PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Statistical Studies of Biomolecular Sequences: Score-Based Methods

Samuel Karlin

Phil. Trans. R. Soc. Lond. B 1994 **344**, 391-402
doi: 10.1098/rstb.1994.0078

Statistical studies of biomolecular sequences: score-based methods

SAMUEL KARLIN

Department of Mathematics, Stanford University, Stanford, California 94305-2125, U.S.A.

SUMMARY

The massive accumulation of DNA and protein sequence data poses challenges and opportunities in terms of interpretation and analysis. This presentation reviews the method of score-based sequence analysis with the objectives of discerning distinctive segments in single sequences and identifying significant common segments in sequence comparisons. A number of new results are described here for both the theory and its applications. These include distributional theory involving several high scoring segments in single sequences, distribution formulas for general scoring régimes in multiple sequence comparisons, bounds for periodic scoring assignments, sensitivity analysis of genome composition and refinements on predicting exons and genes in DNA sequences.

1. INTRODUCTION

Research in molecular biology is generating great volumes of nucleic acid sequence, amino acid sequence, and macromolecular structure data from the genomes of many organisms. Acquisition of these data generally runs considerably ahead of interpretation. Among the objectives of nucleic acid and protein sequence analysis is discovering significant patterns and interpreting them with respect to genomic structure and organization, DNA and RNA processing, gene expression, protein folding, biochemical function and evolution.

Molecular sequence analysis has become an important tool in molecular biology (for example, see books edited by Waterman (1989), Doolittle (1990), and Gribskov & Devereux (1992)). An unusual pattern in a nucleic acid or protein sequence, or a region of strong similarity shared by two or more sequences, putatively correlate with biological function or structure. Statistically distinctive sequence features on the protein level include extremes in certain residue usages, anomalous distributions of charged or other residue types, repetitive sequences (e.g. periodicities, multiplets) and unusual spacings of amino acid types (EGF-like domains, cysteine kringles); for examples, see Brendel *et al.* (1992) and Karlin *et al.* (1992*b*).

Among recent tools for detecting interesting regions in protein sequences are score-based methods. The theory has been developed in two contexts: (i) analysis of a single sequence seeking to identify sequence features that correspond to segments of significant high cumulative score (Karlin & Dembo 1992; for applications, see Karlin & Brendel 1992); and (ii) analysis of multiple sequences seeking to identify common segments having high total similarity score (for applications, see Altschul *et al.* 1990; Green *et al.* 1993). Sequence patterns that can be investigated with the aid of the scoring method include peptides featuring amino acid size anomalies, distinctive hydrophobic sections, regions prone to phosphorylation or glycosylation modifications, and segments of particular secondary structure potential. For single-sequence analysis, scores appropriate for the detection of clusters of certain amino acid types (e.g. charge clusters, transmembrane segments and DNA-binding domains) have been described (Brendel *et al.* 1992; Karlin & Brendel 1992). For sequence comparison, a wide range of scoring régimes have been proposed (Dayhoff *et al.* 1978; Feng *et al.* 1985; Altschul *et al.* 1990; Gonnet *et al.* 1992; Jones *et al.* 1992; Henikoff & Henikoff 1992; Brendel *et al.* 1994).

2. PROBABILITIES OF HIGH SCORING SEGMENTS

In Karlin *et al.* (1990*b*) and Karlin & Dembo (1992) (see also Arratia & Waterman 1985, 1989; Arratia *et al.* 1988, 1990) we presented probabilistic formulas for characterizing statistically significant sequence configurations with respect to a general scoring scheme associated with letter attributes and with varying degrees of similarity in letter matches. The simplest model is as follows: Let $X_1, X_2, \ldots, X_n$ be independent identically distributed letters drawn from a finite alphabet $\{a_i\}_1^r$ such that $\text{Prob}\{X = a_i\} = \text{Prob}\{X = s_i\} = p_i$, $i = 1, 2, \ldots, r$, $p_i > 0$, $\sum p_i = 1$, interpreted as sampling the letter a_i yields a score $X = s_i$ where $r = 4$ for DNA, $r = 20$ for amino acids, $r = 64$ for codons. Theory exists for the more complicated case of Markov-dependent sequences but will not be discussed here (see Karlin & Dembo 1992). Let

$$S_0 = 0, \qquad S_m = \sum_{i=1}^{m} X_i, \qquad m = 1, 2, \ldots,$$

be the cumulative score process. The quantity

Phil. Trans. R. Soc. Lond. B (1994) **344,** 391–402
Printed in Great Britain

$M_n = \sup_{0 \leqslant k \leqslant l \leqslant n}(S_l - S_k)$ corresponds to a segment of the sequence $\{S_m\}_0^n$ with maximal score. The essential assumptions are that $\{S_m\}$ entails a negative drift (negative mean score for each X_i) and some s_i is positive. A parameter fundamental to the limit distribution of M_n is the unique positive root θ^* of the equation $E[\exp(\theta^* X)] = 1$ (E refers to expectation). It is proved (Karlin & Dembo 1992) for n large that

$$\Pr\left[M_n > \frac{\ln n}{\theta^*} + x\right] \approx 1 - \exp\{-K^* e^{-\theta^* x}\}, \qquad (1)$$

with accessible computation for K^* and θ^* (see Karlin & Altschul 1990). The asymptotic formula (1) can be used to establish benchmarks of statistical significance for various distinctive segment features such as hydrophobicity, charge and DNA binding (e.g. Karlin & Brendel 1992). For this purpose we set the right-hand side of (1) to some significance level, for example, $p^* = 0.01$, and solve for $x^* = x(p^*)$. A maximal segment score exceeding $(\ln n/\theta^*) + x^*$ is significant at the p^* level.

The analysis also provides information on the composition of high-scoring segments. For each $y > 0$, let $L(y) = T(y) - K(y)$ be the length of the first segment extending from $K(y) + 1$ to $T(y)$ of aggregate score exceeding y. Let $\boldsymbol{U}_m$ be independent vector random variables where $\boldsymbol{U}_m$ is independent of $X_k, k \neq m$. Form

$$\boldsymbol{W}(y) = \sum_{K(y)+1}^{T(y)} \boldsymbol{U}_k$$

so that $\boldsymbol{W}(y)$ cumulates functionals of the X samples in a high-scoring segment. Then $\boldsymbol{W}(y)/L(y) \to \boldsymbol{u}^*$ as $y \to \infty, \boldsymbol{u}^* = E[\boldsymbol{U}_1 e^{\theta^* X_1}]$ (Dembo & Karlin 1991).

Taking $X_k \in \mathcal{A}$ equal 1 and 0 otherwise, then $W(y)/L(y)$ is the fraction of samples in $\mathcal{A}$ that lie in a high-scoring segment. Thus, over high-scoring segments the relative frequency of score s_i is approximately $q_i = p_i \exp\{\theta^* s_i\}$. It follows that scores defined by

$$s_i = \ln(q_i/p_i), \qquad (2)$$

(a multiplicative scaling of s_i will not change any of the theory or its applications) identify high-scoring segments of target frequencies q_i (Karlin & Altschul 1990; Karlin & Brendel 1992). It should be emphasized that for any segment where $S_l - S_k$ is large, the letter frequencies are biased toward the values $q_i \simeq p_i e^{\theta^* s_i}$, $i = 1, \ldots, r$. Conversely, if the letters in a segment are distributed with frequencies q_i, then with high probability the aggregate score of this segment would be high.

Another statistic that is useful in appraising a given set of scores concerns the length $L(y)$ of a high scoring segment of aggregate score at least y (y large). An asymptotic confidence interval for $L(y)$ is given by

$$L(y) = \frac{y}{\nu} \pm \rho(\alpha)\frac{1}{\nu}\sqrt{\frac{yw}{\nu}}, \qquad (3)$$

where

$$\nu = \sum_{i=1}^{r} p_i s_i e^{\theta^* s_i}, \qquad w = \sum_{i=1}^{r} p_i s_i^2 e^{\theta^* s_i} - \nu^2,$$

and $\rho(\alpha)$ is the normal distribution quantile point for an $\alpha\%$ confidence interval (Dembo & Karlin 1993).

