parameter provides the scale factor by which a score must be multiplied to determine its

Figure 3: The extreme value distribution



The extreme value distribution. The observed distribution (squares) of similarity scores from a comparison of the human glucose transporter sequence `gtr1_human` against each of the $\sim 84,000$ sequences in Swiss-Prot, and the expected (solid line) distribution of scores, based on the extreme value distribution, is shown. Similarity scores were calculated with the Smith-Waterman algorithm, with the `BLOSUM62` scoring matrix and a penalty of -12 for the first residue in a gap and -1 for each additional residue. The y-axis shows the number of Swiss-Prot sequences obtaining the score shown on the x-axis. Three different scales for the similarity scores are shown: z(σ) shows the scores in terms of standard deviations (σ) above the mean (equation 6); λS shows the scale in terms of $\lambda S - \log(Kmn)$ (equation 1); bit shows the *bit* score (equation 4).

probability. For ungapped alignments, λ is the unique positive solution to the equation:

$$\sum_{i,j} p_i p_j \, e^{\lambda s_{i,j}} = 1 \tag{3}$$

λ thus depends both on the scoring matrix ($e^{s_{i,j}}$) and the residue compositions of the two sequences ($p_i p_j$). In some sense, λ can be interpreted as a factor that converts pair-wise match scores to probabilities, so that $e^{-\lambda x}$ is similar to p^l in the coin tossing example. Thus, just as in the coin-tossing case, the expectation of a run of heads (or an alignment run that produces score S) is the product of a space-factor term, Kmn, and a probability term $e^{-\lambda S}$.

The need for a scale factor to convert raw similarity scores into probabilities follows intuitively from the observation that multiplying or dividing every value in a similarity scoring matrix by a constant has no effect on the local alignments that would be produced by that matrix, or on the relative distribution of similarity scores in a library search—the highest scoring sequence will still be the highest, second highest second, etc. Thus it is impossible, without some previous knowledge of the scoring matrix used and the particular scaling of the scoring matrix, to evaluate the statistical significance of a raw similarity score. However, by using a scaled similarity score λS_{raw}, one can readily compare alignments done with any scoring matrix. `blast2.0` (Altschul *et al.*, 1997) and the current `fasta3` comparison program (Pearson, 1999) report the scaled score in terms of a *bit*-score that incorporates the space correction factor K: $S_{bit} = (\lambda S_{raw} - \ln K)/\ln 2$. Thus, substituting in equation 2:

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

Equations 2 and 4 describe the number of times a score $\geq S_{bit}$ would be expected by chance when two random sequences are compared.[2] This expectation can range from a very small value for very high scores (e.g. $S_{bit} = 1000$), to a value that approaches mn when $S = 0$. In a comparison of two average length protein sequences $n = m = 400$, $S_{bit} = 10$ would be expected $E(S_{bit} \geq 10) = mn2^{-S_{bit}} = 156$ times. To estimate the probability $P(S_{bit} \geq 10)$, which must range from 0 to 1, of obtaining at least one score $S \geq x$, we use the Poisson approximation.

[2] More accurately, the statistical model assumes that the two sequences are made up of residues that are independent and identically distributed, *i.i.d.*. The identical distribution assumption can be violated by low complexity regions in proteins and DNA or by strongly biased amino-acid or nucleotide composition.

The Poisson formula describes the probability of an event occuring a specified number of times, based on the average number of times μ it is expected to occur.[3] The Poisson probability of seeing n events when an event is expected μ times on average is: $P(n) = e^{-\mu}\,\mu^n/n!$. In general, we are interested in the probability of seeing the event $\geq n$ times, and in the case of sequence comparisons, we ask for the probability of seeing a high score one or more times ($n \geq 1$). In this case, one can calculate the probability of not seeing the event zero times: $P(n \geq 1) = 1 - P(0)$, so $P(S \geq x) = 1 - P(n = 0) = 1 - e^{-\mu}\mu^0/0!$. Since $\mu = E(S \geq x) = Kmn\,e^{-\lambda x}$ and $\mu^0 = 0! = 1$ the probability of seeing a raw similarity score $S \geq x$ is:

$$P(S \geq x) = 1 - \exp(-\mu) = 1 - \exp(-Kmn\,e^{-\lambda x})$$

as seen earlier in equation 1.

Equation 1 describes the probability of obtaining a similarity score $S \geq x$ in a single pairwise comparison of a query sequence of length m against a library sequence of length n. This equation has the same form as the extreme-value distribution or Gumbel distribution, which is often presented as:

$$P(S \geq x) = 1 - \exp(-e^{-(x-a)/b}) \tag{5}$$

with a providing the "location" of the mode, and b determining the scale, or width, of the distribution. For local similarity scores without gaps, $b = 1/\lambda$ and $a = \ln Kmn/\lambda$. The mean of the extreme-value distribution is $a - b\Gamma'(1)$, where $\Gamma'(1) = -0.577216$ is the first derivative of the gamma function $\Gamma(n = 1)$ with respect to n. The variance is $b^2\pi^2/6$ (Evans *et al.*, 1993). Thus, one can express the probability that an alignment obtains a score z standard deviations above the mean of the distribution of unrelated (or random) sequence scores as:

$$P(Z \geq z) = 1 - \exp(-e^{-(\frac{\pi}{\sqrt{6}}z - \Gamma'(1))}) \tag{6}$$

These equations describe the probability that two sequences would obtain a similarity score by chance in a single comparison. However, in a sequence database search, the highest scoring alignments are identified after a query sequence has been compared with each of the 10–100's of thousands

[3] λ is generally used to denote the characteristic parameter of a Poisson distribution, but we use μ here to avoid confusion with the λ scaling factor and to reinforce the fact that μ is the mean of the Poisson distribution.

of sequences in the database. Thus, in the context of a database search after $D = 100,000$–$500,000$ or more alignments have been scored, the number of times a score is expected to occur, the *expectation value,* is considerably higher:

$$E(S \geq x) = DP(S \geq x) \tag{7}$$

Thus, in 1985, with protein sequence databases containing fewer than $3,000$ sequences, Thus, a similarity score 6 standard deviations above the mean ($z = 6$) has a probability $P(Z \geq 6) < 2.6 \times 10^{-4}$ in a single pairwise comparison. However, in 1985, with $3,000$ entries in the protein sequence database, $E(Z \geq 6) = 3,000 \times 2.6 \times 10^{-4} = 0.77$. Thus, a score 6 standard deviations above the mean should be seen by chance very frequently, and the advice provided with the description of the `fastp` program overestimated statistical significance. Today, with protein sequence databases ranging in size from $100,000$–$500,000$ sequences, a $z = 6$ score would be expected 25 times by chance when searching a $100,000$ sequence database (equations 6 and 7), and $z \geq 12.1$ is required to achieve statistical significance of $E(100,000) \leq 0.01$. (For $E(500,000) \leq 0.01$, $z \geq 13.4$.)

