Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

The rapid accumulation of molecular sequence data has led to an increasing need for fast and versatile computer algorithms and statistical methods for discerning significant patterns and relations within and among sequences. The ability to distinguish what is likely to occur from what is unlikely to occur by chance is important in this context and may help in identifying sequence features for further experimental studies. For example, consider the distribution of charged residues in a protein sequence. A random distribution does not entail nearly even spacings of charged residues but usually shows local fluctuations in charge density. How does one discern significant charge clusters (basic, acidic, or mixed) or significant gaps in the charge distribution? When is a run of charged or uncharged residues (allowing for a few errors) significantly long, and similarly when is a periodic charge pattern like $(+, 0)_n$ or $(-, 0, 0)_n$ significantly long? Here + = {K, R} [one-letter code (*1*)], − = {D, E}, and 0 = {all other amino acids}.

Statistical variation is often recondite, and impressions and intuition can be misleading. For example, the yeast transcriptional activator GCN4 (length N = 281 residues) (*2*) has a total of 36 basic residues (frequency f_+ = 12.8%) and 46 acidic residues (frequency f_- = 16.4%). The COOH-terminal DNA-binding domain of GCN4 contains 15 basic and 7 acidic residues over a length of 46 residues; does this indicate a statistically significant degree of clustering of charged residues? The answer turns out to be borderline (*3*), but as we show below, this region is distinguished by a statistical test specific for the prediction of DNA-binding domains.

Statistical significance and biological significance are not necessarily synonymous. The use of only gross averages eliminates sensitivity to variance and is bound to ignore important parameters of the underlying process. Coupled with the intuition and insights of the experimenter, well-founded statistical analyses are likely to aid in the interpretation of data that bear on the understanding of molecular mechanisms and evolution. Statistical methods can, at their best, extend the potential of a given level of study as well as indicate its limits. Proper statistical procedures raise new questions and suggest new relations. Experiment remains paramount in establishing an hypothesis.

This article discusses three recent statistical methods that may assist in the appraisal of nucleic acid and protein sequence properties (*4*): (i) score-based sequence analysis; (ii) quantile distributions and correlation analysis of amino acid usage; and (iii) genomic heterogeneity assessments by *r*-scan statistics. The first method is a flexible and sensitive way to characterize anomalies in local sequence composition. The second method is appropriate for assessing global compositional biases in proteins relative to a given reference set of sequences. The third method entails analysis of counts and spacings of specific oligonucleotides (such as restriction sites) in genomic sequences. We start with a concise descriptive overview of all three methods, and then continue with elaborations, applications, and possible interpretations of results. Other statistical and computer methods for biomolecular sequence analysis not discussed here include methods for sequence similarity comparisons, analysis of codon bias, determination of consensus sequences, motifs, and profiles, phylogenetic reconstructions, and protein and RNA structure predictions; representative references with annotations are given in (*5*).

S. Karlin is the Robert Grimmett Professor and V. Brendel is research associate at the Department of Mathematics, Stanford University, Stanford, CA 94305.

Descriptive Overview

Score-based sequence analysis. Interesting sequence segments and arrangements can be identified by assigning appropriate scores to the individual residues or to sets of residues of one or several sequences. We discuss the theory in two contexts: (i) analysis of a single protein sequence seeking to identify sequence features that correspond to segments of significantly high cumulative score; and (ii) analysis of multiple sequences seeking to determine evolutionary histories and identify common segments having high total similarity score. In these analyses the segment length is variable. The distribution of the maximal segment score for randomly generated single or multiple protein sequences is available under broad conditions (*6*, *7*). Such results may serve as benchmarks of statistical significance. The results also provide a means for choosing suitable scoring schemes (*7*).

Score-based sequence analysis aimed at locating high-scoring segments can be used to resolve clusters of amino acids with a particular characteristic (such as charge or hydrophobicity) or of a particular type (such as Ser/Thr or Cys). The method can be extended by using two or more scoring regimes simultaneously, for example, in predicting amphipathic α helices. The use of generalized scores [such as PAM matrices that are based on observed amino acid replacements (*8*, *9*) or amino acid classifications (alphabets) that correlate with physiochemical properties] also provides a versatile tool to conduct multiple-sequence similarity comparisons and phylogenetic reconstructions (*10*).

It is instructive to illustrate some natural explicit scoring assignments. We designate the alphabet at hand by $\{a_1, a_2, \ldots, a_r\}$ and the corresponding letter scores by $\{s_1, s_2, \ldots, s_r\}$; for example, for nucleotides, r = 4; for codons, r = 61; for amino acids, r = 20; and for the charge sign of amino acids, r = 3.

1) Scores emphasizing positive charge. For Lys (K) and Arg (R) set s = +2; for Asp (D) and Glu (E), s = −2; and for other amino acids, s = −1; His (H) may be scored in various ways (*7*). Alternatively, we might take for scores the pK value of an amino acid minus 7. In seeking negative charge clusters, interchange the scores as-

signed to {K, R} and {D, E} above.

2) Scores associated with a run of a letter a. Here the score of letter a is set to +1 and the score of all other letters to a sufficiently large negative number. Obviously, only a run of letter a can have positive score.

3) Scores for hydrophobic profiles. In this case one may use the Kyte-Doolittle scale or any of the many other scales that have been proposed for measuring hydrophobicity (*11*).

4) Scores derived from target frequencies. In a class of sequences, suppose the average letter frequencies are $\{p_1, \ldots, p_r\}$. Let $\{q_1, q_2, \ldots, q_r\}$ be a set of target frequencies, which correspond to the composition in representative segments of the type we wish to identify. In many contexts the scores $s_i = \ln(q_i/p_i)$, $i = 1, 2, \ldots, r$ (log-likelihood ratios), are appropriate. Below we exemplify the concept of target frequencies with scoring methods for the detection of transmembrane segments and DNA-binding domains in proteins.

Quantile distributions and correlation analyses of amino acid usage. Detailed knowledge of amino acid usage (the global compositional spectrum) within and between protein sets can provide aids in appraising a particular sequence. For example, if a certain protein is reported to be rich (or poor) in a given amino acid type, one would like to know how unusual this circumstance is among a broad collection of proteins from a similar source. For this purpose, we have introduced the application of quantile distributions of amino acid usage, a method that encompasses a much finer description than mere average and standard deviation estimates (*12*). Explicitly, the quantile $Q(x)$ of a residue type for a given set of proteins is the fraction of proteins for which that residue type occurs with frequency less than the percentage x. Quantile distributions may be calculated for different amino acid alphabets (for example, hydrophobic, charge, and codon groupings) and pertain to the following biological and evolutionary issues. What is the nature of amino acid usage per protein in relation to its function, time of expression, cellular and tissue localization, evolutionary history, or other biological criteria? How does usage of amino acids with particular biochemical or steric attributes correlate? For example, how do the quantile distributions compare between Lys and Arg, both positively charged, between Asp and Glu, both negatively charged, among major hydrophobic amino acids (L, V, I, F, and M), or between the small amino acids, Gly and Ala?

A preview of some results may be useful. Thus, among the acidic residues Glu usage is stochastically larger than Asp usage for all species examined; that is, for any specified frequency of usage there are more proteins that use Glu at or above the specified frequency than there are proteins that use Asp at or above that frequency. Charge compensation is apparently a universal property in that cationic and anionic residue frequencies display significant positive correlation (generally, with a correlation coefficient greater than 0.4) for protein sequences of species as diverse as *Escherichia coli* and human. Other perspectives on compositional biases may relate to the complexity of the biosynthetic pathways for the different amino acids, to relative amino acid abundances, to tRNA availabilities, to aminoacyl synthetase fidelity, and to possible founder effects.

Compositional heterogeneity within and between genomes. In the study of genomic organization, the general problem arises of how to characterize anomalies in the spacings of a specified marker (for example, restriction sites, purine or pyrimidine tracts of certain lengths, and nucleosomes). Similar questions concern the spacings of particular residues (such as Cys) in protein sequences. How does one assess excessive clustering (too many neighboring short spacings), overdispersion (long gaps between markers), or too much regularity (too few short spacings or too few long gaps or both)?

For example, a group of eight DNA adenine methylation (DAM) sites, corresponding to the tetranucleotide GATC, was observed in a 245-bp stretch that included the *E. coli* origin of replication (*13*). If we assume that available DAM sites are distributed at random with a certain frequency around the *E. coli* genome, what is the probability of observing such a cluster somewhere in the *E. coli* genome? A statistical method based on r-scans (sums of r consecutive distances between markers) was introduced in (*14*) for discerning nonrandomness in the distribution of the specified markers in sequence data. The technique is particularly adapted to varying the scale at which inhomogeneities can be detected, from nearest neighbor to more distant interactions.

Score-Based Sequence Analysis

In this section we describe probabilistic formulas for characterizing significant configurations in random letter sequences with reference to specific assignments of letter scores. Of particular interest is the identification and evaluation of the segment of the sequence with maximal additive score, or more generally, of several top-scoring segments. A second set of results deals with the letter composition of high-scoring segments, which in certain contexts provides a method for choosing an appropriate scoring regime.

We designate the alphabet in use by $A = \{a_1, a_2, \ldots, a_r\}$ and the corresponding scores by $S = \{s_1, s_2, \ldots, s_r\}$. Let $X_1, X_2, \ldots, X_N$ be the successive letter scores in a sequence of length N. In the simplest model, the X_i are independently identically distributed with probability distribution $\text{Prob}\{X = s_k\} = p_k$, which may be interpreted to mean that sampling letter a_k (with probability p_k) yields a score s_k. The results we describe have generalizations to a model in which successive letters have a Markov dependence (*15, 16*). Two essential restrictions are imposed on the set of scores: there has to be at least one positive score, and the mean score per letter, $\mu = \Sigma p_i s_i$, has to be negative. If $\mu > 0$, the maximal segment would almost always be the whole sequence, and this is not of interest. In many situations the assumption $\mu < 0$ is intrinsic. Thus, for scores that are derived from a set of "target frequencies" $\{q_i\}$ with score values given by $s_i = \ln(q_i/p_i)$, whenever the frequencies $\{q_i\}$ are not identical to the $\{p_i\}$, then necessarily $\Sigma p_i s_i = \Sigma p_i \ln(q_i/p_i) < 0$.

Statistical theory. Probability measures are available to characterize segments of high aggregate score and the distribution of the number of separate segments of high score. Let $\{S_n\}_0^N$ be the partial sum process of segment scores, that is, $S_0 = 0$, $S_m = \Sigma_{i=1}^m X_i$, $m = 1, 2, \ldots, N$. The quantity $M(N) = \max_{0 \le k < l \le N} (S_l - S_k)$ corresponds to the segment of the sequence with maximal aggregate score. For each value of x, the maximal aggregate score $M(N)$ satisfies

$$\text{Prob}\left\{M(N) > \frac{\ln N}{\lambda^*} + x\right\} \simeq 1 - \exp\{-K^* e^{-\lambda^* x}\} \quad (1)$$

(*17*). Here λ^* (*18, 19*) is the unique positive solution to the equation

$$\sum_{i=1}^{r} p_i e^{\lambda s_i} = 1 \quad (2)$$

and K^* is a parameter that is given by an explicit series expression that can be readily evaluated numerically (*20*). Computer routines that calculate λ^* and K^* are available (*7, 10, 21*).

