ABSTRACT

**Motivation:** Segmental duplications ≥1 kb in length with ≥90% sequence identity between copies comprise nearly 5% of the human genome. They are frequently found in large, contiguous regions known as duplication blocks that can contain mosaic patterns of thousands of segmental duplications. Reconstructing the evolutionary history of these complex genomic regions is a non-trivial, but important task.

**Results:** We introduce parsimony and likelihood techniques to analyze the evolutionary relationships between duplication blocks. Both techniques rely on a generic model of duplication in which long, contiguous substrings are copied and reinserted over large physical distances, allowing for a duplication block to be constructed by aggregating substrings of other blocks. For the likelihood method, we give an efficient dynamic programming algorithm to compute the weighted ensemble of all duplication scenarios that account for the construction of a duplication block. Using this ensemble, we derive the probabilities of various duplication scenarios. We formalize the task of reconstructing the evolutionary history of segmental duplications as an optimization problem on the space of directed acyclic graphs. We use a simulated annealing heuristic to solve the problem for a set of segmental duplications in the human genome in both parsimony and likelihood settings.

**Availability:** Supplementary information is available at http://www.cs.brown.edu/people/braphael/supplements/.

**Contact:** clkahn@cs.brown.edu; braphael@cs.brown.edu.

1 INTRODUCTION

A striking feature of mammalian genomes is the prevalence of segmental duplications or low-copy repeats. Approximately 5% of the human genome consists of segmental duplications ≥1 kb in length with ≥90% sequence identity between copies (Bailey and Eichler, 2006). Segmental duplications account for a significant fraction of the differences between humans and other primate genomes, and are enriched for genes that are differentially expressed between the differences between humans and other primate genomes, and are enriched for genes that are differentially expressed between the species (Blekhman et al., 2009).

Segmental duplications remain an extreme challenge for evolutionary reconstruction, as they are the 'most structurally complex and dynamic regions of the human genome' (Alkan et al., 2009). Human segmental duplications are frequently found within complicated mosaics of duplicated fragments (Bailey and Eichler, 2006). Jiang et al. (2007) produced a comprehensive annotation of this mosaic organization; they derived an 'alphabet' of approximately 11 000 duplicated segments, or duplicons, and delimited 437 duplication blocks or 'strings' of at least 10 (and typically dozens) different duplicons found contiguously on a chromosome. However, the relationships between these annotated duplication blocks are complex and straightforward analysis does not immediately reveal the evolutionary relationships between blocks.

Numerous authors have considered the problem of analyzing relationships between genome sequences that contain duplicated segments. This work falls into roughly two categories. The first focus is the problem of computing genome rearrangement distances, like reversal distance, in the presence of duplicated genes or synteny blocks (El-Mabrouk, 2002; Marron et al., 2004; Sankoff, 1999, for example). However, such rearrangement distances do not model the creation of new duplicates and thus are not well-suited to describe the evolutionary history of segmental duplications in the genome. The second focus is to analyze regions with duplications under 'local' operations like tandem duplications (Chaudhuri et al., 2006; Lajoie et al., 2007, for example). While tandem duplication is undoubtedly important in the generation of duplication blocks, there is strong evidence that an important characteristic of the history of segmental duplications is the frequent duplication and transposition of long segments over large physical distances; as many as 50–60% of segmental duplications were transposed interchromosomally (Bailey and Eichler, 2006). Several general models of rearrangement that allowed for both local operations and duplication–transposition–like operations between different strings were studied by Ergun et al. (2003), but the generality of those models meant that the distances were NP-hard to compute and only approximation algorithms were given.

