The Annals of Statistics
1990, Vol. 18, No. 2, 571–581

STATISTICAL COMPOSITION OF HIGH-SCORING SEGMENTS FROM MOLECULAR SEQUENCES

BY SAMUEL KARLIN[1], AMIR DEMBO AND TSUTOMU KAWABATA

Stanford University, Stanford University and University of Electro-Communications

1. Introduction. Distinguishing features of sequences that are likely or not likely to occur by chance can be important aids for identifying molecular sequence features for experimental manipulation. Two methods for discerning nonrandom aspects of sequences are commonly used: (1) comparisons of the original sequences to corresponding random sequence models and (2) data shuffling protocols [for recent reviews in these respects, see Doolittle (1981), Altschul and Erickson (1985) and Karlin, Ost and Blaisdell (1989)].

We investigate a random model appropriate to the data that provides a benchmark for discerning distributional properties of various data statistics. In the *independence* random model successive letters of a sequence are generated independently so that letter a_j is selected with probability p_j. In the case of proteins (DNA), the p_j are usually specified as the actual amino acid (nucleotide) frequencies in the observed sequence. A random first order Markov model is governed by 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 (dinucleotide) frequencies in a protein (DNA) sequence. For these models, theoretical results (limit distributional properties) have been obtained previously for a variety of sequence statistics including the length of the longest run of a given letter or pattern (allowing for a fixed number or a prescribed fraction of errors), the length of the longest word in a sequence satisfying a specific relationship (e.g., r-fold repeat, dyad symmetry, charge complementarity) and counts and spacings of long repeats [Karlin, Ghandour, Ost, Tavaré and Korn (1983), Karlin and Ost (1985), Arratia and Waterman (1985, 1989), Gordon, Schilling and Waterman (1986), Arratia, Gordon and Waterman (1986, 1990), Foulser and Karlin (1987), Karlin and Ost (1987, 1988), Deheuvals and Devroye (1987), Rootzén (1988) and Karlin, Ost and Blaisdell (1989)]. Several of these analyses have been extended to deal with comparisons within and between multiple sequences, including the ascertainment and statistical characterization of long common words and multidimensional count occurrence distributions for various word relationships [e.g., Karlin and Ghandour (1985), Karlin (1986)].

Received October 1989; revised December 1989.
[1]Supported in part by NIH grants GM39907-02 and GM10452-26 and by NSF grant DMS-86-06244.
AMS 1980 *subject classifications*. Primary 60F05, 60G50.
Key words and phrases. Maximal segmental score and composition conditioned central limit theorem.

In this paper we present new probabilistic formulas for characterizing statistically significant sequence configurations with respect to a general scoring scheme associated with letter attributes and for enabling varying degrees in letter matches. We describe the asymptotic extremal distribution of high aggregate segment scores and the letter composition of high-scoring segments. A number of associated conditional Gaussian central limit laws are also described. We establish strong laws for the longest segment showing a quality q score, that is, an average score per letter of q or more. Theorem 1 allows one to calculate the asymptotic probability that some segment from a random sequence has score greater than any given value. In particular, one can tell when some segment score value occurs in the 1% tail from the distribution of all segment scores. The results have applications in two important contexts: (1) for the analysis of a single protein sequence with the objective of identifying segments with statistically significant high scores of, for example, hydrophobicity (nonaffinity to water), charge strength, glycosylation (sugar attachment) affinity, secondary structure potential and sequence signal motifs; (2) in multiple sequence comparisons for establishing evolutionary histories or protein segments with common function and/or structure.

Scoring assignments for amino acids can reflect on biochemical categorizations such as electrical charge and physical properties (e.g., molecular weight, shape). Other amino acid classifications can relate to evolutionary relationships [Dayhoff, Schwartz and Orcutt (1978)]. We designate the alphabet in use by $A = \{a_1, a_2, \ldots, a_r\}$ and the corresponding letter scores by $S = \{s_1, s_2, \ldots, s_r\}$. For DNA, $r = 4$; for codons, triplets of nucleotides that translate to amino acids, $r = 61$; for the standard amino acids, $r = 20$; and for the charge attributes of amino acids, $r = 3$ (see Karlin, Ost and Blaisdell (1989) on other amino acid alphabets).

