Sequence analysis

Expected gene-order distances and model selection in bacteria

Daniel Dalevi and Niklas Eriksen

1Department of Computing Science and Engineering, Chalmers University of Technology, SE-412 96 Göteborg and 2Department of Mathematical Sciences, Göteborg University and Chalmers University of Technology, SE-412 96 Göteborg, Sweden

Received on October 25, 2007; revised on March 25, 2008; accepted on March 27, 2008

ABSTRACT

Motivation: The evolutionary distance inferred from gene-order comparisons of related bacteria is dependent on the model. Therefore, it is highly important to establish reliable assumptions before inferring its magnitude.

Results: We investigate the patterns of dotplots between species of bacteria with the purpose of model selection in gene-order problems. We find several categories of data which can be explained by carefully weighing the contributions of reversals, transpositions, symmetrical reversals, single gene transpositions and single gene reversals. We also derive method of moments distance estimates for some previously uncomputed cases, such as symmetrical reversals, single gene reversals and their combinations, as well as the single gene transpositions edit distance.

Contact: ner@math.chalmers.se

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Forty years ago Kimura (1968) proposed, in an influential paper, that the majority of all mutant substitutions occurring at the molecular level are selectively neutral. A consequence of his neutral theory (Kimura, 1983) is that sites in DNA sequences that evolve without constraints will have rates of change identical to the actual mutation rate of the organism. If constant, mutations will accumulate at a pace of an evolutionary clock and the number of changes defines the evolutionary distance (Woese, 1987) between organisms. This distance is highly significant for our knowledge on how organisms relate to each other—both at present and in the past.

The early models of evolution treat all types of substitutions equal and assume that all sites in a DNA or protein sequence evolved with the same rate, which would be the case if they were all neutral and under the same mutational pressure. However, as it appears, there are many selective constraints even at positions that are expected to be neutral, such as synonymous codon sites, resulting in a whole spectrum of different rates. At present there exist big families of nested models for maximum likelihood based phylogeny and there are statistical hypothesis tests for selecting among those models (Posada and Buckley, 2004).

A process that would be expected to be neutral, and thus has been less subject to model selection, is the rearrangement of the genes within the genome (Sankoff and Nadeau, 2003). Though there are clusters of genes that are cotranscribed, such as operons (Lawrence, 1997), the gene order should be more or less interchangeable. As it appears, however, there are several selectional constraints that seem to act on bacterial chromosomes. Rebollo et al. (1988) made experiments on Escherichia coli where reversals of segments were performed using a system for in vivo selection of genomic rearrangements. They found several phenotypic constraints. An important discovery was that reversals tend to preserve a gene’s distance to the origin of replication (Niki et al., 2000). Therefore, symmetrical reversals occurring around the origin (ori) or terminus (ter) of replication ought to be much more frequent than others. This can be confirmed with so called dotplots between pairs of closely related bacteria where the gene order is visualized in a graph. Figure 1A illustrates a dotplot of two bacteria with identical gene order and Figure 1B a dotplot resulting from a symmetrical reversal. Successive symmetrical reversals result in a pattern that resembles an ‘X’ and has been called both X-plots and X-files (Eisen et al., 2000; Mackiewicz et al., 2001; Tillier and Collins, 2000).

Another type of gene rearrangements that was shown to be over-represented are single gene reversals and transpositions (Dalevi et al., 2002; Lefebvre et al., 2003; Miklós and Hein, 2005; Sankoff, 2002). Sankoff et al. (2004) found that the distribution of reversal lengths could be approximated by a gamma distribution with shape-parameter $\alpha = 0.60$ and $\beta = 1200$. Attempts have been made to compute distances with length weighted reversals (Bender et al., 2004).

A consequence of a single gene reversal, and other reversals that preserve the direction of transcription relative to replication, is that the gene will swap from leading to lagging strand (or vice versa). This may increase the nucleotide mutation rates because of a mechanism called GC skew (Frank and Lobry, 1999).

In this study we justify the use of different gene-order models by studying the shape and appearance of dotplots acquired from pairwise comparisons of bacteria. We recapitulate minimal and expected distances for the known cases and derive new expected distances for the symmetrical and short reversals together with the minimal distance for short transpositions. We evaluate the performance and link the different
distances to their resulting dotplots. It appears, as in the case of mutations at the nucleotide and protein level, that several models are required to explain the evolution of gene orders in bacteria.