Alignments with Gaps

The statistical analysis of local similarity scores summarized in the previous section was derived for alignments without gaps. Although searches that report only the best local alignment without gaps can perform very well, they do not perform as well as a Smith-Waterman search with modern scoring matrices and appropriate gap penalties (Pearson, 1995). Thus, there is considerable interest in the statistical parameters that describe the distribution of local similarity scores with gaps.

The first implementation of the Smith-Waterman (Smith & Waterman, 1981) algorithm that provided statistical estimates for similarity scores was developed by Collins *et al.* (1992). Although they did not use the extreme-value distribution, they recognized that the number of sequences S_x obtaining a score of $x \geq \bar{x}$, where $\bar{x}$ is the mean similarity score, decreases exponentially. A line fit to $\log(S_x)$, the declining number of scores excluding the top 3%, can be used to extrapolate the expectation of obtaining a high-score. This strategy works reasonably well because the number of sequences that obtain a score predicted from the probability density function of the extreme-value distribution (Fig. 3) has the form: $PDF(S = x) = \lambda Kmne^{-\lambda x}e^{-Kmne^{-\lambda x}}$ The second exponential term does not contribute significantly to the *PDF* when $x > \bar{x}$, so for high scores, the regression becomes: $\log(PDF(S = x)) = \log(\lambda Kmn) - \lambda x$. Collins *et al.* recognized that the highest expected score by chance increased with the length of the query sequence (Collins *et al.*, 1988) but they did not incorporate a length correction into their expectation calculation. The lack of a $\log n_l$ library sequence length correction significantly reduces the sensitivity of the search, as long unrelated sequences can have higher scores, by chance, than shorter related sequences (Pearson, 1995; Pearson, 1998).

Mott (1992) provided the first empirical evidence that the distribution of optimal local similarity scores with gaps could be well approximated by an extreme-value distribution. He considered the an equation of the form: $F(y, m, n, c) = \exp(-e^{-(y-A)/B})$ where $A = a_0 + ca_1 + ca_2 \log(mn)$, $B = cb_1$ and $c = 1/\lambda_{ungapped}$, defined as in equation 3. In this case, a $c = 1/\lambda$ parameter was calculated for sets of sequence pairs with identical compositions. In addition to correcting for the scaling of the $s_{i,j}$ scoring matrix, $c = 1/\lambda$ reflects the amino-acid composition of the two sequences being compared. Unfortunately, estimating λ or c using equation 3 is time-consuming. This approach may improve searches when query sequences have a biased amino-acid composition, but it is not generally available in sequence comparison programs.

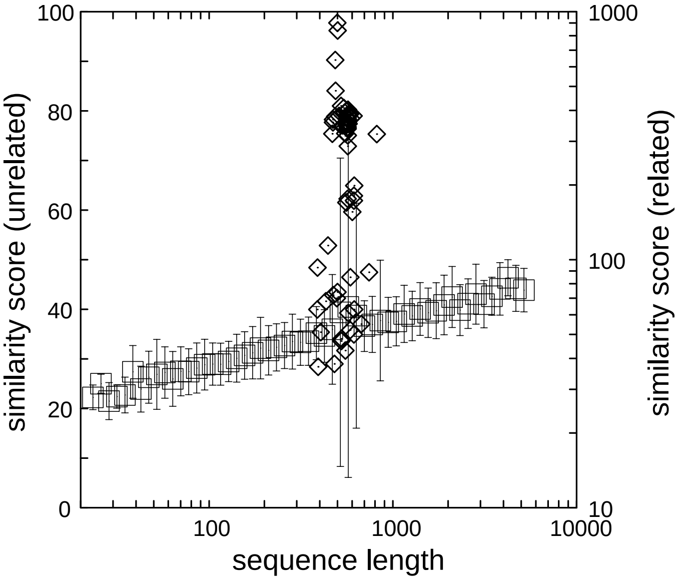
The most widely used estimates for λ and K for searches with gapped alignments are those provided for in the `blast2` and `psi-blast` comparison programs (Altschul *et al.*, 1997). These values are based on maximum-likelihood estimates of λ, K, and H from simulations of random protein sequences of average composition (Altschul & Gish, 1996). H describes the relative entropy, or information content, of a scoring matrix and can be thought of as the average score per aligned residue (Altschul & Gish, 1996). In this case, the parameters of the extreme-value distribution are slightly different:

$$P(S \geq x) = 1 - \exp(-Km'n'e^{-\lambda x}) \qquad (8)$$

where for m, n, the query and library sequence lengths, $m' = m - \log(Kmn)/H$ and $n' = n - \log(Kmn)/H$. By correcting m and n for the expected length of an alignment between two random sequences $\log(Kmn)/H$, the search space term $Km'n'$ is estimated more accurately (Altschul & Gish, 1996).

The `fasta` package of programs estimates the extreme-value parameters from the distribution of similarity scores calculated during the search (Pearson, 1996; Pearson, 1998). This approach is

Figure 4: Empirical estimation of extreme-value parameters



Sequence similarity scores plotted as a function of library sequence length. All the similarity scores calculated in the comparison of `gtr1_human` with an annotated subset of SwissProt ($\sim 24,000$ sequences) are summarized. Scores from unrelated sequences are shown as averages (squares) with standard errors indicated; each score from a related sequence is plotted (diamonds). The unrelated sequence scores are plotted linearly against the left ordinate, the related scores are plotted on a logarithmic scale on the right ordinate.

efficient when scores are available for every sequence in the database, as is the case for a `fasta` or Smith-Waterman search; no additional similarity scores must be calculated and the statistical parameters reflect the true distribution of similarity scores produced with the specific query sequence and the specific sequence database. However, a method that estimates statistical parameters from the actual distribution of similarity scores must avoid including scores from “related” sequences in the estimation sample. This is straightforward in the typical case where a query sequence is compared to $50,000$–$500,000$ sequences in a comprehensive database and fewer than $1,000$ sequences could be related in the worst case. However, these empirical statistical estimates cannot be used when a search is done against a special purpose database that may contain only sequences from a single protein family. For this case, the `fasta` programs provide an option to calculate a similarity score from a shuffled version of each sequence in the database; the distribution of these shuffled scores are then used for parameter estimation (Pearson, 1999).

