*J. Appl. Prob.* **44,** 514–527 (2007)
*Printed in England*


# MARKOV ADDITIVE PROCESSES AND REPEATS IN SEQUENCES

JOHN L. SPOUGE,* *National Library of Medicine*

**Abstract**

Computer analysis of biological sequences often detects deviations from a random model. In the usual model, sequence letters are chosen independently, according to some fixed distribution over the relevant alphabet. Real biological sequences often contain simple repeats, however, which can be broadly characterized as multiple contiguous copies (usually inexact) of a specific word. This paper quantifies inexact simple repeats as local sums in a Markov additive process (MAP). The maximum of the local sums has an asymptotic distribution with two parameters ($\lambda$ and $k$), which are given by general MAP formulas. The general MAP formulas are usually computationally intractable, but an essential simplification in the case of repeats permits $\lambda$ and $k$ to be computed from matrices whose dimension equals the size of the relevant alphabet. The simplification applies to some MAPs where the summand distributions do not depend on consecutive pairs of Markov states as usual, but on pairs with a fixed time-lag larger than one.

*Keywords:* Markov additive process; biological sequence analysis; repeats

2000 Mathematics Subject Classification: Primary 60K15; 60K20; 60G51; 60G70

## 1. Introduction and statement of results

Computer analysis of sequences is a major preoccupation in modern molecular biology. Typically and often implicitly, analysis programs detect deviations from a Markov model of sequence letters, usually a model of order zero, where letters are mutually independent and chosen from a fixed distribution. Real biological sequences often contain repeats, however, which can be broadly characterized as multiple copies (usually inexact) of a particular word. Many repeats flagrantly violate a Markov model of order zero; so analysis programs often flag them for fruitless human scrutiny.

Because repeats can confound sequence analysis programs, ancillary programs have been developed for detecting and masking repeats before performing sequence analysis, e.g. the masking programs SEG [21]–[23], XNU [18], [20], sputnik, DUST, and RepeatMasker, to name a few. Although common sequence repeats in DNA fall into about five classes [19], this paper focuses on simple repeats, which consist of contiguous inexact copies of a specific word. Though a narrow class, simple repeats still pose significant problems in sequence analysis. Thus, some specialized masking programs (e.g. sputnik and DUST) target simple repeats exclusively, while general masking programs usually offer a similar specialized option.

Although simple repeats often impede sequence analysis, it should be noted that sometimes they are of intrinsic biological interest and can constitute an object of study in their own right.

Rigorous statistical results are available for exact repeats (e.g. … sgt/sgt/sgt/sgt…) [2], [12] or exact words [4], [5], [10], [11], [15]. Unfortunately, inexact repeats (e.g. … sgt/sgp/sgt/sgt…)

Received 25 May 2006; revision received 19 February 2007.
* Postal address: National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA. Email address: spouge@ncbi.nlm.nih.gov

are more important biologically [6]. Because few statistical results pertain to inexact repeats, programs usually mask inexact repeats heuristically, without the benefit of rigorous statistical criteria. Accordingly, the next paragraph reformulates the definition of an inexact simple repeat with a view to rigorous results. (Because the rest of the paper is pertinent only to simple repeats, we drop the qualifier 'simple'.)

Let $L := (L_1, L_2, \ldots) \in \mathfrak{L}^\infty$ denote a semi-infinite string (where ':=' denotes a definition), i.e. a sequence on a finite alphabet $\mathfrak{L}$ of size $\#\mathfrak{L}$. Fix $w$ for the rest of the paper, and let $L(i, i+w] := (L_{i+1}, \ldots, L_{i+w})$ be a '$w$-word' within $L$. In addition, let $s: \mathfrak{L} \times \mathfrak{L} \mapsto \mathbb{R}$ be a 'similarity score matrix' on the alphabet $\mathfrak{L}$. Intuitively, $s(a, b)$ is large if the letters $a$ and $b$ symbolize similar objects. (In many applications, $s(a, b) = s(b, a)$ for all $a, b \in \mathfrak{L}$, but asymmetric score matrices are permitted here.) Define the '$w$-repeat score' by $Y_k := s(L_k, L_{k+w})$ with $Y_0 := 0$, the global sums by $S_k := \sum_{i=1}^{k} Y_i$, and the local maxima by $\hat{M}_k := \max_{0 \le i \le j \le k} \sum_{m=i}^{j} Y_m$.

Intuitively, the local maximum $\hat{M}_k$ is large if $L(0, k]$ contains a substring that decomposes into contiguous $w$-words, with each word and its successor being 'similar' according to the matrix $s$. As an example, consider exact repeats, which correspond to a similarity score of $s_\infty(a, b) = 1$ if $a = b$ and $-\infty$ otherwise. Then, $\hat{M}_k + w$ is the length of the longest substring in $L(0, k]$ that decomposes into contiguous exact copies of a $w$-word, possibly terminating with an incomplete copy (e.g. … sgt/sgt/sgt/sg…).

The condition $s(a, b) \in \mathbb{R}$ given implicitly above prohibits $s(a, b) = -\infty$, effectively excluding exact repeats from the mathematical discourse. The exclusion is irrelevant in practice, however, because the theory applies to similarity matrices $s(a, b)$ arbitrarily close to $s_\infty(a, b)$. The mathematical axioms could in fact include exact repeats by taking $s: \mathfrak{L} \times \mathfrak{L} \mapsto \mathbb{R} \cup \{-\infty\}$, but only with some complications (notably, non-uniqueness of the right eigenvector $\boldsymbol{v}$, defined below). For simplicity, therefore, our axioms exclude exact repeats, but with no practical impact.

Some practical consequences of our main results, Theorem 1.1 and Theorem 1.2 below, follow their statement. Theorem 1.1 and Theorem 1.2 make the following probabilistic assumptions, required by the theory of the local maximum $\hat{M}_k$ in a Markov additive process (MAP). (Throughout the paper, the reader may refer to [3] as a general reference for MAPs, Markov chains (MCs), or renewal theory; and to [7], [9] as specific references for the local maximum process in a general MAP.) Let the letters of $L$ be chosen independently from the distribution $\{q_l : l \in \mathfrak{L}\}$. To avoid trivialities, assume $q_l > 0$ for all $l \in \mathfrak{L}$. Assume also that

$$\mu := \sum_{a,b \in L} q_a q_b s(a, b) < 0,$$

and that there is a cycle of letters $a_0, a_1, \ldots, a_m = a_0$ such that $\min_{k=1,\ldots,m} \sum_{i=1}^{k} s(a_{i-1}, a_i) > 0$.

Several definitions are now required. Unless specified otherwise, all vectors are column vectors of dimension $\#\mathfrak{L}$ and all matrices are of dimension $(\#\mathfrak{L}) \times (\#\mathfrak{L})$. Let $a$ and $b$ symbolize the row and column indices of matrices and vectors. For later purposes, think of the letters of $L$ as being produced by a degenerate MC of order one, whose transition matrix $\boldsymbol{P} := \|q_b\|$ has every row equal to its stationary distribution $\boldsymbol{q}^\top := [q_b]$. Consider $\boldsymbol{P}_\theta = \|q_b \exp\{\theta s(a, b)\}\|$, the moment-generating matrix corresponding to the (trivial) random variate $Z_{a,b} = s(a, b)$. Because the degenerate MC is irreducible and aperiodic, the Perron–Frobenius theorem shows that $\boldsymbol{P}_\theta$ has a strictly dominant eigenvalue $\rho(\theta) > 0$ (i.e. $\rho(\theta)$ is the unique eigenvalue of

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

greatest absolute value). The dominant eigenvalue corresponds to strictly positive left and right eigenvectors, $\boldsymbol{u}_\theta$ and $\boldsymbol{v}_\theta$, normalized for convenience so that $\boldsymbol{u}_\theta^\top \boldsymbol{v}_\theta = 1$. On its own (remarks elsewhere notwithstanding [13]), the convexity of the elements of $\boldsymbol{P}_\theta$ as functions of $\theta$ shows that $\rho(\theta) = \sup\{\boldsymbol{u}^\top \boldsymbol{P}_\theta \boldsymbol{v} \colon \boldsymbol{u}^\top \boldsymbol{v} = 1;\ \boldsymbol{u}, \boldsymbol{v} \geq \boldsymbol{0}\}$ is a convex function, and because $\boldsymbol{P} = \boldsymbol{P}_0$ is stochastic, $\rho(0) = 1$.

For $\delta \neq 0$, we have

$$\boldsymbol{u}_\theta(\boldsymbol{P}_{\theta+\delta} - \boldsymbol{P}_\theta)\boldsymbol{v}_\theta \leq \rho(\theta+\delta) - \rho(\theta) \leq \boldsymbol{u}_{\theta+\delta}(\boldsymbol{P}_{\theta+\delta} - \boldsymbol{P}_\theta)\boldsymbol{v}_{\theta+\delta}, \tag{1.1}$$

so in the limit $\delta \to 0$, the continuity of eigenvectors [14, p. 396] yields $\rho'(\theta) = \boldsymbol{u}_\theta(\mathrm{d}\boldsymbol{P}_\theta/\mathrm{d}\theta)\boldsymbol{v}_\theta$. The convexity of $\rho(\theta)$ and the fact that $\rho'(0) = \boldsymbol{q}^\top \|q_b s(a,b)\|\boldsymbol{1} = \mu < 0$ show the existence of

(1) a unique $\theta = \lambda > 0$ such that $\rho(\lambda) = 1$, and

(2) a minimum $0 < \rho := \min_{0<\theta<\lambda} \rho(\theta) < 1$.

