BIOINFORMATICS
Vol. 1 no. 1 2003
Pages 1–7

Detection and validation of single gene inversions

J.F. Lefebvre[1], N. El-Mabrouk[1], E. Tillier[2] and D. Sankoff[3]

[1]Département d'informatique et recherche opérationnelle, Université de Montréal, CP 6128 succ. Centre-ville, Montréal, Québec H3C 3J7 Canada, [2]Ontario Cancer Institute, Princess Margaret Hospital, 620 University Avenue, Suite 703, Toronto, Canada and [3]Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa K1N 6N5, Canada

ABSTRACT

A model of genome evolution through random size inversions must be adapted to take into account an excess of short inversions, particularly those involving a single gene. This is demonstrated through comparison of four pairs of bacterial genomes, using a specially-designed implementation of the Hannenhalli-Pevzner theory, and validated through experimentation on pairs of random genomes matched to the real pairs.

INTRODUCTION.

The algorithmic study of genome rearrangement models (Sankoff and El-Mabrouk, 2002) has focused on minimizing the number of events in an inferred rearrangement history, without cost differentiation within the set of allowed operations. One exception is the DERANGE methodology (Blanchette *et al.*, 1996) for minimizing a weighted sum of inversions and transpositions. More recently, we proposed a general method that allows a choice among equally optimal solutions (i.e., the same minimal number of operations), based on any one of many possible secondary criteria (Ajana *et al.*, 2002).

There is an increasing interest in more detailed study of the rearrangement events, such as the proportion of inversions and translocations (Sankoff *et al.*, 1997, 2000), the position of the operation in the genome (Tillier and Collins, 2000; Ajana *et al.*, 2002; Samonte and Eichler, 2002), or the size of the chromosomal fragment involved (Nadeau and Sankoff, 1998; Sankoff *et al.*, 1997). More specifically, attention has been drawn to the prevalence and significance of short inversions (Dalevi *et al.*, 2002; McLysaght *et al.*, 2000; Sankoff, 2002; Mouse Genome Sequencing Consortium, 1999).

The extensive non-uniqueness represented in optimal reconstructions, as well as their tendency to be more economical than the true history, make it difficult to infer anything definitive about the individual evolutionary events, especially in cases of substantial genomic divergence. In this article we present a new approach to the study of short inversions in particular, taking advantage of a greatly elevated persistence that we demonstrate in their evolutionary signal, compared to that of longer inversions. We apply this to the reconstruction of the evolutionary divergence between relatively closely-related pairs of bacterial genomes, and discover an unexpectedly high number of single-gene inversions.

We adapt the method and software of Ajana *et al.* (2002), and show how to explore the space of optimal reconstructed histories, to detect recurrent patterns among the numerous alternate optima, to determine the conditions under which the algorithm reconstructs the true history, and to demonstrate the validity of the results via constructs with matched simulated inversion histories.

We then discuss the competing possibilities that the single inversions represent a particular evolutionary mechanism with selective functional consequences, that they are the clearest manifestation of a universal tendency toward short inversions as the least disruptive of the gene proximity configuration of a genome, or that they are simply an artifact of incorrect identification of orthologues prior to genome comparison. In the latter case, this methodology becomes a powerful tool for the identification of orthologues by means of gene order considerations, as advocated, for example, in Sankoff (1999).

REARRANGEMENT HISTORY VIA SAFE INVERSIONS

Given two genomes $G = (1, \cdots, n), H = (h_1, \cdots, h_n)$ where† $H = E \times \Pi G$ for some permutation Π of G, and E is some n-vector of 1's and -1's, the Hannenhalli-Pevzner (**HP**) algorithm (Hannenhalli and Pevzner, 1995) computes the minimal number of inversions‡ required to transform G to H, and outputs one minimal sequence of inversions.

The HP algorithm is based on a bicoloured graph

† The $\times$ symbol indicates component-wise multiplication
‡ An inversion transforms $(i_1, \cdots, i_n)$ to $(i_1, \cdots, -i_j, -i_{j-1}, \cdots, -i_h, -i_h, \cdots, i_n)$ for some $1 \leq h \leq j \leq n$.

defined as follows. Replace gene x_i in G by the pair $x_i^t x_i^h$, and similarly in H, except that if $E_i = -1$, then the $-x_i$ is replaced by $x_i^h x_i^t$. The vertices of the graph are just the x_i^t and the x_i^h. The adjacent vertices in G, other than x_i^t and x_i^h from the same x_i, are connected by a red edge, and any two adjacent in H, by a blue edge. The key concept is the decomposition of this graph into disjoint color-alternating cycles, and into **oriented** and **unoriented components** (for this and other details about the breakpoint graph and the HP algorithm, see Setubal and Meidanis (1997)).