By default, the `fasta` programs estimate the location and scale parameters of the extreme-value distribution by fitting a line to the relationship between similarity score and $\log(n_l)$, the library sequence length, by calculating the mean and variance of similarity scores in bins of length $\log(n_l)$ of the library sequence (sequences in each bin differ in length by $\sim$ 10%, Fig. 4). This line provides the location parameter, related to $\log(Kn_l)/\lambda$, and the residual variance ($\widehat{\sigma^2}$) of the $\log(n_l)$-normalized similarity scores can be used to calculate $\lambda = \pi/\sqrt{6\widehat{\sigma^2}}$. Binning similarity scores by $\log(n_l)$ provides a simple strategy for excluding related (high-scoring) sequences from the estimation process. $\log(n_l)$ length bins are initially weighted by the inverse variance of their similarity scores; length bins with very high scores have a high variance. After the initial $\log(n_l)$ regression is performed, bins that continue to have very high score variance are excluded and the number of bins and scores excluded is reported. Typically, this process excludes 0–2 of 50 length bins with about 5% of the library sequences. Once the $\log(n_l)$ regression line and $\widehat{\sigma^2}$, the average residual variance, have been determined, the probability of a single pair-wise similarity score can be calculated using equation 6.

Alternatively, `fasta` provides an option to estimate λ and K using maximum-likelihood, using equation 1. This estimation is similar to that of Mott (1992), but omits the "composition" data c and estimates the λ parameter directly. To avoid the scores from related sequences, the likelihood model implements a censored estimation strategy that excludes the lowest and highest 2.5% of the scores. This approach has the advantage that both K and λ are estimated directly and that there is no assumption that related sequences have well defined lengths. (Families that are globally similar, e.g. globins and cytochome 'c's, have characteristic lengths, but homologous domains, e.g. the EF-hand calcium binding domain, zinc-fingers, or protein kinase domains may be in proteins with very different lengths).

The major difference between the `fasta` programs and the `blast` programs (aside from speed) is the strategy used for estimating statistical significance of similarity scores. While `blast` pre-calculates λ and K from randomly shuffled sequences, `fasta` calculates the extreme value parameters from the actual distribution of similarity scores obtained in a search. Thus, `fasta` must calculate at least an approximate similarity score for every sequence in a database. `blast` is fast because it calculates scores only for sequences that are likely to be homologous. While this strategy works well for protein sequences, it is more problematic for translated-DNA:protein comparisons, where the appropriate statistical model is more difficult to specify.

Pairwise Statistical Significance

The strategies outlined above can be used to estimate the statistical significance of a high similarity score obtained during a database search. If the `blast2.0` (Altschul *et al.*, 1997) λ and K parameters are used as calculated in (Altschul & Gish, 1996), the statistical significance measurement reports the likelihood that a similarity score as good or better would be obtained by two random sequences with "average" amino-acid composition, and lengths similar to the lengths of the sequences that produced the score. However, if either of the two sequences have amino acid compositions significantly different from "average," the statistical significance may be an over or underestimate.

The empirical statistical estimates provided by programs in the `fasta` package (Pearson, 1996; Pearson, 1998) report a slightly different value; the expectation that a sequence with the length and composition of the query sequence would obtain a similarity score against an unrelated sequence drawn at random from the sequence database that was searched. Again, if the query sequence has a slightly biased amino acid composition, e.g. because it is a membrane-spanning protein with several hydrophobic regions, then while the significance of the similarity with respect to "average" composition proteins is accurate, the more biologically important question, the significance of the similarity when compared to unrelated membrane-spanning proteins, may be an overestimate. To address this problem one could use Mott's strategy to include the $c = 1/\lambda_{ungapped}$ composition/scaling parameter in the maximum-likelihood fit, with c calculated using equation 3 for every pairwise comparison in the database. Although the composition calculation can be time consuming, this option is available in the `fasta3` package.

The significance of a specific pairwise similarity score, in the context of the residue distributions in each of the two sequences, the query and library sequence, can also be estimated using a Monte-Carlo approach. The two sequences are compared, and then one or both of the sequences is shuffled hundreds of times to generate a sample of random sequences with the same length and residue composition. Similarity scores are calculated for alignments between the query sequence and each of

Table 1: Statistical Significance Estimates

	qutd_emeni	cit1_ecoli	kgtp_ecoli	yb8g_yeast	λ
gtr1_human:	moderate	distant	very weak	unrelated	
blastp2.0	2.0e-25	1e-05	0.077	2.0	0.2700
ssearch BL50	1.6e-28	6.1e-05	0.014	0.72	0.1544
raw-score	536	199	148	123	
bit-score	127	48	40	35	
% identity	27.1	22.1	24.1	22.1	
ssearch BL62	4.7e-32	1.2e-4	1.3	3.1	0.2584
raw-score	356	120	75	72	
bit-score	138	47	34	33	
% identity	26.9	21.0	27.9	24.1	
ssearch BL62*	2.8e-30	3.2e-4	2.3	5.2	0.2459
bit-score	356	46	33	32	
prss BL50	7.2e-25	6.5e-03	0.0039	92.	
λ	0.1375	0.1237	0.1317	0.1263	
window 20	3.9e-09	0.097	0.21	361.	
λ	0.0653	0.1064	0.1206	0.1110	
BL62	6.6e-30	8.5e-4	0.36	49.	
λ	0.2405	0.2282	0.2343	0.2265	
window 20	2.0e-25	9.8e-03	0.72	92.	
λ	0.2108	0.2011	0.2256	0.2172	

Expectation values are shown for similarity scores between human glucose transporter type 1 (gtr1_human) and three members of the glucose transporter family quinate permease (qutd_emeni); maltose permease (cit1_ecoli); α-ketoglutarate permease (kgtp_ecoli) and a probable non-member a hypothetical yeast protein (yb8g_yeast). The blastp2.0 search was done with the default scoring matrix and gap penalties, BLOSUM62, -12 for the first residue in a gap (-11 gap-open), and -1 for each additional residue (gap-extend). ssearch (Smith-Waterman, Smith & Waterman, 1981,Pearson, 1996 searches used either the default matrix (BLOSUM50, BL50) and gap penalties ($-12/-2$) or the same scoring matrix and gap penalties as the blastp2.0 search (BL62). ssearch statistical estimates were calculated using the default linear-regression method (BL50, BL62) or the maximum likelihood method (BL62*). Both blastp2.0 and ssearch searches examined alignments between sequences with low-complexity regions removed by the seg program (Wootton, 1994). Expectation values are reported in the context of a search of Swiss-Prot (Bairoch & Apweiler, 1996) database ($\sim 84,000$ entries). The λ scaling/composition-factor for each search is shown in the right column.

