*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 generates successive letters of a sequence in an independent fashion such that letter $a_j$ is selected with probability $p_j$. In the case of proteins, the $p_j$ are usually specified as the actual amino acid frequencies in the observed sequence. A random first-order Markov model prescribes $p_{jk}$ as the conditional probability of sampling letter $a_k$ following letter $a_j$. (In this case the $p_{jk}$ would correspond to the observed diresidue frequencies in a protein sequence.) More complex random models could accommodate more elaborate long-range dependencies. 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), kinetic properties (e.g., turnover rates), or associations with secondary structures ($\alpha$-helices, $\beta$-strands, turns, open coils). Amino acid classifications have also been based on the differences between codons (15) and on studies of similar tertiary structures (16). Matching scores can be adduced empirically from studies of evolutionary relationships (17, 18): Dayhoff *et al.* (18) studied groups of closely related proteins from more than 70 superfamilies to construct a statistically based amino acid substitution scoring matrix. Finally, random scores may be used as controls.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: CF, cystic fibrosis.

§To whom reprint requests should be addressed.

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 than the three-letter alphabet and reasonably consistent with it, is the Kyte–Doolittle scale or any of more than 12 alternative scales that have been proposed for hydrophobicity (see refs. 21 and 22).

## Theory for a Single Sequence

We are given an alphabet of letters $A = \{a_1, a_2, \ldots, a_r\}$. For our ends, a random sequence consists of letters sampled independently from $A$ with respective probabilities $\{p_1, p_2, \ldots, p_r\}$. The theorems we describe have generalizations to a random model in which successive letters have a Markov dependence (for proofs and extensions, see refs. 23, 24). Associated with each letter $a_i$ is a score $s_i$. We are primarily interested in the segment of the sequence with greatest aggregate (additive) score, which we will call the *maximal segment*; its score we call the *maximal segment score*. The length of this segment is determined by the data rather than being preset arbitrarily. Traditional profile studies of protein sequences use a fixed scan or window length and keep a record of moving window scores (21). However, no clear criteria for choosing the window length have been proposed and no rigorous significance results are available to date.

We impose two important but reasonable restrictions on the set of scores. First, *we require at least one score to be positive*. Second, *we require the expected score per letter* $E = \sum p_i s_i$ *to be negative*. If, on the contrary, $E$ were positive, the maximal segment would tend to be the whole sequence, and this is not of interest. As discussed below, the case of $E = 0$ is interesting but more recondite.

In many circumstances the assumption $E < 0$ is intrinsic. For example, in the simple case of runs of a letter type (example *ii* of the introduction), clearly $E = -\infty$. In the model of scores calculated using a set of "target frequencies" (example *iii* of the introduction), whenever the frequencies $\{q_i\}$ are not identical to the $\{p_i\}$, then $\sum p_i s_i = \sum p_i \log(q_i/p_i) < 0$ holds automatically. Finally, for any set of scores $\{s_i\}$ with $E$ positive, the modified scores $s_i' = s_i - \alpha E$ with $\alpha > 1$ satisfies $\sum p_i s_i' < 0$. In this case finding a segment with large score using the $\{s_i'\}$ amounts to selecting a segment with score in excess of its statistical mean score by at least the factor $\alpha > 1$.

## 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 $\lambda^*$ 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 $\lambda^*$ and $K^*$ for any valid set of scores and associated probabilities is available from the authors.

The distribution on the right of Eq. 2 is not symmetric but is positively skewed and unimodal with mode at $(\ln K^*)/\lambda^*$. The maximal segment score in the case of zero expected score per letter ($E = 0$) has order growth $M(n) \sim n^{1/2}$ rather than order log $n$. Explicit limit distributional formulas and applications for the mean zero case will be presented elsewhere (23, 24).

The number of "separate" high-scoring segments—i.e., those with scores exceeding $(\ln n)/\lambda^* + x$ where $x$ is a real parameter, and sufficiently far apart—is closely approximated by a Poisson distribution with parameter $K^* \exp\{-\lambda^* x\}$ (24). Thus, the probability of finding $m$ or more distinct segments with score greater than or equal to $S$ is closely approximated by $1 - e^{-y} \sum_{i=0}^{m-1} y^i/i!$, where $y = K^* n \exp\{-\lambda^* S\}$. For $m = 1$ this reduces to Eq. 2. Using this distribution, we can assess whether the count of segments with moderate to high score over a whole protein is unusually high (see examples below).