The problem of minimizing the number of inversions can be reduced to the one of increasing the number of cycles as fast as possible. As an inversion can increase by at most one the number of cycles of a graph (a **good inversion**), the problem is to perform as many good inversions as possible. A **safe inversion** is a good inversion that does not create any new unoriented component. HP proved that an oriented component can be transformed to a set of cycles of size 1 (one red edge and one blue edge) by a sequence of safe inversions. As for unoriented components, some of them can still be solved by safe inversions, whereas others, called **hurdles**, cannot.

For a graph with only oriented components, a sequence of inversions is thus a minimal solution if and only if it contains exclusively safe inversions. In this case, the problem of generating all minimal solutions reduces to the problem of generating all the safe inversions at each step of the sorting procedure. The bottleneck of the HP theory is to generate safe inversions. Different methods have been developed (Hannenhalli and Pevzner, 1995; Berman and Hannenhalli, 1996; Kaplan *et al.*, 1999; Bergeron, 2001; Ajana *et al.*, 2002) that output one or few safe inversions at each step of the HP algorithm, the most efficient one being of linear time complexity (Kaplan *et al.*, 1999). However, there is yet no efficient method to characterize the whole set of safe inversions. In our method, the total time required to find all safe inversions at each step of the HP procedure is in $O(n^3)$.

If, on the other hand, the graph contains hurdles, these have to be transformed into oriented components. This is done by inversions that merge and cut hurdles in a particular way, described in the HP procedure for **clearing** hurdles (Hannenhalli and Pevzner, 1995). This procedure allows some flexibility in choosing which hurdles should be treated first, and how. To find all possible sequences of inversions transforming genome G into H, all ways of clearing hurdles must be considered (Siepel, 2002). In this worst case, this could be prohibitive. However, there may be reasonable constraints on the inversions to be considered. In this paper, for example, we are interested in finding the shortest inversions, which dramatically cuts down on the effort required to clear the hurdles. Moreover, in real genomes and even in systematic simulations, hurdles are almost never encountered.

Our algorithm has two major steps:

1. If the graph contains hurdles, perform a set of inversions clearing the hurdles.
2. Solve each oriented component independently, by choosing safe inversions.

The versatility of this method lies in the many possible strategies for carrying out the second step. For example, if there is a criterion for weighting inversions, we can find a solution of minimal cost.

If we have no particular weighting scheme, we might wish to search for inversions or kinds of inversion that recur in all or most solutions, as in the next section. For this, a set of minimal solutions can be obtained by selecting, at random, one possible safe inversion at each step of the HP procedure. Running the algorithm several times gives rise to several possible solutions. We can then tabulate how many times particular inversions recur in the set of solutions.

Finally, as in the section dedicated to short inversions, we may be motivated to favor inversions of a given size.

PAIRS OF BACTERIAL GENOMES

We chose four pairs of related bacterial genomes with a range of gene sequence and gene order divergence. The protein sequences were obtained from NCBI ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ for the following genomes: *Chlamydia trachomatis* (NC 000117) and *Chlamydophila pneumoniae* AR39 (NC 002179), *Bacillus halodurans* (NC 002570) and *Bacillus subtilis* (NC 000964), *Escherichia coli* K12 (NC 000913 and *Vibrio cholerae* chromosome I (NC 002505), *Streptococcus pyogenes* MGAS8232 (NC 003485) and *Staphylococcus aureus* subsp.*aureus* Mu50 (NC 002758). Orthologous gene pairs were identified as respective significant $E < 0.01$ best hits in all-by-all Fasta analysis.

In comparing genomes by rearrangement distances, the two genomes must first be reduced by removing all genes that are not in common. But in calculating inversion lengths, we restore the genes previously removed from G, i.e., only one of the genomes. This introduces some arbitrary asymmetry into the comparison, but has no statistical effect in this study.

