Bulletin of Mathematical Biology Vol. 51, No. 1, pp. 117–124, 1989.
Printed in Great Britain.

0092–8240/89$3.00+0.00
Pergamon Press plc

PROBABILISTIC MODELS OF GENOME SHUFFLING

■ DAVID SANKOFF* and MARTIN GOLDSTEIN
Centre de recherches mathématiques and Departement de mathématiques et statistique,
Université de Montréal,
C.P.6128, Succursale "A",
Montréal, Canada H3C 3J7

The comparison of entire genomes in evolutionary studies gives rise to alignments characterized by many intersections, or inversions in the order of two fragments in different genomes. To model this, we suggest a random migration process for fragments, and discuss its equilibrium distribution in the case of linear and circular genomes. Simulations are carried out to explore "cut-off" behavior as the process approaches equilibrium. We define a new process to take into account the indistinguishability of two fragments which are adjacent in both genomes being compared. Questions of applicability of these models are discussed.

1. Gene Evolution versus Genome Evolution. Classical models of gene sequence or protein sequence evolution postulate elementary operations of substitution, insertion or deletion of bases or residues, with generalizations to allow contiguous blocks of sequence terms to be substituted, inserted or deleted (Kruskal and Sankoff, 1983). The goal of much sequence comparison theory has been to develop algorithms to reconstruct a parsimonious account, in terms of these elementary operations, of the divergent evolution of two homologous sequences.

In most models, given a set of elementary operations which relates one sequence **s** to another **t**, it is usually possible to express this relationship in terms of an alignment A, which is a set of pairs (s_i, t_j) such that if (s_i, t_j) and (s_h, t_k) are both in A, then $i<h$ if and only if $j<k$. In other words, if we write one sequence above the other, and connect with straight lines each pair of related terms, no two of these lines intersect. An alignment involves sequence terms in three different ways. Terms contained in no pair of A are considered to have been inserted in one sequence or deleted from the other. If (s_i, t_j) is in A and $s_i \neq t_j$, this is the result of the substitution operation; if $s_i = t_j$ there has been no change in this part of the sequence.

The non-intersection property of alignments has been a key element in the development of efficient algorithms for optimizing them. It has proved difficult in theory (Lowrance and Wagner, 1975; Wagner, 1983) and infeasible in practice to allow transpositions into the set of elementary operations or, equivalently, to allow intersecting lines in alignments. The predominance of

* Author to whom correspondence should be addressed.

substitutions, insertions and deletions as evolutionary mechanisms on the level of gene sequences, however, means that the intractability of transpositions is not a major problem.

The comparison of whole genomes is quite different. Here we find that there are many genes or other genome fragments in one organism which have clear counterparts in another organism (as detected by the traditional sequence comparison techniques), but that these fragments are not necessarily in the same order. Writing one (linear) genome above the other and drawing lines between corresponding fragments, we do not have the non-intersection property.

In this paper we discuss such fragment alignments for simple linear and circular genomes. The goal is not to detect nor to optimize alignments but rather to model the evolutionary processes which give rise to them and hence to deduce some of their statistical properties.

2. Properties of Random Permutations. Intersections in fragment alignments may come from several sources—duplication and migration of genes, insertion of external genetic material in both organisms, etc. In this paper we will explore just one mechanism, the migration of a fragment from one position in the genome to another. Since such migration does not generally seem constrained to occur by small steps, we will not impose any such constraint in the model.

The elementary evolutionary operation, then, will be to choose one gene fragment at random in the genome, to remove it from its position, and to insert it at some other point in the genome, also chosen at random. To simplify even further, we assume that there is a well-defined and previously identified set of disjoint fragments making up the genome, all equally eligible to migrate. This is not true, of course, and in Section 4, we will explore a weaker assumption which does not require the fragments to be previously identified.

In the next section we will discuss how the iteration of this mechanism eventually leads to the completely random permutation of the genome fragments. First, however, we investigate some intersection properties of the alignment of two genomes made up of the same fragments, where one is randomly permuted with respect to the other.

2.1. The linear case. We label the fragments in the first genome, in order, $1, 2, \ldots, n$ and those in the second $a_1, a_2, \ldots, a_n$. Then $(a_1, a_2, \ldots, a_n)$ is a random permutation. We define the intersection number (or number of inversions—Feller, 1957, pp. 241–242) Y_n to be the number of pairs (i, j), where $1 \leqslant i < j \leqslant n$ and $a_i > a_j$. The moment generating function of Y_n is:

$$\mathscr{P}_n(s) = \prod_{k=1,n} (1-s^k)/k(1-s),$$

so that:

$$E(Y_n)=n(n-1)/4,$$
$$\mathrm{Var}(Y_n)=n(2n^2+3n-5)/72.$$

These facts allow us, in a context where we know n and can calculate Y_n, to test whether the internal arrangements of two genomes show more parallelism than completely unrelated genomes.