Statistical significance was also estimated by with a Monte-Carlo approach (prss) in which the second sequence was shuffled 1000 times using either a uniform or “window” (-w 20) shuffle. Expectations reported by prss have been multiplied by 84 to reflect the expectation from a search of the $84,000$ entry Swiss-Prot database.

the shuffled sequences. λ and K parameters can then be calculated from this distribution of scores using maximum likelihood, as is done by the `prss` program in the `fasta` package (Pearson, 1996). The `fasta` programs offer two shuffling options: (a) a *uniform* shuffle, in which each residue is randomly repositioned anywhere in the sequence; and (b) a *window* shuffle, in which the sequence is broken into n/w windows (n is the length of the sequence and w is the length of the window, typically 10–20 residues) and the sequence in each window is randomly shuffled. For "average" composition query sequences, both uniform and window-shuffle estimates should be similar to those obtained from a database search. However, for scores of alignments between sequences of biased composition, significance estimates derived from the similarity scores of uniformly shuffled sequences should be more conservative than estimates based on the distribution of unrelated sequences from a comprehensive sequence database (Table 1). Window-shuffle estimates should be even more conservative, particularly if the similarity reflects a local patch of biased amino acid composition that would be homogenized by the uniform shuffling strategy.

Shuffling strategies rely on the assumption that the similarity scores of real unrelated protein sequences behave like the similarity scores of randomly generated sequences. While this is almost always true, some query sequences may have properties that are present in unrelated sequences but not in shuffled sequences. An alternative strategy for estimating λ and K from a comparison of two sequences has been proposed by Waterman and Vingron (1994),[4] based on a strategy they refer to as "Poisson de-clumping". They note that not only are the highest scoring similarity scores from a sequence similarity search extreme-value distributed, but the highest $H_{(1)}$, second highest $H_{(2)}$, $H_{(3)},\ldots,H_{(n)}$ alignment scores from a single pairwise comparison can be used to estimate λ and K, as long as the alignments do not overlap or intersect. An algorithm for calculating the n-best non-intersecting local alignments between two sequences was described by Waterman and Eggert (Waterman & Eggert, 1987), a space-efficient version is available as the `sim` algorithm (Huang *et al.*, 1990). This approach has the advantage that it does not require the use of shuffled sequences, which may have different statistical properties than "natural" protein sequences in some cases, and it calculates λ and K for the pair of sequences, with their specific lengths and residue compositions, rather

[4] p and γ in (Waterman & Vingron, 1994) correspond to $e^{-\lambda}$ and K, respectively in equation 1.

than for an average distribution of library sequences. However, the approach also assumes that for some i, one can assume that $H_{(i)}$ reflects the score of an alignment that occurs by chance, rather than because of homology. This is true for single domain proteins that do not contain internal repeats, but it is not true for proteins containing internal duplications. For example, a comparison of calmodulin with troponin 'C' would produce $H_{(1)},\ldots,H_{(4)}$ which reflect the homology of the four EF-hand calcium binding domains in each sequence, and $H_{(5)},\ldots,H_{(n)}$, which could be used to estimate λ and K. A protein with a dozen copies of a duplicated domain would have more than 100 local alignments with scores that reflect homology.

Accuracy of λ and K

Reliable statistical estimates for similarity scores can dramatically improve the sensitivity of a similarity search, because they provide an accurate quantitative model for the behavior of scores from unrelated sequences. Thus, it is far more informative to state that a pair of distantly related sequences has a similarity score that is expected by chance only once in $10,000$ database searches ($E() < 10^{-4}$) than it is to state that two sequences share 30% identity. Unfortunately, percent-identity remains the most commonly published measure of sequence similarity, despite the fact that identity measures are far less effective than similarity scores that reflect conservative replacements (Schwartz & Dayhoff, 1978; Pearson, 1995; Levitt & Gerstein, 1998). High levels of identity are frequently seen between unrelated sequences over short regions (Kabsch & Sander, 1984) and sequence alignments with less than 25% identity may either be clearly statistically significant (Table 1, `gtr1_human` versus `cit1_ecoli`, `BL62`, $E() < 10^{-9}$) or not significant (`gtr1_human` vs. `yb8g_yeast`, `BL62`, $E() < 0.25$).

Before accurate statistical estimates for local similarity scores were available, it was routine to consider the tradeoffs between a search strategy's "sensitivity," the ability to identify distantly related sequences (to avoid false-negatives), and its "selectivity," not assigning high scores to unrelated sequences (false-positives). With an accurate model for the distribution of similarity scores from unrelated sequences, the threshold for statistical significance (typically 0.02–0.001) sets the selectivity or false-positive rate; a threshold $E() < 0.001$ predicts a false positive every 1000 searches. Thus, a significance threshold of $E() < 0.001$ is expected to produce several false positives when characterizing all the proteins in *E. coli* or yeast ($4,000$ and $6,000$ proteins), and 18 false positives are expected with $E() < 0.001$ when each of the $18,000$ proteins in *C. elegans* is compared to the SwissProt database. However, the conservative strategy of reducing the significance threshold to $0.001/4,000$ for *E. coli*, or $0.001/18,000$ for *C. elegans*, ensures that many homologous proteins will be missed (false negatives).