It is useful to highlight some natural scoring assignments.

1.1. *Scores based on charge.* For the positively charged amino acids lysine and arginine, $s = +2$; for the negatively charged amino acids aspartate and glutamate, $s = -2$; for histidines, $s = 0.04$ (at pH 7.2 in blood serum) or $s = 0.44$ (at pH 6.1 in muscle cells); for other amino acids, $s = -1$.

1.2. *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.

1.3. *Scores derived from target frequencies.* In a random sequence the letters $\{a_1, \ldots, a_r\}$ 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 in Section 3, the scores $s_i = \log(q_i/p_i)$, $i = 1, 2, \ldots, r$ (a log likelihood ratio) are appropriate. Theorem 6 states that in a maximal or high-scoring segment of a random sequence, letter a_i tends to occur with the target frequency $q_i = p_i \exp(\theta^* s_i)$, where θ^* is as described just before Theorem 1, so that all s_i can be expressed as a log likelihood ratio

$s_i = \log(q_i/p_i)$, with the log taken to some suitable base. Thus, since any set of individual scores has an implicit set of target frequencies, the question of what is an appropriate set of scores can be cast as the question of what is an "optimal" set of target frequencies. In the context of molecular sequence analysis, in order to construct the appropriate set of scores we need merely to characterize the letter distributions for the type of region we seek to identify [see Karlin and Altschul (1990) for elaborations on this theme].

1.4. *Scores based on structure alphabets.* Dickerson and Geis (1983) classified amino acids into internal (i), external (e) and ambivalent (a) types, based on experimental and empirical compilations reflecting where certain amino acids tend to be found in protein three-dimensional structures. Specifically, using the one-letter code for amino acids, $i = \{F, I, L, M, V\}$, $e = \{D, E, K, R, H, Q, N\}$, and $a = \{A, C, G, P, S, T, W, Y\}$. This is a good alphabet for studying hydrophobicity. A scoring scheme more refined than the three-letter alphabet and quite consistent with the Kyte-Doolittle scale of hydropathy takes $s = 2$ for I, L, V; $s = 1$ for F, M, A, C; $s = 0$ for G, S, T, W, Y, P; $s = -1$ for N, Q, H, D, E; $s = -2$ for K, R. One can use any of more than 12 alternative scales that have been proposed for hydrophobicity [see von Heijne (1987)].

In the simplest model, the random sequence consists of letters drawn independently from the alphabet A with respective probabilities $\{p_1, p_2, \ldots, p_r\}$. Associated with each letter a_i is a score s_i. We are interested in the segment of the sequence with maximal additive score, the second largest, or the several top-scoring segments. We impose two essential restrictions on the set of scores. Specifically, *we require at least one score to be positive and the expected score per letter* $\mu = \sum p_i s_i$ *to be negative.* If $\mu > 0$, the maximal segment would tend to be the whole sequence, and this is not of interest. The case of $\mu = 0$ is discussed below. In many situations the assumption $\mu < 0$ is intrinsic. For example, in the simple case of runs of a letter type, $\mu = -\infty$. In the model of scores calculated using a set of "target frequencies" (see Section 1.3), whenever the frequencies $\{q_i\}$ are not identical to the $\{p_i\}$, then necessarily $\sum p_i s_i = \sum p_i \log(q_i/p_i) < 0$. Finally, for any set of scores $\{s_i\}$ with μ positive, the modified scores $s_i' = s_i - \alpha\mu$, and $\alpha > 1$, satisfies $\sum s_i' p_i < 0$. In this case the determination of a segment with a large score using of the $\{s_i'\}$ amounts to selecting a segment with score in excess of its statistical mean score (using the $\{s_i\}$) by at least the factor $\alpha > 1$.