The asymptotic formula 1 indicates that $M(N)$ grows as $(\ln N)/\lambda^*$. In practice, we use this result on random sequences of similar overall composition to a given protein sequence to establish benchmarks of statistical significance for various distinctive segment features like hydrophobicity, charge, DNA binding, transactivation, and secondary structure (see below). To this end we set the left-hand side of Eq. 1 equal to some predetermined significance level, for example, $P = 0.01$ or $P = 0.05$, and solve for $x = x(P)$; a maximal segment score exceeding $M_P = (\ln N)/\lambda^* + x(P)$ is significant at the P level.

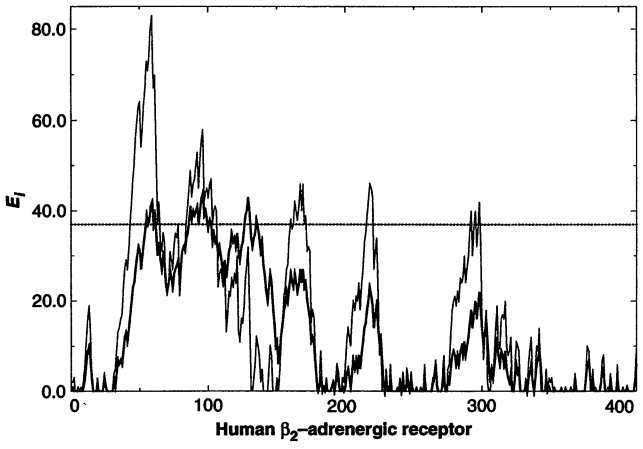

Fig. 1. Identification of high-scoring hydrophobic (thick line) and transmembrane (thin line) segments in the human β_2-adrenergic receptor (*23*). Scoring assignments were as described in text. The dotted line indicates the significance threshold at the 5% level for the hydropathy scores calculated according to Eq. 1. A significant high-scoring segment extends from residue 31 to the peak at residue 96 and corresponds to the first two transmembrane domains (*23*). The other peaks correspond to the remaining transmembrane domains.

Table 1. Scores for transmembrane segments.

Residue	q*	p†	$\text{Log}_2(q/p)$	Score‡
I	0.138	0.051	1.444	6
L	0.205	0.090	1.186	5
V	0.143	0.068	1.070	4
A	0.108	0.066	0.699	3
F	0.067	0.037	0.862	3
M	0.028	0.019	0.543	2
G	0.088	0.069	0.357	1
W	0.018	0.016	0.203	1
C	0.023	0.026	−0.194	−1
Y	0.028	0.035	−0.345	−1
T	0.049	0.067	−0.433	−2
S	0.051	0.076	−0.576	−2
P	0.019	0.055	−1.537	−6
H	0.006	0.023	−2.024	−8
Q	0.008	0.040	−2.314	−9
N	0.008	0.048	−2.611	−10
R	0.003	0.049	−3.976	−16
D	0.003	0.052	−4.025	−16
K	0.003	0.051	−4.031	−16
E	0.003	0.061	−4.187	−17

*Frequencies in the aggregate of annotated transmembrane segments in 980 protein entries of SWISS-PROT Release 21.0 [(*25*); proteins with multiple transmembrane segments excluded]. †Average overall frequencies in the same set of proteins. ‡Values of the previous column multiplied by the scale factor 4 and rounded to the nearest integer.

In many situations there may be natural criteria underlying score assignments. For example, the experimentally derived Kyte-Doolittle scale for measuring hydropathy strength mentioned above is of this kind. In other situations, however, one is confronted with the problem of choosing appropriate individual letter scores. A second theoretical result concerning the composition of high-scoring segments bears directly on this question. It has been proved (*22*) that in random sequences (successive letters independently identically distributed) high-scoring segments have an intrinsic biased composition such that letter type a_i does not occur with the sampling frequency p_i but rather with frequency

$$q_i \approx p_i e^{\lambda^* s_i} \quad (3)$$

Turning this expression around, it follows that scores defined by

$$\{s_i \approx [\ln(q_i/p_i)]/\lambda^*\} \quad (4)$$

identify high-scoring segments of target frequencies $\{q_i\}$. The location and significance of high-scoring segments remains unchanged upon scaling the scores $\{s_i\}$ by any positive factor. Thus, the result in Eq. 4 can be interpreted as follows. Let $\{p_i\}$ be the frequencies of letters in some reference random sequence, and let $\{q_i\}$ be the desirable target frequencies, which are derived from known representatives of the type of region we seek to identify. Then the score for letter a_i should be set proportional to the corresponding log-likelihood ratio $\ln(q_i/p_i)$. It might be emphasized that for any segment covering residues k to l ($l > k$) where $S_l - S_k$ is large and positive, the letter frequencies in this segment are biased toward the values $p_i e^{\lambda^* s_i}$, $i = 1, \ldots, r$, as occurs with the segment of maximal score. Conversely, if the letters in a segment are distributed according to Eq. 3 then, with high probability, the aggregate score of this segment would be very large. Examples are discussed below.

The statistical analysis underlying our discussion involves the notion of an excursion plot. We illustrate this with a hydropathy plot of the human β_2-adrenergic receptor (B2AR) (*23*), a prototype of G protein–coupled receptors. As scores we use a digitized scale corresponding to the hydropathy index of Kyte and Doolittle (*11*) rounded to the closest integer minus 1: 3 (I, V, and L); 2 (F); 1 (C, M, and A); −1 (G); −2 (T, S, W, and Y); −3 (P); −4 (H, E, Q, D, and N); and −5 (K and R). Beginning at the NH_2-terminus of the protein, we successively cumulate the scores as determined by the sequence. Because we are concerned only with high-scoring (positive) segments, we recursively define the excursion scores E_i according to

$$E_0 = 0,\ E_i = \max\{E_{i-1} + s_i,\ 0\},\ i \geq 1 \quad (5)$$

The excursion plot E_i versus i for B2AR is shown in Fig. 1. The value of each excursion is defined to be the peak score [compare with (*17*)]. If the peak score exceeds the critical values $M_{0.05}$ or $M_{0.01}$, then the segment from the beginning of the excursion up to the residue where the peak value is first realized within the excursion is a high-scoring segment, significant at the 5% or 1% level as the case may be. The statistics indicate one strong hydrophobic region in B2AR extending from residue 31 to residue 96 (Fig. 1). This excursion (extending to residue 179) consists of four distinct segments of predominantly hydrophobic residues, which are commonly assigned to the first four transmembrane domains of the receptor. Although these segments are identified as distinct ascends in the profile (Fig. 1), in this case they do not individually score sufficiently highly to be distinguished from chance fluctuations.

Target frequencies: transmembrane domains. Membrane-spanning domains of proteins are usually predicted from peaks in hydropathy plots, although more specific methods are also available (*24*). In Fig. 1 we used the digitized Kyte-Doolittle hydropathy index to predict the transmembrane segments of B2AR. Equation 4 suggests a more specific set of scores. To derive these scores, we established the frequencies q_i of amino acids in the aggregate of 980 annotated transmembrane domains assembled from SWISS-PROT Release 21.0 (*25*), the selection being restricted to proteins with a single transmembrane domain. This restriction was imposed because most annotated transmembrane domains are not firmly established experimentally. Existing prediction algorithms are presumably more likely to identify the correct extent of a transmembrane region if there is only one in the sequence, rather than a succession as in B2AR. Appropriate scores are determined as $\ln(q_i/p_i)$, where p_i are the average overall frequencies of amino acids derived from the same set of proteins (Table 1). Suitably scaled and digitized scores are given in column 4 of Table 1. These scores generally follow the Kyte-Doolittle scale, but charged amino acids score more highly negative because of their severe underrepresentation in the sampled transmembrane regions. Comparison with experimentally established transmembrane regions (when more such data become available) is required to assess whether the scores given in Table 1 yield indeed more accurate predictions. The excursion plot for B2AR gives a curve similar to the one obtained with the hydropathy scores, but with peaks much more pronounced (Fig. 1).

Table 2. Scores for DNA-binding domains.

Residue	q*	p†	$\log_2(q/p)$	Score‡
C	0.033	0.015	1.099	4
R	0.118	0.062	0.940	4
K	0.096	0.057	0.757	3
W	0.015	0.010	0.677	3
Y	0.034	0.027	0.316	1
F	0.037	0.032	0.215	1
I	0.051	0.044	0.201	1
N	0.045	0.043	0.063	0
Q	0.053	0.053	0.024	0
E	0.067	0.067	0.010	0
T	0.054	0.054	−0.020	0
V	0.049	0.053	−0.095	0
L	0.080	0.091	−0.184	−1
H	0.022	0.025	−0.200	−1
A	0.067	0.079	−0.241	−1
M	0.019	0.024	−0.376	−2
G	0.047	0.064	−0.454	−2
S	0.057	0.087	−0.615	−2
D	0.029	0.050	−0.799	−3
P	0.027	0.062	−1.223	−5

*Frequencies in the aggregate of 753 annotated DNA-binding domains assembled from SWISS-PROT Release 21.0 (*25*). †Average overall frequencies in the same set of DNA-binding proteins. ‡Values of the previous column multiplied by the scale factor 4 and rounded to the nearest integer.

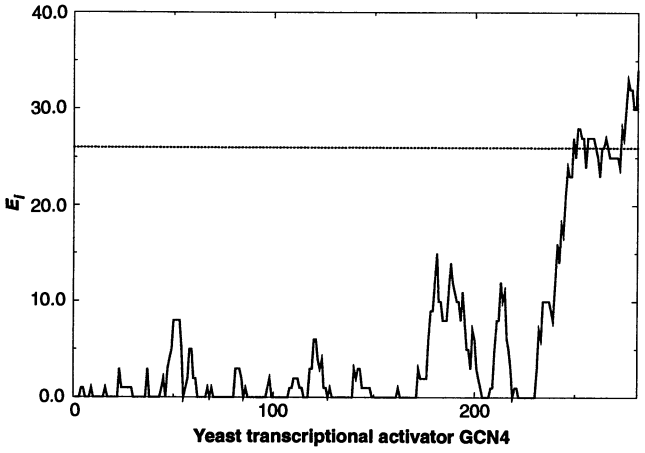

Fig. 2. Identification of high-scoring DNA-binding segments in the yeast transcriptional activator GCN4 (*2*). Scoring assignments were as given in Table 2. The dotted line indicates the significance threshold at the 5% level calculated according to Eq. 1. The high-scoring COOH-terminal sequence contains the DNA-binding function of GCN4 (*2*).