Of the λ and K parameters for the extreme value distribution, the scale parameter λ has the largest effect on the statistical significance estimate. In searches using the BLOSUM62 scoring matrix and gap penalties of $-12/-2$ of a subset of the Swiss-Prot with 50 unrelated protein sequences with lengths ranging from 98–$2,252$ (mean 432 ± 57), maximum likelihood estimates of λ ranged from 0.204–0.304 (mean 0.275) while K ranged from 0.0039–0.062 (mean 0.012). K and λ are strongly correlated; low values of K are found with low values of λ. Around the average values, however, reducing K by a factor of 2 reduces the $E()$ value only 2-fold, (1-bit), but a similar change in statistical significance would occur by reducing λ from 0.275 to 0.268, or about 2.5%. Reducing λ by 20%, which is well within the range of λ's seen after shuffling with `prss` in Table 1, would reduce the statistical significance of a raw score of 100 250-fold, or 8-bits.

Table 1 illustrates the importance of λ on significance estimates for three related and one unrelated sequence. The differences in expectation values reflect differences in estimates for λ and K; for a given scoring matrix (`BLOSUM50` or `BLOSUM62`) the raw similarity scores for each pair-wise comparison (e.g. `gtr1_human:cit1_ecoli`) do not change. The significant differences between the λ values for `BLOSUM50` and `BLOSUM62` reflect the different scaling of the two matrices. `BLOSUM50` is scaled at 0.33 bits per unit raw score, so that a raw score of 148 produces a bit score of $\sim 148/3$ (the actual value for these gap penalties is 0.27 bits/raw-score). `BLOSUM62` is scaled at 0.5 bits/raw-score, with a raw score of 75 giving a bit score of 34.[5]

Two trends are apparent: (1) λ estimates from `prss` shuffled comparisons tend to be smaller than λ estimates from database searches, and (2) λ estimates for local (window) shuffles are somewhat

[5] Because of this different scaling, a gap penalty of $-12/-2$ for `BLOSUM50`, the default with `ssearch`, is equivalent to a gap penalty of $-8/-1$ for `BLOSUM62`.

lower, reducing the significance even further. These decreases in λ are expected because the query and library sequences used in this example have a somewhat biased amino-acid composition; the proteins have multiple transmembrane domains with a bias towards hydrophobic amino-acid residues (Kyte & Doolittle, 1982). Thus, the λ's from `ssearch` are lower than the `blast2.0` λ, because the simulations used to assign λ in `blast2.0` assume an "average" amino-acid composition for both the query and library sequence; the empirical `ssearch` estimates correct for the composition bias of the query sequence, but still reflect the "average" composition of the library sequences. λ's determined by `prss` shuffling are lower still, because `prss` estimates account for the composition bias in both the query and library sequences. Window shuffling in `prss` reduces λ even further, presumably because the highest scoring regions in each pairwise comparison are restricted to sequence patches with the most biased composition. However, despite these differences in λ's, the `ssearch` and `prss` uniform-shuffle significance estimates for the intermediate and distantly related sequence pairs usually agree within a factor of four. Window-shuffled estimates reduce statistical significance much more dramatically, about 2–4 orders of magnitude for moderately and weakly significant similarities.

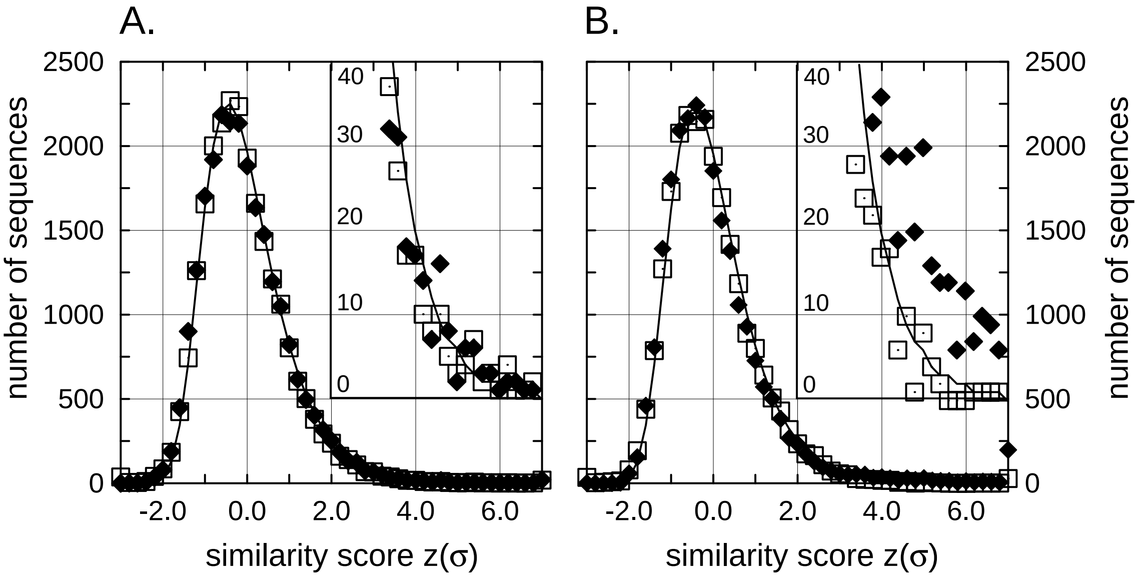
The statistical estimates provided by the `blast2.0` and `fasta` sequence comparison programs are generally robust and reliable. To illustrate the factors affecting significance estimates, we have emphasized the modest differences in λ and $E()$ in Table 1. However, Table 1 illustrates even for sequences with biased amino-acid composition that share 20–25% sequence identity, the significance estimates reported by either `blast2.0` or programs in the `fasta` package are very similar, and consistent with statistical estimates produced by uniform shuffling. Window-based shuffling produces a much more conservative statistical estimate.