2 DATA ANALYSIS
All pairs of bacteria used in this study were downloaded from the NCBI ftp site of bacteria (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Orthologous gene pairs were identified as the best bi-directional hits using BLASTP with an E-value of at least 0.01. Duplicated genes were totally removed from the datasets.

3 CLASSIFICATION OF MODELS
The data analysis described above resulted in a variety of dotplots with several different shapes. Most of these plots can be partitioned into a few categories based on their appearance. This section describes these appearances and how we can generate similar shapes using the known gene-order operators (reversals, transpositions, symmetrical reversals, short transpositions, etc.).

3.1 The whirl
The whirl, shown in Figure 2A, is the common uniform reversal model. The name stems from the dotplot resulting after a few reversals. Some segments appear in an unordered fashion and others as either ascending or descending lines around the center of the plot. The pattern looks like a whirl or a waterspout.

This model is very well explored (Bergeron et al., 2005). This pattern is, however, easily distorted as it does not take many reversals to make it totally unorganized, in which case the model identification fails. One way to separate it from the other models including transpositions, for instance, is to look for hurdles in the breakpoint graph (Section 4.1). These are very rare in the reversal model (Caprara, 1999), but likely to appear if transpositions are used.

Fig. 1. The distance to ori and the direction of transcription, compared to the direction of replication, are preserved under symmetrical reversals. (A), dotplot of two bacteria with identical gene orders. (B), after a symmetrical reversal. The (C) and (D) show the two strands of a circular bacterial chromosome. Two genes are drawn with the direction of transcription. The direction of replication is shown with the arrows on the circles.

Fig. 2. Real data for different pairs of species, each showing the typical appearance of the respective models. A point with coordinates \((k, m)\) indicates that the gene at position \(k\) in the first genome has position \(m\) in the second. In the Supplementary Material, these dotplots are color-coded to separate genes with positive and negative orientation, respectively. The squares highlight short reversals, which are common in (B). The models depicted are: (A) the whirl (Bordetella bronchiseptica/Bordetella parapertussis), (B) the X-model (Chlamydia trachomatis/Chlamydomphila pneumoniae AR39), (C) the fat X-model (Mycobacterium bovis/Mycobacterium leprae), (D) the zipper (Escherichia coli CFT073/Shigella dysenteriae) and (E) the cloud (Bacillus halodurans/Bacillus subtilis).

3.2 The X-model
The X-model, shown in Figure 2B, has emerged quite strongly in recent years (e.g. Eisen et al., 2000; Mackiewicz et al., 2001; Tillier and Collins, 2000). The dotplot is shaped like an ‘X’ when the origin or terminus of replication is placed at the origin of the graph. As shown in Figure 2B, such a graph can easily be obtained by performing successive symmetrical reversals. Contrary to the whirl, the ‘X’ pattern is stable and remains visible no matter how many symmetrical reversals are applied. This is one reason why it occurs so often in data comparisons of bacterial species.

The dotplot of the Chlamydia does not only display an ‘X’. There are also many spots of red in the blue region and vice versa. These cannot be explained by symmetrical reversals and are most likely due to single gene reversals. When they are frequent we name the model the spotted X.

3.3 The fat X-model
A perfect X-model has a sharp ‘X’ in the dotplot. But many dotplots show a wider ‘X’. This could to some extent be explained by: (1) the difference in size between genes,
(2) the removal of genes that are not similar enough to be paired with their orthologs in the other genome and (3) symmetrical reversals are not perfectly symmetrical. The width displayed in Figure 2C, however, is hard to explain from these sources. We name this model the fat X-model. Using simulations we could confirm that this pattern is produced if the symmetrical reversals are accompanied by short transpositions. These transpositions remove a fairly short segment and reinserts it at another location. These segments should not be longer than about 5% of the genome length.

3.4 The zipper

The zipper is characterized by a long line with short perpendicular lines along it. It may appear on its own or in combination with symmetrical reversals, as in Figure 2D. We could easily confirm with simulations that this pattern is created using short reversals, up to about 5% of the genome length, equally distributed along the genome. Since the reversals are quite short, the pattern is visible for a longer time than the whirl. The difference between a zipper and, for example, the X-model is mainly in which regions the reversals are allowed to occur.