3. APPLICATIONS OF FORMULAS (1) AND (2)

There are natural score assignments for certain sequence features. Examples include:

1. Scores emphasizing positive charge. For the positively charged amino acids lysine and arginine, $s = +2$; for the negatively charged amino acids aspartate and glutamate, $s = -2$; for other amino acids, $s = -1$.
2. Scores for hydrophobic profile. One can use the Kyte-Doolittle scale or any of the many other scales that have been proposed for assessing hydrophobicity (see von Heijne 1978, Chapter 5; see also Brendel *et al.* 1992).
3. Scores derived from target frequencies. Let $\{q_1, q_2, \ldots, q_r\}$ be a set of desirable target frequencies of the letter types, and $\{p_1, \ldots, p_r\}$ the average letter frequencies. In certain contexts, the scores $s_i = \log(q_i/p_i), i = 1, 2, \ldots, r$ are appropriate since in a high-scoring segment letter a_i tends to occur with the target frequency $q_i = p_i \exp(\theta^* s_i)$, see formula (2).

Example of a mixed charge cluster

Assign the score values $s = 2$ for the acidic amino acids aspartate and 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 a frequency of charged amino acids of 20.1%. The maximal scoring segment is located at positions 238–291 (contains 11 basic and 14 acidic residues) of aggregate score 21 with probability p^* of achieving this level or higher by formula (1) less than 0.008. This maximal scoring segment of charge concentration is postulated to be functionally important for the keratin protein. The keratin protein also contains two significantly long uncharged segments at positions 42–152 and 518–586.

4. SOME EXTENSIONS OF SCORE-BASED METHODOLOGY

(*a*) *Distributional properties for sums of high-scoring segments*

Applications of the scoring method often concern the sum of the r highest segment scores. This measure is appropriate when there may be several distinct segments of a given type within a DNA or protein sequence (e.g. multiple purine tracts, several transmembrane segments, multiple charge clusters). For sequence comparisons, the existence of insertions or deletions can break an alignment into several pieces. The sum of the scores of these pieces can be an appropriate measure of sequence similarity. Denote the r-highest distinct scoring segments of the model (1)

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

as $M_n^{(1)} = M_n, M_n^{(2)}, \ldots, M_n^{(r)}$. It is convenient to deal with the centered segment scores

$$S_n^{(i)} = M_n^{(i)} - \frac{\log nK^*}{\theta^*} \qquad i = 1, 2, \ldots, r. \tag{4}$$

It can be proved that the joint limiting density of (4) is

$$f(x_1, \ldots, x_r) = (\theta^*)^r \exp\{-e^{-\theta^* x_r}\} e^{-\theta^*(x_1 + x_2 + \ldots + x_r)}, \tag{5}$$

defined on the domain $x_r \leqslant x_{r-1} \leqslant \ldots \leqslant x_1$ (Karlin & Altschul 1993). For the variable

$$\bar{S}_{n,r} = \sum_{i=1}^{r} S_n^{(i)}$$

integrating out the accessory variables from (5) we deduce the limit density ($n \to \infty$) of $\bar{S}_{n,r}$ as

$$f_r(x) = (\theta^*)^r \frac{e^{-\theta^* x}}{r!(r-2)!} \int_0^\infty y^{r-2} \exp\{-e^{-(\theta^*/r)(x-y)}\} dy, \tag{6}$$

or equivalently

$$\lim_{n\to\infty} \Pr\{M_n^{(1)} + \ldots + M_n^{(r)} - r\frac{\log n K^*}{\theta^*} = x\} = f_r(x).$$

From the joint distribution formula the distribution of the r-th distinct highest segment score is easily derived having the density

$$\Pr\left\{M_n^{(r)} - \frac{\log nK^*}{\theta^*} = x\right\} = g_r(x) = \frac{\theta^*}{(r-1)!} e^{-r\theta^* x} \exp\{-e^{-\theta^* x}\}, \quad -\infty < x < \infty, \text{ for } n \text{ large.}$$

(*b*) *Periodic scoring schemes*

Periodic scoring schemes are appropriate for detecting amphipathic helices and other periodic sequence structures. We illustrate the period-2 model. Suppose a letter sequence is generated by scores $\{s_i\}_1^r$ following the probabilities $\{p_i\}_1^r$ at the even-sequence positions and scores $\{s_i'\}_1^r$ with probabilities $\{p_i'\}_1^r$ at the odd-sequence positions. In line with the essential requirements in our sequence analysis of a negative mean score per letter, we assume $\sum(s_i p_i + s_i' p_i') < 0$ and at least for one pair i, j, s_i and s_j' are positive. Denote $s_i + s_j' = s_{ij}$ and $p_i p_j' = \pi_{ij}$. Determine θ^* such that

$$\sum_{i,j=1}^{r} \pi_{ij} e^{\theta^* s_{ij}} = \left(\sum_{i=1}^{r} p_i e^{\theta^* s_i}\right)\left(\sum_{j=1}^{r} p_j' e^{\theta^* s_j'}\right) = 1 \text{ and } K^*$$

(as in formula (1)) corresponding to the scores $(s_i + s_j')$ occurring with probabilities π_{ij}. For this model the probability law for the maximal segment score M_n obeys the inequalities

$$\Pr\left\{M_n \geqslant \frac{\log n}{\theta^*} + x\right\} \leqslant \max\left\{1 - \sum_{i=1}^{r} p_i \exp\{-K^* e^{-\theta^*(x - s_i)}\}, \; 1 - \sum_{i=1}^{r} p_i' \exp\{-K^* e^{-\theta^*(x - s_i')}\}\right\}$$

(*c*) *Distribution theory for high scoring segments allowing for a limited number of insertions or deletions*

For a single sequence, consider the maximal scoring segment of length Δ_n (n refers to the length of the original sequence). We know that

$$\Delta_n / \log n \to \frac{1}{\theta^* w^*}, \qquad w^* = \sum_{i=1}^{r} s_i p_i e^{\theta^* s_i} > 0,$$

so that the length Δ_n is of order $\log n$. The empirical frequencies of the different letters over the segment Δ_n follow the target frequencies $p_i e^{s_i \theta^*}$. We delete from the maximal segment k occurrences of the letter $a^* = a_{i^*}$ whose score value satisfies $-s^* = \max(-s_i)$, i.e. s^* has the most negative value. For any finite k (or even infinite k of smaller order than $\log n$ (e.g. $k \leqslant \log\log n$) the number of occurrences of a_{i^*} in the maximal segment is asymptotically $p_{i^*} \exp(\theta^* s^*) \log n$. The score $M_n^* = M_n - ks^*$ gives the maximal score of an interval allowing k deletions. Since

$$\Pr\left\{M_n < \frac{\log n}{\theta^*} + x\right\} \approx e^{-K^* e^{-\theta^* x}},$$

then

$$\Pr\left\{M_n^* = M_n - ks^* < \frac{\log n}{\theta^*} + x - ks^*\right\} \approx e^{-K^* e^{-\theta^* x}},$$

or

$$\Pr\left\{M_n^* < \frac{\log n}{\theta^*} + y\right\} \approx \exp\{-K^* e^{-ks^*} e^{-\theta^* y}\},$$

and statistical significance for maximal scoring segments allowing k deletions can be evaluated via the last formula. Corresponding constructions can be used in sequence matching allowing for at most k aligned letter deletions.