As part of the Perron–Frobenius theorem, the eigenvalue $\rho(\lambda) = 1$ corresponds to a strictly positive right eigenvector $\boldsymbol{v} = [v_a] := \boldsymbol{v}_\lambda$. (The eigenvector $\boldsymbol{v}$ is unique within a multiplicative factor, and any nonzero multiple of $\boldsymbol{v}$ serves present purposes. The uniqueness can be lost, however, if the possibility $s(a,b) = -\infty$ were permitted.) Let $\boldsymbol{V} := \mathrm{diag}(\boldsymbol{v}) := \|v_a \delta_{a;b}\|$, where the Kronecker delta $\delta_{a;b} = 1$ for $a = b$ and 0 otherwise. The 'associated transition matrix'

$$\boldsymbol{P}^* := \|p^*_{a,b}\| := \|v_a^{-1} v_b q_b \exp\{\lambda s(a,b)\}\| = \boldsymbol{V}^{-1}\boldsymbol{P}_\lambda \boldsymbol{V} \tag{1.2}$$

is stochastic, i.e. $p^*_{a;b} \geq 0$ and $\boldsymbol{P}^*\boldsymbol{1} = \boldsymbol{V}^{-1}\boldsymbol{P}_\lambda \boldsymbol{V}\boldsymbol{1} = \boldsymbol{V}^{-1}\boldsymbol{P}_\lambda \boldsymbol{v} = \boldsymbol{V}^{-1}\boldsymbol{v} = \boldsymbol{1}$. Thus, $\boldsymbol{P}^*$ defines transitions for some 'associated MC' on $\mathfrak{L}$. The associated MC has stationary distribution $(\boldsymbol{\pi}^*)^\top = (\boldsymbol{u}_\lambda)^\top \boldsymbol{V}$, and its mean score per step is $\mu_* := \sum_{(a,b)} \pi^*_a p^*_{a;b} s(a,b)$. The convexity of $\rho(\theta)$ shows that $\mu_* = \rho'(\lambda) > 0$ at $\rho(\lambda) = 1$.

Consider a sequence of independent, identically distributed random variates $\{\tilde{X}_k \colon k = 1, 2, \dots\}$ and their global sums $\tilde{S}_k := \sum_{i=1}^k \tilde{X}_i$, under three possible marginal distributions: $\hat{\mathrm{P}}\{\tilde{X} = s(a,b)\} = q_a q_b$ (with expectation $\hat{\mathrm{E}}$); $\tilde{\mathrm{P}}_\alpha\{\tilde{X} = s(a,b)\} = q_a q_b v_b / r$ (with expectation $\tilde{\mathrm{E}}_\alpha$); and $\tilde{\mathrm{P}}_\beta\{\tilde{X} = s(a,b)\} = \pi^*_a p^*_{a;b} v_b^{-1}/r_*$ (with expectation $\tilde{\mathrm{E}}_\beta$). The distribution $\hat{\mathrm{P}}$ appears in Theorem 1.1 and $\tilde{\mathrm{P}}_\alpha$ and $\tilde{\mathrm{P}}_\beta$ (which are normalized to probability distributions by the constants $r$ and $r_*$) appear in Theorem 1.2.

The argument following (1.1) demonstrates the existence of a unique $\hat{\lambda} > 0$ satisfying $\hat{\mathrm{E}} \exp(\hat{\lambda}\tilde{X}) = \sum_{(a,b)} q_a q_b \exp\{\hat{\lambda} s(a,b)\} = 1$, called 'the $\lambda$ for gapless alignment' in bioinformatics [8].

**Theorem 1.1.** *If $y, m \to \infty$ so that $\mathrm{e}^{-\lambda y} m$ tends to a finite, nonzero constant, then*

$$\lim_{y\to\infty} \mathrm{P}\{\hat{M}_m \geq y\} = (k_w \mathrm{e}^{-\lambda y})m,$$

*where the pre-factor $k_w$ is defined in Section 3 along with an efficient simulation for estimating it. (Note: $\lambda$ does not depend on $w$.) In addition, $0 < \lambda \leq \hat{\lambda}$, with equality if and only if $\sum_{(b)} q_b \exp\{\lambda s(a,b)\} = 1$ for all $a \in \mathfrak{L}$ or $\sum_{(a)} q_a \exp\{\lambda s(a,b)\} = 1$ for all $b \in \mathfrak{L}$.*

A plot illustrating the equation $\lim_{y\to\infty} \mathrm{P}\{\hat{M}_m \geq y\} = (k_w \mathrm{e}^{-\lambda y})m$ in Theorem 1.1 is given in Figure 1.

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

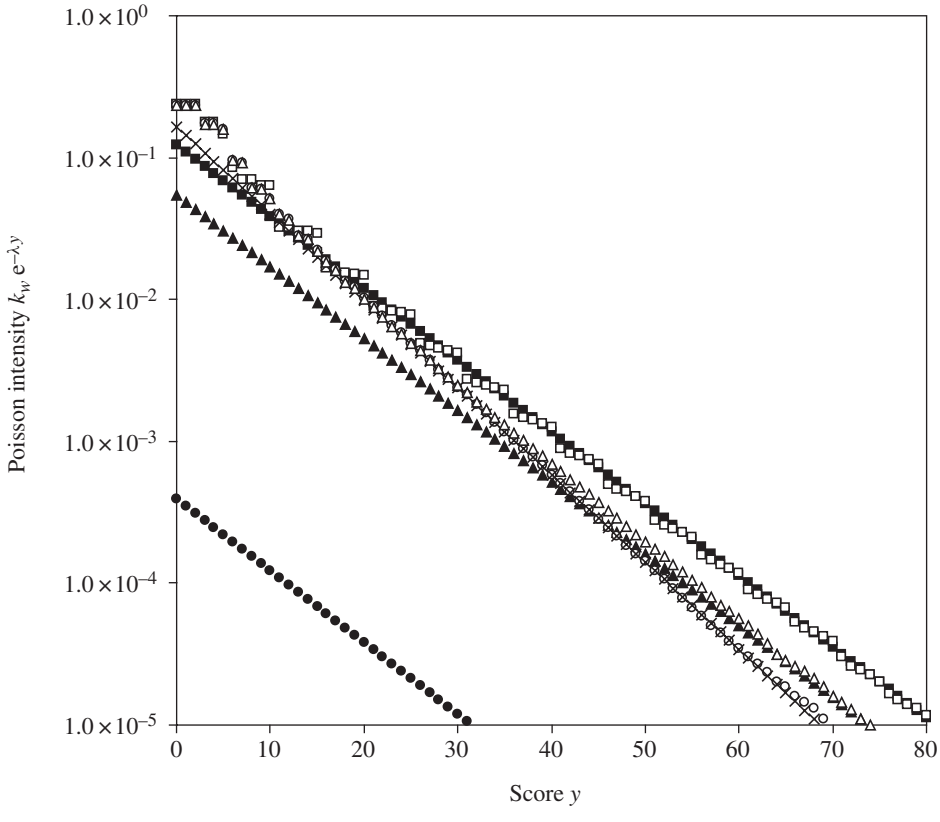


FIGURE 1: A comparison between the density (open symbols) and the asymptotic Poisson intensity $k_w e^{-\lambda y}$ (black symbols) for repeat word-lengths $w = 1$ (squares); $w = 8$ (triangles); and $w = 64$ (circles). For comparison, the asymptotic Poisson intensity, $\hat{k}e^{-\hat{\lambda} y}$, for ungapped alignment of two independent sequences is also plotted (crosses).

Figure 1 illustrates simulation results for the binary alphabet $\mathfrak{L} = \{0, 1\}$, a uniform distribution $q_0 = q_1 = 0.5$, and a similarity score matrix with elements $s(0, 0) = 5$, $s(1, 1) = 2$, and $s(0, 1) = s(1, 0) = -8$. The simulation generated a single sequence of length $m = 10^7$, for which the Ruzzo–Tompa algorithm defined maximal segments of $w$-repeats [17]. The number of maximal segments $(i, j]$ with $S_{(i,j]} := \sum_{k=i+1}^{j} Y_k \geq y$ was divided by $m = 10^7$ to yield the density of maximal segments with $S_{(i,j]} \geq y$. For $w = 1$ (squares), the density agrees with repeat asymptotics above the score $y = 10$. For $w = 8$ (triangles), the density agrees with alignment asymptotics above the score $y = 10$ before crossing over to repeat asymptotics above the score $y = 50$. For $w = 64$ (circles), the density agrees with alignment asymptotics throughout the plot area. The crossover behavior shown for $w = 8$ can be explained as follows. Until the score $y$ exceeds the sum of $w = 8$ similarity scores $s(a, b)$, each letter is matched with only one other letter (as in the alignment of two independent sequences), not with two other letters (as in the middle of long repeats).

**Remark 1.1.** In matrix notation, the final alternatives are $\boldsymbol{P}_\lambda \mathbf{1} = \mathbf{1}$ and $\boldsymbol{q}^\top \boldsymbol{P}_\lambda = \boldsymbol{q}^\top$.

