Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 2264–2268, March 1990
Evolution

Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes

(sequence alignment/protein sequence features)

SAMUEL KARLIN† AND STEPHEN F. ALTSCHUL‡§

†Department of Mathematics, Stanford University, Stanford, CA 94305; and ‡National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Contributed by Samuel Karlin, December 26, 1989

ABSTRACT **An unusual pattern in a nucleic acid or protein sequence or a region of strong similarity shared by two or more sequences may have biological significance. It is therefore desirable to know whether such a pattern can have arisen simply by chance. To identify interesting sequence patterns, appropriate scoring values can be assigned to the individual residues of a single sequence or to sets of residues when several sequences are compared. For single sequences, such scores can reflect biophysical properties such as charge, volume, hydrophobicity, or secondary structure potential; for multiple sequences, they can reflect nucleotide or amino acid similarity measured in a wide variety of ways. Using an appropriate random model, we present a theory that provides precise numerical formulas for assessing the statistical significance of any region with high aggregate score. A second class of results describes the composition of high-scoring segments. In certain contexts, these permit the choice of scoring systems which are "optimal" for distinguishing biologically relevant patterns. Examples are given of applications of the theory to a variety of protein sequences, highlighting segments with unusual biological features. These include distinctive charge regions in transcription factors and protooncogene products, pronounced hydrophobic segments in various receptor and transport proteins, and statistically significant subalignments involving the recently characterized cystic fibrosis gene.**

Nucleic acid and protein sequence analysis has become an important tool for the molecular biologist. Determining what is likely or unlikely to occur by chance may help in identifying sequence features of interest for experimental study. A pattern of potential interest in a protein sequence might be an unusual local concentration of charged residues or of potential glycosylation sites; a region of high similarity shared by two or more sequences might be evidence of evolutionary homology or of common function.

Statistical methods for evaluating sequence patterns can be based on theoretical models or on permutation reconstructions of the observed data (refs. 1–4; for a recent review on patterns in DNA and amino acid sequences and their statistical significance, see ref. 5). Here we use a "random" model appropriate to the data to provide a benchmark for analyzing various data statistics. The *independence* random model pendencies. For these models, theoretical results (distributional properties) have previously been obtained for a variety of sequence statistics such as the length of the longest run of a given letter or pattern (allowing for a fixed number of errors), the length of the longest word (oligonucleotide, peptide) in a sequence satisfying a prescribed relationship (e.g., *r*-fold repeat, dyad pairing), and counts and spacings of long repeats (5–14). Several of these analyses have been extended to deal with comparisons within and between multiple sequences, including the identification and statistical evaluation of long common words and multidimensional count occurrence distributions for various word relationships (e.g., refs. 5, 7, 8, 12). One limitation to the applicability of these results has been their inability to allow for properties or mismatches that vary in degree. For example, in describing the charge or hydrophobicity of amino acid residues, it would be more informative to use different score levels, and when comparing sequences one may wish to count a mismatch between isoleucine and valine differently than a mismatch between glycine and tryptophan.

In this paper we describe a rigorous statistical theory that provides explicit formulas for characterizing significant sequence configurations with reference to a general scoring scheme. In particular, we determine the distribution of high aggregate segment scores and the distribution of the number of separate segments of significantly high score. A second class of results deals with the letter composition of high-scoring segments, which in certain contexts provides a method for choosing suitable scoring schemes. We will discuss the theory in two primary contexts: (*i*) the analysis of a single protein sequence with the objective of identifying segments with statistically significant high scores for hydropathy strength, charge concentration, size profile, phosphorylation potential, or secondary structure propensity; (*ii*) multiple sequence comparisons for establishing evolutionary histories or protein segments with common function and/or structure.

Scoring assignments for nucleotides or amino acids may arise from a variety of considerations. Scoring criteria can be provided by biochemical properties (e.g., charge, hydrophobicity), physical properties (e.g., molecular weight, shape),

Henceforth, we designate the alphabet in use by $\{a_1, a_2, \ldots, a_r\}$ and the corresponding letter scores by $\{s_1, s_2, \ldots, s_r\}$. For nucleotides, $r = 4$; for purines versus pyrimidines, $r = 2$; for codons, $r = 61$; for the standard amino acids, $r = 20$; for an amino acid chemical classification [aliphatic, aromatic, . . . (see ref. 19)], $r = 8$; and for the charge attributes of amino acids, $r = 3$. It is useful to describe some concrete natural scoring assignments.