3.5 The cloud

While the structure of the dotplot in almost all models disappears in a cloud as time progresses, we may also have a cloud superposed on a structure, for instance an ‘X’ as in Figure 2E. While the ‘X’ is a bit tortuous, it is still clearly visible. What, then, is the origin of the cloud?

We believe this pattern is created using a few symmetrical reversals and a lot of single gene transpositions; simulations using these operations create a cloudy background without really disturbing the primary pattern.

In Figure 2E, it is clear that some parts of the genome omit and receive genes more frequently than others. While we cannot give a full explanation to why this happens, and which parts are more prone to these mutations, we suspect that this is a key towards understanding the nature of gene-order mutations.

It is important to realize that the cloud can be superposed on virtually any other pattern. In fact, single gene transpositions are present in most other examples presented in Figure 2, but to a lesser extent. It seems that these may be as frequent as single gene reversals, which at present have received much more attention in the literature.

4 EVOLUTIONARY DISTANCES OF THE MODELS

There are many ways to measure the distance between two gene orders. We model the genomes as signed, circular permutations. Thus, the genome [1 3 −2] can also be written [3 −2 1] (rotated) or even [−3 −1 2] (flipped over, which causes all genes to change strand relative to the viewer). The simplest way to compare two genomes is to look for some measure of dissimilarity, such as the number of adjacent genes in one of the genomes that are not adjacent with the same orientation in the other. This is known as the number of breakpoints and is easily computed. For instance, the number of breakpoints between [1 2 3 4 5] and [1, −3 −2, −4, 5] is 3 (marked in the second genome). We may also define the distance as the minimal number of operations needed to transform one of them into the other. We refer to this as an edit distance. The computability of these distances depends heavily on which operations we allow.

The operations we consider here are the most common operations in bacterial gene-order permutations, namely reversals (or inversions) and (block) transpositions. A reversal consists of reversing the order of a segment of genes, as well as changing their signs. A transposition consist of moving a segment of genes to another location in the genome, and possibly performing a reversal on that segment. That combination is known as an inverted transposition or transversion. We use the term transposition for both kinds of transpositions.

A third distance measure option comes from combining the two approaches above. Starting with two identical gene orders and applying random operations to one of them, the expected number of breakpoints (or any other distance such as edit distances) between them generally increases in a non-linear fashion with the number of operations we apply. We wish to deduce the number of applied operations from the number of breakpoints using this function. The method of moments estimate of the number of operations applied is obtained by feeding the number of breakpoints into the inverse of the expected breakpoint distance after \( t \) operations. These estimates are sometimes known as expected distances.

In gene orders of closely related species, both the edit distances and the method of moments give good estimates of the true number of operations. In distantly related species, though, the edit distances tend to underestimate the true number of operations, in contrast to the method of moments, which gives relevant information over longer time spans (Eriksen and Hultman, 2004).

In this section, we recapitulate old and derive new results for edit and expected distances of signed gene-order permutations. These models are usually too simple to explain the full evolutionary scenario with respect to the dotplots in Section 3, but they can still provide, either on their own or in combination, useful estimates of the bacterial evolutionary distance.

4.1 The whirl: uniform reversal distribution

Among the classical results in this area of research is the formula for the reversal distance by Hannenhalli and Pevzner (1999),

\[
d_{ev}(\pi) = n - c(\pi) - h(\pi) + f(\pi),
\]

where \( n \) is the number of genes and \( c(\pi) \) the number of cycles in the associated breakpoint graph. For most genomes created using reversals, the other terms (hurdles and fortress) are both equal to zero (Caprara, 1999). The computation of the reversal distance can be made in linear time and an up-to-date summary of all relevant aspects is given in Bergeron et al. (2005). As noted before, the presence of many hurdles indicates that the gene order was not generated purely by reversals.

