Sequence Similarity

Jaap Heringa, *VU University Amsterdam, Amsterdam, The Netherlands*

Advanced article

Article Contents

- Comparative Sequence Analysis
- Sequence Alignment
- Statistics of Alignment Similarity Scores

Online posting date: 15th July 2008

Sequence similarity is a measure of an empirical relationship between sequences. A common objective of sequence similarity calculations is establishing the likelihood for sequence homology: the chance that sequences have evolved from a common ancestor. A similarity score is therefore aimed to approximate the evolutionary distance between a pair of nucleotide or protein sequences. Many implementations for measuring sequence similarity exist, where a general aim is to infer structural or functional characteristics of an unannotated molecular sequence.

Comparative Sequence Analysis

Comparative sequence analysis is a common first step in the analysis of sequence–structure–function relationships in protein and nucleotide sequences. In the quest for knowledge about the role of a certain unknown protein in the cellular molecular network, comparing the query sequence with the many sequences in annotated protein sequence databases often leads to useful suggestions regarding the protein's three-dimensional (3D) structure or molecular function. This extrapolation of the properties of sequences in public databases that are identified as 'neighbours' by sequence analysis techniques has arguably led to the putative characterization (annotation) of more sequences than any other single technology during the last three decades. Although progress has been made, the direct prediction of a protein's structure and function is still a major unsolved problem in molecular biology. Since the advent of the genome sequencing projects and subsequent rapid expansion of sequence databases, the method of indirect inference of molecular function by comparative sequence techniques has only gained in significance. Many current research projects are aiming to improve the sensitivity of sequence comparison techniques, which requires high-performance computing given the current and rapidly growing database sizes.

Sequence Alignment

Although many properties of nucleotide or protein sequences can be used to derive a similarity score, for example, nucleotide or amino acid composition, isoelectric point or molecular weight, the vast majority of sequence similarity calculations are based upon an alignment between two sequences from which a similarity score is inferred. Ideally, the alignment matches the nucleotide or amino acid sequences from either sequence according to their evolutionary descent from a common ancestor, with conserved and corresponding mutated residues at matched positions and inserted/deleted fragments intervening at proper sequence positions. Often, however, evolution has led to very widely diverged sequences such that at the primary sequence level the ancestral ties have become blurred beyond recognition, leading in many cases to biologically incorrect alignments. Another confounding issue is the fact that an increasing number of cases are identified of non-orthologous displacement, where enzymes carrying out an identical function in different organisms belong to entirely different protein families, and thus are not expected to show any sequence similarity. Examples of non-orthologous displacement include ornithine decarboxylase in *Escherichia coli* and *Saccharomyces cerevisiae*, where the isozymes speF and speC are responsible for this function in *E. coli* and share the same structure comprising three domains (ornithine decarboxylase *N*-terminal 'wing' domain, pyridoxal phosphate (PLP)-dependent transferase and ornithine decarboxylase *C*-terminal domain), whereas the corresponding enzyme spe1 in *S. cerevisiae* is a two-domain protein with entirely different domain structures (PLP-binding barrel and alanine racemase-like domain). In general, sequence alignment techniques are aimed at recognizing divergent evolution by mutation, including changes of gene structure by gene fusion or fission. However, the techniques are not able to trace evolutionary cases of horizontal gene transfer or functional displacement of one gene by another within a genome. **See also**: Alignment: Statistical Significance; Evolutionary Distance; Evolution: Convergent and Parallel Evolution; Gene Fusion; Northern Blotting and RNA Detection; Proteins: Mutational Effects in

Techniques for pairwise alignment

Many methods for the calculation of sequence alignments have been developed, of which implementations of the dynamic programming (DP) algorithm (Needleman and

ELS subject area: Evolution and Diversity of Life

How to cite:
Heringa, Jaap (July 2008) Sequence Similarity. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0005317.pub2

Wunsch, 1970; Smith and Waterman, 1981) are considered the standard in yielding the most biologically relevant alignments. The DP algorithm requires a *scoring matrix*, which is an evolutionary model in the form of a symmetrical 4×4 nucleotide or a 20×20 amino acid exchange matrix. Each matrix cell approximates the evolutionary propensity for the mutation of one nucleotide or amino acid type into another. The DP algorithm also relies on the specification of *gap penalties*, which model the relative probabilities for the occurrence of insertion/deletion events. Normally, a gap opening and extension penalty is used for creating a gap and each extension, respectively (*affine* gap penalties), so that the chance for an insertion/deletion depends linearly upon the length of the associated fragment. Given an exchange matrix and gap penalty values, which together are commonly called the *scoring scheme*, the DP algorithm is guaranteed to produce the highest scoring alignment of any pair of sequences: the *optimal alignment*. **See also**: Dynamic Programming; Substitution Matrices

