DEDUCING STRINGS FROM SUB-STRING COUNTS – A PROBLEM IN COMPUTATIONAL BIOLOGY

by

Martin Dyer
Alan Frieze
Stephen Suen

Department of Mathematics
Carnegie Mellon University
Pittsburgh, PA 15213

Research Report No. 93–158

October 1993

Deducing strings from sub-string counts - a problem in computational biology

Martin Dyer* Alan Frieze†
Stephen Suen
Department of Mathematics,
Carnegie Mellon University,
Pittsburgh PA15213, U.S.A.

October 12, 1993

1 Introduction

The following is an abstraction of a problem occuring in the sequencing of (fragments of) DNA molecules.

Let Σ be a fixed alphabet with s letters, and ξ be a string chosen uniformly at random from Σ^m, where m is an integer. Let $\ell = \lfloor \log_s(m^2/2c) \rfloor$, where $c > 0$ is a constant. For each $\sigma \in \Sigma^{\leq m}$ let $N(\sigma, \xi)$ denote the number of occurrences of σ as a sub-string of ξ. Let $\mathcal{N}_\ell(\xi) = (N(\sigma_1, \xi), N(\sigma_2, \xi), \ldots, N(\sigma_\tau, \xi))$ where $\tau = s^\ell$ and $\sigma_1, \sigma_2, \ldots, \sigma_\tau$ is some enumeration of Σ^ℓ.

*Research done while visiting Carnegie Mellon University in Spring 1993. Permanent address: School of Computer Studies, University of Leeds, Leeds, U.K.
†Supported by NSF grant CCR-9002435 and CCR9225008.

We say that ξ is $\ell - recoverable$ if $\xi' \in \Sigma^m, \xi \neq \xi'$ implies $\mathcal{N}_\ell(\xi) \neq \mathcal{N}_\ell(\xi')$.

Our main result is

Theorem 1

$$\lim_{m\to\infty} \mathbf{Pr}(\xi \textit{ is } \ell\textit{-recoverable}) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}(2\lambda)^k}{k!(k+1)!},$$

where $\lambda = (s-1)c$.

As explained later it is easy to tell whether ξ is ℓ-recoverable and recover ξ from $\mathcal{N}_\ell(\xi)$ if it is.

It is of some interest to compare the result of Theorem 1 with the following *information theoretic* lower bound. Since $|N(\sigma,\xi)| \leq m$ for all σ, we see that there are at most m^τ different values of $\mathcal{N}_\ell$. Thus to have a significant number of ℓ-recoverable strings we need $m^\tau \geq s^m$ or $\tau \geq m/\log_s m$, and the theorem tells us that this lower bound is approximately the square root of the real answer.

We now explain the relevance of this result to sequencing DNA fragments. First of all, a chromosome can be thought of as a string of some 10^8 letters over the alphabet of nucleotides {A,G,C,T}. The primary aim of the human genome project is to determine the strings defined by human chromosomes.

The method of Sequencing by Hybridization (Bains and Smith [2], Lysov et al [6], Drmanac et al [3], Pevzner et al [9], Pevzner [7]) involves a two-dimensional matrix of immobilised oligonucleotides (short strings, length ℓ). Once a DNA fragment ξ is *hybridized* with the matrix one can determine which ℓ-tuples occur. With great difficulty one can perhaps tell if an ℓ-tuple occurs more than once. One hopes that this is enough information to

determine ξ exactly. Our theorem shows that the number of oligonucleotides needs to grow like m^2 in order for there to be any reasonable chance of this to be true. It is interesting to note that if ℓ, m are such that there is a reasonable chance of reconstruction by this method, then it is unlikely that any string appears three or more times. Thus one could reasonably replace more than once by two.

See Alizadeh, Karp, Newberg and Weisser [1] or Karp [5] for surveys of computational problems related to DNA sequencing.