Here, we present a novel formulation of the problem of computing an evolutionary history for a set of segmental duplications that are organized in duplication blocks. We represent evolutionary relationships between a set of duplication blocks as a directed acyclic graph (DAG), and we formalize the evolutionary reconstruction problem as an optimization over the space of DAGs. We present two different methods for scoring a DAG: one based on parsimony and one based on likelihood. The parsimony score for a DAG is a straightforward extension of 'duplication distance', a measure introduced by some of us (Kahn and Raphael, 2008, 2009) that describes the most parsimonious sequence of duplicate operations needed to construct a given target string. The likelihood score for a DAG is the product of the likelihood scores for each of the duplication blocks, where a duplication block's likelihood is derived by computing the weighted ensemble of all possible duplication scenarios that could have generated it. We describe how to compute the partition function of the ensemble efficiently using a dynamic program that generalizes the duplication distance (i.e. parsimony score) recurrence. Deriving a probabilistic model from a dynamic program this way is analogous to the
As noted above, an important characteristic of segmental duplications that these reconstructions to the analysis of Jiang relationships between segmental duplications. of earlier analysis but also reveals additional and more subtle duplication that we use to compute the likelihood score for an evolutionary Definition for X target string. to construct Y by copying and pasting substrings of X into an initially empty to a target string Y is the minimum number of duplicate operations needed the characters of Y is a sequence of substrings of X. Definition j 2We note that the duplication distance between a pair of strings is not formally 2 methods Here, we present two methods for determining the optimality of an evolutionary relationship between a pair of duplication blocks—one based on a parsimony criterion and one based on a likelihood criterion. In Sections 2.1 and 2.2, we describe the parsimony-based model of segmental duplication that is based on duplication distance, introduced in Kahn and Raphael (2008, 2009). Next, we present a novel probabilistic model of segmental duplication that we use to compute the likelihood score for an evolutionary relationship between a pair of duplication blocks.

2.1 A model of segmental duplication

As noted above, an important characteristic of segmental duplications that distinguishes them from other types of repeats is that they are frequently transposed across large genomic distances from their respective ancestral loci. Kahn and Raphael (2008, 2009), we modeled the process in which a duplication block, a composite of many duplicons, is built by copying strings of duplicons from other duplication blocks. In particular, we define the basic ‘copy-paste’ operation as follows.

**Definition 2.1** A duplication operation, \( b_{j,p}(X) \), copies a substring \( X Sl \) of a source string \( X \) and pastes it into a target string at position \( p \). Specifically, if \( X=x_1...x_m \) and \( Y=y_1...y_n \), then \( Z=b_{j,p}(X)(x_1...x_j...x_{j+p-1})y_{p+1}...y_n \).

**Definition 2.2** The duplication distance \( d(X,Y) \) from a source string \( X \) to a target string \( Y \) is the minimum number of duplication operations needed to construct \( Y \) by copying and pasting substrings of \( X \) into an initially empty target string.

A subsequence is distinguished from a substring because the characters of a subsequence need not be contiguous. Given a string \( X \), a subsequence \( S \) of \( X \) can be expressed as an increasing list of indices of \( X \). For example, for \( X=\text{abcde} \), the subsequence \( S=(1,3,5) \) is the string ace.

**Definition 2.3** Two subsequences \( S=(i_1,...,i_k) \) and \( T=(j_1,...,j_l) \) of a string \( X \) overlap if either (i) there exist indices \( i,j \), with \( 1 \leq i < j \) such that \( i \leq j \), or (ii) there exist indices \( i,j \), with \( 1 \leq i < j \), and a \( f \), with \( 1 \leq f < j \) such that either \( i < j < f \) or \( e < f < j \).

Given a source/target pair \( X,Y \), any sequence of duplicate operations of the form \( b_{i_1,p_1}(X)...b_{i_k,p_k}(X) \) that generates \( Y \) from \( X \) uniquely partitions the characters of \( Y \) into non-overlapping subsequences corresponding to characters that were copied conjointly from \( X \).

**Definition 2.4** A source string \( X \), a generator \( \Psi_X=(X_{k,1},...,X_{k,k}) \) is a sequence of substrings of \( X \). 3In Kahn and Raphael (2008, 2009), we also considered reverse duplications in which the copied substring is inverted before being inserted into the target. We note that all of our definitions and algorithms presented here can be similarly augmented but we omit the details.

2.2 Parsimony

In Kahn et al. (2010), we describe a polynomial-time algorithm to compute the duplication distance from \( X \) to \( Y \). We use duplication distance to measure the similarity between a pair of duplication blocks by counting the number of operations needed to generate \( Y \) from \( X \) in a simplest or most-parsimonious scenario.

While the parsimony assumption is attractive from a theoretical perspective and can produce useful biological insight, it might be overly restrictive, particularly when there are many different optimal or nearly optimal solutions. Consider, for example, the strings \( X=\text{abcde} \), \( Y=\text{deabc} \), \( T=\text{deabcd} \), and \( g=\text{hijkl} \). The duplication distance, \( d(X,Y) \), is 13 and there is a single feasible generator with this optimum length. However, there are 989 possible feasible generators for \( Y \), 119 of which have length 14, just slightly suboptimal.