2. The statistical model and the limit distribution for the maximal segment score. Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. random variables based on observations from a finite alphabet $\{a_i\}_1^r$ such that

$$\Pr\{X = s_i\} = p_i, \qquad i = 1, 2, \ldots, r,\ p_i > 0, \sum p_i = 1,$$

is interpreted in the manner that sampling the letter a_i yields a score s_i. Let

$\{S_m\}_1^n$, $S_0 = 0$, be the partial sum process. The quantity

$$M(n) = \sup_{0 \le k \le l \le n} (S_l - S_k) \tag{1}$$

corresponds to a segment of the sequence $\{S_m\}_1^n$ with maximal score. The only assumptions on the process $\{S_m\}_1^n$ used are

$$E[e^{\theta X}] < \infty \quad \text{for real } -\theta_1 < \theta < \theta_2, (\theta_1, \theta_2 \text{ positive})$$

$$\lim_{\theta \downarrow -\theta_1} E[e^{\theta X}] = \lim_{\theta \uparrow \theta_2} E[e^{\theta X}] = +\infty, \tag{2}$$

$$\mu = E[X] < 0, \quad \text{so that } \{S_m\} \text{ entails a negative drift}$$

and

$$\Pr\{X > 0\} > 0$$

(see later for discussion of the case $\mu = 0$). To facilitate the analysis of (1), the distributional properties of

$$T(y) = \inf\{n: M(n) > y\} \tag{3}$$

are germane and of independent interest.

A parameter fundamental to the limit distribution of $M(n)$ and $T(y)$ is the *unique positive* root θ^* of the equation $E[e^{\theta X}] = 1$, well defined by virtue of the negative mean of X.

THEOREM 1 [Iglehart (1972) and Karlin, Dembo and Kawabata (1990)]. *Under the conditions* (2), *when X is nonlattice*

$$\lim_{n \to \infty} \Pr\left\{M(n) - \frac{\ln n}{\theta^*} \le x\right\} = \exp\{-K^* e^{-\theta^* x}\}, \tag{4}$$

where

$$K^* = \frac{\exp\left\{-2 \sum_{k=1}^{\infty} \frac{1}{k} \left(E\left[e^{\theta^* S_k}; S_k < 0\right] + \Pr\{S_k \ge 0\}\right)\right\}}{\theta^* E[Xe^{\theta^* X}]} \tag{5}$$

[*the series of* (5) *converge geometrically fast*]. *For the case that X is a lattice variable of span* δ, (4) *is replaced by*

$$\begin{aligned} \exp\{-K_+ e^{-\theta^* x}\} &\le \liminf_{n \to \infty} \Pr\left\{M(n) - \frac{\ln n}{\theta^*} < x\right\} \\ &\le \limsup_{n \to \infty} \Pr\left\{M(n) - \frac{\ln n}{\theta^*} < x\right\} \\ &\le \exp\{-K_- e^{-\theta^* x}\}, \end{aligned} \tag{6}$$

where

$$K_- = \frac{\theta^* \delta}{e^{\theta^* \delta} - 1} K^*, \qquad K_+ = \frac{\theta^* \delta}{1 - e^{-\theta^* \delta}} K^*.$$

Iglehart (1972) obtained the result of this theorem in the nonlattice case.

Some molecular sequence examples illustrating the use of (4)–(6) are given in Section 4.

The calculation of K^* is much simplified for the score values $\{-m, \ldots, -1, 0, 1\}$ occurring with respective probabilities $\{p_{-m}, \ldots, p_{-1}, p_0, p_1\}$. In this case

$$K_- = (e^{-\theta^*} - e^{-2\theta^*}) E[Xe^{\theta^* X}].$$

For the score assignments $\{-1, 0, 1, \ldots, m\}$ occurring with probabilities $\{p_{-1}, p_0, p_1, \ldots, p_m\}$, we have

$$K_- = \frac{(e^{-\theta^*} - e^{-2\theta^*})(E[X])^2}{E[Xe^{\theta^* X}]}.$$

By virtue of equality for the events $\{T(y) < n\} = \{M(n) > y\}$, we deduce Theorem 2 on the basis of Theorem 1.

THEOREM 2. *For $\mu < 0$, then*

$$\lim_{y \to \infty} \Pr\{T(y)\exp\{-\theta^* y\} \leq t\} = 1 - \exp\{-K^* t\}, \qquad t > 0, \tag{7}$$

and this limit law holds regardless of whether X is lattice or nonlattice.

The results of Theorems 1 and 2 combined with the inherent monotonicity of $M(n)$ and $T(y)$ imply the strong laws of Corollary 1.