2 Proof of Theorem 1

Given $\mathcal{N}_\ell$ we can define a (multi-)digraph $G = G(\mathcal{N}_\ell)$ as follows: the vertex set of G is $[s]^{\ell-1}$ and if $x = x_1x_2\dots x_{\ell-1}, y = y_1y_2\dots y_{\ell-1}$ then there is no edge (x,y) unless $x_2 = y_1, x_3 = y_2, \dots, x_{\ell-1} = y_{\ell-2}$ in which case there are precisely $N(x_1x_2\dots x_{\ell-1}y_{\ell-1}, \xi)$ edges from x to y. As already observed by Pevzner [7], ξ is ℓ-recoverable if and only if G has a unique Euler path, up to the order in which parallel edges are traversed. We will find the limiting probability that this is the case. We first show that **whp** (i.e. with probability 1-o(1) as $m \to \infty$) no vertex of G has out-degree 3 or more and so G is rather simple.

Lemma 1 *Let ξ be chosen randomly from Σ^m. Let $\mathcal{E}_0$ be the event*

$$\{\exists \zeta \in [s]^{\ell-1} : N(\zeta, \xi) \geq 3\}.$$

Then

$$\mathbf{Pr}(\mathcal{E}_0) = o(1).$$

Proof If $\xi = \xi_1\xi_2\ldots\xi_m$ let $\xi[i,j] = \xi_i\xi_{i+1}\ldots\xi_j$ for $1 \le i \le j \le m$. Let $\mathcal{E}_{i,j,k}$ denote the event $\{\xi[i,i+\ell-2] = \xi[j,j+\ell-2] = \xi[k,k+\ell-2]\}$ for $(i,j,k) \in I = \{(i,j,k) : 1 \le i < j < k \le m-\ell+2\}$. Now divide I into $I_1 = \{(i,j,k) \in I : \max\{j-i, k-j\} > \ell-2\}$ and $I_2 = I \setminus I_1$. If $(i,j,k) \in I_1$ then $\mathbf{Pr}(\mathcal{E}_{i,j,k}) = s^{-2(\ell-1)}$ and if $(i,j,k) \in I_2$ then $\mathbf{Pr}(\mathcal{E}_{i,j,k}) \le s^{-\ell+1}$ suffices. Clearly $|I_2| = O(m\ell^2)$ and so

$$\begin{aligned}
\mathbf{Pr}(\mathcal{E}_0) &\le \sum_{(i,j,k)\in I_1} \mathbf{Pr}(\mathcal{E}_{i,j,k}) + \sum_{(i,j,k)\in I_2} \mathbf{Pr}(\mathcal{E}_{i,j,k}) \\
&= O(m^3 s^{-2\ell}) + O(m\ell^2 s^{-(\ell-1)}) \\
&= O(m^{-1}) + O((\log m)^2/m) \\
&= o(1).
\end{aligned}$$

□

For each pair of positions $1 \le i < j \le m-\ell+2$ on ξ, let $I_{i,j}$ be the indicator for the event $\{\xi[i,i+\ell-2] = \xi[j,j+\ell-2$ and $(i = 1 \vee (\xi_{i-1} \ne \xi_{j-1})\}$. Write X as the sum of these indicator functions. Then

$$\begin{aligned}
\mathbf{E}[X] &= (m-\ell+1)s^{-(\ell-1)} + \binom{m-\ell+1}{2} s^{-(\ell-1)}(1-s^{-1}) \qquad (1) \\
&\approx \frac{m^2}{2} s^{-\ell}(s-1) \\
&\approx (s-1)c.
\end{aligned}$$

The first term in the RHS of (1) corresponds to $i = 1$ and the second to $i \ge 2$.

We would next like to prove a Poisson limit theorem for X. The following lemma provides the basis for subsequent calculations:

Lemma 2 *A pair of indices will be denoted by* $u = (i_u, j_u)$ *where* $i_u < j_u$. *Let* $\mathcal{A} = \{u : j_u - i_u > 5\ell\}$. *(5 is taken for convenience rather than minimality.)*

(a) Let

$$\mathcal{E}_1 = \{\exists u : \xi[i_u, i_u + 2\ell] = \xi[j_u, j_u + 2\ell]\},$$

$$\mathcal{E}_2 = \{\exists u \in \bar{\mathcal{A}} : I_u = 1\},$$

and

$$\mathcal{E}_3 = \{I_{1,m-\ell+2} = 1\}.$$

Then $\mathbf{Pr}(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3) = o(1)$.

