## Acknowledgments

I would like to thank Dr. Mike Sternberg for support and advice. I would also like to thank Drs. Marketa Zvelebil, Chris Rawlings, and Prof. Tom Blundell. This work was supported by the Science and Engineering Research Council (UK) and The Imperial Cancer Research Fund.

# [26] Genomic Divergence through Gene Rearrangement

*By* DAVID SANKOFF, ROBERT CEDERGREN, AND YVON ABEL

## Introduction

Measures of similarities and distances among nucleotide or amino acid sequences have been used to find related regions in long sequences, to test for homology, to assess phylogenetic and functional relationships, and to estimate divergence time between pairs of evolutionarily related sequences. This last use especially invokes a model, possibly implicit and including a random component, of sequence change through nucleotide (or amino acid) replacement and insertion or deletion of single nucleotides or small blocks of contiguous nucleotides. Modeling genetic events by these mechanisms, however, cannot reflect the more macroscopic processes of evolutionary divergence of organisms, such as duplication, inversion, and transposition (shuffling) of parts of the genome. Although little comprehensive data are available as yet to study evolution at the level of entire genomes, the megasequencing projects now being set up will be producing genomic sequences in the near future.

In this chapter we discuss simple probabilistic models for genome shuffling introduced by Sankoff and Goldstein[1] and apply them to the assessment of relationships among a number of bacterial genomes. Lacking complete nucleotide sequences at this level, we assess our methodology on genetic map data. See Nadeau and Taylor[2] and Sakharov and Valeev[3] for comparable approaches.

[1] D. Sankoff and M. Goldstein, *Bull. Math. Biol.* **51,** 117 (1988).

[2] J. H. Nadeau and B. A. Taylor, *Proc. Natl. Acad. Sci. U.S.A.* **81,** 814 (1984).

[3] E. A. Sakharov and A. K. Valeev, *Dokl. Akad. USSR* **301,** 1213 (1988).

## Shuffling Models

In the simplest model of random genome shuffling, we assume that the genome consists of *n* fragments, linearly disposed. These fragments may consist of genes, entire operons, or other larger or smaller regions of the genome. In the absence of evidence to the contrary, we assume that each of these fragments has the same probability per unit time of being transposed elsewhere in the genome, even for genomes of very different sizes. To ensure this we postulate that shuffling events occur at regular time intervals, inversely proportional to *n* (i.e., the rate of shuffling events is proportional to *n*), and each event sees one fragment chosen at random, moved to some other randomly chosen point on the genome between two other fragments, or to one end, and inserted there. The same process can also be used to model shuffling of a circular genome, though in this case we need not worry about the possibility of moving to the end of the genome.

More general models, which we do not discuss here, would favor relatively short-range fragment migration according to some probability distribution over distance along the genome. Such models would allow incorporation of empirically obtained parameters on rates of transposition, the distribution of transposition hot spots, and the tendency of contiguous fragments to be transposed as a unit.

## Measures of Divergence

For models of sequence divergence through replacement, insertion, and deletion, the usual way of measuring sequence similarity or difference is to write one sequence above the other as in Fig. 1 and to draw a series of trace lines connecting pairs of terms, one in each sequence, such that no two lines cross and such that an optimality criterion is satisfied. Both elements of a pair of terms connected in the trace are inferred to originate in the same term in the ancestral sequence. The optimality criterion basically sums the similarity or difference scores of each pair of terms connected by the trace lines, plus a score for each unconnected term, i.e., insertion or deletion.

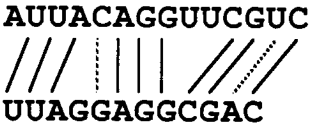

FIG. 1. Trace between two sequences implying two replacement mutations (dotted lines) and four insertions and deletions (unconnected terms).

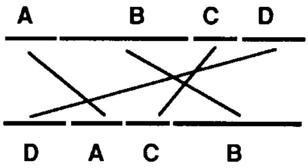

FIG. 2. Trace between two genomes consisting of four fragments. There are four intersections in this trace.