(*d*) *Large exceedances in vector scores*

In vector scoring of sequences, successive positions are vectors $\boldsymbol{X}_i \in \boldsymbol{R}^d$ (Euclidean d-dimension) with components of different attributes in the sequence position. For example, for protein sequences, the components could be simultaneous charge, hydrophobicity, and steric measurements of the amino acids. High quality segments correspond to indices $K(y)$ and $T(y)$ of the sequence such that $S_{T(y)} - S_{K(y)}$ first attains a multivariate score corresponding to a set A. Consider the first segment of S_n which hits a rare set yA with y large (yA consists of all vectors $y\boldsymbol{a}, \boldsymbol{a} \in A$). Thus, if A requires that the minima of each coordinate are at least c, then a multivariate high score is achieved provided over the segment all the cumulative scores for each coordinate is at least yc. For detailed developments of the above theory, see Dembo *et al.* (1994*a*).

(*e*) *Sequence matching and significant average segment score (SAS value)*

In DNA and protein sequence matching, segment scores are of the form

$$\sum_{l=1}^{t} F(a_{i+l}, a_{j+l}'),$$

where a_i is the i-th letter in the first sequence, a'_j is the j-th letter in the second sequence and $F(x, y)$ is the score for the letter pair (x, y). For amino acid sequence comparisons, a wide range of scoring régimes have been proposed. The PAM, BLOSUM, and SISS scoring matrices will identify high scoring segments common to different protein sequences. The PAM similarity scores (Dayhoff *et al.* 1978; Jones *et al.* 1992) have been developed from considerations of evolutionary amino acid replacements in homologous genes from different species. The BLOSUM score matrices were constructed centering on blocks of functional motifs from various protein classes (Henikoff & Henikoff 1992). The SISS scoring scheme is based on screening of statistically significant long segments among protein sequences (Brendel *et al.* 1994). All segment pairs with scores significant at the 1% level can be identified (i.e. those with probability less than 0.01 of attaining a score at least as high for a segment pair in random sequences of the same lengths and amino acid frequencies). One way of scoring global similarity between two protein sequences is as follows. For each pair of protein sequences, the significant average similarity score (SAS value) is the maximal value with respect to all consistently ordered high (significant) scoring segments (overlaps are eliminated) calculated by summing these segment scores and dividing by the minimal length of the two protein sequences (Karlin *et al.* 1994). How should one interpret SAS values? As the score for amino acid identities with the PAM 120 matrix average, about 5.3, a SAS value of 2.00 generally reflects about 30%–40% identify, a SAS value of 3.00 corresponds to about 50%–60% identify, and a SAS value exceeding 4.00 carries at least 75% identify.

We illustrate the method of SAS values on the major virion glycoprotein B (VGLB) sequence available in 11 herpesvirus genomes. Herpesviruses are widespread in vertebrate species, sharing several moderately to well-conserved genes based on amino acid identity comparisons (e.g. DNA polymerase, major capsid protein, VGLB) even though they exhibit a dramatic variation in mean G+C genomic frequency ranging from 35% to 75% (Honess 1984). The herpesviruses are commonly perceived to be of ancient origin, at least 300 million years old. On the basis of biological characteristics, tissue tropism, genomic organization and amino acid block identities, the herpesviruses are classified into α, β and γ types. The α-herpesviruses include EHV1 (equine herpesvirus 1, host horse), HSV1 (herpes simplex virus 1, human), VZV (varicello-zoster virus, human), PRV1 (Pseudorabies virus 1, pig), MDV (Marek disease virus, turkey), BHV1 (bovine herpesvirus 1, cow), BHV2 (bovine herpesvirus 2, cow); the β-herpesviruses include HHV6 (human herpesvirus 6) and HCMV (human cytomegalovirus); the γ-herpesviruses include EBV (Epstein-Barr virus, human) and HVS (herpes saimiri virus, green monkey).

The length of the (VGLB) gene averages about 900 residues, slightly less in HVS, and only a fragment of length 259 residues was available from HHV6. Among the α sequences the SAS values were in good agreement, see table 1. The within γ-class SAS value was 2.00, whereas the between γ and α sequence

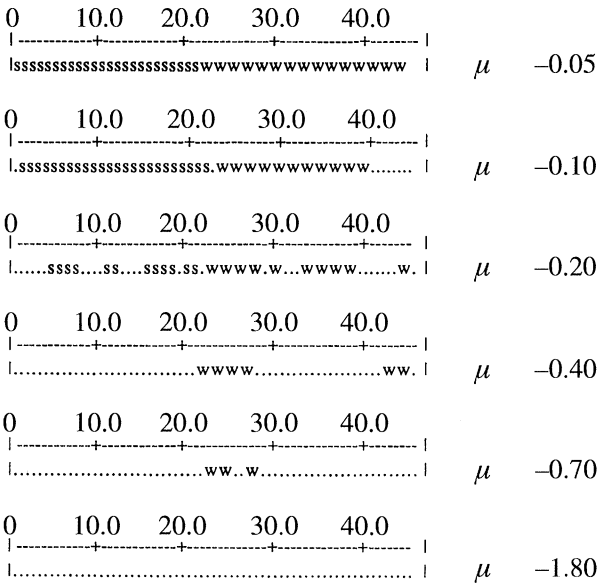

Figure 1. Map of strong and weak segments in lambda phage for various score stringencies. The figure describes schematically the strong and weak segments of the lambda (λ) phage genome (45 815 b.p.). It is clear and originally demonstrated experimentally (Inman 1966) that λ divides into two halves, one relatively G+C% rich (about 55%) and the other G+C% poor (about 45%); see Karlin *et al.* (1992*a*) for further discussion.

Table 1. *SAS-scores for glycoprotein B of herpesviruses*

VGLB	Length	EBV	HVS	HCMV	HHV6f	EHV1	HSV1	VZV	PRV1	MDV	BHV1	BHV2
EBV	(857)	—										
HVS	(808)	2.00	—									
HCMV	(906)	1.13	1.09	—								
HHV6f	(259)	1.22	1.06	1.48	—							
EHV1	(980)	0.81	0.89	0.77	0.86	—						
HSV1	(904)	0.93	0.82	0.87	0.70	2.52	—					
VZV	(868)	0.88	0.80	0.85	0.65	2.76	2.53	—				
PRV1	(913)	0.88	0.68	0.78	0.79	2.83	2.62	2.89	—			
MDV	(865)	0.80	0.88	0.90	0.69	2.28	2.54	2.54	2.20	—		
BHV1	(928)	0.69	0.72	0.68	0.80	2.57	2.39	2.78	2.84	2.36	—	
BHV2	(917)	0.86	0.83	0.86	0.73	2.41	3.56	2.59	2.50	2.48	2.26	—

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Table 2. *Score sensitivity analysis of three two-letter alphabets*

(1. Stringent S segments of EBV tend to occur with short tandem repeat sequences predominantly associated with the latent genes of EBV. The only stringent W segment in EBV abut 5′ the 3 kb repeats. The second most concentrated A+T region is the ori-P (origin of latent replication) region of EBV about 60% A+T, but this region does not yield any significant stringent W segments (cf. Karlin 1986).

2. The paramount high-scoring R segments are also associated with two primary latent (EBNA 1 and EBNA 2) genes of EBV. The only stringent Y segments are part of the 3 kb repeats.

3. In the amino-keto alphabet, again the only rich M and K segments are connected with repeat elements of latent genes.

Interpretations and implications of this sensitivity analysis applied to all the human herpes virus genomes are discussed in Karlin & Cardon (1994).)