Wang and Warnow pioneered the area of expected distances and their contributions are reviewed in (Wang and Warnow, 2005). Based on their results, Eriksen (2002) presented a close approximation of the method of moments reversal distance estimate from the breakpoint distance

\[
t(b) = \log \left(1 - \frac{b}{e^{\left(\frac{b}{n-b+2}\right)}}\right).
\]
where \( n \) is the number of genes and \( b \) the number of breakpoints in the genome \( \pi \). In addition, there is a (more complicated) formula for the expected reversal distance after \( t \) reversals, whose inverse (which has to be computed numerically) gives the expected number of reversals given the reversal distance (Eriksen and Hultman, 2004). These two methods give comparable results within the model, but may differ if the data does not adhere perfectly to the model.

### 4.2 The X-model: symmetrical reversals

In this model, we assume an axis of replication that divides the genome into two equally long halves. A reversal is symmetrical if exactly half of the genes are located on either side of the axis. This model cannot sort all permutations, but for those that can be sorted, the distance is exactly half the number of breakpoints.

In computing the expected number of breakpoints after \( t \) reversals, we start with the identity order \([1, 2, 3, \ldots, n]\). By circular symmetry, we need only keep track of gene 2 and see if it ends up next to gene 1. This is described by a Markov chain with \( 2n-2 \) states, one for each available position and orientation of gene 2. The symmetrical restriction results in only two active states, corresponding to genes 1 and 2 being on the same or opposite sides of the axis of replication. Let the total number of possible symmetrical reversals be \( m \), equaling \( (n-1)/2 \) for odd \( n \) and either \( n/2 \) or \( n/2-1 \) for even \( n \), depending on whether the axis of replication is through or between genes. Then, only one reversal can divide or unite a pair of originally neighboring genes, giving the transitions matrix

\[
M = \begin{pmatrix}
  m-1 & 1 & 0 \\
  1 & m-1 & 0 \\
  0 & 1 & m-1
\end{pmatrix}.
\]

What we need to compute is (see Eriksen, 2002)

\[
b(t) = n \left( 1 - \frac{\sum \lambda_i^2 j_i^2}{m^2} \right),
\]

where \( \lambda_i \) are the eigenvalues of \( M \), \( j_i \) the first entries of the corresponding normed eigenvectors and the \( m \) in the denominator is the common row sum in the matrix, that is the number of available operations. With eigenvalues \( m \) and \( m-2 \), we plug \( m=n/2 \) into our formula to get

\[
b(t) = \frac{n}{2} \left( 1 - \left( 1 - \frac{2}{n} \right)^t \right)
\]

which proves this theorem.

**Theorem 4.1.** The method of moments estimate of the reversal distance based on \( b \) breakpoints in a symmetrical reversals model is

\[
t(b) = \frac{\log \left( 1 - \frac{2}{n} \right)}{\log \left( 1 - \frac{n}{2} \right)},
\]

where \( n \) is the number of genes.

### 4.3 The spotted X: single gene reversals

The single gene reversals are those that only alter the sign of a gene. Most permutations cannot be sorted using this model, but for those that have a solution, the minimal number of single gene reversals is given by the number of negative elements.

In the method of moments estimate, restricting the model to single gene reversals results in four states, corresponding to all combinations of positive and negative orientations of genes 1 and 2. It is easy to verify that the transition matrix is

\[
M = \begin{pmatrix}
  n-2 & 1 & 1 & 0 \\
  1 & n-2 & 0 & 1 \\
  1 & 0 & n-2 & 1 \\
  0 & 1 & 1 & n-2
\end{pmatrix}.
\]

In this case, the eigenvalues are \( n, n-2 \) and \( n-4 \) and by computing the eigenvectors we get this theorem.

**Theorem 4.2.** The expected number of breakpoints in a genome with \( n \) genes after \( t \) single gene reversals taken from a uniform distribution is

\[
b(t) = \frac{n}{4} \left( 3 - 2 \left( 1 - \frac{2}{n} \right)^t \right) - \left( 1 - \frac{4}{n} \right)^t.
\]

We have not found an analytical inverse of this function. However, we can compute the method of moments estimate from the reversal edit distance.

**Theorem 4.3.** The method of moments estimate \( t(r) \) of the single gene reversal distance from the single gene reversal edit distance \( r \) is

\[
t(r) = \frac{\log \left( 1 - \frac{2}{n} \right)}{\log \left( 1 - \frac{1}{n} \right)},
\]

where \( n \) is the number of genes.