Evaluating statistical estimates

The inference of homology (common ancestry) from statistically significant similarity rests on two assertions: (1) that similarity scores, calculated with optimal (Smith-Waterman) or heuristic (`blast` or `fasta`) algorithms using common scoring matrices (`PAM250`, `BLOSUM62`) and gap penalties follow the extreme value distribution; and (2) that the behavior of similarity scores for random sequences holds as well for real, unrelated, protein sequences. This second assertion is critical—an accurate statistical theory for similarity scores of random sequences is of little value if real sequences have properties that distinguish their scores from those of random sequences. It seems reasonable that real protein sequences might have statistical properties that distinguish them from random sequences; of the $20^{400} = 10^{520}$ potential sequences of length 400 that could be generated at random from 20 amino-acids, fewer than 10^5–10^8 unrelated sequences are thought to exist in nature, and many structural biologists would argue that there are fewer than 10^3 distinct protein folds (Brenner *et al.*, 1997). Real protein sequences are constrained to fold into a compact three dimensional structure with a physiological function; the fact that such a large fraction (typically 50–80%) of the sequences in most organisms can be found in other distantly related organisms suggests that the folding constraint substantially restricts the universe of protein sequences; it is far easier to produce a new protein sequence by duplicating an old one than by producing a sequence *de novo*. Thus, it would not be surprising to learn that the folding/function constraint produced real protein sequences whose similarity scores behave differently from those of random protein sequences.

The reliability of statistical estimates can be evaluated both by (1) comparing the observed distribution of sequence similarity scores obtained in a search with the expected extreme value distribution and (2) examining the expectation value for the highest scoring non-homologous sequence. Fig. 5 shows the distribution of sequence similarity scores for two query sequences, an “average” protein sequence, `pyre_colgr`, orotate phosphoribosyltransferase, and a protein sequence with a biased amino-acid composition, `prio_atepa`, major prion precursor. Two sets of similarity scores are shown for each sequence. One set shows the scores obtained when all the amino-acid residues in the library are examined; the second shows the scores when low-complexity sequences, or regions of sequenced with a reduced or biased amino acid composition, are removed (Wootton, 1994). With the “average” protein, the distributions of the “complete” database scores and “high-complexity” scores are indistinguishable. With the prion protein (Fig. 5B), there is some difference in the central portion of the distribution, but the greatest differences are seen for the highest scoring sequences, where there are typically 2–3 times as many “raw” scores as expected between 4–5 standard deviations above the mean, and 5–10-times as many scores as expected from 5–7 standard deviations above the mean. This effect of biased composition is largely removed by searching against a “`seg`’ed” database that

Figure 5: Distribution of sequence similarity scores



Distribution of library sequence similarity scores in searches with (A) an "average" protein sequence, `pyre_colgr` , and (B) a sequence with biased amino-acid composition, `prio_atepa`. Filled diamonds show the distribution of similarity scores that include all the residues in every sequence; open squares show the distribution of scores when low-complexity regions are removed with the `pseg` program (Wootton, 1994). The solid line shows the expected distribution of scores predicted from the size of the database and the extreme-value distribution (equ. 6). The x-axis reports similarity scores scaled in standard deviations above the mean. Searches were done with the `ssearch33` program (Smith-Waterman) using the `BLOSUM62` scoring matrix with gap penalties of -12 for the first residue in a gap and -2 for each additional residue. λ and K were estimated by maximum likelihood (`-z 2` option, Pearson, 1999).

has had low complexity regions removed.

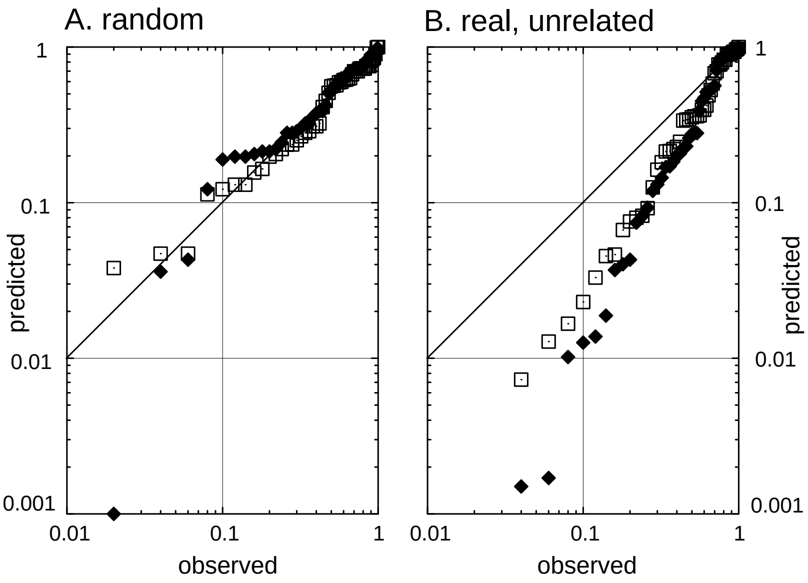
The effect of biased composition is seen more dramatically by looking at the number of very high-scoring sequences and the expectation value of the highest scoring unrelated sequence. When `prio_atepa` is used as a query, 198 library "raw" sequence scores have $z \geq 7.0$;[6] this is reduced to 28, 26 of which are related to the query, when the "`seg`'ed" database is used. Likewise, when the "raw" sequences are examined, the highest scoring unrelated sequence is a glycine-rich cell wall protein that obtains an expectation value of $E() < 10^{-8}$ and there are 90 unrelated sequences with $10^{-8} \leq E() \leq 0.01$. In contrast, with the "`seg`'ed" database the highest scoring unrelated sequence has an expectation value of $E() = 0.012$ and the second highest unrelated sequence has $E() = 0.99$.

Reliable statistical estimates—statistics that estimate $E() < 0.02$ about 2% of the time—allow much more sensitive searches. If an investigator can have confidence that an unrelated sequence will obtain a score of $E() < 0.001$ about once in $1,000$ searches, $E() < 0.001$ can be used to reliably infer homology. However, if unrelated sequences sometimes obtain $E() < 0.001$ by chance, a more conservative threshold may be adopted, e.g. $E() < 10^{-6}$ or even $E() < 10^{-10}$. While using a very stringent threshold for statistical significance ensures that one will rarely infer homology when the proteins are unrelated, it also ensures that moderately distant evolutionary relationships will be missed. Thus, both the `fasta` and `blast` developers have given high priority to the accuracy of the statistical estimates, particularly for the highest scoring unrelated sequences (Brenner *et al.*, 1998).

When evaluating the quality of statistical estimates for high scoring unrelated sequences, it is important to examine real protein sequences, whose properties may differ from randomly generated sequences. Fig. 6 summarizes the highest scoring unrelated sequence similarity scores obtained when query sequences from 50 randomly selected Pfam protein families were used to search a database of sequences with carefully annotated evolutionary relationships. Searches were done with either random sequences generated from the 50 Pfam family queries, or with the queries themselves, against either a "raw" protein sequence database, or one with low complexity regions removed. Fig. 6A shows that even when random sequences are used to search the database, similarity scores can be

[6] Similarity scores 7 standard deviations above the mean have an expectation value $E() < 1.7$ for this database of $23,981$ sequences.