*Theorem 1* allows one to calculate explicitly the probability that some segment from a random sequence has score greater than any given value. In particular, one can tell when the segment score value occurs in the 1% or 5% tail of the distribution. This at least provides a benchmark for assessing the statistical significance of high-scoring segments. As discussed below, the theorem can also be generalized to apply to pairwise or multiple sequence comparison, allowing the significance of certain sequence alignments to be evaluated.

## Composition of the Maximal Segment

In many cases there may be natural criteria underlying score assignments. In other situations, however, one is confronted with the problem of choosing appropriate individual letter scores. A second theorem concerning the composition of high-scoring segments bears directly on this question.

Suppose we wish to search a set of proteins for regions characterized by an unusual amino acid composition but by no other easily described feature. For example, a transmembrane region might consist preferentially of hydrophobic residues but have no other obvious structure. We would like a set of individual letter scores to distinguish transmembrane regions. Which scores are best suited to this task?

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 distinguish the true segments. Therefore, a scoring scheme can be "optimal" for identifying a particular sort of region only if there is no statistical difference between the composition of high-scoring chance segments and the composition of similarly scoring true segments. It is this observation that makes the following theorem concerning the *composition* of high-scoring chance segments relevant to the selection of scores.

THEOREM 2. *As the length of a random sequence grows without bound, the frequency of letter* $a_i$ *in any sufficiently high-scoring segment approaches* $p_i \exp\{\lambda^* s_i\}$ *with probability* 1. *In particular, this is true for the maximal segment.*

*Theorem 2* states that in a maximal or high-scoring segment of a random sequence, letter $a_i$ tends to occur with the frequency $q_i = p_i \exp\{\lambda^* s_i\}$. Notice that the $s_i$ can all be written in the form

$$s_i = \ln(q_i/p_i)/\lambda^*. \quad [3]$$

In other words, the score associated with each letter is the logarithm to some base of $q/p$, where $p$ is the frequency with which the letter appears by chance (i.e., its frequency in a typical sequence), and $q$ is the letter's implicit *target frequency*. Since any set of individual scores has an implicit set of target frequencies, the question of what is an "optimal" set of scores can be recast into the question of what is an "optimal" set of target frequencies.

As we argued above, the best target frequencies to choose are simply those found in the sort of region we seek to identify. So to construct an appropriate set of scores, we need merely to characterize the letter distribution among such regions. The score for letter $a_i$ can then be set equal to the corresponding "log-likelihood ratio" $\log(q_i/p_i)$.

Unfortunately, we may not always have a good model of the type of segment we consider of interest. As discussed earlier, individual scores can arise from a variety of considerations, and we may know only that we seek regions rich in a certain sort of residue. However, the realization that *any* set of individual scores has an implicit target distribution can still guide our choice of reasonable scores.

## Sequence Comparisons

A basic problem in biological sequence comparison is, given two or more nucleic acid or protein sequences, to find similar segments in each. For protein sequences, one approach is to define a *score matrix* for aligning pairs of amino acids (16–18). Relatively similar amino acids receive various positive scores, while dissimilar amino acids receive negative scores. Alignments can also allow for insertions/deletions (gaps). Algorithms for finding high-scoring subalignments (i.e., alignments of segments from two or more sequences) have been used widely by molecular biologists (25–29). A natural question is, when are such subalignments statistically significant?

A variation of *Theorem 1* applies to sequence alignments when gaps are not allowed. Consider two independent random 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 $\Sigma_{i,j}\, p_i p'_j s_{ij}$ is *negative* and that there is some probability of a positive score. The number $\lambda^*$ 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 rates, *Theorem 1* holds for the maximal scoring segmental alignment, but with $n$ replaced by the product $mn$. Without these restrictions, *Theorem 1* overestimates the probability of high maximal subalignment scores $M$ so that the evaluation of statistical significance is conservative. For large $x$, we have

$$\text{Prob}\left\{ M > \frac{\ln nm}{\lambda^*} + x \right\} \leq K^* e^{-\lambda^* x}. \quad [5]$$

Thus any alignment of segments from two sequences has an unusually high score (statistically significant at the 1% level) if $M$ exceeds $(\ln nm)/\lambda^* + x_0$, where $x_0$ is determined so that $K^* \exp\{-\lambda^* x_0\} = 0.01$.

