Short inversions and conserved gene clusters

David Sankoff

Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale
Centre-Ville, Montréal, Québec, Canada H3C 3J7

Received on September 3, 2001; revised on September 6, 2001; accepted on October 24, 2001

ABSTRACT

Motivation: Two independent sets of recent observations on newly sequenced microbial genomes pertain to the prevalence of short inversion as a gene order rearrangement process and to the lack of conservation of gene order within conserved gene clusters. We propose a model of inversion where the key parameter is the length of the inverted fragment.

Results: We show that there is a *qualitative* difference in the pattern of evolution when the inversion length is small with respect to the cluster size and when it is large. This suggests an explanation of the lack of parallel gene order in conserved clusters and raises questions about the statistical validity of putative functionally selected gene clusters if these have only been tested against inappropriate null hypotheses.

Contact: sankoff@poste.umontreal.ca

INTRODUCTION

A striking difference in the patterns of genomic evolution between higher eukaryotes and more primitive organisms, including prokaryotes and yeast, emerges from examination of genomic sequences. Among those animals, or plants, that have been the subject of comparative mapping studies, the close evolutionary relationships between two species is manifested by relatively long *conserved segments*, regions of the chromosome with identical gene content and linear gene order in both (Nadeau and Taylor, 1984) In contast, two closely related prokaryotes typically share many *gene clusters*, sets of genes in close proximity to each other, but not necessarily contiguous nor in the same order in both genomes. These clusters, lacking the conservation of linear order, have been explained by functional selection (Overbeek *et al.*, 1999), operon formation (Bork *et al.*, 2000), horizontal transfer (Lawrence and Roth, 1996) and other evolutionary processes affecting gene content and gene order (Kolsto, 1997; Mushegian and Koonin, 1996; Huynen and Bork, 1998; Tamames *et al.*, 1997). This function- and selection-based discussion, however, lacks a well-developed neutralist null hypothesis against which to test its claims (Durand and Sankoff, 2002). For genomes containing thousands of genes, patterns of similar clusters could be remnants of ancestral configurations, not yet fully disrupted by regular evolutionary processes of genome rearrangement.

A separate set of observations concerns the greatly elevated frequency of inversions of short segments (containing one or a few genes) in genomic sequence in the evolution of microbial genomes (Dalevi *et al.*, 2000) and in lower eukaryotes as well (McLysaght *et al.*, 2000; Seoighe *et al.*, 2000). This contrasts to genetic and physical map patterns in plants and animals where such inversions seem less prevalent.

The observations about inversion length suggest a way of explaining the segment/cluster difference between higher eukaryotes and smaller genomes. We introduce a neutral probabilistic model of genome rearrangement where the lone parameter is the difference between d, the physical size of a gene cluster, and x, the scope of the elementary rearrangement operation. We focus on the effects of rearrangements in expanding the interval containing a fixed set of genes. We find a discontinuity at the point $d = x$ in how the model evolves over time. When $x > d$, the pattern is that of conserved segments, within which gene order is preserved. When $x < d$, the tendency is toward conserved clusters, within which gene order is shuffled. This ties together the microbial clusters versus higher eukaryote segments contrast and the reports of high rates of short inversion (small x) in sequenced genomes.

A CONTINUOUS ANALOG

For analytical simplicity, we construct continuous analogs of the genome and of gene order rearrangements.

We will study the case of unsigned (no strandedness information) circular genomes, though a parallel development could be made for signed genomes, or for multichromosomal genomes. The circularity serves to evoke the microbial genome; the only mathematical consequence in our analysis is to avoid having to deal with 'end effects'. The key notion will be the interval occupied by a cluster of genes.

To model rearrangement processes, we will first study inversions of fixed length. This allows us to highlight an essential discontinuity inherent in the model. We will then generalize to an arbitrary distribution of inversion lengths.

THE MODEL

Consider a circular genome of unit circumference and of an inversion operation of length $x \ll \frac{1}{2}$ placed at random on the circle. We will focus on what happens to the interval occupied by a cluster of genes, originally of length $d \ll 1$. If one of the endpoints of the inversion falls within the interval of length d and the other outside, then the inversion will change the length of the interval to d', as shown in Figure 1. Inversions, both of whose endpoints fall inside the interval or both fall outside the interval, do not affect the length of the interval.