**Proof.** Let \( r(t) \) be the expected single gene reversal edit distance after \( t \) uniformly distributed random single gene reversals, and \( p(t, k) \) the probability that the single gene reversal edit distance after \( t \) steps is \( k \). Then,

\[
r(t) = \sum_k k p(t, k)
\]

\[
= \sum_k k \left[ \left( 1 - \frac{k-1}{n} \right) p(t-1, k-1) + \frac{k+1}{n} p(t-1, k+1) \right]
\]

\[
= \sum_k \left( k - 2 \left( \frac{k}{n} \right) + 1 \right) p(t-1, k)
\]

\[
= 1 + \left( 1 - \frac{2}{n} \right) r(t-1).
\]

Initial value \( r(0) = 0 \) now gives

\[
r(t) = \sum_{j=1}^t \left( 1 - \frac{2}{n} \right)^{j-1} = \frac{n}{2} \left( 1 - \left( 1 - \frac{2}{n} \right)^t \right),
\]

which gives the theorem.
Considering the clear division between the breakpoint growth for symmetrical and single gene reversals, respectively, we should consider what happens when we mix the two models. For instance, if we apply the proportion \( p \) of single gene reversals and the proportion \( 1-p \) of one axis (A) and two axes (B) symmetrical reversals, respectively. While increasing \( p \) move \( b(t) \) from one axis symmetrical reversals straight to single gene reversals, the combination of single gene reversals and two axes symmetrical reversals has more breakpoints than the respective pure models for most \( p \). One should note that the behavior of these functions does not depend on the number of genes—except for small numbers, the graphs are very similar if the abscissa is scaled proportionally. We chose 40 genes in order to make computations such as the expected transposition distance run quickly.

Fig. 3. Expected number of breakpoints after \( t \) reversals in models with symmetrical reversals (around one or two symmetry axes) and single gene reversals, using a genome of 40 genes. We have used the proportion \( p \) of single gene reversals and the proportion \( 1-p \) of one axis (A) and two axes (B) symmetrical reversals, respectively. While increasing \( p \) move \( b(t) \) from one axis symmetrical reversals straight to single gene reversals, the combination of single gene reversals and two axes symmetrical reversals has more breakpoints than the respective pure models for most \( p \). One should note that the behavior of these functions does not depend on the number of genes—except for small numbers, the graphs are very similar if the abscissa is scaled proportionally. We chose 40 genes in order to make computations such as the expected transposition distance run quickly.

Also, with three axes the number of breakpoints grows almost faster growth than either taken separately. In Figure 3, we have plotted the expected number of breakpoints of different models using a genome of 40 genes. These models include all reversals, single gene reversals, reversals symmetrical about one axis, two and three axes through adjacent genes and finally one axis between two genes plus one or two axes between one or both of these adjacent genes.

Interestingly, if we pose the similar question for single gene reversals and symmetrical reversals about two axes; the answer is ‘neither’. Combining these two models gives in general faster growth than either taken separately. In Figure 3, we have plotted the expected number of breakpoints of different values of \( p \).

4.4 The fat X and the zipper: reversal combinations

Assuming symmetrical reversals about one sharp axis is often too simple. In fact, when Ajana et al. (2002) investigated the asymmetric coefficient of applied reversals, they found that while for some genomes these reversals had a very low asymmetric coefficient (thus being very symmetrical), it was usually strictly greater than zero. A reversal with a positive asymmetric coefficient can be modeled using an axis that differs from the axis of replication. A situation with small asymmetric coefficients would thus be modeled by allowing several axes, which are close to each other.

The transition matrix of a model with several axes is obtained by adding the full \( 2n-2 \) state transition matrices for each axis. Examining the eigenvalues and eigenvectors of the resulting transition matrices, no clear patterns emerge to help us derive a closed formula, but we can still plot \( b(t) \). For comparison, we have done this for different sets of axes (Fig. 4). In addition, we have \( b(t) \) for the models presented above.

Fig. 4. Expected number of breakpoints after \( t \) reversals in different models, using a genome of 40 genes. These models include all reversals, single gene reversals, reversals symmetrical about one axis, two and three axes through adjacent genes and finally one axis between two genes plus one or two axes between one or both of these adjacent genes.