Target frequencies: DNA-binding domains. As a second example of how to choose scores best suited for the identification of specific domains in proteins we consider scores for the prediction of DNA-binding domains. Several DNA-binding motifs have been described, such as helix-turn-helix, zinc fingers, homeodomain, basic regions juxtaposed with a leucine zipper, and helix-loop-helix (HLH); other DNA-binding domains are known with no apparent similarity to the listed structural motifs (*26*). It may seem ambitious to pool all different DNA-binding motifs and on top of that ignore all positional information (like the spacings between cysteines in zinc fingers) and yet hope to recover some distinguishing property of DNA-binding domains. A common feature of DNA-binding domains, however, is their primarily basic character. Thus, a compositional prescreening of a query sequence should at least target possible regions for further inspection as potential DNA-binding domains. Proceeding as in the previous example, we determined the frequencies q_i for the aggregate of 753 DNA-binding domains assembled from SWISS-PROT feature table annotations (*25*) and derived scores according to Eq. 4 (Table 2). The scores reflect the relative overrepresentation of cysteines (mostly due to the zinc finger motifs) and of basic residues, whereas the helix-breaking residue proline is relatively underrepresented.

The partial sum plot for the yeast transcriptional activator GCN4 (*2*) is given in Fig. 2. The COOH-terminal residues 231 to 281 form a highly significant high-scoring segment, and this is indeed the experimentally determined DNA-binding domain [a basic region–leucine zipper structure (*2*)]. The high-scoring segment of GCN4 is devoid of cysteines, the high score being due mainly to positively charged residues. The association of DNA-binding domains with statistically significant charge clusters was discussed previously (*3*). Those earlier studies used fixed window size screening of sequences and binomial models to evaluate significance (*27*). Among 1307 human protein entries in SWISS-PROT of lengths at least 200 residues there are 76 sequences with annotated DNA-binding domains. The scoring method identifies 53 of these at the 5% significance level and also predicts DNA-binding segments in an additional 70 proteins (including a number of unannotated known DNA-binding proteins and false positives). Thus, the scoring method performs quite nicely as a preliminary probe for possible DNA-binding activity of a protein sequence.

Scores for hypercharge runs. As mentioned in the introduction, the evaluation of runs of a particular letter type is also encompassed by the scoring method. The choice of scores is determined by the target length and purity of the runs. For example, to detect very long charge runs one might choose scores 1 (KRED) and −4 (all other residues). Significant hypercharge runs were found to occur in many nuclear autoantigens and are speculated to play a role in autoimmunity (*28*). Significance thresholds are also available by comparison with Markov chain models (*29*).

Identification of amphipathic helices. We illustrate the versatility of scoring assignments with a brief discussion of how the method could be used to identify likely amphipathic helices in protein sequences. To this end we would define subsequences of the given protein derived by projecting (in all possible phases) the sequence onto an α-helical wheel and scoring residues separately on either side of the assumed helix (that is, residues 1, 4, 5, 8, 12, 15, 16, 19, 22, and so forth for one side, and residues 3, 6, 7, 10, 13, 14, 17, 21, 24, 25, and so forth for the other side). Thus, for each phase, both subsequences defined as described are screened with two scoring schemes, one emphasizing hydrophobicity and the other emphasizing hydrophilicity. The scoring schemes used could be the digitized Kyte-Doolittle hydropathy values for the hydrophobic side, and the negative of these values for the hydrophilic side. Alternatively, scores could be derived from target frequencies assembled from crystallographically established amphipathic helices. The latter approach would have the advantage of including helix-forming propensity biases among the residues. Two high-scoring segments, one indicating a strongly hydrophobic stretch and the other a strongly hydrophilic stretch, that overlap with respect to the original sequence would locate a region of amphipathic character.

Applications to sequence comparisons. The method of score-based sequence analysis has also been applied to the problem of establishing statistical significance for sequence comparisons (*7*, *10*). Let two independent random sequences of lengths N and N' consist of letter types a_i drawn with probabilities $\{p_1, p_2, \ldots, p_r\}$ and $\{p'_1, p'_2, \ldots, p'_r\}$, respectively. For a given alignment and position, letter type a_i in sequence 1 would be paired with letter type a_j in sequence 2, with an associated score s_{ij}. Under certain restrictions, Eq. 1 holds with N replaced by NN', λ* replaced by the unique positive root of the equation $\Sigma_{i,j=1}^{r} p_i p'_j e^{\lambda s_{ij}} = 1$, and K* computed by an accessible formula (*7*). The choice of scores is generally germane to an evolutionary model of protein relatedness, high scores given to identities of rare amino acids and negative scores associated with residues of least substitutability. Most commonly used is the protein comparison matrix of Dayhoff *et al.* (*8*). Arguments similar to those justifying scores as in Eq. 4 can be given in favor

of the Dayhoff scores over other assignments (9). Computer programs to screen a query sequence rapidly against a database for significant high-scoring segments are currently in wide use (*10*).

Quantile Distributions of Amino Acid Usage

The residue usage of specific protein sets has been the subject of a number of comparative studies (*30*). All of these comparative studies have centered on average residue usages for different protein collections. Our methods are founded on quantile distributions, stochastic ordering relations, and correlation analysis applied to different amino acid classifications.

Quantile distributions. For each residue type and protein class $\mathcal{C}$, let the quantity $y = Q(x)$ be the fraction of proteins in $\mathcal{C}$ that carry the residue type at a frequency at most x. Thus, x_m yielding $y = 1/2$ indicates the median usage value among the proteins of $\mathcal{C}$, and the interquartile range $[x_{0.25}, x_{0.75}]$ is defined by the usage frequency points of $Q(x_{0.25}) = 0.25$ and $Q(x_{0.75}) = 0.75$. A quantile distribution $\tilde{Q}(\cdot)$ is said to be stochastically larger than the quantile distribution $Q(\cdot)$ if $\tilde{Q}(x) < Q(x)$ for all x. This relation implies that the mean usage corresponding to the quantile distribution $\tilde{Q}(x)$ exceeds the mean usage corresponding to the quantile distribution $Q(x)$ and, more generally, each monotone transformation on levels of usage is similarly ranked (see, for example, Fig. 3).

Correlations of residue usage. A standard measure of concordance is the cross (Pearson) correlation coefficient which, however, can be readily confounded by outlier observations or data set biases or both effects. The Kendall Tau correlation coefficient is less affected in this way. Accordingly, for each residue type pair (X, Y) we ascertain the frequencies $p_i(X)$ and $p_i(Y)$ of these residue types in the i^{th} protein sequence of $\mathcal{C}$. For each pair of sequences, indexed i and j, we determine

$$\tau_{ij} = \begin{cases} +1 \text{ if } & [p_i(X) - p_j(X)]\,[p_i(Y) - p_j(Y)] > 0 \\ -1 \text{ if } & [p_i(X) - p_j(X)]\,[p_i(Y) - p_j(Y)] < 0 \\ 0 \text{ if either } & p_i(X) = p_j(X) \text{ or } p_i(Y) = p_j(Y) \end{cases} \quad (6)$$

The Kendall Tau association measure is

$$\tau(X,Y) = \frac{\Sigma_{i \neq j} \tau_{ij}}{[n(n-1) - 2t]^{1/2}[n(n-1) - 2s]^{1/2}} \quad (7)$$

where n is the number of sequences in $\mathcal{C}$, and t and s are the numbers of ties among pairs of $\{p_i(X)\}$ and $\{p_i(Y)\}$, respectively. Clearly $-1 \leq \tau \leq 1$, and $\tau = 1$ or -1 if and only if the values of $\{p_i(X)\}$ and $\{p_i(Y)\}$ exhibit a completely concordant or discordant ordering, respectively. A value $|\tau| \geq 0.25$ for two independent random orderings (with $n \geq 250$) has a probability <0.01 of occurring (*31*). An adjusted value of τ can be used to account for the constraint that the frequencies add to 1 (*12*).

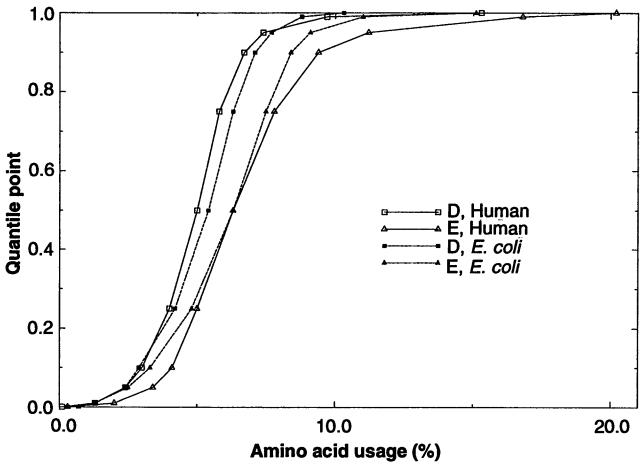

Fig. 3. Quantile distributions for acidic residues in human and *E. coli* proteins. Protein sets were as described in the legend to Table 1. Glutamate (E) is stochastically larger than aspartate (D) in both human and *E. coli*, and also in all other species examined. Graphically, this means that the glutamate quantile curves lie entirely to the right of the corresponding curves for aspartate.

Amino acids of most and least frequent usage for various species. The most frequently used amino acid (in terms of mean and median values) in almost all species is Leu, although in *E. coli* Ala is a virtual tie. The least frequently used amino acid is, on average, Trp in the eukaryotic species and in the viruses, whereas Cys is the least frequent residue in the prokaryotes *E. coli* and *Bacillus subtilis*. Cysteine, on average, is used at most 1% in the unicellular species *E. coli*, *B. subtilis*, and yeast compared to more than 2% in the higher eukaryotes. Cysteine usage entails quantile distributions that markedly deviate between human and *E. coli*. Nearly 10% of the *E. coli* proteins compared to about 5% of the human proteins are devoid of Cys residues (including many ribosomal proteins and regulatory proteins functioning in mRNA processing). At the high extreme, *E. coli* lacks cysteine-rich proteins (99% quantile = 3.8%), whereas the human collection carries many such proteins (99% quantile = 7.4%). The dearth of cysteine-rich proteins in *E. coli* presumably reflects the near absence of extracellular proteins, whereas the human collection involves many cysteine-rich secreted proteins, such as blood-clotting factors and proteins of the complement series, as well as an assortment of glycoproteins bearing a variety of disulfide-bonding patterns, such as epidermal growth factor–like domains and cysteine kringles (*32*).