Applying our algorithm to the four pairs of bacterial genomes, repeating each comparison ten times with different random choices of safe inversion, we arrive at the distribution of inversion sizes summarized in Figure 1.

To control for possible biases in our algorithm, we also show in Figure 1 the analogous results based on simulated pairs of genomes, generated by inversions of random size, positioned randomly along the genome length. The

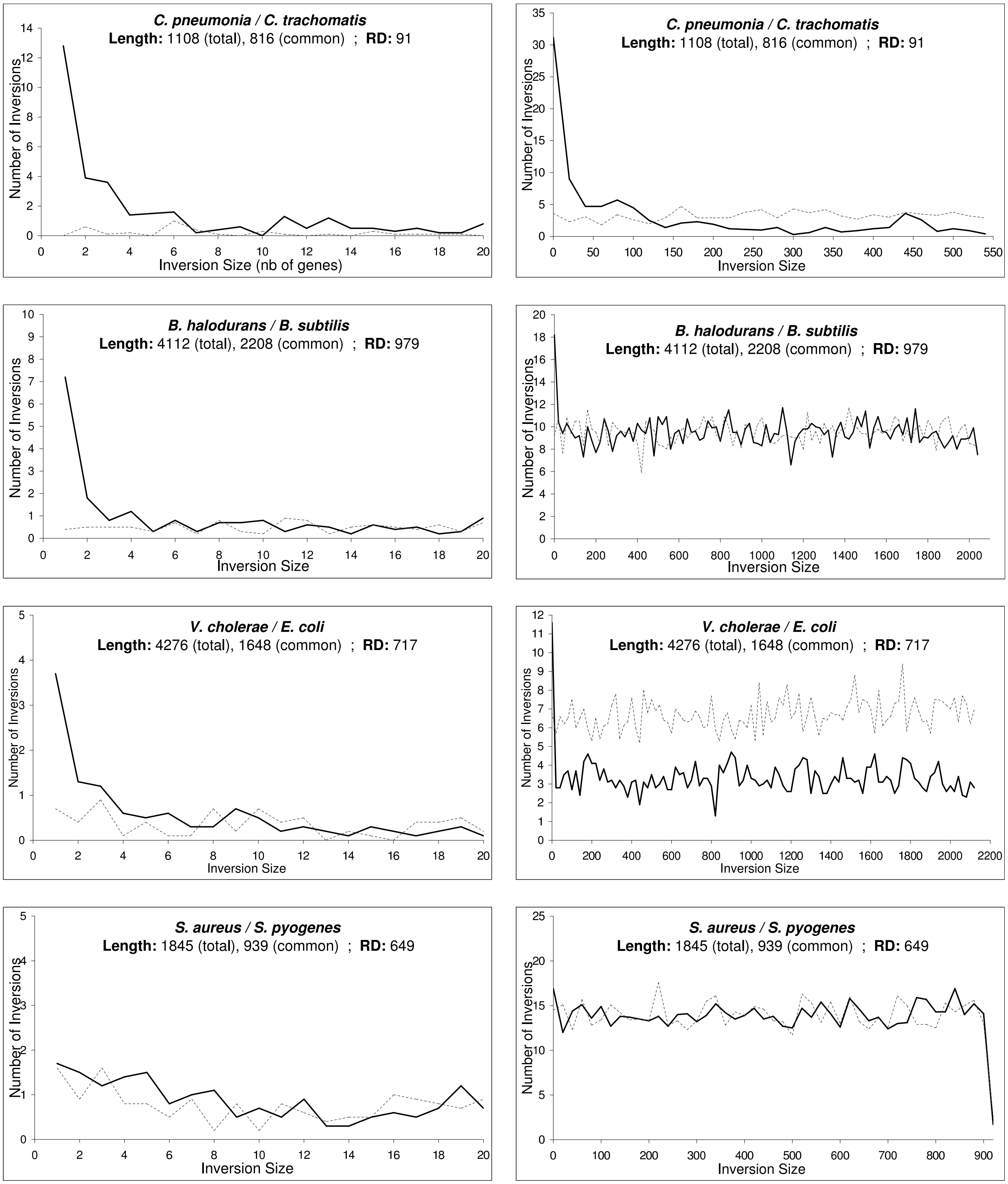

Fig. 1. Frequency (vertical axis) of inversion sizes (horizontal axis) inferred in comparisons of bacterial genomes. The solid line indicates the result of analyzing real genomes, the dotted line the genomes generated by randomly-placed inversions (reversals) of random size. For each pairwise comparison, the number of inversions (reversal distance, RD) and the length of the genomes (total genes as counted in both genomes and genes in common) are given. The left-hand diagrams focuses on small inversions, the right one the whole range. Results averaged over ten runs of the algorithm.

number of inversions were chosen to give approximately the same number of *inferred* inversions as in the real genome pair.