2.2. The circular genome. In aligning two circular genomes, the calculation of the number of intersections is not so easily carried out.

We can write $1, 2, \ldots, n$ equally spaced around a circle, opposite $a_1, a_2, \ldots, a_n$ on a concentric circle as in Fig. 1. Then if $a_i=1$, we draw a line from 1 to a_i either clockwise or counter-clockwise, a choice which does not present itself in the case of linear genomes. This choice will in general have consequences for the number of intersections, as illustrated in Fig. 1, where there are one or two intersersections depending on which direction is chosen for the line from 1 to the outer circle to $a_2=1$ on the inner circle.

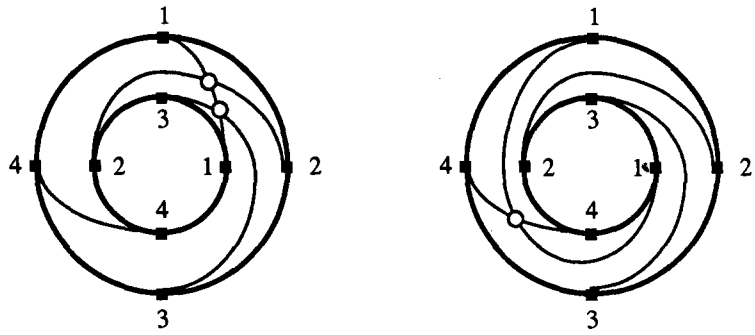

Figure 1. Two representations of the alignment of a pair of circular genomes. Open dots indicate intersections.

In general, it might seem to be most reasonable to choose the clockwise direction to join j to $a_k=j$ if $k-j\leqslant n/2 \pmod n$ and counter-clockwise if $k-j\geqslant n/2 \pmod n$, though this rule does not decide for the case $k-j=n/2 \pmod n$ and does not always minimize the number of intersections, as is clear from Fig. 1.

Another difference between linear and circular genomes is that in the latter case, if we are presented with a pair of genomes to compare, the choice of a_i to occupy the "twelve o'clock" position on the inner circle, opposite the 1 on the outer circle, is essentially arbitrary. As a result of the clockwise–counter-

clockwise rule, the number of intersections can be manipulated by rotating the inner concentric circle as in Fig. 2.

A reasonable solution to this problem might be to choose the rotation of the inner circle so as to minimize the number of intersections. This solution, however, renders the analytic study of Y_n quite difficult.

Were the rotation chosen at random, however, it can be shown that $E(Y_n) \approx n^2/6$ rather than the $n^2/4$ as in the case of linear genomes.

3. Shuffling Models. To the extent that the limiting distribution of the number of intersections is known, we can test whether a given alignment represents a detectibly early stage of evolutionary divergence between the two genomes or whether any observed parallelism in fragment sequence is no more than random coincidence.

Indeed, if we knew the trajectory of the expected number of intersections over time in an alignment of an ancestral genome and its descendant, this would provide us with a calibration for estimating divergence time based on number of intersections observed. Figure 3 depicts this trajectory as estimated from 1000 simulations of the process described at the beginning of Section 2, when the number of fragments is 100.

From this figure, we can see that when the number of intersections observed is 2000, say, then the estimated divergence times is about 130 (shuffling events) though the upper and lower error curves cross 2000 at 100 and at 170, respectively.

The variability in divergence times estimates is worse for longer times and leads to the question of when we should no longer make use of this type of inference. This suggests investigating the "cut-off" behavior of this shuffling process which is similar to those studied by Aldous and Diaconis (1986; 1987). These authors have shown that in a large class of card-shuffling methods, a good deal of non-randomness remains in the order of the n cards in the deck until a certain cut-off point is nearly reached (typically $n \log n$ shuffles), at which point this non-randomness disappears rather rapidly. The amount of non-randomness is measured by the variation distance $\sum|Q_k - Q|$ between the probability distribution Q_k on the space of permutations after k shuffles and Q, the uniform distribution on this space.