The theorem also generalizes in the natural way to the comparison of more than two sequences. The theorem must be used with some caveats because the random model for protein sequences upon which it relies is only a reference standard. It is best used for showing that the scores of certain subalignments can be explained by chance alone.

While it has been proved only for the single-sequence case, *Theorem 2* is conjectured to hold for multiple-sequence comparisons. It has been proved for the special case in which the scores for all aligned pairs of identical residues are positive, and all other scores are $-\infty$ (14). We assume below that an analog of *Theorem 2* holds for multiple-sequence comparison.

We shall consider the concept of "optimal" protein comparison scores, assuming we are restricted to seeking subalignments lacking gaps and to assigning scores only to the $20 \times 20 = 400$ pairs of amino acids. Over the years many different such score matrices have been proposed, using a wide variety of rationales (see ref. 16). We wish such a matrix to differentiate as accurately as possible those subalignments similar by chance from those similar by descent and those emerging through convergent evolution.

In brief, given a random protein model, *any* score matrix can be specified up to a constant factor by its implicit target distribution for paired amino acids. The composition of high-scoring subalignments of a set of random sequences will approach this distribution. If this composition is distinguishable from that found among similarly scoring subalignments that represent related segments, then a different matrix would better separate the two classes. Thus, the optimal target frequencies for amino acid pairs is simply that found among properly aligned related, but not strongly related, proteins. This is just the set of frequencies Dayhoff *et al.* (18) estimated in constructing their widely used protein comparison matrix (PAM-250). It is possible to criticize their method for calculating target frequencies (30), but our argument supports their basic log-likelihood approach.

While these observations do not imply that the PAM-250 score matrix cannot be improved, they do suggest that the most direct approach to obtaining better matrices is through the refined estimation of random and target distributions. In

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 $s_{ij} = \log(q_{ij}/p_ip_j)$.

**Examples**

A broad-ranging study applying the statistical theory of the paper will be presented elsewhere. Representative examples of maximal scoring segments with scores based on charge, hydropathy, cysteine clusters, and amino acid similarity are given below.

(*a*) For high-scoring mixed charge segments (of basic and acidic residues) we use the scoring scheme $s = 2$ for aspartate, glutamate, lysine, arginine, and histidine and $s = -1$ for the other amino acids.

(*i*) Human 67-kDa keratin cytoskeletal type II [length $n =$ 643 amino acids, the frequency of charged amino acids, $f(s = 2) = 20.1\%$]. Maximal segment positions 238–291 (contains 11 + and 14 − residues), aggregate score 21; probability $P$ of a score of this level or greater $< 0.008$. Second highest distinct segment, 427–463, score 14; this region is part of a mixed charge cluster in the sense of refs. 31 and 32 at the positions 427–491, containing 10 + and 16 − residues. The existence of two or more separate statistically significant charge clusters in a protein sequence is rare, found in less than 3% of all protein sequences of more than 10,000 examined (33). The keratin protein also encodes two significantly long uncharged segments at positions 42–152 and 518–586.

(*ii*) Human c-jun, nuclear transcription factor and protooncogene product [$n = 331$, $f(s = 2) = 20.2\%$]. Maximal segment 246–285, score 29, $P < 2 \times 10^{-4}$. c-jun, a member of the AP1 family of transcription factors (34), features a significant mixed charge cluster at positions 246–285, containing 10 + and 7 − residues. This charge centers on a positive charge cluster (positions 252–273) involving 12 + and 1 − residues which has sequence similarity to the DNA-binding domain of GCN4 (35). Preceding the charge cluster, c-jun contains a region of 85 residues involving only one positive and one negative charge. The charged region in c-jun (jun-A in mouse) is highly conserved in jun-B and jun-D (36), which also preserve the uncharged central portion without amino acid identity (37).

(*b*) High-scoring acidic charge segments: score assignments, $s = 2$ for aspartate and glutamate; $s = -2$ for lysine and arginine; s = −1 otherwise.

*Drosophila* zeste protein, nuclear transcription factor [$n =$ 575, $f(s = 2) = 9.4\%$, $f(s = -2) = 8\%$]. Maximal segment 194–209, score 11, $P \approx 0.0037$; this segment contains 10 acidic and no basic residues. The zeste protein involves an unusual charge distribution featuring multiple charge clusters of positive, negative, and mixed sign, respectively, as well as a long uncharged region. The uncharged region is abundant with glutamine and alanine, one of several structures of a regulatory activating domain (38). While functional domains of the zeste protein have not been delineated, it is known to regulate in *Drosophila* embryogenesis *Ubx* (Ultrabithorax), white, and DPP (decapentaplegic complex) gene expression and is thought to interact with other protein factors in mediating transactivation and transvection (39).