Figure 6: Predicted and Observed Statistical Significance



Quantile-quantile plot of expectation values for searches with (A) 50 random sequences and (B) 50 real protein sequences for which the highest scoring unrelated sequence is known. Searches were performed against a "raw" annotated protein sequence database (filled diamonds) and the same database with low complexity regions removed (open squares). For each search, the highest score (A) or highest scoring unrelated (B) sequence was recorded, and converted from an expectation ($E()$) to a probability of obtaining that $E()$ using the Poisson formula $p(E) = 1 - e^{-E}$. Each set of 50 probabilities was sorted from lowest to highest and plotted. The 50 query sequences was chosen from 50 randomly selected PFAM families (Sonnhammer *et al.*, 1997) with 25 or more members. The random sequences were obtained by shuffling the 50 real PFAM derived sequences. Searches were done using the Smith-Waterman algorithm (`ssearch33`) using the default scoring matrix (`BLOSUM50`) and gap penalties ($-12/-2$) with regression-scaled (binned) statistical estimates.

much higher than expected (and $E()$-values much lower than expected) if low complexity regions are present in the sequence database. Thus, when 50 random sequences were used, the lowest $E()$-value was 0.006 from a match between a randomly shuffled human histone H1 (`h10_human`) and other histone H1 sequences. This may simply reflect the fact that it is difficult to randomly shuffle a sequence that is 30% lysine. However, when low complexity regions are removed, the observed and expected distributions of $E()$-values agree extremely well.

When real sequences are used as the query, the statistical estimates are not as accurate, even when low complexity regions are removed. Most of the time however, the estimates are not far off. The log/log plot in Fig. 6 emphasizes the searches that obtained the lowest $E()$-values for unrelated sequences, but 80% of the query sequences had expectation values $E() > 0.1$ (low by a factor of 2), and 90% had $E() > 0.02$ (low by a factor of 5) when low-complexity sequences were removed from the database. In the search of the "`seg`'ed" database, again the most "significant" unrelated similarity score was involved alignments with `h10_human`. In the search against the raw database, this alignment had an $E() < 0.002$ (low by factor of 10), against a "`seg`'ed" database the score was even lower, $E() < 0.0006$. Histone H1 has an exceptionally biased amino-acid composition, which cannot be completely corrected for by removing low complexity regions from the database. However, for the vast majority of query sequences (80–90%), unrelated sequences will have expectation values within a 2–5 of their true frequency in database searches. Thus, thresholds of statistical significance between $0.001 < E() < 0.01$ against "`seg`'ed" sequence databases will be reliable with rare exceptions.

The observation that the statistical significance estimates ($E()$-values) from similarity searches with real, unrelated sequences are 2–5-fold less conservative than those obtained for genuinely random sequences suggests that to a large extent, real, unrelated protein sequences have many of the same statistical properties as random sequences. The major difference between real protein sequences and random sequences seems to be the independent, identically distributed (*i.i.d.*) assumption for amino-acid residue positions. In real, unrelated sequences, unusual amino-acid compositions are distributed in low-complexity clumps. The `seg` program, which masks out these regions, removing them from the similarity score calculation, can reduce the effect of clumps with biased composition, but not eliminate it. Fortunately, the deviation from *i.i.d* is modest in 80% of protein sequences. Other than the biased composition effect, no other property of “real” protein sequences has been identified that distinguishes them from sequences built from picking amino-acids from a probability distribution at random.

Summary—exploiting statistical estimates

The inference of homology from statistically significant sequence similarity is one of the most reliable conclusions a scientist can draw. Indeed, the vast majority of bacterial, *C. elegans*, and *Drosophila* genes are annotated largely on the basis of statistically significant sequence similarity shared by other proteins with known structures or functions. This trend is certain to continue as sequence databases become more comprehensive.

While the inference of homology from significant sequence similarity is reliable—sequences that share much more similarity than expected by chance share a common ancestor—the inference tells us much more about structure than function. Without exception, sequences that share statistically significant similarity share significant structural similarity. However, homologous proteins need not perform the same, or even similar functions. Functional inferences are most reliable when based on assignments of *orthology*. *Orthologous* sequences are sequences that differ because of species differences. This contrasts with *paralogous* sequences, which are produced by gene duplication events. While *homology* can be demonstrated by sequence similarity, an inference of *orthology* is best supported by phylogenetic analysis, which is considerably more challenging computationally. In addition, many proteins are built from evolutionarily independent domains with different structures and functions. The inference of homology is transitive—if protein A is homologous to B and B is homologous to C, even if A and C do not share significant similarity—but it is critical that such inferences be limited to the domain to which they apply. There is great concern that incorrect functional assignments are greatly reducing the value of sequence database annotations because functional assignments are inappropriately extended to new family members based on a correct, but functionally uninformative, inference of homology.

Statistical significance estimates, whether as expectation ($E()$)-values or bit scores, are far more informative than the most commonly used measure of sequence similarity, percent identity. It has been known for more that 20 years (Dayhoff *et al.*, 1978) that percent identity is much less effective than measures of similarity that distinguish biochemically similar and dissimilar amino acids, and that recognize that some amino-acids mutate far more rapidly than others. Moreover, high sequence identity is expected over very short regions by chance in unrelated sequences that share no structural similarity (Kabsch & Sander, 1984). Thus, the inference of homology should always be based on statistically significant sequence similarity using an appropriate scoring matrix (Altschul, 1991).

However, once homology has been established, measures of statistical significance are not good measures of evolutionary distance. Two sequences that have diverged by the same amount, and thus share the same average levels of sequence similarity, can have very different similarity scores, with very different levels of statistical significance, depending on their lengths. For example, two members of the orotate phosphoribosyltransferase family, `pyre_colgr` and `pyre_klula` that share 48.5% identity over 223 amino-acid residues have similarity scores $S_{bit} = 161$ with $E() < 10^{-39}$, while two members of the 2-times longer glucose transporter family with slightly lower identity (47.4% over 502 amino-acids) obtain a similarity score of 308 bits with $E() < 10^{-82}$. Thus, similarity scores and expectation values must be adjusted when comparing among different length protein sequences if they are used as surrogates for evolutionary divergence.