To simulate this behavior, if it exists for our model, we defined a sub-set of permutations which includes all except relatively few, "non-random", permutations. This set M includes all those permutations where the number of intersections, when aligned with $1, \ldots, n$ is at least $E(Y_n) - \mathrm{Var}(Y_n)^{1/2}$. We reasoned that the distance of the distribution Q_k from Q would be closely related to the proportion of sample paths leading to permutations which still have an abnormally small number of intersections after k shuffles. Cut-off of non-randomness around k shuffling events would then be reflected in a

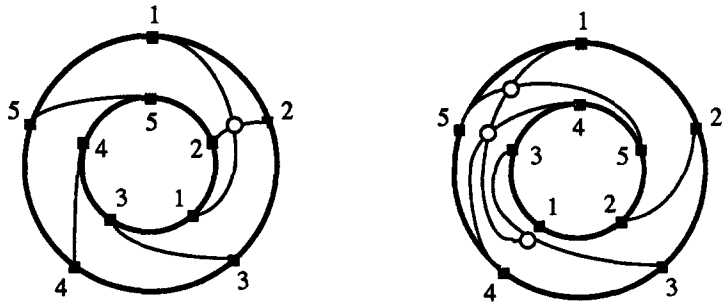

Figure 2. Change in number of intersections imposed by the clockwise–counter-clockwise rule after rotation of inner circle.

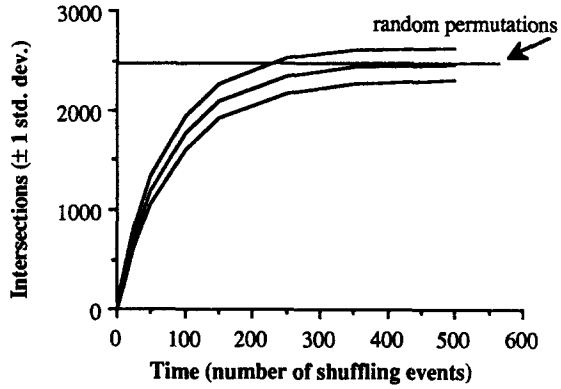

Figure 3. Increase in the expected number of intersections over time.

relatively rapid increase in the proportion of paths in M before and after k shuffles. Figure 4 depicts this phenomenon for the simulation experiment described above.

This result is as expected: no paths enter M before about one hundred events, and by 300 events, almost all paths are in M. Consider now the point at which half of the paths are in M. This is also the point at which, about half of the time, we will no longer be able to distinguish the current arrangement from a purely random permutation by a one-tailed, one standard deviation test of the sort mentioned in Section 2.1. How does the position of this "cut-off point" vary with n? We repeated the simulation experiment for various values of n to produce the results in Fig. 5. It is clear from the least squares fit depicted in the figure that the position of the cut-off point is proportional to $n \log n$.

4. How Many Fragments? In the previous sections we have taken for granted that the set of genome fragments which *have* migrated or which *could have* migrated is known. In fact, this is not the case. When we compare two genomes

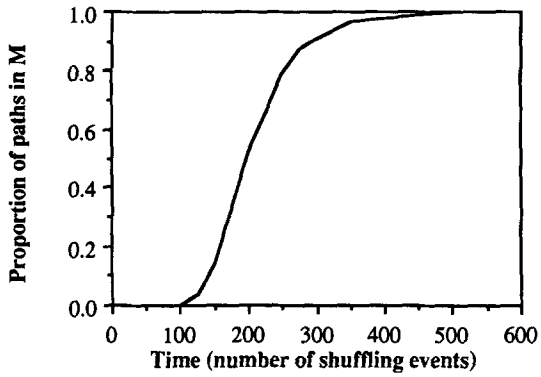

Figure 4. Simulated cut-off behaviour for genome shuffling model.

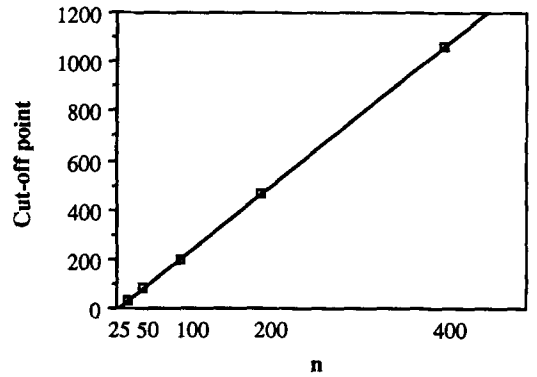

Figure 5. Location of cut-off as a function of n. Abcissa scaled as $n \log n$.

and notice that there is a rearrangement of the order of some sections in one compared to the other, we have no way of knowing (in general) whether each fragment which seems to have migrated is indivisible, in the sense that no sub-fragment could have migrated by itself. Similarly we have no way of knowing whether long sections which have remained in the same configuration in both genomes have not been rearranged because of functional or structural reasons, or whether each is composed of a number of independent fragments, and not enough time has elapsed for random migrations to have occurred in sufficient number.