Table 2*a*. *High-scoring strong (S) and weak (W) nucleotide segments (sensitivity analysis)*

		mean score per letter (μ)	significant segment	length	comments
Epstein-Barr virus (EBV)	S rich	−1	50 608–52 102	1494	most of 12 × 125 b.p. repeats
(length 172 282 b.p.)			70 400–70 516	116	structural virion gene BDLF1:
G+C% = 59.7					tegument, overlaps 9 × 15 b.p. repeats
(Baer *et al.* 1984)			100 131–100 231	100	part of 10×'15 b.p.' repeat in EBNA3
			170 146–170 294	149	in terminal repeats
			170 684–170 833	149	in terminal repeats
			171 207–171 356	149	in terminal repeats
			171 745–171 894	149	in terminal repeats
		−1.5	50 614–52 082	1468	
			70 400–70 516	116	
		−2	70 400–70 516		
		−4	none		
	W rich	−1	11 519–11 605	86	5′ proximal to 3 kb repeats
			11 855–11 950	95	
		−1.5	11 855–11 909	54	
		−2	none		

Table 2*b*. *High-scoring purine (R) and pyrimidine (Y) segments*

		mean score per letter (μ)	significant segment	length	comments
EBV	R rich	−1	49 525–49 580	55	part of EBNA 2
(R = A+G freq. = 49.2%)			108 044–108 108	84	part of EBNA 1
			108 211–109 089	878	includes long *ala gly* tract in EBNA 1
			169 105–169 194	89	part of LMP (latent membrane protein)
		−1.5	108 211–108 932	721	
		−2	108 216–108 932	716	
		−4	none		
	Y rich	−1	12 077–12 138	61	in 3 kb repeats
			15 149–15 210	61	in 3 kb repeats
			⋮		
		−1.5	none		

Table 2*c*. *High-scoring amino (M) and keto (K) segments*

		mean score per letter (μ)	significant segment	length	comments
EBV	M rich	−1.0	48 678–48 802	124	14 of CCCCCACCA repeats of EBNA 2
M% = 50.2%		−10	48 689–48 757	68	
	K rich	−1.0	90 498–90 611	113	in BLLF 1 contained in six copies of 21 b.p. approximate repeats
		−2.0	none		

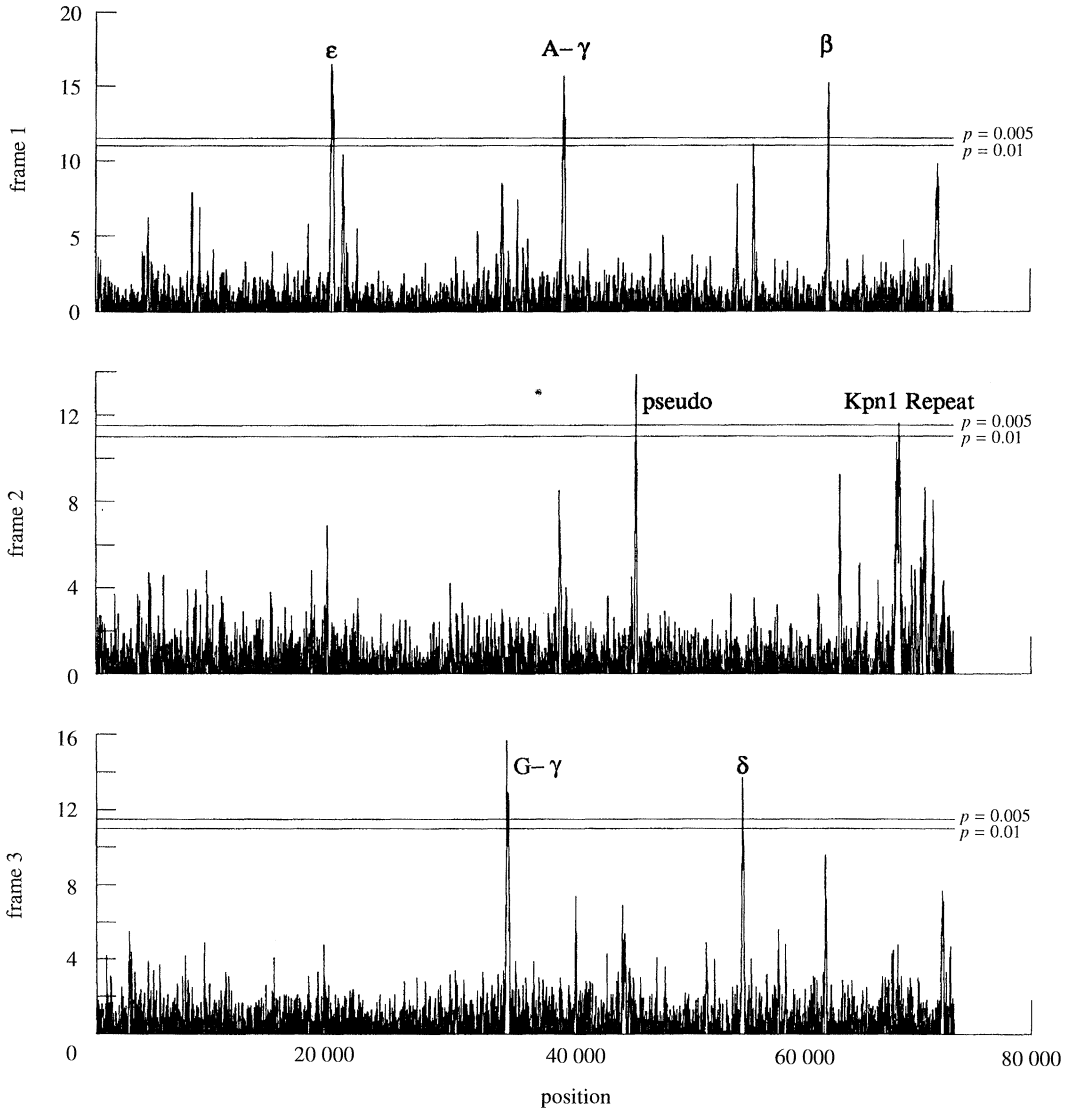

Figure 2. Plots of excursion scores for human beta globin region on chromosome 11 (73 326 b.p.). Each plot shows excursion scores for one reading frame: frames 1–3 are in the direction $5' \rightarrow 3'$ as presented in Genbank, frames 4–6 are in the direction $3' \rightarrow 5'$. The ordinate of each plot is determined by the recursion $E_0 = 0, E_i = \max\{E_{i=1} + X_i, 0\}$, for scores X_i at position $i = 1, 2, \ldots$, traversing the sequence. Thus, when the score sums become non-positive, an excursion has ended and cumulative scores start from zero for the next excursion. Excursions are regarded as high scoring segments (HSS) if they exceed the 0.01 significance level of the maximal segment score.

comparisons yielded SAS values in the range 0.68–0.93. The comparison of γ sequences with the HCMV and HHV6 representatives of β sequences gave higher SAS values, around 1.10. The SAS value for the β sequences versus each α sequence yielded SAS values in the range 0.65–0.90, about the same as for a γ sequence compared to an α sequence. The comparisons with respect to most homologous proteins (including VGLB) tend to produce SAS values among the α-herpesviruses significantly greater than the SAS value (Karlin *et al.* 1994) within the β-herpesviruses (HHV6 versus HCMV) and within the γ-herpesviruses (EBV versus HVS). The diminished SAS values for the within β-class protein comparisons relative to the within α- and γ-herpesviruses suggest that HHV6 and HCMV separated earlier than did the other herpesviruses. The higher SAS values among the α-herpesviruses supports the hypotheses that α-herpesviruses are of more recent ancestry. For elaborations on this example, see Karlin *et al.* (1994).