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

Define $\eta\mathbb{Z} := \{\cdots -\eta, 0, \eta, \dots\}$ for $\eta > 0$ and $0\mathbb{Z} := \mathbb{R}$. Then, the set of scores $\mathbb{S} := \{s(a,b); a, b \in \mathfrak{L}\}$ has lattice span $\delta := \sup\{\eta \geq 0 \colon \mathbb{S} \subseteq \eta\mathbb{Z}\}$. Define $\theta_{-\delta} := \delta^{-1}(1 - \mathrm{e}^{-\delta\lambda})$ if $\delta > 0$ and $\theta_0 := \lambda$ if $\delta = 0$, along with $r := \sum_{(b)} q_b v_b$ and $r_* := \sum_{(a,b)} \pi_a^* p_{a;b}^* v_b^{-1} = \sum_{(b)} \pi_b^* v_b^{-1}$. The product $rr_*$, which is independent of the normalization of $\boldsymbol{v}$, satisfies

$$rr_* = (\boldsymbol{q}^\top \boldsymbol{v})(\boldsymbol{\pi}^*)^\top \boldsymbol{V}^{-1}\mathbf{1} = (\boldsymbol{q}^\top \boldsymbol{V}1)(\boldsymbol{\pi}^*)^\top \boldsymbol{V}^{-1}\mathbf{1} = \boldsymbol{q}^\top \boldsymbol{V}[\mathbf{1}(\boldsymbol{\pi}^*)^\top]\boldsymbol{V}^{-1}\mathbf{1} \leq \boldsymbol{q}^\top \boldsymbol{V}\boldsymbol{V}^{-1}\mathbf{1} = 1, \tag{1.3}$$

because the Perron–Frobenius theorem shows that the stochastic matrix $\mathbf{1}(\boldsymbol{\pi}^*)^\top$ has dominant eigenvalue 1. The equality $rr_* = 1$ holds if and only if either $\boldsymbol{V}^{-1}\mathbf{1}$ is proportional to the right eigenvector $\mathbf{1}$ of $\mathbf{1}(\boldsymbol{\pi}^*)^\top$ or $\boldsymbol{q}^\top \boldsymbol{V}$ is proportional to the left eigenvector $(\boldsymbol{\pi}^*)^\top = \boldsymbol{u}_\lambda^\top \boldsymbol{V}$ of $\mathbf{1}(\boldsymbol{\pi}^*)^\top$. Thus, $\boldsymbol{P}_\lambda \mathbf{1} = \mathbf{1}$ (i.e. $\boldsymbol{v} = v\mathbf{1}$ for some $v \neq 0$) or $\boldsymbol{q}^\top \boldsymbol{P}_\lambda = \boldsymbol{q}^\top$, so Theorem 1.1 also shows that $rr_* = 1$ if and only if $\lambda = \hat{\lambda}$.

Let $\tilde{\rho}_\alpha(\theta) := \tilde{\mathrm{E}}_\alpha \exp(\theta \tilde{X}) = r^{-1}\boldsymbol{q}^\top \boldsymbol{P}_\theta \boldsymbol{v}$. Then, $\tilde{\rho}_\alpha(\lambda) = r^{-1}\boldsymbol{q}^\top \boldsymbol{v} = 1 = \tilde{\rho}_\alpha(0)$, so the argument following (1.1) demonstrates that $0 < \tilde{\rho}_\alpha(\underline{\theta}_\alpha) := \min_{0<\theta<\lambda} \tilde{\rho}_\alpha(\theta) < 1$, the uniqueness of the minimum defining $\underline{\theta}_\alpha$. (Because $\tilde{\rho}_\alpha(\theta) \leq r^{-1}\boldsymbol{q}^\top \boldsymbol{v}\rho(\theta) = \rho(\theta)$, the use of $\tilde{\rho}_\alpha$ instead of $\rho$ strengthens some of the inequalities given below.)

Given the sums $\{\tilde{S}_k\}$, define their first (weak) descending ladder epoch (DLE) by $\tilde{\alpha} := \min\{m > 0 \colon \tilde{S}_m \leq 0\}$ and their first strict ascending ladder epoch (SALE) by $\tilde{\beta} := \min\{m > 0 \colon \tilde{S}_m > 0\}$. Let the sequence $\{\tilde{s}_{\alpha,w} \colon w = 0, 1, \dots\}$ satisfy the recursion

$$\tilde{s}_{\alpha,w} := \tilde{\mathrm{E}}_\alpha[1 - \exp(\lambda \tilde{S}_{\tilde{\alpha}});\ \tilde{\alpha} = w] + \sum_{m=0}^{w-1} \tilde{s}_{\alpha,m} \tilde{\mathrm{P}}_\alpha\{\tilde{\alpha} = w - m\}, \tag{1.4}$$

with $\tilde{s}_{\alpha,0} := 0$. The renewal theorem yields $\lim_{w\to\infty} \tilde{s}_{\alpha,w} = \tilde{s}_\alpha := \tilde{\mathrm{E}}_\alpha[1 - \exp(\lambda \tilde{S}_{\tilde{\alpha}})]/\tilde{\mathrm{E}}_\alpha \tilde{\alpha}$.

Similarly, $\tilde{\rho}_\beta(\theta) := \tilde{\mathrm{E}}_\beta \exp(-\theta \tilde{X}) = r_*^{-1}(\boldsymbol{\pi}^*)^\top \boldsymbol{V}^{-1} \boldsymbol{P}_{\lambda-\theta}\mathbf{1}$ satisfies $\tilde{\rho}_\beta(\lambda) = 1 = \tilde{\rho}_\beta(0)$, so $0 < \tilde{\rho}_\beta(\underline{\theta}_\beta) := \min_{0<\theta<\lambda} \tilde{\rho}_\beta(\theta) < 1$. (Note that $\tilde{\rho}_\beta(\theta) = r_*^{-1}(\boldsymbol{\pi}^*)^\top \boldsymbol{V}^{-1} \boldsymbol{P}_{\lambda-\theta}\mathbf{1} \leq \rho(\lambda - \theta)$, again strengthening some of the inequalities given below.) Let the sequence $\{\tilde{s}_{\beta,w} \colon w = 0, 1, \dots\}$ satisfy the recursion

$$\tilde{s}_{\beta,w} := \tilde{\mathrm{E}}_\beta[1 - \exp(-\lambda \tilde{S}_{\tilde{\beta}});\ \tilde{\beta} = w] + \sum_{m=0}^{w-1} \tilde{s}_{\beta,m} \tilde{\mathrm{P}}_\beta\{\tilde{\beta} = w - m\}, \tag{1.5}$$

with $\tilde{s}_{\beta,0} := 0$, so as above, $\lim_{w\to\infty} \tilde{s}_{\beta,w} = \tilde{s}_\beta := \tilde{\mathrm{E}}_\beta[1 - \exp(-\lambda \tilde{S}_{\tilde{\beta}})]/\tilde{\mathrm{E}}_\beta \tilde{\beta}$.

Define $\tilde{k}_w := \tilde{s}_{\alpha,w}\tilde{s}_{\beta,w}/(\theta_{-\delta}\mu_*)$, and note that $\tilde{k} := \tilde{s}_\alpha \tilde{s}_\beta/(\theta_{-\delta}\mu_*) = \lim_{w\to\infty} \tilde{k}_w$. Theorem 1.2 below uses the error bounds $\varepsilon_{\alpha,w} := 2\{\rho_\alpha(\underline{\theta}_\alpha)\}^w \rho(\underline{\theta}_\alpha)/\{1 - \rho(\underline{\theta}_\alpha)\}$ and $\varepsilon_{\beta,w} := 2\{\rho_\beta(\underline{\theta}_\beta)\}^w \rho(\lambda - \underline{\theta}_\beta)/\{1 - \rho(\lambda - \underline{\theta}_\beta)\}$, which vanish geometrically with $w$.

**Theorem 1.2.** *Under the conditions of Theorem 1.1, if* $0 < \varepsilon_{\alpha,w} < \tilde{s}_{\alpha,w}$ *and* $0 < \varepsilon_{\beta,w} < \tilde{s}_{\beta,w}$,

$$1 \leq \frac{(rr_*)^w \tilde{k}_w}{k_w} \leq (1 - \varepsilon_{\alpha,w}\tilde{s}_{\alpha,w}^{-1})^{-1}(1 - \varepsilon_{\beta,w}\tilde{s}_{\beta,w}^{-1})^{-1}, \tag{1.6}$$

*so the relative error in replacing* $k_w$ *by its upper bound* $(rr_*)^w \tilde{k}_w$ *vanishes geometrically with a computable bound. The foregoing also proves* $\lim_{w\to\infty}(rr_*)^{-w} k_w = \lim_{w\to\infty} \tilde{k}_w = \tilde{k}$.