Usage of charged amino acids. Table 3 displays the quantile points for the levels y = 0, 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99, and 1 of charge attributes relative to the class of all available proteins (*25*) of the human, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *E. coli*, and *B. subtilis* species and for all open reading frames (ORFs) of the major human herpes viruses [herpes simplex virus type 1 (HSV1), varicella-zoster virus (VZV), cytomegalovirus (CMV), and Epstein-Barr virus (EBV)]. The usage of charged residues over protein sets displays many intriguing features. The 10 to 90% quantile points of negatively (D+E) and positively (K+R) charged amino acids are both fairly well conserved across species. Although the aggregate average protein positive charge (K+R) content is approximately constant across species (~11.5%), K versus R usage varies significantly. In *E. coli*, the median frequencies of the basic amino acids are not strongly disparate (K 4.6% and R 5.7%) despite the difference in size of their codon complements (2 versus 6). In the human set, R is less frequent (median usage 5.3%) compared to the *E. coli* set. Concomitantly, K is a relatively abundant residue among human proteins (median usage 5.8%). The difference may in part result from CpG suppression that constrains codon usage in vertebrates but not in prokaryotes (*33*) and may also be affected by genomic compositional biases [human genomic DNA being A+T-rich, ~60%, whereas the *E. coli* genome is balanced, ~50% A+T (*34*)]. For the four major human herpes viruses K is broadly underused compared to human host proteins (that is, many herpes proteins use K at a frequency below the human 5% quantile point) and R is broadly overused (many herpes proteins use R at a frequency above the human 95% quantile point). The median and mean use of acidic residues (D+E) is nearly invariant across species, ~11.8%, composed from E, on average 6.4%, and D, on average, 5.4%.

The central range (corresponding to the 0.10 to 0.90 quantile levels) of the quantile distributions of total charge are largely concordant in all species examined. The extreme range (corresponding to the 0.01 to 0.99 quantile levels) shows much greater difference between species (apparently not simply due to variation in sample size): human, 29.1% ($x_{0.99} - x_{0.01}$; sample size

751 proteins); *Drosophila*, 33.2% (227 proteins); yeast, 24.1% (431 proteins); *E. coli*, 22.1% (710 proteins); and *B. subtilis*, 25.4% (135 proteins). Note the similar range values for the unicellular species compared to the larger values for the higher eukaryotes. It is intriguing that total charge is reduced by more than 2% on average for all major human herpes virus ORFs relative to the human protein set. Across all pro- and eukaryotic species studied, including the protein sets of Table 3 as well as chicken, *Xenopus*, *Caenorhabditis elegans*, maize, *Arabidopsis thaliana*, and *Neurospora crassa*, the median and mean net charge among proteins is slightly negative (about −0.5%). In contrast, the human herpes virus ORFs consistently carry a slightly positive average net charge, about +0.3%.

A manifest positive correlation underlies positive and negative charge content of a general protein (Table 4). In all species glutamate (E) and aspartate (D) usage are significantly positively correlated (Table 5), that is, proteins with more E tend to have more D, and vice versa. In contrast to the acidic residues, the usage of K versus R tends to be uncorrelated or slightly negatively correlated, with the median frequency of positive charge approximately constant for all species (about 11.5%). Interestingly, in all species E is stochastically larger than D (Fig. 3). No stochastic ordering occurs between basic (K+R) and acidic (D+E) usage (see Fig. 4 for human proteins).

Stochastic ordering of amino acid usage. For each amino acid type we compare the quantile distributions of the human versus the *E. coli* protein set, which are of about equal sample size (751 and 710 protein sequences, respectively) with representation of many different protein classes. The following stochastic orderings prevail: (i) *E. coli* is stochastically larger than human for the amino acid quantile distributions of L, I, V, A, M, R, and (R+K), emphasizing the major hydrophobics (except for the aromatic F); (ii) human is stochastically larger than *E. coli* for the amino acid quantile distributions of C, P, S, and E; and (iii) no definite stochastic ordering between human and *E. coli* is seen for the amino acid quantile distributions of G, F, Y, K, T, Q, N, D, W, H, and E+D. This large number of stochastic orderings is surprising in view of the long divergence time between *E. coli* and human.

Correlation analysis of amino acid usage. Tables 4 and 5 reveal three major tendencies: (i) the property of charge compensation is reflected in the high correlation of basic versus acidic residue numbers per protein, which is consistent with the approximate neutrality of proteins; (ii) positive correlations exist between functionally and structurally similar amino acids, including most pairs of hydrophobic amino acids and pairs of aromatic amino acids (primarily those having high values in the Dayhoff substitutability matrix); and (iii) the high negative correlation of strong codon group amino acids (A, G, and P; translated exclusively from SSN codons, where S stands for the strongly bonding bases cytosine and guanine, and N stands for any base) versus the weak codon group amino acids (F, I, K, M, N, and Y, translated exclusively from WWN codons, where W stands for the weakly bonding bases adenine and thymidine). How can one explain the large negative correlation of strong versus weak codon group amino acids? Possibly this condition reflects on strongly hydrogen-bonding base pair regions alternating with weakly hydrogen-bonding base pair regions of the genome, where compartments of high G+C content carry relatively more strong and relatively fewer weak codon type amino acids, and vice versa in high A+T compartments. Genomic inhomogeneity consisting of strong and weak patches of 200 to 1000 kb in length (isochores) is well established in vertebrate species (and prob-

Table 3. Quantile distributions of charge types in different species and viral protein sets. For example, 75% of human proteins have a frequency of positively charged amino acids that does not exceed 12.5%; Min, minimum, and Max, maximum. Protein sets were compiled from SWISS-PROT Release 17 (*25*). The *Drosophila* and yeast sets contain proteins from *Drosophila melanogaster* and *Saccharomyces cerevisiae* only. Highly similar sequences were culled for redundancies (*59*). Also, sequence biases based on structural or functional properties were reduced by selecting only a few representatives from multigene families (such as globins and collagens). The complete genomic viral sets correspond to the lists of known and putative ORFs as proposed by McGeoch *et al.* (*60*) for herpes simplex virus (HSV1), Davison and Scott (*61*) for varicella-zoster virus (VZV), Chee *et al.* (*62*) for cytomegalovirus (CMV), and Baer *et al.* (*63*) for the Epstein-Barr virus (EBV). Protein set sizes: human (751), *Drosophila* (227), yeast (431), *E. coli* (710), *B. subtilis* (135), HSV1 (69), VZV (64), CMV (115), and EBV (72). From all sets, sequences shorter than 200 residues were excluded in order to reduce statistical fluctuations.

Organism	Quantile distribution											
	Min	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99	Max	Mean
						Positively charged amino acids (K+R)						
Human	3.9	5.4	7.3	8.1	9.3	10.8	12.5	14.3	15.7	23.0	30.2	11.1
Dros.	4.3	4.8	6.9	7.6	9.2	10.6	12.9	14.5	15.8	24.2	27.0	11.1
Yeast	3.1	5.1	7.6	8.7	10.1	11.7	13.5	14.9	15.9	19.3	23.4	11.8
E. coli	3.0	5.0	6.1	7.4	9.0	10.4	11.8	13.1	13.8	16.2	20.5	10.4
B. sub.	4.9	5.0	6.9	7.9	9.8	11.3	13.5	15.0	16.0	17.4	18.7	11.5
CMV	4.2	5.4	5.8	7.2	9.2	10.8	12.2	14.4	16.5	18.4	19.4	10.7
EBV	2.6	2.6	5.1	6.7	8.3	9.6	11.1	12.2	13.6	14.9	14.9	9.7
HSV	6.9	6.9	8.0	8.1	8.9	9.9	11.5	12.7	14.0	17.3	17.3	10.3
VZV	7.0	7.0	7.2	7.7	8.7	9.8	11.2	12.3	14.3	16.6	16.6	10.1
						Negatively charged amino acids (D+E)						
Human	0.4	4.9	6.8	8.0	9.6	11.4	13.2	15.2	17.7	22.9	26.2	11.6
Dros.	2.0	4.7	5.8	7.1	9.2	11.0	12.8	15.2	17.7	20.2	34.9	11.2
Yeast	3.5	5.4	7.7	9.1	10.8	12.4	14.2	16.1	17.5	21.6	27.5	12.6
E. coli	2.0	3.5	5.1	6.7	10.0	11.9	13.3	14.6	15.4	17.3	20.3	11.4
B. sub.	2.8	3.1	4.1	8.5	12.1	14.0	15.4	16.8	17.8	18.5	21.3	13.2
CMV	2.7	3.1	4.1	5.3	7.1	9.8	11.4	13.1	15.0	18.5	18.8	9.5
EBV	3.1	3.1	4.4	5.5	7.6	9.6	11.1	12.8	13.3	17.3	17.3	9.4
HSV	4.0	4.0	6.3	7.2	8.7	10.1	11.1	12.3	13.7	15.1	15.1	9.9
VZV	3.1	3.1	5.5	7.4	8.6	10.0	11.3	12.4	13.4	18.6	18.6	9.9
						Total charge (K+R+E+D)						
Human	6.3	11.2	15.3	17.1	19.4	22.1	25.3	29.2	32.1	40.3	50.0	22.7
Dros.	7.0	11.0	13.1	15.3	18.6	21.9	25.7	29.3	32.2	44.2	56.6	22.4
Yeast	7.3	12.6	16.5	18.6	21.8	24.3	26.9	29.9	32.4	36.7	42.9	24.4
E. coli	7.1	9.0	11.5	14.7	19.6	22.4	24.9	26.7	28.2	31.1	33.3	21.8
B. sub.	9.4	9.4	11.2	15.7	22.9	25.4	28.5	30.6	32.8	34.8	35.8	24.7
CMV	9.6	9.9	11.7	13.5	17.1	20.8	23.4	26.2	27.3	30.1	30.9	20.2
EBV	9.0	9.0	11.5	13.0	17.3	19.5	21.3	23.3	24.5	28.3	28.3	19.1
HSV	12.9	12.9	14.7	16.4	18.2	19.9	22.5	23.9	25.6	29.5	29.5	20.2
VZV	10.7	10.7	13.2	15.4	17.8	20.1	22.2	24.5	26.4	30.4	30.4	20.0
						Net charge (K+R−D−E)						
Human	−14.6	−8.8	−5.1	−3.7	−2.2	−0.7	1.0	2.9	4.9	7.9	26.6	−0.5
Dros.	−13.2	−8.4	−5.1	−3.6	−1.9	−0.3	1.3	3.2	4.8	12.3	22.3	−0.1
Yeast	−15.3	−10.6	−6.0	−4.3	−2.3	−0.7	0.9	2.6	3.8	7.0	10.3	−0.8
E. coli	−10.1	−7.1	−4.2	−3.7	−2.6	−1.5	0.4	2.1	3.1	7.6	12.1	−1.0
B. sub.	−7.5	−6.7	−5.8	−4.9	−3.7	−2.4	0.2	2.8	3.6	4.7	5.7	−1.7
CMV	−7.1	−6.7	−3.1	−2.0	−1.2	0.6	3.0	5.2	10.4	13.3	13.8	1.2
EBV	−10.6	−10.6	−5.0	−3.7	−1.7	0.0	2.2	3.6	6.3	9.7	9.7	0.3
HSV	−5.2	−5.2	−2.8	−2.0	−1.2	0.2	1.7	3.5	4.9	9.6	9.6	0.4
VZV	−9.9	−9.9	−4.8	−2.2	−1.0	0.3	1.5	2.8	4.2	8.6	8.6	0.2

ably extant in all eukaryotes) (35). In contrast, *E. coli* has no discernable isochore structure with a nearly balanced composition [~50% G+C; (36)] and, consistently, its proteins show only a small negative correlation of strong versus weak codon group amino acid usage (see Table 4).