5. GENOMIC SENSITIVITY ANALYSIS

Local clustering in nucleotide sequences can be explored by scoring statistics at different levels of sensitivity. To illustrate, consider the task of identifying segments of significantly high A+T content. We assign a positive integer score s_w to weakly bonding bases (A and T) and a negative integer score s_s to strongly bonding bases (C and G) such that the expected score per nucleotide $\mu = p_w s_w + p_s s_s$ is negative, where p_w and p_s are the frequencies of A+T and C+G in the sequence, respectively. Different values for μ can be attained by adjusting s_w and s_s appropriately, thus tuning the sensitivity of the method. For example, for $\mu = -0.5$ high-scoring segments would be long clusters of A+T

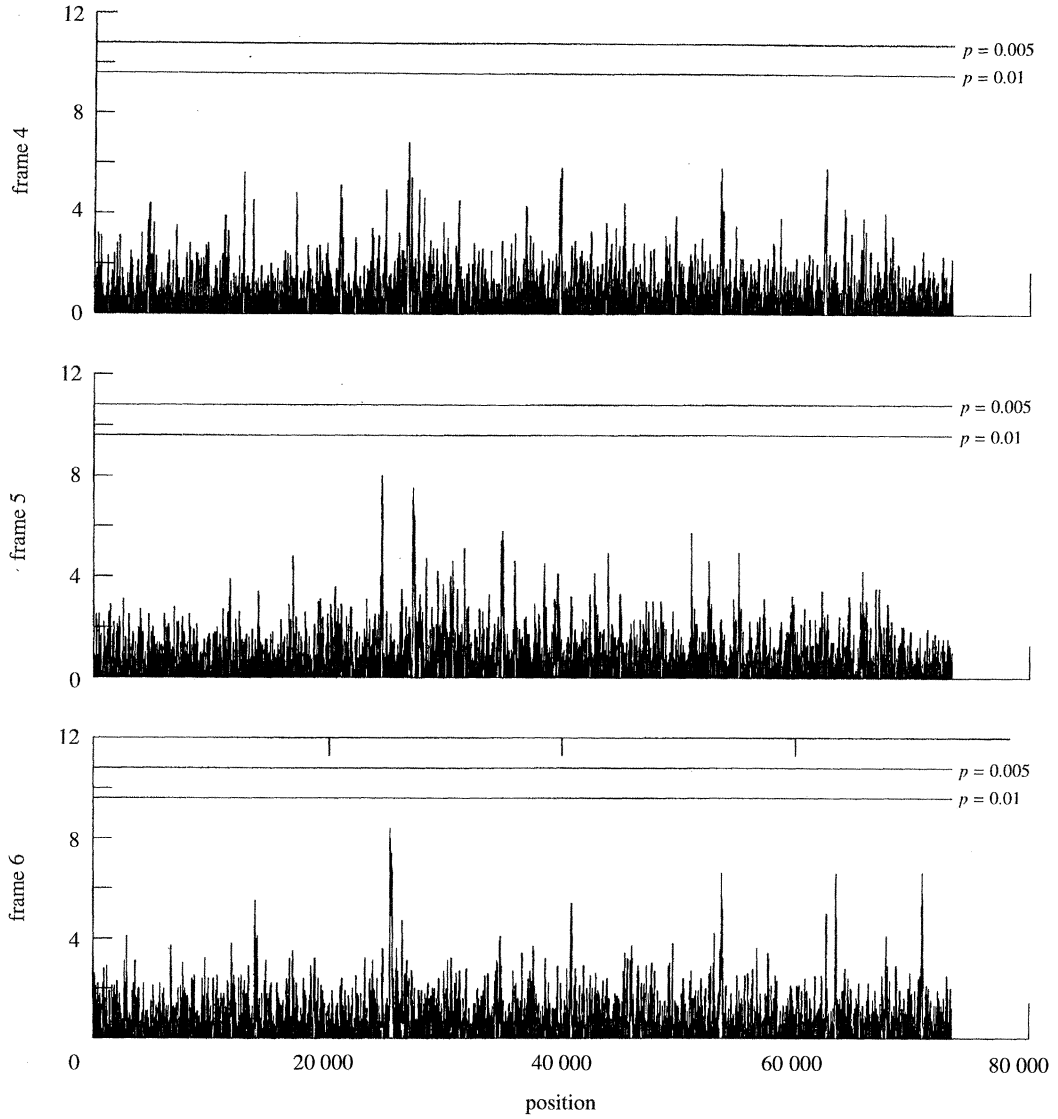

Figure 2. Continued.

(with a modest proportion of G + C bases interspersed), whereas $\mu = -10$ would tend to select mostly runs of A + T (allowing at most one C or G per 10 weakly bonding bases). In this way the scoring statistic provides a versatile tool that allows evaluation of significance of clustering both in terms of runs and long compartments (isochores).

We offer some examples of the score sensitivity analysis applied to the bacteriophage genome of lambda (λ) and to the human herpesvirus genome of Epstein-Barr virus (EBV). We will deal with the three two-letter alphabets strong (S) versus weak (W) bases: purine (R) versus pyrimidine (Y) where R = {G or A}, Y = {C or T}; keto {K} versus amino (M), K = {T or G}, M = {A or C}. To illustrate, consider a DNA sequence $\mathcal{S}$ and a given two-letter alphabet, say S versus W. We seek to identify high scoring C + G segments with increasing stringency. Accordingly, we set scores 1 for S bases and $-s$ for W bases with s calculated such that the expected score per nucleotide $1fs - sf_W = -\mu(\mu = 1,1.5,2,2.5,4,10)$ where $f_s = f_{G+C}$ is the frequency of strong bases in the sequence and f_W is the corresponding frequency of weak bases. For each specified μ, we determine all segments (of minimal length 50 b.p.) with a high score at the 0.99 significance level, i.e. the probability of observing such a segment score in a random sequence of the same S, W base composition is less than 0.01, cf. figure 1, see also table 2. The sensitivity analysis in the {S,W} alphabet is symmetric (synonymous) for a strand and its complementary strand. However, a purine-rich stretch corresponds to a pyrimidine-rich stretch of the other strand and similarly for the K versus M alphabet. This is true because in general, including the examples treated here, the purine frequency for sequences $\geqslant 50$ kb tend to be very close to 50% and the same for the amino frequency. On the other hand, the S (G + C) frequencies tend to be highly variable, for example, ranging in prokaryotic and eukaryotic organisms from 20% to 80%.

6. SCORE-BASED PREDICTIONS OF EXONS

Identification of all genes and the construction of genetic maps for many genomes (human, model organisms, e.g. mouse, *Drosophila*, *C. elegans*, yeast, viruses) is a major