In models where fragment migration is the mechanism of evolutionary divergence, there will be exactly *n* trace lines connecting two genomes of common ancestry, and these will necessarily cross, as in Fig. 2, unless no net evolution has occurred. In fact, the more fragments have moved, the more intersections there tend to be. Thus, we can count the number of such intersections and use this as an indicator of the extent of divergence. Note that, in contrast to gene-level models involving replacement, insertion, and deletion, in transposition models we assume that we know which fragments in one genome are related to which ones in the other, so that there is no necessity of finding the optimal trace. There is no "alignment problem"; we know the true trace.

For circular genomes, the trace is constructed by drawing two concentric circles, and then connecting corresponding fragments by trace lines proceeding in a clockwise or counterclockwise direction within the ring between the circles, as in Fig. 3. Note that, in contrast to linear genomes, even though we may know which fragments correspond to each other in the two circular genomes, there are many ways of constructing the trace because of the possibility of choosing the clockwise or counterclockwise direction for each connecting line. As we shall see, this leads to a problem

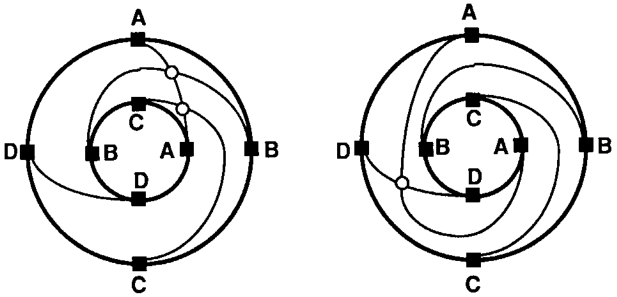

FIG. 3. Two ways of constructing a trace for the same pair of circular genomes. Filled squares represent midpoints of fragments A, B, C, and D. Open circles indicate intersections.

of optimal alignment, though not of the same sort as in the gene-level models.

Where few transposition events differentiate between two related genomes, most of the fragments will be in the same linear (or circular) order in the two sequences, but as the number of migrations increase it will become increasingly difficult to find a long subset of fragments that are in the same order in the two. Thus, another way of assessing the similarity between two genomes is to find the largest subset of the fragments which are in the same order in both genomes. The size of this subset, in comparison to $n$, is an indicator of the relationship between the two genomes. This indicator does not depend on how the trace is constructed.

## Initial Behavior and Limit Theory

Sankoff and Goldstein investigated some of the properties of the number of intersections in a genome consisting of $n$ fragments, under the random shuffling model.[1] In the linear model, in each of the first few shuffling operations, the number of intersections can be expected to increase by somewhat less than $3n/8$ on the average. This may be seen by considering first a fragment at one end of the genome. When it is moved randomly, this will give rise to between 0 (if it is put back into its original position) and $n - 1$ intersections (if it is moved to the opposite end of the genome). The average will be $(n - 1)/2$. For a fragment at the center of the genome, its movement will result in between 0 and $(n - 1)/2$ intersections (assuming $n$ is odd), for an average of $(n - 1)/4$. Thus, the movement of a randomly chosen fragment will result in a number of intersections midway between the two extreme cases, namely $3(n - 1)/8$. Because the rate of shuffling events has also been assumed linear with $n$, we may expect the number of intersections to increase at an initial rate proportional to $n^2$. After many shuffles (of the order of $n \log n$), it can be proved that the number of intersections will approach $n(n - 1)/4$.