Two main types of alignment are generally distinguished: global and local alignment. *Global alignment* (Needleman and Wunsch, 1970) denotes an alignment over the full length of both sequences, which is an appropriate strategy to follow when two sequences are similar or have roughly the same length. A special form of global alignment is *semiglobal alignment*, where alignment extends over full-length sequences, but with zero gap cost for end gaps preceding or following a sequence. This is an appropriate strategy for cases such as, for example, aligning a single gene against a genome, or a single-domain protein sequence against a multidomain protein sequence with one of its domains being homologous to the single-domain sequence. However, some sequences may show similarity limited to a motif or a domain only, while remaining sequence stretches may be essentially unrelated. In such cases, global alignment may well misalign the related fragments, as these become overshadowed by the unrelated sequence portions that the global method attempts to align, possibly leading to a score that would not allow the recognition of any similarity. An appropriate technique that addresses this issue is *local alignment* (Smith and Waterman, 1981). The first method for local alignment, often referred to as the Smith–Waterman (SW) algorithm (Smith and Waterman, 1981), is in fact a minor modification of the DP algorithm for global alignment. The algorithm selects the best-scoring subsequence from each sequence and provides their alignment, thereby disregarding the remaining sequence fragments. If no knowledge about the relationship of two sequences is available, it is generally advisable to align selected fragments of either sequence that have retained most of the evolutionary signal. Later elaborations of the local alignment algorithm include methods that generate a number of suboptimal local alignments in addition to the optimal pairwise alignment (e.g. Waterman and Eggert, 1987). **See also**: Global Alignment; Sequence Alignment; Smith–Waterman Algorithm

Calculating alignment scores

Since the DP algorithm essentially models the alignment of two sequences as a Markov process, where the amino acid matches are considered independent, the product of the probabilities for each match within an alignment should be taken. Since many of the scoring matrices contain exchange propensities converted to logarithmic values (*log-odds*), the alignment score can be calculated by summing the log-odd values corresponding to matched residues minus appropriate gap penalties:

$$S_{a,b} = \sum_{l} s(a_i, b_j) - \sum_{k} N_k \mathrm{gp}(k)$$

where the first summation is over the exchange values associated with l matched residues and the second is over each group of gaps of length k, with N_k the number of gaps of length k and $\mathrm{gp}(k)$ the associated gap penalty. In case affine gap penalties are used (see above), $\mathrm{gp}(k) = p_o + kp_e$, where p_o and p_e are the penalties for gap opening and extension, respectively. A consequence of the widely used affine gap penalty scheme is that long gaps required, for example, to span an inserted domain B in aligning a two-domain sequence AC (where A and C represent domains) with a three-domain sequence ABC, are often too costly, so that such sequences become misaligned.

The method Ngila (Cartwright, 2007) allows the application of biologically more realistic gap penalties of the form $\mathrm{gp}(k) = p_o + kp_e + p_c \ln k$, where p_o, p_e and p_c are the penalties for gap opening, extension and concave extension, respectively. Setting p_e to zero yields the so-called *concave* gap penalty regime, which is more amenable to inserting large gaps than the affine scheme.

Sequence database searching

A typical application to infer knowledge for a given query sequence is to compare it with all sequences in an annotated sequence database. Unfortunately, the DP algorithm is too slow for repeated searches over large databases, and may take multiple CPU hours for a single query sequence on a standard workstation. Although some special hardware has been designed to accelerate the DP algorithm, this problem has triggered the development of several heuristic algorithms that represent shortcuts to speed up the basic alignment procedure. Typically, these algorithms employ a quick scanning step to discard putatively unrelated database sequences. The relatively few remaining sequences are then subjected to a more rigorous comparison to assess their biological relatedness.

Fast heuristic methods

The homology searching tool FASTA (Pearson and Lipman, 1988) was the first heuristic for fast sequence comparison. The method incorporates a quick filtering step that approximates local alignment using hashing, a computational indexing technique. Database sequences that are retained are then locally aligned to the query sequence

using full DP. The popular BLAST suite (Altschul *et al*., 1990) includes a number of routines for searching of all combinations of nucleotide or protein sequences against a nucleotide or protein database. For example, it enables the routine searching of a protein query sequence against an annotated nonredundant (NR) database of about 4 million protein sequences. The currently most widely used method to scour sequence databases for homologies, PSI-BLAST (Altschul *et al*., 1997), is an extension of the BLAST technology in that it can iteratively rescan the sequence database. It does so by means of a profile representing the query sequence that is constructed using database hit sequences from a preceding iteration. A number of updates to the PSI-BLAST technique were effected (Schäffer *et al*., 2001), the most significant of which was making the position-specific scoring matrix (PSSM) in PSI-BLAST take compositional biases of the query and database sequences into account. Przybylski and Rost (2007) extended the PSI-BLAST technology by regularizing the NR database. In their approach, each database sequence is replaced by a consensus sequence based upon neighbouring database sequences. Thus the constructed consensus database led to improved PSI-BLAST searches, whereas the PSI-BLAST method itself did not need any modifications.