Assessment of Genomic Heterogeneity and the Statistics of *r*-Scans

Genomic heterogeneity is widely recognized. For example, mammalian coding regions tend to be G+C-rich, as opposed to yeast coding regions, which are A+T-rich (37). Other forms of heterogeneity include CpG suppression prominent in vertebrate genomes (33), HTFII islands (38), the widespread underrepresentation of the dinucleotide TA in nuclear DNA (39), dispersed *Alu* sequences and satellite centromeric tandem repetitive DNA (40), and characteristic telomeric sequences (41). Thus, genomic heterogeneity occurs broadly and on different scales.

Questions about spacings of a marker array and general issues of sequence heterogeneity led us to a statistical consideration of the cumulative lengths of r consecutive fragments (called r-fragments or r-scans; for example, r = 1, 2, 3, 5, 10), where a (single) fragment length is the distance between two consecutive marker sites. In particular, we focus on the lengths of the k (for example, k = 1, 2, 3) longest and the k shortest r-fragments as appropriate statistics for detecting cases of significant clumping, significant overdispersion, or excessive regularity in the spacings of the marker. The use of sums of r consecutive fragment lengths, rather than single (r = 1) fragment lengths, provides greater sensitivity for detecting unusual spacings in the marker array. The r-fragment statistics are also more tolerant of measurement errors and less affected by statistical fluctuations compared to single fragment lengths (42, 43).

Minimal and maximal spacings. Consider a sequence of length N and a specified array of n markers randomly distributed in the sequence. These occurrences induce $n + 1$ spacings (U_0, U_1, . . ., U_n), where U_0 is the distance before the first occurrence, U_i is the distance from the i^{th} occurrence of the marker to the $i + 1^{st}$ occurrence, and U_n is the distance after the last occurrence. Distances are scaled such that the distance between immediately adjacent markers equals $1/N$. Our statistical analysis focuses on the extremal spacings $m^* = \min\{U_0, U_1, \ldots, U_n\}$ and $M^* = \max\{U_0, U_1, \ldots, U_n\}$. The following classical exact probability calculations for independent uniformly distributed sites on the unit interval can help in the analysis of the spacings of a marker (44):

$$F(a) = \text{Prob}\{m^* < a\} = 1 - [1 - (n+1)a]^n, \quad \text{for } 0 < a \le \frac{1}{n+1} \tag{8}$$

$$G(b) = \text{Prob}\{M^* \ge b\} = 1 - \sum_{i=0}^{n+1} \binom{n+1}{i} (-1)^i [\delta(1 - ib)]^n, \quad \text{for } 1 > b \ge \frac{1}{n+1} \tag{9}$$

where $\delta = 1$ if $ib < 1$ and $\delta = 0$ otherwise. More generally, let $P_k(x)$ denote the probability that k of the $n + 1$ fragments are of lengths less than x; then

$$P_k(x) = \binom{n+1}{k} \sum_{i=0}^{k} \binom{k}{i} (-1)^i \{\delta[1 - (n + 1 + i - k)x]\}^n \tag{10}$$

where $\delta = 1$ if $(n + 1 + i - k)x) < 1$ and $\delta = 0$ otherwise. Equation 10 allows the analysis of the data in terms of spacings other than the extremes $k = 0$ (m^*) and $k = n + 1$ (M^*).

The evaluation of an extremal minimum at the 1% significance level rests on the determination of a^* such that $F(a^*) = 0.01$. For an observed m^* smaller than a^*, the minimum spacing is considered significantly small. Similarly, the largest gap is considered statistically significant if the observed M^* exceeds b^*, where b^* satisfies $G(b^*) = 0.01$. For an observed m^* too large [$m^* \ge c^*$, where $F(c^*) = 0.99$] or an observed M^* too small [$M^* < d^*$, where $G(d^*) = 0.99$] or both, the spacings are considered to be overly regular. Equations 8 and 9 apply to n sites sampled uniformly over a linear sequence; when the sites are sampled equally likely on a circular sequence, the formulas are to be modified by replacing n with $n - 1$. The formulas are practical for n small or of moderate size. For n large, we use the asymptotic probability calculations set forth in Eqs. 11 and 12 discussed below.

Examples of anomalous residue spacings in some protein sequences. We discuss examples from two protein families, human histones and representative HLH DNA-binding proteins. The maximum spacing between any two adjacent basic residues (K, R, or H) in histone H2A is 29 (between positions 42 and 71), and, on the basis of the random sampling model (Eq. 9) for the given composition of H2A, such a large gap length occurs with probability ≤0.01. The second largest spacing (length 19 between residues 99 and 118) is also statistically significant. Similarly, histone H3 has a significantly large maximal spacing between basic residues of length 30 (positions 83 to 113). In contrast, histones H1, H2B, and H4, all

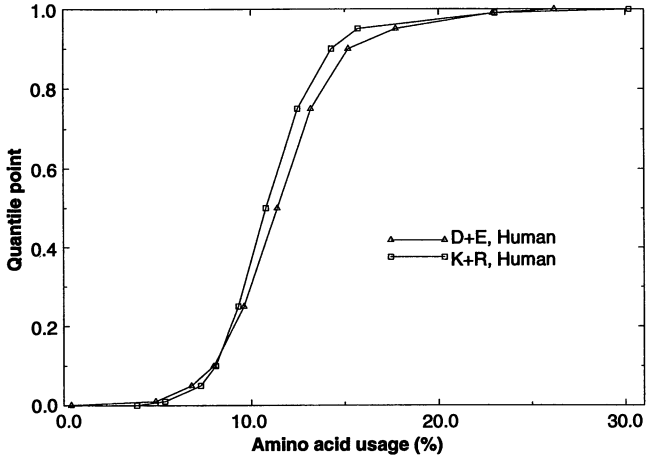

Fig. 4. Quantile distributions for basic (K+R) and acidic (D+E) residues in human proteins. The protein set was as described in the legend to Table 1. There is no stochastic ordering between the two residue types.

Table 4. Compositional correlations between amino acid charge and codon group types. Correlations were calculated according to Eq. 7 for the protein sets described in the legend to Table 3. Positive and neutral are negatively correlated, but not significantly; strong and weak codon type amino acids versus intermediate types are not significantly correlated.

Pair	Organism								
	Human	*Dros.*	Yeast	*E. coli*	*B. sub.*	HSV	VZV	CMV	EBV
Charge (+/−)*	0.560	0.632	0.492	0.614	0.595	0.230	0.300	0.220	0.260
Codon (strong/weak)†	−0.405	−0.337	−0.343	−0.030	−0.173	−0.530	−0.560	−0.530	−0.410

*Charge alphabet: positive, K+R; negative, D+E; and neutral, all others. †Codon type alphabet: strong codon type, {A, G, P}; weak codon type, {F, I, K, M, N, Y}; and intermediate, all others.

similarly high in positively charged residues, do not show such distributional anomalies. The unusual spacing of basic residues in H2A and H3 may be associated with nucleosome formation and stability.

In some cases, the two largest spacings occur in tandem (forming a long 2-scan, see below) and thus indicate an unusually long region essentially devoid of a particular residue type. We give two examples from the HLH protein family (45). The 710-residue *Drosophila* daughterless protein displays a marked difference in charge between the NH_2-terminal and COOH-terminal halves of the protein: of the 21 lysines, none occur from residues 26 to 271, inclusively, nor from residues 273 to 470 (highly significant maximal spacings), and there is a 271-residue stretch free of glutamates (of which there are 26 overall in the sequence) beginning at residue 52. Similarly, the spacing analysis indicates a significantly long stretch free of glycines at the COOH-terminus of the human transcription factor E2-α (E12 gene product), although the glycine frequency (13.3%) is actually very high. No anomalous amino acid spacings of any type are evident in most other HLH proteins, including the *Drosophila* achaete-scute, the mouse myoD, and the Myc protein families.

Spacings of palindromes in genomic sequences. The genome of coliphage λ is composed of 48,502 bp. Assays on λ-phage reveal two half genomic sections, one C+G-rich (~54% frequency) and the other A+T-rich (~55%), whereas overall the nucleotides A, T, C, and G occur with close to equal frequencies (46). The distribution of each of the 64 6-bp palindromes around the λ-genome was tested for clustering, overdispersion, or persistent regularity (47). Of the individual 6-palindromes, four involved significantly long gaps (overdispersion): CAGCTG (15 copies), maximum gap M* = 21,299 bp; CATATG (7 copies), M* = 36,001 bp; CCTAGG (2 copies), M* = 48,428 bp; and CTGCAG (28 copies), M* = 14,057 bp. Clumping was revealed for a single palindrome: CCTAGG (2 copies), m^* = 74 bp. Extremal tests on the distribution of 6-bp palindromes in the linear coliphage T7 genome (~40 kb) did not reveal a single 6-bp palindrome with anomalous spacings.

The extreme rarity of the tetranucleotide CTAG in λ-phage (14 occurrences) and *E. coli* (frequency ≈0.02%) and of the DAM site GATC in phage T7 (6 occurrences) prompted us to investigate more closely the distribution of these tetranucleotides in these three organisms. CTAG is missing in the left half of the λ-genome in a segment of 24,743 bp and occurrences concentrate in three clusters in the right half. Eight are located in noncoding regions or at stop codons, four in open reading frames of undetermined expression, one in the CI gene near the carboxyl end, and one in gene S (affecting cell lysis). Thus, the distribution of CTAG sites in λ is highly nonrandom. The distribution of DAM sites in T7 is not unusual in any way [for more details, see (47)]. In *E. coli*, CTAG occurs relatively more frequently in the rRNA genes than elsewhere. The low frequency of CTAG persists in general bacterial genomes entailing clustering of CTAG in the 16S and 23S ribosomal RNA genes (47). Is it possible that CUAG sites are nucleation or anchor points in the assembly of the ribosomal complex?

r-Scan statistics. For a given set of single spacings $\{U_0, U_1, \ldots, U_n\}$, r-scans are formed according to $R_i = \sum_{j=i}^{i+r-1} U_j$, $i = 0, 1, \ldots, n - r + 1$. To study the distribution of the markers in a sequence, we compare the distribution of $\{R_i\}$ calculated under a theoretical model with the observed distribution of r-fragment lengths. The extreme-valued r-scans (largest and smallest) are of particular use: $M_k^{(r)}$ = length of k^{th} largest r-fragment and $m_k^{(r)}$ = length of k^{th} smallest r-fragment. To detect clustering among markers, we examine all r-scans and ascertain whether the minimum is especially small with respect to the postulated theoretical distribution of markers. Similarly, in deciding whether some successive markers are excessively dispersed, we check the maximum length among r-scans to see whether it is especially large. Conversely, when the minimum r-scan length is especially large or the maximum r-scan length is especially small or both, then the spacings of the marker are assessed to be excessively regular.