(b) For $u, v \in \mathcal{A}$ *with* $u \neq v$, $\mathbf{E}(I_u I_v) \leq s^{-2(\ell-1)}$.

Proof (a)

$$\begin{aligned} \mathbf{Pr}(\mathcal{E}_1) &\leq m^2 s^{-2\ell} \\ &= o(1). \end{aligned}$$

Since there are fewer than $5m\ell$ pairs such that $j_u - i_u \leq 5\ell$ we have

$$\begin{aligned} \mathbf{Pr}(\mathcal{E}_2) &\leq 5m\ell s^{-\ell} \\ &= o(1). \end{aligned}$$

Clearly $\mathbf{Pr}(\mathcal{E}_3) = s^{-\ell} = o(1)$.

(b) We show that $\mathbf{Pr}(I_v = 1 \mid I_u = 1) \leq s^{\ell-1}$ and deduce the result from $\mathbf{E}(I_u I_v) = \mathbf{Pr}(I_v = 1 \mid I_u = 1)\mathbf{Pr}(I_u = 1)$.

Assume $i_v \geq i_u$. Note first that if $i_v - i_u = j_v - j_u > 0$ then either I_u and I_v are independent or $I_v I_u = 0$. For in the latter case, if $I_u = 1$ then $\xi_{j_v-1} = \xi_{i_v-1}$ which implies $I_v = 0$.

Condition on $I_u = 1$ and let $\mathcal{B}_k = \{\xi_{i_v+k} = \xi_{j_v+k}\}, 0 \le k \le \ell - 2$. Suppose first that $j_v + k \notin [j_u, j_u + \ell - 2]$. Then $\mathcal{B}_k$ is independent of I_u and $\mathcal{B}_{k'}, k' \neq k$ and $\mathbf{Pr}(\mathcal{B}_k) = s^{-1}$.

On the other hand let $K = \{k : j_v + k = j_u + k^* \in [j_u, j_u + \ell - 2]\}$. Suppose $k \in K$ and $I_u = 1$. Then

$$\begin{aligned} \mathbf{Pr}(\mathcal{B}_k \mid I_u) &= \mathbf{Pr}(\xi_{j_u+k^*} = \xi_{i_v+k} \mid \xi_{i_u+k} = \xi_{j_u+k^*}) \\ &= s^{-1}. \end{aligned}$$

Also, as k runs through K, k^* runs through distinct values. Hence the events $\mathcal{B}_k, k \in K$ are also conditionally independent. □

Let $X' = \sum_{u \in \mathcal{A}} I_u$. Then Lemma 2(a) and its proof show that

$$X' = X \qquad \textbf{whp}$$

and

$$\mathbf{E}(X') = \mathbf{E}(X) + o(1).$$

For $u \in \mathcal{A}$, write $p_u = \mathbf{E}[I_u]$. Then using a theorem of Suen [10] which is similar to a theorem in Suen [11], we have for any $\theta \in [0, 1]$,

$$\left| \mathbf{E}[\prod_{i \in \mathcal{A}} (1 - \theta I_u)] - \prod_{i \in \mathcal{A}} (1 - \theta p_u) \right| \le \prod_{i \in \mathcal{A}} (1 - \theta p_u) \left(\exp\left(\sum_{e \in \mathcal{H}} y(\theta, e) \right) - 1 \right),$$

where $\mathcal{H}$ is the set of pairs $u, v \in \mathcal{A}$, with $u \neq v$, such that I_u and I_v are not independent (or simply, $\mathbf{E}[I_u I_v] \neq \mathbf{E}[I_u]\mathbf{E}[I_v]$), and for $e = \{u, v\}$,

$$y(\theta, e) = 2\theta^2 (\mathbf{E}[I_u I_v] + p_u p_v) \prod_{w \in N(u,v)} (1 - \theta p_k)^{-1},$$

with $N(u, v)$ equal to the set $w \in \mathcal{A}$ such that I_w is not independent of I_u or I_v. Note first that