We note that in three of the four comparisons, there is a distinct surfeit of short inversions in the pairs of real genomes, particularly of inversions of single genes. From simple probabilistic arguments, we can predict about one single-gene inversion per genome comparison, if the genomes are randomly permuted, though this is not entirely the case here. Note as well the apparently uniform distribution of inversion length, as can be predicted for random genomes.

DECAY OF EVOLUTIONARY SIGNAL AS A FUNCTION OF INVERSION SIZE

We may ask to what extent the optimal reconstructed histories actually reflect the true evolutionary history. It is well-known that past a threshold of θn, where n is the number of genes and θ is in the range of $\frac{1}{3}$ to $\frac{2}{3}$, the *number* of operations begins to be underestimated by edit operation-based inferences. But whether any signal is conserved as to the actual individual operations themselves, and which ones, is a different question.

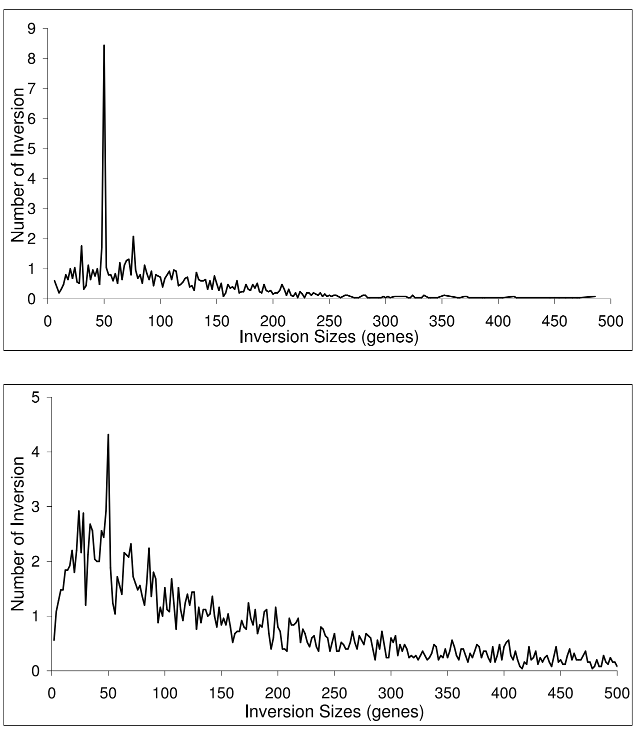

Fig. 2. Frequency (vertical axis) of inversion sizes (horizontal axis) inferred by the algorithm for random genomes obtained by performing i inversions of size $l = 50$. The figure on the top is for $i = 80$ and the bottom one is for $i = 200$.

We carried out the following test: For a genome of size $n = 1000$, we generated i inversions of size $l = 5, 10, 15, 20, 50, 100, 200$ at random, and then reconstructed the optimal inversion history, for a range of values of i. Typically, for small enough values of i, the algorithm reconstructs the true inversion history, although inversions that do not overlap may be reconstructed in any order. There are other sources of non-uniqueness that do not obscure the i generating inversions, either. Depending on l, however, above a certain value of i, the reconstructed inversions manifest a range of sizes, as illustrated in Figure 2, reflecting the decay of the "evolutionary" signal.

For each l, we calculated

$r_l = \min\{i|\text{reconstruction has at least } 5\% \text{ error}\}$

and

$s_l = \max\{i|\text{reconstruction has at most } 95\% \text{ error}\}$,

where any inversion having length different from l is considered to be an error. Figure 3 plots r and s as a function of l and shows how quickly the detailed evolutionary signal decays for large inversions. Nevertheless, we note that for very small inversions, there is a clear signal preserved long after longer ones have been completely obscured, at least in this experimental context.