COROLLARY 1.

$$\frac{\theta^* M(n)}{\ln n} \to 1 \quad \text{a.s. } as\ n \to \infty$$

and

$$\frac{\ln T(y)}{\theta^* y} \to 1 \quad \text{a.s. } as\ y \to \infty.$$

The sample paths of $\{S_m\}$ divide the time frame $(0, n)$ into successive excursions of the nonnegative axis:

$$K_0 = 0, \qquad K_\nu = \min\{k: k \geq K_{\nu-1} + 1, S_k - S_{K_{\nu-1}} < 0\}, \qquad \nu = 1, 2, \ldots . \tag{8}$$

A concomitant of Theorem 1 is that the asymptotic ($n \to \infty$) distribution of the number of separate excursions attaining a score in excess of $\ln n/\theta^* + x$ is Poisson with parameter $K^* e^{-\theta^* x}$.

The analysis of $M(n)$ and $T(y)$ for $E[X] = \mu = 0$ with $\sigma^2 = E[X^2] > 0$ leads to different limit laws. In this case the growth rate of $M(n)$ is of order $\sqrt{n}$

rather than of order $\log n$, while that of $T(y)$ is y^2 instead of $\exp(\theta^* y)$. More precisely, we have the following.

THEOREM 1′. *For $\mu = 0$, then*

$$\lim_{n\to\infty} \Pr\{M(n) \le \sqrt{n}\, x\} = \sum_{k=0}^{\infty} (-1)^k \frac{(\sigma\sqrt{2})^k}{x^k} \frac{1}{\Gamma(1 + k/2)}.$$

The right side is a Mittag–Leffler function of order $\frac{1}{2}$.

3. The length and composition of high-scoring segments. For each $y > 0$ and K_ν, $\nu = 0, 1, 2, \ldots$ delineated in (8), we define

$$T_\nu(y) = \min\{m\colon m > K_\nu, \text{ and either } S_m - S_{K_\nu} < 0 \text{ or } S_m - S_{K_\nu} > y\}, \tag{9}$$
$$\nu = 0, 1, 2, \ldots .$$

Let

$$L_\nu(y) = T_\nu(y) - K_\nu. \tag{10}$$

THEOREM 3 [Dembo and Karlin (1990)]. *For $\mu < 0$ let the integer-valued random variable ν^* correspond to the start of the first excursion, where the event T_ν is realized by the condition $S_{T_\nu} - S_{K_\nu} > y$. Then,*

$$w^* \frac{L_{\nu^*}(y)}{y} \to 1 \quad a.s.\ as\ y \to \infty,\ w^* = E[Xe^{\theta^* X}], \tag{11}$$

and the following central limit theorem holds:

$$\frac{y - w^* L_{\nu^*}(y)}{\sqrt{L_{\nu^*}(y)}} \to N(0, v^*) \quad \text{in distribution as } y \to \infty \tag{12}$$

and $v^* = E[X^2 e^{\theta^* X}] - (E[Xe^{\theta^* X}])^2$.

The limit law (12) can be expressed as a conditional central limit law. Indeed, let $\mathscr{E}(y)$ be the event that the first exit from the interval $[0, y]$ of the partial sum process $\{S_m\}$ upcrosses the barrier y. Let $L(y)$ be the time duration of this event. Then (12) is synonymous with [cf. Siegmund (1975)]

$$E\left[\exp\left\{\frac{iz[y - w^* L(y)]}{\sqrt{L(y)}}\right\} \middle| \mathscr{E}(y)\right] \to \exp\{-z^2 v^*/2\} \quad \text{as } y \to \infty. \tag{13}$$

In the mean zero case $\mu = 0$, we have the Laplace transform limit,

$$\lim_{y\to\infty} E\left[\exp\left\{-\frac{s\sigma^2}{2}\frac{L(y)}{y^2}\right\} \middle| \mathscr{E}(y)\right] = \frac{\sqrt{s}}{\sinh\sqrt{s}},$$

where $\sigma^2 = E[X^2]$.

The following generalization of Theorem 3 contains information on the composition of high-scoring segments. Let $\mathbf{U}_m$ be i.i.d. d-vector random variables; $\mathbf{U}_m$ can depend on X_m but is independent of X_k, $k \neq m$. We form

$$\mathbf{W}_{\nu^*}(y) = \sum_{k=K_{\nu^*}+1}^{T_{\nu^*}(y)} \mathbf{U}_k \tag{14}$$

[see (8)–(10) concerning notation], so $\mathbf{W}_{\nu^*}(y)$ cumulates suitable functionals of the X samples in a high-scoring (reaching a level more than y) excursion.

