http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

Introduction

To assess whether a given alignment constitutes evidence for homology, it helps to know how strong an alignment can be expected from chance alone. In this context, "chance" can mean the comparison of (i) real but non-homologous sequences; (ii) real sequences that are shuffled to preserve compositional properties [1-3]; or (iii) sequences that are generated randomly based upon a DNA or protein sequence model. Analytic statistical results invariably use the last of these definitions of chance, while empirical results based on simulation and curve-fitting may use any of the definitions.

The statistics of global sequence comparison

Unfortunately, under even the simplest random models and scoring systems, very little is known about the random distribution of optimal global alignment scores [4]. Monte Carlo experiments can provide rough distributional results for some specific scoring systems and sequence compositions [5], but these can not be generalized easily. Therefore, one of the few methods available for assessing the statistical significance of a particular global alignment is to generate many random sequence pairs of the appropriate length and composition, and calculate the optimal alignment score for each [1,3]. While it is then possible to express the score of interest in terms of standard deviations from the mean, it is a mistake to assume that the relevant distribution is normal and convert this *Z*-value into a *P*-value; the tail behavior of global alignment scores is unknown. The most one can say reliably is that if 100 random alignments have score inferior to the alignment of interest, the *P*-value in question is likely less than 0.01. One further pitfall to avoid is exaggerating the significance of a result found among multiple tests. When many alignments have been generated, e.g. in a database search, the significance of the best must be discounted accordingly. An alignment with *P*-value 0.0001 in the context of a single trial may be assigned a *P*-value of only 0.1 if it was selected as the best among 1000 independent trials.

The statistics of local sequence comparison

Fortunately statistics for the scores of local alignments, unlike those of global alignments, are well understood. This is particularly true for local alignments lacking gaps, which we will consider first. Such alignments were precisely those sought by the original BLAST database search programs [6].

A local alignment without gaps consists simply of a pair of equal length segments, one from each of the two sequences being compared. A modification of the Smith-Waterman [7] or Sellers [8] algorithms will find all segment pairs whose scores can not be improved by extension or trimming. These are called high-scoring segment pairs or HSPs.

To analyze how high a score is likely to arise by chance, a model of random sequences is needed. For proteins, the simplest model chooses the amino acid residues in a sequence independently, with specific background probabilities for the various residues. Additionally, the expected score for aligning a random pair of amino acid is required to be negative. Were this not the case, long alignments would tend to have high score

independently of whether the segments aligned were related, and the statistical theory would break down.

Just as the sum of a large number of independent identically distributed (i.i.d) random variables tends to a normal distribution, the maximum of a large number of i.i.d. random variables tends to an extreme value distribution [9]. (We will elide the many technical points required to make this statement rigorous.) In studying optimal local sequence alignments, we are essentially dealing with the latter case [10,11]. In the limit of sufficiently large sequence lengths *m* and *n*, the statistics of HSP scores are characterized by two parameters, *K* and *lambda*. Most simply, the expected number of HSPs with score at least *S* is given by the formula

$$E = Kmn\, e^{-\lambda S} \qquad (1)$$

We call this the *E*-value for the score *S*.

This formula makes eminently intuitive sense. Doubling the length of either sequence should double the number of HSPs attaining a given score. Also, for an HSP to attain the score *2x* it must attain the score *x* twice in a row, so one expects *E* to decrease exponentially with score. The parameters *K* and *lambda* can be thought of simply as natural scales for the search space size and the scoring system respectively.

Bit scores

Raw scores have little meaning without detailed knowledge of the scoring system used, or more simply its statistical parameters *K* and *lambda*. Unless the scoring system is understood, citing a raw score alone is like citing a distance without specifying feet, meters, or light years. By normalizing a raw score using the formula

$$S' = \frac{\lambda S - \ln K}{\ln 2} \qquad (2)$$

one attains a "bit score" *S'*, which has a standard set of units. The *E*-value corresponding to a given bit score is simply

$$E = mn\, 2^{-S'} \qquad (3)$$

Bit scores subsume the statistical essence of the scoring system employed, so that to calculate significance one needs to know in addition only the size of the search space.