**Remark 1.2.** As a consequence of Theorem 1.1, Theorem 1.2, and the argument following (1.3), $\lambda = \hat{\lambda}$ if and only if $rr_* = 1$ if and only if $\lim_{w\to\infty} k_w = \tilde{k}$ if and only if $\boldsymbol{P}_\lambda \mathbf{1} = \mathbf{1}$ or

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

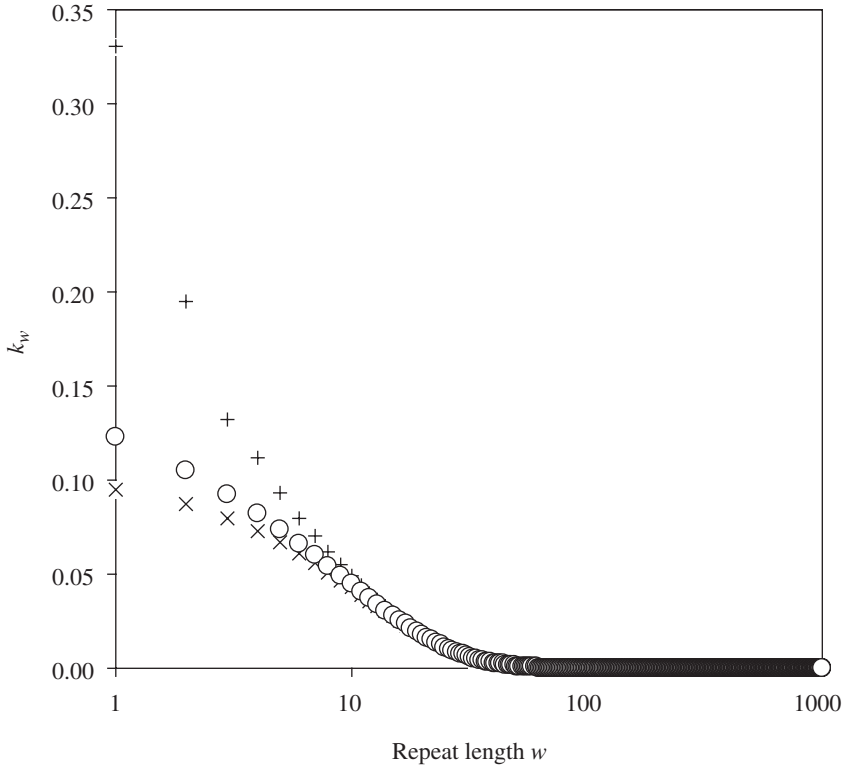


FIGURE 2: A plot illustrating the approximation $k_w \approx (rr_*)^w \tilde{k}_w$ in Theorem 1.2. Simulation results were obtained under the same conditions as in Figure 1: a binary alphabet $\mathfrak{L} = \{0, 1\}$, a uniform distribution $q_0 = q_1 = 0.5$, and a similarity score matrix with elements $s(0, 0) = 5$, $s(1, 1) = 2$, and $s(0, 1) = s(1, 0) = -8$. The simulation described in Section 3 generated estimates for $k_w$ ($w = 1, \ldots, 10^3$) (open circles) from a single random sequence of length $m = 10^7$. The upper points (+) are the approximating upper bound $(rr_*)^w \tilde{k}_w$; the lower points (×) are the approximations $(rr_*)^w \tilde{k}$, where $\tilde{k} = \lim_{w\to\infty} \tilde{k}_w$.

$\boldsymbol{q}^\top \boldsymbol{P}_\lambda = \boldsymbol{q}^\top$. In addition, let $\boldsymbol{q}^\top$ be uniform (i.e. $\boldsymbol{q} = (\#\mathfrak{L})^{-1}\mathbf{1}$) and the score matrix symmetric, $\boldsymbol{P}_\lambda \mathbf{1} = \mathbf{1}$. Then, $\boldsymbol{q}^\top \boldsymbol{P}_\lambda = (\boldsymbol{P}_\lambda \boldsymbol{q})^\top = \boldsymbol{q}^\top$, so that both $\boldsymbol{P}_\lambda \mathbf{1} = \mathbf{1}$ and $\boldsymbol{q}^\top \boldsymbol{P}_\lambda = \boldsymbol{q}^\top$ hold, yielding $\boldsymbol{\pi}^* = \boldsymbol{q}$. Thus, $\tilde{\mathrm{P}}_\alpha = \hat{\mathrm{P}}$ and $\tilde{\mathrm{P}}_\beta\{\tilde{X} = s(a, b)\} = q_a q_b \exp\{\hat{\lambda} s(a, b)\}$, so that $\tilde{k} = \hat{k}$, where $\hat{k}$ is the so-called '$k$ for gapless alignment' in bioinformatics [8].

The simplifications in Remark 1.2 have the following import. Most matrices used for the nucleotide alphabet $\mathfrak{L} = \{A, C, G, T\}$ are 'match-mismatch scores', i.e. for some $\{s_-, s_+\}$ and any $a, b \in \mathfrak{L}$, $s(a, a) = s_+ > 0$ and $s(a, b) = s_- < 0$ for $a \neq b$. For any match-mismatch score, if the letters of $\mathfrak{L}$ are uniformly distributed ($q_b = (\#\mathfrak{L})^{-1}$ for all $b \in \mathfrak{L}$), Remark 1.2 pertains. Now, some masking programs find repeats by aligning a sequence to itself and assuming heuristically that the resulting self-alignment scores are distributed as though the two identical aligned sequences were in fact independent, e.g. in the case of gapless alignment some masking programs assume $\lambda = \hat{\lambda}$. If the letters do not follow a uniform distribution (as in some clinically important organisms with GC- or AT- rich genomes, e.g. mycobacterium tuberculosis or malaria), Theorem 1.1 and Theorem 1.2 suggest that the naïve substitution of $\hat{\lambda}$ for $\lambda$ yields inaccurate approximations for random repeat frequencies (see Figure 1).

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

A computer program calculating the tail probabilities (p-values) from Theorem 1.1 is presently being used in a study of nonglobular protein domains and will be made available upon publication of that study. Readers interested in practical applications of Theorem 1.1 and Theorem 1.2 can examine a related bioinformatics article [1], which appeared while this article was under review. Using the present notation, Achaz *et al.* [1] suggested the use of $S_{\text{best}-\text{repeat}} := \max_w \hat{M}_m$ as a repeat statistic. They assumed that the asymptotic distribution

$$\lim_{y\to\infty} \mathrm{P}\{\max_w \hat{M}_m \geq y\} = (k\mathrm{e}^{-\lambda y})m$$

and estimated $\lambda$ with crude Monte Carlo sampling of $\max_w \hat{M}_m$. However, their analysis did not examine contributions to $k$ from different word-lengths $w$. Interestingly, in random sequences, all word-lengths $w$ have the same exponential rate $\lambda$, but the small word-lengths $w$ can dominate the statistic $\max_w \hat{M}_m$ (see Figure 1 and Figure 2).

The layout of this paper is as follows. In Section 2 we describe the MAP pertinent to repeats and includes an essential simplification that permits the practical application of general MAP theory to repeats. In Section 3 we present a formula for the pre-factor $k_w$. Because of the formula's complexity, $k_w$ must be determined by simulation, if the approximation in Theorem 1.2 is not accurate enough. In Section 3 we suggest a change of measure to increase the statistical efficiency when simulating $k_w$. Finally, in Section 4 we present a proof of Theorem 1.2.

## 2. The $w$-repeat Markov additive process

Define $A_k := L(k, k+w]$, so that $\{A_k \in \mathfrak{W}\colon k = 0, 1, \ldots\}$ is an MC on the state-space $\mathfrak{W} := \mathfrak{L}^w$. The transitions $A_{k-1} \to A_k$ are $L(k-1, k+w-1] \to L(k, k+w]$, with each letter $L_{k+w}$ chosen independently from the distribution $\{q_l : l \in \mathfrak{L}\}$. The MC permits transitions from one word to another, only if the suffix of length $w-1$ of the first word equals a prefix of the second word. Index the rows and columns of vectors and matrices over $\mathfrak{W}$ by $a = (a_1, \ldots, a_w)$ and $b = (b_1, \ldots, b_w)$, the indexes $a$ and $b$ ranging over $\mathfrak{W} := \mathfrak{L}^w$ in the same order. Unless specified otherwise, all vectors are column vectors of dimension $\#\mathfrak{W} := (\#\mathfrak{L})^w$ and all matrices are of dimension $(\#\mathfrak{W}) \times (\#\mathfrak{W})$. Define the generalized Kronecker delta $\delta_{a_2,\ldots,a_w;b_1,\ldots,b_{w-1}} := 1$ if $(a_2, \ldots, a_w) = (b_1, \ldots, b_{w-1})$ and 0 otherwise. The MC $\{A_k\}$ has stationary distribution $\bar{\boldsymbol{\pi}}^\top := [\bar{\pi}_b] := [q_{b_1} \ldots q_{b_w}]$ and transition matrix

$$\bar{\boldsymbol{P}} := \|\bar{p}_{a;b}\| := \|\bar{p}_{a_1,\ldots,a_w;b_1,\ldots,b_w}\| = \|\delta_{a_2,\ldots,a_w;b_1,\ldots,b_{w-1}} q_{b_w}\|,$$

of dimension $(\#\mathfrak{W}) \times (\#\mathfrak{W})$. Let $\mathrm{P}_{\bar{\boldsymbol{\gamma}}}$ and $\mathrm{E}_{\bar{\boldsymbol{\gamma}}}$ respectively denote the MC probability measure and expectation corresponding to the initial $1 \times (\#\mathfrak{W})$ distribution $\bar{\boldsymbol{\gamma}}^\top$.

For the $w$-repeat score

$$Y_k := s(L_k, L_{k+w}), \qquad Y_0 := 0,$$

we can verify that $\{(A_k, Y_k)\colon k = 0, 1, \ldots\}$ satisfies the properties of a MAP, as follows. First, as noted above, $(A_k)$ is an MC. Second, conditioned on the complete MC $(A_k)$, the distributions of $Y_k$ are independent. (In fact, each $Y_k := s(L_k, L_{k+w})$ is conditionally deterministic.) Third, the marginal distribution of $Y_k$ is determined by the transition $A_{k-1} \to A_k$, $A_{k-1}$ determines $L_k$ and $A_k$ determines $L_{k+w}$. With $S_k := \sum_{i=1}^k Y_i$, we call $\{(A_k, S_k)\}$ the '(inexact simple) $w$-repeat MAP'.