THEOREM 4 [Dembo and Karlin (1990)]. *For $\mu < 0$, $\{\mathbf{U}_k, X_k\}$ i.i.d. $(d+1)$-tuples, $\mathbf{W}_{\nu^*}(y)$ defined in (14) and ν^* the index of the first excursion where a level exceeding y is attained before crossing to the negative real line, then*

$$\frac{\mathbf{W}_{\nu^*}(y)}{L_{\nu^*}(y)} \to \mathbf{u}^* \quad a.s.\ as\ y \to \infty, \qquad \mathbf{u}^* = E[\mathbf{U}e^{\theta^* X}], \tag{15}$$

and the following conditional central limit theorem holds:

$$\frac{1}{\sqrt{L_{\nu^*}(y)}} \sum_{k=K_{\nu^*}+1}^{T_{\nu^*}(y)} (\mathbf{U}_k - \mathbf{u}^*) \to N(\mathbf{0}, \Sigma^*) \quad \textit{in distribution}, \tag{16}$$

where $\Sigma^* = \|\sigma_{ij}^*\|$, $\sigma_{ij}^* = E[(U_i - u_i^*)(U_j - u_j^*)e^{\theta^* X}]$.

By specializing to the indicator function

$$U_k = \begin{cases} 1 & \text{if } X_k \in A \ (A \text{ a Borel set in the range of } X), \\ 0 & \text{otherwise,} \end{cases}$$

then $W_{\nu^*}(y)/L_{\nu^*}(y) = \mu(A; y)$ is the fraction of samples in A for a high excursion reaching higher than level y. For this special case Theorem 4 translates to the following.

THEOREM 5. *For a high-scoring segment the empirical distribution of samples $\mu(A; y)$ satisfies*

$$\mu(A; y) \to \mu^*(A) = E\left[I_A(X)e^{\theta^* X}\right], \quad a.s.\ as\ y \to \infty, \tag{17}$$

where $I_A(X) = 1$ for $X \in A$, 0 for $X \notin A$ and

$$\sqrt{L(y)}\,[\mu(A; y) - \mu^*(A)] \to N\left(0, (\sigma^*)^2\right),$$

where $(\sigma^*)^2 = E[(I_A(X))^2 e^{\theta^* X}] - (E[I_A(X)e^{\theta^* X}])^2$.

In the case where X assumes only discrete scores $\{s_i\}_1^r$ with corresponding probabilities $\{p_i\}_1^r$, then for high-scoring segments ($y \to \infty$) the relative fre-

quency of score s_i is approximated with probability $p_i e^{\theta^* s_i}$. We state this formally as follows.

THEOREM 6. *Let* $\Pr\{X = s_i\} = p_i$, $\Sigma p_i = 1$. *The frequency of letter* a_i *in any sufficiently high-scoring segment approaches* $p_i \exp(\theta^* s_i)$ *with probability* 1. *In particular, this is true for the segment of maximal excursion.*

A special case of Theorem 6 was considered in Arratia, Morris and Waterman (1988).

We describe in conclusion a strong law for long *quality* q segments. A quality q segment for the index range $\{k$ to $l\}$ of the partial sum process based on $\{X_i\}$ satisfies the condition

$$\frac{1}{l-k} \sum_{i=k+1}^{l} X_i \geq q,$$

so that the average score from $k+1$ to l is at least q. We consider only feasible q levels satisfying $E[X] < q < \max X$. In particular, a quality q segment yields a larger average score per letter.

For a sequence of length n, let $R_n^{(q)}$ be the *longest* quality q segment. The stipulations (2) on $\{X_i\}$ apply. The following theorem holds.