P-values

The number of random HSPs with score >= *S* is described by a Poisson distribution [10,11]. This means that the probability of finding exactly *a* HSPs with score >=*S* is given by

$$e^{-E}\frac{E^{a}}{a!} \qquad (4)$$

where E is the E-value of S given by equation (1) above. Specifically the chance of finding zero HSPs with score >=S is e^{-E}, so the probability of finding at least one such HSP is

$$P = 1 - e^{-E} \qquad (5)$$

This is the P-value associated with the score S. For example, if one expects to find three HSPs with score >= S, the probability of finding at least one is 0.95. The BLAST programs report E-value rather than P-values because it is easier to understand the difference between, for example, E-value of 5 and 10 than P-values of 0.993 and 0.99995. However, when $E < 0.01$, P-values and E-value are nearly identical.

Database searches

The E-value of equation (1) applies to the comparison of two proteins of lengths m and n. How does one assess the significance of an alignment that arises from the comparison of a protein of length m to a database containing many different proteins, of varying lengths? One view is that all proteins in the database are *a priori* equally likely to be related to the query. This implies that a low E-value for an alignment involving a short database sequence should carry the same weight as a low E-value for an alignment involving a long database sequence. To calculate a "database search" E-value, one simply multiplies the pairwise-comparison E-value by the number of sequences in the database. Recent versions of the FASTA protein comparison programs [12] take this approach [13].

An alternative view is that a query is *a priori* more likely to be related to a long than to a short sequence, because long sequences are often composed of multiple distinct domains. If we assume the *a priori* chance of relatedness is proportional to sequence length, then the pairwise E-value involving a database sequence of length n should be multiplied by N/n, where N is the total length of the database in residues. Examining equation (1), this can be accomplished simply by treating the database as a single long sequence of length N. The BLAST programs [6,14,15] take this approach to calculating database E-value. Notice that for DNA sequence comparisons, the length of database records is largely arbitrary, and therefore this is the only really tenable method for estimating statistical significance.

The statistics of gapped alignments

The statistics developed above have a solid theoretical foundation only for local alignments that are not permitted to have gaps. However, many computational experiments [14-21] and some analytic results [22] strongly suggest that the same theory applies as well to gapped alignments. For ungapped alignments, the statistical parameters can be calculated, using analytic formulas, from the substitution scores and the background residue frequencies of the sequences being compared. For gapped alignments, these parameters must be estimated from a large-scale comparison of "random" sequences.

Some database search programs, such as FASTA [12] or various implementation of the Smith-Waterman algorithm [7], produce optimal local alignment scores for the comparison

of the query sequence to every sequence in the database. Most of these scores involve unrelated sequences, and therefore can be used to estimate *lambda* and *K* [17,21]. This approach avoids the artificiality of a random sequence model by employing real sequences, with their attendant internal structure and correlations, but it must face the problem of excluding from the estimation scores from pairs of related sequences. The BLAST programs achieve much of their speed by avoiding the calculation of optimal alignment scores for all but a handful of unrelated sequences. The must therefore rely upon a pre-estimation of the parameters *lambda* and *K*, for a selected set of substitution matrices and gap costs. This estimation could be done using real sequences, but has instead relied upon a random sequence model [14], which appears to yield fairly accurate results [21].

Edge effects

The statistics described above tend to be somewhat conservative for short sequences. The theory supporting these statistics is an asymptotic one, which assumes an optimal local alignment can begin with any aligned pair of residues. However, a high-scoring alignment must have some length, and therefore can not begin near to the end of either of two sequences being compared. This "edge effect" may be corrected for by calculating an "effective length" for sequences [14]; the BLAST programs implement such a correction. For sequences longer than about 200 residues the edge effect correction is usually negligible.

The choice of substitution scores

The results a local alignment program produces depend strongly upon the scores it uses. No single scoring scheme is best for all purposes, and an understanding of the basic theory of local alignment scores can improve the sensitivity of one's sequence analyses. As before, the theory is fully developed only for scores used to find ungapped local alignments, so we start with that case.