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

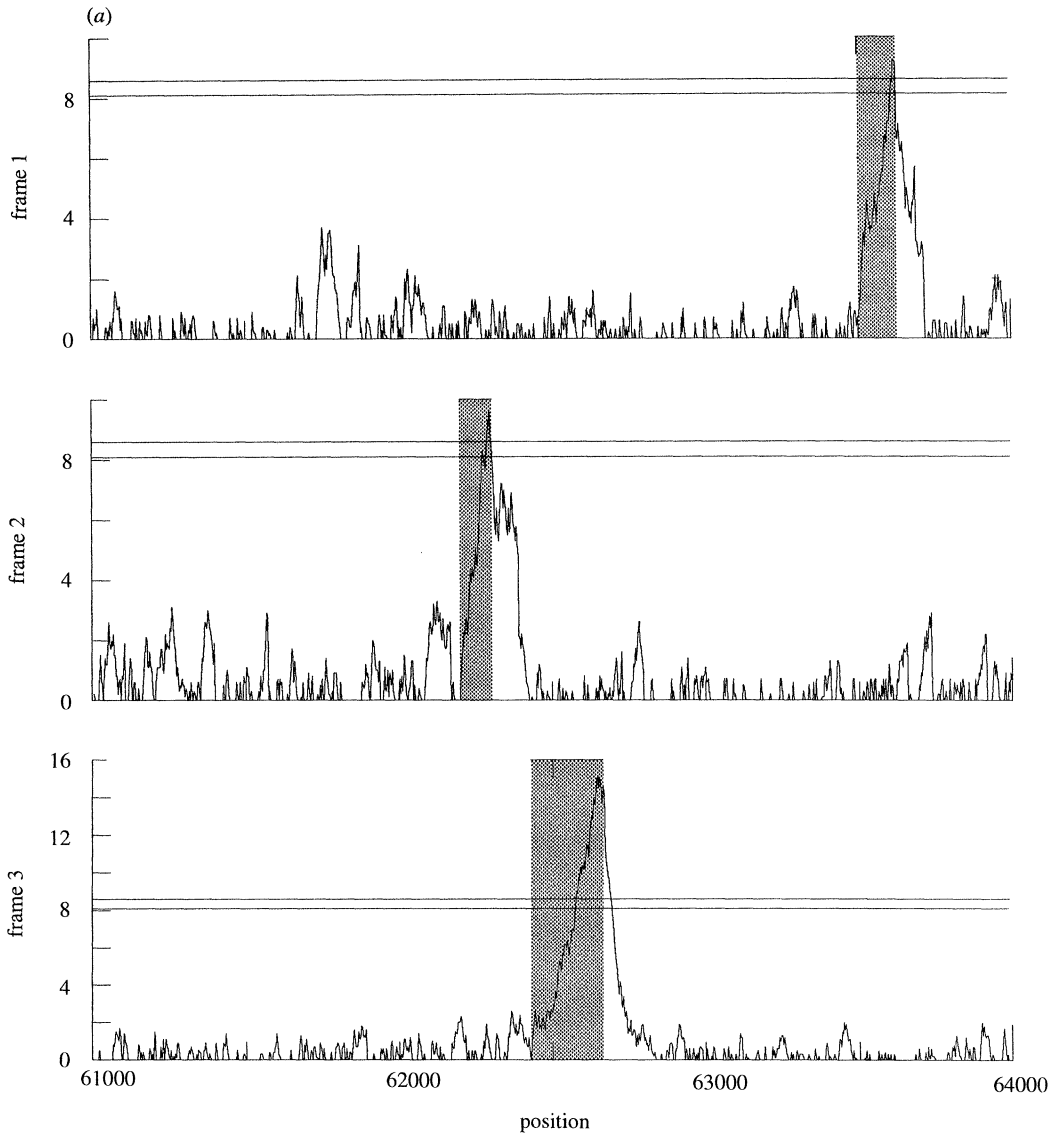

Figure 3. (*a*) An enlargement of figure 2 for the interval of positions 61 000–64 000. (*b*) An enlargement of figure 2 for the interval of positions 54 000–57 000.

objective of the human genome project. This is a formidable task because of the phenomena of split genes in eukaryotes. A natural scoring scheme is the log odds ratio $s_\nu = \ln(q_\nu/p_\nu), \nu = 1, 2, \ldots, 64$, defined for all codons where the target frequencies q_α are determined from a data base of known genes in an appropriate class and p_α are based on all triplet frequencies across all frames in the sequences under study. High-scoring segments flanked with splice signals and possibly other transcriptional control elements will be considered potential exons.

Several recent algorithms have demonstrated some success in predicting coding regions in genomic DNA. Those in most widespread use include the programs: GRAIL (Uberbacher and Mural 1991), GeneModeler (Fields & Soderlund 1990), GeneID (Guigo *et al.* 1992), SORFIND (Hutchinson & Hayden 1992), GeneParser (Snyder & Stormo 1993) and Markov chain methods (e.g. Borodovsky & McIninch 1993), all of which combine the two classical methodologies of 'gene search by signal' and 'gene search by content' that emerged from early attempts to identify contiguous coding regions in prokaryotes reviewed in von Heijne (1987). Search by signal methods rely on sequence motifs such as promoters, start and stop codons and splice site motifs to predict candidate genes, whereas search by content methods exploit features in coding regions such as codon usage (Staden 1990), local compositional complexity (Konopka & Owens 1990), and exon and intron length distributions. The signal and content information typically is combined according to some predefined rules or by connectionist artificial intelligence techniques with various weighting and training schemes. In rule-based procedures (e.g. GeneModeler, GeneID), candidate genes include all predicted regions meeting the criteria of the particular set of rules. The connectionist algorithms judge predictive accuracy according to similarity of the candidate gene to the training set, with predictions often ranked in order of confidence (e.g. confidence levels of 'Excellent', 'Good', or 'Marginal', in GRAIL, levels 1–5 in SORFIND). Related

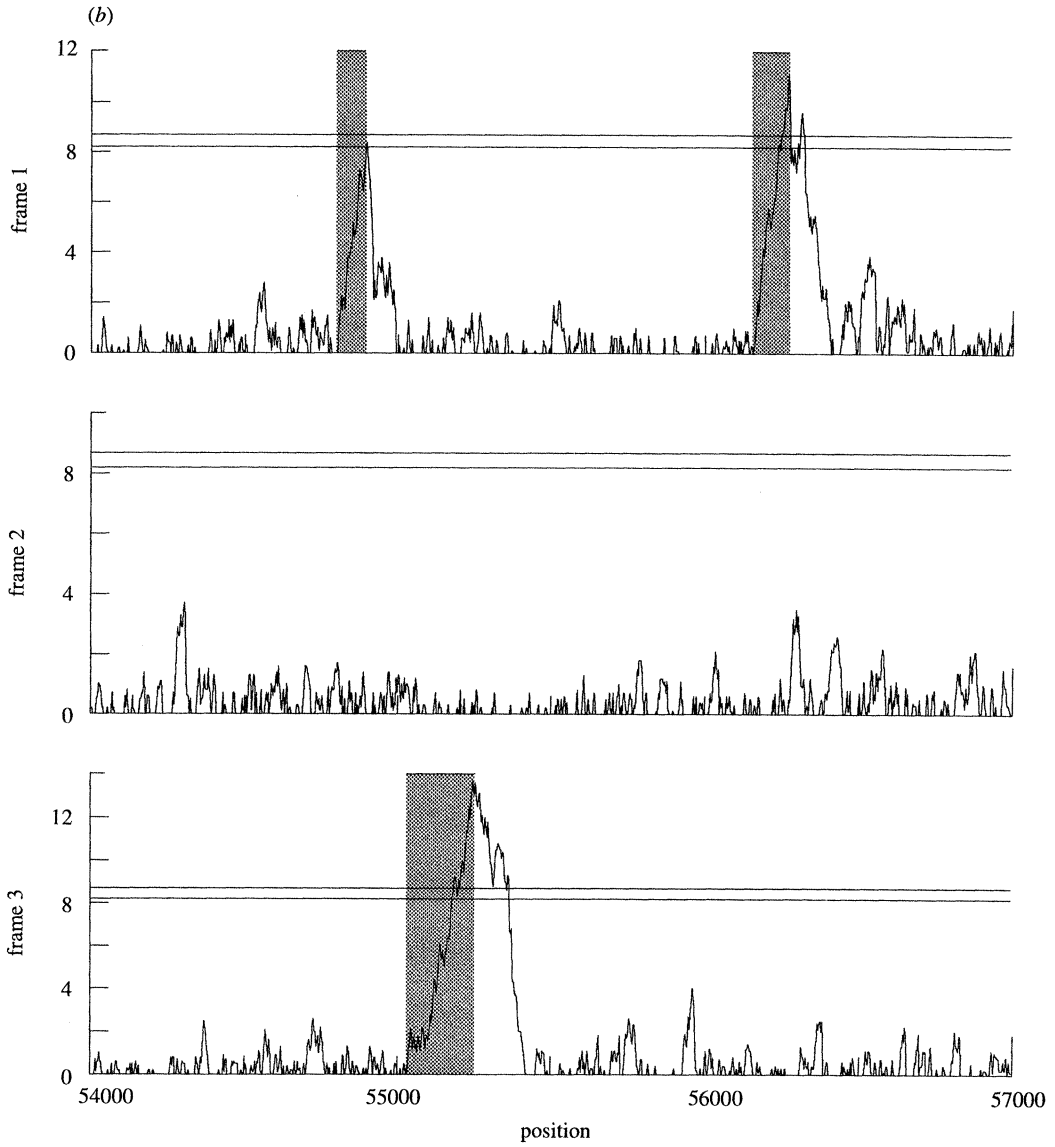

Figure 3. Continued.

artificial neural network training schemes are applied in Brunak *et al.* (1991) and Engelbrecht *et al.* (1992).