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

To apply MAP theory for the local maximum, note the following. First,

$$\begin{aligned}\bar{\mu} := \mathrm{E}_{\bar{\pi}}\, Y_1 &= \sum_{a,b\in\mathfrak{L}} q_{a_1}\ldots q_{a_w}\delta_{a_2,\ldots,a_w;b_1,\ldots,b_{w-1}} q_{b_w} s(a_1, b_w)\\ &= \sum_{a_1,b_w\in\mathfrak{L}} q_{a_1} q_{b_w} s(a_1, b_w)\\ &= \mu < 0.\end{aligned}$$

Second, there exists a cycle of states $A_0, A_1, \ldots, A_m = A_0$ such that $\min_{k=1,\ldots,m} S_k > 0$. MAP theory immediately yields the limit

$$\lim_{y\to\infty} \mathrm{P}\{\hat{M}_m \geq y\} = (k_w \mathrm{e}^{-\bar{\lambda}y})m$$

given in Theorem 1.1, where the formula for the pre-factor $k_w$ is delayed to Section 3, and the exponential rate $\bar{\lambda}$ is defined immediately below. The main task in the rest of this section is to identify $\bar{\lambda}$ as $\lambda$ in Theorem 1.1.

General MAP theory gives the exponential rate $\bar{\lambda}$, as follows. The moment-generating matrix for the (trivial) random variate $s(a_1, b_w)$ is given by

$$\bar{\boldsymbol{P}}_\theta = \|\delta_{a_2,\ldots,a_w;b_1,\ldots,b_{w-1}} q_{b_w} \exp\{\theta s(a_1, b_w)\}\|. \tag{2.1}$$

As discussed in the introduction, because the MC $(A_k)$ is irreducible and aperiodic, the Perron–Frobenius theorem states that $\bar{\boldsymbol{P}}_\theta$ has a strictly dominant eigenvalue $\bar{\rho}(\theta) > 0$ (i.e. $\bar{\rho}(\theta)$ is the single eigenvalue with greatest absolute value), with $\bar{\rho}(\theta)$ corresponding to a strictly positive right eigenvector, which is unique within a multiplicative factor. Because

$$\bar{\rho}(\theta) = \max\{\bar{\boldsymbol{u}}^\top \bar{\boldsymbol{P}}_\theta \bar{\boldsymbol{v}} \colon \bar{\boldsymbol{u}}^\top \bar{\boldsymbol{v}} = 1;\, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}} \geq \boldsymbol{0}\}$$

is the maximum of functions that are convex in $\theta$, it also is a convex function. Because $\bar{\boldsymbol{P}} = \bar{\boldsymbol{P}}_0$ is stochastic, $\bar{\rho}(0) = 1$. Because $\bar{\mu} < 0$, there exists a unique $\theta = \bar{\lambda} > 0$ such that $\bar{\rho}(\bar{\lambda}) = 1$ (see the introduction).

Although (2.1) leads to a definition of $\bar{\lambda}$, direct application of the definition is usually impractical because the relevant matrices have $(\#\mathfrak{W})^2 = (\#\mathfrak{L})^{2w}$ elements and grow exponentially with $w$. In fact, general MAP formulas usually defy practical implementation because of their sheer complexity. Here, however, the $w$-repeat MAP yields an essential simplification for a Markov sequence of order zero, permitting a practical computation of $\bar{\lambda}$ from matrices of dimension $(\#\mathfrak{L}) \times (\#\mathfrak{L})$.

As motivation for the simplification, a $w$-repeat has $w$ 'phases' with each phase corresponding to a letter position within the repeating $w$-word [23]. The $w$-repeat MAP $\{(A_k, S_k)\}$ therefore decomposes into $w$ 'phased MAPs' corresponding to $\{(A_{i+kw}, \sum_{j=0}^{k} Y_{i+jw}) \colon k = 0, 1, \ldots\}$ $(i = 0, \ldots, w-1)$. Intuitively, because the $w$-repeat MAP $\{(A_k, S_k)\}$ can be reconstructed from the phased MAPs, the phased MAPs determine $\bar{\lambda}$. The two paragraphs preceding Theorem 1.1 (which the reader should now review) contain several definitions pertinent to the phased MAP.

In the following, context and subscripts distinguish the row and column indices $a = (a_1, \ldots, a_w)$ and $b = (b_1, \ldots, b_w)$ from the coordinates they contain. We now proceed to identify $\bar{\lambda}$ in the $w$-repeat MAP as $\lambda$ in the phased MAPs.

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

Consider the $w$-repeat MAP and the column vector $\bar{\boldsymbol{v}} = [v_{a_1} \dots v_{a_w}]$ of dimension $\#\mathfrak{W}$. Note that $\bar{\boldsymbol{v}}$ is a right eigenvector of $\bar{\boldsymbol{P}}_\lambda$ ($\lambda$, not $\bar{\lambda}$) with eigenvalue 1:

$$\begin{aligned}
\bar{\boldsymbol{P}}_\lambda \bar{\boldsymbol{v}} &= \|\delta_{a_2,\dots,a_w;b_1,\dots,b_{w-1}} q_{b_w} \exp\{\lambda s(a_1, b_w)\}\| [v_{a_1} \dots v_{a_w}] \\
&= \left[ \sum_{(c_1,c_2,\dots,c_w)} \delta_{a_2,\dots,a_w;c_1,\dots,c_{w-1}} q_{c_w} \exp\{\lambda s(a_1, c_w)\} v_{c_1} \dots v_{c_w} \right] \\
&= \left[ v_{a_2} \dots v_{a_w} \sum_{(c_w)} q_{c_w} \exp\{\lambda s(a_1, c_w)\} v_{c_w} \right] \\
&= [v_{a_1} \dots v_{a_w}] \\
&= \bar{\boldsymbol{v}}.
\end{aligned}$$

Let $\bar{\boldsymbol{V}} := \operatorname{diag}(\bar{\boldsymbol{v}}) := \|\bar{v}_a \delta_{a;b}\|$. In analogy with the phased MAPs, the associated transition matrix for the $w$-repeat MAP is given by

$$\begin{aligned}
\bar{\boldsymbol{P}}^* := \bar{\boldsymbol{V}}^{-1} \bar{\boldsymbol{P}}_\lambda \bar{\boldsymbol{V}} &= \|v_{a_1}^{-1} v_{b_w} \delta_{a_2,\dots,a_w;b_1,\dots,b_{w-1}} q_{b_w} \exp\{\lambda s(a_1, b_w)\}\| \\
&= \|\delta_{a_2,\dots,a_w;b_1,\dots,b_{w-1}} p^*_{a_1;b_w}\|.
\end{aligned} \tag{2.2}$$

Because $\bar{\boldsymbol{P}}^* = \|\bar{p}^*_{a,b}\|$ is similar to $\bar{\boldsymbol{P}}_\lambda$, it has the same eigenvalues. Because $\bar{\boldsymbol{P}}^*$ is stochastic, $\bar{\rho}(\lambda) = 1$. Because only one $\bar{\lambda} > 0$ can satisfy the equation $\bar{\rho}(\bar{\lambda}) = 1$, the equality $\bar{\lambda} = \lambda$ follows.

The stochastic matrix $\bar{\boldsymbol{P}}^*$ has stationary distribution $(\bar{\boldsymbol{\pi}}^*)^\top = [\bar{\pi}^*_b] := [\pi^*_{b_1} \dots \pi^*_{b_w}]$:

$$\begin{aligned}
(\bar{\boldsymbol{\pi}}^*)^\top \bar{\boldsymbol{P}}^* &= [\pi^*_{b_1} \dots \pi^*_{b_w}] \|\delta_{a_2,\dots,a_w;b_1,\dots,b_{w-1}} p^*_{a_1;b_w}\| \\
&= \left[ \sum_{(c)} \pi^*_{c_1} \dots \pi^*_{c_w} \delta_{c_2,\dots,c_w;b_1,\dots,b_{w-1}} p^*_{c_1;b_w} \right] \\
&= \left[ \pi^*_{b_1} \dots \pi^*_{b_{w-1}} \sum_{(c_1)} \pi^*_{c_1} p^*_{c_1;b_w} \right] \\
&= (\bar{\boldsymbol{\pi}}^*)^\top.
\end{aligned} \tag{2.3}$$

Let $\mathrm{P}^*_{\bar{\boldsymbol{\gamma}}}$ and $\mathrm{E}^*_{\bar{\boldsymbol{\gamma}}}$ respectively denote the probability measure and expectation corresponding to the initial distribution $\bar{\boldsymbol{\gamma}}^\top$ and transition matrix $\bar{\boldsymbol{P}}^*$. In conjunction with the (deterministic) variates $Y_k := s(L_k, L_{k+w})$, the matrix $\bar{\boldsymbol{P}}^*$ defines an 'associated MC' and an 'associated MAP' for the $w$-repeat MAP.

For future reference, $\bar{\boldsymbol{P}}_\theta / \rho(\theta)$ (because of its dominant right eigenvector $\bar{\boldsymbol{v}}_\theta > \boldsymbol{0}$) is similar to a stochastic matrix, by analogy to $\bar{\boldsymbol{P}}_\lambda / \rho(\lambda)$ and (2.2). Consequently, $\bar{\boldsymbol{P}}_\theta / \rho(\theta)$ has dominant eigenvalue 1, so $\bar{\rho}(\theta) = \rho(\theta)$ for all $0 \le \theta \le \lambda$.