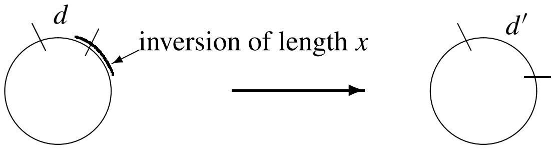

Fig. 1. Effect of inversion of length x in lengthening a cluster interval from d to d'

To calculate these changes we first state the following lemma, illustrated in the close-up in Figure 2.

LEMMA 1. *The probability that random intervals of lengths $x < 1$ and $d < 1$ overlap partially (neither is completely contained in, or completely disjoint from the other) equals* $2 \min[x, d]$.

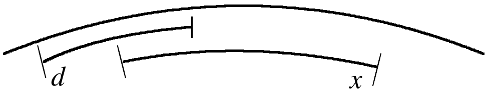

Fig. 2. Overlap of inversion of length x with cluster of length d.

Based on Lemma 1, the expression for the probability density of d' has two forms, depending on whether x or d is smaller:

$$p(d') = (1-2x)\delta(d) + 2xU[d, d+x], \text{ for } x \le d \quad (1)$$
$$p(d') = (1-2d)\delta(d) + 2dU[x, d+x], \text{ for } x > d \quad (2)$$

where $\delta(d)$ is a unit point mass at d and $U[s,t]$ is the uniform distribution on the interval $[s,t]$. These two densities are depicted in Figure 3 (a) and (b). The means for densities (1) and (2) are given by (3) and (4), repectively. These two densities are depicted in Figure 3 (a) and (b):

$$\mu = d + x^2, \text{ for } x \le d \quad (3)$$
$$\begin{aligned}\mu &= d + x^2 - (x-d)^2 \\ &= d + 2xd - d^2, \text{ for } x > d \quad (4)\end{aligned}$$

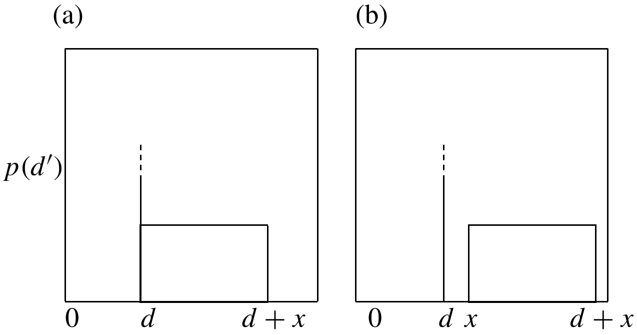

Fig. 3. Two forms of the density of $p(d')$: (a) $x \le d$. (b) $x > d$.

Note that intervals cannot shrink as an effect of inversion in this continuous model. In discrete reality, of course, such shrinkage is possible, and depends on the position of the genes within the interval and on where the endpoint of the inversion falls among these genes. Nevertheless, for intervals containing a good number r of relatively closely packed genes, and for x not trivially small, i.e. not too small with respect to $\frac{d}{r}$, interval expansion will greatly predominate over shrinkage.

EVOLUTION OF THE MODEL.

In the small inversion case, as we apply further random inversions of size x to the genome the cluster interval continues to expand, from d to $d', d'', \cdots, d^{(k)}$, etc. Denoting by μ_k the expected size of $d^{(k)}$, the interval after k inversions, we have

$$\mu_k = d + kx^2, \text{ for } x \le d \quad (5)$$

For very large inversions, say where x is two or more orders of magnitude greater than d, we can approximate (4) by

$$\begin{aligned}\mu &= d + 2xd - d^2 \\ &\approx d + 2xd \\ &= d(1+2x) \quad (6)\end{aligned}$$

and

$$\mu_k \approx d(1+2x)^k, \quad (7)$$

at least until the cluster interval grows large enough to invalidate approximation (6), i.e. until $d^{(k)}$ begins to be comparable in size to $2x$.

A MEASURE OF TOTAL EVOLUTION

It is clear that a very small inversion perturbs the global genomic structure less than a very large one does. It is not clear, however, how to quantify this. Many measures of genome divergence based on gene order have been proposed (Sankoff and El-Mabrouk, 2001), but there is little precedent for comparing divergences across contexts as different as the short inversion and the very long inversion models we are studying. A natural way in this context might be based on the average expansion of a cluster interval of size d in one genome when examined in the other, where d is the length of a typical cluster of interest.