A computational effective adaptation of the BLAST technique is the method BLAT, the BLAST-like alignment tool (Kent, 2002). BLAT performs rapid messenger ribonucleic acid/deoxyribonucleic acid (mRNA/DNA) and cross-species protein alignments. Using a different technology it is about 500 times faster than BLAST when used for mRNA/DNA alignments and 50 times faster for protein alignments at sensitivity settings typically used when comparing vertebrate sequences. When BLAT is applied to DNA sequences, an index of the entire genome with a size of approximately 1 GB is built up in memory. The index consists of all nonoverlapping 11-mers except for those heavily involved in repeats. It is used to delineate putatively homologous areas, which are then loaded into memory for further detailed alignment. The BLAT routine for protein sequences works in a similar manner but uses 4-mers rather than 11-mers. Although the protein index generated is larger than that for DNA, it can be stored in memory on modern workstations.

The standard implementation of BLAT quickly finds sequences of 95% similarity or more, and length ≥ 40 bases when applied to DNA. However, more divergent regions at longer length or more similar regions at shorter lengths can be missed. When applied to search protein sequences, BLAT finds sequences $\geq 80\%$ sequence identity and of length 20 amino acids or more. In practice, DNA BLAT can be applied for homology searches covering primate genomes, whereas protein BLAT is amenable to land vertebrates.

Extended sequence comparison techniques

Owing to advances in computational performance, procedures for homology searching have been developed based on more computationally intense formalisms such as the hidden Markov modelling-based tools SAM-T2K (Kevin *et al*., 2001; Karplus *et al*., 1998) and HMMER2 (Eddy, 1998). Other improvements to the alignment of distant sequences have been achieved using alternative approaches. Yu *et al*. (2003) showed that the use of organism-specific or alignment set-specific background frequencies for contextual readjustment of the standard amino acid exchange weights provide a more sensitive and biologically accurate way to align sequences. Alternatively, structural and/or homologous sequence information can be incorporated into the alignment process to help identify the distant relations between sequences. The advantage of using related sequence information has led to numerous profile–profile alignment methods that apply different profile-scoring schemes (e.g. Yona and Levitt, 2002; Edgar and Sjölander, 2004; von Ohsen *et al*., 2004; Tomii and Akiyama, 2004; Jaroszewski *et al*., 2005; Simossis *et al*., 2005). Most of these profile–profile alignment approaches have mainly been used for sequence database searching (local pairwise alignment), where a popular application has been to use profile–profile comparisons for aligning a profile derived from a query multiple alignment with a number of profiles describing a collection of different protein families. **See also**: BLAST Algorithm; FASTA Algorithm; Hidden Markov Models; Multiple Alignment; Profile Searching

Databases and annotation

A database search can be performed for a nucleotide or amino acid sequence against an annotated database of nucleotide (e.g. EMBL, GenBank, DDBJ) or protein sequences (e.g. SwissProt, PIR, TrEMBL, GenPept, NR-NCBI, NR-ExPasy). Also the GSS, EST, STS or HTGS nucleotide databases can be scrutinized to find homologies, gain insight into expression data or locate a gene on the genome map. The NR-NCBI database is compiled by the NCBI (National Center for Biotechnology Information) as an NR protein sequence database for BLAST searches. It contains a total of about 4 million nonidentical sequences from GenBank CDS translations, PDB, Swiss-Prot, PIR and PRF.

The success of sequence similarity searches depends crucially on the quality and coverage of the sequence database used. The quality of the data, especially nucleic acid sequences, has improved as a result of modern sequencing techniques, while the amount of raw sequence data is increasing very rapidly. However, functional characterization is also critically reliant on the annotation of the data. Since inferring and experimentally determining the annotations represent a bottleneck in the generation of the data, there is a rapidly widening gap between sequence and annotation data. This is reflected in the fact that many sequences have 'unknown' as functional annotation, whereas an increasing number of sequences, especially those originating from bacterial genomes, feature annotations such as 'conserved hypothetical'. Conserved hypothetical database sequences have homologues, usually in other organisms, but none of

these homologues have known functions. Nonetheless, if conserved hypothetical homologues exist, this provides confidence that the sequence considered is truly a gene. **See also**: Comparative Human Genomics

