Adv. Appl. Prob. **38,** 1071–1097 (2006)
Printed in Northern Ireland
© *Applied Probability Trust* 2006

A SIMPLE MAXIMIZATION MODEL INSPIRED BY ALGORITHMS FOR THE ORGANIZATION OF GENETIC CANDIDATES IN BACTERIAL DNA

ANDREW G. HART,* **
SERVET MARTÍNEZ * *** AND
LEONARDO VIDELA,* **** *Universidad de Chile*

Abstract

We propose a simple model for interaction between gene candidates in the two strands of bacterial DNA (deoxyribonucleic acid). Our model assumes that 'final' genes appear in one of the two strands, that they do not overlap (in bacteria there is only a small percentage of overlap), and that the final genes maximize the occupancy rate, which is defined to be the proportion of the genome occupied by coding zones. We are more concerned with describing the organization and distribution of genes in bacterial DNA than with the very hard problem of identifying genes. To this end, an algorithm for selecting the final genes according to the previously outlined maximization criterion is proposed. We study the graphical and probabilistic properties of the model resulting from applying the maximization procedure to a Markovian representation of the genic and intergenic zones within the DNA strands, develop theoretical bounds on the occupancy rate (which, in our view, is a rather intractable quantity), and use the model to compute quantities of relevance to the *Escherichia coli* genome and compare these to annotation data. Although this work focuses on genomic modelling, we point out that the proposed model is not restricted to applications in this setting. It also serves to model other resource allocation problems.

Keywords: Constrained optimization; renewal process; Markov process

2000 Mathematics Subject Classification: Primary 60J27; 92D20

1. The model and the criterion

In the two strands of bacterial DNA (deoxyribonucleic acid), one distinguishes between 'possible' coding zones and noncoding zones. We will call a possible coding zone a gene candidate and any noncoding region an intergenic zone or gap. Each strand can be read according to any one of three reading frames. Current automated systems for gene annotation typically select gene candidates in each reading frame by using methods such as hidden Markov models [5], [6] or interpolated Markov models [7]. Afterwards, the final candidates for genes in bacterial DNA are selected so as to be nonoverlapping according to some criteria which

Received 4 September 2006; revision received 21 September 2006.

* Postal address: Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático, UMR 2071 CNRS-UCHILE, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Casilla 170-3, Correo 3, Santiago, Chile.
** Email address: ahart@dim.uchile.cl
*** Email address: smartine@dim.uchile.cl
**** Email address: lvidela@dim.uchile.cl

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

include deletion of small gene candidates and comparison with genes that have been previously annotated and made available in databases.

Here we propose a simple scheme for modelling the organization of genes based on a maximization principle which, in its first incarnation, does not take into account external information from databases or explicit knowledge of gene structure. This work is not primarily concerned with the identification of individual genes, which is the case with the automated, state-of-the-art gene annotation systems mentioned above. Instead, we are interested in the overall organization and distribution of genetic information throughout the genome.

We begin by introducing some notation. Let us denote the primary and complementary strands by 1 and -1 respectively. The three reading frames on the first strand are labelled 1, 2, and 3 and those on the second strand are labelled -1, -2, and -3. Each strand is modelled as a continuous half-line and the six ways of reading the DNA in a sequentially linear fashion are enumerated by $i \in \mathcal{R} = \{1, 2, 3, -1, -2, -3\}$. In each reading frame $i \in \mathcal{R}$, the line is partitioned into gene candidates, marked as 1-zones, and gaps, marked as 0-zones. The 1- and 0-zones are formed by a countable number of disjoint intervals and we assume, for technical reasons, that these intervals are left closed and right open, that is, of the form $[a, b)$, with $a < b$. The reversed version of the process has intervals that are left open and right closed, that is, of the form $(a, b]$.

For each reading frame $i \in \mathcal{R}$, $E^i = (E^i_t : t \in \mathbb{R})$ denotes the $\{0, 1\}$-valued random process, and the corresponding family of gene candidates is denoted by $\mathcal{I}^i$. By definition, $\{E^i_t = 1\} = \bigcup_{I \in \mathcal{I}^i} I$. The set of all gene candidates, regardless of the strand to which they belong, can be written $\mathcal{I} = \bigcup_{i \in \mathcal{R}} \mathcal{I}^i$. Let $I \in \mathcal{I}$ be an arbitrary gene candidate. We denote the start and end points of I by $s(I)$ and $f(I)$, respectively, so $I = [s(I), f(I))$, and we define $\boldsymbol{i}(I)$ to identify strand I with its reading frame: $\boldsymbol{i}(I) = i$ if $I \in \mathcal{I}^i$. In the event that I is contained in several $\mathcal{I}^i$, then $\boldsymbol{i}(I)$ is viewed as taking all these values.

It is clear that the candidates for genes in the three reading frames on the two strands can overlap, but the final organization of genes should be such that no overlap occurs, to the extent that this is observed in bacterial DNA. Consequently, the final organization can also be coded as a 0–1 process. We denote this process by $(E^*_t : t \in \mathbb{R})$. Its family of genes, denoted by $\mathcal{I}^*$, is the class of connected components of $E^*_t = 1$, so $\{E^*_t = 1\} = \bigcup_{I \in \mathcal{I}^*} I$. This family of genes satisfies the condition that every one of its elements, $I \in \mathcal{I}^*$, is a gene candidate in one of the strands, that is, $\mathcal{I}^* \subseteq \mathcal{I}$.

However, this condition does not identify a particular organization. For instance, if two gene candidates in different strands overlap, then a choice must be made such that at most one of these candidates appears as a final gene. We propose a maximization criterion which requires that the coding zone occupy as much of the DNA's length as possible. We proceed by describing this criterion.

Let K be a connected component of the set $\{t \geq 0 : \sum_{i \in \mathcal{R}} E^i_t \geq 1\}$. In other words, K is 'covered' by a series of overlapping gene candidates and contains no gaps. Any such interval K will be called an *island*. Furthermore, we shall use $s(K)$ and $f(K)$ to denote its start and, respectively, end points in the same way we use $s(I)$ and $f(I)$ to denote the start and end points of a gene candidate. As islands and gene candidates are both merely intervals, this apparent abuse of notation should not cause any confusion. Now, the set of all islands is denoted by $\mathcal{K}$. From the condition $\mathcal{I}^* \subseteq \mathcal{I}$, we know that each gene $I \in \mathcal{I}^*$ is contained in some island. Fix some island K and let $\mathcal{I}^i(K) = \{I \in \mathcal{I}^i : I \subseteq K\}$ be the set of gene candidates on strand $i \in \mathcal{R}$ included in K. Also define $\mathcal{I}(K) = \bigcup_{i \in \mathcal{R}} \mathcal{I}^i(K)$. We shall denote the sum of the lengths of a family of sets $\tilde{\mathcal{I}} \subseteq \mathcal{I}(K)$ by $\Sigma(\tilde{\mathcal{I}}) := \sum_{I \in \tilde{\mathcal{I}}} |I|$. The maximization criterion is then as follows:

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

the final set of genes, $\mathcal{I}^* = \bigcup_{K \in \mathcal{K}} \mathcal{I}^*(K)$, is such that, for each island K, the class of sets $\mathcal{I}^*(K)$ solves the constrained optimization problem

$$\Sigma(\mathcal{I}^*(K)) = \Sigma^{*M}(\mathcal{I}(K)),$$
$$\text{where } \Sigma^{*M}(\mathcal{I}(K)) := \max\{\Sigma(\tilde{\mathcal{I}})\colon \tilde{\mathcal{I}} \subseteq \mathcal{I}(K) \text{ is a disjoint family}\}. \quad (1)$$

We will assume that the length, $f(K) - s(K)$, of every island K is finite and that there are only a finite number of gene candidates in each finite interval, almost everywhere. Hence, the maximization problem reduces to a finite deterministic problem over each island.

Section 2 describes the algorithm for selecting final genes that maximize the occupancy rate while not overlapping. The behaviour and distribution of gene candidates, assuming that their placement is Markovian in nature, is then studied in Section 3. There we introduce a quantity called the occupancy rate, which may be loosely defined as the proportion of the genome area occupied by the coding zone. Properties of islands and results concerning the occupancy rate and the disposition of final genes are also given. We have no closed-form expression for the occupancy rate, but Section 5 provides a number of useful theoretical bounds on it. A graphical comparison between the various upper and lower bounds discussed is also included. In Section 7, we apply the model to the *Escherichia coli* K-12 genome and perform a number of numerical experiments to study the efficacy of the model in practice. Although our focus is on global organization of coding zones, the results of this section also serve to illustrate the performance of the model as a gene identification tool. To conclude, we make some final remarks in Section 8.

Although this article focuses on genomic modelling, we should point out that the proposed model is not restricted to applications in this setting. It could also serve to model any resource allocation problem in which there are a number of competitors demanding to use contiguous portions of the available resource and for which the aim is to maximize resource usage.

2. Maximizing the coding zone

Before beginning, we remark that the notation and discussion in both this and the preceding section hold for any finite set $\mathcal{R}$, not only for $\mathcal{R} = \{1, 2, 3, -1, -2, -3\}$. To work within the most general and flexible framework, we shall take $\mathcal{R}$ to be an arbitrary finite set with $N := |\mathcal{R}| \geq 2$. As $R \subset \mathbb{Z}$, we will impose the usual ordering '$\leq$' of the integers on it.

In this section, we describe an algorithm for selecting gene candidates which maximizes the amount of space occupied over an island. The maximization procedure is based on the dynamic programming paradigm and is linear in both its time and space (memory) complexity.

To simplify the discussion in this section, we will restrict our attention to a single, fixed island K. We will write $\mathcal{J}^i := \mathcal{I}^i(K)$, $i \in \mathcal{R}$, and $\mathcal{J} := \mathcal{I}(K) = \bigcup_{i \in \mathcal{R}} \mathcal{J}^i$, which are families of intervals in K. The fact that K is an island can be expressed as follows: K is an interval and $K = \bigcup_{I \in \mathcal{J}} I$.

We shall define the neighbourhood of a gene candidate I to be $\mathcal{N}(I) := \{J \in \mathcal{J}\colon J \cap I \neq \varnothing\}$. Suppose that $n := |\mathcal{J}|$ is the number of gene candidates. We shall order the gene candidates in $\mathcal{J}$ according to the following relation:

$$I \preceq_{\mathrm{f}} J \quad \text{if and only if} \quad f(I) < f(J);\ f(I) = f(J) \text{ and } s(I) > s(J);$$
$$\text{or } s(I) = s(J),\ f(I) = f(J), \text{ and } \boldsymbol{i}(I) \leq \boldsymbol{i}(J).$$

Thus, $\mathcal{J} = \{I_1, I_2, \dots, I_n\}$, where $I_j \preceq_{\mathrm{f}} I_k$, $1 \leq j \leq k \leq n$. The relation '$\preceq_{\mathrm{f}}$' is a total order which arranges the candidates primarily according to their end points, secondly according to

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

their lengths, and finally according to the strands on which they appear. Next define $t_j := s(I_j)$ and $T_j := f(I_j)$, for $j = 1, \dots, n$, whence $I_j = [t_j, T_j)$. Observe that

$$I_i \cap I_k \neq \varnothing \quad \Longrightarrow \quad I_j \cap I_k \neq \varnothing \qquad \text{for } 1 \leq i \leq j \leq k \leq n. \tag{2}$$

In fact, $T_j \in [t_k, T_k]$ because in the contrary case $T_i \leq T_j < t_k$ we would be led to conclude that $I_i \cap I_k = \varnothing$, contradicting the assumption.

Define $\bar{J}_0 = \varnothing$ and set $\bar{J}_m := \{I_1, I_2, \dots, I_m\}$ for $m = 1, 2, \dots, n$.

Lemma 1. *Let $m \geq 1$. Then there exists an l_m, $0 \leq l_m \leq m$, such that $\bar{J}_m \cap \mathcal{N}(I_m) = \bar{J}_m \setminus \bar{J}_{l_m}$.*

Proof. By definition, the result is trivially true for $m = 1$, so fix $m \geq 2$ and let $l = \min\{l' : I_{l'} \in \bar{J}_m \cap \mathcal{N}(I_m)\}$. First, observe that $I_l \cap I_m \neq \varnothing$ and $\bar{J}_{l-1} \cap \mathcal{N}(I_m) = \varnothing$. Then set $l_m = l - 1$, from which the result follows by fixing $i = l - 1$ and $k = m$ and applying relation (2) with $j = l, l+1, l+2, \dots, m$.

Next, let w_m be the value of any solution to the problem

$$\max\{\Sigma(\tilde{\mathcal{I}}) : \tilde{\mathcal{I}} \subseteq \bar{J}_m \text{ is a disjoint family}\}, \tag{3}$$

for all m, $0 \leq m \leq n$. The following algorithm solves (1).

Algorithm 1. 0. *Let $\hat{w}_0 = 0$ and $\hat{w}_1 = |I_1|$. Set $\hat{J}_0 = \varnothing$ and $\hat{J}_1 = \{I_1\}$. Also, define $\hat{f}_1 = 1$.*
1. *For each $m = 2, 3, \dots, n$, let l_m be the index such that $\mathcal{N}(I_m) \cap \bar{J}_m = \bar{J}_m \setminus \bar{J}_{l_m}$.*
2. *Recursively compute the following quantities: $\hat{w}_m = \max\{\hat{w}_{m-1}, \hat{w}_{l_m} + |I_m|\}$.*
3. *Recursively compute the optimal solutions corresponding to the $\hat{w}_m$, according to the following decision scheme.*

(a) *If $\hat{w}_{m-1} > \hat{w}_{l_m} + |I_m|$ then set $\hat{J}_m = \hat{J}_{m-1}$ and $\hat{f}_m = 0$.*

(b) *If $\hat{w}_{m-1} < \hat{w}_{l_m} + |I_m|$ then set $\hat{J}_m = \hat{J}_{l_m} \cup \{I_m\}$ and $\hat{f}_m = 1$.*

(c) *If $\hat{w}_{m-1} = \hat{w}_{l_m} + |I_m|$ then arbitrarily choose $\hat{J}_m$ to be $\hat{J}_{m-1}$ or $\hat{J}_{l_m} \cup \{I_m\}$, respectively setting $\hat{f}_m = 0$ or $\hat{f}_m = 1$.*

4. *Take the final gene selection to be $\hat{J} := \hat{J}_n$. The area occupied by this choice of genes is $\hat{w} := \hat{w}_n$.*

Proposition 1. *For all $m \geq 1$, $\hat{J}_m$ is an optimal solution over the subset $\bar{J}_m$ and $\hat{w}_m$ is the area occupied by this solution. In other words, $\hat{w}_m = \Sigma(\hat{J}_m) = \Sigma^{*M}(\bar{J}_m) = w_m$. Thus, Algorithm 1 produces an overall optimal solution $\hat{J} = \hat{J}_n$ and the amount of space it occupies in the DNA sequence is given by $\hat{w} = \hat{w}_n$.*

Proof. We proceed by induction.