Using this notion, we can investigate both models in the same framework and ask, for example, the number of inversions i it would take for an interval to grow to size h in the two cases, i.e. to evolve to the same extent. In the small inversions case, we substitute h for μ_i in (5)

$$i \approx \frac{h-d}{x^2}, \text{ for } x \leq d. \tag{8}$$

In the case of very large inversions, from (7)

$$i \approx \frac{\log h - \log d}{\log(1+2x)}, \text{ for } x >> d. \tag{9}$$

Comparing (8) where x is small and x^2 is extremely small, so that i represents a very large number of inversions, with (9) where x is large, so that i is a represents a relatively small number of inversions, gives us a clue as to why small genomes diverge differently from large ones. This will be discussed below.

INFERENCE

We remarked above on the transition from the long inversion to the short inversion regime, which should happen as we approach the point $d(1+2x)^k = x$. After this point, i.e. for higher values of k, the interval should grow much more slowly.

At a fixed point in time, i.e. at a fixed but unknown value of k, a plot of the average value of $\frac{d^{(k)}}{d}$ as a function of d, as in Figure 4, should reflect a high constant value ($\approx (1+2x)^k$) for small values of d (the 'initial' segment) and a low value ($\approx 1+\frac{kx^2}{d}$) for large values of $d \geq x$ (the 'final' segment). For intermediate values of d, the interval will have been expanding at the high rate from $k=1$ until the transition to the short inversion regime was reached (the 'transitional' segment). To the extent that left-hand end of the final segment can be discerned in the data, we can estimate

$$\hat{x} = \min\left\{d | \frac{d^{(k)}}{d} \approx 1 + \frac{kx^2}{d}\right\}. \tag{10}$$

The value $d = d^*$ at the right-hand end of the initial segment corresponds to intervals which, with the k-th inversion, are near the point of transition, i.e.

$$d^* = \max\{d | d(1+2x)^k \leq x\}, \tag{11}$$

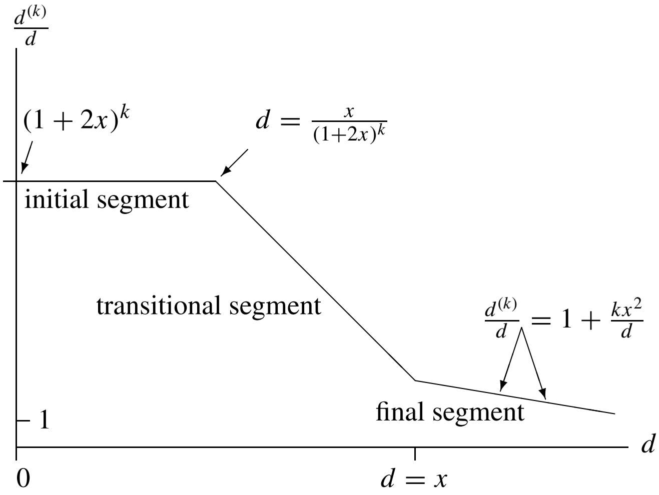

Fig. 4. Factor of expansion after k inversions of length x as a function of initial interval length d.

so that we can estimate

$$\hat{k} = \frac{\log \hat{x} - \log d^*}{\log(1+2\hat{x})}. \tag{12}$$

RANDOM INVERSION LENGTHS

Let f be an arbitrary density on $[0, l]$, where $l \ll \frac{1}{2}$. Then the generalization of (1) and (2) takes the form:

$$p(d') = \delta(d)\{1 - 2[F(d)\mu(f,d) + (1-F(d))d]\} + 2[F(d') - F(d'-d)], \tag{13}$$

where $\delta(d)$ is a unit point mass at d, F is the cumulative probability corresponding to the density f, and $\mu(f,d)$ is the mean of f conditioned on $x < d$. Of course formula (13) depends heavily on the specific nature of f, but its relation to (1) and (2) is relatively easy to state. In the short-inversion case, if f is non-zero only for $x < d$, we can show $E(d') = d + X$, where X is small and does not depend on d, while for long inversions, where f is concentrated on $x >> d$, then $E(d')$ behaves as $d(1+Y)$. This is essentially the same as (3) and (6). Where the mass of f is partly on one side of d and partly on the other, the resulting behavior will be a weighted combination of the two patterns.

INTERPRETING THE DIFFERENCE BETWEEN SHORT AND LONG INVERSION REGIMES

It can be seen that in the long inversion regime , only the (relatively few) inversions which expand the interval actually have an endpoint within the interval and thus have a potential effect on its internal gene order. On the other hand, where inversions are short, a good proportion of the inversions which do not expand the interval have *both* endpoints in the interval and thus can change the internal gene order. Thus for the same overall amount of evolution, measured by growth in cluster interval length, we can expect far more internal changes in gene order with short inversions than with long ones.