A large number of different amino acid substitution scores, based upon a variety of rationales, have been described [23-36]. However the scores of any substitution matrix with negative expected score can be written uniquely in the form

$$S_{ij} = \left(\ln \frac{q_{ij}}{p_i\, p_j}\right) / \lambda \qquad (6)$$

where the q_{ij}, called target frequencies, are positive numbers that sum to 1, the p_i are background frequencies for the various residues, and *lambda* is a positive constant [10,31]. The *lambda* here is identical to the *lambda* of equation (1).

Multiplying all the scores in a substitution matrix by a positive constant does not change their essence: an alignment that was optimal using the original scores remains optimal. Such multiplication alters the parameter *lambda* but not the target frequencies q_{ij}. Thus, up

to a constant scaling factor, every substitution matrix is uniquely determined by its target frequencies. These frequencies have a special significance [10,31]:

> A given class of alignments is best distinguished from chance by the substitution matrix whose target frequencies characterize the class.

To elaborate, one may characterize a set of alignments representing homologous protein regions by the frequency with which each possible pair of residues is aligned. If valine in the first sequence and leucine in the second appear in 1% of all alignment positions, the target frequency for (valine, leucine) is 0.01. The most direct way to construct appropriate substitution matrices for local sequence comparison is to estimate target and background frequencies, and calculate the corresponding log-odds scores of formula (6). These frequencies in general can not be derived from first principles, and their estimation requires empirical input.

The PAM and BLOSUM amino acid substitution matrices

While all substitution matrices are implicitly of log-odds form, the first explicit construction using formula (6) was by Dayhoff and coworkers [24,25]. From a study of observed residue replacements in closely related proteins, they constructed the PAM (for "point accepted mutation") model of molecular evolution. One "PAM" corresponds to an average change in 1% of all amino acid positions. After 100 PAMs of evolution, not every residue will have changed: some will have mutated several times, perhaps returning to their original state, and others not at all. Thus it is possible to recognize as homologous proteins separated by much more than 100 PAMs. Note that there is no general correspondence between PAM distance and evolutionary time, as different protein families evolve at different rates.

Using the PAM model, the target frequencies and the corresponding substitution matrix may be calculated for any given evolutionary distance. When two sequences are compared, it is not generally known a priori what evolutionary distance will best characterize any similarity they may share. Closely related sequences, however, are relatively easy to find even will non-optimal matrices, so the tendency has been to use matrices tailored for fairly distant similarities. For many years, the most widely used matrix was PAM-250, because it was the only one originally published by Dayhoff.

Dayhoff's formalism for calculating target frequencies has been criticized [27], and there have been several efforts to update her numbers using the vast quantities of derived protein sequence data generated since her work [33,35]. These newer PAM matrices do not differ greatly from the original ones [37].

An alternative approach to estimating target frequencies, and the corresponding log-odds matrices, has been advanced by Henikoff & Henikoff [34]. They examine multiple alignments of distantly related protein regions directly, rather than extrapolate from closely related sequences. An advantage of this approach is that it cleaves closer to observation; a disadvantage is that it yields no evolutionary model. A number of tests [13,37] suggest that the "BLOSUM" matrices produced by this method generally are superior to the PAM matrices for detecting biological relationships.

DNA substitution matrices

While we have discussed substitution matrices only in the context of protein sequence comparison, all the main issues carry over to DNA sequence comparison. One warning is that when the sequences of interest code for protein, it is almost always better to compare the protein translations than to compare the DNA sequences directly. The reason is that after only a small amount of evolutionary change, the DNA sequences, when compared using simple nucleotide substitution scores, contain less information with which to deduce homology than do the encoded protein sequences [32].

Sometimes, however, one may wish to compare non-coding DNA sequences, at which point the same log-odds approach as before applies. An evolutionary model in which all nucleotides are equally common and all substitution mutations are equally likely yields different scores only for matches and mismatches [32]. A more complex model, in which transitions are more likely than transversions, yields different "mismatch" scores for transitions and transversions [32]. The best scores to use will depend upon whether one is seeking relatively diverged or closely related sequences [32].