Basis step: Clearly, $\hat{w}_0 = \Sigma(\hat{J}_0) = \Sigma(\varnothing) = 0 = \Sigma^{*M}(\bar{J}_0) = w_0$ and $\hat{w}_1 = \Sigma(\hat{J}_1) = \sigma(\{I_1\}) = |I_1| = \Sigma^{*M}(\bar{J}_1) = w_1$.

Inductive step: To verify the correctness of Algorithm 1, we need only realize that steps 2 and 3 deal with two mutually exclusive cases, namely

- I_m does not belong to a disjoint family yielding the solution to (3) and
- I_m belongs to a disjoint family yielding the solution to (3).

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

In the first case, the algorithm arrives at a solution to (3) for $\bar{J}_m$ by solving the same problem for $\bar{J}_{m-1}$. In the second case, the optimal solution is obtained by ruling out the gene candidates that overlap I_m and combining I_m with an optimal solution up to the last gene candidate that has index smaller than m and does not overlap I_m. Thus, given $\hat{w}_k = w_k$ for all $k = 0, 1, \dots, m-1$, steps 2 and 3 are guaranteed to produce a $\hat{J}_m$ satisfying $\Sigma(\hat{J}_m) = \hat{w}_m = w_m$.

Remark 1. The sequence $\hat{f}_m$, $m = 1, 2, \dots, n$, is neither needed nor used by Algorithm 1. We have defined it in order to facilitate subsequent discussion of the algorithm. Points m at which $\hat{f}_m = 1$ are important because they indicate when it is necessary to combine the current candidate I_m with a previous solution $\hat{J}_{l_m}$ in order to obtain an optimal solution over $\bar{J}_m$.

Now let $[a, b)$ be a finite interval. We define $\mathcal{I}^i([a, b)) = \{I \cap [a, b) \colon I \in \mathcal{I}^i\}$, $i = 1, -1$, and $\mathcal{I}([a, b)) = \mathcal{I}^1([a, b)) \cup \mathcal{I}^2([a, b))$. Observe that only pieces of genes can appear at the extremes a and b. Now denote by $\mathcal{I}^*([a, b))$ any solution to problem (1) over the interval $[a, b)$; that is, $\mathcal{I}^*([a, b))$ is a subset of nonoverlapping elements of $\mathcal{I}([a, b))$ such that $\Sigma(\mathcal{I}^*([a, b))) = \Sigma^{*M}(\mathcal{I}([a, b)))$.

Proposition 2. *Optimal solutions to (1) over the interval* $[a, c)$, $a < c$, *possess the following subadditive property:* $\Sigma^{*M}(\mathcal{I}([a, c))) \le \Sigma^{*M}(\mathcal{I}([a, b))) + \Sigma^{*M}(\mathcal{I}([b, c)))$ *for* $a \le b \le c$.

Proof. The inequality follows easily from the fact that, when considering the basis set $[a, c)$, the maximization problem (1) for the class of sets $\mathcal{I}([a, c))$ contains at least all the constraints of the same problem over $[a, c)$, but for the class of sets $\mathcal{I}([a, b)) \cup \mathcal{I}([b, c))$.

3. The Markovian hypothesis

In this section, we shall examine the probabilistic properties of the process obtained by applying the maximization algorithm described in the preceding section to two strands of DNA modelled as independent telegraph processes. So, we fix $\mathcal{R} = \{1, -1\}$ and $N = 2$.

3.1. The continuous processes

Consider the product process $(E_t = (E_t^1, E_t^{-1}) \colon t \ge 0)$ taking values in $\{0, 1\}^2$. Let $(\Omega, \mathcal{B}, \mathrm{P})$ be a probability space. The points $\omega \in \Omega$ are identified with the trajectories of the product process $(E_t(\omega) \colon t \ge 0) := ((E_t^1(\omega), E_t^{-1}(\omega)) \colon t \ge 0)$ that are right continuous with left limits (in the reversed version they are left continuous with right limits). We shall use $W_t = \int_0^t \mathbf{1}_{\{E_s^* = 1\}}\,\mathrm{d}s$ to denote the amount of bacterial DNA occupied by final genes in the region $[0, t)$.

Then we define the occupancy rate, M^*, to be the asymptotic proportion of the total genome length that is occupied by coding regions, that is, $M^* := \lim_{t\to\infty} W_t/t$.

Lemma 2. *Assume that the process* (E_t) *is stationary and ergodic. Then the limiting proportion of the space occupied is the well-defined quantity* $M^* = \lim_{t\to\infty} \mathrm{E}(W_t)/t$, P*-almost everywhere* (P *-a.e.).*

Proof. The result follows from Proposition 2 and the subadditive ergodic theorem (see [3, pp. 40–42]) applied to W_t.

We will assume that (E_t) is a Markov process and that the marginals $E^i(t)$, $i = 1, -1$, are independent and equally distributed. In this case, there is a unique final configuration of genes satisfying the maximization criterion a.e. We encode the states of (E_t) in $\{0, 1\}^2$ as $0 = (0, 0)$, $1 = (1, 0)$, $-1 = (0, 1)$, and $2 = (1, 1)$, and order them as written.

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

The generator, $Q^i = (q^i_{jl} \colon j, l \in \{0, 1\})$, of each of the marginal processes E^i is given by $q^i_{01} = \nu = -q^i_{00}$ and $q^i_{10} = \mu = -q^i_{11}$ for some ν and μ, $0 < \nu, \mu < \infty$. The lengths of the gene candidates and gaps in either of the strands are thus exponentially distributed with means $1/\mu$ and $1/\nu$, respectively (we denote their distributions by $\text{Exp}(\mu)$ and $\text{Exp}(\nu)$, respectively). We set $\rho := \nu/\mu$. The stationary distribution, $\pi = (\pi_0, \pi_1)$, of each marginal is given by $\pi_0 = 1 - \theta^*$ and $\pi_1 = \theta^*$, where $\theta^* \equiv \theta^*(\rho) := \rho/(1 + \rho)$.

Next, the generator, $Q^E = (q^E_{ij} \colon i, j \in \{0, 1, -1, 2\})$, of the process E_t is

$$Q^E = \begin{pmatrix} -2\nu & \nu & \nu & 0 \\ \mu & -(\mu+\nu) & 0 & \nu \\ \mu & 0 & -(\mu+\nu) & \nu \\ 0 & \mu & \mu & -2\mu \end{pmatrix}.$$

We can construct a version of the process $(E_t = (E^1_t, E^{-1}_t))$ with $t \in \mathbb{R}$. To this end, take $(N^{i,e}_t \colon i \in \{1, -1\},\ e \in \{0, 1\})$ to be four independent Poisson processes in $\mathbb{R}$, with rates given by ν if $e = 0$ and by μ if $e = 1$. We denote the marks (renewal times) of these processes in $\mathbb{R}$ by $(\mathcal{X}^{i,e}_l \colon l \in \mathbb{Z})$. A trajectory of the process $(E_t = (E^1_t, E^{-1}_t))$ with initial state $E_0 = (e^1_0, e^{-1}_0)$ is given by the following construction. First, we define sequences of random times as

$$\begin{aligned} \boldsymbol{x}^i_0 &= 0, \\ \boldsymbol{x}^i_{k+1} &= \inf\{\mathcal{X}^{(k+e^i_0)[2],i}_l > \boldsymbol{x}^i_k \colon l \in \mathbb{Z}\}, \qquad k \geq 0, \\ \boldsymbol{x}^i_{k-1} &= \sup\{\mathcal{X}^{(k+e^i_0)[2],i}_l < \boldsymbol{x}^i_k \colon l \in \mathbb{Z}\}, \qquad k \leq 0, \end{aligned}$$

where we set $n[2] = 0$ if n is even and $n[2] = 1$ if n is odd (in other words, $n[2] = n \pmod 2$). Then, for $t \geq 0$, we define $E^i_t = (k + e^i_0)[2]$, where k is the unique value such that $t \in [\boldsymbol{x}^i_k, \boldsymbol{x}^i_{k+1})$. For $t < 0$, we define $E^i_t = (k + e^i_0)[2]$ if $t \in [\boldsymbol{x}^i_{k-1}, \boldsymbol{x}^i_k)$, where $k \leq 0$ is the unique value such that $t \in [\boldsymbol{x}^i_{k-1}, \boldsymbol{x}^i_k)$. This produces the desired construction. Observe that $E^i_t = e^i_0$ for $t \in [\boldsymbol{x}^i_{-1}, \boldsymbol{x}^i_1)$.

Remark 2. The previous construction can be interpreted in the genomic model as follows. As observed, there are many initial codons for potential genes (corresponding to the marks $\{\mathcal{X}^{i,1}_l\}$ where a 1-zone could potentially start) and final codons (corresponding to the marks $\{\mathcal{X}^{i,0}_l\}$ where a 0-zone could potentially start). This is further elaborated upon in Section 7.1.

The above construction has a number of inherent symmetries.

Reversibility. The law of $(E_t \colon t \geq 0)$ is the same as that of $(E_{-t} \colon t \geq 0)$. Therefore, the reversed process, $({}^{\mathrm{r}}E_t := E_{-t} \colon t \in \mathbb{R})$, is distributed in the same way as $(E_t \colon t \in \mathbb{R})$ (but using the reversed version, which is left continuous with right limits).

Lumping. States 1 and -1 may be exchanged with each other; that is, $({}^{\mathrm{e}}E_t \colon t \in \mathbb{R})$ is distributed in the same way as $(E_t \colon t \in \mathbb{R})$, where ${}^{\mathrm{e}}E_t = E_t$ if $E_t \in \{0, 2\}$ and ${}^{\mathrm{e}}E_t = -E_t$ if $E_t \in \{1, -1\}$.

The stationary distribution, Π^E, of E_t is

$$\Pi^E_0 = \frac{1}{(1+\rho)^2}, \qquad \Pi^E_1 = \frac{\rho}{(1+\rho)^2} = \Pi^E_{-1}, \qquad \Pi^E_2 = \frac{\rho^2}{(1+\rho)^2}.$$

As expected from reversibility, the generator $Q^{{}^{\mathrm{r}}E} = (q^{{}^{\mathrm{r}}E}_{ij} := q^E_{ji}\Pi^E_j/\Pi^E_i \colon i, j \in \{0, 1, -1, 2\})$ of the reversed Markov process, ${}^{\mathrm{r}}E_t$, is the same as Q^E.

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

Lemma 3. *The occupancy rate $M^* \equiv M^*(\rho)$ depends only on ρ and is given by*

$$M^*(\rho) := \lim_{t\to\infty} \frac{1}{t}\,\mathrm{E}(W_t) = \lim_{t\to\infty} \frac{1}{t} W_t \quad \mathrm{P}\text{-}a.e. \tag{4}$$

Proof. Since $M^*(\nu, \mu) = M^*(c\nu, c\mu)$ for all $c > 0$, we immediately see that M^* depends only on ρ. Now let us show (4). If E_0 is distributed in the same way as Π^E, then (E_t) is a stationary ergodic process and from Lemma 2 we obtain the result. A coupling argument shows that this quantity is independent of the starting point, that is, if (E_t) starts from a point $E_0 \in \{0, 1, -1, 2\}$ (or any distribution on this set), then (4) also holds.

The process $(\tilde{E}_t := E_t^1 + E_t^{-1})$, $t \ge 0$, is a Markov process with state space $\{0, 1, 2\}$. This process is the 'lumped' process associated with (E_t), that is, where the states 1 and -1 are grouped together into a single state labelled 1. Note that, according to the state space encoding for (E_t) we have chosen to use, we have $\tilde{E}_t = |E_t|$. Let $Q^{\tilde{E}} = (q_{ij}^{\tilde{E}} : i, j \in \{0, 1, 2\})$ be its generator and $\Pi^{\tilde{E}} = (\Pi_i^{\tilde{E}} : i \in \{0, 1, 2\})$ its stationary distribution. By the lumping property, we have

$$q_{ij}^{\tilde{E}} = q_{ij}^{E} \quad \text{if } j \in \{0, 2\}, \qquad q_{i1}^{\tilde{E}} = q_{i1}^{E} + q_{i,-1}^{E} \quad \text{if } i \in \{0, 2\}, \qquad q_{11}^{\tilde{E}} = q_{11}^{E},$$
$$\Pi_i^{\tilde{E}} = \Pi_i^{E} \quad \text{if } i \in \{0, 2\}, \qquad \Pi_1^{\tilde{E}} = \Pi_1^{E} + \Pi_{-1}^{E}.$$

The process $(\tilde{E}_t)$ is also reversible in time; that is, the law of $({}^{\mathrm{r}}\tilde{E}_t := \tilde{E}_{-t} : t \in \mathbb{R})$ is the same as the law of $(E_t : t \in \mathbb{R})$ (but using the reversed version, because intervals $[a, b)$ are transformed into intervals $(-b, -a]$). From reversibility (or direct computation), the generator of the reversed process, $Q^{\mathrm{r}\tilde{E}}$, is the same as $Q^{\tilde{E}}$.

Next let $(B_n : n \in \mathbb{Z})$ be a sequence of independent, identically distributed Bernoulli$(\frac{1}{2}, \frac{1}{2})$ random variables taking values in $\{1, -1\}$. The trajectories of the process (E_t) can be recovered from those of $(\tilde{E}_t)$ and (B_n), in the manner described below. First, $E_t = 2$ if $\tilde{E}_t = 2$ and $E_t = 0$ if $\tilde{E}_t = 0$. In the complementary region, we set $\{t \in \mathbb{R} : \tilde{E}_t = 1\} = \bigcup_{l\in\mathbb{Z}} J_l$, where $(J_l : l \in \mathbb{Z})$ is an ordered sequence of disjoint intervals. Then the remainder of the process is constructed as follows: $E_t = 1$ if $t \in J_l$ and $B_l = 1$ and $E_t = -1$ if $t \in J_l$ and $B_l = -1$. We notice that when (E_t) is time invariant, the marginals (E_t^i) and $(\tilde{E}_t)$ are also time invariant.