To assess clustering, we use the theoretical probability that the k^{th} smallest r-fragment [length $m_k^{(r)}$] would be as small or smaller than those observed if markers were distributed randomly (for example, sampled uniformly over the long sequence). The following asymptotic formula holds for n large (48):

$$Pr\left\{m_k^{(r)} < \frac{x}{n^{1+1/r}}\right\} \approx 1 - \sum_{i=0}^{k-1} \frac{\lambda^i}{i!} e^{-\lambda}, \lambda = \frac{x^r}{r!} \quad (11)$$

With x chosen so that the right side of Eq. 11 is equal to 0.01, we declare the observed $m_k^{(r)}$ too small if it is less than $x/n^{1+1/r}$.

To assess overdispersion, we use the theoretical probability that the k^{th} largest r-fragment [length $M_k^{(r)}$] would be as large or larger than those observed if markers were in fact located randomly. The asymptotic formula in this case is (48)

Table 5. Amino acid usage correlations. Correlations were calculated according to Eq. 7 for the protein sets described in the legend to Table 3. Amino acid pairs exhibiting rank correlations higher than 0.2 or lower than −0.2 in at least two of the eukaryotic species or in both prokaryotic species were selected for display. Correlation coefficients between −0.2 and +0.2 are considered nonsignificant and are therefore not shown.

Amino acid pair	Organism				
	Human	*Dros.*	Yeast	*E. coli*	*B. sub.*
	Strongly positively correlated amino acid pairs				
A/G	0.285	0.240	0.524	0.294	0.467
A/V			0.437	0.274	0.406
C/H				0.317	0.323
D/E	0.321	0.447	0.341	0.263	0.282
D/K	0.357	0.431	0.248	0.346	0.352
E/K	0.457	0.501	0.494	0.378	0.387
E/R		0.348	0.282	0.402	0.415
F/I	0.345	0.541			
F/Y	0.345	0.332	0.327	0.306	0.313
G/V			0.522	0.377	0.481
I/V	0.316	0.431			
	Strongly negatively correlated amino acid pairs				
A/Y	−0.220	−0.308		−0.330	−0.220
D/P	−0.248	−0.489			
E/F				−0.234	−0.265
E/G	−0.259	−0.340		−0.216	−0.269
E/W				−0.298	−0.386
F/P	−0.247	−0.245			
G/K	−0.262	−0.305			−0.330
G/Q			−0.279	−0.283	−0.260
G/R				−0.303	−0.388
I/P	−0.415	−0.424			
K/P	−0.338	−0.317			
L/N				−0.227	−0.327
V/Y				−0.201	−0.323

$$Pr\left\{M_k^{(r)} > \frac{1}{n}[\ln(n) + (r-1)\ln(\ln n) + x]\right\} \approx 1 - \sum_{i=0}^{k-1} \frac{\mu^i}{i!} e^{-\mu}, \quad \mu = \frac{e^{-x}}{(r-1)!} \quad (12)$$

With x chosen so that the right side of Eq. 12 is equal to 0.01, we declare the observed $M_k^{(r)}$ too large at the 1% significance level when it exceeds $\frac{1}{n}[\ln(n) + (r-1)\ln(\ln n) + x]$.

To detect too much regularity, we use the theoretical probabilities that $m_k^{(r)}$ is especially large or $M_k^{(r)}$ is especially small or both, calculated on the basis of Eqs. 11 and 12.

The r-scan process is a moving sum process derived from the original first-order process and therefore tends to smooth out random fluctuations. Sums of r contiguous fragments have a coefficient of variation (sample standard deviation divided by mean) inversely proportional to $\sqrt{r}$, rendering r-scans quite sensitive statistics in detecting clustering. For example, the method of r-scans was particularly useful for dissecting the heterogeneities inherent in the Kohara *E. coli* physical map data ascertained by complete digestion with eight enzymes that recognize 6-bp segments (Bam HI, Bgl I, Eco RI, Eco RV, Hind III, Kpn I, Pst I, and Pvu II) (*49*). In the Kohara map there are several sources of error in the data, especially measurement errors that were generated by recording restriction sites to the closest 100 bp (rounding off). Stretches of the map were absent because of difficulties in resolving all of the fragments on autoradiograms. Moreover, the physical map appears to contain much fewer restriction sites than occur in the genome. In fact, after screening about 1.43×10^6 bp of available nonredundant *E. coli* sequences (*50*) for restriction sites, we extrapolate that (with the exception of Bam HI) the *E. coli* genome contains more sites than were mapped by a factor ranging between about 1.1 (Kpn I) and 1.9 (Eco RV) (*14*). The sequence data revealed greater disparity for frequent cutters than for sparse cutters, suggesting that the differences are due mainly to undetected small fragment lengths in construction of the physical map. Because of the digitization of restriction sites in units of 100 bp (*50*), detection of clustering by r-scans of low order ($r = 1$ or 2) was precluded (under this scheme a minimum site separation of 0 units would not be unlikely). For each restriction enzyme used, the histogram of fragment lengths were examined and the data (apart from very small sizes) followed an exponential density consistent with a homogeneous distribution of sites. However, an application of the statistics of r-fragments ($r = 10$) revealed for $m_1^{(10)}$, $m_2^{(10)}$, and $m_3^{(10)}$ a significant cluster corresponding to 13 Pst I sites beginning at map position 2074.8 kb and spanning 13 kb (*14, 51*). This cluster is not significant based on r-fragment lengths with $r = 5$. Thus, by varying r, organization on different scales can be discriminated.

Cluster of DAM sites in the ori-C region of E. coli? DAM sites are important regulatory signals composed of the tetranucleotide GATC (*52*). These sequences serve in part to distinguish the template strand (fully methylated) from the newly synthesized strand (unmethylated) during semiconservative replication and repair (*53*). These sites are also associated with genes involved in the SOS response, transposon function, and bacteriophage infection (*53*). How do these functions affect the distribution of DAM sites?

In the 245-bp sequence that defines the minimal ori-C region of *E. coli* there are eight DAM sites (*13*). In a 350-bp stretch flanking the ori-C region there are an additional 12 DAM sites. Many of these sequences are conserved in the origins of replication of other enterobacteria. Do the eight DAM methylation sites observed in a stretch of 245 bp that includes the *E. coli* origin of replication or that are joined with the additional 12 DAM sites located in the flanking 350 bp or both represent a statistically significant cluster? We apply the formula 12 first in the case of $r = 7$ and then in the case of $r = 19$, where n is the number of DAM sites throughout the *E. coli* genome. In the 1.4×10^6 bp of available *E. coli* sequences (*50*), the GATC frequency is 0.0044. On this basis, we extrapolate about $n = 0.0044 \times 4.7 \times 10^6 \approx 20{,}680$ DAM sites over the entire genome. Now, $\exp\{-x^7/7!\} = 0.99$ when $x = 1.75$. Hence, the critical value for $m_1^{(7)}$ is $(4.7 \times 10^6)(1.75/n^{8/7}) \approx 96$ bp. Thus, for a random sequence of the composition and length of *E. coli*, a segment of at most 96 bp in length that contains eight occurrences of the DAM site would be a statistically significant cluster at the 1% level. The same formula shows that the presence of eight DAM sites in a stretch of 245 bp somewhere in the *E. coli* genome would occur with probability ~0.06. Thus, the observed concentration of DAM sites in the ori-C region of *E. coli* is not statistically significant. However, repeating the calculation with $r = 19$ we find that a segment of 1,068 bp in length or less containing 20 DAM sites presents a statistically significant cluster.

Distributions of the tetranucleotide CTAG *in human herpes virus genomes.* The frequency of CTAG is significantly low in all bacterial sequences and substantially low in many eukaryotic DNA sequence sets, including *Drosophila*, chicken, *C. elegans*, CMV, HSV1, and adenovirus, and below average in virtually all sequence collections examined (*47*). Application of the r-scan statistics ($r = 1$, 3, 5, and 10) to study the distribution of CTAG sites in the major human herpes viruses gave the following results: (i) CMV (genome size ≈ 230 kb) contains a total of 341 CTAG sites (frequency = 0.0015). A significant cluster of CTAG occurs starting at position 91832 with 11 copies (10-scan) of CTAG over a stretch of 1064 bp (probability < 0.01). It is noteworthy that the region 91800 to 93500 is distinguished as the lytic origin of replication of CMV (*54*). Is it possible that these sites help in suitable protein binding for the formation of the preinitiation complex effecting replication? (ii) The EBV B-95 strain (genome size ≈ 172 kb) contains 342 CTAG sites (frequency = 0.0020). The most significant cluster of CTAG sites in EBV measured by 5-scans occurs at position 53082, extending for 255 bp. This region overlaps the EBV lytic origin of replication (*55*); and (iii) The neurotropic herpes viruses HSV1 and VZV, both substantially low in CTAG counts, have no significant clusters or gaps of CTAG as measured by r-scans.

Prospects and Limitations

An intrinsic problem of the application of statistics to biological data is the lack of firm correspondence between statistical and biological significance. Opting for stringent statistical criteria one errs on account of missing biologically relevant patterns, whereas relaxing the criteria produces a deluge of false positives. Probably the best choice is a mixture of empirical and theoretical analyses with emphasis on robustness as discussed below. A limitation of both the score statistic (formulas 1 and 3) and the r-scan statistic (formulas 11 and 12) in this context is that the given probability estimates hold only asymptotically as N (length of the sequence) and n (number of markers) are large. Rates of convergence are generally not available, and the error terms may be larger than the tail probabilities being estimated. Thus, the probability estimates have to be interpreted with caveats, and their main usefulness is to provide benchmarks for further analysis.

Application of the statistics is also restricted by assumptions underlying the statistical theory. For example, the formulas for the r-scan statistic were presented for a uniform distribution of markers, which may be appropriate in some situations and not in others. Extensions of the theory to a nonuniform distribution of markers are partly available (*43*). Application of the scoring statistic in sequence comparisons presupposes a not strongly dissimilar composition of the sequences being compared (*7, 56,*

57). Implementations of the method for this purpose (*10*) do not accommodate gaps in alignments, and applications to multiple-sequence comparisons remain computationally difficult.