Comparing DNA or protein sequences

The actual pairwise comparison of sequences can take place at the nucleotide or peptide level. However, the most effective way to compare sequences is at the peptide level (Pearson, 1996), which requires that nucleotide sequences must first be translated in all six reading frames followed by comparison with each of these conceptual protein sequences. Although mutation, insertion and deletion events take place at the DNA level, there are several reasons why comparing protein sequences can reveal more distant relationships: (1) Many mutations within DNA are synonymous, which means that these do not lead to a change of the corresponding amino acids. As a result of the fact that most evolutionary selection pressure is exerted on protein sequences, synonymous mutations can lead to an overestimation of the sequence divergence if compared at the DNA level. (2) The evolutionary relationships can be more finely expressed using a 20×20 amino acid exchange table than using exchange values among four nucleotides. (3) DNA sequences contain noncoding regions, which should be avoided in homology searches. Note that the latter is still an issue when using DNA translated into protein sequences through a codon table. However, a complication arises when using translated DNA sequences to search at the protein level because frameshifts can occur, leading to stretches of incorrect amino acids in the wrongly transcribed product and possible elongation of sequences due to missed stop codons. On the contrary, frameshifts typically result in stretches of highly unlikely and distant amino acids, which can be used as a signal to trace their occurrence. **See also**: Mutation Rates: Evolution

Similarity versus homology

Many times the term 'homologous sequence' is used when in fact a sequence should only be referred to as similar to a given reference sequence (May, 2001). Whereas sequence similarity is a quantification of an empirical relationship of sequences expressed using a gradual scale, the term 'homology' denotes an inference in that the presence of a common ancestor between the sequences and hence divergent evolution is assumed, leading to orthologous genes. This means that homology is a qualitative state; that is, a pair of sequences is homologous or not. As protein tertiary structures are more conserved during evolution than their coding sequences, homologous sequences are assumed to share the same protein fold. Although it is possible in theory that two proteins evolve different structures and functions from a common ancestor, this situation cannot be traced so that such proteins are seen as unrelated. However, numerous cases exist of homologous protein families where subfamilies with the same fold have evolved distinct molecular functions. The term 'homology' is often used in practice when two sequences have the same structure or function, although in the case of two sequences sharing a common function this ignores the possibility that the sequences are analogues resulting from convergent evolution, now often referred to as nonorthologous displacement (see earlier). Unfortunately, it is not straightforward to infer homology from similarity as enormous differences exist between sequence similarities within homologous families. Many protein families of common descent comprise members that share pairwise sequence similarities, which are only gradually higher than those observed between unrelated proteins. This region of uncertainty has been characterized to lie in the range 15–25% sequence identity (Doolittle, 1981) (see later), and is commonly referred to as the 'twilight zone'. There are even some known examples of homologous proteins with sequence similarities below the randomly expected level given their amino acid composition (Pascarella and Argos, 1992). As a consequence, it is impossible to prove using sequence similarity that two sequences are not homologous. **See also**: Protein Homology Modeling

The similarity score for two sequences can be calculated from their alignment using the above formula for $S_{a,b}$, such that it depends on the actual scoring matrix and gap penalties used. Sequence similarity has also been calculated as a fraction of a maximal score possible for two sequences using a normalized scoring matrix and by normalizing the raw alignment score by the length of the shorter sequence (Abagyan and Batalov, 1997).

Sequence similarity versus identity

Numerous studies into protein sequence relationships evaluate sequence alignments using a simple binary scheme of matched positions being identical or nonidentical. Sequence identity is normally expressed in the percentage identical residues found in a given alignment, where normalization can be performed using the length of the alignment or the shorter sequence. The scheme is simple and does not rely on an amino acid exchange matrix. However, if two proteins are said to share a given percentage in sequence identity, this is based on a sequence alignment, which will have been almost always constructed using an amino acid exchange matrix and gap penalty values, so that sequence identity cannot be regarded as being independent of sequence similarity. Using sequence identity as a measure, Sander and Schneider (1991) estimated that if two protein sequences are longer than 80 residues, they could relatively safely be assumed to be homologous whenever their sequence identity is 25% or more. Another commonly used notion is that if two sequences share more than 50% sequence identity, their enzymatic function will be the same (Rost, 2002). Contrary to this notion, however, it has been estimated that 70% of pair fragments with above 50% sequence identity might not have a completely identical function (Rost, 2002). An example is *Bacillus subtilus* exodeoxyribonuclease (SwissProt code exoa_bacsu) and rat

DNA-lyase (SwissProt code ape1_rat), where the sequences share 57% identity over 122 alignment positions, leading to a very significant BLAST E-value of 1.6×10^{-96}, but yet fulfil different functions (DNA degradation and repair, respectively). Despite its popularity and use in empirical rules as above, the use of sequence identity percentages is not optimal for homology searches (Abagyan and Batalov, 1997). As a result no major sequence comparison methods employ sequence identity scores in deriving statistical significance estimates. **See also**: Gene Families

