Sequence Determination from Overlapping Fragments: A Simple Model of Whole-Genome Shotgun Sequencing

Bernard Derrida* and Thomas M. A. Fink†

Laboratoire de Physique Statistique, Ecole Normale Supérieure, 75231 Paris Cedex 05, France and Institute for Theoretical Physics, University of California, Santa Barbara, California 93106-4030

(Received 3 July 2001; published 28 January 2002)

Assembling fragments randomly sampled from along a sequence is the basis of whole-genome shotgun sequencing, a technique used to map the DNA of the human and other genomes. We calculate the probability that a random sequence can be recovered from a collection of overlapping fragments. We provide an exact solution for an infinite alphabet and in the case of constant overlaps. For the general problem we apply two assembly strategies and give the probability that the assembly puzzle can be solved in the limit of infinitely many fragments.

DOI: 10.1103/PhysRevLett.88.068106 PACS numbers: 87.14.Gg, 87.15.Aa

The problem of sequencing the human and other genomes is at the origin of a large number of technical and theoretical challenges. Considerable progress in practical sequencing techniques has allowed the recent completion of the human genome [1,2]. Theoretical effort has focused on, among other things, sequence alignment [3], long-range correlations [4], and distinguishing between coding and noncoding regions [5]. Many of these questions can be formulated in terms of random spin chains or systems with quenched disorder and have therefore created a lot of interest among physicists [6,7].

Here we study a simple model related to the problem of reconstructing an unknown sequence by assembling fragments. We consider a random, uncorrelated sequence S of length L, composed of characters drawn from an alphabet of size A. This sequence S might be made of letters as in DNA (e.g., ACTTAATG...) or of bits as in a binary sequence (e.g., 11001000...). For convenience we take S to have periodic boundary conditions, so that it forms a ring. We do not know the sequence itself or its length but only the identity of M short fragments of length $\ell < L$, sampled at random from along the sequence. In this paper we address the following question: What is the probability $P(M, L, \ell, A)$ that knowledge of the M fragments enables us to reconstruct the original, native sequence S?

This question is at the heart of whole-genome shotgun sequencing [8,9], a technique developed by Celera Genomics which proved crucial in accelerating the map of the human and other genomes. Shotgun sequencing involves breaking identical copies of a single strand of DNA into pieces and sequencing the fragments, which must then be assembled in their correct order. To accomplish this it is necessary to collect enough fragments such that some regions are sampled redundantly. These multiply sampled regions, or overlaps, must be long enough to allow recognition of which pairs of fragments are neighbors along the original DNA molecule. Shotgun fragments are typically narrowly distributed about a fixed length [8], so we take in our model all pieces to be of constant length ℓ. Because DNA is directed, the fragments cannot be flipped.

Since in our model the characters in S are uncorrelated, we do not take into account repeat regions present in real DNA. Moreover, we do not consider sequencing errors or scaffold structure [8].

Clearly, if the number of fragments M is too small (if $M\ell < L$), we can be sure that some regions of S are not covered by the fragments, thereby making it impossible to recover S in its entirety. Moreover, even if $M\ell > L$, there remains a nonzero probability that the M randomly selected fragments do not cover the whole sequence.

But whether or not the sequence is completely covered by the fragments is not the full story. Additionally, the fragments must be long enough for the information contained in their overlaps to be sufficient to reassemble them. For instance, even if all fragments of length 2 from the sequence 001011 are known (00, 01, 10, 01, 11, 10), they do not give a prescription for a unique sequence; the pieces could also be assembled into 001101.

Knowing the M fragments, one can construct the matrix of their overlaps $q_{\alpha,\beta}$, where $q_{\alpha,\beta}$ is the maximal length over which the head of fragment α coincides with the tail of fragment β. In this matrix, nonzero elements have two possible origins: either the same region in S has been covered by two fragments (we call this region a *native* overlap) or two fragments in S overlap by chance (we call this a *non-native* overlap). Of course, it is not *a priori* known which are which, and our problem may be posed as distinguishing one set from the other.