We finish our discourse on Theorem 1.1 with a proof of its final statement. Returning to the notation used in the introduction for the phased MAP (no over-bars), consider the strictly convex function $\hat{f}(\theta) := \boldsymbol{q}^\top \boldsymbol{P}_\theta \boldsymbol{1} - 1$, which satisfies $\hat{f}(0) = \hat{f}(\hat{\lambda}) = 0$. From $\boldsymbol{q}^\top \boldsymbol{1} = 1$, it follows that $\hat{f}(\lambda) = \boldsymbol{q}^\top \boldsymbol{P}_\lambda \boldsymbol{1} - 1 \le \rho(\lambda) - 1 = 0$, so $0 < \lambda \le \hat{\lambda}$. Equality $\hat{f}(\lambda) = 0$ holds if and only if $\boldsymbol{P}_\lambda \boldsymbol{1} = \boldsymbol{1}$ or $\boldsymbol{q}^\top \boldsymbol{P}_\lambda = \boldsymbol{q}^\top$.

## 3. The simulation of the pre-factor $k_w$

After some preliminary definitions, (3.2) in this section gives the pre-factor $k_w$ used in Theorem 1.1. Equation (3.2) is elegant in its generality, but it is rarely computationally tractable.

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

However, if (as in Section 2) the right eigenvector $\bar{\boldsymbol{v}}$ and the distribution of the associated MAP have been determined, a general simulation technique (given after (3.2)) can be used to determine $k_w$.

Consider the global sums $(S_k : k = 0, 1, \ldots)$, and define inductively their descending ladder epochs (DLEs) $\alpha_0 := 0$ and $\alpha_{i+1} := \min\{m > \alpha_i : S_m \le S_{\alpha_i}\}$, with $\alpha := \alpha_1$ (where $\min \varnothing := \infty$). The states of successive DLEs in the original MAP constitute an MC with the transition matrix $\bar{\boldsymbol{P}}^{(\alpha)} := \|\mathrm{P}_a\{A_\alpha = b\}\|$ and some stationary distribution $(\bar{\boldsymbol{\pi}}^{(\alpha)})^\top = (\bar{\boldsymbol{\pi}}^{(\alpha)})^\top \bar{\boldsymbol{P}}^{(\alpha)}$.

Define $(\alpha) := \{\alpha_k : k = 0, 1, \ldots\}$ (the set of DLEs); $\widehat{\alpha}_j := \max\{k \in (\alpha) : k < j\}$ (the DLE preceding $j$); and $\Delta_\alpha S_j = \sum_{k=\widehat{\alpha}_j+1}^{j} Y_k$ (the segmental sum between $j$ and the DLE preceding $j$). The DLEs generate a Markov renewal process $\{(A_{\alpha_k}, \alpha_k) : k = 0, 1, \ldots\}$. Reward the interval $(\alpha_{k-1}, \alpha_k]$ with $\bar{v}_a\{1 - \exp(\lambda \Delta_\alpha S_{\alpha_k})\}$ if $A_{\alpha_k} = a$. By the Markov version of the renewal-reward theorem [16, p. 132], the limiting reward per unit time is

$$\lim_{j\to\infty} \mathrm{E}_a\left[\bar{v}_{A_j}\{1 - \exp(\lambda \Delta_\alpha S_j)\};\ j \in \alpha)\right] = \frac{\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\alpha)}}[\bar{v}_{A_\alpha}\{1 - \exp(\lambda S_\alpha)\}]}{\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\alpha)}}\,\alpha}, \tag{3.1}$$

for every starting state $A_{\alpha_0} = A_0 = a$.

Likewise, the global sums have strict ascending ladder epochs (SALEs) $\beta_0 := 0$ and $\beta_{i+1} := \min\{m > \beta_i : S_m > S_{\beta_i}\}$, with $\beta := \beta_1$. The states of successive SALEs in the associated MAP constitute an MC with the transition matrix $\bar{\boldsymbol{P}}^{(\beta)} := \|v_a^{-1} v_b \,\mathrm{E}_a[\exp(\lambda S_\beta);\ A_\beta = b,\ \beta < \infty]\|$ (proved stochastic by the optional stopping theorem from the theory of the Wald martingale [9]) and some stationary distribution $(\bar{\boldsymbol{\pi}}^{(\beta)})^\top = (\bar{\boldsymbol{\pi}}^{(\beta)})^\top \bar{\boldsymbol{P}}^{(\beta)}$.

A 'negative associated MAP' can be constructed from the associated transition matrix $\boldsymbol{P}^*$ and the variates $Y_k = -s(L_k, L_{k+w})$. The (strict) DLEs of the negative associated MAP lead to the analogs $(\beta)$, $\widehat{\beta}_j$, and $\Delta_\beta S_j$ of the quantities $(\alpha)$, $\widehat{\alpha}_j$, and $\Delta_\alpha S_j$ above, along with an analog of (3.1).

Dembo and Karlin [9] give several formulas relating to the pre-factor $k_w$, all summarized compactly by the following:

$$\begin{aligned} k_w &= \frac{\mu}{\theta_{-\delta}} \frac{\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\beta)}}[\bar{v}_{A_0^{-1}}\{\exp(\lambda S_\beta) - 1\};\ \beta < \infty]}{\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\beta)}}\{S_\beta \bar{v}_{A_\beta} \bar{v}_{A_0}^{-1} \exp(\lambda S_\beta);\ \beta < \infty\}} \frac{\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\alpha)}}[\bar{v}_{A_\alpha}\{1 - \exp(\lambda S_\alpha)\}]}{\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\alpha)}}\, S_\alpha} \\ &= \frac{1}{\theta_{-\delta}\mu_*} \frac{\mathrm{E}^*_{\bar{\boldsymbol{\pi}}^{(\beta)}}[\bar{v}_{A_\beta}^{-1}\{1 - \exp(-\lambda S_\beta)\}]}{\mathrm{E}^*_{\bar{\boldsymbol{\pi}}^{(\beta)}}\beta} \frac{\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\alpha)}}[\bar{v}_{A_\alpha}\{1 - \exp(\lambda S_\alpha)\}]}{\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\alpha)}}\,\alpha}, \end{aligned} \tag{3.2}$$

where $\mathrm{E}^*$ is defined after (2.3). The second equality in (3.2) is justified by theorems pertaining to a change of measure and by Markov versions of Wald's identity, i.e. $\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\alpha)}} S_\alpha = \mu\, \mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\alpha)}} \alpha$ and $\mathrm{E}^*_{\bar{\boldsymbol{\pi}}^{(\beta)}} S_\beta = \mu_*\, \mathrm{E}^*_{\bar{\boldsymbol{\pi}}^{(\beta)}} \beta$. Although written in the notation of the $w$-repeat MAP, (3.2) holds for any MAP and has no specialized simplifications.

In (3.2), expectations over the initial distributions $\bar{\boldsymbol{\pi}}^{(\alpha)}$ and $\bar{\boldsymbol{\pi}}^{(\beta)}$ sum at least $\#\mathfrak{W} = (\#\mathfrak{L})^w$ terms, typically too many for exhaustive enumeration, even if the distributions $\bar{\boldsymbol{\pi}}^{(\alpha)}$ and $\bar{\boldsymbol{\pi}}^{(\beta)}$ could be determined. However, simulations can efficiently evaluate $k_w$ to an arbitrary accuracy.

To estimate the factors involving $\bar{\boldsymbol{\pi}}^{(\alpha)}$ in (3.2), simulate the original MAP distribution, i.e. independent letters with distribution $\{q_l : l \in \mathfrak{L}\}$. The initial state $A_0$ can be chosen according to the stationary distribution $\bar{\boldsymbol{\pi}} = [q_{a_1} \ldots q_{a_w}]$ of the original MAP, which provides a ready surrogate for $\bar{\boldsymbol{\pi}}^{(\alpha)}$. The MC underlying the $w$-repeat MAP is assumed to be ergodic, thus, the ergodic theorem proves that, wp1-$\bar{\boldsymbol{P}}$ (i.e. with probability 1 under $\bar{\boldsymbol{P}}$), the average

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

of $\bar{v}_{A_j}\{1 - \exp(\lambda\Delta_\alpha S_j)\}$ over successive DLEs $j \in (\alpha)$ in the simulation approaches the right side of (3.1). Thus, a single realization estimates the ratio common to (3.1) and (3.2).

To estimate the factors involving $\bar{\boldsymbol{\pi}}^{(\beta)}$ in (3.2), simulate $w$ independent letter sequences $(A_i^*, A_{i+w}^*, A_{i+2w}^*, \ldots)$ $(i = 0, \ldots, w-1)$ from the associated MC, whose transition probabilities $\{p_{a;b}^*\}$ are given in (1.2). The initial states $A_i^*$ $(i = 0, \ldots, w-1)$ can be chosen according to the stationary distribution $\boldsymbol{\pi}^*$ of the associated MC. Now, interlace the sequences into a single letter sequence $(A_0^*, A_1^*, A_2^*, \ldots)$, which has the same transition probabilities as the associated MC for $w$-repeats. (The letter sequence has initial distribution $\bar{\boldsymbol{\pi}}^*$, which provides a ready surrogate for the stationary distribution $\bar{\boldsymbol{\pi}}^{(\beta)}$.) As above, wp1-$\bar{\boldsymbol{P}}^*$, the average of $\bar{v}_{A_\beta}^{-1}\{1 - \exp(-\lambda S_\beta)\}$ over successive (strict) DLEs $j \in (\beta)$ in the simulation of the negative associated MAP approaches the ratio involving $\mathrm{E}^*_{\bar{\boldsymbol{\pi}}^{(\beta)}}$ in the final expression of (3.2).