For long inversions, then, the model evolves over time to produce a pattern of conserved segments, within which gene order is preserved. For short inversions, the tendency is toward conserved clusters, within which gene order is shuffled. This ties together the microbial clusters versus higher eukaryote segments contrast and the reports of high rates of short inversion (small x) in sequenced genomes.

CONCLUSIONS

The simplifications inherent in our model enable us to discern a qualitative difference in the patterns of evolution under regimes of short and long inversions, with respect to gene clusters of a given size. At the same time, these simplifications preclude direct application to real data for the development of tests and other inferential procedures, though our procedures may well be meaningful in estimating an average or 'effective' inversion length.

Nevertheless, in the light of our results, it is clear that significance tests for clusters of genes in close proximity in two genomes against the null hypothesis of random gene ordering, cannot be justified simply by superficial observations of lack of conserved gene order. From the comparative mapping experience in eukaryotes, we might expect long chromosomal segments with conserved gene order in two genomes to become shuffled into shorter and shorter segments with the passage of time. In this context, the lack of any short segments could well be indicative of highly divergent genomes, randomly scrambled with respect to each other. On the other hand, under the short inversion regime, which may be more appropriate for microbial and lower eukaryote genomes, we expect extreme scrambling at the local level (gene order) together with considerable conservation of mid-level structure (gene clusters). Appropriate null hypotheses should thus take into account this conserved mid-level structure in random evolution models, in testing whether the conserved excess of gene clustering is significant enough to warrant conclusions such as the functional association between neighbouring genes.

Further work on the model will require returning to the discrete origins of the problem, and to characterize the transition between the two regimes in this context. Only then can useful tests of clustering be considered. The discrete model will also allow for interval shrinkage, strandedness and other aspects.

In the meantime, more focused statistics on the length of conserved clusters in pairs of sequenced genomes will provide the motivation for the development of analytical tools and tests.

ACKNOWLEDGMENTS

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. The author is a Fellow of the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

REFERENCES

Bork,P., Snel,B., Lehmann,G., Suyama,M., Dandekar,T., Lathe III,W. and Huynen,M. (2000) Comparative genome analysis: exploiting the context of genes to infer evolution and predict function. In Sankoff,D. and Nadeau,J.H. (eds), *Comparative Genomics*. Kluwer Academic Press, Dordrecht, pp. 281–294.

Dalevi,D., Eriksen,N., Eriksson,K. and Andersson,S. (2000) *Genome comparison: The number of evolutionary events separating* C. pneumoniae *and* C. trachomatis, Technical report, University of Uppsala.

Durand,D. and Sankoff,D. (2002) Tests for gene clustering. In Myers,G., Hannenhalli,S., Istrail,S., Pevzner,P. and Waterman,M. (eds), *RECOMB 2002 Proceedings of the Sixth Annual International Conference on Computational Biology*. ACM, New York, pp. 144–154.

Huynen,M.A. and Bork,P. (1998) Measuring genome evolution. *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.

Kolsto,A.B. (1997) Dynamic bacterial genome organization. *Mol. Microbiol.*, **24**, 241–248.

Lawrence,J. and Roth,J. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**, 1843–1860.

McLysaght,A., Seoighe,C. and Wolfe,K.H. (2000) *High frequency of inversions during eukaryote gene order evolution*, Comparative Genomics, Sankoff,D. and Nadeau,J.H. (eds), Kluwer Academic Press, Dordrecht, pp. 47–58.

Mushegian,A.R. and Koonin,E.V. (1996) Gene order is not conserved in bacterial evolution. *Trends Genet.*, **12**, 289–290.

Nadeau,J.H. and Taylor,B.A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA*, **81**, 814–818.

Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Sankoff,D. and El-Mabrouk,N. (2002) Genome rearrangement. In Jiang,T., Xu,Y. and Zhang,M.Q. (eds), *Current Topics in Computational Biology*. MIT Press, Cambridge, MA, pp. 135–155.

Seoighe,C., Federspiel,F., Jones,T., Hansen,N., Bivolarovic,V., Surzycki,R., Tamse,R., Komp,C., Huizar,L., Davis,R., Scherer,S., Tait,E., Shaw,D., Harris,D., Murphy,L., Oliver,K., Taylor,K., Rajandream,M.-A., Barrell,B. and Wolfe,K. (2000) Prevalence of small inversions in yeast gene order evolution. *Proc. Natl Acad. Sci. USA*, **97**, 14427–14432.

Tamames,J., Casari,G., Ouzounis,C. and Valencia,A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.*, **44**, 66–73.