(i) *Scores based on charge.* For lysine and arginine, $s = +1$; for aspartate and glutamate, $s = -1$; for histidine $s = 0.04$ (at pH 7.2 in blood serum) or $s = 0.44$ (at pH 6.1 in muscle cells); for other amino acids, $s = 0$. Alternatively, we might take s to be the pK value of an amino acid minus 7.

(ii) *Scores associated with a run of a particular letter type* a. Here we set the score of letter a to $+1$ and the score of all other letters to $-\infty$. Obviously, only a run of the letter a can have positive score.

(iii) *Scores derived from target frequencies.* In a random sequence the letters are sampled with probabilities $\{p_1, \ldots, p_r\}$, respectively. Let $\{q_1, q_2, \ldots, q_r\}$ be a set of desirable "target frequencies" of the letter types. In certain contexts that will be discussed below, the scores $s_i = \log(q_i/p_i)$, $i = 1, 2, \ldots, r$, (resembling a likelihood ratio) are appropriate.

(iv) *Scores based on structure alphabets.* Dickerson and Geis (20) classified amino acids into internal (i), external (e), and ambivalent (a) types. This is a good alphabet for studying hydrophobicity. An associated scoring scheme, more refined

[redacted]

Limit Distribution for Maximal Segment Scores

To assess the statistical significance of high-scoring segments, we need to know the probability distribution for maximal segment scores from a random sequence of length n. *Theorem 1* provides an answer to this question. All the results described below make use of a key number λ^* which is the *unique positive* solution to the equation

$$\sum_{i=1}^{r} p_i \exp\{\lambda s_i\} = 1. \qquad [1]$$

Note that $\lambda = 0$ also solves the equation.

For a sequence of length n, let $M(n)$ denote the maximal segment score. It can be proved that $M(n)$ is of the order $(\ln n)/\lambda^*$ (24). Subtracting this centering value from $M(n)$, we can ask what is the limiting probability distribution for $\tilde{M}(n) = M(n) - (\ln n)/\lambda^*$.

THEOREM 1. *The random variable* $\tilde{M}(n)$ *(the centered maximal segment score) has the close approximating distribution*

$$Prob\{\tilde{M}(n) > x\} \approx 1 - exp\{-K^* e^{-\lambda^* x}\}. \qquad [2]$$

A formula for K^*, given in the appendix, is a rapidly converging series. A subroutine in the C programming language that calculates λ^* and K^* for any valid set of scores and associated probabilities is available from the authors.

Imagine we have experimentally identified a large collection of transmembrane regions. If we give positive scores to hydrophobic and negative scores to hydrophilic residues, these regions are likely to be the highest scoring segments of their respective proteins. However, it is possible that many other proteins that contain no transmembrane regions will have equally high-scoring segments merely by chance. Is there any way better to separate by score the true transmembrane segments from the illusory ones?

Suppose that there is some statistical difference between the respective amino acid frequencies of "true" and high-scoring "chance" segments. For instance, glycine might occur more frequently among the true segments. In this case, increasing the score for glycine would tend better to distin-

dom sequences with letter probabilities $\{p_1, \ldots, p_r\}$ and $\{p'_1, \ldots, p'_r\}$, respectively. The pair of letters a_i of the first sequence and a_j of the second sequence occurs with probability $p_i p'_j$. Let the score for such a pairing be s_{ij}. We assume, as previously, that the expected pair score $\sum_{i,j} p_i p'_j s_{ij}$ is *negative* and that there is some probability of a positive score. The number λ^* is determined (compare with Eq. 1) as the unique positive solution of the equation

$$\sum_{i,j} p_i p'_j \exp\{\lambda s_{ij}\} = 1. \quad [4]$$

Subject to the restriction that the probability distributions $\{p_i\}$ and $\{p'_j\}$ for the two sequences are not too dissimilar and that the sequence lengths m and n grow at roughly equal

analogy to the one-sequence transmembrane example discussed previously, one can start by examining related sets of segments from a variety of protein superfamilies to estimate the amino acid substitution frequencies $\{q_{ij}\}$ found as the result of evolution over substantial periods of time. Then, using individual amino acid frequencies $\{p_i\}$ from the same set of proteins, one can calculate the "log-odds" scores $s_{ij} = \log(q_{ij}/p_i p_j)$.

score exceeding $\ln n/\lambda^* = \ln 1320/0.94 = 7.6$, which for the Poisson distribution with parameter $K^* = 0.337$ has probability of occurrence $P \approx 0.0050$.

(ii) Zeste protein [$n = 575, f(s = 2) = 11.0\%, f(s = -2) = 9.4\%$], maximal segment 78–86, score 12, $P \approx 0.0040$; this is part of a positive charge cluster, residues 78–128, containing 18 basic and 3 acidic residues; see ref. 40.