For exon prediction, the scoring method offers a distinct advantage over extant methods by providing a means to evaluate the statistical significance of predicted coding regions rather than broad classifications of predictive accuracy. The scoring approach for exon prediction makes use of 'target frequencies' of codons ($q_\alpha, \alpha = 1, \ldots, 64$) drawn from a database of known genes in a particular species or subclass, and triplet frequencies (p_α) across all frames in the sequences at hand. Scores for each codon are given by $\ln(q_\alpha/p_\alpha)$, a likelihood ratio as suggested by equation (2). The algorithm identifies segments of high aggregate score, or high-scoring segments (HSS), with probability of occurrence $\leqslant 0.01$ in any reading frame of the sequence of interest. Relevant statistical formulae for these assessments are given in equations (1) and (2). As the HSS are determined entirely by the data, no segment length restrictions are imposed, except that we exclude HSSs of length $\leqslant 60$ nucleotides (20 codons) for conservative prediction. Stop codons are virtually always excluded from the predicted exons because their large negative scores (owing to low target frequencies) virtually prevent them from meeting the aggregate score threshold for significance.

As in other procedures of this genre, we combine 'content' information (from the scoring approach) with 'signal' information. High-scoring segments are extended in both directions until a start codon (5′) or stop codon (3′) is located in-frame. This extension begins 3–5 codons within each terminus of the predicted segment. If the HSS remains significant at a lower threshold (e.g. 0.15), the segment is considered a strong exon candidate. Otherwise, the nearest three splice sites (AG 5′ of the segment; GU on the 3′ side) are located irrespective of frame, again beginning 3–5 codons within the segment, and the segment score is re-evaluated for significance. With sufficient significance, the extension is proposed as a true exon. The methods also exploit established patterns and motifs of amino acids, the relative abundance of local secondary structures in regulatory regions and relative paucity thereof in coding sequences, and complexity analysis. Another device

will rely on evaluations of translations of DNA sequences in all frames and analysis of hits in comparing against the available protein data bank.

Prediction of exons in β-globulin sequence

We have evaluated the human β-globin region as a test of the scoring approach. The published sequence for this region spans 73.3 kb and contains five genes (ε, Gγ,Aγ,δ,β), each with three exons. The sequence also includes a pseudogene, several Alu repeats, and a Kpn1 (LINE) sequence. Figure 2 shows the exons predicted by the scoring method, which include exact or overlapping portions of 12 of the known exons. One of the pseudogene exons is also detected by the method; the other two pseudogene exons in this sequence contain several stop codons and are correctly discriminated against. An enlargement (figure 3*a*) of the region 54 800–58 900 selects virtually the precise exons (up to at most one codon) of the beta globin gene. Figure 3*b* does the same for the ε-globin gene. Similar fine tuning identify the almost exact exons of the Gγ, Aγ, and δ genes (not shown).

7. MATHEMATICS OF SEQUENCE MATCHING WITH GENERAL SCORES

In DNA and protein sequence matching, let $F(x,y)$ be the score for the letter pair (x,y). For the longest perfect match with few errors, and the longest quality q match (% matching exceeding q), the formulas of Karlin & Ost (1985, 1988), Arratia *et al.* (1986, 1990), Arratia & Waterman (1989) determine the asymptotic distribution of the maximal matching intersequence segment, at least when the underlying probability laws of the sequences are similar enough. The maximal segment score allowing shifts (variable i,j) is

$$M_n = \max_{\substack{0 \leqslant i,j \leqslant n-\Delta \\ \Delta>0}} \left\{ \sum_{l=1}^{\Delta} F(X_{i+l}, Y_{j+l}) \right\}.$$

The two sequences are assumed to be independent: $X_1, \ldots, X_n$ following the distribution law μ_X and $Y_1, \ldots, Y_n$ following the distribution law μ_Y, where μ_X and μ_Y refer to probabilities on the alphabet $\sum_X$ and $\sum_Y$ of the sequences $\{X_i\}$ and $\{Y_j\}$, respectively. We assume that the expected score per letter pair is negative and there is positive probability of attaining some positive pair score, in which case $M_n \to \infty$ corresponds to rare events. The growth asymptotics of M_n is characterized in the following theorem.

Theorem I. (Dembo *et al.* 1994*b*). There exists a finite and positive constant $\gamma^*(\mu_X,\mu_Y)$ depending on μ_X and μ_Y such that $M_n/\ln n \to \gamma^*(\mu_X,\mu_Y)$. The calculation of the constant γ^* involves relative entropy functionals described below.

Let $\sum = \sum_X \times \sum_Y$ be the alphabet of letter pairs, and ν and μ typical probability measures on $\sum$. The relative entropy of ν with respect to μ is denoted by $H(\nu|\mu)$, and for $\sum = \{b_1, \ldots, b_N\}$ is given by the formula:

$$H(\nu|\mu) = \sum_{i-1}^{N} \nu(b_i) \log \frac{\nu(b_i)}{\mu(b_i)}.$$

Let

$$H^*(\nu|\mu_X,\mu_Y) = \max\left\{\frac{1}{2}H(\nu|\mu_X \times \mu_Y), H(\nu_X|\mu_Y), H(\nu_Y|\mu_Y)\right\},$$

where ν_X and ν_Y denote the marginals of ν on $\sum_X$ and $\sum_Y$, respectively. We abbreviate $H^*(\nu)$ for $H^*(\nu|\mu_X,\mu_Y)$. The expected score per letter pair with respect to the probability law ν is denoted by

$$E_\nu[F] = \sum_{(a,b)} F(a,b)\nu(a,b).$$

Define $J(\nu) = E_\nu(F)/H^*(\nu)$. The constant γ^* is

$$\gamma^* = \gamma^*(\mu_X,\mu_Y) = \max_\nu J(\nu).$$

There is, in analogy with the score of the maximal segment score for a single sequence, the special measure

$$\alpha^*(a,b) = \mu_X(a)\mu_Y(b)e^{\theta^* F(a,b)}, \tag{7}$$

where θ^* is the unique positive value satisfying the equation $E_{\mu_X\times\mu_Y}[e^{\theta^* F}] = 1$. It can be proved that always

$$\frac{1}{\theta^*} \leqslant \gamma^*(\mu_X,\mu_Y) \leqslant \frac{2}{\theta^*}.$$

Under the condition

$$\frac{1}{2}H(\alpha^*|\mu_x \times \mu_y) > \max\{H(\alpha_x^*|\mu_x), H(\alpha_y^*|\mu_y)\}, \tag{8}$$

the following holds:

Theorem II. (Dembo *et al.* 1994*c*). Assuming (8), then $\gamma^* = 2/\theta^*$, and

$$\Pr\left\{M_n \leqslant \frac{\log n^2}{\theta^*} + x\right\} \to \exp\{-K^* e^{-\theta^* x}\} \qquad n \to \infty, \tag{9}$$

where K^* is the constant associated with a one sequence (of length n^2) evaluation of the maximal scoring segment (equation (1)). The length Δ_n of the maximal score segment is

$$\approx \frac{1}{\theta^* E_{\alpha^*}(F)}(\log n^2).$$

The empirical match distribution on Δ_n is that of the probability measure α^* defined in (7).

For symmetric scores ($F(x,y) = F(y,x)$) and $\mu_X = \mu_Y$, the condition (8) holds provided $F(x,y)$ is not of the form $s(x) + s(y)$. The condition (8) seems to be of wide scope. In all cases, the right-hand side of (9) provides an upper bound.

I acknowledge extensive helpful discussions and comments on the manuscript from Dr B. E. Blaisdell, Dr V. Brendel, Dr L. R. Cardon and Dr A. Dembo. Research reported in this paper was supported in part by NIH Grants 5R01GM10452-30 and 8R01HG00335-06 and NSF Grant DMS-9106974.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. 1990 Basic local alignment search tool. *J. molec. Biol.* **215**, 403–410.