4. Simulation using a jump chain

4.1. Simulation of the process

In this subsection, we shall see how the process $(E_t : t \in \mathbb{R})$ can be simulated using a single discrete-time renewal process together with a collection of mutually independent random processes which are themselves sequences of independent random variables.

A point process $\boldsymbol{T} = (T_j : j \in \mathbb{Z})$ in $\mathbb{Z}$ is a (doubly) infinite and strictly increasing sequence, and we fix $T_1 = \inf\{T_i \ge 0 : i \in \mathbb{Z}\}$. Let $p = (p_n : n \ge 1)$ be a probability vector. The process $\boldsymbol{T}$ is a stationary renewal process with interarrival distribution p if $T_{j+1} - T_j$, $j \in \mathbb{Z}$, are independent, $T_{j+1} - T_j \sim p$ for $j \ne 0$, the mean $m(p) = \sum_{j\ge1} jp_j$ is finite, $\mathrm{P}\{T_1 - T_0 = n\} = np_n/m(p)$ for $n \ge 1$, and $\mathrm{P}\{T_1 = n\} = \sum_{j>n} p_j/m(p)$ for $n \ge 0$. This process is stationary, that is, $\boldsymbol{T} - a = (T_j - a : j \in \mathbb{Z}) \sim \boldsymbol{T}$ for all $a \in \mathbb{Z}$, where '$\sim$' denotes equality in distribution.

For $\theta \in (0, 1)$ we shall use Geom(θ) to denote a geometric distribution with parameter θ; e.g. $S \sim$ Geom(θ) if $\mathrm{P}\{S = n\} = (1 - \theta)\theta^{n-1}$ for $n \ge 1$. A stationary renewal process $\boldsymbol{T}$ with interarrival distribution Geom(θ) is called a geometric(θ) stationary renewal process. In

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

this case, $\mathrm{P}\{T_1 = n\} = (1-\theta)\theta^n$ for $n \geq 0$ and $\mathrm{P}\{T_0 = -n\} = (1-\theta)\theta^{n-1}$ for $n \geq 1$. Let $B = (B_n : n \in \mathbb{Z})$ be a Bernoulli$(\theta, 1-\theta)$ sequence taking values in $\{0, 1\}$, that is, such that $\mathrm{P}\{B_n = 0\} = \theta = 1 - \mathrm{P}\{B_n = 1\}$. Then $\boldsymbol{T} = (T_j : j \in \mathbb{Z}) = \{n \in \mathbb{Z} : B_n = 1\}$ is a geometric(θ) renewal process. Conversely, if $\boldsymbol{T}$ is a geometric(θ) stationary renewal process, then $(B_n : n \in \mathbb{Z})$ defined by $B_n = \mathbf{1}_{\{n \in \boldsymbol{T}\}}$ is a Bernoulli$(\theta, 1-\theta)$ sequence on $\{0, 1\}$. Thus, T_1 and T_0 are independent.

Definition 1. Let $\boldsymbol{T} = (T_j : j \in \mathbb{Z})$ be a point process on $\mathbb{Z}$. We define the following pair of point processes on $\mathbb{Z}$:

$$\hat{\boldsymbol{T}} = (\hat{T}_j : j \in \mathbb{Z}), \quad \text{where } \hat{T}_j = \begin{cases} T_j + 1 & \text{if } j > 0, \\ T_j & \text{if } j \leq 0, \end{cases}$$

$$\boldsymbol{T}^* = (T_j^* : j \in \mathbb{Z}), \quad \text{where } T_j^* = \begin{cases} T_j - T_1 & \text{if } j > 0, \\ T_{j-1} - T_0 & \text{if } j \leq 0. \end{cases}$$

Therefore, in the trajectories of $\hat{\boldsymbol{T}}$ we add an extra nonrenewal point (0) at time 0 and shift the trajectory over the positive integers one step to the right, and in the trajectories of $\boldsymbol{T}^*$ we delete the interval $[T_0, T_1)$, shifting the trajectory on the right of T_1 so as to start at the origin while shifting the trajectory on the left of T_0 to start at -1. In particular, note that $T_0^* = 1$.

Next let $\boldsymbol{T}$ be a geometric(θ) stationary renewal process, and let Ξ be a Bernoulli$(\alpha, 1-\alpha)$ random variable with values in $\{0, 1\}$, independent of $\boldsymbol{T}$. Then the point process $\boldsymbol{T}^\alpha = (T_j^\alpha : j \in \mathbb{Z})$ defined by $\boldsymbol{T}^\alpha = (1-\Xi)\hat{\boldsymbol{T}} + \Xi\boldsymbol{T}^*$ is characterized as follows: $\boldsymbol{T}^\alpha$ has independent increments, $T_{j+1}^\alpha - T_j^\alpha \sim \mathrm{Geom}(\theta)$ for $j \neq 0$, T_0^α is independent of T_1^α, $T_1^\alpha \sim (1-\alpha)\delta_0 + \alpha\,\mathrm{Geom}(\theta)$, and $-T_0^\alpha \sim \mathrm{Geom}(\theta)$. In fact, in this geometric case, it suffices to observe that $\boldsymbol{T}^\alpha = \{n \in \mathbb{Z} : B_n = 1\}$, where $(B_n : n \in \mathbb{Z})$ is a sequence of independent Bernoulli random variables with values in $\{0, 1\}$ such that $\mathrm{P}\{B_n = 1\} = 1 - \theta$ for $n \neq 0$ and $\mathrm{P}\{B_0 = 1\} = 1 - \alpha$. Observe that if $\alpha = \theta$ then $\boldsymbol{T}^\alpha \sim \boldsymbol{T}$ is a geometric(θ) stationary renewal process. This is the only case where $\boldsymbol{T}^\alpha$ is a stationary renewal process.

Proposition 3. *The process $E = (E^1, E^{-1})$ starting from a distribution*

$$\kappa = (\kappa_i : i \in \{0, 1, -1, 2\})$$

that is symmetric in $\{1, -1\}$ (that is, $\kappa_1 = \kappa_{-1}$) can be reconstructed from the following mutually independent random elements:

- *$\boldsymbol{T} = (T_j : j \in \mathbb{Z})$, a geometric$(\theta^*)$ stationary renewal process;*
- *$B = (B_j : j \in \mathbb{Z})$, a Bernoulli$(\frac{1}{2}, \frac{1}{2})$ sequence on $\{1, -1\}$;*
- *Γ, a Bernoulli$(2\kappa_1/(2\kappa_1 + \kappa_2), \kappa_2/(2\kappa_1 + \kappa_2))$ random variable on $\{1, 2\}$;*
- *Ξ, a Bernoulli$(\kappa_0, 1 - \kappa_0)$ random variable on $\{0, 1\}$; and*
- *$Z = (Z^j = (Z_n^j : n \in \mathbb{Z}) : j \in \{0, 1, 2\})$, a family of three independent random sequences exponentially distributed such that $Z_n^0 \sim \mathrm{Exp}(2\nu)$, $Z_n^1 \sim \mathrm{Exp}(\mu + \nu)$, and $Z_n^2 \sim \mathrm{Exp}(2\mu)$, $n \in \mathbb{Z}$.*

Remark 3. In the case where $\kappa_0 = 1 - \theta^*$, the random variable Ξ is not necessary in the construction.

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

Proof of Proposition 3. The set of jump times, $S = \{S_n : n \in \mathbb{Z}\}$, of the process (E_t) is given by $S_n := \inf\{t > S_{n-1} : E_t \neq E_{t-}\}$, $n \geq 1$, where we fix $S_1 = \inf_n\{S_n \geq 0\}$. The jump chain $(X_n := E_{S_n} : n \in \mathbb{Z})$ is a discrete Markov chain whose transition matrix, $P^X = (p^X_{ij} : i, j \in \{0, 1, -1, 2\})$, is

$$P^X = \begin{pmatrix} 0 & \dfrac{1}{2} & \dfrac{1}{2} & 0 \\ \dfrac{1}{1+\rho} & 0 & 0 & \dfrac{\rho}{1+\rho} \\ \dfrac{1}{1+\rho} & 0 & 0 & \dfrac{\rho}{1+\rho} \\ 0 & \dfrac{1}{2} & \dfrac{1}{2} & 0 \end{pmatrix}.$$

Its stationary distribution, π^X, is given by $\pi^X_0 = 1/(2(1+\rho))$, $\pi^X_1 = \frac{1}{4} = \pi^X_{-1}$, and $\pi^X_2 = \rho/(2(1+\rho))$. The chain (X_n) is endowed with symmetries analogous to those of (E_t). Thus, it is reversible in time: the laws of $({}^{\mathrm{r}}X_n : n \geq 0)$ and $(X_n : n \geq 0)$ are the same, where ${}^{\mathrm{r}}X_n := X_{-n}$, $n \in \mathbb{Z}$. Also, the lumping property holds; that is, we can exchange state 1 with state -1.

The jump times of the process $(\tilde{E}_t)$ are the S_n, the same as those of (E_t), and the jump chain $(\tilde{X}_n := \tilde{E}_{S_n} : n \geq 0)$ is a discrete Markov chain whose transition matrix, $P^{\tilde{X}} = (p^{\tilde{X}}_{ij} : i, j \in \{0, 1, 2\})$, is the lumped chain of P^X. Trajectories of (X_n) can be retrieved from those of $(\tilde{X}_n)$ and the Bernoulli sequence $B = (B_n)$ by setting $X_n = 2$ if $\tilde{X}_n = 2$, $X_n = 0$ if $\tilde{X}_n = 0$, $X_n = 1$ if $\tilde{X}_n = 1$ and $B_n = 1$, and $X_n = -1$ if $\tilde{X}_n = 1$ and $B_n = -1$. We note that the process $(\tilde{X}_n)$ is reversible in time in the same sense as (X_n). The Markov chain $X = (X_n : n \in \mathbb{Z})$ starts from $X_0 \sim E_0 \sim \kappa$ and the lumped chain, $\tilde{X} = (\tilde{X}_n := \tilde{E}_{S_n} : n \in \mathbb{Z})$, starts from $\tilde{X}_0 \sim \kappa'$, with $\kappa'_i = \kappa_i$ for $i = 0, 2$ and $\kappa'_1 = \kappa_1 + \kappa_{-1} = 2\kappa_1$.

Now let $Z^j = (Z^j_n : n \in \mathbb{Z})$, $j \in \{0, 1, 2\}$, be sequences as described in the statement of the proposition. Then Z^j_n is distributed in the same way as the sojourn time of the process $\tilde{E}$ in state $j \in \{0, 1, 2\}$. From the chain X and the family of sequences Z, we can retrieve E as follows. For $t \in [0, Z^{|X_0|}_0)$, we set $E_t = X_0$. For general nonnegative times, we set $E_t = X_n$ for $t \in [\sum_{k=0}^{n} Z^{|X_k|}_k, \sum_{k=0}^{n+1} Z^{|X_k|}_k)$. The construction is similar for negative times: for $t \in [-Z^{|X_0|}_{-1}, 0)$, we set $E_t = X_0$ and, for all other $t < 0$, we set $E_t = X_n$ for $t \in [\sum_{k=-n-1}^{-1} Z^{|X_{k+1}|}_k, \sum_{k=-n}^{-1} Z^{|X_{k+1}|}_k)$.

Hence, E can be reconstructed from $\tilde{X}$, B, and Z. It thus suffices to show that $\tilde{X}$ can be reconstructed using $\boldsymbol{T}$ and Ξ. To this end, we define an intermediate process, $\hat{X}$. Its trajectories are defined by deleting all the 0s from the trajectories of $\tilde{X}$, as follows. Fix $\hat{X}_0 = \tilde{X}_a$, where $a = 0$ if $\tilde{X}_0 \neq 0$ and $a = -1$ otherwise (in the latter case, $\hat{X}_0 = 1$ necessarily). Then copy $(\tilde{X}_j : j > a)$ to the right of 0 and $(\tilde{X}_j : j < a)$ to the left of 0 and delete all the 0s. This gives a well-defined process $\hat{X} = (\hat{X}_n : n \in \mathbb{Z})$. Note that $\hat{X}$ is a 2-memory homogeneous Markov chain, taking values in $\{1, 2\}$, such that the 2s are isolated, its sample paths contain no more than two consecutive 1s, and its transition probabilities are given by

$$\mathrm{P}\{\hat{X}_{n+1} = 2 \mid \hat{X}_n = 1,\ \hat{X}_{n-1} = 2\} = \theta^* = 1 - \mathrm{P}\{\hat{X}_{n+1} = 1 \mid \hat{X}_n = 1,\ \hat{X}_{n-1} = 2\},$$

$$\mathrm{P}\{\hat{X}_{n+1} = 2 \mid \hat{X}_n = 1,\ \hat{X}_{n-1} = 1\} = 1 = \mathrm{P}\{\hat{X}_{n+1} = 1 \mid \hat{X}_n = 2\}.$$

The starting distribution of $\hat{X}$ is given by

$$\mathrm{P}\{\hat{X}_0 = 1,\ \hat{X}_1 = 1\} = \kappa_0, \qquad \mathrm{P}\{\hat{X}_0 = 2,\ \hat{X}_1 = 1\} = \kappa_2, \qquad \mathrm{P}\{\hat{X}_0 = 1,\ \hat{X}_1 = 2\} = 2\kappa_1.$$

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

$\tilde{X}$ can be reconstructed from $\hat{X}$ by reversing the method used to construct $\hat{X}$. Indeed, it suffices to insert a 0 between any two consecutive 1s in each trajectory of $\hat{X}$, as follows. If $(\hat{X}_0, \hat{X}_1) = (1, 1)$ then we set $(\tilde{X}_{-1}, \tilde{X}_0, \tilde{X}_1) = (1, 0, 1)$ and if $(\hat{X}_0, \hat{X}_1) \neq (1, 1)$ then we set $(\tilde{X}_0, \tilde{X}_1) = (\hat{X}_0, \hat{X}_1)$. Then, moving to the right starting from $\hat{X}_1$ and moving to the left starting from $\hat{X}_0$, we insert a 0 between every pair of consecutive 1s. This procedure recovers $\tilde{X}$.

To finish, it suffices to show that $\hat{X}$ can be reconstructed from $\boldsymbol{T}$, Ξ, and Γ. Let $\alpha = 1 - \kappa_0$. First we use $\boldsymbol{T}$ and Ξ to construct $\boldsymbol{T}^\alpha$ and then we construct $\hat{X}$ from $\boldsymbol{T}^\alpha$ and Γ. If $\alpha = \theta^*$ then we do not need the random variable Ξ, because $\boldsymbol{T}^\alpha \sim \boldsymbol{T}$. We proceed as follows. If T_j^α and T_{j+1}^α are consecutive elements in $\boldsymbol{T}$, we create a 'j-block' $121 \cdots 21$ of length $L_j = 2(T_{j+1}^\alpha - T_j^\alpha) - 1$. Then we sequentially order these j-blocks according to j. The only thing left to consider is the starting point. If $T_1^\alpha = 0$ (that is, if $0 \in \boldsymbol{T}^\alpha$), then we position the 1-block in such a way that it finishes at time 0. For $T_1^\alpha > 0$, if $\Gamma = 2$ then we place the 1-block so as to finish at $2T_1^\alpha - 1$, and if $\Gamma = 1$ then we place the 1-block so as to finish at $2T_1^\alpha$. Therefore, $\mathrm{P}\{(\hat{X}_0, \hat{X}_1) = (1, 1)\} = 1 - \alpha = \kappa_0$ and

$$\frac{\mathrm{P}\{(\hat{X}_0, \hat{X}_1) = (1, 2)\}}{\mathrm{P}\{(\hat{X}_0, \hat{X}_1) = (2, 1)\}} = \frac{2\kappa_1}{\kappa_2}.$$

This completes the proof.

We have noted that the unique initial value, $\kappa_0 = 1 - \alpha$, for which $\boldsymbol{T}^\alpha$ is stationary is $\kappa_0 = \pi_0 = 1 - \theta^*$, the value of the stationary distribution of E^i in state 0, which differs from Π_0, the value of the stationary distribution of E in state 0. In the sequel, we assume that $\kappa_0 = \pi_0$; therefore, $\boldsymbol{T}^\alpha$ can be supposed to be equal to $\boldsymbol{T}$, a geometric(θ^*) stationary renewal process.

In the next section, we shall use the ideas introduced in Proposition 3 to stochastically simulate an island and to calculate a lower bound based on renewal arguments.

4.2. Simulation of the islands and genes

Recall that islands are the connected zones where $\tilde{E}_t \geq 1$, that is, where there is a gene candidate present on one or both strands. For each $I \in \mathcal{I}$, we shall use $\boldsymbol{K}(I)$ to denote the island K containing I. Let L be the length of some arbitrary island K. From the discussion in Sections 3.1 and 4, it should be clear that L can be decomposed as follows:

$$L = \sum_{i=1}^{S} Z_i^1 + \sum_{i=1}^{S-1} Z_i^2,$$

where S, $(Z_i^1 : i \geq 1)$, and $(Z_i^2 : i \geq 1)$ are independent random variables. Here Z_i^1 and Z_i^2 respectively represent the times spent in states 1 and 2 by $(\tilde{E}_t)$ during a visit. Thus,

$$Z_i^1 \sim \mathrm{Exp}(\mu + \nu), \qquad Z_i^2 \sim \mathrm{Exp}(2\mu).$$

By considering $P^{\tilde{E}}$, the generator of $(\tilde{X}_t)$, we can see that $S \sim \mathrm{Geom}(\theta^*)$ with $\theta^* = \rho/(1+\rho)$, so $\mathrm{P}\{S = k\} = (1 - \theta^*)\theta^{*k-1}$ for $k = 1, 2, \dots$.

The moment generating function of L is $m_L(t) = m_1(t) g_S(m_2(t) m_1(t))$, where $m_1(t) = (\mu + \nu)/(\mu + \nu - t)$ and $m_2(t) = 2\mu/(2\mu - t)$ are the moment generating functions of the Z_i^1s and the Z_i^2s, respectively, and $g_S(t) = 1/(1 + \rho - \rho t)$ is the probability generating function of S. Hence,

$$m_L(t) = \frac{2\mu^2 - \mu t}{2\mu^2 - 3\mu t - \nu t + t^2}$$

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

and from this we may calculate the mean and variance of L:

$$\mathrm{E}(L) = m'(0) = \frac{1}{\mu}\left(1 + \frac{\rho}{2}\right), \qquad \operatorname{var}(L) = m''(0) - m'(0)^2 = \mathrm{E}(L)^2 + \frac{\rho}{2\mu^2}.$$

Next, let R be a random variable independent of S, $(Z_i^1 : i \geq 1)$, and $(Z_i^2 : i \geq 1)$ and satisfying $R \sim \operatorname{Geom}(\frac{1}{2})$, whence $\mathrm{P}\{R = k\} = (\frac{1}{2})^k$ for $k \geq 1$. We shall write $S \wedge R := \min\{S, R\}$. It can be shown that $S \wedge R \sim \operatorname{Geom}(\theta^*/2)$. We have the following elementary result.

Lemma 4. (i) *The event* $\{S \wedge R = S\}$ *is independent of* S.

(ii) *The random variable* $S - S \wedge R$ *is independent of* $S \wedge R$, *and* $S - S \wedge R$ *conditioned on the event* $\{S > S \wedge R\}$ *is* $\operatorname{Geom}(\theta^*)$*-distributed.*

(iii) *The random variable* $S \wedge R$ *is independent of the event* $\{S > S \wedge R\}$.

Observe that $\xi := \mathrm{P}\{S \wedge R < S\}$ can be explicitly evaluated:

$$\xi = \frac{\theta^*}{2 - \theta^*} = \frac{\rho}{2 + \rho}. \tag{5}$$

Define $C = \mathbf{1}_{\{S > S \wedge R\}}$. This is independent of $S \wedge R$ and distributed as a Bernoulli(ξ) random variable ($C \sim \operatorname{Ber}(\xi)$), that is, $\mathrm{P}\{C = 1\} = \xi = 1 - \mathrm{P}\{C = 0\}$.

It is easy to verify the following decomposition result.

Lemma 5. *The random variable* $Y = \sum_{i=1}^{S \wedge R} Z_i^1 + \sum_{i=1}^{(S \wedge R) - 1} Z_i^2 + C Z_{S \wedge R}^2$ *satisfies* $Y \sim \operatorname{Exp}(\mu)$.

5. Bounds on the occupancy rate

5.1. Bounds

As obtaining an analytic expression for the quantity $M^*(\rho)$ has proven to be exceedingly difficult, we shall discuss upper and lower bounds on it in this section. However, before doing so, we shall digress for a moment to establish a relationship between the (general) occupancy rate and the occupancy rate of an island. Let us suppose that $E_0 = 0$, that is, both DNA strands start with a gap. We denote the islands starting after the point 0 by $K(i)$, $i \geq 0$.

Let $\mathcal{L}$ be the occupancy of an arbitrary island K after having applied the maximization algorithm. That is,

$$\mathcal{L} \sim \int_{s(K)}^{f(K)} \mathbf{1}_{\{E_s^* = 1\}} \, \mathrm{d}s.$$

Now, the mean area occupied by the final genes in an island K is

$$\mathcal{L}^*(\rho) = \mathrm{E}\left(\int_{s(K)}^{f(K)} \mathbf{1}_{\{E_s^* = 1\}} \, \mathrm{d}s\right). \tag{6}$$

Notice that since $\int_{s(K(i))}^{f(K(i))} \mathbf{1}_{\{E_s^* = 1\}} \, \mathrm{d}s$, $i \geq 1$, are independent and identically distributed random variables, from the law of large numbers we obtain

$$\mathcal{L}^*(\rho) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n} \int_{s(K(i))}^{f(K(i))} \mathbf{1}_{\{E_s^* = 1\}} \, \mathrm{d}s \qquad \mathrm{P}\text{-a.e.} \tag{7}$$

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

Proposition 4. *The following relation holds:*

$$M^*(\rho) = \frac{2\mathcal{L}^*(\rho)\mu\rho}{(1+\rho)^2}.$$

Proof. We start from $E_0 = 0$. For each $t > 0$, let us define two random variables U_t and V_t, where

$$U_t = \begin{cases} \sup\{0 \le s \le t \colon E_s \ne 0\}, & E_t = 0, \\ \sup\{0 \le s \le t \colon E_s = 0\}, & E_t \ne 0, \end{cases} \qquad V_t = \begin{cases} \inf\{s \ge t \colon E_s \ne 0\}, & E_t = 0, \\ \inf\{s \ge t \colon E_s = 0\}, & E_t \ne 0. \end{cases}$$

Since $\lim_{t\to\infty} U_t/t = 1 = \lim_{t\to\infty} V_t/t$, we find that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^{U_t} \mathbf{1}_{\{E_s^*=1\}} \,\mathrm{d}s) = M^*(\rho) = \lim_{t\to\infty} \frac{1}{t} \int_0^{V_t} \mathbf{1}_{\{E_s^*=1\}} \,\mathrm{d}s \quad \mathrm{P}\text{-a.e.} \tag{8}$$

For each $t > 0$, let D_t denote the number of islands that are completely contained inside $[0, t)$. From (8), we have

$$\begin{aligned} M^*(\rho) &= \lim_{t\to\infty} \frac{1}{t} \sum_{i=0}^{D_t-1} \int_{s(K(i))}^{f(K(i))} \mathbf{1}_{\{E_s^*=1\}} \,\mathrm{d}s \\ &= \lim_{t\to\infty} \frac{D_t}{t} \frac{1}{D_t} \sum_{i=0}^{D_t-1} \int_{s(K(i))}^{f(K(i))} \mathbf{1}_{\{E_s^*=1\}} \,\mathrm{d}s \quad \mathrm{P}\text{-a.e.} \end{aligned}$$

As $\lim_{t\to\infty} D_t = \infty$ P-a.e., we obtain

$$M^*(\rho) = \lim_{t\to\infty} \frac{D_t}{t} \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \int_{s(K(i))}^{f(K(i))} \mathbf{1}_{\{E_s^*=1\}} \,\mathrm{d}s \quad \mathrm{P}\text{-a.e.}$$

From (7), we obtain

$$\mathcal{L}^*(\rho) = M^*(\rho) \lim_{t\to\infty} \frac{t}{D_t}. \tag{9}$$

It remains to compute $\lim_{t\to\infty} t/D_t$. The quantity (D_t) is an alternating renewal process which counts the number of islands entirely enclosed in $[0, t]$, where each renewal is made according to the law of $R = R_1 + R_2$, the sum of two independent random variables: R_1 is the sojourn time at 0 and R_2 is the sojourn time at an island. From renewal theory, we have

$$\lim_{t\to\infty} \frac{t}{D_t} = \mathrm{E}(R_1) + \mathrm{E}(R_2).$$

Now, $R_1 \sim \operatorname{Exp}(2\nu)$, so $\mathrm{E}(R_1) = 1/(2\nu)$. Furthermore, from above we have $\mathrm{E}(R_2) = (1/\mu)(1 + \rho/2)$. Thus,

$$\lim_{t\to\infty} \frac{t}{D_t} = \frac{1}{2\nu} + \frac{1}{\mu}\left(1 + \frac{\rho}{2}\right) = \frac{1}{2\mu\rho}(1+\rho)^2. \tag{10}$$

Finally, substituting (10) into (9) yields the desired result.

Next we shall investigate bounds on $M^*(\rho)$. We begin by providing a simple upper bound.

Proposition 5. *We have the upper bound*

$$M^*(\rho) \le \tilde{M}(\rho) := \frac{\rho(2+\rho)}{(1+\rho)^2}.$$

Proof. Define $\tilde{W}_t = \int_0^t \mathbf{1}_{\{\tilde{E}_s>0\}}\,\mathrm{d}s$. Clearly $\tilde{W}_t \ge W_t$, so $\tilde{M}(\rho) \ge M^*(\rho)$. By ergodicity,

$$\tilde{M}(\rho) := \lim_{t\to\infty} \frac{1}{t}\tilde{W}_t = \int_0^t \mathrm{E}(\mathbf{1}_{\{\tilde{E}_s>0\}})\,\mathrm{d}s = \Pi_1^{\tilde{E}} + \Pi_2^{\tilde{E}} = \frac{\rho(2+\rho)}{(1+\rho)^2} \quad \mathrm{P}\text{-a.e.}$$

The question of finding a good lower bound is more difficult. Trivially, we have

$$M^*(\rho) \ge \max\left\{\frac{\rho}{1+\rho}, \frac{2\rho}{(1+\rho)^2}\right\},$$

because $\theta^* = \rho/(1+\rho)$ is the equilibrium proportion of space occupied by gene candidates on a single DNA strand and $\Pi_1^E + \Pi_{-1}^E = 2\rho/(1+\rho)^2$ is the equilibrium proportion of DNA where only one of the strands is occupied by gene candidates. Note that $\theta^* \ge \Pi_1^E + \Pi_{-1}^E$ if and only if $\rho \ge 1$.

Let us define

$$\hat{M}(\rho) = \frac{32\rho + 80\rho^2 + 84\rho^3 + 42\rho^4 + 8\rho^5}{16 + 72\rho + 130\rho^2 + 118\rho^3 + 54\rho^4 + 10\rho^5}. \tag{11}$$

Note that $\hat{M}(\rho) \ge 2\rho/(1+\rho)^2$ for every $\rho \in [0,\infty)$.

Proposition 6. *We have the lower bound*

$$M^*(\rho) \ge \max\left\{\frac{\rho}{1+\rho}, \hat{M}(\rho)\right\}.$$

Proof. Since $M^*(\rho) \ge \rho/(1+\rho)$, it suffices to show that $M^*(\rho) \ge \hat{M}(\rho)$.

Let Y be the length of the first gene candidate, I, of a certain island. Let us consider the representation given by Lemma 5. If $C = 0$ then this gene candidate terminates at the end of the island. On the other hand, if $C = 1$ then, at the point where the gene candidate I ends, there will already be another gene candidate, I', present on the opposite strand; we shall denote its length by Y'. By V and V' we denote the sum of the lengths of gene candidates in the same strand as I that are strictly contained in the region covered by gene candidate I' and, respectively, the sum of the lengths of the gene candidates in the same strand as I' that are strictly contained in the region demarcated by gene candidate I.

Let S, S', S'', R, R', and R'' be mutually independent random variables also independent of the families of independent random variables $(Z_i^1 : i \ge 1)$, $(Z_i^2 : i \ge 1)$, $({Z'}_i^1 : i \ge 1)$, and $({Z'}_i^2 : i \ge 1)$. Here $S \sim S' \sim S'' \sim \mathrm{Geom}(\theta^*)$, $R \sim R' \sim R'' \sim \mathrm{Geom}(\frac{1}{2})$, $Z_i^1 \sim {Z'}_i^1 \sim \mathrm{Exp}(\mu+\nu)$, and $Z_i^2 \sim {Z'}_i^2 \sim \mathrm{Exp}(2\mu)$. Define $C = \mathbf{1}_{\{S>S\wedge R\}}$, $C' = \mathbf{1}_{\{S'>S'\wedge R'\}}$, and $C'' = \mathbf{1}_{\{S''>S''\wedge R''\}}$, which are independent $\mathrm{Ber}(\xi)$-distributed random variables.

Let $N = S \wedge R$ and $N' = S' \wedge R'$. From Lemma 4, $N \sim N' \sim \mathrm{Geom}(\theta^*/2)$. From Lemma 5, we have

$$Y = \sum_{i=1}^{N} Z_i^1 + V', \qquad \text{where } V' = \sum_{i=1}^{N-1} Z_i^2 + C Z_N^2,$$

$$Y' = \sum_{i=1}^{N'} {Z'}_i^1 C + V, \qquad \text{where } V = \left(\sum_{i=1}^{N'-1} {Z'}_i^2 + C' {Z'}_{N'}^2\right) C.$$

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

Hence, on $\{C = 1\}$,

$$\begin{aligned} Y + V > Y' + V' &\iff \sum_{i=1}^{N} Z_i^1 > \sum_{i=1}^{N'} {Z'}_i^1, \\ Y' + V' > Y + V &\iff \sum_{i=1}^{N'} {Z'}_i^1 > \sum_{i=1}^{N} Z_i^1. \end{aligned}$$

From this relation and from the independence between (N, N') and (C, C', C''), $(Z_i^1 : i \geq 1)$, $(Z_i^2 : i \geq 1)$, $({Z'}_i^1 : i \geq 1)$, and $({Z'}_i^2 : i \geq 1)$, we obtain

$$\begin{aligned} \mathrm{P}\{\{Y + V > Y' + V'\} \cap \{C = 1,\ C' = 1\}\} &= \frac{\xi^2}{2}, \\ \mathrm{P}\{\{Y' + V' > Y + V\} \cap \{C = 1,\ C' = 1,\ C'' = 1\}\} &= \frac{\xi^3}{2}. \end{aligned} \tag{12}$$

On the other hand, we also find that

$$\begin{aligned} &(Y + V)\,\mathbf{1}_{\{Y+V>Y'+V'\}}\,\mathbf{1}_{\{C=1\}} \\ &\quad = \Bigg(\sum_{i=1}^{N} Z_i^1 + \sum_{i=1}^{N} Z_i^2 + \sum_{i=1}^{N'-1} {Z'}_i^2 + B'{Z'}_{N'}^2\Bigg)\,\mathbf{1}_{\{\sum_{i=1}^{N} Z_i^1 > \sum_{i=1}^{N'} {Z'}_i^1\}}\,\mathbf{1}_{\{C=1\}}, \end{aligned} \tag{13}$$

$$\begin{aligned} &(Y' + V')\,\mathbf{1}_{\{Y'+V'>Y+V\}}\,\mathbf{1}_{\{C=1\}} \\ &\quad = \Bigg(\sum_{i=1}^{N'} {Z'}_i^1 + \sum_{i=1}^{N'-1} {Z'}_i^2 + B'{Z'}_{N'}^2 + \sum_{i=1}^{N} Z_i^2\Bigg)\,\mathbf{1}_{\{\sum_{i=1}^{N} Z_i^1 > \sum_{i=1}^{N'} {Z'}_i^1\}}\,\mathbf{1}_{\{C=1\}}\,. \end{aligned} \tag{14}$$

Hence,

$$(Y + V)\,\mathbf{1}_{\{Y+V>Y'+V'\}}\,\mathbf{1}_{\{C=1\}} \sim (Y' + V')\,\mathbf{1}_{\{Y'+V'>Y+V\}}\,\mathbf{1}_{\{C=1\}}\,. \tag{15}$$

Recall that $\mathcal{L}$ was defined to be the occupancy of an arbitrary island after having applied the maximization algorithm. Let $\mathcal{L}'$ and $\mathcal{L}''$ be two independent copies of $\mathcal{L}$ also independent of S, S', S'', R, R', R'', $(Z_i^1 : i \geq 1)$, $(Z_i^2 : i \geq 1)$, $({Z'}_i^1 : i \geq 1)$, and $({Z'}_i^2 : i \geq 1)$. The renewal equation gives us the following stochastic domination relation (where '$\geq_{\mathrm{st}}$' means 'stochastically larger than'):

$$\begin{aligned} \mathcal{L} \geq_{\mathrm{st}}\ & Y\,\mathbf{1}_{\{C=0\}} \\ &+ ((Y + V)\,\mathbf{1}_{\{Y+V>Y'+V'\}} + (Y' + V')\,\mathbf{1}_{\{Y'+V'>Y+V\}})\,\mathbf{1}_{\{C=1,\,C'=0\}} \\ &+ (Y + V + \mathcal{L}')\,\mathbf{1}_{\{Y+V>Y'+V'\}}\,\mathbf{1}_{\{C=1,\,C'=1\}} \\ &+ (Y' + V')\,\mathbf{1}_{\{Y'+V'>Y+V\}}\,\mathbf{1}_{\{C=1,\,C'=1,\,C''=0\}} \\ &+ (Y' + V' + \mathcal{L}'')\,\mathbf{1}_{\{Y'+V'>Y+V\}}\,\mathbf{1}_{\{C=1,\,C'=1,\,C''=1\}}\,. \end{aligned}$$

Therefore, by using (12) and following the definition of $\mathcal{L}^*(\rho)$ in (6), we obtain

$$\begin{aligned} \mathcal{L}^*(\rho)(1 - \tfrac{1}{2}\xi^2(1 + \xi)) \geq\ & \mathrm{E}(Y\,\mathbf{1}_{\{C=0\}}) \\ &+ \mathrm{E}([(Y + V)\,\mathbf{1}_{\{Y+V>Y'+V'\}} + (Y + V)\,\mathbf{1}_{\{Y'+V'>Y+V\}}]\,\mathbf{1}_{\{C=1\}}). \end{aligned}$$

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

Using (15), we conclude that

$$\mathcal{L}^*(\rho)(1-\tfrac{1}{2}\xi^2(1+\xi)) \geq \mathrm{E}(Y\,\mathbf{1}_{\{C=0\}}) + 2\,\mathrm{E}((Y+V)\,\mathbf{1}_{\{Y+V>Y'+V'\}}\,\mathbf{1}_{\{C=1\}}). \tag{16}$$

Let us compute the first term on the right-hand side. We have

$$\begin{aligned}\mathrm{E}(Y\,\mathbf{1}_{\{C=0\}}) &= \mathrm{E}\biggl(\biggl(\sum_{i=1}^{N} Z_i^1 + \sum_{i=1}^{N-1} Z_i^2 + BZ_N^2\biggr)\mathbf{1}_{\{C=0\}}\biggr)\\ &= (1-\xi)(\mathrm{E}(Z_1^1)\,\mathrm{E}(N) + \mathrm{E}(Z_1^2)\,\mathrm{E}(N-1)).\end{aligned}$$

From the relations $\mathrm{E}(N) = \sum_{k\geq 1}\mathrm{P}\{S\wedge R \geq k\} = 2/(2-\theta^*)$, $\theta^* = \rho/(1+\rho)$, and $\xi = \rho/(2+\rho)$, we obtain

$$\mathrm{E}(Y\,\mathbf{1}_{\{C=0\}}) = (1-\xi)\biggl(\frac{2}{(2-\theta^*)(\mu+\nu)} + \frac{\theta^*}{2\mu(2-\theta^*)}\biggr) = \frac{4+\rho}{\mu(2+\rho)^2}. \tag{17}$$

Note that

$$\mathrm{E}(Y\,\mathbf{1}_{\{C=0\}}) > \mathrm{E}(Z_1^1\,\mathbf{1}_{\{Z_1^2>Z_1^1\}}) = p_{1,0}^X\,\mathrm{E}(Z_1^1) = \frac{1}{\mu(1+\rho)^2},$$

the right-most expression being the length of a gene candidate which finishes before any gene candidate appears in the opposite strand.

To compute the second term on the right-hand side of (16), we need the following well-known relation, in which $a \geq 0$ and $\lambda > 0$ are parameters:

$$\begin{aligned}\int_{\sum_{i=1}^n x_i>a,\,x_i\geq 0,\,i=1,\dots,n}\biggl(\sum_{i=1}^n x_i\biggr)^r \lambda^n \exp\biggl(-\lambda\sum_{i=1}^n x_i\biggr)\mathrm{d}x_1\cdots\mathrm{d}x_n&\\ = \frac{(n+r-1)!}{(n-1)!}\lambda^{-r}\sum_{l=0}^{n+r-1}\frac{\lambda^l a^l}{l!}\mathrm{e}^{-\lambda a}&.\end{aligned} \tag{18}$$

For $a=0$, this reduces to

$$\int_{x_i\geq 0,\,i=1,\dots,n}\biggl(\sum_{i=1}^n x_i\biggr)^r \lambda^n \exp\biggl(-\lambda\sum_{i=1}^n x_i\biggr)\mathrm{d}x_1\cdots\mathrm{d}x_n = \frac{(n+r-1)!}{(n-1)!}\lambda^{-r}. \tag{19}$$

Now we can compute the second term on the right-hand side of (16). In (13) we sum over the variables $N=n$, $N'=m$, $Z_i^1 \in (x_i, x_i+\mathrm{d}x_i]$, $Z_i^2 \in (y_i, y_i+\mathrm{d}y_i]$, $Z'^1_i \in (x'_i, x'_i+\mathrm{d}x'_i]$, and $Z'^2_i \in (y'_i, y'_i+\mathrm{d}y'_i]$. Thus, we obtain the following expression, where

$$\begin{aligned}L_{n,m} = \int (\mu+\nu)^m \exp\biggl(-(\mu+\nu)\sum_{i=1}^m x'_i\biggr)&\\ \times\biggl(\xi A\biggl(\sum_{i=1}^m x'_i, m, n\biggr) + (1-\xi)A\biggl(\sum_{i=1}^m x'_i, m-1, n\biggr)\biggr)\,\mathrm{d}x'_1\cdots\mathrm{d}x'_m&\end{aligned}$$

and the variables x_i, x'_i, y_i, and y'_i always take positive values:

$$\mathrm{E}((Y+V)\,\mathbf{1}_{\{Y+V>Y'+V'\}}\,\mathbf{1}_{\{C=1\}}) = \xi\biggl(1-\frac{\theta^*}{2}\biggr)^2\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\biggl(\frac{\theta^*}{2}\biggr)^{n+m-2}L_{n,m}.$$

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

In $L_{n,m}$,

$$A(a,l,n)=\int_{\sum_{i=1}^{n}x_i>a}(\mu+\nu)^n\exp\biggl(-(\mu+\nu)\sum_{i=1}^{n}x_i\biggr)G\biggl(\sum_{i=1}^{n}x_i,l,n\biggr)\,\mathrm{d}x_1\cdots\mathrm{d}x_n,$$

and here

$$\begin{aligned}G\biggl(\sum_{i=1}^{n}x_i,l,n\biggr)&=\int(2\mu)^{n+l}\exp\biggl(-2\mu\biggl(\sum_{i=1}^{l}y_i'+\sum_{i=1}^{n}y_i\biggr)\biggr)\\&\qquad\times\biggl(\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}y_i+\sum_{i=1}^{l}y_i'\biggr)\,\mathrm{d}y_1'\cdots\mathrm{d}y_l'\,\mathrm{d}y_1\cdots\mathrm{d}y_n\\&=\sum_{i=1}^{n}x_i+\frac{n+l}{2\mu}.\end{aligned}$$

Hence,

$$\begin{aligned}A\biggl(\sum_{i=1}^{m}x_i',l,n\biggr)&=\int_{\sum_{i=1}^{n}x_i>\sum_{i=1}^{m}x_i'}\biggl(\sum_{i=1}^{n}x_i\biggr)(\mu+\nu)^n\exp\biggl(-(\mu+\nu)\biggl(\sum_{i=1}^{n}x_i\biggr)\biggr)\,\mathrm{d}x_1\cdots\mathrm{d}x_n\\&\quad+\frac{n+l}{2\mu}\int_{\sum_{i=1}^{n}x_i>\sum_{i=1}^{m}x_i'}(\mu+\nu)^n\exp\biggl(-(\mu+\nu)\biggl(\sum_{i=1}^{n}x_i\biggr)\biggr)\,\mathrm{d}x_1\cdots\mathrm{d}x_n.\end{aligned}$$

From (18), we find that

$$\begin{aligned}A\biggl(\sum_{i=1}^{m}x_i',l,n\biggr)&=\biggl(\frac{n}{\mu+\nu}\sum_{r=0}^{n}\frac{(\sum_{i=1}^{m}x_i')^r(\mu+\nu)^r}{r!}+\frac{n+l}{2\mu}\sum_{r=0}^{n-1}\frac{(\sum_{i=1}^{m}x_i')^r(\mu+\nu)^r}{r!}\biggr)\\&\quad\times\exp\biggl(-(\mu+\nu)\sum_{i=1}^{m}x_i'\biggr).\end{aligned}$$

From the above expression, and by using (19), we obtain $L_{n,m}=C_{n,m}+D_{n,m}+E_{n,m}$, where

$$C_{n,m}=\frac{n2^{-m}}{\mu+\nu}\sum_{r=0}^{n}\frac{(m+r-1)!}{(m-1)!\,r!}2^{-r},\qquad D_{n,m}=\frac{n2^{-m}}{2\mu}\sum_{r=0}^{n-1}\frac{(m+r-1)!}{(m-1)!\,r!}2^{-r},$$

$$E_{n,m}=\frac{(m-1+\xi)2^{-m}}{2\mu}\sum_{r=0}^{n-1}\frac{(m+r-1)!}{(m-1)!\,r!}2^{-r}.$$

By using

$$\begin{aligned}\sum_{n=r}^{\infty}n\biggl(\frac{\theta^*}{2}\biggr)^{n-1}&=\frac{2r}{2-\theta^*}\biggl(\frac{\theta^*}{2}\biggr)^{r-1}+\frac{4}{(2-\theta^*)^2}\biggl(\frac{\theta^*}{2}\biggr)^{r},\\\sum_{n=r+1}^{\infty}\biggl(\frac{\theta^*}{2}\biggr)^{n-1}&=\frac{2}{2-\theta^*}\biggl(\frac{\theta^*}{2}\biggr)^{r},\end{aligned}$$

we obtain

$$
\begin{aligned}
&\mathrm{E}((Y+V)\,\mathbf{1}_{\{Y+V>Y'+V'\}}\,\mathbf{1}_{\{C=1\}})\\
&\quad=\frac{\xi(2-\theta^*)^2}{8(\mu+\nu)}\sum_{m=1}^{\infty}\left(\frac{\theta^*}{4}\right)^{m-1}\left(\frac{J(m,\theta^*)}{2-\theta^*}+\frac{4K(m,\theta^*)}{(2-\theta^*)^2}\right)\\
&\qquad+\frac{\xi(2-\theta^*)^2}{16\mu}\sum_{m=1}^{\infty}\left(\frac{\theta^*}{4}\right)^{m-1}\left(\frac{\theta^*J(m,\theta^*)}{2(2-\theta^*)}+\frac{2K(m,\theta^*)}{2-\theta^*}+\frac{2\theta^*K(m,\theta^*)}{(2-\theta^*)^2}\right)\\
&\qquad+\frac{\xi(2-\theta^*)^2}{16\mu}\sum_{m=1}^{\infty}(m-1+\xi)\left(\frac{\theta^*}{4}\right)^{m-1}\frac{2K(m,\theta^*)}{2-\theta^*},
\end{aligned}
$$

where

$$
J(m,\theta^*)=\sum_{r=0}^{\infty}\binom{m+r-1}{r}r\left(\frac{\theta^*}{4}\right)^{r-1},\qquad K(m,\theta^*)=\sum_{r=0}^{\infty}\binom{m+r-1}{r}\left(\frac{\theta^*}{4}\right)^{r}.
$$

Now, it is known that, for every $m\geq 1$,

$$
\sum_{r=0}^{\infty}\binom{m+r-1}{r}\left(1-\frac{\theta^*}{4}\right)^{m}\left(\frac{\theta^*}{4}\right)^{r}=1,
$$

because it is the total mass of a Pascal distribution with parameter pair $(m, 1-\theta^*/4)$. Thus, $K(m,\theta)=(1-\theta^*/4)^{-m}$. On the other hand,

$$
\begin{aligned}
\sum_{r=0}^{\infty}\binom{m+r-1}{r}r\left(\frac{\theta^*}{4}\right)^{r-1}&=\sum_{r=1}^{\infty}\frac{(m+r-1)!}{(r-1)!\,m!}m+\left(\frac{\theta^*}{4}\right)^{r-1}\\
&=m\sum_{r=0}^{\infty}\binom{m+1+r-1}{r}\left(\frac{\theta^*}{4}\right)^{r},
\end{aligned}
$$

so $J(m,\theta^*)=mK(m+1,\theta^*)=m(1-\theta^*/4)^{-(m+1)}$. Putting all these elements together, we obtain

$$
\begin{aligned}
\mathrm{E}((Y+V)\,\mathbf{1}_{\{Y+V>Y'+V'\}}\,\mathbf{1}_{\{C=1\}})={}&\frac{\xi}{2(\mu+\nu)(2-\theta^*)}+\frac{\xi}{(\mu+\nu)(2-\theta^*)}+\frac{\xi\theta^*}{8\mu(2-\theta^*)}\\
&+\frac{\xi\theta^*}{4\mu}+\frac{\xi\theta^*}{4\mu(2-\theta^*)}+\frac{\xi(4-\theta^*)}{8\mu(2-\theta^*)}-\frac{\xi(1-\xi)}{4\mu}.
\end{aligned}
$$

Applying (5), we obtain

$$
\mathrm{E}((Y+V)\,\mathbf{1}_{\{Y+V>Y'+V'\}}\,\mathbf{1}_{\{C=1\}})=\frac{1}{\mu(2+\rho)^2}\left(\frac{3\rho}{2}+\frac{3\rho^2}{8}+\frac{\rho^2(2+\rho)}{4(1+\rho)}+\frac{\rho(4+3\rho)}{8}-\frac{\rho}{2}\right).
$$

Therefore, from (16) and (17), and since

$$
(1-\tfrac{1}{2}\xi^2(1+\xi))^{-1}=\frac{(2+\rho)^3}{8+12\rho+5\rho^2},
$$

we obtain

$$\mathcal{L}^*(\rho) \geq \frac{2+\rho}{\mu(8+12\rho+5\rho^2)}\left(4+3\rho+\frac{3\rho^2}{4}+\frac{\rho^2(2+\rho)}{2(1+\rho)}+\frac{\rho(4+3\rho)}{4}\right).$$

On the other hand, we know from (6) that $M^*(\rho) = \mathcal{L}^*(\rho)2\mu\rho/(1+\rho)^2$. Hence,

$$\begin{aligned} M^*(\rho) & \\ &\geq \frac{\rho(2+\rho)(16(1+\rho)+12\rho(1+\rho)+3\rho^2(1+\rho)+2\rho^2(2+\rho)+\rho(4+3\rho)(1+\rho))}{2(1+\rho)^3(8+12\rho+5\rho^2)} \\ &= \hat{M}(\rho). \end{aligned}$$

This completes the proof.

This lower bound has an asymptote at $\frac{4}{5}$ as $\rho \to \infty$, and $\hat{M}(1) = 0.615$. It can be checked that there exists a $\bar{\rho} \in (2, 3)$ such that

$$\hat{M}(\rho)\begin{cases} \geq \dfrac{\rho}{1+\rho} & \text{if } \rho \leq \bar{\rho}, \\ \leq \dfrac{\rho}{1+\rho} & \text{if } \rho \geq \bar{\rho}. \end{cases}$$

5.2. Discussion concerning loss networks

We note that the bound $\hat{M}(\rho)$ in (11) satisfies

$$\hat{M}(\rho) \leq \underline{M}(\rho) := \frac{2\rho}{1+2\rho}.$$

The quantity on the right-hand side comes from a loss network model which we briefly discuss.

Consider the process $(\underline{E}_t : t \in \mathbb{R})$ with state space $\{0, 1, -1\}$ and generator $Q^{\underline{E}}$ given by

$$Q^{\underline{E}} = \begin{pmatrix} -2\nu & \nu & \nu \\ \mu & -\mu & 0 \\ \mu & 0 & -\mu \end{pmatrix}.$$

The stationary distribution of $(\underline{E}_t)$ is given by $\pi_0^{\underline{E}} = 1/(1+2\rho)$ and $\pi_1^{\underline{E}} = \rho/(1+2\rho) = \pi_{-1}^{\underline{E}}$. Hence, $\pi_1^{\underline{E}} + \pi_{-1}^{\underline{E}} = 2\rho/(1+2\rho)$ is the proportion of time that $\underline{E}$ spends away from state 0. The process $(\underline{E}_t)$ may be viewed as a loss network in which there are two routes (1 and -1) sharing a single link of unit capacity. See [2] for more information about loss networks. Calls (that is, gene candidates) 'arrive' according to the Poisson streams $N^{1,i}$, $i \in \{1, -1\}$, and are accepted or rejected according to whether the link is free (no gene candidates currently present on either strand) or busy (a gene candidate is present on one of the strands).

It might be thought that the process $(\underline{E}_t)$ is equivalent to scanning through each island K from $s(K)$ to $f(K)$ and culling those gene candidates which overlap any previously accepted gene candidates. However, this is not the case. In the loss network setting, the length of a call (gene candidate) remains exponential with mean $1/\mu$, but in our setting the time between calls is not exponentially distributed. Since there may or may not be another gene candidate present on the opposite strand at the point where the most recently accepted gene candidate terminates, the time (gap) between calls has the same law as $BY + (1-B)(X+Y)$, where $B \sim \text{Ber}(\xi)$ (a Bernoulli random variable with parameter $\xi = \rho/(2+\rho)$), $Y \sim \text{Exp}(\nu)$, and $X \sim \text{Exp}(\mu)$ are independent random variables. Moreover, as simulations show, the quantity $\underline{M}(\rho)$ is neither a lower nor an upper bound, but does seem to closely approximate the empirical estimate of $M^*(\rho)$.

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

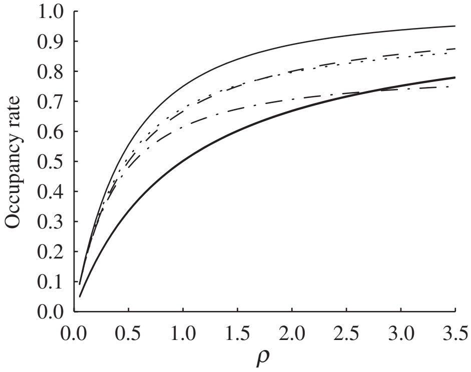

FIGURE 1: Bound on the final gene selection made under the maximization criterion. See the text.

5.3. Simulations of bounds

In order to more easily visualize the occupancy rate and the bounds derived above, we performed a series of simulation experiments. Fixing $\mu = 1$ and a strand length of $T = 1000$, we considered values of $\rho = \nu/\mu = \nu$ in the range $[0, 3.5]$. For each ρ, we generated multiple pairs of DNA strands. The maximality criterion was then applied to each such pair and the average of the empirical occupancy rates, $(1/T)W_t$, was calculated. The results of these experiments appear in Figure 1.

The solid line in Figure 1 illustrates the theoretical upper bound $\tilde{M}(\rho)$ obtained in Proposition 5 by considering the proportion of the strand length covered by islands. The dotted line is the occupancy rate, $M^*(\rho)$, estimated from the numerical simulations.

In contrast, the dash–dot line corresponds to the theoretical lower bound, $\hat{M}(\rho)$, while the double line plots the single-strand occupancy rate, $\theta^*(\rho) = \rho/(1+\rho)$. The lower bound posited in Proposition 6 appears as the combination of the dash–dot line and the double line. In the figure it can be seen that $\hat{M}(\rho)$ provides a tighter lower bound than $\theta^*(\rho)$ for values of ρ less than 2.75, while the converse is true for values of ρ greater than 2.8. Naturally, $\hat{M}(\rho)$ and $\theta^*(\rho)$ cross at some point in the interval $(2.75, 2.8)$.

We also plot the quantity, $\underline{M}(\rho)$, derived from a loss network as the dashed line. It is evident from the graph that $\underline{M}(\rho)$ yields a 'relatively' tight approximation for the empirical estimate of $M^*(\rho)$, always falling within 3.15% of $M^*(\rho)$. More precisely, $|M^*(\rho)/\underline{M}(\rho) - 1| \leq 0.0315$ for all $\rho \geq 0$. It can be seen that $\underline{M}(\rho)$ underestimates $M^*(\rho)$ for small ρ ($\rho \leq 1.7$) and overestimates $M^*(\rho)$ for large ρ ($\rho \geq 1.75$).

6. Properties of optimal selection

Now we shall study the stochastic properties arising from the application of Algorithm 1 to a realization of E. For a (continuous-time or discrete-time) process $C = (C_t : t \in \mathbb{R})$, we write $\mathcal{B}_s^C := \sigma(\{C_r : r \leq s\})$ for the σ-algebra of the process C up to time s. If D is an event and $\mathcal{A}$ is a σ-algebra, we shall use $\mathcal{A} \vee D$ to mean the σ-algebra generated by $\mathcal{A}$ and D.

Recall that the process $E = (E^1, E^{-1})$, as well as all continuous-time processes derived from it, is defined for all times in $\mathbb{R}$. Let $t \in \mathbb{R}$. We denote by $\boldsymbol{I}_{(t)}^i$ the first gene candidate appearing in strand i among those finishing after time t, so $s(\boldsymbol{I}_{(t)}^i) = \inf\{s(I) : f(I) > t,\ I \in \mathcal{I}^i\}$. Hence, if $E_t^i = 1$ then $t \in \boldsymbol{I}_{(t)}^i$, and if $E_t^i = 0$ then $\boldsymbol{I}_{(t)}^i$ is the first gene candidate on strand i starting

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

after time t. For $\tilde{E}_t = 0$, we define $\boldsymbol{I}_{(t)}$ by $\boldsymbol{I}_{(t)} = \boldsymbol{I}^i_{(t)}$ if $s(\boldsymbol{I}^i_{(t)}) < s(\boldsymbol{I}^{-i}_{(t)})$. Analogously, we denote by $\boldsymbol{I}^*_{(t)}$ the first gene appearing in $\mathcal{I}^*$ among those that finish after t, so if $E^*_t = 1$ then $t \in \boldsymbol{I}^*_{(t)}$, and if $E^*_t = 0$ then $\boldsymbol{I}^*_{(t)}$ is the first gene in $\mathcal{I}^*$ appearing after t.

Let $(I^*_n \colon n \in \mathbb{Z})$ be the family of final genes $\mathcal{I}^*$ arranged in increasing order according to their starting positions, and fix $I^*_0 = \boldsymbol{I}^*_{(0)}$.

Remark 4. Following a gap (a (0, 0)-zone), one of the first gene candidates appearing on each strand must belong to the final configuration, that is, $\tilde{E}_t = 0$ implies that $\boldsymbol{I}^1_{(t)} \in \mathcal{I}^*$ or $\boldsymbol{I}^{-1}_{(t)} \in \mathcal{I}^*$. Obviously, in the case $\boldsymbol{I}^1_{(t)} \cap \boldsymbol{I}^{-1}_{(t)} \neq \varnothing$, only one of the two genes can belong to $\mathcal{I}^*$.

Symmetry conditions imply that, following a gap, the next gene candidate to appear in the final configuration is equally likely to belong to either one of the two strands. Similarly, we see that if we consider any point near the end of a gene candidate I after which (point) there is no overlapping gene candidate on the other strand, then a gap must immediately follow I and the next gene candidate will belong to either one of the two strands with equal likelihood. Thus, we have the following result.

Lemma 6. $\mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^i \mid \mathcal{B}^E_t \vee \{\tilde{E}_t = 0\}\} = \mathrm{P}\{X_{n+1} = i \mid \mathcal{B}^{\tilde{X}}_n \vee \{\tilde{X}_n = 0\}\} = \frac{1}{2}$, $i = 1, -1$.

The situation is radically different if, at the point where a gene candidate ends, there is a gene candidate present on the other strand. In this case, we introduce the crucial parameter

$$\chi := \mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^* \mid \tilde{E}_t = 0\},$$

which is the probability that the first gene candidate that appears following a gap region is a gene. This probability satisfies the following bound.

Proposition 7. $\chi > \frac{1}{2} + 1/(2(1+\rho))$.

Proof. Let $i = \boldsymbol{i}(\boldsymbol{I}_{(t)})$ be the strand containing $\boldsymbol{I}_{(t)}$, and let $\eta = s(\boldsymbol{I}^{-i}_{(t)})$. Then

$$\mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^*,\ \tilde{E}_t = 0\} = \mathrm{P}\{\eta \geq f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0\} + \mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^*,\ \eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0\}.$$

Since $[\eta, f(\boldsymbol{I}_{(t)}))$ is strictly contained in $\boldsymbol{I}_{(t)}$, we have

$$\mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^*,\ \eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0\} > \mathrm{P}\{[\eta, f(\boldsymbol{I}_{(t)})) \in \mathcal{I}^*([\eta, \infty)),\ \eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0\}.$$

Therefore, by conditioning on η and observing that $\{\eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0\} \in \mathcal{B}^E_\eta$, we can use the strong Markov property to obtain

$$\begin{aligned}
&\mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^*,\ \eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0\} \\
&\quad > \mathrm{P}\{\eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0)\,\mathrm{P}\{[\eta, f(\boldsymbol{I}_{(t)})) \in \mathcal{I}^*([\eta, \infty)) \mid \eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0\} \\
&\quad = \mathrm{P}\{\eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0\}\,\mathrm{E}(\mathrm{P}\{[\eta, f(\boldsymbol{I}_{(t)})) \in \mathcal{I}^*([\eta, \infty)) \mid \mathcal{B}^E_\eta\} \mid \eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0) \\
&\quad = \mathrm{P}\{\eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0\}\,\mathrm{E}(\mathrm{P}_{E_\eta}\{\boldsymbol{I}_{(0)} \in \mathcal{I}^*\} \mid \eta < f(\boldsymbol{I}_{(t)}), \tilde{E}_t = 0) \\
&\quad = \mathrm{P}\{\eta < f(\boldsymbol{I}_{(t)}),\ \tilde{E}_t = 0\}\,\mathrm{P}\{\boldsymbol{I}_{(0)} \in \mathcal{I}^* \mid s(\boldsymbol{I}^i_{(0)}) = 0 \text{ for all } i\},
\end{aligned}$$

where we use $\mathrm{P}_e\{\cdot\}$ to denote $\mathrm{P}\{\cdot \mid E_0 = e\}$. Since the first gene candidate in each strand starts at 0 and the lengths of both candidates are independent, identically distributed exponential

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

random variables with parameter μ, one of them must be in the final configuration and $\mathrm{P}\{\boldsymbol{I}_{(0)} \in \mathcal{I}^* \mid s(\boldsymbol{I}^i_{(0)}) = 0 \text{ for all } i\} = \frac{1}{2}$. Hence,

$$\begin{aligned}\mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^* \mid \tilde{E}_t = 0\} &> \mathrm{P}\{\eta \geq f(\boldsymbol{I}_{(t)}) \mid \tilde{E}_t = 0\} + \tfrac{1}{2}\mathrm{P}\{\eta < f(\boldsymbol{I}_{(t)}) \mid \tilde{E}_t = 0\} \\ &= \tfrac{1}{2}(1 + \mathrm{P}\{\eta \geq f(\boldsymbol{I}_{(t)}) \mid \tilde{E}_t = 0\}).\end{aligned}$$

Now $\mathrm{P}\{\eta \geq f(\boldsymbol{I}_{(t)}) \mid \tilde{E}_t = 0\} = p^{\tilde{X}}_{10} = 1/(1+\rho)$, and the proof is complete.

Lemma 7. $\mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^* \mid \mathcal{B}^E_t \vee \{\tilde{E}_t = 0\}\} = \chi$.

Proof. Set $I = \boldsymbol{I}_{(t)}$ and let $\Omega_1 \subseteq \{\tilde{E}_t = 0\}$ belong to $\mathcal{B}^E_t$. Since $\boldsymbol{I}_{(t)} \in \mathcal{I}^*$ is $\sigma(E_t : t \geq s(I))$ -measurable, from the Markov property we find that $\boldsymbol{I}_{(t)} \in \mathcal{I}^*$ and Ω_1 are conditionally independent with respect to the event $\{\tilde{E}_t = 0\}$, and the result follows.

Lemma 8. *Defining* $\|K\|$ *to be the number of gene candidates contained within an island* K*, we find that* $\mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^* \mid \|K(\boldsymbol{I}_{(t)})\| \geq 2,\ \tilde{E}_t = 0\} > \frac{1}{2}$.

Proof. We have

$$\begin{aligned}&\mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^* \mid \tilde{E}_t = 0\} \\ &\quad = \mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^* \mid \tilde{E}_t = 0,\ \|K(\boldsymbol{I}_{(t)})\| = 1\}\,\mathrm{P}\{\|K(\boldsymbol{I}_{(t)})\| = 1 \mid \tilde{E}_t = 0\} \\ &\qquad + \mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^* \mid \tilde{E}_t = 0,\ \|K(\boldsymbol{I}_{(t)})\| \geq 2\}\,\mathrm{P}\{\|K(\boldsymbol{I}_{(t)})\| \geq 2 \mid \tilde{E}_t = 0\}.\end{aligned}$$

Since $\mathrm{P}\{\|K(\boldsymbol{I}_{(t)})\| = 1 \mid \tilde{E}_t = 0\} = 1/(1+\rho)$, the first term on the right-hand side of this equation becomes

$$\mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^* \mid \tilde{E}_t = 0,\ \|K(\boldsymbol{I}_{(t)})\| = 1\}\,\mathrm{P}\{\|K(\boldsymbol{I}_{(t)})\| = 1 \mid \tilde{E}_t = 0\} = \frac{1}{1+\rho}.$$

This is because any island formed by a single gene candidate necessarily belongs to the optimal solution.

On the other hand, $\mathrm{P}\{\|K(\boldsymbol{I}_{(t)})\| \geq 2 \mid \tilde{E}_t = 0\} = \rho/(1+\rho)$ and, so, the second term may be written as $q\rho/(1+\rho)$, with $q := \mathrm{P}\{\boldsymbol{I}_{(t)} \in \mathcal{I}^* \mid \|K(\boldsymbol{I}_{(t)})\| \geq 2,\ \tilde{E}_t = 0\}$. The result then follows by applying Proposition 7 and solving for q.

Before proceeding to the next result, let us pause for a moment to define some necessary notation. We shall use K to denote the island containing $I \in \mathcal{I}^*$ and we shall use I^+ to stand for the successor to I in $\mathcal{I}^*$.

From Algorithm 1, recall that $\hat{J}_m$ is a solution to (3) with $\Sigma(\hat{J}_m) = \hat{w}_m$ for all $m = 1, 2, \ldots, n$, n being the number of gene candidates contained in the island K. By convention, we shall set $\hat{f}_{n+1} = 0$. Then we define

$$F^+(K) := \{f(I_m) \colon \hat{f}_m = 1 \text{ and } \hat{f}_{m+1} = 0,\ m = 1, 2, \ldots, n\}$$

and $f^*(K) := \max F^+(K)$.

Each point in $F^+(K)$ corresponds to the end point of some gene candidate; that is, if $t \in F^+(K)$ then there exists an $m \in \{1, 2, \ldots, n\}$ such that $t = f(I_m)$. If $t = f^*(K)$ then Algorithm 1 ensures that $\hat{J}_{m'} = \hat{J}_m$ for all $m' > m$. So, we will suppose that $t \neq f^*(K)$. Since there are only two strands and I_m belongs to an island, each I_m possesses some important characteristics. First, according to the ordering '$\preceq_{\mathrm{f}}$', the interval I_m terminates no later than

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

I_{m+1}, that is, $t \le f(I_{m+1})$. Second, I_{m+2} must lie after I_m on the same strand and cannot finish at any point prior to $f(I_{m+1})$. Hence, $\hat{J}_{m+2}$ will comprise the candidates from $\hat{J}_m$ plus I_{m+2}. In a similar way, provided that $m+3 \le n$, we can see that $\hat{J}_{m+3}$ must contain all the members of $\mathcal{J}_m$ plus I_{m+2} and/or I_{m+3}. Thus, $\hat{J}_m$ is also a subset of $\hat{J}_{m+3}$. Now, for $m' > m+3$, Algorithm 1 ensures that $\hat{J}_{m'}$ will include either all the candidates in $\hat{J}_{m+2}$ or all those in $\hat{J}_{m+3}$. Therefore, any solution $\hat{J}_m$ to (3) such that $f(I_m) \in F^+(K)$ is a subset of the solutions of all problems of size larger than m. In other words, $\hat{J}_m \subseteq \hat{J}_{m'}$ for all $m' = m+1, m+2, \dots, n$. On the other hand, it is *not* always true that $\hat{J}_{m'} \subseteq \hat{J}_m$ for all $m' = 1, 2, \dots, m-1$. To see this, it suffices to consider $\hat{J}_m$ and $\hat{J}_{m-1}$ and note that $I_m \cap I_{m-1} \ne \varnothing$ since $f(I_m) \in F^+(K)$.

This means that, given a point $t \in F^+(K)$ corresponding to the end point of I_m, decisions made in order to obtain the optimal selection of genes over the interval $[s(K), t)$ do not in any way influence or affect the optimal choice of genes within the interval $[t, f(K))$. In the Markovian model we have proposed, the optimal solutions $\mathcal{I}^*([s(K), t))$ and $\mathcal{I}^*([t, f(K)))$ are in fact conditionally independent given any such point t. Thus, we have $\mathcal{I}^*(K) = \mathcal{I}^*([s(K), t)) \cup \mathcal{I}^*([t, f(K)))$, and equality holds in Proposition 2 with $a = s(K)$, $b = t$, and $c = f(K)$.

In fact, we can elucidate the structure of solutions to the maximization problem (1) even further. Let t and t', with $t \le t'$, be two 'consecutive' elements of $F^+(K)$ corresponding to the end points of I_k and $I_{k'}$, respectively. The stipulation that t and t' be consecutive means that $F^+(K) \cap (t, t') = \varnothing$. Then $\boldsymbol{i}(I) = \boldsymbol{i}(J)$ for all $I, J \in \mathcal{I}^*([t, t'))$. Alternatively, we can say that $\mathcal{I}^*([t, t')) = \mathcal{I}^i([t, t'))$ for some $i \in \{1, -1\}$. To see this, first note that $\hat{f}_k = 1$, $\hat{f}_{k+1} = 0$, and $\hat{f}_m = 1$ for all $m = k+2, \dots, k'$. Now let $m = k'$ and note that $I_{k'} \in \hat{J}_{k'}$. Since $\hat{f}_m = 1$, we have $\hat{J}_m = \hat{J}_{l_m} \cup \{I_m\}$, so I_{l_m} immediately precedes I_m in $\hat{J}_m$. However, by definition of l_m, I_{l_m} must belong to the same strand as I_m. By replacing m with l_m, we can recursively repeat this argument and show that all gene candidates contained in the interval $[t, t')$ lie on the same strand.

The next results make use of the equivalence $\{\tilde{E}_{f(I)} = 0\} \iff s\{E^{-\boldsymbol{i}(I)}_{f(I)} = 0\}$, which holds for every $I \in \mathcal{I}$.

Lemma 9. *Let $I \in \mathcal{I}^*$. Then*

$$\mathrm{P}\{\boldsymbol{i}(I^+) = \boldsymbol{i}(I) \mid \mathcal{B}^E_{f(I)} \vee \{\tilde{E}_{f(I)} = 0\}\} = \tfrac{1}{2}, \quad (20)$$

$$\mathrm{P}\{\boldsymbol{i}(I^+) = \boldsymbol{i}(I) \mid \mathcal{B}^E_{f(I)} \vee \{\tilde{E}_{f(I)} = 1,\ f(I) = f^*(K(I))\}\} = \tfrac{1}{2}, \quad (21)$$

$$\mathrm{P}\{\boldsymbol{i}(I^+) = \boldsymbol{i}(I) \mid \mathcal{B}^E_{f(I)} \vee \{\tilde{E}_{f(I)} = 1,\ f(I) \notin F^+(K(I))\}\} = 1, \quad (22)$$

$$\mathrm{P}\{\boldsymbol{i}(I^+) = \boldsymbol{i}(I) \mid \mathcal{B}^E_{f(I)} \vee \{\tilde{E}_{f(I)} = 1,\ f(I) \in F^+(K(I)),\ f(I) \ne f^*(K(I))\}\} = \chi. \quad (23)$$

Proof. Relation (20) is straightforward to verify and implies (21) since the condition that $\tilde{E}_{f(I)} = 1$ and $f(I) = f^*(K)$ means that there is an empty region (gap) appearing after I and immediately preceding I^+.

Next, suppose that I is such that $f(I) \notin F^+(K(I))$. Also, let t and t' be consecutive elements of $F^+(K(I))$ such that $I \in [t, t']$. From the discussion following Lemma 7, we know that if t and T' are two successive elements in $F^+(K(I))$, then all the gene candidates in $\mathcal{I}^*([t, t'))$ lie on the same strand. As a result, $I^+ \in [t, t')$, $\boldsymbol{i}(I) = \boldsymbol{i}(I^+)$, and we see that (22) holds.

Finally, let us show that the last relation, (23), holds. Set $i := \boldsymbol{i}(I)$. Since $E^{-i}_{f(I)} = 1$, there is a gene candidate L on strand $-i$ containing the point $f(I)$. Thus, L cannot appear in the final configuration $\mathcal{I}^*$, because L intersects I and $I \in \mathcal{I}^*$. Now set $J = \boldsymbol{I}^i_{(f(I))}$ and let $\Omega_1 \subseteq \{\tilde{E}_t = 1,\ f(I) \in F^+(K(I)),\ f(I) \ne f^*(K(I))\}$ and $\Omega_1 \in \mathcal{B}^E_{f(I)}$. Since $J \in \mathcal{I}^*$

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

is $\sigma(E_t : t \geq s(J))$-measurable, from the Markov property we find that $J \in \mathcal{I}^*$ and Ω_1 are conditionally independent with respect to the event $\{\tilde{E}_t = 1,\ f(I) \in F^+(K(I)),\ f(I) \neq f^*(K(I))\}$. Therefore,

$$\begin{aligned}\mathrm{P}\{\boldsymbol{i}(I^+) = i \mid \mathcal{B}^E_{f(I)} \vee \{\tilde{E}_{f(I)} = 1,\ f(I) \in F^+(K(I)),\ f(I) \neq f^*(K(I))\}\}\\ = \mathrm{P}\{\boldsymbol{i}(I^+) = i \mid \tilde{E}_{f(I)} = 1,\ f(I) \in F^+(K(I)),\ f(I) \neq f^*(K(I))\}.\end{aligned}$$

Since $L \notin \mathcal{I}^*$, we have

$$\begin{aligned}\mathrm{P}\{\boldsymbol{i}(I^+) = i \mid \tilde{E}_{f(I)} = 1,\ f(I) \in F^+(K(I)),\ f(I) \neq f^*(K(I))\}\\ = \mathrm{P}\{\boldsymbol{I}_{(f(I))} \in \mathcal{I}^* \mid \tilde{E}_{f(I)} = 0\} = \chi,\end{aligned}$$

and the result follows.