Score-based statistics present a refinement over alphabet-based statistics in that degrees of matching a certain property can be incorporated. The method is quite flexible, and recent extensions of the theory allow for neighbor dependencies (*15, 16*), random scores, and vector scores (scoring, for example, simultaneously size and hydropathy attributes of amino acids). An appropriate choice of scores for a particular task is not necessarily obvious a priori and may require a judicious amount of experimentation. Similarly, in using *r*-scans one has to discern suitable values for *r* that will provide sensitivity to anomalies in the distribution of markers at the desired level of sequence organization.

Large-scale sequencing projects are well under way. Besides posing formidable problems of how to organize these data in accessible ways, what does this large amount of data bode for statistical sequence analysis? Because of the large number of multiple tests performed whenever the whole database is considered, virtually anything but the most extreme feature is to be expected to occur by chance alone. In particular, this applies to the evaluation of weak sequence similarities. In attempting to surmount this problem one might have to rely much more on multiple sequence alignments, which exploit the fact that sufficiently long words shared by several sequences would be more robust against statistical fluctuation. This is an area of much current investigation, and several new methods have been advanced (*10, 58*).

In our view, the role of statistics in sequence analysis is primarily exploratory and interactive with the data, generating new questions and lines of experimental investigation. Rather than fitting models to biomolecular sequences with the purpose of statistical hypothesis testing, the analysis of the extreme tails of distributions derived from random sequences can provide benchmarks for the selection of sequences, parts of sequences, or sequence features to concentrate on for further study. Essential in this approach is the use of a mixture of different statistics and interaction with the data and the experimenter. The conclusions drawn from such analyses rely on robustness of the results. Here robustness includes sensitivity of the statistics to outliers due to sampling biases, concordance among several different measures that examine the data in different ways, and consistency among independently sampled data sets. There are also many challenging problems related to the classification of protein and DNA sequences with reference to function, structure, subcellular localization and expression, phylogenetic relations, and other biological criteria. Statistical stratification of the databases can aid in these tasks as more sequences become available. From this perspective, the accumulation of sequence data should continue to open many possibilities for empirical and theoretical research.

REFERENCES AND NOTES

1. The one-letter code is A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
2. I. A. Hope and K. Struhl, *Cell* **46**, 885 (1986); SWISS-PROT (*25*) file GCN4$YEAST.
3. V. Brendel and S. Karlin, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 5698 (1989).
4. Although our primary interests derive from studies of biomolecular sequences, the concepts and methods described in this article can be adapted to sequence comparisons with respect to human speech and text collations, with respect to bird songs and general musical scores, and in many contexts of computer science including string editing, comparison of computer files, coding theory, and information theory. For extensive references and other discussions on these topics, see D. Sankoff and J. B. Kruskel, Eds., *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparisons* (Addison Wesley, Reading, MA, 1983).
5. Recent volumes on the topic of sequence analysis are B. S. Weir, Ed., *Statistical Analysis of DNA Sequence Data* (Dekker, New York, 1983); G. von Heijne, *Sequence Analysis in Molecular Biology* (Academic Press, San Diego, 1987); M. S. Waterman, Ed., *Mathematical Methods for DNA Sequences* (CRC Press, Boca Raton, FL, 1989); G. Bell and T. Marr, Eds., *Computers and DNA* (Longman, London, 1989), vol. 7; R. Doolittle, Ed., *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences, Methods Enzymol.* **183** (1990); see also D. Sankoff and J. B. Kruskel (*4*). Statistical methods and mathematical analyses have also been applied to sequence comparisons [see, for example, J. G. Reich, H. Drabsch, A. Daumler, *Nucleic Acids Res.* **12**, 5529 (1984); M. S. Waterman, L. Gordon, R. Arratia, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 1239 (1987); S. Karlin, P. Bucher, V. Brendel, S. F. Altschul, *Annu. Rev. Biophys. Biophys. Chem.* **20**, 175 (1991)]; to genome contig models [E. S. Lander and M. S. Waterman, *Genomics* **2**, 231 (1988); R. Arratia, E. S. Lander, S. Tavaré, M. S. Waterman, *ibid.* **11**, 806 (1991); E. Barrilot, J. Dausset, D. Cohen, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 3917 (1991)]; to studies of protein structure-function motifs and profiles [M. Gribskov, A. D. McLachlan, D. Eisenberg, *ibid.* **84**, 4355 (1987); R. F. Smith and T. F. Smith, *ibid.* **87**, 118 (1990); H. O. Smith, T. M. Annau, S. Chandrasegaran, *ibid.*, p. 826; J. U. Bowie, R. Lüthy, D. Eisenberg, *Science* **253**, 164 (1991)]; to determination of consensus sequences [M. S. Waterman, R. Arratia, D. J. Galas, *Bull. Math. Biol.* **46**, 515 (1984); G. D. Stormo, *Annu. Rev. Biophys. Biophys. Chem.* **17**, 241 (1988)]; to Markov chain fitting of sequences [for example, B. E. Blaisdell, *J. Mol. Evol.* **21**, 278 (1985); G. J. Phillips, J. Arnold, R. Ivarie, *Nucleic Acids Res.* **15**, 2611 (1987); E. E. Stückle, C. Emmerich, U. Grab, P. J. Nielson, *ibid.* **18**, 6641 (1990)]; and to prediction of RNA secondary structure [for example, J. A. Jaeger, D. H. Turner, M. Zuker, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 7706 (1989)].
6. S. Karlin, A. Dembo, T. Kawabata, *Ann. Stat.* **18**, 571 (1990).
7. S. Karlin and S. F. Altschul, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2264 (1990).
8. M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt, in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, DC, 1978), vol. 5, pp. 345–352.
9. S. F. Altschul, *J. Mol. Biol.* **219**, 555 (1991).
10. _____, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990); S. F. Altschul and D. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 5509 (1990).
11. The hydropathy index of J. Kyte and R. F. Doolittle [*J. Mol. Biol.* **157**, 105 (1982)] associates: I, 4.5; V, 4.2; L, 3.8; F, 2.8; C, 2.5; M, 1.9; A, 1.8; G, −0.4; T, −0.7; S, −0.8; W, −0.9; Y, −1.3; P, −1.6; H, −3.2; E, −3.5; Q, −3.5; D, −3.5; N, −3.5; K, −3.9; and R, −4.5. In a typical hydropathy plot these scores are averaged over a small fixed window size and the resulting moving average is plotted along the sequence. Hydrophobic segments show up as positive peaks in the profile. The question is as to what constitutes a significant peak, higher or broader or both than what would occur due to chance fluctuations.
12. S. Karlin, B. E. Blaisdell, P. Bucher, unpublished results; S. Karlin and P. Bucher, unpublished results.
13. S. Krawiec and M. Riley, *Microbiol. Rev.* **54**, 502 (1990); J. W. Zyskind, in *The Bacterial Chromosome*, K. Drlica and M. Riley, Eds. (American Society for Microbiology, Washington, DC, 1990), pp. 269–278.
14. S. Karlin and C. Macken, *Nucleic Acids Res.* **19**, 4241 (1991).
15. A. Dembo and S. Karlin, *Ann. Prob.* **19**, 1737 (1991); *ibid.*, p. 1756.
16. S. Karlin and A. Dembo, *Adv. Appl. Prob.* **24**, 113 (1992).
17. It is natural to divide the realizations of $\{S_m\}_1^N$ into successive excursion epochs such that the time frame $K_{\nu-1} + 1$ to K_ν ($\nu = 1, 2, \ldots$; $K_0 = 0$) commences from score zero at index $K_{\nu-1}$, proceeds with additive scores for $k = K_{\nu-1} + 1, K_{\nu-1} + 2, \ldots, K_\nu - 1$, and terminates at the first index K_ν for which $\Sigma_{i=K_{\nu-1}+1}^{K_\nu} X_i \leq 0$. Let the variable $Q_\nu = max_{K_{\nu-1} \leq k \leq K_\nu} (S_k - S_{K_{\nu-1}})$ be the maximal aggregate score attained during the ν^{th} excursion epoch. $M(N)$ is the maximum among the peak levels of each excursion, that is, $M(N) = \max(Q_1, Q_2, \ldots)$. The distribution of Q_ν can be calculated with exact exponential tail probability rates (*16*). These facts can be used to derive formula 1 of the text by appealing to the theory of asymptotic extremal distributions for successive independent random variables [(*6, 16*); see also J. Galambos, *The Asymptotic Theory of Extreme Order Statistics* (Wiley, New York, 1978)]. The result of formula 1 can also be applied in characterizing the asymptotic maximal waiting time distribution for the single-server queue (GI/G/1) [D. Iglehart, *Ann. Math. Stat.* **43**, 627 (1972)] and for insurance risk models and traffic flow [S. Assmussen, *Adv. Appl. Prob.* **14**, 143 (1982)]. The Markov chain extension of formula 1 is developed in (*16*).
18. The parameter λ^* is known as the conjugate or dual exponent associated with a process of partial sums of independently identically distributed real random variables [for example, (*7, 19*)]. Its Markov chain analog is set forth in (*15, 16*); see also P. Ney and E. Nummelin, *Ann. Prob.* **15**, 561 (1987).
19. W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1968), vol. 2.
20. Let S_k be the sum of the scores of k independently sampled letters. With $E[X]$ denoting the expected value of the random variable X, further define

$$A = \sum_{k=1}^{\infty} \frac{1}{k} E\left[e^{\lambda^* S_k}; S_k < 0\right] = \sum_{s_i < 0} p_i e^{\lambda^* s_i} + \frac{1}{2} \sum_{s_i + s_j < 0} p_i p_j e^{\lambda^*(s_i + s_j)}$$

$$+ \ldots, \text{define } B = \sum_{k=1}^{\infty} \frac{1}{k} Pr\{S_k \geq 0\} = \sum_{s_i \geq 0} p_i + \frac{1}{2} \sum_{s_i + s_j \geq 0} p_i p_j$$

$+ \ldots$, and let $C = E[S_1 e^{\lambda^* S_1}] = \sum_{i=1}^{r} p_i s_i e^{\lambda^* s_i}$.