Statistics of Alignment Similarity Scores

Sequence alignment methods are essentially pattern search techniques, leading to an alignment with a similarity score even in case of absence of any biological relationship ('garbage in, garbage out'). Although similarity scores of unrelated sequences are essentially random, they can behave like 'real' scores and, for example, like the latter are correlated with the length of the sequences compared. Particularly in the context of database searching, it is important to know what scores can be expected by chance and how scores that deviate from random expectation should be assessed. If within a rigid statistical framework a sequence similarity is deemed statistically significant, this provides confidence in deducing that the sequences involved are in fact biologically related. As a result of the complexities of protein sequence evolution and distant relationships observed in nature, any statistical scheme will invariably lead to situations where a sequence is assessed as unrelated while it is in fact homologous (*false negative*), or the inverse, where a sequence is deemed homologous while it is in fact biologically unrelated (*false positive*). A relatively frequent cause of erroneous transfer of annotation is based on similarity found over relatively short sequence regions and/or similarity based on different domains in multidomain structures (Rost, 2002).

The derivation of a general statistical framework for evaluating the significance of sequence similarity scores has been a major task. However, a rigid framework has not been established for global alignment, and has only partly been completed for local alignment.

Expected values for global similarity scores

Sequence similarity values resulting from global alignments are known to grow linearly with the sequence length (**Figure 1**), although the growth rate has not been determined. Also, the exact distribution of global similarity scores is yet unknown, and only numerical approximations exist, providing only a rough bound on the expected random scores. As the variance of the global similarity score has not been determined either, most applications derive a sense of the score by using shuffled sequences. Shuffled sequences retain the composition of a given real sequence

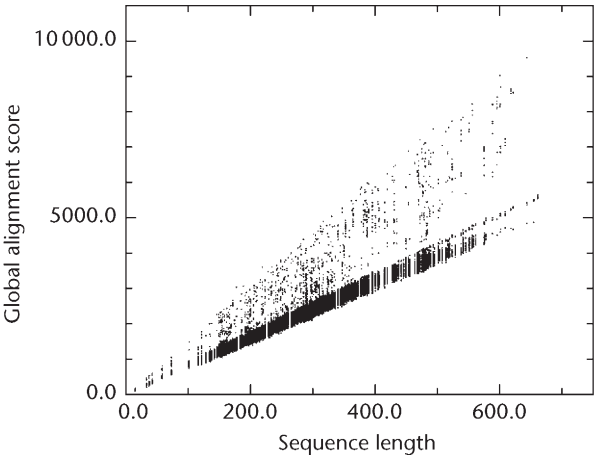

Figure 1 Distribution of 66 066 global similarity scores, derived from pairwise global alignments over an artificial database of sequences derived using a random mutation and insertion/deletion protocol, versus the length of the shortest sequence in each pairwise alignment. The alignments were effected using the PRALINE method (Heringa, 1999, 2002), where the alignment scores were calculated using the BLOSUM62 matrix and gap penalty values of 12 and 1 for gap initiation and extension, respectively. A clearly linear lower band of alignment scores of unrelated sequences is visible. The correlation coefficient of the random scores within the lower band is 0.99 while the slope of the regression line is 7.864. Also the higher scores of putatively related sequences above the lower band are correlated: the correlation coefficient is 0.98 and the linear regression line slope is 12.50. Random and real scores were separated by the line $y = 9.334x$.

but have a permuted order of nucleotides or amino acids. The distribution of similarity scores over a large number of such shuffled sequences often approximates the shape of the Gaussian distribution, which is therefore taken to represent the underlying random distribution. Using the mean (m) and standard deviation (σ) calculated from such shuffled similarity scores, each real score S can be converted to the z-score using

$$z - \text{score} = \frac{S - m}{\sigma}$$

The z-score measures how many standard deviations the score is separated from the mean of the random distribution. In many studies, a z-score > 6 is taken to indicate a significant similarity.

Expected values for local similarity scores

Statistics of local alignments without gaps

A rigid statistical framework for local alignments without gaps has been derived for protein sequences following the work by Karlin and Altschul (1990), who showed that the optimal local ungapped alignment score grow linearly with the logarithm of the product of sequence lengths of two considered random sequences:

$$S \sim \frac{\ln(n \cdot m)}{\lambda}$$

where n and m are the lengths of two random sequences, and λ a scaling parameter that depends on the scoring

matrix used and the overall distribution of amino acids in the database. Specifically, λ is the unique solution for x in the equation

$$\sum_{i,j} P_i P_j e^{S_{i,j}x} = 1$$

where summation is done over all amino acid pairs, p_i represents the background probability (frequency) of residue type i and $s_{i,j}$ the scoring matrix.