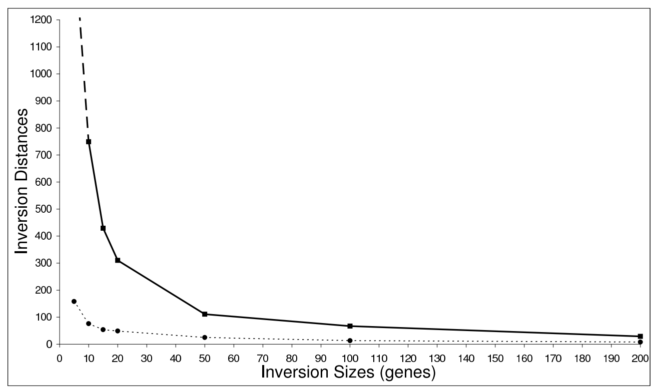

Fig. 3. The solid line corresponds to the plots for s, and the dotted line to the plots for r (see text above).

IMPOSING A BIAS TOWARDS SHORT INVERSIONS

The possibility of more accurately reconstructing small inversions suggests that we learn something by incorporating a bias in our algorithm towards the choice of small inversions. We may thus over-estimate the frequency of such inversions, but this can be controlled by using the same technique on pairs of randomly generated genomes differing by the same number of inversions. The results are shown in Figure 4. Note that the forced choice of smallest inversion removes most of the non-uniqueness of the solution, so that multiple runs do not add any information.