Gap scores

Our theoretical development concerning the optimality of matrices constructed using equation (6) unfortunately is invalid as soon as gaps and associated gap scores are introduced, and no more general theory is available to take its place. However, if the gap scores employed are sufficiently large, one can expect that the optimal substitution scores for a given application will not change substantially. In practice, the same substitution scores have been applied fruitfully to local alignments both with and without gaps. Appropriate gap scores have been selected over the years by trial and error [13], and most alignment programs will have a default set of gap scores to go with a default set of substitution scores. If the user wishes to employ a different set of substitution scores, there is no guarantee that the same gap scores will remain appropriate. No clear theoretical guidance can be given, but "affine gap scores" [38-41], with a large penalty for opening a gap and a much smaller one for extending it, have generally proved among the most effective.

Low complexity sequence regions

There is one frequent case where the random models and therefore the statistics discussed here break down. As many as one fourth of all residues in protein sequences occur within regions with highly biased amino acid composition. Alignments of two regions with similarly biased composition may achieve very high scores that owe virtually nothing to residue order but are due instead to segment composition. Alignments of such "low complexity" regions have little meaning in any case: since these regions most likely arise by gene slippage, the one-to-one residue correspondence imposed by alignment is not valid. While it is worth noting that two proteins contain similar low complexity regions, they are best excluded when constructing alignments [42-44]. The BLAST programs employ the SEG algorithm [43] to filter low complexity regions from proteins before executing a database search.

References

[1] Fitch, W.M. (1983) "Random sequences." J. Mol. Biol. 163:171-176. (PubMed)
[2] Lipman, D.J., Wilbur, W.J., Smith T.F. & Waterman, M.S. (1984) "On the statistical significance of nucleic acid similarities." Nucl. Acids Res. 12:215-226. (PubMed)
[3] Altschul, S.F. & Erickson, B.W. (1985) "Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage." Mol. Biol. Evol. 2:526-538. (PubMed)
[4] Deken, J. (1983) "Probabilistic behavior of longest-common-subsequence length." In "Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison." D. Sankoff & J.B. Kruskal (eds.), pp. 55-91, Addison-Wesley, Reading, MA.
[5] Reich, J.G., Drabsch, H. & Daumler, A. (1984) "On the statistical assessment of similarities in DNA sequences." Nucl. Acids Res. 12:5529-5543. (PubMed)
[6] 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. (PubMed)
[7] Smith, T.F. & Waterman, M.S. (1981) "Identification of common molecular subsequences." J. Mol. Biol. 147:195-197. (PubMed)
[8] Sellers, P.H. (1984) "Pattern recognition in genetic sequences by mismatch density." Bull. Math. Biol. 46:501-514.
[9] Gumbel, E. J. (1958) "Statistics of extremes." Columbia University Press, New York, NY.
[10] 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.(PubMed)
[11] Dembo, A., Karlin, S. & Zeitouni, O. (1994) "Limit distribution of maximal non-aligned two-sequence segmental score." Ann. Prob. 22:2022-2039.
[12] Pearson, W.R. & Lipman, D.J. (1988) Improved tools for biological sequence comparison." Proc. Natl. Acad. Sci. USA 85:2444-2448. (PubMed)
[13] Pearson, W.R. (1995) "Comparison of methods for searching protein sequence databases." Prot. Sci. 4:1145-1160. (PubMed)
[14] Altschul, S.F. & Gish, W. (1996) "Local alignment statistics." Meth. Enzymol. 266:460-480. (PubMed)
[15] Altschul, S.F., Madden, T.L., Schäffer, 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. (PubMed)
[16] Smith, T.F., Waterman, M.S. & Burks, C. (1985) "The statistical distribution of nucleic acid similarities." Nucleic Acids Res. 13:645-656. (PubMed)
[17] Collins, J.F., Coulson, A.F.W. & Lyall, A. (1988) "The significance of protein sequence similarities." Comput. Appl. Biosci. 4:67-71. (PubMed)
[18] Mott, R. (1992) "Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores." Bull. Math. Biol. 54:59-75.
[19] Waterman, M.S. & Vingron, M. (1994) "Rapid and accurate estimates of statistical significance for sequence database searches." Proc. Natl. Acad. Sci. USA 91:4625-4628. (PubMed)