Infinite alphabet.—The simplest version of the sequence recovery problem is the limiting case of an infinite alphabet ($A \to \infty$). In this case all nonzero overlaps $q_{\alpha,\beta}$ are native. As long as the entire sequence is covered and all nearest-neighbor fragments have overlap ≥ 1, the sequence S can be recovered. Accordingly, the recovery probability can be expressed as a covering problem *with pieces of length* $\ell - 1$, that is,

$$P(M,L,\ell,\infty) = C(M,L,\ell-1), \tag{1}$$

where $C(M,L,\ell)$ is the probability that M pieces of length ℓ completely cover (with or without overlap) a sequence of length L.

The calculation of $C(M,L,\ell)$, not presented here, can be done exactly. One first calculates the probability that, having chosen M fragments from L possible fragments, M_0 of them are different. Then one calculates the probability that the sum σ of $M_0 - 1$ random integers uniformly distributed between 1 and ℓ is such that $L - \ell \le \sigma \le L - 1$. We find

$$C(M,L,\ell) = \sum_{M_0=1}^{M} \sum_{j=1}^{M_0} (-1)^{M_0-j} \frac{j^M (M_0-1)!}{j!\,(M_0-j)!\,L^{M-1}} \times \frac{1}{2\pi i} \oint \frac{dz}{z^{L+1}} \left(\frac{z-z^{\ell+1}}{1-z}\right)^{M_0} \tag{2}$$

(Fig. 2 below), which, at least in the continuous limit ($L,\ell \to \infty$ at fixed $\frac{\ell}{L}$), is a known result [10]. For large γ and M_0, $C \simeq \exp(-M_0 e^{-\gamma})$ since covering is achieved when the end point of each fragment is covered by other fragments (here M_0 may be replaced by M provided $M \ll L$).

Constant native overlaps.—Another simplified version of the recovery problem is the case in which the M fragments are not randomly sampled but equally spaced with a constant overlap b between consecutive fragments (Fig. 1). In this case $M = \frac{L}{\ell-b}$. To avoid correlations between characters in different overlapping regions, we consider only the case $b \le \frac{\ell}{2}$. Since all neighboring pairs of fragments must have precisely overlap b, only in the case of two or more identical overlaps (out of a total of $N = A^b$ realizations) can a non-native sequence be constructed.

The probability that *no* two of the M overlaps are identical is the same as the solution to the well-known birthday problem [11]: if M pieces are selected (with replacement) from N objects, the probability that no two are identical is $Q(0) = \frac{N}{N}\frac{N-1}{N}\cdots\frac{N-M+1}{N}$, which, in the regime $1 \ll M \ll N$, yields

$$Q(0) \simeq \exp[-M^2/(2N)]. \tag{3}$$

The probability $Q(k)$ that there are k pairs of identical overlaps, with the remaining $M - 2k$ overlaps distinct, is $Q(k) = \frac{N(N-1)\ldots(N-M+k+1)M(M-1)\ldots(M-2k+1)}{2^k k! N^M}$. For $M \sim N^{1/2}$, this becomes

$$Q(k) \simeq \frac{1}{2^k k!}\left(\frac{M^2}{N}\right)^k Q(0). \tag{4}$$

One could also calculate the probability that some overlaps are repeated more than twice [11], but they do not contribute significantly as long as $M \ll N^{2/3}$.

Now we need the probability that M fragments containing k identical pairs can be uniquely assembled. Since the $M - 2k$ unique overlaps do not allow rearrangements, we can disregard them, as shown in Fig. 1. The number of ways that the $2k$ remaining overlaps can be paired is $\frac{(2k)!}{2^k k!}$.