An important contribution for fast sequence database searching has been the realization (Karlin and Altschul, 1990; Dembo and Karlin, 1991; Dembo *et al*., 1994) that local similarity scores of ungapped alignments follow the *extreme value distribution* (EVD) (Gumbel, 1958). The computational advantage of exploiting the EVD for statistical significance scores was used for the first time in BLAST (Altschul *et al*., 1990). This distribution is unimodal but not symmetrical like the normal distribution, because the right-hand tail at high scoring values falls off more gradually than the lower tail, reflecting the fact that a best local alignment is associated with a score that is the maximum out of a great number of independent alignments (**Figure 2**).

Following the EVD, the probability of a score S to be larger than a given value x can be calculated as

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

where $\mu = (\ln Kmn)/\lambda$, and K a constant that can be estimated from the background amino acid distribution and scoring matrix. (See Altschul and Gish (1996) for a collection of values for λ and K over a set of widely used scoring matrices.) Using the equation for μ, the probability for S becomes

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

In practice, the probability $P(S \geq x)$ is estimated using the approximation $1 - \exp(-e^{-x}) \approx e^{-x}$, which is valid for large values of x. This leads to a simplification of the equation for $P(S \geq x)$:

$$P(S \geq x) \approx e^{-\lambda(x-\mu)} = Kmne^{-\lambda x}$$

The lower the probability for a given threshold value x, the more significant the score S.

Despite the usefulness of the above statistical estimates in recognizing sequence similarity, it should be noted that they do not judge the distribution of similarity along the sequences, which is a crucial aspect in assessing homology, and can correspond to a single domain in a multidomain protein sequence or to a single motif within a domain.

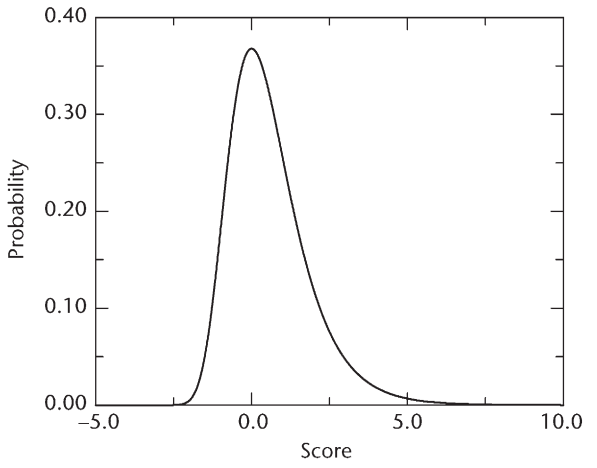

Figure 2 Probability density function for the extreme value distribution (EVD) resulting from parameter values $\mu = 0$ and $\lambda = 1$, where μ is the characteristic value and λ the decay constant.

Statistics of local alignments with gaps

Although similarities between sequences can be detected reasonably well using methods that do not allow insertions/deletions in aligned sequences, it is clear that insertion/deletion events play a major role in divergent sequences. This means that accommodating gaps within alignments of distantly related sequences is important for obtaining an accurate measure of similarity. Unfortunately, a rigorous statistical framework as obtained for gapless local alignments has not been conceived for local alignments with gaps. However, although it has not been proven analytically that the distribution of S for gapped alignments can be approximated with the EVD, there is accumulated evidence that this is the case. For example, for various scoring matrices, gapped alignment similarities have been observed to grow logarithmically with the sequence lengths (Arratia and Waterman, 1994). Other empirical studies have shown it to be likely that the distribution of local gapped similarities follows the EVD (Smith *et al*., 1985; Waterman and Vingron, 1994), although an appropriate downward correction for the effective sequence length has been recommended (Altschul and Gish, 1996). The distribution of empirical similarity values can be obtained from unrelated biological sequences (Pearson, 1998). Fitting of the EVD parameters λ and K (see earlier) can be performed using a linear regression technique (Pearson, 1998), although the technique is not robust against outliers which can have a marked influence. Maximum likelihood estimation (Mott, 1992; Lawless, 1982) has been shown to be superior for EVD parameter fitting and, for example, is the method used to parameterize the gapped BLAST method (Altschul *et al*., 1997). However, when low gap penalties are used to generate the alignments, the similarity scores can lose their local character and assume more global behaviour, such that the EVD-based probability estimates are not valid anymore (Arratia and Waterman, 1994).