An immediate consequence of this lack of observability is that when we label the fragments in one genome $1, \ldots, n$ and those in the other genome $a_1, \ldots, a_n$, we cannot have that:

$$a_{i+1} = a_i + 1, \tag{1}$$

because in this case we would have no evidence that the i^{th} and $i+1^{\text{st}}$ fragments in the first genome, and a_i and a_{i+1} in the second, do not consist simply of one single fragment.

Since we do not know the "true" value of n, we cannot know whether Y, the number of intersections, is particularly high or not. For example, for the observed permutation (3, 2, 1), we have $Y=3$, which is maximal if $n=3$. However, it could be that the "true" number of possibly movable fragments is $n=9$, but instead of observing (3, 4, 5, 6, 7, 8, 9, 2, 1) where $Y=15$, relatively small compared to $E(Y_9)=18$, we can only observe (3, 2, 1), and conclude incorrectly that the two genomes conserve no parallelism whatsoever.

This leads to the following class of inference problems: We hypothesize that there is an "underlying" n and a permutation on $1, \ldots, n$. Wherever equation (1) holds, i and $i+1$ are replaced by a single fragment labelled i, and similarly a_i and a_{i+1} replaced by a_i, (and the rest renumbered $i+1 \leftarrow i+2$, $i+2 \leftarrow i+3, \ldots, n-1 \leftarrow n$). If equation (1) holds for some other i, the same procedure is carried out again, and so on. Thus, we observe only some $m \leqslant n$ and some permutation on $1, \ldots, m$ where equation (1) is excluded. How do we estimate n? Does it help to take Y into account? For completely random permutations, this is a relatively minor problem since $E(m) \approx n-1$. For the random shuffling model of Section 3, however, the behavior of $E(m)$ over time is as complex as that of $E(Y)$.

5. *Discussion.* While the familiar techniques of gene sequence comparison will be indispensible for detecting and evaluating similar fragments in two or more genomes, the comparison of *entire* genomes requires evolutionary models which transcend the "microscopic" processes of nucleotide substitution, insertion and deletion. Here we have discussed only a restricted class of such models where unconstrained genome fragment migration is the key "macroscopic" process of genome evolution. The formalization of this process leads to a number of known and new problems in probability theory.

We have chosen to focus on a particular measure of the degree of evolution, namely the number of intersections in the alignment of the two genomes. Other measures might be of equal or greater interest; for example, the length of the longest sub-set of genome fragments which have the same linear order in the two genomes. For random permutations, the limiting behavior of this statistic is known (Logan and Shepp, 1977).

More realistic models would, of course, have to take into account other important processes of genome evolution, notably gene duplication. We might wish as well in certain contexts to impose some non-uniform distribution on the distance along the genome a migrating fragment travels.

The case of the circular genome is more difficult than the linear one, chiefly because of the possibility of rotating one genome relative to the other to

simplify the alignment. For large values of n, however, this difficulty may not be serious.

The simulations of the shuffling model confirm that it belongs to the Aldous–Diaconis class, and provide a rationale for the choice of a particular cut-off point for divergence time estimates. It remains to prove the cut-off behavior analytically.

Finally, the increase in the number of "observable" fragments over time, and its relation to the increase in the number of intersections is an important question to study if the models we are proposing are to have real applications.

We thank Yvon Abel for carrying out the simulations, David Aldous for discussions on shuffling models and Samuel Karlin for information about inversions. This work was supported in part by operating and infrastructure grants to D.S. from the National Science and Engineering Research Council of Canada. D.S. is a fellow of the Canadian Institute for Advance Research.

REFERENCES

Aldous, D. and P. Diaconis. 1986. "Shuffling Cards and Stopping Times." *Am. Math. Mont.* **93**, 333–348.

——— and ———. 1987. "Strong Uniform Times and Random Walks." *Adv. Appl. Math.* **8**, 69–97.

Feller, W. 1957. *An Introduction to Probability Theory and Its Applications.* New York: Wiley.

Kruskal, J. B. and D. Sankoff. 1983. "An Anthology of Algorithms and Concepts for Sequence Comparison." *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*, D. Sankoff and J. B. Kruskal (Eds), pp. 265–310.

Logan, B. F. and L. A. Shepp. 1977. "A Variational Problem for Young Tableaux." *Adv. Math.* **26**, 206–222.

Lowrance, R. and R. A. Wagner. 1975. "An Extension of the String-to-String Correction Problem." *J. Assoc. Comput. Mach.* **22**, 177–183.

Wagner, R. A. 1983. "On the Complexity of the Extended String-to-String Correction Problem." *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*, D. Sankoff and J. B. Kruskal (Eds), pp. 215–235.

Received for publication 1 July 1988