## 4. The approximation of the pre-factor $k_w$

Recall the probability space with measures $\tilde{\mathrm{P}}_\alpha$ and $\tilde{\mathrm{P}}_\beta$ described immediately before Theorem 1.1. Let $[j-w, j] := \{j-w, \ldots, j\}$, and define '$w$-quantities' analogous to $(\alpha)$, $\widehat{\alpha}_j$, and $\Delta_\alpha S_j$ above, but relevant to $[j-w, j]$, namely, the set of $w$-DLEs

$$(\alpha_{[j-w,j]}) := \{i \in [j-w, j] \colon S_i \le \min\{S_{j-w}, \ldots, S_{i-1}\}\}, \tag{4.1}$$

(where $\min \varnothing := \infty$, so $j - w \in (\alpha_{[j-w,j]})$); $\widehat{\alpha}_{[j-w,j]} := \max\{k \in (\alpha_{[j-w,j]}) \colon k < j\}$ (the $w$-DLE preceding $j$); and $\Delta_\alpha S_{[j-w,j]} = \sum_{k=\widehat{\alpha}_{[j-w,j]}+1}^{j} Y_k$ (the segmental sum between $j$ and the $w$-DLE preceding $j$). In the probability space corresponding to measure $\tilde{\mathrm{P}}_\alpha$, define analogs for the above quantities: substitute $\tilde{X}_k$ for $Y_k := s(L_k, L_{k+w})$ throughout the definitions, and embellish the analogs with over-tildes to indicate their provenance, e.g. $(\tilde{\alpha})$, $\Delta_\alpha \tilde{S}_{[j-w,j]}$, etc.

Motivated by a desire to approximate the value in (3.1), consider the corresponding $w$-quantity

$$\begin{aligned}
&\lim_{j\to\infty} \mathrm{E}_a[\bar{v}_{A_j}\{1 - \exp(\lambda\Delta_\alpha S_{[j-w,j]})\};\ j \in (\alpha_{[j-w,j]})] \\
&\quad = \lim_{j\to\infty} \mathrm{E}_a\left[\left(\prod_{i=j+1}^{j+w} v_{L_i}\right)\{1 - \exp(\lambda\Delta_\alpha S_{[j-w,j]})\};\ j \in (\alpha_{[j-w,j]})\right] \\
&\quad = \mathrm{E}_{\tilde{\pi}}\left[\left(\prod_{i=w+1}^{2w} v_{L_i}\right)\{1 - \exp(\lambda\Delta_\alpha S_{[0,w]})\};\ w \in (\alpha_{[0,w]})\right] \\
&\quad = r^w \tilde{\mathrm{E}}_\alpha[1 - \exp(\lambda\Delta_\alpha \tilde{S}_w);\ w \in (\tilde{\alpha})],
\end{aligned} \tag{4.2}$$

where the equalities are justified as follows. The first equality reflects the definition of $\bar{v}_{A_j}$. The second holds because the second expression is the limit of a bounded function of the variates $\{A_k \colon k = j-w, \ldots, j\}$, and the function's distribution is determined by the distribution of $A_{j-w}$, which is stationary in the limit. The Markov property justifies the final equality, because $\bar{\boldsymbol{\pi}}^\top = [q_{b_1} \ldots q_{b_w}]$. (The quantities $r$ and $\tilde{\mathrm{E}}$ are defined after Theorem 1.1 in the introduction.) The quantity $\tilde{\mathrm{E}}_\alpha[1 - \exp(\lambda\Delta_\alpha \tilde{S}_w);\ w \in (\tilde{\alpha})]$ satisfies (1.4), and is referred to as $\tilde{s}_{\alpha,w}$ in Theorem 1.2. Note that the $w$-quantities are determined by the Markov states $\{A_k \colon k = j-w, \ldots, j\}$, which permitted removal of Markov dependency in the final expression of (4.2).

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

Let an indicator random variate $\mathbf{1}\{\bullet\} = 1$ if the event $[\bullet]$ occurs and 0 otherwise. Define the difference

$$\begin{aligned} D_{w,j} &:= \{1 - \exp(\lambda \Delta_\alpha S_{[j-w,j]})\}\, \mathbf{1}\{j \in (\alpha_{[j-w,j]})\} - \{1 - \exp(\lambda \Delta_\alpha S_j)\}\, \mathbf{1}\{j \in (\alpha)\} \\ &= \{1 - \exp(\lambda \Delta_\alpha S_{[j-w,j]})\}\, \mathbf{1}\{j \in (\alpha_{[j-w,j]}) \setminus (\alpha)\} \\ &\quad + \{\exp(\lambda \Delta_\alpha S_j) - \exp(\lambda \Delta_\alpha S_{[j-w,j]})\}\, \mathbf{1}\{j \in (\alpha)\}, \end{aligned} \tag{4.3}$$

(noting that $\alpha$, $\Delta_\alpha S_j$, etc., depend implicitly on $w$). Now, $\Delta_\alpha S_{[j-w,j]} \le 0$ on $[j \in (\alpha_{[j-w,j]})]$, so the first term is nonnegative and no more than $\mathbf{1}\{j \in (\alpha_{[j-w,j]}) \setminus (\alpha)\}$. Moreover, because $(\alpha) \subseteq (\alpha_{[j-w,j]})$, the event $j \in (\alpha)$ implies $\Delta_\alpha S_j = \Delta_\alpha S_{[j-w,j]}$ unless the DLE prior to $j$ satisfies $\widehat{\alpha}_j < j - w$, in which case $\Delta_\alpha S_{[j-w,j]} < \Delta_\alpha S_j \le 0$. The second term is therefore nonnegative and no more than $\mathbf{1}\{j \in (\alpha)\}\, \mathbf{1}\{\widehat{\alpha}_j < j - w\}$. Thus, $0 \le D_{w,j}$ and

$$\begin{aligned} D_{w,j} &\le \mathbf{1}\{j \in (\alpha_{[j-w,j]}) \setminus (\alpha)\} + \mathbf{1}\{j \in (\alpha)\}\, \mathbf{1}\{\widehat{\alpha}_j < j - w\} \\ &\le 2\, \mathbf{1}\{S_j \le S_{j-w} \text{ and } \min\{S_0, S_1, \ldots, S_{j-w-1}\} < S_{j-w}\}. \end{aligned} \tag{4.4}$$

Let the event in the final indicator be $F$. Because the MC equilibrates, for any initial state $a$, bounded convergence yields

$$0 \le \lim_{j \to \infty} \mathrm{E}_a[\bar{v}_{A_j} D_{w,j}] = \mathrm{E}_{\bar{\pi}}[\bar{v}_{A_j} D_{w,j}] \le 2\, \mathrm{E}_{\bar{\pi}}[\bar{v}_{A_j}; F]. \tag{4.5}$$

Let $\bar{\boldsymbol{\delta}}_c := [\delta_{c,a}]$ be the column vector of 0s with a single 1 at the $c$th coordinate. The eigenvalue $\bar{\rho}(\theta) = \rho(\theta)$ satisfies the inequality $\bar{\boldsymbol{u}}^\top \bar{\boldsymbol{P}}_\theta \bar{\boldsymbol{w}} \le \rho(\theta) \bar{\boldsymbol{u}}^\top \bar{\boldsymbol{w}}$ for all $\bar{\boldsymbol{u}}, \bar{\boldsymbol{w}} \ge \mathbf{0}$ (see the paragraph following (2.3)). Set $\bar{\boldsymbol{w}} = \bar{\boldsymbol{\delta}}_a$ and $\bar{\boldsymbol{u}} = \bar{\boldsymbol{\pi}}$ to derive $\bar{\boldsymbol{\pi}}^\top \bar{\boldsymbol{P}}_\theta \bar{\boldsymbol{\delta}}_a \le \rho(\theta) \bar{\pi}_a$, so $\bar{\boldsymbol{\pi}}^\top \bar{\boldsymbol{P}}_\theta \le \rho(\theta) \bar{\boldsymbol{\pi}}^\top$. Iteration then yields $\bar{\boldsymbol{\pi}}^\top \bar{\boldsymbol{P}}_\theta^k \le \rho^k(\theta) \bar{\boldsymbol{\pi}}^\top$. For $k = 0, \ldots, j - w - 1$ and $\theta \ge 0$, therefore,

$$\begin{aligned} \mathrm{P}_{\bar{\pi}}\{S_{j-w} - S_k > 0;\, A_{j-w} = a\} &\le \mathrm{E}_{\bar{\pi}}[\exp\{\theta(S_{j-w} - S_k)\};\, A_{j-w} = a] \\ &= \bar{\boldsymbol{\pi}}^\top \boldsymbol{P}_\theta^{j-w-k} \bar{\boldsymbol{\delta}}_a \\ &\le \rho^{j-w-k}(\theta) \bar{\pi}_a, \end{aligned} \tag{4.6}$$

where the first inequality is of the Chernoff type; the second equality rewrites the expectation in matrix notation; and the third inequality follows from $\bar{\boldsymbol{\pi}}^\top \bar{\boldsymbol{P}}_\theta^k \le \rho^k(\theta) \bar{\boldsymbol{\pi}}^\top$.