The effect on changing the eigenvalues is most important for the larger ones, so for \( p = 1/2 \), this is close to the formula for single gene reversals.
as fast as for uniform reversals. There seems to be nothing gained by introducing more axes—instead we can use the plain reversal model.

The same is true for the fat X-model induced by short transpositions and the zipper. In both these cases, the number of breakpoints increases almost as fast as in the uniform transposition and reversal models, respectively. Thus, these models are fully adequate for computing method of moments estimates. Since the allowed operations in the fat X-models are just subsets of the uniform models, but seemingly not simpler to handle, we can also use the uniform models for computing edit distances.

4.5 The cloud: single gene transpositions

Transpositions seem less tractable to edit distance computation than reversals. There is no polynomial time algorithm for computing the edit distance of transpositions.

**Single gene transpositions**, moving a single gene to any other position in the genome and possibly changing its orientation (that is including reversed transpositions on one gene), are sufficient to sort a genome. It is also quite easy to compute the distance to the identity, at least if we view single gene reversals as a special case of single gene transpositions.

**Definition 1.** An increasing subsequence in a permutation π is a sequence \( i_1 < i_2 < \cdots < i_k \) such that \( \pi_{i_1} < \pi_{i_2} < \cdots < \pi_{i_k} \).

**Theorem 4.4.** The number of single gene transpositions needed to transform a signed permutation \( \pi \) to the identity is the number of genes minus the length of a longest increasing subsequence of positive or negative genes in \( \pi \). For circular genomes, the starting point of the longest increasing subsequence is arbitrary.

**Proof.** It is easy to see that if we take an element that is not part of an increasing subsequence and insert it at an appropriate position, we can make this increasing subsequence one element longer. On the other hand, inserting this element somewhere else will not make the subsequence longer, and moving an element that belongs to the subsequence will only make it shorter. Thus, each single gene transposition can only prolong the subsequence by at most one.

Conversely, since each element that is not part of a longest increasing subsequence can be inserted at its proper position relative to the subsequence, we find that we do not need more transpositions than we have elements outside this subsequence. Also, for circular genomes, any increasing subsequence will do, regardless of its starting point.

The longest increasing subsequence can be computed in polynomial time, using for instance the Hunt–Szymanski algorithm (Hunt and Szymanski, 1977), which runs in \( O(n \log \log n) \). Having circular genomes, to find an optimal starting point it is enough to iteratively try the position of the minimal number not found in any longest increasing subsequence starting elsewhere. We consider only the positive genes, but must also repeat the process for the positive genes of the reversed genome. The time needed is thus not more than \( O(n^2 \log \log n) \).

![Fig. 5. Expected number of breakpoints after \( r \) reversals in the uniform reversal model, the single gene transposition model and the uniform transposition model (40 genes). We find that the single gene transposition model resembles the uniform transposition model much more than the uniform reversal model.](image)

For expected distances, it seems equally hard to compute the eigenvalues of the transition matrices of single gene transpositions as for transpositions and reversed transpositions. However, numerical computations of \( b(t) \) for single gene transpositions show that it is very close to the uniform transpositions case (Fig. 5). Thus, genomes scrambled with single gene transpositions have significantly more breakpoints than those scrambled with a comparably sized set of reversals, whether symmetrical, single gene or neither.

5 CONCLUSIONS AND DISCUSSION

Dotplots of gene orders between pairs of bacterial genomes show specific patterns that reflect which events have occurred since the species diverged from their last common ancestor. We show how these patterns likely were obtained by carefully weighing the contributions of reversals, transpositions, symmetrical reversals, single gene transpositions and single gene reversals. We recapitulate known results required to infer the minimal and expected evolutionary distances and derive new results for the undescribed cases.

The software GERMS (http://www.math.chalmers.se/~ner/germs.html) was developed to discriminate between the different gene order permutations and to propose a likely model under which it sorts the data (Andreen, M., manuscript in preparation). It uses a neural network trained on artificial data for selecting the models and outputs expected distances with greedily obtained sorting scenarios. Preliminary results on a subset of genomic data indicate that the most common models are the different X’s and to some extent the zipper (see the Supplementary Material). The cloud is also present at various degrees in almost all pairs of genomes.

Conflict of Interest: none declared.
REFERENCES