Statistics of database searches

To be useful in sequence database searches, the above framework for comparing a pair of random sequences should be adapted to multiple pairwise comparisons. Here, it becomes important to establish the probability for a given query sequence to have a significant similarity with at least one of the database sequences. A *p-value* is the probability

of seeing at least one unrelated score S greater than or equal to a given score x in a database search over n sequences. This probability has been demonstrated to follow the Poisson distribution (Waterman and Vingron, 1994):

$$P(x, n) = 1 - e^{-nP(S \geq x)}$$

where n is the number of sequences in the database. In addition to the p-value, some database search methods employ the *expectation value* (or *E-value*) of the Poisson distribution, which is defined as the expected number of nonhomologous sequences with scores greater than or equal to a score x in a database of n sequences:

$$E(x, n) = nP(S \geq x)$$

For example, if the E-value of a matched database sequence segment is 0.01, then the expected number of random hits with score $S \geq x$ is 0.01, which means that this E-value is expected by chance only once in 100 independent searches over the database. However, if the E-value of a hit is five, then five fortuitous hits with $S \geq x$ are expected within a single database search, which renders the hit not significant. Database searching is commonly performed using an E-value in between 0.1 and 0.001. Low E-values decrease the number of false positives in a database search, but increase the number of false negatives such that the sensitivity (see below) of the search is lowered.

Evaluating sequence database searches

A few useful measures are commonly used to measure the accuracy of sequence database search methods over an annotated NR database. The *sensitivity* of a search is defined as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP is the number of true positives and FN the number of false negatives, which reflects the fraction of true hits found relative to the total number of sequences in the database that are homologous to the query. The sensitivity reflects to what extent the method is able to identify distantly related sequences. In many studies this measure is also referred to as *coverage*. The *specificity* (or *selectivity*) is defined as

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

which denotes the fraction of entries correctly excluded as hits, and hence measures the avoidance of unrelated hits. Yet another widely used measure is the *positive predictive value* (PPV), defined as

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

which measures the proportion of true homologues within all sequences designated by the search tool as related. In practical database searches, there is a trade-off between sensitivity and specificity: The more the p-values or E-values are relaxed to allow more distantly related sequences to be found, the more likely it becomes that chance hits infiltrate the search. Moreover, even if a very statistically significant similarity is encountered, problems remain. For example, if high similarity is found over only a portion of the sequences, the sequences may each contain multiple domains and share a single homologous domain only (see above), so that only an aspect of the overall function might be inferred. In iterative homology searches, often carried out using the aforementioned method PSI-BLAST (Altschul *et al*., 1997), protein sequences containing more than one structural domain can be problematic in that they cause the search to terminate prematurely or lead to an 'explosion' of common domains (George and Heringa, 2002). For example, the occurrence in the query sequence of a common and conserved protein domain such as the tyrosine kinase domain, which is then hit many times in the database, can obscure weaker but also relevant matches to other domain types (George and Heringa, 2002), particularly when the E-value is set to only include strong hits. Conversely, when multidomain sequences with the same sequential order of domains as in the query sequence are found initially during an iterative search, homologues with different domain combinations might well be missed due to early convergence of the search.

Low-complexity sequences

To reduce the chance of including spurious hits, some database search engines, such as PSI-BLAST (Altschul *et al*., 1997), scan query sequences for the presence of so-called *low-complexity regions* comprising biased residue compositions such as repeat-rich, coiled-coil or transmembrane regions (Wooton and Federhen, 1996). These are then excluded from alignment to limit the inclusion of false-positive hits due to database sequence matches with these regions. However, the occurrence of database sequences with low-complexity regions can still cause an explosion of false positives in iterative homology searches (George and Heringa, 2002). Sharon *et al*. (2005) presented a model that corrects BLAST E-values for low-complexity sequences without the need of complexity filtering.

Despite recent improvements of search techniques, complications such as above illustrate that automatic biological evaluation of homology searches in genomic pipelines remains elusive. **See also**: Alignment: Statistical Significance; Bioinformatics in Genome Sequencing Projects; Protein Homology Modeling; Sequence Alignment; Similarity Search

References

Abagyan RA and Batalov S (1997) Do aligned sequences share the same fold? *Journal of Molecular Biology* **273**: 355–368.

Altschul SF and Gish W (1996) Local alignment statistics. In: Doolittle RF (ed.) *Methods in Enzymology*, vol. 266, pp. 460–480. San Diego, CA: Academic Press.

Altschul SF, Gish W, Miller W, Meyers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Altschul SF, Madden TL, Schäffer AA *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Arratia R and Waterman MS (1994) A phase transition for the sore in matching random sequences allowing depletions. *Annals of Applied Probability* **4**: 200–225.

Cartwright RA (2007) Ngila: global pairwise alignments with logarithmic and affine gap costs. *Bioinformatics* **23**: 1427–1428.

Dembo A and Karlin S (1991) Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. *Annals of Probability* **19**: 1737.