In Fig. 3, it can be seen that the choice of clockwise versus counterclockwise for a trace line can affect the number of intersections, and that the choice of the shortest route (i.e., $<180°$) does not minimize the number of intersections. Thus, the two A fragments give rise to two intersections when connected by the shortest route and only one when connected by the more circuitous route. Nevertheless, it seems reasonable to make the convention that all trace lines travel no more than 180° (this is always possible). In this case, the initial increase per shuffle should be of the order of $3n/16$, using the same reasoning as in the linear case, so that the initial rate per unit time should also be quadratic in $n$. The asymptotic expectation can be shown to be of the order of $n^2/6$. Note, however, that this latter

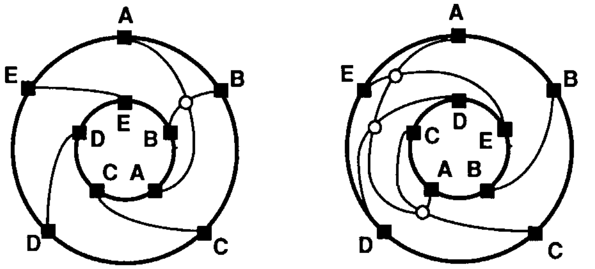

FIG. 4. Changing number of intersections as one genome is rotated with respect to the other.

value is calculated for the trace of two thoroughly shuffled circular genomes randomly rotated with respect to each other. However, by slowly rotating one genome while holding the other fixed, and examining how this changes the trace (as some clockwise lines get too long, i.e., travel more than 180°, and are thus replaced by shorter counterclockwise ones, and/or vice versa), we inevitably find some traces with a lot more intersections and some with less, as in Fig. 4.

Which rotation should we use to calculate the number of intersections? Perhaps the most natural choice is the one which minimizes the number of intersections in the trace since, in preasymptotic behavior, the minimizing rotation will presumably be close to the original homologous alignment of the two genomes. The minimizing value in the completely randomized case would be hard to predict analytically, but simulations are quite feasible, as discussed below.

As for the largest ordered fragment subset, the initial few shuffles should each reduce its size by about 1, in both the circular and linear models. The limiting value for the expected size of this subset is $2\sqrt{n}$ in the linear model[4] and is very close to the same value in the case of random circular genomes.

## Data

There are of course few data sets at present representing entire genomes. On the nucleic acid level, these are confined to viruses and some mitochondria and chloroplasts. It is thus necessary to make use of other kinds of data in developing and testing our methodology. In this study, we identify the genome with a linear (or, rather, circular) set of genetic

[4] D. Sankoff and S. Mainville, *in* "Time Warps, String Edits and Macromolecules" (D. Sankoff and J. B. Kruskal, eds.), pp 363–365. Addison-Wesley, Reading, MA, 1983.

markers of bacteria:[5] *Escherichia coli, Salmonella typhimurium, Bacillus subtilis, Caulobacter crescentus,* and *Pseudomonas aeruginosa.*

The shuffling of the relative positions of the markers in one organism with respect to the corresponding markers in the other reflects the genomic history of transposition. There are a number of problems associated with the use of these kinds of data. The most serious has to do with the comparability of the different databases. There is little difficulty in identifying corresponding markers on the maps of *E. coli* and *Salmonella,* but whether similarly labeled markers in other genomes represent homologous genes is not always clear.[6] Likewise, two homologous markers may be labeled in slightly, or even completely, different ways in two different genomes. Thus, we systematically examined the functional descriptors of the genes in order to construct a normalized labeling across all five organisms. Identical labels for clearly unrelated genes in two organisms, caused by orthographic coincidence, were altered to avoid artifactual correspondences. Conversely, homologous genes differently labeled in two experimental traditions were relabeled to reflect this homology. Markers missing from either the map or the list of descriptions were discarded, and orthographic inconsistencies were regularized.

Functionality, of course, especially as inferred from the brief descriptions accompanying the genetic maps, is not always a reliable guide to homology, but it seems clear that, at least in a statistical sense, the labels in our normalized data base[7] constitute a better reflection of marker identities and differences across organisms than the uncorrected maps.

The fact that there may not be many comparable markers in two genomes as inferred from the four-letter marker labels contained in the genetic map data base led us to experiment as well with the first three letters only, given the possibility that slightly different labeling conventions might obscure genuine homologies, either because these are unknown or because they cannot be inferred from the functional descriptors in the original data. Thus, we repeat all calculations once requiring the first three letters of the marker labels to be identical in order to determine "homology" and once requiring all four symbols to be identical.