[20] Waterman, M.S. & Vingron, M. (1994) "Sequence comparison significance and Poisson approximation." Stat. Sci. 9:367-381.
[21] Pearson, W.R. (1998) "Empirical statistical estimates for sequence similarity searches." J. Mol. Biol. 276:71-84. (PubMed)
[22] Arratia, R. & Waterman, M.S. (1994) "A phase transition for the score in matching random sequences allowing deletions." Ann. Appl. Prob. 4:200-225.
[23] McLachlan, A.D. (1971) "Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c-551." J. Mol. Biol. 61:409-424. (PubMed)
[24] Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) "A model of evolutionary change in proteins." In "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3 (ed. M.O. Dayhoff), pp. 345-352. Natl. Biomed. Res. Found., Washington, DC.
[25] Schwartz, R.M. & Dayhoff, M.O. (1978) "Matrices for detecting distant relationships." n "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3 (ed. M.O. Dayhoff), p. 353-358. Natl. Biomed. Res. Found., Washington, DC.
[26] Feng, D.F., Johnson, M.S. & Doolittle, R.F. (1984) "Aligning amino acid sequences: omparison of commonly used methods." J. Mol. Evol. 21:112-125. (PubMed)
[27] Wilbur, W.J. (1985) "On the PAM matrix model of protein evolution." Mol. Biol. Evol. 2:434-447. (PubMed)
[28] Taylor, W.R. (1986) "The classification of amino acid conservation." J. Theor. Biol. 119:205-218. (PubMed)
[29] Rao, J.K.M. (1987) "New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters." Int. J. Peptide Protein Res. 29:276-281.
[30] Risler, J.L., Delorme, M.O., Delacroix, H. & Henaut, A. (1988) "Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix." J. Mol. Biol. 204:1019-1029. (PubMed)
[31] Altschul, S.F. (1991) "Amino acid substitution matrices from an information theoretic perspective." J. Mol. Biol. 219:555-565. (PubMed)
[32] States, D.J., Gish, W. & Altschul, S.F. (1991) "Improved sensitivity of nucleic acid database searches using application-specific scoring matrices." Methods 3:66-70.
[33] Gonnet, G.H., Cohen, M.A. & Benner, S.A. (1992) "Exhaustive matching of the entire protein sequence database." Science 256:1443-1445. (PubMed)
[34] Henikoff, S. & Henikoff, J.G. (1992) "Amino acid substitution matrices from protein blocks." Proc. Natl. Acad. Sci. USA 89:10915-10919. (PubMed)
[35] Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) "The rapid generation of mutation data matrices from protein sequences." Comput. Appl. Biosci. 8:275-282. (PubMed)
[36] Overington, J., Donnelly, D., Johnson M.S., Sali, A. & Blundell, T.L. (1992) "Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds." Prot. Sci. 1:216-226. (PubMed)
[37] Henikoff, S. & Henikoff, J.G. (1993) "Performance evaluation of amino acid substitution matrices." Proteins 17:49-61. (PubMed)
[38] Gotoh, O. (1982) "An improved algorithm for matching biological sequences." J. Mol. Biol. 162:705-708. (PubMed)
[39] Fitch, W.M. & Smith, T.F. (1983) "Optimal sequence alignments." Proc. Natl. Acad. Sci. USA 80:1382-1386.
[40] Altschul, S.F. & Erickson, B.W. (1986) "Optimal sequence alignment using affine gap costs." Bull. Math. Biol. 48:603-616. (PubMed)

[41] Myers, E.W. & Miller, W. (1988) "Optimal alignments in linear space." Comput. Appl. Biosci. 4:11-17. (PubMed)
[42] Claverie, J.-M. & States, D.J. (1993) "Information enhancement methods for large-scale sequence-analysis." Comput. Chem. 17:191-201.
[43] Wootton, J.C. & Federhen, S. (1993) "Statistics of local complexity in amino acid sequences and sequence databases." Comput. Chem. 17:149-163.
[44] Altschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. (1994) "Issues in searching molecular sequence databases." Nature Genet. 6:119-129. (PubMed)