(*c*) High-scoring basic charge segments: score assignments, $s = 2$ for lysine, arginine, and histidine; $s = -2$ for aspartate and glutamate; $s = -1$ otherwise.

(*i*) *Drosophila* sodium ion channel protein [$n = 1320$, $f(s = +2) = 10.2\%$, $f$(s = −2) = 9.9%]. Maximal segment 930–943, score 10, $P \approx 0.034$. There are three separate segments of 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.

(*iii*) U1 70-kDa small nuclear ribonucleoprotein is a prime factor of the spliceosome ensemble, acting mainly at the 5′ donor intron site (41) [$n = 614$, $f(s = +2) = 25.1\%$, $f(s = -2) = 18.5\%$]. Maximal segment 407–483, score 37, $P < 2 \times 10^{-4}$.

(*d*) Strong hydrophobic segments: score assignments, $s = +1$ for isoleucine, leucine, valine, phenylalanine, methionine, cysteine, and alanine; $s = -1$ for glycine, serine, threonine, tryptophan, tyrosine, and proline; $s = -2$ for arginine, lysine, aspartate, glutamate, histidine, asparagine, and glutamine.

(*i*) *Drosophila* engrailed, participates during embryogenesis in control of anterior–posterior segment determination [$n = 552$, $f(s = +1) = 31.7\%$, $f(s = -1) = 31.9\%$]. Maximal segment 63–88, score 17, $P \approx 1.8 \times 10^{-5}$; this segment is rich in alanine; second maximal segment 232–243, score 10, $P <$ 0.015; the long stretch 302–394 contains a single charged residue but is abundant in serine and glutamine (see refs. 38 and 40).

(*ii*) Human c-mas, angiotensin receptor protein [$n = 325$, $f(s = +1) = 46.8\%$, $f(s = -1) = 29.8\%$]. Maximal segment 186–212, score 15, $P \approx 0.080$. c-mas is substantially hydrophobic, possessing seven potential transmembrane segments reminiscent of rhodopsin channel proteins (42). c-mas at its carboxyl terminus has a strong mixed charge cluster (40).

(*iii*) Cystic fibrosis (CF) gene product identified in ref. 43 [$n = 1480$, $f(s = +1) = 41.6\%$, $f(s = -1) = 26.8\%$]. Maximal segment 986–1029, score 21, $P \approx 0.0010$; second maximal segment 859–884, score 17, $P \approx 0.0105$. The latter region is preceded by an acidic charge cluster at positions 819–838. Sequence comparisons in ref. 43 project the CF gene product as structurally similar to a membrane-associated transport protein.

(*e*) Cysteine cluster: score assignments, $s = 5$ for cysteine and $s = -1$ otherwise.

Human thrombomodulin, participates in down-regulating thrombin levels of the coagulation pathway [$n = 575$, $f(s = 5) = 8.5\%$]. Maximal segment 404–427 (contains six cysteine residues), score 12; $P \approx 0.91$. The high $P$ value indicates that there are no striking cysteine clusters in this protein relative to the high cysteine frequency 8.5%.

(*f*) Sequence comparisons, identifications are of maximal subalignments; score assignments, PAM-250 matrix, see ref. 18.

(*i*) Human phenobarbital-inducible cytochrome P450, fragment 331 residues, compared with alkane-inducible yeast cytochrome P450 (see ref. 44 for a recent review on cytochrome P450s); the maximal subalignment occurs at positions 265–297 in first sequence and positions 39–71 in second sequence, score 62, $P \approx 0.010$.

(*ii*) CF gene product (1480 residues); maximal internal similarity comparison aligns positions 497–586 with positions 1295–1384, score of this subalignment is 120, $P < 10^{-5}$. This subalignment is consistent with the third internal similarity region reported in ref. 43.

(*iii*) A complete data base similarity search for the CF protein sequence gave a significant alignment with molybdenum transport protein chlD of *Escherichia coli* (B26871 of the Protein Identification Resource data base, length 300), the maximal subalignment here involves the 48 residues in the CF protein at positions 540–587 and the residues of the molybdenum transport protein at positions 121–168, score 99, $P <$ $10^{-4}$. This subalignment was not reported in ref. 43.