Figure 4 shows that the discrepancy between the real

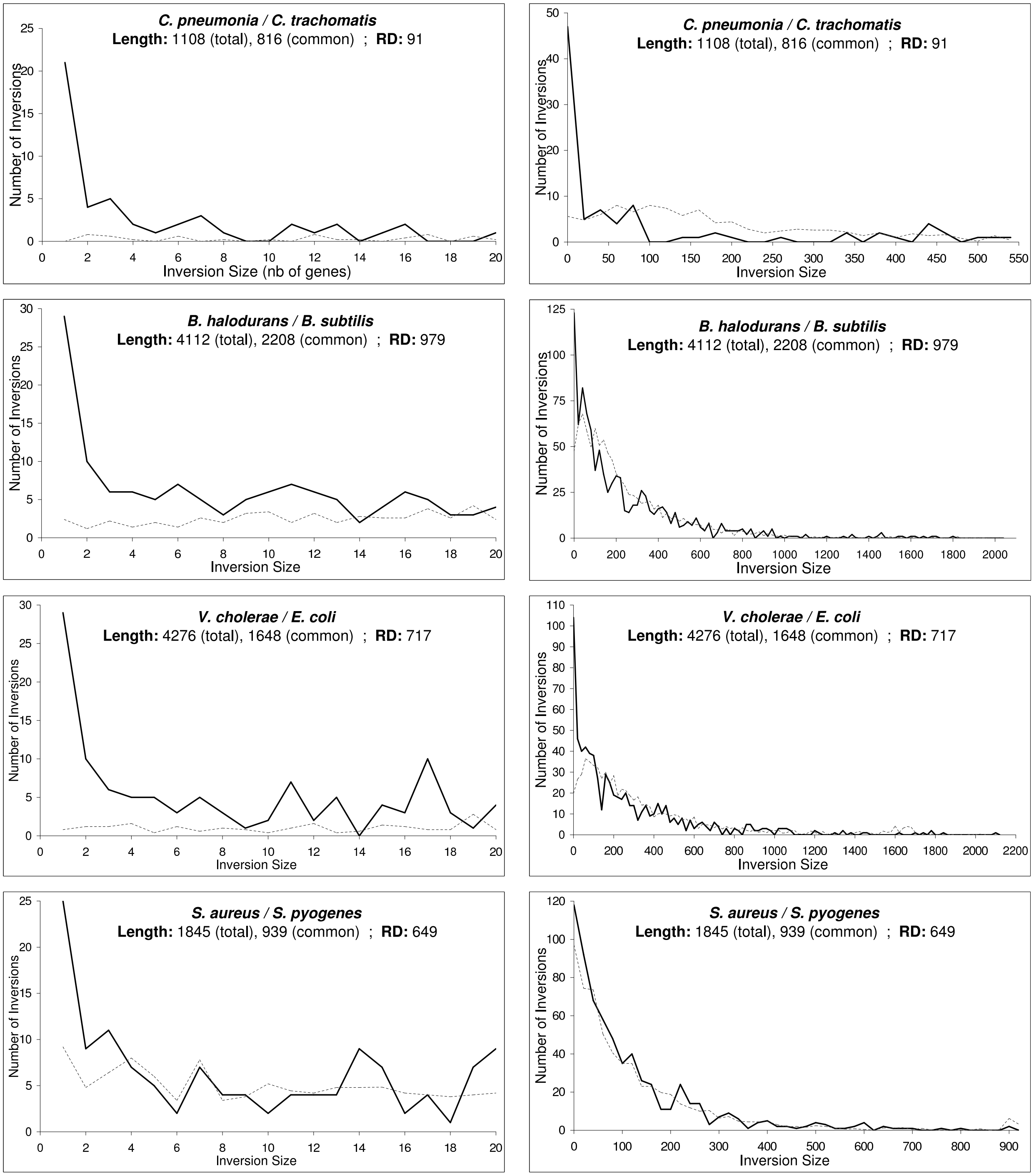

Fig. 4. Results from the same analysis as in Figure 1, but where at each step of the algorithm, the shortest safe inversion is chosen.

data and the simulated genomes becomes much clearer under the shortest inversion regime, and shows up even in the *Staphylococcus-Streptococcus* comparison.

The fact that random comparisons, despite the bias away from long inversions, do not show any appreciable increase in inversions of size 1, or short inversions ($l < 20$) in general[§], whereas the comparisons of real genomes do, suggests that the additional single-gene and other short

[§] with the exception of the relatively highly scrambled *Staphylococcus-Streptococcus* comparison

Genomes	single-gene inversions	b -c d	b I -c I d	u -c v	u I -c I v
C.p - C.t	21	14	4	1	2
B.h - B.s	29	8	4	7	10
V.c - E.c	29	8	8	6	7
S.a - S.p	25	2	4	12	7

Table 1. Genomic context of genes undergoing single-gene inversions. The inverted gene is denoted by c. The context b -c d indicates that c has changed sign from G to H but retains at least one immediate neighbour. The context u -c v indicates that c is not adjacent to any of the same genes in G and H. I indicates genes present in G but not in H.

inversions we have picked up through the introduction of the bias in our algorithm do in general reflect genuine events.

INFERRED VERSUS "OBSERVED" INVERSIONS

Some inversions are predictable through examination of the two genomes: if $\cdots abcde \cdots$ appears in one genome and $\cdots ab$ -$cde \cdots$ in the other, then no analysis is necessary to show that there has been a single-gene inversion. Table 1 contains an analysis of the contexts of the genes inferred to have occurred in single-gene inversions.

In the case of inversions that are not immediately evident by direct comparison of the two genomes, there are at least two kinds of explanation.

One possibility is that they represent genes that are not only inverted, but also moved by some other mechanism (called transposition or translocation) to completely different contexts in at least one of the two genomes. Since our algorithm uses inversions only, to invert and move a gene to a new position without disrupting its original context requires two inversions, the first to move the gene (and one side of its context) to the new position, and the second to restore the the original context, minus the gene in question. To the extent that this explanation is valid, it does not detract from our discovery of high rates of single-gene inversion. That single genes are so frequently displaced from their transcriptional context is the biologically interesting aspect of the rearrangement phenomenon we are studying.

The second possibility is that some of the genes that must be inverted in transforming G to H may have been identified as orthologues erroneously. It would not require a high rate of error to produce an artifactual result of this magnitude. While this possibility merits further examination on a gene-by-gene basis, it could account for only a fraction of the inversions we identified. First, it is only applicable to genes in the last two columns of Table 1. Second, of all the 104 single-gene inversions we studied, only 14 cases had a secondary hit in the FASTA identification of orthologs with sequence identity within 5% of that for the best hit, of which only 10 cases were in the last two columns of Table 1. Thus at worst, as many as 10% of the apparent inversions were due to potentially mistaken ortholog identification. Finally, the results in the previous section show that the excess of single-gene inversions is but the extreme case of a general tendency towards shorter inversions in the comparison of real genomes.

CONCLUSIONS

The single-gene inversions may represent a particular evolutionary mechanism with selective functional consequences. They may allow a gene to obtain transcriptional independence from its erstwhile operon, or to take advantage of new or altered functionality or to participate in a different pathway through a more appropriate genomic positioning.