Arratia, R., Gordon, L. & Waterman, M. 1986 An extreme value theory for sequence matching. *Ann Stat.* **14**, 971–994.

Arratia, R., Gordon, L. & Waterman, M.S. 1990 An Erdös Rényi law in distribution for coin tossing and sequence matching. *Ann. Stat.* **18**, 539–570.

Arratia, R. & Waterman, M.S. 1989 The Erdös Rényi strong law for pattern matching with a given proposition of mismatches. *Ann. Prob.* **17**, 1152–1169.

Arratia, R., Morris, P. & Waterman, M.S. 1988 Stochastic scrabble: large deviations for sequences with scores. *J. appl. Prob.* **25**, 106–119.

Arratia, R. & Waterman, M.S. 1985 An Erdös Rényi law with shifts. *Adv. Math.* **55**, 13–23.

Baer, R., Bankier, A.T., Biggin, M.D. *et al* 1984 DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature, Lond.* **310**, 207–211.

Bairoch, A. & Boechmann, B. 1991 The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **20**, 2019–2022.

Borodovsky, M. & McIninch, J. 1993 Recognition of genes in DNA sequence with ambiguities. *Biosystems* **30**, 161–171.

Brendel, V., Bucher, P., Nourbakhsh, I.R., Blaisdell, B.E. & Karlin, S. 1992 Methods and algorithms for statistical analysis of protein sequences. *Proc. natn. Acad. Sci. U.S.A.* **89**, 2002–2006.

Brendel, V., Ladunga, I. & Karlin, S. 1994 Score assignments for amino acid substitutions in protein comparisons. *J. molec. Biol.* (Submitted.)

Brunak, S., Engelbrecht, J. & Knudsen, S. 1991 Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. molec. Biol.* **220**, 49–64.

Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. 1978 A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (ed. M. O. Dayhoff), vol. 5, suppl. 3, pp. 345–352. Washington: Natl. Biomed. Res. Found.

Dembo, A. & Karlin, S. 1991 Strong limit theorems of empirical functionals for large exceedences of partial sums of i.i.d. variables. *Ann. Prob.* **19**, 1737–1755.

Dembo, A. & Karlin, S. 1993 Central limit theorems of partial sums for large segmental values. *Stoch. Process. Appl.* **45**, 259–271.

Dembo, A., Karlin, S. & Zeitouni, O. 1994*a* Large exceedances of vector scores. *Ann. Appl. Prob.* (In the press.)

Dembo, A., Karlin, S. & Zeitouni, O. 1994*b* Critical phenomena for sequence matching with scoring. *Ann. Prob.* (In the press.)

Dembo, A., Karlin, S. & Zeitouni, O. 1994*c* Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.* (In the press.)

Doolittle, R. (ed.) 1990 Molecular evolution: computer analysis of protein and nucleic acid sequences. *Meth. Enzymol.* **183**.

Engelbrecht, J., Knudsen, S. & Brunak, S. 1992 G+C-rich tract in 5′ end of human introns. *J. molec. Biol.* **227**, 108–113.

Feng, D.F., Johnson, M.S. & Doolittle, R.F. 1985 Aligning amino acid sequences: comparison of commonly used methods. *J. molec. Evol.* **21**, 112–125.

Fields, C.A. & Soderlund, C.A. 1990 gm: a practical took for automating DNA sequence analysis. *CABIOS* **6**, 263–270.

Gonnett, G.H., Cohen, M.A. & Benner, S.A. 1992 Exhaustive matching of the entire protein sequence database. *Science, Wash.* **256**, 1443–1445.

Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. & Claverie, J.M. 1993 Ancient conserved regions in new gene sequences and the protein databases. *Science, Wash.* **259**, 1711–1716.

Gribskov, M.R. & Devereux, J. (eds) 1992 *Sequence analysis primer*. New York: Freeman.

Guigo, R., Knudsen, S., Drake, N. & Smith, T. 1992 Prediction of gene structure. *J. molec. Biol.* **226**, 141–157.

Henikoff, S. & Henikoff, J.G. 1992 Amino acid substitution matrices from protein blocks. *Proc. natn. Acad. Sci. U.S.A.* **89**, 10915–10919.

Honess, R.W. 1984 Herpes simplex and "the Herpes complex": diverse observations and a unifying hypothesis. *J. gen. Virol.* **65**, 2077–2107.

Hutchinson, G.B. & Hayden, M.R. 1992 The prediction of exons through an analysis of spliceable open reading frames. *Nucl. Acids Res.* **20**, 3453–3462.

Inman, R.B. 1966 A denaturation map of the λ phage DNA molecule determined by electron microscopy. *J. molec. Biol.* **18**, 464–476.

Jones, D.T., Taylor, W.R. & Thornton, J.M. 1992 The rapid generation of mutation data matrices from protein sequences. *Comput. appl. Biosci.* **8**, 275–282.

Karlin, S. 1986 Significant potential secondary structures in the Epstein-Barr virus genome. *Proc. natn. Acad. Sci. U.S.A.* **83**, 6915–6919.

Karlin, S. & Altschul, S.F. 1990 Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. natn. Acad. Sci. U.S.A.* **87**, 2264–2268.

Karlin, S. & Altschul, S.F. 1993 Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. natn. Acad. Sci. U.S.A.* **90**, 5873–5877.

Karlin, S., Blaisdell, B.E. & Schachtel, G.A. 1990 Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus: data and hypotheses. *J. Virol.* **64**, 4264–4273.

Karlin, S. & Brendel, V. 1992 Chance and statistical significance in protein and DNA sequence analysis. *Science, Wash.* **257**, 39–49.

Karlin, S., Brendel, V. & Bucher, P. 1992*b* Significant similarity and dissimilarity in homologous proteins. *Molec. Biol. Evol.* **9**, 152–167.

Karlin, S., Burge, C. & Campbell, A.M. 1992*a* Over- and under-representation of short oligonucleotides in DNA in DNA sequences. *Proc. natn. Acad. Sci. U.S.A.* **89**, 1358–1362.

Karlin, S. & Cardon, L. 1994 Computational DNA sequence analysis. *A. Rev. Microbiol.* **48**. (In the press.)

Karlin, S. & Dembo, A. 1992 Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv. appl. Prob.* **24**, 113–140.

Karlin, S., Dembo, A. & Kawabata, T. 1990 Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.* **18**, 571–581.

Karlin, S., Mocarski, E.M. & Schachtel, G.A. 1994 Molecular evolution of herpesviruses: genomic and protein sequence comparisons. *J. Virol.* **68**, 1886–1902.

Karlin, S. & Ost, F. 1985 Maximal segmental match length among random sequences from a finite alphabet. In *Proc. Berkeley Conf. in honor of J. Neyman and J. Kiefer* (ed. L. M. LeCam & R. A. Olshen). Belmont, California: Wadsworth.

Karlin, S. & Ost, F. 1988 Maximum length of common words among random letter sequences. *Ann. Prob.* **16**, 535–563.

Konopka, A.K. & Owens, J. 1990 Complexity charts can be used to map functional domains in DNA. *Gen. Anal. Tech. Appl.* **7**, 35–38.

Snyder, E.E. & Stormo, G.D. 1993 Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucl. Acids Res.* **21**, 607–613.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Staden, R. 1990 Finding protein coding regions in genomic sequences. *Meth. Enzymol.* **183**, 163–180.

Uberbacher, E.C. & Mural, R.J. 1991 Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. natn. Acad. Sci. U.S.A.* **88**, 11261–11265.

von Heijne, G. 1987 *Sequence analysis in molecular biology: treasure trove or trivial pursuit?* New York: Academic Press.

Waterman, M.S. (ed) 1989 *Mathematical methods for DNA sequence*. Boca Raton, Florida: C.R.C. Press.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES