



Distribution of words with a predefined range of mismatches to a DNA probe in bacterial genomes

O. Michael Melko^{1,*}, § and Arcady R. Mushegian^{1,2}

¹Stowers Institute for Medical Research, Kansas City, MO 64110, USA and

²Department of Microbiology, Molecular Genetics and Immunology, University of Kansas Medical Center, Kansas City, KS 66160, USA

Received on March 25, 2003; revised on June 11, 2003; accepted on July 22, 2003

ABSTRACT

Motivation: Hybridization of oligonucleotides with longer nucleotide sequences is an essential step in nucleic acid biosynthesis *in vitro* and *in vivo*, in oligonucleotide-based diagnostics, and in therapeutic applications of oligonucleotides. A major factor determining sensitivity and selectivity of hybridization is the number of base pair mismatches that occur in an ungapped alignment of the oligonucleotide (probe) and a longer sequence (target).

Results: The *k*-distance match count between the probe and the target is defined as the number of ungapped alignments between the two sequences that have exactly *k* mismatches, and the *k*-neighbor match count is defined as the sum of the *j*-distance match counts for *j* between 0 and *k*. We derive a novel formula for the probability of a *k*-distance match. This formula is based on the assumption that the target is strand-symmetric Bernoulli text (i.e. nucleotides are independently, identically distributed in the target and satisfy Chargaff's second parity rule). Our model predicts that the GC-content in both the probe and the target significantly affects the match count expectation. The ratio of *k*-neighbor match counts in two distinct genomes for a given probe is a measure of its specificity. We calculated such ratios for pairs of bacterial genomes with different combinations of length, GC-content and phylogenetic distance. Examination of the extreme values of these ratios indicates that probes with a high discriminative power exist for each tested pair.

Contact: omm@stowers-institute.org

Supplementary information: Stowers Institute Technical Report No. 0002, C++ source code, *Mathematica* notebooks and other information is available at <http://www.stowers-institute.org/labs/bioinformatics/omm/index.htm>

INTRODUCTION

An understanding of nucleotide sequence hybridization is fundamental to the study of any process involving DNA or

RNA synthesis, function, or turnover *in vitro* and *in vivo*. In particular, the utility of molecular diagnostic probes and therapeutic nucleotides is determined by the extent to which their hybridization to a target can be controlled. A good formalism for identifying such probes and predicting their behavior will involve a statistical model for the distribution of words in target sequences, and an understanding of the relation of base composition, order and mismatch location to probe hybridization. Such a formalism is currently unavailable.

This paper focuses on aspects of genomic sequences relevant to the problem of finding useful probes. Specifically, we first look at the degree of randomness of the distribution of words in bacterial genomes, then we study the distribution of *k*-neighbor matches between a probe and a target genome, and finally, we undertake a comparative investigation of *k*-neighbor matches in pairs of bacterial genomes in order to find good probe candidates. Before continuing, we describe the context of our investigation more formally.

Suppose that we have *r* probes \mathbf{q}_i of a fixed length *m* ($1 \leq i \leq r$), and *s* target nucleotide sequences \mathbf{T}_j of variable length ($1 \leq j \leq s$) in an assay. Let σ_i denote the total signal from probe \mathbf{q}_i , and let σ_{ij} denote the signal produced by hybridization of \mathbf{q}_i to \mathbf{T}_j . If c_j is the concentration of \mathbf{T}_j and n_i is the 'noise' associated with σ_i , we have the following system of linear equations

$$\sum_{j=1}^s \sigma_{ij} c_j + n_i = \sigma_i \quad (1)$$

The quantities c_j and n_i are the unknowns in this system, and we would like to solve for c_j (or at least determine if $c_j > 0$). Before we can solve this system, we must estimate the signals σ_{ij} and the noise n_i .

The terms σ_{ij} , n_i can be described thermodynamically, at least in principle, and the corresponding signal-to-noise ratio depends on environmental factors, such as ambient temperature and salt concentrations (Bonnet *et al.*, 1999; SantaLucia, 1998). In the sequel, we assume that the noise to be negligible (i.e. $n_i = 0$). The intensity of the hybridization signal σ_{ij} at a given temperature depends on two factors: the number of

*To whom correspondence should be addressed at

§Present address: Northern State University, 1200 S. Jay Street, NSU Box 713, Aberdeen, SD 57401, USA

sites on \mathbf{T}_j at which \mathbf{q}_i can bind to produce a hybrid, and the stability of the resulting hybrid. Note that, in general, \mathbf{q}_i could have highly stable binding sites on \mathbf{T}_j for $i \neq j$, thus producing strong signals at those sites. This does not necessarily impede the solution of (1); we simply require that the matrix of signals σ_{ij} satisfies an appropriate regularity condition (such as invertibility, if $r = s$). We will consider the more stringent condition that the cross-hybridization signals σ_{ij} are weak for $i \neq j$, when compared with σ_{ii} , making the system (1) approximately diagonal. One step toward insuring the latter is the identification of probes \mathbf{q}_i , which do not have any highly complementary sites on targets other than \mathbf{T}_i .

The stability of a particular hybrid is affected by the degree of complementarity (i.e. the number and location of base pair mismatches), base composition and base order (Kaderali and Schliep, 2002). The combined effect of these parameters on the stability of the hybrids has not been systematically assessed (Bonnet *et al.*, 1999; Pozhitkov and Tautz, 2002), and the description of hybridization is further complicated by the fact that the probe, target and hybrid all can form complicated secondary structures. We ignore these issues in the current work, and further assume that only the number, but not the location, of mismatches affects the stability. Estimation of σ_{ij} is then a matter of determining (i) the distribution of ungapped alignments between the oligonucleotide and the target that have k base pair mismatches ($k \leq m$) and (ii) the stability (or melting temperature) of a given hybrid with k base pair mismatches.

In what follows, we study the distribution of k -distance and k -neighbor match counts in bacterial genomes. For the case of two bacterial genomes, we also give empirical evidence that probes can be found with weak cross-hybridization signals. Our method can probably be generalized to larger sets of genomes.

RESULTS AND DISCUSSION

The simplest statistical model of a nucleotide sequence is that of uniform Bernoulli text. This model assumes that sequences arise from independent, identically distributed and equiprobable random drawings of characters from an alphabet. It is often used to derive estimates for the expectation and variance of match counts between a probe and a target (Pevzner, 2000). We will see that estimates of expectation based on this model are inaccurate when the GC-content of a nucleotide sequence deviates substantially from 50%.

Two popular ways to generalize this model are, first to abandon the assumption that the random drawings are equiprobable, and second, to use a Markov model (Prum *et al.*, 1995). Here, we present a model based on the first approach. While the space of probability measures on the alphabet of nucleotides is naturally a three-dimensional simplex, our essential idea is to reduce the degrees of freedom in this space from three to one by invoking Chargaff's second parity rule (Forsdyke,

2002). This rule asserts that the probability of occurrence of individual nucleotides ξ_i , $i \in \mathcal{A} := \{A, G, C, T\}$ satisfy the equations $\xi_A = \xi_T$ and $\xi_C = \xi_G$, in a (long enough) *single-stranded DNA fragment*. The second parity rule has been observed to hold, to a good approximation, in actual nucleotide sequences. This can be explained, at least in part, by frequent strand inversion recombination events in the evolution of genomes (Fickett *et al.*, 1992), although other mechanisms generating parity also seem to be required (Forsdyke, 2002; Baisnée *et al.*, 2002). Thus the degrees of freedom in the nucleotide frequency space are reduced to 1, with three missing values easily computed if the frequency of one nucleotide is known. We will say that Bernoulli text is *strand-symmetric* if its underlying probability measure ξ satisfies this rule. In this case we will also say that ξ is strand-symmetric.

Thus, if \mathbf{T} is one strand of a long nucleotide sequence, we expect to find the *relative character frequencies* $\xi_i^*(\mathbf{T}) := n_i(\mathbf{T}) / \sum_{i \in \mathcal{A}} n_i(\mathbf{T})$ to satisfy $\xi_A^*(\mathbf{T}) \approx \xi_T^*(\mathbf{T})$ and $\xi_C^*(\mathbf{T}) \approx \xi_G^*(\mathbf{T})$, where $n_i(\mathbf{T})$ denotes the number of occurrences of the character i in \mathbf{T} . In what follows, we estimate the probabilities ξ of \mathbf{T} by setting $\xi_A = \xi_T = [\xi_A^*(\mathbf{T}) + \xi_T^*(\mathbf{T})]/2$ and $\xi_C = \xi_G = [\xi_C^*(\mathbf{T}) + \xi_G^*(\mathbf{T})]/2$.

Definition and properties of the perturbed binomial distribution

Throughout this paper, random variables will be denoted by upper case roman characters, such as X , and their outcomes will be denoted by the corresponding lower case characters. Sequences of random variables will be set in upper case boldface, e.g. $\mathbf{X} = X_1 X_2 X_3 \dots$, and lower case boldface will denote a corresponding sequence of possible outcomes.

We denote the number of character mismatches between two words \mathbf{x}, \mathbf{y} of length m (m -words) by $\delta(\mathbf{x}, \mathbf{y})$, and we say that \mathbf{x} is a k -distance match of \mathbf{y} if $\delta(\mathbf{x}, \mathbf{y}) = k$, resp., \mathbf{x} is a k -neighbor of \mathbf{y} if $\delta(\mathbf{x}, \mathbf{y}) \leq k$.

Let m be a non-negative integer, suppose that $0 \leq \rho \leq 1$, and set $\hat{\rho} = 1 - \rho$. Then the (standard) binomial distribution is given by $\{b_k(m, \rho)\}_{k=0}^m$, where

$$b_k(m, \rho) := \binom{m}{k} \rho^{m-k} \hat{\rho}^k \quad (2)$$

and $\binom{m}{k}$ denotes the binomial coefficient $m! / (k!(m-k)!)$. If \mathbf{X} is uniform Bernoulli text of length m , and if \mathbf{y} is a word of the same length, the probability of a k -distance match between \mathbf{X} and \mathbf{y} is given by $p[\delta(\mathbf{X}, \mathbf{y}) = k] = b_k(m, 1/4)$.

More generally, the *perturbation parameter* η is defined to be the unique number satisfying the relations

$$\xi_C = \xi_G = \frac{1}{4}(1 - \eta), \quad \xi_A = \xi_T = \frac{1}{4}(1 + \eta) \quad (3)$$

where $\xi = \{\xi_A, \xi_T, \xi_C, \xi_G\}$ is a strand-symmetric probability measure. Note that $-1 \leq \eta \leq 1$. The *perturbed binomial*

distribution $g_k(m, \eta, c)$ is then defined by the formula

$$g_k(m, \eta, c) = h(m, \eta, c) u_k(m, \eta, c) \quad (4)$$

where c, k, m are integers with $0 \leq c, k \leq m$, and

$$h(m, \eta, c) = \frac{1}{4^m} (1 - \eta)^c (1 + \eta)^{m-c}$$

$$u_k(m, \eta, c) = \sum_{i=0}^{m-k} \binom{c}{i} \binom{m-c}{m-k-i} v_k(i, \eta, c)$$

$$v_k(i, \eta, c) = \left(\frac{3 + \eta}{1 - \eta} \right)^{c-i} \left(\frac{3 - \eta}{1 + \eta} \right)^{k+i-c}$$

Now, if \mathbf{X} is strand-symmetric Bernoulli text of length m with perturbation parameter η , it follows that[†] the probability $p[\delta(\mathbf{X}, \mathbf{y}) = k] = g_k(m, \eta, c)$, where \mathbf{y} is a given m -word over \mathcal{A} and c is the GC-content of \mathbf{y} . Observe that $g_k(m, 0, c) = b_k(m, 1/4)$, and note further that, if X is a non-negative integer-valued random variable with distribution $\{g_k(m, \eta, c)\}_{k=0}^m$, then the expectation and variance of X are given by[†]

$$E(X) = E_0 \left[1 - \frac{1}{3} \left(1 - \frac{2c}{m} \right) \eta \right]$$

$$V(X) = V_0 \left[1 + \frac{2}{3} \left(1 - \frac{2c}{m} \right) \eta - \frac{1}{3} \eta^2 \right] \quad (5)$$

where $E_0 := 3m/4$ and $V_0 := 3m/16$ are the expectation and variance, respectively, for a random variable with distribution $\{b_k(m, 1/4)\}_{k=0}^m$. In particular, if $X = \delta(\mathbf{X}, \mathbf{y})$, then the first equation in (5) gives the expected number of matches between \mathbf{X} and \mathbf{y} , and the second gives the variance in match counts. These equations show that the perturbed binomial distribution differs significantly from the binomial distribution for values of η significantly different from 0.

The *cumulative perturbed distribution* is defined by the formula $G_k(m, \eta, c) := \sum_{j=0}^k g_j(m, \eta, c)$. It is clear that $p[\delta(\mathbf{X}, \mathbf{y}) \leq k] = G_k(m, \eta, c)$. Thus, $G_k(m, \eta, c)$ is the probability that strand-symmetric Bernoulli text \mathbf{X} with length m and perturbation parameter η is a k -neighbor of a word \mathbf{y} with length m and GC-content c .

Expectation formula for k -distance and k -neighbor match counts

In what follows, suppose the probe \mathbf{q} has length m and GC-content c , the target \mathbf{T} has length $n > m$, and set $n' = n - m + 1$. Let $\chi_{\mathbf{q}}^k(\mathbf{T})$ denote the number of k -distance matches between \mathbf{q} and \mathbf{T} , i.e., $\chi_{\mathbf{q}}^k(\mathbf{T})$ is the random variable corresponding to the total number of times \mathbf{q} occurs in \mathbf{T} with exactly k -base pair mismatches. We denote the

number of k -neighbor matches by $\theta_{\mathbf{q}}^k(\mathbf{T})$. Clearly, $\theta_{\mathbf{q}}^k(\mathbf{T}) = \sum_{j=0}^k \chi_{\mathbf{q}}^j(\mathbf{T})$. The expectation of $\chi_{\mathbf{q}}^k$ is given by the formula[†]

$$E[\chi_{\mathbf{q}}^k(\mathbf{T})] = n' g_k(m, \eta, c) \quad (6)$$

and the expectation of $\theta_{\mathbf{q}}^k(\mathbf{T})$ is given by

$$E[\theta_{\mathbf{q}}^k(\mathbf{T})] = n' G_k(m, \eta, c) \quad (7)$$

Thus, the expectation (or mean over a sample) of these match counts does not depend on the correlation of overlapping words. This correlation does affect the variance of match counts; however, and makes the calculation of the variance difficult. Calculation of the variance is beyond the scope of this paper (but see the discussion below).

We can also calculate the expectation for k -distance and k -neighbor match counts when the probe is not known beforehand. This result should be useful for estimating the average value of match counts in a target text \mathbf{T} over a sample of words drawn from a source text \mathbf{S} . Here, \mathbf{S} and \mathbf{T} are assumed to be strand-symmetric Bernoulli texts, with perturbation parameters η_S and η_T , respectively.

Let $\mathbf{Q} = \mathbf{S}[i, i + m - 1]$ denote the m -word in \mathbf{S} starting at i , and let $\chi_{\mathbf{Q}}^k$, resp. $\theta_{\mathbf{Q}}^k$, denote the k -distance match counts, resp. k -neighbor match counts, between the random texts \mathbf{Q} and \mathbf{T} . Then

$$E[\chi_{\mathbf{Q}}^k(\mathbf{T})] = n' \sum_{c=0}^m b_c(m, \hat{\rho}_S) g_k(m, \eta_T, c) \quad (8)$$

$$E[\theta_{\mathbf{Q}}^k(\mathbf{T})] = n' \sum_{c=0}^m b_c(m, \hat{\rho}_S) G_k(m, \eta_T, c) \quad (9)$$

where $\hat{\rho}_S = (1 + \eta_S)/2$.

Distribution of words by GC-content in some bacterial genomes

We now assess the accuracy of the strand-symmetric Bernoulli model in predicting the distribution of words of fixed length by GC-content in bacterial genomes. This will be done by means of a metric on probability measures which is defined as follows.

Recall that the space of probability measures Ξ_m on a finite set of cardinality $m + 1$ is an m -dimensional simplex in \mathbb{R}^{m+1} . A natural metric on Ξ_m is $d(\alpha, \beta) := \|\alpha - \beta\|/\sqrt{2}$, where $\|\cdot\|$ denotes Euclidean distance, and $\alpha, \beta \in \Xi_m$. The maximum possible distance between probability measures in this metric is 1. We use it to measure how close one probability measure is to another.

In the rest of this paper, we restrict our discussion to words that are 25 bases in length. Since the optimum length of oligomers in molecular probe design is a fiercely debated issue, it is noteworthy that we obtained results similar to

[†]For mathematical proofs and other details, see the technical report under supplementary information.

Table 1. Accuracy of the strand-symmetric Bernoulli model for some bacterial genomes and human chromosomes

Nucleotide sequence (T)	Abbrev.	$ T $	$ \xi_A^* - \xi_T^* $	$ \xi_C^* - \xi_G^* $	η	$d(b, \nu)$
<i>Clostridium perfringens</i>	<i>C.perf</i>	3 031 430	0.014898	0.009073	0.428679	0.024
<i>Escherichia coli K12-MG1655</i>	<i>E.coli</i>	4 639 221	0.000271	0.000573	-0.015778	0.025
<i>Haemophilus influenzae KW20</i>	<i>H.infl</i>	1 830 138	0.001848	0.001796	0.236994	0.014
<i>Mycobacterium leprae TN</i>	<i>M.lepr</i>	3 268 203	0.001593	0.003515	-0.155935	0.008
<i>Mycobacterium tuberculosis</i>	<i>M.tube</i>	4 403 836	0.000060	0.001481	-0.312188	0.005
<i>Neisseria meningitidis MC58</i>	<i>N.meni</i>	2 272 351	0.000638	0.004147	-0.030557	0.046
<i>Pseudomonas aeruginosa PAO1</i>	<i>Pa.aeru</i>	6 264 403	0.002743	0.005756	-0.331114	0.011
<i>Staphylococcus aureus Mu50</i>	<i>S.aure</i>	2 878 040	0.001981	0.001020	0.342368	0.008
<i>Streptococcus pneumoniae R6</i>	<i>S.pneu</i>	2 038 615	0.000776	0.001167	0.205680	0.017
<i>Streptococcus pyogenes M1</i>	<i>S.pyog</i>	1 852 441	0.003058	0.003639	0.229730	0.008
Human chromosome 1	H.chr1	256 422 225	0.000030	0.000011	0.165258	—
Human chromosome X	H.chrX	151 672 893	0.000462	0.000098	0.212066	—
Human chromosome Y	H.chrY	58 368 225	0.003901	0.001373	0.218692	—

Species name, its abbreviation, and sequence length are shown in the first three columns. The next two columns give the difference in A/T content, G/C content, respectively, and show the accuracy of Chargaff’s second parity rule. Perfect strand-symmetry would occur if both of these numbers were 0. The sixth column gives the corresponding perturbation parameter η , and the last column gives the distance $d(b, \nu)$ between the predicted and observed distribution of words in **T** by GC-content. Human chromosomes are included to show the apparent dependence of $|\xi_A^* - \xi_T^*|$ and $|\xi_C^* - \xi_G^*|$ on sequence length. Bacterial sequences were retrieved from the Genome Division of GenBank (Benson *et al.*, 2003). Human chromosome sequences were obtained from UC Santa Cruz at <http://genome.ucsc.edu/downloads.html>.

those discussed in this section for various word lengths, up to 100 bp[‡].

We denote the *relative GC-content* of a nucleotide sequence **T** by $\rho = \xi_C^* + \xi_G^*$. If **T** is indeed Bernoulli text, we expect the distribution of words in **T** to follow $b := \{b_c(m, \hat{\rho})\}_{c=0}^{25}$, where $\hat{\rho} = 1 - \rho = (1 + \eta)/2$. To determine the actual distribution of words by GC-content, we counted the number w_c of words of length $m = 25$ in **T** with GC-content c . This was done for all values of c in the range $0 \leq c \leq m$, and the resulting distribution was taken to be $\nu := \{\nu_c\}_{c=0}^{25}$, where $\nu_c = w_c/n'$. We then calculated the distance $d(b, \nu)$ for all bacteria in our study.

Relevant data for the complete bacterial genomes and extensively sequenced human chromosomes considered here are listed in Table 1. Note that the range of η is significant: $-0.331114 \leq \eta \leq 0.428679$. These bounds are close to the extremes observed in all sequenced bacterial genomes[§]. Representative results for the genome of *Pseudomonas aeruginosa* are shown in the inset in Figure 1. Note that, although $d(b, \nu)$ is small, a χ^2 test shows the deviation between b and ν to be statistically significant.

Predicted and observed k -distance match counts

We now show that Equation (6) gives a good approximation to the expected number of k -distance matches between a probe **q** and a target **T**. It follows that this expectation is sensitive to the GC-content of both **q** and **T**. The latter assertion is perhaps best illustrated by an example.

Suppose that **T** is strand-symmetric Bernoulli text with the same length and perturbation parameter as the *Staphylococcus aureus* genome. Suppose, also, that we have a probe **q** of length $m = 25$ with GC-content c . We want to see how the expected number of k -distance matches between **q** and **T** depends on k and c . First, let’s consider the distribution $g_k(m, \eta, c)$. The solid curves in Figure 1 show how $g_k(m, \eta, c)$ changes as a function of c ; the left-most curve corresponds to $g_k(m, \eta, 0)$ and the right-most curve to $g_k(m, \eta, m)$. This figure illustrates how $g_k(m, \eta, c)$ deviates from the standard binomial distribution (the dashed curve) when η is substantially different from 0. Now let’s consider some specific values for the expectation. Suppose that $k = 10$, then the expectation for $c = 25$ is $n' g_{10}(25, \eta, 25) \approx 3$, whereas $n' g_{10}(25, \eta, 0) \approx 12\,162$, where $n' = 2\,878\,026$. In this case, the standard binomial distribution gives an expectation of 493.

In order to compare the expectation formula with actual genome sequences, we tabulated k -distance matches between a given target and a representative sample of probes, and then computed the average. This was done by generating a collection of $m + 1$ sets of probes $\{Q_c\}_{c=0}^m$, where each set Q_c contains only words of length m with identical GC-content c . The words in each set Q_c were generated using *Mathematica* (Wolfram Research, Champaign, IL) as follows: first, a sequence $l_0 = \{1, \dots, 1, 0, \dots, 0\}$ of c consecutive 1’s followed by $m - c$ consecutive 0’s was generated. Then, given a sequence l_n with the same number of 1’s and 0’s, a new sequence l_{n+1} of this type was obtained by applying a random permutation to l_n . This process was applied recursively, starting with l_0 , to generate 10^4 random binary sequences. A uniform random number generator was then

[‡]See the *Mathematica* notebooks under supplementary information.

[§]See <http://www.cbs.dtu.dk/services/GenomeAtlas/Bacteria>

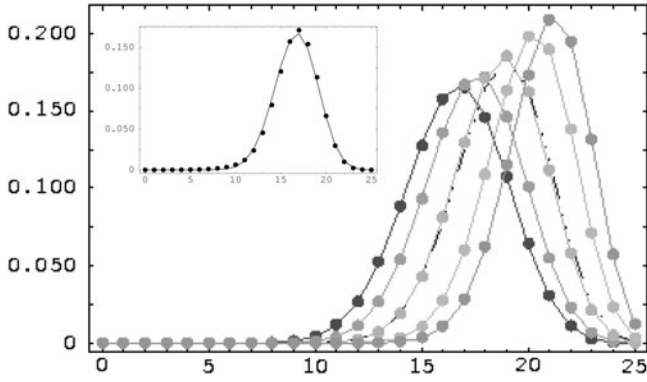


Fig. 1. Curves are instances of the perturbed binomial distribution $g_k(25, 0.342368, c)$ presented as a function of k . From left to right, they correspond to probes of GC-content $c = 0, 5, 12, 20, 25$. The dashed curve is $b_k(25, 1/4)$. Dots represent the averages \bar{v}_c for $S.aureus$ defined in the discussion on predicted and observed k -distance match counts. *Inset:* The curve shows $b_c(25, \hat{\rho})$ presented as a function of c . Here, $\hat{\rho} = (1 + \eta)/2$, where η is the perturbation parameter for $P.aeruginosa$. Dots represent the actual distribution ν of words in $P.aeruginosa$ by GC-content as described in the discussion on distribution of words by GC-content. The distance between these distributions is $d(b, \nu) \approx 0.011$.

used to substitute an A or T for an occurrence of 0 , and a C or G for an occurrence of 1 .

We then took the nucleotide sequence of a bacterial genome as our target text \mathbf{T} , and counted the number $n_c^k(\mathbf{q})$ of words of length m in \mathbf{T} with distance k from a particular $\mathbf{q} \in \mathcal{Q}_c$. This was done for each $0 \leq k \leq m$ and $\mathbf{q} \in \mathcal{Q}_c$. The average \bar{v}_c^k of $v_c^k(\mathbf{q}) := n_c^k(\mathbf{q})/n'$ over \mathcal{Q}_c was then calculated, where $n' = n - m + 1$, and n is the length of \mathbf{T} . Finally, we calculated the distance $d(\bar{v}_c, g_c)$, where $\bar{v}_c := \{\bar{v}_c^k\}_{k=0}^m$ and $g_c := \{g_k(m, \eta, c)\}_{k=0}^m$.

These calculations were carried out for $c = 0, 5, 12, 20, 25$, for the bacterial genomes listed in Table 1. The results in all cases were similar, and are illustrated by dots in Figure 1 for the case of $S.aureus$. The smallest distances in each case occurred at $c = 12$ (which corresponds to a GC-content in \mathbf{q} of roughly 50%), and increases as $|c - 12|$ increases. In particular, we found that $3.6 \times 10^{-5} \leq d(\bar{v}_{12}, g_{12}) \leq 2.1 \times 10^{-2}$, and that $3.5 \times 10^{-4} \leq d(\bar{v}_0, g_0), d(\bar{v}_{25}, g_{25}) \leq 2.1 \times 10^{-2}$.

The deviation of the sample mean \bar{v}_c^k from the expectation g_c^k in units of standard error is given by $|\bar{v}_c^k - g_c^k|/(s_c^k/100)$, where s_c^k is the sample standard deviation of v_c^k over \mathcal{Q}_c . In most cases[†], this deviation is statistically significant, ranging as high as 10^3 .

Extreme-value statistics for k -neighbor ratios

The theory developed so far provides us with a means of estimating the number of k -neighbor matches between a probe \mathbf{q} and a target sequence \mathbf{T} . We now describe a method for finding probes that are specific to a source genome \mathbf{S} . To that end,

we define the (*normalized source-target*) k -neighbor ratio of \mathbf{q} to be the quantity

$$\psi_{\mathbf{q}}^k(\mathbf{S}, \mathbf{T}) := \frac{|\mathbf{S}|}{|\mathbf{T}|} \cdot \frac{\theta_{\mathbf{q}}^k(\mathbf{T})}{\theta_{\mathbf{q}}^k(\mathbf{S})},$$

where $|\mathbf{U}|$ denotes the length of a sequence \mathbf{U} , and \mathbf{S} and \mathbf{T} are strand-symmetric Bernoulli texts. The factor $|\mathbf{S}|/|\mathbf{T}|$ is used to normalize the ratio, so that the difference in length of the two sequences does not affect its value.

We want to characterize extreme values of this ratio statistic over some sample set. A theoretical characterization would require knowledge of higher moments of the distribution of k -neighbor match counts. Since this is currently unavailable, we proceed empirically.

Let \mathcal{S} denote a sample set of oligonucleotides of fixed length m drawn from \mathbf{S} , let $\mathcal{I}_{\mathcal{S}}^k(\mathbf{S}, \mathbf{T}) := \inf\{\psi_{\mathbf{q}}^k(\mathbf{S}, \mathbf{T}) \mid \mathbf{q} \in \mathcal{S}\}$ denote the infimum of $\psi_{\mathbf{q}}^k(\mathbf{S}, \mathbf{T})$ over \mathcal{S} , and let $k_0 := \sup\{k \mid \mathcal{I}_{\mathcal{S}}^k(\mathbf{S}, \mathbf{T}) = 0\}$. A *source-specific probe* $\mathbf{q}_0 \in \mathcal{S}$ is, by definition, one that satisfies $\psi_{\mathbf{q}_0}^{k_0}(\mathbf{S}, \mathbf{T}) = 0$; we can think of k_0 as the distance from \mathbf{q}_0 to \mathbf{T} . If k_0 relatively large, we expect the cross-hybridization signals in (1) to be low. The word \mathbf{q}_0 is then specific to the source \mathbf{S} , and the signal strength will be a function of the number of k -neighbors of \mathbf{q}_0 in \mathbf{S} .

In our comparison of bacterial genomes, we chose \mathcal{S} to be a 1% uniform sample of the source genome \mathbf{S} , and set the word length to $m = 25$, as before. Table 2 shows that the quantity k_0 turns out to be fairly large with small variation ($7 \leq k_0 \leq 9$). These results might be improved upon by a search over all source-derived words, rather than just a sample.

It is also interesting to study the statistic $r_1 := \sup\{\mathcal{G}_{\mathcal{S}}^k(\mathbf{S}, \mathbf{T}) \mid 0 \leq k \leq m\}$, where $\mathcal{G}_{\mathcal{S}}^k(\mathbf{S}, \mathbf{T}) := \sup\{\psi_{\mathbf{q}}^k(\mathbf{S}, \mathbf{T}) \mid \mathbf{q} \in \mathcal{S}\}$. We define k_1 to be the value of k at which r_1 is assumed. It turns out[†] that $k_1 \leq k_0 + 1$, and $10 \leq r_1 \leq 300$ for the cases studied. Expectedly, r_1 depends on the difference of the perturbation parameters (or equivalently, GC-content) of the source and target. If the perturbation parameters for the source and target are far apart, the value of r_1 is rather large. The exception to this rule is the source-target pair ($E.coli, N.meningitidis$), for which r_1 is quite large even though the corresponding perturbation parameters are nearly equal. If \mathbf{q}_1 is a word for which $\psi_{\mathbf{q}_1}^{k_1}(\mathbf{S}, \mathbf{T}) = r_1$, then \mathbf{q}_1 will be more specific to the target \mathbf{T} than the source \mathbf{S} , even though it was derived from the source.

Search algorithm complexity and implementation

We have written a C++ program, implementing a dynamic-programming search algorithm similar to one described in Landau and Vishkin (1986), except that we did not allow insertions or deletions. We will refer to it as the *k-distance match* algorithm. The input consists of a source sequence \mathbf{S} , a target sequence \mathbf{T} , a probe of length m and a sampling step-size d . The corresponding output is a list of ordered pairs,

Table 2. Extreme-value statistics for source–target pairs of bacterial genomes

S	T	Genome length	GC content	Phylogenetic distance	$k_0(\mathbf{S}, \mathbf{T})$	$k_0(\mathbf{T}, \mathbf{S})$	η_S	η_T
<i>S.aure</i>	<i>C.perf</i>	+	+	+	8	8	0.342368	0.428679
<i>S.pyog</i>	<i>S.pneu</i>	+	+	+	8	8	0.229730	0.205680
<i>M.tube</i>	<i>M.lepr</i>	–	+	+	7	8	–0.312188	–0.155935
<i>S.pyog</i>	<i>S.aure</i>	–	+	+	8	8	0.229730	0.342368
<i>Paeru</i>	<i>H.infl</i>	–	–	+	9	8	–0.331114	0.236994
<i>S.pyog</i>	<i>H.infl</i>	+	+	–	8	8	0.229730	0.236994
<i>N.meni</i>	<i>E.coli</i>	–	+	–	7	8	–0.030557	–0.015778
<i>S.aure</i>	<i>Paeru</i>	–	–	–	8	9	0.342368	–0.331114
<i>S.pyog</i>	<i>Paeru</i>	–	–	–	8	9	0.229730	–0.331114

The first two columns are source–target pairs, as described in the discussion on extreme-value statistics. Abbreviations for species’ names are as in Table 1. The next three columns compare some characteristics of \mathbf{S} and \mathbf{T} . Similarity in a characteristic is denoted with a +, and the corresponding criteria are as follows: genome length is + if the length of the smaller genome is at least 90% of the larger one; GC-content is + if $|\eta_S - \eta_T| < 1/4$; phylogenetic distance is + if the two bacteria are in the same subdivision. The sixth and seventh columns give the corresponding k_0 statistic for a source sample of 1%. Note that $k_0(\mathbf{S}, \mathbf{T})$ is fairly large relative to the probe size of 25, and that it is not sensitive to differences in structure and phylogeny. The perturbation parameters η_S , resp., η_T for \mathbf{S} , resp., \mathbf{T} are included for convenience.

indexed by $0 \leq l \leq |\mathbf{S}|/d$, containing a location dl in \mathbf{S} and a sublist of $m + 1$ numbers $\{n_0, \dots, n_m\}$ where n_k is the number of k -distance matches between $\mathbf{S}[dl, \dots, dl + m - 1]$ and \mathbf{T} . An explicit list of locations in the target where a k -distance match occurs together with the k -distance match count and the matching strings in the source and target can also be produced by the program. The latter option requires a substantial increase in memory as the number of k -distance matches becomes large.

We now give pseudocode for a simplified version of the k -distance match algorithm. The actual code uses modular arithmetic to reduce memory usage, but including the details would only obfuscate the basic idea. First, define $\varepsilon: \mathcal{A} \times \mathcal{A} \rightarrow \{0, 1\}$ by the rule $\varepsilon(x, y) = 1$ if $x = y$, and $\varepsilon(x, y) = 0$ otherwise. We further define V to be an m -vector of integers, and L to be an $m \times n$ -matrix of integers. We will call L the *scoring matrix*. For notational convenience, let $\mathbf{q} = \mathbf{S}[dl, \dots, dl + m - 1]$ for some l . The arrays are then initialized as follows: $V[j] = 0, 0 \leq j < m$; $L[i][0] = \varepsilon(\mathbf{q}[i], \mathbf{T}[0]), 0 \leq i < m$; $L[0][j] = \varepsilon(\mathbf{q}[0], \mathbf{T}[j]), 0 \leq j < m$. The main loop of our algorithm is then

```

for  $1 \leq j \leq n$  do
  for  $1 \leq i < m$  do
     $L[i][j] \leftarrow L[i - 1, j - 1] + \varepsilon(\mathbf{q}[i], \mathbf{T}[j])$ 
  if  $m \leq j$  then
     $V[L[i][j]] \leftarrow V[L[i][j]] + 1$ 

```

When execution is complete, we will have $V[j] = \chi_{\mathbf{q}}^j(\mathbf{T})$. This process is repeated for each l in the range $0 \leq l \leq |\mathbf{S}|/d$.

We now add a word about the ‘post-processing’ phase, which was done with *Mathematica*. To find $k_0(\mathbf{S}, \mathbf{T})$ for a given source–target pair, we ran the process described above twice: first with \mathbf{S} as the target, and second with \mathbf{T} as the target. (The source in both cases is \mathbf{S} .) The results were imported into

Mathematica and used to construct a list of ratios $\psi_{\mathbf{q}}^k(\mathbf{S}, \mathbf{T})$ for all $\mathbf{q} \in \mathcal{S}$. Then $k_0(\mathbf{S}, \mathbf{T})$ was computed using built-in *Mathematica* functions.

The heart of the k -distance match algorithm is a character-by-character comparison (represented here by the function ε), hence the time complexity of a single probe search is $O(m \cdot |\mathbf{T}|)$. The time complexity of a search over the entire sample set of probes is $O(m \cdot (|\mathbf{S}|/d) \cdot |\mathbf{T}|)$. An $m \times 2$ matrix was used to store the dynamic array data, hence the space complexity is $O(2m + |\mathbf{S}| + |\mathbf{T}|)$. The search takes approximately 1 h on a single CPU of a Compaq GS80 server at $m = 25, |\mathbf{S}|/d = 10^4$ and $|\mathbf{T}| = 10^6$.

A faster serial algorithm (Gusfield, 1997, p. 200) has the best-case time complexity $O(k \cdot |\mathbf{T}|)$ for a single probe search. This is an advantage if one is only interested in values of $k \ll m$. Since we were interested in tabulating all the k -distance match counts for $0 \leq k \leq m$, we were unable to exploit this advantage.

CONCLUDING REMARKS AND FURTHER DIRECTIONS

We have studied aspects of word statistics in genomic sequences relevant to the problem of molecular probe design discussed in the introduction. What follows is a summary of related research that may prove useful in extending our results.

Lippert *et al.* (2002) study the distribution of exact matches of m -words between two sequences, which they call the D_2 statistic. This statistic is essentially $\sum_{\mathbf{Q} \in \mathcal{S}} \chi_{\mathbf{Q}}^k(\mathbf{T})$ in our notation, with $k = 0$. The authors derive a formula for the expectation of D_2 , give an upper bound for the variance, and show that, when $m < \log(n)/6$, D_2 is approximately a normal distribution, and when $m > 2 \log(n)$, it’s approximately a compound Poisson distribution. For $n = 10^6$ (the order of magnitude of the size of bacterial genomes), these asymptotic

regimes are given by $m \leq 2$ and $m \geq 28$, respectively. Thus, the latter approximation may apply to the distribution of exact m -word matches between two bacterial genomes for $m \geq 28$. It may be possible to extend the approach in Lippert *et al.* (2002) to the k -distance and k -neighbor match counts considered here, thus giving us a way to estimate their variance.

The expectation formula (6) can be found in Régnier and Szpankowski (1998), except that the probability is not specified there. In addition, an exact formula for the variance, and approximate formulas for the distribution of match counts are given. Their approach is based on a characterization of approximate pattern occurrences by means of formal languages; they use probability-generating functions over these languages to calculate moments. Régnier (2000) extends these results to the Markov case, and considers both overlapping and non-overlapping match counts. The formula for the variance of the match count statistic in Régnier and Szpankowski (1998) is computationally intractable, in general, because it requires calculating a correlation matrix that grows very rapidly with the length of the probes. For example, the matrix corresponding to all 10-neighbors of a given 25-word has over 5×10^{13} entries. We have also observed that (6) can be derived from an application of the ergodic theorem to the Bernoulli shift operator described in Billingsley (1965).

This study suggests that the distribution of words of moderate length in bacterial genomes can be modeled by Bernoulli text. In contrast, Pevzner *et al.* (1989a,b); Prum *et al.* (1995) invoked Markov models to describe the distribution of short words (2–6 nt). A comparison of the Bernoulli and Markov models over a range of word sizes may therefore be interesting, and provide new insights.

Strand-symmetry in nucleotide sequences seems to follow the law of large numbers, i.e. when $|\mathbf{T}| \rightarrow \infty$, both $|\xi_A^*(\mathbf{T}) - \xi_T^*(\mathbf{T})| \rightarrow 0$ and $|\xi_C^*(\mathbf{T}) - \xi_G^*(\mathbf{T})| \rightarrow 0$ (Table 1). For example, the human chromosome 1 is about 50 times longer than most bacterial genomes, and the corresponding differences in the relative frequencies is significantly smaller.

We assumed the distribution of nucleotides in a sequence \mathbf{T} to be homogeneous. This is reported to be the case for prokaryotes and lower eukaryotes (Forsdyke, 2002), although prokaryotes also exhibit some variation in GC-content, GC-skew $[\xi_C^*(\mathbf{T}) - \xi_G^*(\mathbf{T})]/[\xi_C^*(\mathbf{T}) + \xi_G^*(\mathbf{T})]$, and AT-skew $[\xi_A^*(\mathbf{T}) - \xi_T^*(\mathbf{T})]/[\xi_A^*(\mathbf{T}) + \xi_T^*(\mathbf{T})]$.

These skews (Jensen *et al.*, 1999), which measure the deviation of \mathbf{T} from strand-symmetry, are typically within 4%, and are not correlated with GC-content. They are, however, correlated with the direction of replication and/or gene orientation (Baisnée *et al.*, 2002). Isochores, i.e. regions with a characteristic GC-content, have been observed in higher eukaryotic genomes (Bernardi, 1993; Forsdyke, 2002). These facts suggest modeling each isochores separately, and studying the effect of GC-skew and AT-skew on the accuracy of our model.

The nearest-neighbor model is widely used to describe the thermodynamics of hybridization of nucleic acids (SantaLucia, 1998; Kaderali and Schliep, 2002), and the free energies of dinucleotide hybridization at 37°C given in SantaLucia (1998) could be used to estimate the free energy of oligonucleotide hybridization. Hence, it is possible to derive a hybridization energy profile along the source and target of a given probe. Ultimately, this should allow us to estimate the signals σ_{ij} defined in 1 and thereby determine the specificity of a probe derived by our method.

Little is known about the relationship between the oligonucleotide probes which have low k -neighbor match counts in the target, and regions in the source genome that encode such highly source-specific probes. In a preliminary study, we noticed that protein-coding genes and intergenic regions in the source genome are equally likely to give rise to such probes, but, among the protein-coding genes, there seems to be a bias toward certain functional classes of genes (unpublished data). Understanding of evolutionary and functional constraints on gene sequences may facilitate better selection of molecular probes for molecular diagnostics and gene therapy.

ACKNOWLEDGEMENTS

The authors thank Shouguang Jin and Weihong Tan for introduction to the exciting world of molecular beacons, Mike Coleman and Earl Glynn for valuable discussions on algorithm design and for programming help, and Ognen Duzlevski for source code review.

REFERENCES

- Baisnée, P.F., Hampson, S. and Baldi, P. (2002) Why are complementary DNA strands symmetric? *Bioinformatics*, **18**, 1021–1033.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2003) GenBank. *Nucleic Acid Res.*, **31**, 23–27.
- Bernardi, G. (1993) The vertebrate genome: isochores and evolution. *Mol. Biol. Evol.*, **10**, 186–204.
- Billingsley, P. (1965) *Ergodic Theory and Information*. John Wiley & Sons, Inc., New York.
- Bonnet, G., Tyagi, S., Libchbaer, A. and Kramer, F.R. (1999) Thermodynamic basis of the enhanced specificity of structured DNA probes. *Proc. Natl Acad. Sci. USA*, **96**, 6171–6176.
- Fickett, J.W., Torney, D.C. and Wolf, D.R. (1992) Base compositional structure of genomes. *Genomics*, **13**, 1056–1064.
- Forsdyke, D.R. (2002) Symmetry observations in long nucleotide sequences: a commentary on the discovery note of Qi and Cuticchia, **18**, 215–217.
- Gusfield, D. (1997) *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, Cambridge.
- Jensen, L.J., Friis, C. and Ussery, D.W. (1999) Three views of microbial genomes. *Res. Microbiol.*, **150**, 773–777.
- Landau, G.M. and Vishkin, U. (1986) Efficient Parallel and Serial String Matching. *Computer Science Department Technical Report 221*, Tel Aviv University.
- Lippert, R.A., Huang, H. and Waterman, M.S. (2002) Distributional regimes for the number of k -word matches between

- two random sequences. *Proc. Natl Acad. Sci. USA*, **99**, 13980–13989.
- Kaderali,L. and Schliep,A. (2002) Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, **18**, 1340–1349.
- Pevzner,P.A. (2000) *Computational Molecular Biology, An Algorithmic Approach*. MIT Press, Cambridge, MA.
- Pevzner,P.A., Borodovsky,M.U. and Mironov,A.A. (1989a) Linguistics of nucleotide sequences I: the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J. Biomol. Struct. Dyn.*, **6**, 1013–1026.
- Pevzner,P.A., Borodovsky,M.U. and Mironov,A.A. (1989b) Linguistics of nucleotide sequences II: stationary words in genetic texts and the zonal structure of DNA. *J. Biomol. Struct. Dyn.*, **6**, 1027–1038.
- Pozhitkov,A.E. and Tautz,D. (2002) An algorithm and program for finding sequence specific oligo-nucleotide probes for species identification. *BMC Bioinformatics*, **3**.
- Prum,B., Rodolphe,F. and de Truckheim,E. (1995) Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *J. R. Statist. Soc. B*, **57**, 205–220.
- Régnier,M. (2000) A unified approach to word occurrence probabilities. *Discrete Appl. Math. (Special Issue on Computational Biology)*, **104**, 259–280.
- Régnier,M. and Szpankowski,W. (1998) On the approximate pattern occurrences in a text. *Compression and Complexity of SEQUENCES 1997*, IEEE Computer Society, pp. 253–264.
- SantaLucia,J., Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.