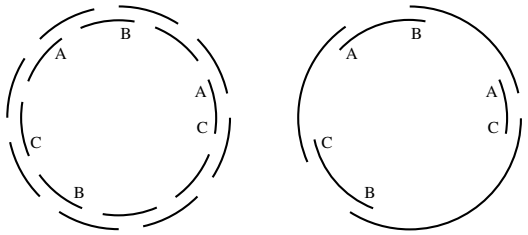

FIG. 1. Left: fragments drawn from a native sequence with constant overlaps. Only pairs of identical overlaps (labeled A, B, and C) allow possible rearrangements. Right: equivalent diagram after fixing nonrepeated overlaps.

Whether or not we are able to uniquely recover S from the remaining $2k$ pieces depends on the precise order of their $2k$ overlaps along S. As soon as among the k pairs two pairs are entangled (e.g., $ABAB$), one can reorder the fragments in a non-native way. Indeed, if in the original sequence we have Ap_1B, Bp_2A, Ap_3B, Bp_4A, where the p_i indicate the nonoverlapping interiors of the pieces, we can reorder them as Ap_1B, Bp_4A, Ap_3B, Bp_2A. Therefore we need to count the number of ways c_k that there is no entanglement between any two pairs (this question appears in many contexts, in particular in the enumeration of meanders [12]). Clearly $c_1 = 1$, $c_2 = 2$ and, with the convention $c_0 = 1$, one can show by recursion that $c_{k+1} = \sum_{i=0}^{k} c_{k-i}c_i$, which is satisfied by the Catalan numbers $c_k = \frac{(2k)!}{k!(k+1)!}$. These are the number of permutations of the $2k$ overlaps for which S can be recovered, out of a total of $\frac{(2k)!}{2^k k!}$. So if M overlapping fragments contain k pairs of identical overlaps, S can be uniquely recovered with probability

$$P_k^{\text{unif}} = 2^k/(k+1)!. \tag{5}$$

Summing the product of (4) and (5) over k, with $N = A^b$, yields the probability of recovering S; it is

$$P^{\text{unif}}(M, M(\ell-b), \ell, A) \simeq \exp\left(-\frac{M^2}{2A^b}\right) \times \sum_{k=0}^{\infty} \frac{1}{k!\,(k+1)!}\left(\frac{M^2}{A^b}\right)^k. \tag{6}$$

Note that, as suggested by the birthday problem, (6) changes from 0 to 1 at $b_c \sim 2\log_A M$. For $b < b_c$, there are many pairs of identical overlaps and recovery is unlikely, whereas for $b > b_c$ the number of repetitions is small and recovery is very probable.

Native versus non-native overlaps.—The key to determining the native sequence is the ability to distinguish between native and non-native overlaps $q_{\alpha,\beta}$. Recovery is difficult when native and non-native overlaps are comparable in length, but becomes easier as the native overlaps dominate. In particular, recovery should be straightforward whenever the smallest native overlap $q_{\min}^{\text{nat}}$ is larger than the largest non-native overlap $q_{\max}^{\text{non}}$.

The probability that $q_{\min}^{\text{nat}} \ge q$ is simply the probability that M pieces of length $\ell - q$ cover S. In terms of the

covering probability (2), this is $C(M, L, \ell - q)$. In the range $1 \ll M \ll L$, $q_{\min}^{\text{nat}}$ has typical value

$$q_{\min}^{\text{nat}} \simeq \max\left[\ell - \frac{L}{M}\ln M, 0\right]. \quad (7)$$

Calculating the exact distribution of $q_{\max}^{\text{non}}$, however, is much more difficult. This will become apparent below in our estimation of the probability r_n that a non-native overlap has length n.

Let t_n be the probability that the tail of one random fragment coincides with the head of another (both supposed to be long) over a length n. Let r_n be the probability that these two fragments have overlap n, i.e., the tail of one coincides with the head of the other over a *maximal* length n. For an alphabet of size A, $t_n = \frac{1}{A^n}$. Surprisingly, an analytic derivation of the r_n is less straightforward than one might expect. This may be appreciated by noting that $t_{i,j} = t_i t_j$ but $t_{i,j,k} \neq t_i t_j t_k$ (where $t_{i,j}$ is the probability that the two fragments have coincidence i and j, etc.). In particular, for $i < j$, one always has $t_{i,j,j+1} = t_{j,j+1} \neq t_i t_{j,j+1}$.