Using Lemma 9 and Proposition 7, we obtain the following result.

Corollary 1. *Let $I \in \mathcal{I}^*$ and let I^+ be its successor in $\mathcal{I}^*$. Then* $\mathrm{P}\{\boldsymbol{i}(I^+) = \boldsymbol{i}(I)\} > \frac{1}{2}$.

7. An application to *Escherichia coli*

In this section, we consider the optimization procedure in the context of a complete, annotated DNA sequence. We shall see that the optimization paradigm proposed is able to capture various statistical properties we observe in bacterial DNA. Our empirical studies have been done using the well-documented *Escherichia coli* K-12 organism. The genome of this prokaryote comprises $n = 4\,639\,221$ nucleotide base pairs. The raw DNA sequence was sourced from the *E. coli* Genome Project at the Laboratory of Genetics, University of Wisconsin-Madison. The annotation was taken from GenBank, updated as of 7 March 2003.

7.1. Basic statistics

In preparation for seeing how the maximization model performs with *E. coli* data, we applied a rather naive procedure to the raw DNA sequence of the *E. coli* genome in order to generate gene candidates for each of the six reading frames. Note that what we have been calling gene candidates are commonly referred to as open reading frames by biologists. Each strand of DNA is a chain of nucleotides and is represented as a sequence of letters from the alphabet $\alpha = \{\mathrm{A, C, G, T}\}$. Let START and STOP be subsets of α^3 containing the *start codons* and *stop codons*, respectively. The genetic code possesses three immutable stop codons: STOP = {TAA, TAG, TGA}. In addition, the genome of the *E. coli* organism makes use of five start codons: START = {ATG, ATT, CTG, GTG, TTG}. We note that the primary strand, S^1, and the complementary strand, S^{-1}, are read in opposite directions, so S^1_j pairs with S^{-1}_{n-j+1} for $j = 1, 2, \dots, n$ under the pairing rules A ⇔ T and C ⇔ G.

To find candidate genes in the ith reading frame on the primary strand, we examine the strand in reading order at points $\{i, i+3, i+6, \dots\}$ until we find a triplet (codon) belonging to START. Having found a potential start codon, we record its position, t, and then continue searching, for a member of STOP. Upon finding a stop codon, we note the position, T, of the last letter composing it. The first candidate is then $[t, T]$. The next candidate is found by repeating this procedure from position $T + 1$ on the strand, and so on. This procedure generates $\mathcal{I}^1$, $\mathcal{I}^2$, and $\mathcal{I}^3$. When finding gene candidates on the complementary strand, a little care must be taken to represent them using the same frame of reference used on the primary strand. Thus, positions are labelled $n, n-1, n-2, \dots, 1$ when the strand is viewed according to its 5′–3′

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

Table 1: Some statistics concerning the genome of the *E. coli* K-12 bacteria.

Strand	Frame	n_c	Empirical average candidate length (bp)	Empirical average gap length (bp)	Candidate occupancy rate (%)	Candidate K–S statistic	Gap K–S statistic
1	1	30 842	96.66	55.75	64.26	0.12	0.11
1	2	30 658	97.12	56.19	64.18	0.13	0.11
1	3	30 718	97.63	55.38	64.64	0.12	0.11
−1	−1	30 338	97.53	57.38	63.78	0.13	0.11
−1	−2	30 950	96.52	55.36	64.39	0.13	0.11
−1	−3	30 505	98.55	55.19	64.80	0.13	0.10

reading order, and offsets from position n to the starting point of each reading frame must be taken into account. With these changes in mind, the procedure outlined above also produces $\mathcal{I}^{-1}$, $\mathcal{I}^{-2}$, and $\mathcal{I}^{-3}$.

Table 1 summarizes various statistics for each of the six reading frames. It shows the number of gene candidates found, n_c, the mean lengths (expressed in base pairs (bp)) of the candidates and the gaps separating them, and the proportion of the reading frame occupied by candidates. In addition, the last two columns of the table show Kolmogorov–Smirnov (K–S) statistics which test the gene candidate and gap length distributions against exponential hypotheses.

It is clear from the Kolmogorov–Smirnov statistics in Table 1 that neither the gene candidates nor the gap lengths are exponentially distributed. Burge and Karlin [1], as well as Lukashin and Borodovsky [4], have made the same observations in eukaryotic genome sequences.

7.2. Comparison between gene annotation and final genes

As noted earlier, we have used a complete annotation of *E. coli* K-12 sourced from GenBank and dated 7 March 2003. According to the annotation, the genome of this organism has 4266 annotated genes or coding sequences which occupy 87.48% of its base pairs. Henceforth, we will purposefully abuse the correct genetics terminology and use 'CDS' to refer to a gene which is recognized as a coding sequence in the *E. coli* K-12 annotation. We are partly justified in doing so because CDSs correspond to genes in prokaryotic organisms like bacteria. We shall continue to use 'final gene' to mean any gene candidate our model accepts as a valid gene under the maximization criterion. In a very small percentage of the *E. coli* genome (0.32% of the base pairs), the CDSs in the annotation do, in fact, overlap. Although this contradicts the primary requirement of the maximization criterion, it is nevertheless a very small proportion of the total genome. We recognize that this small portion of the genome will be inaccessible to our model in its current form, and we shall not consider it for the time being.

It is important to note that the set of gene candidates $\mathcal{I}$ excludes many of the coding sequences recognized as genes in the *E. coli* K-12 annotation. We shall use $\mathcal{C}$ to denote the set of those annotated genes that also belong to $\mathcal{I}$. After applying the candidate extraction process to the *E. coli* data, we found that $\mathcal{C}$ only contains 2848 (66.76%) of the 4266 annotated genes. Clearly this imposes a limit on the number of annotated genes that will appear in the optimal solution found by the maximization procedure. The scheme we have used to construct $\mathcal{I}$ from the raw DNA sequence is simplistic and, in line with the concept of maximizing the proportion of the genome devoted to genic zones, is biased towards finding larger gene candidates. There are two classes of potential genes that the selection scheme entirely overlooks. The first of these

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

are those potential genes whose start codon is contained in a member of $\mathcal{I}$. The second class is those genes which undergo a frame shift such that their start codons appear in one reading frame while their end codons appear in another. The initial candidate selection process is far from ideal, but it nevertheless serves to illustrate an application of the maximization model to real-world data. Suffice it to say that the automatic identification of potential genes remains one of the difficult problems in gene annotation.

Having implemented the maximization algorithm from Section 2, we proceeded to apply it to the whole set of gene candidates. In our experiment, each of the six reading frames in the *E. coli* DNA sequence corresponds to a so-called 'strand' in the description of the maximization algorithm. Using the maximization criterion we obtained 15 336 final genes, occupying 95.49% of the total genome. The percentage of final genes that were found to be CDSs (the positive predictive value of the model) was 13.43%. In other words, 86.57% of the genes identified by the model were false positives. On the other hand, the proportion of CDSs which were counted among the final genes selected was 48.27%.

The sensitivity and specificity of the maximization procedure as a method for identifying annotated genes can be calculated. The model's sensitivity, expressed as a percentage, is 72.30%, while its specificity is 92.67%. It is worth noting that, while the mean length of a coding sequence in the genome annotation data is 954.95 nucleotides, the mean length of a final gene as determined by the model is only 288.86 base pairs. Also, as was mentioned above, annotated CDSs account for 87.48% of the genome, but final genes occupy 95.49%. The model tends to include too many final genes, which results in the overestimation of the occupancy rate. Also, given the small mean length of final genes, it would appear that the final genes are generally too small. It is known that bacterial genomes contain very few CDSs which are less than 100 base pairs in length. *E. coli* has 13 such CDSs. In light of this, it seems natural to consider only those gene candidates containing 100 or more base pairs in order to understand where the model is not performing satisfactorily. We performed a series of experiments in which gene candidates whose lengths were less than a certain threshold value were eliminated from all the reading frames before applying the maximization algorithm. We considered threshold values in the range from 100 to 250 in steps of size 1. Figures 2 and 3 summarize our findings.

The solid line in Figure 2 shows the proportion of the genome length covered by final genes in the model's maximal solution, $M^*(\rho)$. Observe that the model overestimates the occupancy rate for smaller threshold values and underestimates it for larger ones. The dashed line in the graph shows that the percentage of final genes that are annotated in the GenBank data (i.e. that are CDSs) steadily increases as the threshold value increases. This is as expected. On the other hand, the proportion of annotated genes that are also final genes (the dotted line) confines itself to a narrow band between 48.59% and 50.23%, peaking at a threshold value of 206–208 base pairs. This suggests that the proportion of CDSs that constitute final genes in the model appears to be fairly insensitive to changes in the threshold value.

Figure 3 plots the sensitivity (the percentage of candidates in $\mathcal{C}$ classified as final genes (dashed line)) and specificity (the percentage of candidates in $\mathcal{I} \setminus \mathcal{C}$ not classified as final genes (dotted line)) over the range of threshold values, in addition to the proportion, $100|\mathcal{C}|/|\mathcal{I}|$ (solid line), of annotated genes contained in the gene candidates submitted to the maximization procedure. The sensitivity increases due to the fact that smaller gene candidates overlapping a CDS which cause that CDS to be rejected as a final gene are themselves removed as candidates by the threshold criterion. The observed decrease in specificity occurs because eliminating candidates below a threshold length causes the number of true negatives to diminish quickly relative to the number of false positives.

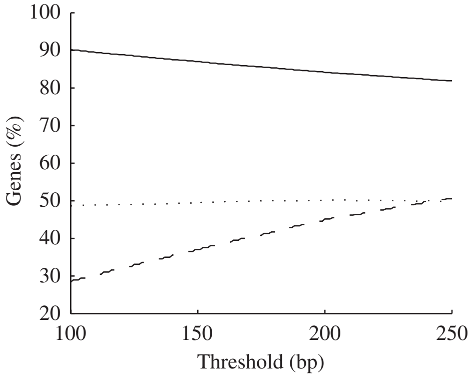

FIGURE 2: Comparison of actual genes (CDSs obtained from annotation data) and final genes (as predicted by the model) when genes smaller than a specified threshold value are excluded. See the text.

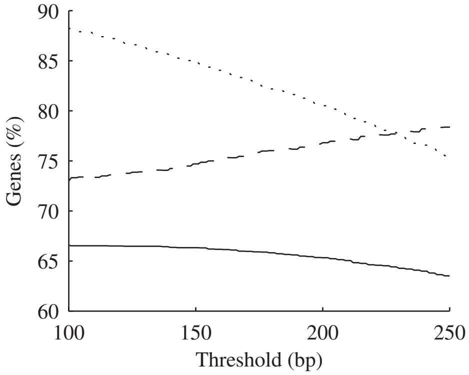

FIGURE 3: Sensitivity and specificity of the model when genes smaller than a specified threshold are excluded. See the text.

Removing gene candidates whose lengths are below a certain threshold value has illustrated that the model is reasonably good at identifying large genes but is fairly poor at identifying small genes.

8. Concluding remarks

In this work, we have proposed a global scheme for modelling the placement of genes in bacterial genomes. Efficient algorithms for implementing the suggested maximization criterion have been devised. We have studied the probabilistic properties of the simplified two-strand case (ignoring reading frames altogether). Even in this restricted case, the model succeeds in capturing some of the statistical properties of the genome. For example, the lower bound established in Lemma 8 holds in the case of *E. coli*. Consider islands containing at least two gene candidates. Let k_2 be the number of such islands. Next let k_2^1 be the number of these islands whose first gene candidate is a CDS. Then $k_2^1/k_2 = 0.6337 > 0.5$, as predicted.

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press

We have implemented some simple variations of this model. In one of these, we eliminated all the gene candidates that were contained within gene candidates on other strands. We have not included this study here because the addition of this constraint does not alter the theoretical analysis and, moreover, does not result in a significant improvement to the fit of the data.

In future, the authors plan to remove as many of the independence assumptions as possible and expand the probabilistic results to encompass six reading frames. Also, we have yet to consider the maximization criterion as a useful adjunct to assist existing gene identification techniques such as the hidden Markov model scheme at the core of R'HOM [5], [6] or the interpolated Markov model employed by the GLIMMER system [7]. Finally, it will be necessary to apply the model to a range of bacterial genomes other than *E. coli* K-12 in order to more fully assess the utility of the model we have presented.

Acknowledgements

This work was supported by the Millennium Nucleus in Information and Randomness P04-069-F, MSI. The authors would like to thank Raphael Hauser for discussions and for pointing out an error in Section 5. For discussions and support, thanks also go to the members of the Laboratory of Bioinformatics and Mathematics of the Genome in the CMM.

References

[1] Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Molec. Biol.* **268,** 78–94.

[2] Kelly, F. P. (1991). Loss networks. *Ann. Appl. Prob.* **1,** 319–378.

[3] Krengel, U. (1985). *Ergodic Theorems* (De Gruyter Stud. Math. **6**). Walter De Gruyter, Berlin.

[4] Lukashin, A. V. and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115.

[5] Nicolas, P. (2003). Mise au point et utilisation de modèles de Markov cachées pour l'etude des séquences d'ADN. Doctoral Thesis, Université d'Evry.

[6] Nicolas, P. and Muri-Majoube, F. (2001). R'HOM. Programs to segment DNA sequences into homogeneous regions. Tech. Rep., Université d'Evry. Available at http://genome.jouy.inra.fr/ssb/rhom/rhom_doc/rhom_doc.html.

[7] Salzberg, S. L., Delcher, A. L., Kasif, S. and White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26,** 544–548.

https://doi.org/10.1017/S0001867800001452 Published online by Cambridge University Press