THEOREM 7. *Let* $R_n^{(q)}$ *be the length of the longest quality* q *segment. Let* $u(\theta) = \log E[e^{\theta X}] - \theta q$ *and determine* $\hat{\theta}$ *as the unique minimum of* $u(\theta)$ (*necessarily* $0 < \hat{\theta}$). *Then*

$$\lim_{n \to \infty} \left[\frac{-u(\hat{\theta}) R_n^{(q)}}{\ln n} \right] = 1 \quad a.s.$$

All the results described above extend to the situation where X_i is generated as a Markov chain subject to mild positivity conditions [see Karlin, Dembo and Kawabata (1990) and Dembo and Karlin (1990) for details]. The methods of proof rely heavily on martingale theory particularly exploiting the Wald martingale family in its Markov chain setting, fluctuation theory for partial sums of i.i.d. random variables and their extensions to Markov chains (appropriate Wiener–Hopf factorizations) and semi-Markov renewal limit laws for the excess random variable in exceeding a high level.

4. Examples. Theorem 1 allows one to calculate the asymptotic probability that some segment from a random letter sequence has score exceptionally high. In particular, one can tell when the $M(n)$ value occurs in the 1% or 5%

tail of its distribution. This at least provides a benchmark for assessing the statistical significance of high-scoring segments. A wide-ranging empirical study applying the statistical theory described in this paper will be published elsewhere [see also Karlin and Altschul (1990)]. We illustrate here the ideas and results with two examples from protein sequences, one displaying high-scoring concentrated charge segments and the other containing a high-scoring hydrophobic interval.

4.1. We wish to identify a high-scoring mixed charge region prominent with basic and acidic amino acids. To this end, we stipulate a scoring assignment such that $s = 2$ for the acidic amino acids aspartate, glutamate, and for the basic amino acids lysine, arginine and histidine, but the score -1 for all other amino acids. Consider the human keratin (found in fingernails and hair) 67K cytoskeletal type II protein (length 643 amino acids) with its frequency of charged amino acids of 20.1%. The maximal scoring segment is located at positions 238–291 of aggregate score 21, probability P^* of achieving this level or higher by formula (4) is less than 0.008, the second-highest distinct scoring segment is located at positions 427–463, score 14; $P^* \approx 0.15$. These segments of charge concentrations are postulated to be functionally important for the keratin protein [cf. Brendel and Karlin (1989), Karlin, Blaisdell, Mocarski and Brendel (1989)].

4.2. Score assignments for hydrophobic attributes (we use the one-letter code for amino acids): $s = +1$ for I, L, V, F, M, C, A; $s = -1$ for G, S, T, W, Y, P and $s = -2$ otherwise.

The recently characterized cystic fibrosis (CF) protein [Riordan et al. (1989)] is composed of 1480 amino acids, and in this case the frequency $s = +1$ is 41.6% and frequency of $s = -1$ is 26.8%. The maximal segment occurs at positions 986–1029, score 21, $P^* < 0.001$, and the second maximal segment is found at 859–884, score 17, $P^* < 0.012$. The latter region is preceded by an acidic charge cluster at positions 819–838. Sequence comparisons in Riordan et al. (1989) project CF as structurally similar to a membrane associated transport protein; the existence of these pronounced hydrophobic segments is consistent with characteristics of an integral membrane protein.

Acknowledgments. We thank R. Olshen and D. Siegmund for valuable comments and references relating to the paper.

REFERENCES

ALTSCHUL, S. F. and ERICKSON, B. W. (1985). Significance of nucleotide sequence alignments: A method for random sequence permutations that preserves dinucleotide and codon usage. *Molecular Biol. Evol.* **2** 526–538.

ARRATIA, R. and WATERMAN, M. S. (1985). An Erdös–Rényi law with shifts. *Adv. in Math.* **55** 13–23.

ARRATIA, R. and WATERMAN, M. S. (1989). The Erdös–Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.* **17** 1152–1169.

ARRATIA, R., GORDON, L. and WATERMAN, M. S. (1986). An extreme value theory for sequence matching. *Ann. Statist.* **14** 971–993.

ARRATIA, R., GORDON, L. and WATERMAN, M. S. (1990). The Erdös–Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.* **18** 539–570.

ARRATIA, R., MORRIS, P. and WATERMAN, M. S. (1988). Stochastic scrabble: Large deviations for sequences with scores. *J. Appl. Probab.* **25** 106–119.

BRENDEL, V. and KARLIN, S. (1989). Association of charges clusters with functional domains of cellular transcription factors. *Proc. Nat. Acad. Sci. U.S.A.* **86** 5698–5702.