## Appendix

To give a general formula for the parameter $K^*$, we need to develop some notation. Let $S_k$ be a random variable representing the sum of the scores of $k$ independently chosen letters. Let $E(X)$ be the expected value of the random variable $X$—i.e., the sum of $x$ Prob$(X = x)$ over all values $x$ that $X$ can attain. Let $E(X; X > 0)$ denote the same sum, but taken only over possible values of $X$ that are greater than 0.

We need to define the constant $C^*$.

$$C^* = \frac{\exp\left\{-2\sum_{k=1}^{\infty}\frac{1}{k}\left(E[e^{\lambda^* S_k}; S_k < 0] + \text{Prob}(S_k \geq 0)\right)\right\}}{\lambda^* E[S_1 e^{\lambda^* S_1}]}.$$

Then the parameter $K^*$ of *Theorem 1* in the text is bounded between

$$K^- = C^*\left(\frac{\lambda^*\delta}{e^{\lambda^*\delta} - 1}\right), \quad K^+ = C^*\left(\frac{\lambda^*\delta}{1 - e^{-\lambda^*\delta}}\right),$$

where $\delta$ is the smallest span of score values. When all scores are integers with greatest common divisor 1, then $\delta = 1$. A rigorous statement of Eq. 2 is that for $n$ large (in practical terms $n \geq 150$ suffices)

$$1 - \exp\{-K^- e^{-\lambda^* x}\} \leq \text{Prob}\left\{M(n) > \frac{\ln n}{\lambda^*} + x\right\} \leq 1 - \exp\{-K^+ e^{-\lambda^* x}\}. \quad [6]$$

Using $K^+$ for $K^*$ always provides a conservative estimate of statistical significance. The infinite series for $C^*$ converges geometrically fast, so that only a small number of terms are needed to get a good estimate of $K^*$ (i.e., $K^+$).

For certain sets of scores the formulas above can be simplified to give closed-form expressions for $\lambda^*$ and $K^*$. For example, if score 1 occurs with probability $p$, score 0 with probability $r$, and score $-1$ with probability $q$ where $q > p$, then $\lambda^* = \ln(q/p)$ and $K^* = (q - p)^2/q$.

*Theorem 1* allows explicit calculation of the probability that some segment from a random sequence has score less than or equal to any given value. For example, consider a specific protein of length $n$ and set of amino acid scores. We wish to calculate the level below which 99% of the maximal segment scores for random sequences with similar composition and length will fall. First, we take the amino acid probabilities for a random protein model directly from the protein at hand. From these probabilities and the given scores, we can calculate the parameters $K^*$ and $\lambda^*$ as described above. Solving the equation $\exp\{-e^{-\lambda^* x}\} = 0.99$ for $x$ yields $x = [-\ln \ln(1/0.99)]/\lambda^*$. Any segment with score greater than $[(\ln n + \ln K^*)/\lambda^*] + x$ is then considered significant at the 99% level.

We appreciate helpful discussions and comments on the manuscript from Drs. Edwin Blaisdell, Volker Brendel, Philipp Bucher, and David Lipman. S.K. was supported in part by National Institutes of Health Grants GM39907-02 and GM10452-26 and National Science Foundation Grant DMS86-06244.