$$p_u \le s^{-(\ell-1)},$$

and

$$|N(u,v)| \leq 2m\ell.$$

Thus,

$$\prod_{w \in N(u,v)} (1-\theta p_w)^{-1} = 1 + o(1),$$

uniformly for all (u,v). Also, it is clear that $|\mathcal{H}| \leq 2m^3\ell$, and so

$$\sum_{\{u,v\} \in \mathcal{H}} p_u p_v = o(1).$$

Also Lemma 2(b) shows

$$\begin{aligned} \sum_{\{u,v\} \in \mathcal{H}} \mathbf{E}[I_u I_v] &\leq 2m^3 \ell s^{-2(\ell-1)} \\ &= o(1), \end{aligned} \tag{2}$$

in which case

$$\left| \mathbf{E}[\prod_{i \in \mathcal{A}} (1-\theta I_u)] - \prod_{i \in \mathcal{A}} (1-\theta p_u) \right| = o(1), \qquad \theta \in [0,1].$$

Since

$$\mathbf{E}[\prod_{i \in \mathcal{A}} (1-\theta I_u)] = \mathbf{E}[(1-\theta)^{X'}],$$

it follows that X', and hence X, converges in distribution to a Poisson variable with parameter $\lim_{m \to \infty} \mathbf{E}[X'] = \lambda = (s-1)c$.

We now assume that $\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3$ does not occur and that there are k pairs of maximal common substrings in ξ of lengths at least $\ell - 1$. We may also assume that $I_u I_v = 0$ for $\{u,v\} \in \mathcal{H}$ (see (2)). Thus the common substrings of length at least $\ell - 1$ will not overlap each other. Let $\mathcal{E}$ be the union of events we have so far excluded. We may regard these k pairs of substrings as

pairs of labelled markers $m_1, m_2, \ldots, m_k$ on ξ. Let $\mathcal{F}$ be the event that there are two pairs of markers which occur as $\ldots, m_a, \ldots m_b, \ldots, m_a, \ldots, m_b, \ldots$ in the order from left to right of ξ. Note that if $\mathcal{F}$ occurs, then it is not possible to determine the order of the two (necessarily different) substrings in ξ between the two occurrences of m_a and m_b.

Pevzner [8] has shown that if neither $\mathcal{E}_3$ nor $\mathcal{F}$ occur then ξ is ℓ-recoverable (proving a conjecture of Ukkonen [12]).

We next need to find the probability of $\mathcal{F}$ given k pairs of markers. There are $(2k)!/2^k$ distinct orderings of the markers $m_1, m_2, \ldots, m_k$. Let C_k denote the Catalan number giving the number of well formed strings of k parentheses (,) (see for example Graham, Knuth, Patashnik [4]). There are $k!C_k$ ways of placing the markers so that $\mathcal{F}$ does not occur. To see this map a sequence of markers in which $\mathcal{F}$ does not occur into a sequence of parentheses by replacing the first occurrence of an m_i by a (and the second occurrence by a). If $\mathcal{F}$ does not occur then the sequence of (,)'s is well formed. This is easily proved by induction on k where the inductive step involves removing an innermost repeatred pair. Conversely, given a well formed sequence of (,)'s, one can produce $k!$ sequences of the markers in which $\mathcal{F}$ does not occur. Here we assign markers to parentheses so that if (is assigned m_a then the) receiving m_a must appear later in the sequence. This is again easily proved by induction on k. The inductive step involves looking at an innermost pair (,). If this is assigned a pair m_a, m_a then we use induction. If this is assigned $m_a, m_b, a \neq b$ then the other m_a must follow and the other m_b must precede these two, causing $\mathcal{F}$ to occur.