Alternatively, single-gene inversions may simply be the clearest manifestation of a universal tendency towards short inversions as the least disruptive of the gene proximity configuration, and attendant functionality, of a genome. In Sankoff (2002), we argued that a predisposition for such inversions in small genomes might explain the prevalence of "gene clusters" found across many sequenced genomes in microorganisms, in contrast to the shuffled "conserved segments" pattern characteristic of the higher eukaryotes.

Finally, apparent single-gene inversions may be artifacts of incorrect identification of orthologues prior to genome comparison. In this case, our methodology becomes a powerful tool for the identification of orthologues by means of gene order considerations, as advocated, for example, in Sankoff (1999). Gene-by-gene study is underway to investigate this question.

ACKNOWLEDGMENTS

Research supported by grants from the Natural Sciences and Engineering Research Council (NSERC), and the *Fonds québécois de recherche sur la nature et les technologies*. DS holds the Canada Research Chair in Mathematical Genomics. NE-M, ET and DS are affiliated with the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

REFERENCES

Ajana,Y., Lefebvre,J.F. and Tillier,E. (2002) Exploring the set of all minimal sequences of reversals -An application to test the replication-directed reversal hypothesis. *Lecture Notes in Computer Science, Algorithms in Bioinformatics, Second International Workshop, WABI*. Guigó,R. and Gusfield,D. Eds., Springer Verlag, **2452**, pp. 300–315.

Bergeron,A. (2001) A very elementary presentation of the Hannenhalli-Pevzner theory. *Lecture Notes in Computer Science, CPM'01*. Springer Verlag, pp. 106–117.

Berman,P. and Hannenhalli,S. (1996) Fast sorting by reversals. *Lecture Notes in Computer Science, CPM'96*. Springer Verlag, pp. 168–185.

Blanchette,M., Kunisawa,T. and Sankoff,D. (1996) Parametric genome rearrangement. *Gene-Combis*, **172**, 11–17.

Dalevi,D.A., Eriksen,N., Eriksson,K. and Andersson,S.G. (2002) Measuring genome divergence in bacteria: a case study using chlamydian data. *Journal of Molecular Evolution*, **55**, 24–36.

Hannenhalli,S. and Pevzner,P.A. (1995) Transforming cabbage into turnip. *Proc. 27th Annu. ACM-SIAM Symp. Theory of Comp.*, 178–189.

Kaplan,H., Shamir,R. and Tarjan,R.E. (1999) Faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal on Computing*, **29:3**, 880–892.

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, N, pp. 47–58.

Mouse Genome Sequencing Consortium, (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Nadeau,J.H. and Sankoff,D. (1998) The lengths of undiscovered segments in comparative maps. *Mammalian Genome*, **9**, 491–495.

Samonte,R.V. and Eichler,E. (2002) Segmental duplications and the evolution of the primate genome. *Nature Reviews Genetics*, **3**, 65–72.

Sankoff,D. (1999) Genome rearrangement with gene families. *Bioinformatics*, **15**, 909–917.

Sankoff,D. (2002) Short inversions and conserved gene clusters. *Bioinformatics*, **18**, 1305–1308.

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

Sankoff,D., Ferretti,V. and Nadeau,J.H. (1997) Conserved segment identification. *Journal of Computational Biology*, **4**, 559–565.

Sankoff,D., Parent,M.N. and Bryant,D. (2000) Accuracy and robustness of analyses based on numbers of genes in observed segments. *Comparative Genomics*. Sankoff,D. and Nadeau,J.H. Eds., Kluwer Academic Press, Dordrecht, N, pp. 299–306.

Sankoff,D., Parent,M.N., Marchand,I. and Ferretti,V. (1997) On the Nadeau-Taylor theory of conserved chromosome segments. *Lecture Notes in Computer Science, CPM'8th Annual Symposium*. Apostolico,A. and Hein,J. Eds., Springer Verlag, **1264**, pp. 262–274.

Setubal,J. and Meidanis,J. (1997) *Introduction to Computational Molecular Biology*. PWS Publishing Company, pp. 215–244.

Siepel,A.C. (2002) An algorithm to find all sorting reversals. *RECOMB'02*. ACM, pp. 281–290.

Tillier,E.R.M. and Collins,R. (2000) Genome rearrangement by replication-directed translocation. *Nature Genetics*, **26**, 195–197.