Independent of labeling considerations, maps of different genomes have different sets of markers, if only due to more extensive research on one organism compared to another. This, however, presents no difficulty within the framework of our model. By discarding, in each pairwise com-

[5] S. J. O'Brien, ed., "Genetic Maps." Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1987.

[6] K. E. Sanderson and J. R. Roth, *Microbiol. Rev.* **52,** 485 (1988).

[7] R. Cedergren, Y. Abel, and D. Sankoff, unpublished work.

parison of genomes, all markers not present in either one or the other, we arrive at a value of $n$ which is perfectly appropriate for that particular comparison. Our model has been constructed, especially with regard to the rates of occurrence of transposition events, so that whatever we can infer from the $n$ comparable markers does not depend on what has happened to the markers which have been discarded. This is actually an advantage of using genetic maps in this type of study rather than physical maps or complete sequences. One of the problems in applying shuffling models to sequence-level genomic data is the difficulty of identifying the fragments. It is not usually obvious where to place the boundaries around a region of the sequence which can potentially be transposed, unless it is already located in different places within the genome in a number of related organisms. This becomes feasible only with large bodies of comparative data. The genetic markers, however, are generally independent, discrete entities, though in some cases the location of one marker may be tied functionally to the position of another.

Another problem has to do with gene duplication, represented in this data set by repeated labels in the same genome. Gene duplication should not be widespread in bacterial genomes, however, so that most apparent duplications in our data are probably due to similarity in three-letter marker labels of nonhomologous genes within the same operon. Nevertheless, in this exercise, an ad hoc solution to the duplicate label problem is to pick one marker at random out of each set of repetitions independently in each organism prior to counting intersections or finding largest ordered subsets. This may be repeated a few times to produce an average score.

## Computational Methods

Not all the genomes we are studying are known to be circular; nevertheless, for the purposes of testing our methods, we treat them all in the same way. We compare the five genomes two at a time. As a first step we list each of the genetic markers in each genome. We retain only those which occur in both, according to either a three-symbol or a four-symbol criterion for matching their labels. If there are duplicate labels in one genome, one is chosen at random to correspond with that label in the other genome. If there are duplicate labels in both genomes, as many of such random correspondences are set up as possible. This protocol gives us a set of $n$ markers. The analysis to follow is repeated for 10 such sets, created by different random choices among the duplicate labels, and the scores of the following calculations are averaged over the 10.

As mentioned above, given two circular genomes derived from the same ancestor by a random shuffling process, it is not necessarily clear how they were aligned (rotationally) at the moment of their divergence. Thus, we repeat the following analysis using every possible rotation of one genome with respect to the other.

A trace line is defined for each corresponding pair of markers in the two genomes. Because of the circular configuration, each trace line can be drawn clockwise or counterclockwise; we always choose the shorter path (< 180°, as in Fig. 4). A routine then enumerates the number of intersecting trace lines in computing time quadratic with *n*. The search for the largest subset of markers with the same order in the two genomes is carried out by a dynamic programming algorithm similar to that used for matching nucleotide or amino acid sequences. It also requires quadratic time. The rotation with the smallest average intersection score is inferred to reflect the original alignment of the two genomes. In further studies, another approach might be to use the largest ordered subset of markers as a basis for the alignment, since this subset is likely to contain just those markers which have not undergone shuffling.

## Simulations

Figure 5 portrays the difference between $n^2/6$ (the limiting approximation for the expected number of intersections under random rotation), the simulated expected value (based on 1000 samples), and the simulated minimum rotation expected value (based on 1000 samples for *n* less than 60, on 100 samples for *n* less than 150, and 10 samples for larger *n*). Both curves are approximately linear, confirming that, as a proportion of $n^2/6$, both discrepancies tend toward zero. In addition, the standard deviation of the simulated number of intersections also tends toward zero as a function of $n^2/6$.