Then the K^* appearing in formula 1 of the text is given by $K^* = Fe^{-2A-B}/(\lambda^* C)$, where $F = 1$ for nonlattice scores and $F = \lambda^*\delta/(1 - e^{-\lambda^*\delta})$ for lattice scores with δ being the smallest span of score values (that is, all scores can be written as multiples of δ, $|s_i| = i\delta$) (*6, 7*). The use of an estimate of K^* that is larger than the correct value increases the x that is determined from formula 1 in the text to give the 1% significance level threshold for $M(N)$. Approximations to K^* that use only a finite number of terms in the above series will accordingly be conservative. Simple expressions for λ^* and K^* are available for certain scoring schemes (*6, 16*). For example, let score 1 occur with probability p, score 0 with probability r, and score -1 with probability $q > p$; then $\lambda^* = \ln(q/p)$ and $K^* = (q - p)^2/q$. For scores $\{-m, \ldots, -1, 0, 1\}$ we have $K^* = (1 - e^{-\lambda^*})C$, and for scores $\{-1, 0, 1, \ldots, m\}$ one obtains $K^* = (1 - e^{-\lambda^*})(E[S_1])^2/C$. In general, calculations are considerably simplified for integer scores, and in biomolecular sequence analysis applications we tend to use digitized scales; see (*7, 10, 21*).

21. V. Brendel, P. Bucher, I. R. Nourbakhsh, B. E. Blaisdell, S. Karlin, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 2002 (1992).
22. The proof of Eq. 3 exploits decisively the properties of martingale processes, especially the Wald family of martingales for sums of independent random variables [(*15*); see also, for example, (*19*)].
23. B. K. Kobilka *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 46 (1987); SWISS-PROT (*25*) file B2AR$HUMAN.
24. G. D. Fasman and W. A. Gilbert, *Trends Biochem. Sci.* **15**, 89 (1990).
25. A. Bairoch and B. Boeckmann, *Nucleic Acids Res.* **19**, 2247 (1991).
26. R. Schleif, *Science* **241**, 1182 (1988); P. J. Mitchell and R. Tjian, *ibid.* **245**, 371 (1989).
27. For a predetermined segment, the occurrences of charged residues are considered as outcomes of Bernoulli trials with probability of success equal to $f = f_+ + f_-$. For a segment of length W and charge count c, the normalized count $\hat{c} = (c - Wf)/[Wf(1 - f)]^{1/2}$ is compared with the standard Gaussian curve of the normal density function. If $\hat{c}$ falls in the extreme upper tail of the curve then a significantly high charge count is ascertained. In order to avoid excessive statistical fluctuations it is necessary to apply the above formula only to segment lengths W of ~25 to 75 residues and longer. Experience has shown that threshold values of 4.0 for sequences <500 residues, 4.5 for sequences between 500 and 1000 residues, and 5.0 for longer sequences are appropriate [S. Karlin, B. E. Blaisdell, E. S. Mocarski, V. Brendel, *J. Mol. Biol.* **205**, 165 (1989)]. Computer programs to identify charge clusters in protein sequences are available (*21*).
28. V. Brendel, J. Dohlman, E. B. Blaisdell, S. Karlin, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 1536 (1991).
29. For a sequence of length N and a letter occurring with frequency f, the probability of observing a run of this letter of length exceeding $L = \ln N/(-\ln f) + z$ is asymptotically at most $1 - \exp\{-(1 - f)f^z\}$. Setting this probability equal to 0.01 we obtain z and L corresponding to the length of runs significant at the 1% level. Formulas for estimating the significance of runs with errors and of periodic patterns (for example, charge occurring every second or third residue) are given in S. Karlin, B. E. Blaisdell, V. Brendel [*Methods Enzymol.* **183**, 388 (1990)]; see also (*57*).
30. For example, N. Sueoka, *Cold Spring Harbor Symp. Quant. Biol.* **26**, 35 (1961); H. Nakashima, K. Nishikawa, T. Ooi, *J. Biochem.* **99**, 153 (1986); R. F. Doolittle, in *Of URFS and ORFS* (University Science, Mill Valley, CA, 1986), pp. 55–59; P. McCaldon and P. Argos, *Proteins* **4**, 99 (1988); G. D'Onofrio, D. Mouchiroud, B. Aissani, C. Gautier, G. Bernardi, *J. Mol. Evol.* **32**, 504 (1991); T. Ikemura, K. Wada, S. Aota, *Genomics* **8**, 207 (1990).
31. M. Hollender and D. A. Wolfe, *Nonparametric Statistical Methods* (Wiley, New York, 1973); Y. M. Bishop, S. E. Feinberg, P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice* (MIT Press, Cambridge, MA, 1975).
32. R. F. Doolittle, D. F. Feng, M. S. Johnson, *Nature* **307**, 558 (1984); R. F. Doolittle *Trends Biochem. Sci.* **10**, 233 (1985).
33. J. Josse, A. D. Kaiser, A. Kornberg, *J. Biol. Chem.* **236**, 864 (1961).
34. A. L. Lehninger, *Biochemistry* (Worth, New York, 1975), p. 861; H. S. Shaprio, in *CRC Handbook of Biochemistry and Molecular Biology*, G. D. Fasman, Ed. (CRC Press, Cleveland, OH, 1976), vol. 3, pp. 241–281.
35. G. Bernardi *et al.*, *Science* **228**, 953 (1985); D. Mouchirond, C. Bautier, G. Bernardi, *J. Mol. Evol.* **28**, 7 (1988).
36. C. L. Schildkraut, J. Marmur, P. Doty, *J. Mol. Biol.* **4**, 430 (1962); J. Marmur and P. Doty, *ibid.* **5**, 109 (1962).
37. T. H. Jukes and V. Bhushan, *J. Mol. Evol.* **24**, 39 (1986); G. A. Schachtel, P. Bucher, E. S. Mocarski, B. E. Blaisdell, S. Karlin, *ibid.* **33**, 483 (1991).
38. A. P. Bird, *Nature* **321**, 209 (1986).
39. R. Nussinov, *J. Biol. Chem.* **256**, 8458 (1981); *J. Theor. Biol.* **125**, 219 (1981); S. Ohno, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 9630 (1988); E. Beutler, T. Gelbart, J. Han, J. A. Koziol, B. Beutler, *ibid.* **86**, 192 (1989); C. G. Kozhukhin and P. A. Pevzner, *Comput. Appl. Biosci.* **7**, 39 (1991); C. Burge, A. Campbell, S. Karlin, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1358 (1992).
40. B. H. Howard and K. Sakamoto, *New Biol.* **2**, 759 (1990); J. Jurka and A. Milosavljevic, *J. Mol. Evol.* **32**, 105 (1991); G. L. G. Mikolas, in *Molecular Evolutionary Genetics*, G. R. McIntyre, Ed. (Plenum, New York, 1985), pp. 241–321; H. F. Willard and J. S. Wade, *Trends Genet.* **3**, 192 (1990).
41. E. H. Blackburn, *Nature* **350**, 569 (1991).
42. Applications of scan-statistics analysis for a fixed-length sliding window pertain to phenomena such as clusters of disease in time, generalized birthday proximities, and rth nearest-neighbor problems. Early work on scan statistics focused mainly on exact formulae [see the bibliographic compilation of J. I. Naus, *Int. Stat. Rev.* **47**, 47 (1979)] exploiting calculations of coincidence probabilities in diffusion stochastic processes. More recent distributional studies of scan statistics concentrate largely on bounds and approximations [for example, S. Wallenstein and N. Neff, *Stat. Med.* **6**, 197 (1987); J. Glaz, *J. Am. Stat. Assoc.* **84**, 560 (1989)]. The asymptotic results reported here are based on the powerful Chen-Stein method of Poisson approximations [R. Arratia, L. Goldstein, L. Gordon, *Ann. Prob.* **17**, 9 (1989); *Stat. Sci.* **5**, 403 (1990); and (*43*)]. The theory extends to the case where the marker sites $\{X_i\}$ are generated in a Markov-dependent manner (*43*).
43. A. Dembo and S. Karlin, *Ann. Appl. Prob.* **2**, 304 (1992).
44. W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, ed. 3, 1968), vol. 1; S. Karlin and H. Taylor, *A Second Course in Stochastic Processes* (Academic Press, New York, 1981), chap. 13.
45. C. Murre, P. S. McCaw, D. Baltimore, *Cell* **56**, 777 (1989).
46. R. B. Inman, *J. Mol. Biol.* **18**, 464 (1966).
47. S. Karlin, C. Burge, A. W. Campbell, *Nucleic Acids Res.* **20**, 1363 (1992).
48. N. Cressie, *Aust. J. Stat.* **19**, 132 (1977); L. Holst, *J. Appl. Prob.* **17**, 284 (1980); and (*43*).
49. Y. Kohara, K. Akiyama, K. Isoro, *Cell* **50**, 495 (1987).
50. K. Rudd *et al.*, *Nucleic Acids Res.* **19**, 637 (1991).
51. G. A. Churchill, D. L. Daniels, M. S. Waterman, *ibid.* **18**, 589 (1990).
52. N. Sternberg, *J. Bacteriol.* **164**, 490 (1985).
53. R. McMacken, L. Silver, C. Georgopoulos, in Escherichia coli and Salmonella typhimurium *Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (American Society for Microbiology, Washington, DC, 1987), pp. 564–612.
54. F. M. Hamzeh, P. S. Lietman, W. Gibson, G. S. Hayward, *J. Virol.* **64**, 6184 (1990); D. G. Anders and S. M. Punturieri, *ibid.* **65**, 931 (1991); M. J. O. Masse, S. Karlin, G. A. Schachtel, E. S. Mocarski, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 5246 (1992).
55. W. Hammerschmidt and B. Sugden, *Cell* **55**, 427 (1988).
56. R. Arratia and M. S. Waterman, *Adv. Math.* **55**, 13 (1985).
57. S. Karlin, F. Ost, B. E. Blaisdell, in *Mathematical Methods for DNA Sequences*, M. Waterman, Ed. (CRC Press, Boca Raton, FL, 1989), pp. 133–157.
58. M. Vingron and P. Argos, *Comput. Appl. Math. Biosci.* **5**, 115 (1989); G. D. Schuler, S. F. Altschul, D. J. Lipman, *Proteins* **9**, 180 (1991); M. van Heel, *J. Mol. Biol.* **220**, 877 (1991); M.-Y. Leung, B. E. Blaisdell, C. Burge, S. Karlin, *ibid.* **211**, 1367 (1991); M. Zuker, *ibid.* **221**, 403 (1991).
59. V. Brendel, in *Math. Compt. Modelling* **16** (No. 6/7) 37 (1992).
60. D. J. McGeoch *et al.*, *J. Gen. Virol.* **69**, 1531 (1988).
61. A. J. Davison and J. E. Scott, *ibid.* **67**, 1759 (1986).
62. M. S. Chee *et al.*, *Curr. Top. Microbiol. Immunol.* **154**, 125 (1990).
63. R. Baer *et al.*, *Nature* **310**, 207 (1984).

64. We thank J. Brauman, D. Brutlag, C. Burge, A. Campbell, S. Henikoff, E. Mocarski, R. Sapolsky, L. Stryer, and M. Zuker for helpful comments. Supported in part by NIH grants HG00335-04 and GM10452-29 and NSF grant DMS86-06244.