Within the approximation that all the $t_{i,j,k,\ldots}$ are factorizable, that is, $t_{i,j,k,\ldots} \simeq A^{-i-j-k-\cdots}$, we find

$$r_n^{\text{approx}} = \frac{1}{A^n} - \frac{1}{A^{2n}}\frac{1}{A-1} + \frac{1}{A^{3n}}\frac{1}{A-1}\frac{1}{A^2-1} - \frac{1}{A^{4n}}\frac{1}{A-1}\frac{1}{A^2-1}\frac{1}{A^3-1} + \ldots \quad (8)$$

by the standard formula $r_n = t_n - \sum_{n_1>n} t_{n,n_1} + \sum_{n_2>n_1>n} t_{n,n_1,n_2} - \ldots$. These approximate values are reasonably close but clearly not equal to numerical extrapolations of data obtained by an exhaustive enumeration of small strings (Table I).

Although we could not calculate the r_n's exactly, it is clear that, for large n, $r_n \simeq t_n = A^{-n}$. We can then estimate the typical value $q_{\max}^{\text{non}}$, or at least a bound on it, by writing $q_{\max}^{\text{non}} \lesssim n$ for $M^2/(2A^n) \sim 1$ (this estimate neglects the correlations between pairs of non-native overlaps). We find

$$q_{\max}^{\text{non}} \lesssim 2\ln M/\ln A. \quad (9)$$

From (7) and (9), the condition $q_{\min}^{\text{nat}} > q_{\max}^{\text{non}}$ is satisfied if

$$\ell > \left(\frac{L}{M} + \frac{2}{\ln A}\right)\ln M, \quad (10)$$

TABLE I. The probability r_n that two random fragments have overlap n, for $A = 2$ and $A = 4$. The r_n^{extrap} are our best numerical estimates, whereas the r_n^{approx} are given by (8).

	$A = 2$		$A = 4$	
n	r_n^{extrap}	r_n^{approx}	r_n^{extrap}	r_n^{approx}
0	0.267787	0.288788	0.687748	0.688538
1	0.300420	0.288788	0.230237	0.229512
2	0.198919	0.192525	0.061264	0.061203
3	0.112161	0.110014	0.015548	0.015544
4	0.059285	0.058674	0.003901	0.003901
5	0.030446	0.030284	0.000976	0.000976

which, in terms of $\gamma = M\ell/L$ (called redundancy of coverage in [13]), may be written $M < e^{\gamma}$ for $M \ll L$.

Assembly strategies.—We now come back to our original question concerning fragments independently sampled from a finite alphabet native sequence, a situation mirroring shotgun sequencing. How is shotgun data put together in practice? Because the number of ways of ordering the fragments grows as $M!$, an exhaustive search over all possible solutions quickly becomes impractical. Instead, the many overlapping fragments are assembled according to one of a variety of proprietary heuristic algorithms [14].

The assembly puzzle can be naturally mapped onto a traveling salesman problem (TSP) [15]. To each permutation π of the fragments we associate an energy E_π defined as $M\ell$ minus the sum of the overlaps of fragments consecutive in π. The assembly puzzle is soluble if and only if the ground state energy $E_{\min} = L$ and is nondegenerate.

Inspired by TSP greedy algorithm heuristics, we investigate two simple assembly strategies G1 and G2 below. An elaboration of G1, in which fragment cleanliness (absence of repeat regions, reliability of bases, etc.) is also considered, is the basis of a real assembly technique [14].