For any $0 < \theta < \lambda$, return to (4.5) to find that

$$\begin{aligned} \mathrm{E}_{\bar{\pi}}[\bar{v}_{A_j}; F] &\le 2 \sum_{k=0}^{j-w-1} \mathrm{E}_{\bar{\pi}}[\bar{v}_{A_j}; S_j \le S_{j-w} \text{ and } S_k < S_{j-w}] \\ &= 2 \sum_{k=0}^{j-w-1} \sum_{a \in \mathfrak{W}} \mathrm{E}_{\bar{\pi}}[\bar{v}_{A_j}; S_{j-w} - S_j \ge 0 \mid A_{j-w} = a] \\ &\quad \times \mathrm{P}_{\bar{\pi}}[S_{j-w} - S_k > 0;\, A_{j-w} = a] \\ &\le 2 \sum_{k=0}^{j-w-1} \rho^{j-w-k}(\theta) \sum_{a \in \mathfrak{W}} \bar{\pi}_a\, \mathrm{E}_a[\bar{v}_{A_w} \exp(\theta S_w);\, S_w \le 0] \\ &\le 2\, \mathrm{E}_{\bar{\pi}}[\bar{v}_{A_w} \exp(\theta S_w)] \frac{\rho(\theta)}{1 - \rho(\theta)} \\ &= r^w \{\tilde{\rho}_\alpha(\theta)\}^w \frac{\rho(\theta)}{1 - \rho(\theta)}, \end{aligned} \tag{4.7}$$

https://doi.org/10.1239/jap/1183667418 Published online by Cambridge University Press

with the following justification. The first inequality follows because of Boole's union-sum inequality applied to $[\min\{S_0, S_1, \ldots, S_{j-w-1}\} < S_j] = \bigcup_{k=0}^{j-w-1} [S_k < S_j]$. The second equality follows from conditioning on the event $[A_{j-w} = a]$, and the fact that the future ($S_{j-w} - S_j$ and $A_j$) is conditionally independent of the past ($S_{j-w} - S_k$ and states prior to $A_{j-w}$). The third inequality follows from an application of (4.6) and using stationarity to translate the origin of time from epoch 0 to epoch $j - w$, thereby replacing $S_{j-w} - S_j$ with $-S_w$. The fourth inequality follows by summing the geometric series ($0 < \rho(\theta) < 1$) and by dropping any restriction to the event $[S_w \le 0]$. The relation $\mathrm{E}_{\bar{\pi}}[\bar{v}_{A_w} \exp(\theta S_w)] = \{r\tilde{\rho}_\alpha(\theta)\}^w$ yields the fifth equality.

The convenient but effective choice of $\theta = \underline{\theta}_\alpha$ in the final expression of (4.7) yields $\lim_{j\to\infty} \mathrm{E}_a[\bar{v}_{A_j} D_{w,j}] \le r^w \varepsilon_{\alpha,w}$, so by bounded convergence as $j \to \infty$, (3.1)–(4.7) give

$$0 \le r^w \tilde{s}_{\alpha,w} - \frac{\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\alpha)}}[\bar{v}_{A_\alpha}\{1 - \exp(\lambda S_\alpha)\}]}{\mathrm{E}_{\bar{\boldsymbol{\pi}}^{(\alpha)}}\, \alpha} \le r^w \varepsilon_{\alpha,w}. \tag{4.8}$$

Denote the second term in (4.8) by $s_{\alpha,w}$ to yield $1 \le r^w \tilde{s}_{\alpha,w}/s_{a,w} \le (1 - \varepsilon_{\alpha,w}\tilde{s}_{\alpha,w}^{-1})^{-1}$ for $0 < \varepsilon_{\alpha,w} < \tilde{s}_{\alpha,w}$.

In (3.2), the factors corresponding to the associated MAP can be treated similarly. For the associated MAP, (4.8) corresponds to

$$0 \le r_*^w \tilde{s}_{\beta,w} - \frac{\mathrm{E}^*_{\bar{\boldsymbol{\pi}}^{(\beta)}}[\bar{v}_{A_\beta}^{-1}\{1 - \exp(-\lambda S_\beta)\}]}{\mathrm{E}^*_{\bar{\boldsymbol{\pi}}^{(\beta)}}\, \beta} \le r_*^w \varepsilon_{\beta,w}. \tag{4.9}$$

Equation (4.9) is easily proved by considering the negative associated MAP again, where in an obvious notation, $\boldsymbol{P}^*_\theta = \boldsymbol{V}^{-1}\boldsymbol{P}_{\lambda-\theta}\boldsymbol{V}$. Replace $\rho^{j-w-k}(\theta)\bar{\pi}_a$ in (4.6) by $\rho^{j-w-k}(\lambda - \theta)\bar{\pi}^*_a$. The arguments above, applied to the negative associated MAP, then produce a factor of

$$\mathrm{E}^*_{\bar{\pi}^*}[\bar{v}_{A_w}^{-1} \exp(-\theta S_w)] = \{r_*\tilde{\rho}_\beta(\theta)\}^w$$

in (4.7), replacing $r^w\{\tilde{\rho}_\alpha(\theta)\}^w$ in (4.7) eventually leads to (4.9).

Denote the second term in (4.9) by $s_{\beta,w}$ to yield $1 \le r_*^w \tilde{s}_{\beta,w}/s_{\beta,w} \le (1 - \varepsilon_{\beta,w}\tilde{s}_{\beta,w}^{-1})^{-1}$ for $0 < \varepsilon_{\beta,w} < \tilde{s}_{\beta,w}$. Together, the error bounds in (4.8) and (4.9) for the original and negative associated MAP yield (1.6), which concludes the proof of Theorem 1.2.

## Acknowledgements

It is my pleasure to acknowledge conversations with Dr John Wootton, Dr Yonil Park, Mr Kannan Tharakaraman, Ms Aleksandra Safronova, and Dr Sergey Sheetlin, the latter two having contributed to computer code for evaluating simple repeats. This research was supported by the Intramural Research Program of the NIH.

## References

[1] Achaz, G. *et al.* (2007). Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics* **23,** 119–121.
[2] Arratia, R., Martin, D., Reinert, G. and Waterman, M. S. (1996). Poisson process approximation for sequence repeats, and sequencing by hybridization. *J. Comput. Biol.* **3,** 425–463.
[3] Asmussen, S. (2003). *Applied Probability and Queues*. Springer, New York.
[4] Biggins, J. D. (1987). A note on repeated sequences in Markov chains. *Adv. Appl. Prob.* **19,** 739–742.
[5] Biggins, J. D. and Cannings, C. (1987). Markov renewal processes, counters and repeated sequences in Markov chains. *Adv. Appl. Prob.* **19,** 521–545.

[6] Boeva, V., Regnier, M., Papatsenko, D. and Makeev, V. (2006). Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics* **22,** 676–684.
[7] Dembo, A. and Karlin, S. (1991). Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of Markov variables. *Ann. Prob.* **19,** 1756–1767.
[8] Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Nat. Acad. Sci. USA* **87,** 2264–2268.
[9] Karlin, S. and Dembo, A. (1992). Limit distributions of maximal segmental score among Markov dependent partial sums. *Adv. Appl. Prob.* **24,** 113–140.
[10] Karlin, S. and Ghandour, G. (1985). Comparative statistics for DNA and protein sequences. Multiple sequence analysis. *Proc. Nat. Acad. Sci. USA* **82,** 6186–6190.
[11] Karlin, S. and Ghandour, G. (1985). Comparative statistics for DNA and protein sequences. Single sequence analysis. *Proc. Nat. Acad. Sci. USA* **82,** 5800–5804.
[12] Karlin, S. *et al.* (1983). New approaches for computer analysis of nucleic-acid sequences. *Proc. Nat. Acad. Sci. USA* **80,** 5660–5664.
[13] Kingman, J. F. C. (1961). A convexity property of positive matrices. *Quart. J. Math. Oxford* **12,** 283–284.
[14] Lancaster, P. and Tismenetsky, M. (1985). *The Theory of Matrices*. Academic Press, New York.
[15] Regnier, M. (2000). A unified approach to word occurrence probabilities. *Discrete Appl. Math.* **104,** 259–280.
[16] Ross, S. (1996). *Stochastic Processes*. John Wiley, New York.
[17] Ruzzo, W. L. and Tompa, M. (1999). *A linear time algorithm for finding all maximal scoring subsequences*. In *Proc. Seventh Internat. Conf. Intelligent Systems Molec. Biol.*, Heidelberg, pp. 234–241.
[18] Shpaer, E. G. *et al.* (1996). Sensitivity and selectivity in protein similarity searches: a comparison of Smith–Waterman in hardware to BLAST and FASTA. *Genomics* **38,** 179–191.
[19] Smit, A., Hubley, R. and Green, P. (1996). RepeatMasker. Available at http://www.repeatmasker.org.
[20] States, D. J. and Agarwal, P. (1996). Compact encoding strategies for DNA sequence similarity search. In *Proc. Internat. Conf. Intelligent Systems Molec. Biol.*, Heidelberg, pp. 211–217.
[21] Wootton, J. C. (1994). Sequences with 'unusual' amino acid composition. *Current Opinion Structural Biol.* **4,** 413–421.
[22] Wootton, J. C. and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17,** 149–163.
[23] Wootton, J. C. and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol.* **266,** 554–571.