Dembo A, Karlin S and Zeitouni O (1994) Limit distributions of maximal non-aligned two-sequence segmental score. *Annals of Probability* **22**: 2022.

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry. *Science* **214**: 149–159.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* **14**: 755–763.

Edgar RC and Sjölander K (2004) A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics* **20**: 1301–1308.

George RA and Heringa J (2002) Protein domain identification and improved sequence searching using PSI-BLAST. *Proteins – Structure Function and Genetics* **48**: 672–681.

Gumbel EJ (1958) *Statistics of Extremes*. New York, NY: Columbia University Press.

Heringa J (1999) Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Computers and Chemistry* **23**: 341–364.

Heringa J (2002) Local weighting schemes for protein multiple sequence alignment. *Computers and Chemistry* **26**: 459–477.

Jaroszewski L, Rychlewski L, Li Z, Li W and Godzik A (2005) FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Research* **33**: W284–W288.

Karlin S and Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the USA* **87**: 2264–2268.

Karplus K, Barrett C and Hughey R (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**: 846–856.

Kent WJ (2002) BLAT – The BLAST-like alignment tool. *Genome Research* **12**: 656–664.

Kevin K, Karchin R, Barrett C *et al*. (2001) What is the value added by human intervention in protein structure prediction? *Proteins: Structure, Function, and Genetics* **45**(S5): 86–91.

Lawless JF (1982) *Statistical Models and Methods for Lifetime Data* pp. 141–202. New York, NY: Wiley.

May AC (2001) Related problems. *Nature* **413**: 453.

Mott R (1992) Maximum-likelihood estimation of the statistical distribution of Smith–Waterman local sequence similarity scores. *Bulletin of Mathematical Biology* **54**: 59–75.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**: 443–453.

von Ohsen N, Sommer I, Zimmer R and Lengauer T (2004) Arby: automatic protein structure prediction using profile–profile alignment and confidence measures. *Bioinformatics* **20**: 2228–2235.

Pascarella S and Argos P (1992) A data bank merging related protein structures and sequences. *Protein Engineering* **5**: 121–137.

Pearson WR (1996) Effective protein sequence comparison. In: Doolittle RF (ed.) *Methods in Enzymology*, vol. 266, pp. 227–258. San Diego, CA: Academic Press.

Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. *Journal of Molecular Biology* **276**: 71–84.

Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA* **85**: 2444–2448.

Przybylski D and Rost B (2007) Consensus sequences improve PSI-BLAST through mimicking profile–profile alignments searches. *Nucleic Acids Research* **35**(7): 2238–2246.

Rost B (2002) Enzyme function is less conserved than anticipated. *Journal of Molecular Biology* **318**: 595–608.

Sander C and Schneider R (1991) Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins – Structure Function and Evolution* **9**: 56–68.

Schäffer AA, Aravind L, Madden TL *et al*. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research* **29**: 2994–3005.

Sharon I, Birkland A, Chang K, El-Yaniv R and Yona G (2005) Correcting BLAST e-values for low-complexity segments. *Journal of Computational Biology* **12**: 978–1001.

Simossis VA, Kleinjung J and Heringa J (2005) Homology-extended sequence alignment. *Nucleic Acids Research* **33**: 816–824.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* **147**: 195–197.

Smith TF, Waterman MS and Burks C (1985) The statistical distribution of nucleic acid similarities. *Nucleic Acids Research* **13**: 645.

Tomii K and Akiyama Y (2004) FORTE: a profile–profile comparison tool for protein fold recognition. *Bioinformatics* **20**: 594–595.

Waterman MS and Eggert M (1987) A new algorithm for best subsequences alignment with applications to the tRNA–rRNA comparisons. *Journal of Molecular Biology* **197**: 723–728.

Waterman MS and Vingron M (1994) Rapid and accurate estimates of statistical significance for sequence data base searches. *Proceedings of the National Academy of Sciences of the USA* **91**: 4625.

Wooton JC and Federhen S (1996) Analysis of compositionally biased regions in sequence databases. In: Doolittle RF (ed.) *Methods in Enzymology*, vol. 266, pp. 554–571. San Diego, CA: Academic Press.

Yona G and Levitt M (2002) Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. *Journal of Molecular Biology* **315**: 1257–1275.

Yu YK, Wootton JC, Altschul SF *et al*. (2003) The compositional adjustment of amino acid substitution matrices. *Proceedings of the National Academy of Sciences of the USA* **100**: 15688–15693.

Further Reading

Doolittle RF (ed.) (1996) *Methods in Enzymology*, vol. 266, p. 711. San Diego, CA: Academic Press.

Higgins D and Taylor WR (eds) (2000) *Bioinformatics: Sequence, Structure and Databanks*, p. 249. Oxford: Oxford University Press.