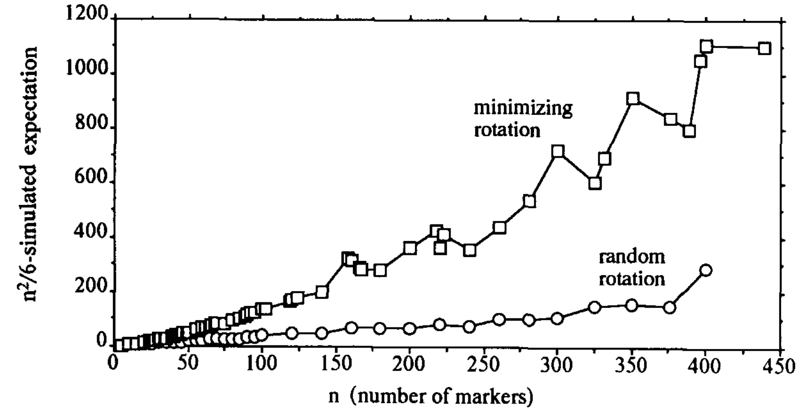

FIG. 5. Difference between $n^2/6$ and simulated number of intersections.

## Comparisons

Figure 6 shows the number of intersections in an optimally rotated alignment of each pair of genomes, divided by the simulated expected value under the null hypothesis of completely randomly shuffled genomes, as a function of *n*, number of markers in common to each pair (based on the first three symbols of the label). The error bars represent the standard deviation of the predicted number of intersections as estimated in the simulations, plus the standard deviation of the observed number of intersections as estimated from the 10 sets of markers with randomly matched duplicate labels, all divided by the simulated expectation.

The relationship between *E. coli* and *Salmonella* can be seen clearly in the figure with the number of intersections being only 20% of that expected under complete randomization. This is true despite an inversion of a large segment of the genome that is evident when comparing one organism with the other, probably contributing a large portion of the intersections.

The other comparisons, except some of those involving *B. subtilis,* also show fewer intersections (i.e., a greater relationship) than randomly shuffled genomes, though only those involving *Caulobacter* are convincing.

Figure 7 compares the four-symbol intersection rates with the three-symbol ones of Fig. 6.

In all cases, again with the exception of comparisons involving *B. subtilis,* the increased confidence in marker homology when all four letters of the corresponding labels must be identical is reflected in a much greater level of relationship as detected through a decrease in the normalized rate of intersection. It should be noted that the error ranges for the four-symbol comparisons are distinctly less than those for three-symbol comparisons

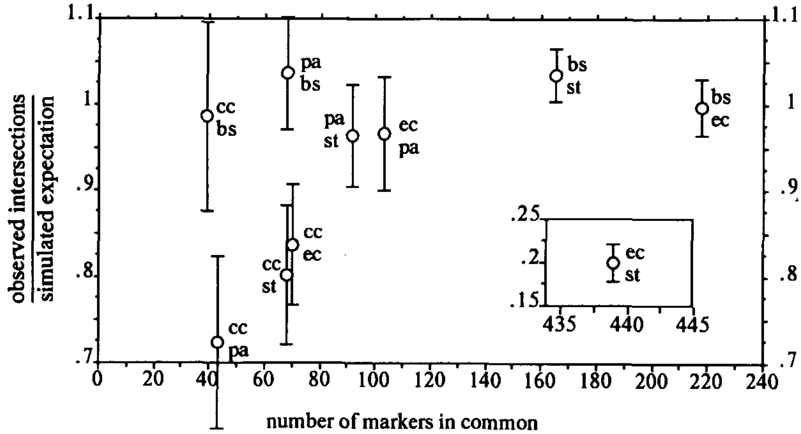

FIG. 6. Pairwise comparisons of genomes, ec, *Escherichia coli;* st, *Salmonella typhimurium;* bs, *Bacillus subtilis;* cc, *Caulobacter crescentus;* pa, *Pseudomonas aeruginosa.*