In the first greedy strategy (G1), a random fragment is selected, and to its right is concatenated that fragment from the $M - 1$ remaining with the greatest overlap. To the right of this piece is joined the optimal fragment from the $M - 2$ remaining, and so on, until one fragment remains. This and the string are concatenated at both ends and the resulting sequence is the candidate sequence.

The second strategy (G2), described in [16], is similar to the first one but allows the formation of disjoint strings. From the pool of M fragments, that pair with the greatest overlap is concatenated and thereafter considered as a single fragment. This is then repeated with the pool of $M - 1$ fragments, and so on, until two fragments remain. These are joined at both ends and the result forms the candidate sequence.

If G1 or G2 are faced with identical overlaps, we proceed by flipping a coin. As we want the probability that S

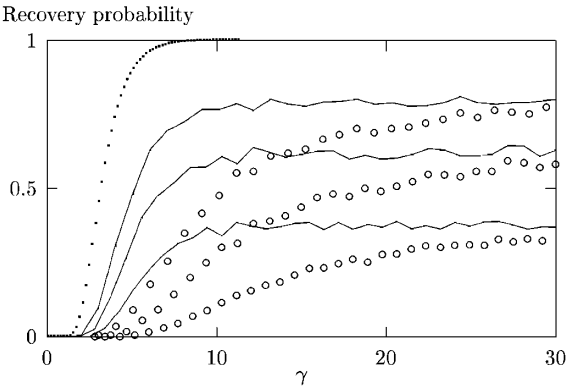

FIG. 2. Probability of recovering the exact sequence as a function of $\gamma = M\ell/L$ using G1 (circles) and G2 (lines). Curves are shown for (from left) $\ell = 13, 12, 11$, with $L = 64$ and $A = 2$. The covering probability (2) (dots) is shown for comparison with $\ell = 12$ and $L = 64$.

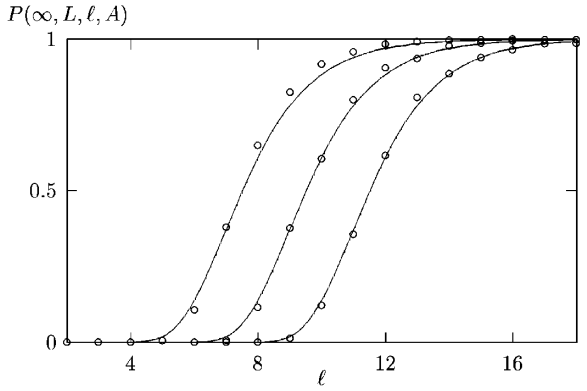

FIG. 3. The points represent asymptotes ($M \to \infty$) for the G2 curves shown in Fig. 2 and others, as a function of ℓ with $A = 2$. The solid lines are given by [12] for (from left) $L = 16, 32, 64$.

can be reconstructed independent of chance, we repeat G1 (or G2) many times (10–100) and consider S to be successfully recovered only if it is assembled without exception from the same M fragments.

G1 and G2 are compared in Fig. 2 as a function of redundancy of coverage $\gamma = \frac{M\ell}{L}$. While initially G2 outperforms G1, for large γ the two curves display identical asymptotic behavior. Notably, the chance of making a mistake does not vanish as $\gamma \to \infty$.

In the limit of large γ, all L possible fragments are sampled, and the only source of assembly errors is the presence of two or more identical regions of length $\ell - 1$. Therefore $P(\infty, L, \ell, A)$ reduces to the probability $Q(L, \ell - 1, A)$ that no two regions of length $\ell - 1$ in S are the same [17]. This again looks like a birthday problem, and one could believe that Q is given by (3) with $M = L$ and $N = A^{\ell-1}$. But in this case $b = \ell - 1 \geq \frac{\ell}{2}$ and the overlaps are no longer independent—they contain all together $L(\ell - 1)$ characters but only L of them can be freely chosen.