Thus, the probability of $\bar{\mathcal{F}}$ conditional on having k pairs of markers is

$$k!\binom{2k}{k}\frac{1}{k+1}\frac{2^k}{(2k)!} = \frac{2^k}{(k+1)!}.$$

Hence,

$$\begin{aligned}
\mathbf{Pr}(\xi \text{ is } \ell\text{-recoverable}) &= \sum_{k=0}^{\infty} \mathbf{Pr}(\bar{\mathcal{F}}|X=k,\bar{\mathcal{E}})\mathbf{Pr}(X=k,\bar{\mathcal{E}}) + O(\mathbf{Pr}(\mathcal{E})) \\
&= \sum_{k=0}^{\infty} \frac{2^k}{(k+1)!}\mathbf{Pr}(X=k) + o(1) \\
&= \sum_{k=0}^{\infty} \frac{2^k}{(k+1)!}\mathbf{Pr}(X'=k) + o(1)).
\end{aligned}$$

Since the moment generating function of X' converges to that of a Poisson variable with parameter $\lambda = (s-1)c$, it follows that

$$\mathbf{Pr}(\xi \text{ is } \ell\text{-recoverable}) \to \sum_{k=0}^{\infty} \frac{\mathrm{e}^{-\lambda}(2\lambda)^k}{k!(k+1)!}.$$

This completes the proof of Theorem 1.

Remark: the above result can be generalised to non-uniform sampling. Suppose $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_s\}$ and let ξ be generated one symbol at a time, with each symbol chosen independently of previous symbols. Let $\mathbf{Pr}(\xi_j = \sigma_i) = p_i, 1 \leq i \leq s$, and $1 \leq j$. We ignore the trivial case in which there is an i such that $p_i = 1$. Suppose $\alpha = p_1^2 + p_2^2 + \cdots + p_s^2$ and $\beta = p_1^3 + p_2^3 + \cdots + p_s^3$. Then the previous analysis can be pushed through with λ (in the statement of Theorem 1) replaced by $(\alpha^{-1} - 1)c$. (s in the RHS of (1) is replaced by α^{-1} and the RHS of (2) becomes $O(m^3\ell\sum_{j=1}^{\ell}\alpha^{2(\ell-j)}\beta^j) = O(m^3\ell^2(\alpha^{2\ell} + \beta^\ell)) = o(1)$ since $\beta < \alpha^{1.5+\epsilon}$ for some fixed $\epsilon > 0$.)

Acknowledgement: We would like to thank Pavel Pevzner for his valuable comments.

References

[1] F.Alizadeh, R.M.Karp, L.A.Newberg and D.K.Weisser, *Physical mapping of chromosomes: a combinatorial problem in molecular biology*, Proceedings of 4'th Annual ACM-SIAM Symposium on Discrete Algorithms (1993) 371-381.

[2] W.Bains and G.C.Smith, *A novel method for nucleic acid sequence determination*, Journal of Theoretical Biology 135 (1988) 303-307.

[3] R.Dramanac, L.Labat, I Brukner and R. Crkvenjakov, *Sequencing of megabase plus DNA by hybridization: theory of a method*, Genomics 4 (1989) 114-?.

[4] R.L.Graham, D.E.Knuth and O.Patashnik, *Concrete Mathematics*, Addison-Wesley, 1989.

[5] R.M.Karp, *Mapping the genome: some combinaatorial problems arising in molecular biology*, Proceedings of the 25'th Annual ACM Symposium on Theory of Computing (1993) 278-285.

[6] Y.P.Lysov, V.L.Florentiev, A.A.Khorlin, K.R.Khrapko, V.V.Shik and A.D.Mirzabekov, *???*, Doklady Acad. Sci. of USSR 303 (1988) 1508-1511.

[7] P.A.Pevzner, *L-tuple DNA sequencing: computer analysis*, Journal of Biomolecular Structure and Dynamics 7 (1989) 63-73.

[8] P.A.Pevzner, DNA physical mapping and Eulerian cycles in coloured graphs, to appear.

[9] P.A.Pevzner, Y.P.Lysov, K.R.Khrapko, A.V.Belyavsky, V.L.Florentiev and A.D.Mirzabekov, *Improved chips for sequencing by hybridization*, Journal of Biomolecular Structure and Dynamics 9 (1991) 399-410.

[10] W.C.S. Suen, *Ph.D. Dissertation*, University of Bristol, 1985.

[11] S.Suen, *A correlation inequality and a Poisson limit theorem for nonoverlapping balanced subgraphs of a random graph*, Random Structures and Algorithms 1 (1990) 231-242.

[12] E.Ukkonen, *Approximate string matching with q-grams and maximal matching*, Theoreetical Computer Science 92 (1992) 191-211.