1. Doolittle, R. F. (1981) *Science* **214**, 149–159.
2. Karlin, S., Ghandour, G., Ost, F., Tavaré, S. & Korn, L. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 5660–5664.
3. Fitch, W. M. (1983) *J. Mol. Biol.* **163**, 171–176.
4. Altschul, S. F. & Erickson, B. W. (1985) *Mol. Biol. Evol.* **2**, 526–538.
5. Karlin, S., Ost, F. & Blaisdell, B. E. (1989) in *Mathematical Methods for DNA Sequences*, ed. Waterman, M. S. (CRC, Boca Raton, FL), pp. 133–157.
6. Karlin, S. & Ost, F. (1985) *Proceedings of Berkeley Conference in Honor of J. Neyman and J. Kiefer*, eds. LeCam, L. M. & Olshen, R. A. (Wadsworth, Monterey, CA), Vol. 1.
7. Arratia, R. & Waterman, M. S. (1985) *Adv. Math.* **55**, 13–23.
8. Arratia, R. & Waterman, M. S. (1986) *Ann. Prob.* **13**, 1236–1249.
9. Gordon, L., Schilling, M. F. & Waterman, M. S. (1986) *Prob. Theor. Relat.* **72**, 279–287.
10. Arratia, R., Gordon, L. & Waterman, M. S. (1986) *Ann. Stat.* **14**, 971–993.
11. Foulser, D. & Karlin, S. (1987) *Stochastic Proc. Appl.* **24**, 203–224.
12. Karlin, S. & Ost, F. (1987) *Adv. Appl. Prob.* **19**, 293–351.
13. Karlin, S. & Ost, F. (1988) *Ann. Prob.* **16**, 535–563.
14. Arratia, R., Morris, P. & Waterman, M. S. (1988) *J. Appl. Prob.* **25**, 106–119.
15. Feng, D. F., Johnson, M. S. & Doolittle, R. F. (1985) *J. Mol. Evol.* **21**, 112–125.
16. Risler, J. L., Delorme, M. O., Delacroix, H. & Henaut, A. (1988) *J. Mol. Biol.* **204**, 1019–1029.
17. McLachlan, A. D. (1971) *J. Mol. Biol.* **61**, 409–424.
18. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington), Vol. 5, pp. 345–352.
19. Mahler, H. R. & Cordes, E. H. (1966) *Biological Chemistry* (Harper & Row, New York).
20. Dickerson, R. E. & Geis, I. (1983) *Hemoglobin: Structure, Function, Evolution and Pathology* (Cummings, Menlo Park, CA).
21. von Heijne, G. (1987) *Sequence Analysis in Molecular Biology* (Academic, San Diego, CA).
22. Bishop, M. J. & Rawlings, C. J. (1987) *Nucleic Acid and Protein Sequence Analysis: A Practical Approach* (IRL, Oxford).
23. Dembo, A. & Karlin, S. (1990) *Ann. Prob.*, in press.
24. Karlin, S., Dembo, A. & Kawabata, T. (1990) *Ann. Stat.*, in press.
25. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48**, 443–453.
26. Smith, T. F. & Waterman, M. S. (1981) *Adv. Appl. Math.* **2**, 482–489.
27. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
28. Sellers, P. H. (1984) *Bull. Math. Biol.* **46**, 501–524.
29. Karlin, S., Morris, M., Ghandour, G. & Leung, M.-Y. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 841–845.
30. Wilbur, W. J. (1985) *Mol. Biol. Evol.* **2**, 434–447.
31. Karlin, S., Blaisdell, B. E., Mocarski, E. S. & Brendel, V. (1989) *J. Mol. Biol.* **205**, 165–177.
32. Karlin, S., Blaisdell, B. E. & Brendel, V. (1990) *Methods Enzymol.* **183**, 388–402.
33. Karlin, S. (1990) in *Proceedings of the Sixth Conversation in Biomolecular Stereodynamics*, eds. Sarma, R. H. & Sarma, M. H. (Adenine, Guilderland, NY), Vol. 2, pp. 171–180.
34. Rauscher, F. J., III, Cohen, D. R., Curran, T., Bos, T. J., Vogt, P. K., Bohmann, D., Tjian, R. & Franza, R. B., Jr. (1988) *Science* **240**, 1010–1016.
35. Vogt, P. K., Bos, T. J. & Doolittle, R. F. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 3316–3319.
36. Ryder, K., Leu, L. F. & Nathans, D. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 1487–1491.
37. Karlin, S. & Brendel, V. (1990) *Oncogenes* **5**, 85–95.
38. Mitchell, P. J. & Tjian, R. (1989) *Science* **245**, 371–378.
39. Pirrotta, V., Manet, E., Hardon, E., Bickel, S. E. & Benson, M. (1987) *EMBO J.* **6**, 791–799.
40. Brendel, V. & Karlin, S. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 5698–5702.
41. Theissen, H., Etzerodt, M., Reuter, R., Schneider, C., Lottspeich, R., Argos, P., Luehrmann, R. & Philipson, L. (1986) *EMBO J.* **5**, 3209–3217.
42. Jackson, T. R., Blair, L. A. C., Marshall, J., Goedert, M. & Hanley, M. R. (1988) *Nature (London)* **335**, 437–440.
43. Riordan, J. R., Rommens, J. M., Keren, B., Alan, N., Romahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J. L., Drumm, M. L., Iannuzzi, M. C., Collins, F. S. & Tsui, L. C. (1989) *Science* **241**, 1066–1071.
44. Gonzalez, F. J. (1989) *Pharmacol. Rev.* **40**, 243–288.