DAYHOFF, M. O., SCHWARTZ, R. M. and ORCUTT, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* **5** 345–352. National Biomedical Research Foundation, Washington, D.C.

DEHEUVALS, P. and DEVROYE, L. (1987). Limit laws of Erdös–Rényi–Shepp type. *Ann. Probab.* **15** 1363–1386.

DEMBO, A. and KARLIN, S. (1990). Limit distributions of empirical functionals for large exceedences of partial sum of i.i.d. variables. Unpublished manuscript.

DICKERSON, R. E. and GEIS, I. (1983). *Hemoglobin Structure, Function, Evolution and Pathology.* Benjamin Cummings, Menlo Park, Calif.

DOOLITTLE, R. F. (1981). Similar amino acid sequences: Chance or common ancestry. *Science.* **214** 149–159.

FOULSER D. and KARLIN, S. (1987). Maximal success runs for semi-Markov processes. *Stochastic Process. Appl.* **24** 203.

GORDON, L., SCHILLING, M. F. and WATERMAN, M. S. (1986). An extreme value theory for long head runs. *Probab. Theory Related Fields* **72** 279–287.

IGLEHART, D. (1972). Extreme values in the GI/G/1 queue. *Ann. Math. Statist.* **43** 627–635.

KARLIN, S. (1986). Comparative analysis of structural relationships in DNA and protein sequences. In *Evolutionary Processes and Theory* (S. Karlin and E. Nevo, eds.) 329. Academic, Orlando.

KARLIN, S. and ALTSCHUL, S. F. (1990). New methods for assessing statistically significant molecular sequence features using general scoring schemes. *Proc. Nat. Acad. Sci. U.S.A.* **87** 2264–2268.

KARLIN, S. and GHANDOUR, G. (1985). Comparative statistics for DNA and protein sequences: Multiple sequence analysis: *Proc. Nat. Acad. Sci. U.S.A.* **82** 6186–6190.

KARLIN, S., DEMBO, A. and KAWABATA, T. (1990). Limit distribution of maximal segmental score among partial sums. In preparation.

KARLIN, S. and OST, F. (1985). Maximal segmental match length among random sequences from a finite alphabet. *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) **1** 225–243. Wadsworth, Monterey, Calif.

KARLIN, S. and OST, F. (1987). Counts of long aligned word matches among random letter sequences. *Adv. in Appl. Probab.* **19** 293–351.

KARLIN, S. and OST, F. (1988). Maximal length of common words among random letter sequences. *Ann. Probab.* **16** 535–563.

KARLIN, S., BLAISDELL, B. E., MOCARSKI, E. S. and BRENDEL, V. (1989). A method to identify distinctive charge configurations in protein sequences, with application to human herpesvirus polypeptides. *J. Mol. Biol.* **205** 165–177.

KARLIN, S., GHANDOUR, G., OST, F. TAVARÉ, S. and KORN, L. J. (1983). New approaches for computer analysis of nucleic acid sequences. *Proc. Nat. Acad. Sci. U.S.A.* **80** 5660–5664.

KARLIN, S. OST, F. and BLAISDELL, B. E. (1989). In *Mathematical Methods for DNA Sequences.* (M. S. Waterman, ed.) 133–158. CRC Press, Boca Raton, Fla.

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. and

TSUI, L. C. (1989). Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science* **241** 1066–1071.

ROOTZÉN, H. (1988). Maxima and exceedances of stationary Markov chains. *Adv. in Appl. Probab.* **20** 371–390.

SIEGMUND, D. (1975). The time until ruin in collective risk theory. *Mitt. Verein. Schweiz. Versicherungsmath.* **2** 157–166.

VON HEIJNE, G. (1987). *Sequence Analysis in Molecular Biology*. Academic, San Diego.

SAMUEL KARLIN
DEPARTMENT OF MATHEMATICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

AMIR DEMBO
DEPARTMENTS OF MATHEMATICS
AND STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

TSUTOMU KAWABATA
DEPARTMENT OF COMMUNICATIONS SYSTEMS
UNIVERSITY OF ELECTRO-COMMUNICATIONS
1-5-1, CHOFUGAOKA
CHOFU-SHI, TOKYO, 182
JAPAN