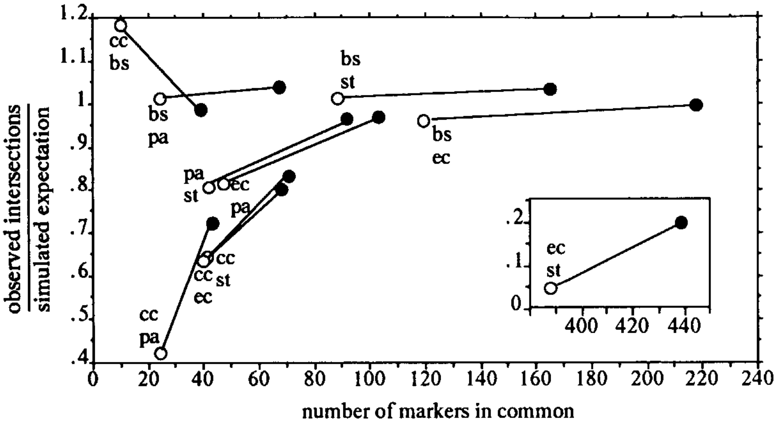

FIG. 7. Normalized intersection rates using three-symbol (filled dots) and four-symbol (open dots) correspondences.

for similar values of $n$, since there is little or no variation due to random correspondence of duplicate labels.

Both the three-symbol and four-symbol analyses permit us to make the same phylogenetic inferences. *E. coli* and *Salmonella* are closely related, *Caulobacter* and *Pseudomonas* are fairly closely related, and the first pair are more closely related to the second pair than either is to *B. subtilis*. In addition, the results suggest that *Pseudomonas* has been evolving at a faster rate than *Caulobacter*. None of these inferences is contrary to what is known about bacterial phylogeny.

The size of the largest marker subset with a common order in the two genomes, divided by $2\sqrt{n}$, gives the identical phylogenetic picture as does the normalized rate of intersection.

## Discussion

The genetic maps currently available contain too few markers in common to enable us to test our methods on organisms which are evolutionarily very divergent. The closely related (around 115 million years) *E. coli* and *Salmonella* genomes have many of the same markers mapped, and the relationships among these and *Caulobacter* and *Pseudomonas* can be assessed, but between all of these and *B. subtilis*, representing a time depth of about 800 million years, our measures show no detectable difference from a completely shuffled model. It does seem important, however, to develop this type of methodology in anticipation of the quantity of experimental results which are expected in the next few years.

Further mathematics and/or simulations are needed to understand the behavior of our measures based on optimal rotations, instead of under the (usually false) hypothesis that we know the original alignment of the two genomes.

Normalized rate of intersection and largest common ordered subset measures of divergence through rearrangement should be compared for sensitivity and accuracy to indices based on the least number of rearrangements necessary to convert one genome into another.[8]

Much work is needed to collate the labels for markers used in distantly related genomes.

Empirical and theoretical research on fragment transposition distances is needed to produce better models of this phenomenon. Generalizing the term "synteny" used in chromosomal genetics to encompass the notion of proximate fragments in one genome tending to be close together as well on another would give an appropriate label to this field of study.

## Acknowledgments

This work was supported by individual operating grants and to David Sankoff and Robert Cedergren, as well as an infrastructure grant and a CRAY computer time allotment, from the National Sciences and Engineering Research Council of Canada. David Sankoff and Robert Cedergren are Fellows in the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

[8] D. Sankoff, *Bul. Int. Stat. Inst.* **53(3),** 461 (1989).

# [27] Multiple Sequence Comparison

*By* DAVID J. BACON and WAYNE F. ANDERSON

## Introduction

One reason for performing amino acid sequence comparisons is to discover structural and/or functional similarities among proteins. Because structural similarity may be present in proteins that do not exhibit a strong sequence similarity (e.g., see Matthews *et al.*[1]), one would like to be able to recognize the structural resemblance even when the sequences are very different.

This chapter addresses the problem of finding weak similarities or distant relationships among proteins for which only the sequences are known. Comparing just two sequences at a time by current methods does

[1] B. W. Matthews, M. G. Grutter, W. F. Anderson, and S. J. Remington, *Nature (London)* **290,** 334 (1981).