This review of sequence similarity statistics has focused on protein sequence comparison for two reasons: (1) protein sequence comparison is far more sensitive than DNA sequence comparison—the evolutionary look-back time for protein sequences is typically 5–10-times greater than that for DNA sequences (Pearson, 1997). Moreover, protein databases are more compact, so that more rigorous algorithms can be used for similarity searching. (2) In addition, DNA sequences are well known to have higher-order sequence dependence due to codon bias and simple-sequence repeat regions. Because of the small nucleotide alphabet and the possible translation of normal complexity DNA sequences into low complexity protein sequences, it is much more difficult to detect and correct for deviations from *i.i.d* in DNA sequences. Thus, in general, statistical estimates from protein sequence comparisons are more reliable than the similar comparisons with DNA.

Our understanding of the statistical properties of biological sequences has improved dramatically over the past decade, so that most sequence similarity searching methods now include reliable statistical estimates. However, there is still room for improvement, as more searches are done with more complex queries e.g. profiles, position specific scoring matrices (Altschul *et al.*, 1997), and three-dimensional sequence-structure alignments, whose statistical properties on real sequences are not well understood. Fortunately, there is no shortage of data that can be used to develop and validate new statistical approaches.

Acknowledgements

We thank Stephen Altschul for his critical reading of this chapter and his helpful explanations and comments.

References

Allison, T. J., Wood, T. C., Briercheck, D. M., Rastinejad, F., Richardson, J. P. & Rule, G. S. (1998). Crystal structure of the RNA-binding domain from transcription termination factor ρ. *Nat. Struc. Biol.* **5**, 352–356.

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

Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460–480.

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.

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

Andrade, M., Casari, G., de Daruvar, A., Sander, C., Schneider, R., Tamames, J., Valencia, A. & Ouzounis, C. (1997). Sequence analysis of the *Methanococcus jannaschii* genome and the prediction of protein function. *Comput. Appl. Biosci.* **13**, 481–483.

Arratia, R., Gordan, L. & Waterman, M. S. (1990). The Erdos-Renyi law in distribution, for coin tossing and sequence matching. *Ann. Stat.* **18**, 539–570.

Arratia, R., Gordon, L. & Waterman, M. S. (1986). An extreme value theory for sequence matching. *Ann. Stat.* **14**, 971–993.

Bairoch, A. & Apweiler, R. (1996). The Swiss-Prot protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids. Res.* **24**, 21–25.

Brenner, S. E., Chothia, C. & Hubbard, T. J. (1997). Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol,* **7**, 369–376.

Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA,* **95**, 6073–6078.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sulton, G. G., Blake, J. A., Fitzgerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J.-F., Adams, M. D., Reisch, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H.-P., Fraser, C. M., Smith, H. O., Woese, C. R. & Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science,* **273**, 1058–1073.

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

Evans, M., Hastings, N. & Peacock, B. (1993). *Statistical Distributions, 2nd ed.* Wiley, New York, NY.

Fitch, W. M. (1966). An improved method of testing for evolutionary homology. *J. Mol. Biol.* **16**, 9–16.

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

Huang, X., Hardison, R. C. & Miller, W. (1990). A space-efficient algorithm for local similarities. *Comp. Appl. Biosci.* **6**, 373–381.

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.

Kabsch, W. & Sander, C. (1984). On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. USA,* **81**, 1075–1078.

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., Bucher, P., Brendel, V. & Altschul, S. F. (1991). Statistical methods and insights for protein and DNA sequences. *Annu. Rev. of Biophys. Biophys. Chem.* **20**, 175–203.

Kent, G. C. (1992). *Comparative Anatomy of the Vertebrates*. Mosby Year Book, Inc., St. Louis.

Koonin, E. V. (1997). Big time for small genomes. *Genome Res,* **7**, 418–421.

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.

Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA,* **95**, 5913–5920.

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

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.

Neurath, H., Walsh, K. A. & Winter, W. P. (1967). Evolution of structure and function of proteases. *Science,* **158**, 1638–1644.

Orengo, C. A., Swindells, M., Michie, A., M.J. Zvelebil, P. D., Waterfield, M. & Thornton, J. (1995). Structural similarity between the pleckstrin homology domain and verotoxin: The problem of measuring and evaluating structural similarity. *Prot. Sci.* **4**, 1977–1983.

Owen, R. (1843). *Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals*. Longman, Brown, and Green Co., London.

Owen, R. (1866). *On the Anatomy of Vertebrates*. Longmans, Green and Co., London.

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

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227–258.

Pearson, W. R. (1997). Identifying distantly related protein sequences. *Comp. Appl.Biosci.* **13**, 325–332.

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

Pearson, W. R. (1999). Flexible similarity searching with the FASTA3 program package. In *Bioinformatics Methods and Protocols*, (Misener, S. & Krawetz, S. A., eds), pp. 185–219. Humana Press Totowa, NJ.

Rawlings, N. D. & Barrett, A. (1993). Evolutionary families of peptidases. *Biochem. J.* **290**, 205–218.

Sanderson, M. & Hufford, L., eds (1996). *Homoplasy: The Recurrence of Similarity in Evolution*. Academic Press, New York.

Schwartz, R. M. & Dayhoff, M. (1978). Matrices for detecting distant relationships. In *Atlas of Protein Sequence and Structure*, (Dayhoff, M., ed.), vol. 5, supplement 3, pp. 353–358. National Biomedical Research Foundation Silver Spring, MD.

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

Sonnhammer, E. L., Eddy, S. R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins,* **28**, 405–420.

States, D. J., Gish, W. & Altschul, S. F. (1991). Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *METHODS: A companion to Methods in Enzymology,* **3**, 66–70.

Waterman, M. S. (1995). *Introduction to computational biology*. Chapman and Hall, London.

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.

Waterman, M. S. & Vingron, M. (1994). Rapid and accurate estimates of the statistical significance for sequence database searches. *Proc. Natl. Acad. Sci. USA,* **9192**, 4625–4628.

Watson, H. C. & Kendrew, J. (1961). Comparison between the amino-acid sequences of sperm whale myoglobin and of human haemoglobin. *Nature,* **190**, 670–672.

Wootton, J. C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computers and Chemistry,* **18**, 269–285.