Heuristic arguments, too long to outline here, lead us to conjecture that Q is identical to (3) apart from a factor of $(A - 1)/A$ in the exponential, that is,

$$P(\infty, L, \ell, A) = Q(L, \ell - 1, A) \simeq \exp\left(-\frac{(A-1)L^2}{2A^\ell}\right), \tag{11}$$

and is therefore significantly higher than (3) would suggest. As shown in Fig. 3, Eq. (11) predicts the asymptotic behavior of our two assembly strategies (as well as, were we able to perform one, an exhaustive search).

Discussion.—For an idealized shotgun analysis of a typical human chromosome ($L \simeq 10^8$, $\ell \simeq 500$, $A = 4$), complete sequence coverage (2) is likely for redundancy of coverage $\gamma = M\ell/L \gtrsim 15$. Our condition for sequence assembly (10) leads to the same estimate because for this range of parameters the second term in the right-hand side of (10) is negligible. This estimate is comparable to values used in the laboratory and in shotgun sequencing models, e.g., $\gamma = 10, 10, 7.5$ [8,9,13]. It should be kept in mind, however, that all of these rely on different definitions of "recovery"; ours requires that the probability of no errors is close to unity, whereas others allow a small fraction of errors (neglecting repeat regions).

Throughout this Letter we have assumed a random native sequence S. Real DNA, however, contains long-range correlations [4], in particular a preponderance of identical repeated regions [8,14]. As we have seen, identical regions in S are the dominant source of errors once S is likely to be covered. Therefore we would expect the probability of uniquely recovering a real biological sequence to be diminished and (10) to be modified.

In addition to the introduction of correlations in S, a number of extensions of the present work offer increased realism but are not considered here. These include sequencing errors in the fragments themselves and DNA polymorphisms [8], polydispersity of the fragment lengths, and partial recovery of S. Of theoretical interest is the calculation of the minimal length over all permutations of M fragments *not* drawn from a native sequence but chosen at random, which could be studied using techniques developed for the TSP problem [15].

The authors thank E. Brunet, J. L. Lebowitz, and L. Shepp for useful discussions and the ITP, Santa Barbara, for hospitality. This research was supported in part by the National Science Foundation under Grant No. PHY99-07949.

*Electronic address: derrida@lps.ens.fr
†Electronic address: fink@lps.ens.fr

[1] E. S. Lander *et al.*, Nature (London) **409**, 860 (2001).
[2] J. C. Venter *et al.*, Science **291**, 1304 (2001).
[3] Martin Vingron and Michael S. Waterman, J. Mol. Biol. **235**, 1 (1994).
[4] A. Arneodo *et al.*, Phys. Rev. Lett. **74**, 3293 (1995).
[5] Pedro Bernaola-Galván *et al.*, Phys. Rev. Lett. **85**, 1342 (2000).
[6] T. Hwa and M. Lassig, Phys. Rev. Lett. **76**, 2591 (1996).
[7] T. Hwa, Nature (London) **399**, 17 (1999).
[8] James L. Weber and Eugene W. Myers, Genome Res. **7**, 401 (1997).
[9] J. Craig Venter *et al.*, Science **280**, 1540 (1998).
[10] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1966), Vol. II, p. 28.
[11] R. Arratia, L. Goldstein, and L. Gordon, Stat. Sci. **5**, 403 (1990).
[12] P. Di Francesco, O. Golinelli and E. Guitter, Commun. Math. Phys. **186**, 1 (1997).
[13] Eric S. Lander and Michael S. Waterman, Genomics **2**, 231 (1988).
[14] Ting Chen and Steven S. Skiena, Bioinformatics **16**, 494 (2000).
[15] J. Vannimenus and M. Mezard, J. Phys. (Paris) Lett. **45**, L1145 (1984).
[16] M. S. Waterman, *An Introduction to Computational Biology* (Chapman and Hall, London, 1995).
[17] J. L. Guibas and A. M. Odlyzko, Probab. Theory Relat. Fields **53**, 241 (1980).