Because the space of all possible feasible generators is very large, a probabilistic model might give very low probability to an optimal parsimony solution. Thus, in the next section, we present a probabilistic model of segmental duplication that considers the weighted ensemble of all feasible generators for a source/target string pair.

2.3 The partition function

For a given source string \( X \) and positive integer \( k \), we consider the space of all length-ke generators \( \Psi_X \). We define a probability distribution on the collection of generators by defining \( Pr(\Psi_X) \propto \omega(\Psi_X) \) where \( \omega(\Psi_X) \) is the ‘score’, or weight, assigned to a generator, and we compute the partition function \( Z(\Psi_X) \) of the weighted ensemble of all possible length-ke generators \( \Psi_X \). Given a source string \( X \) and a target string \( Y \), we define the event \( F \) to be the event of choosing a length-ke generator that is feasible for \( Y \) from the space of length-ke generators. We define a probabilistic model for segmental duplications that, given a target string \( Y \), assigns a probability to \( F \). Pr\[F(\Psi_X) \mid Y \] is the weighted ensemble of all possible length-ke generators that are feasible for \( Y \), normalized by the
2.7 Therefore, in order to compute the partition function \( Z^\omega \) and corresponding Young tableaux. If \( \omega \) define as Kahn et al. define as proof of correctness due to space considerations. In this section, we present the final ingredient necessary to compute the partition function \( Z^\omega \) of feasible generators \( \Psi^\omega \) that are for feasible \( Y \) satisfies the following recurrence.

\[
N^\omega_i(Y) = \sum_{j=0}^{\omega} N^\omega_j(Y, i) \quad \text{satisfies the following recurrence.}
\]

\[
N^\omega_0(Y, i) = 1 \quad \text{if } Y = Y_i^1 \quad \text{otherwise,}
\]

\[
N^\omega_1(Y, i) = N^\omega_1(Y_1, i) + \sum_{j=1}^{\omega} \sum_{l=0}^{\omega} N^\omega_l(Y_1^j, i).
\]

Here, the term \( N^\omega_i(Y, j) \) represents the number of feasible generators \( \Psi^\omega \) with length \( k \) given that the character \( y_j \) is generated by a substring of \( X \) starting at \( x_i \).

We compute the restricted partition function \( Q^\omega_i(Y) \) efficiently by first counting the number of feasible generators, namely \( N^\omega_i(Y, j) \) and scoring each generator appropriately by \( \sigma(k, |F|) \).

**Theorem 2.9.** Let \( X = x_1 \ldots x_n \) and \( Y = y_1 \ldots y_n \) be source target string pair and let \( k \) be a positive integer. The restricted partition function \( Q^\omega_i(Y) \) satisfies the following.

\[
Q^\omega_i(Y) = N^\omega_i(Y, j) \cdot \sigma(k, |F|).
\]

The recurrence given in Lemma 2.8 can be computed in time \( O(|Y|^2 |\mu(Y)| |\mu(X)|) \) where \( \mu(Y) \) (resp. \( \mu(X) \)) is the maximum multiplicity of any character that appears in \( Y \) (resp. \( X \)), so computing \( Q^\omega_i(Y) \) takes the same time. We include a proof of correctness in Supplementary Section 1.1.

## 3 ALGORITHM

Here, we formalize the problem of computing a segmental duplication evolutionary history for a set of duplication blocks in the human genome with respect to either a parsimony or likelihood criterion.

### 3.1 Maximum parsimony and maximum likelihood evolutionary histories

The input to our problem is the set of duplication blocks found in the human genome, each represented as a signed string on the alphabet of duplicons. Our goal is to compute a putative duplication history that accounts for the construction of all of the duplication blocks. We assume that the ancestral genome is devoid of segmental duplications. A duplication history is a sequence of duplicate events that first builds up a set of seed duplication blocks by duplicating and aggregating duplicons from their ancestral loci and then successively constructs the remaining duplication blocks by duplicating substrings of previously constructed blocks.

We observed in Kahn and Raphael (2008) strong evidence that many of the duplication blocks identified by Jiang et al. (2007) had been constructed through the duplication and aggregation of substrings of duplicons from several other blocks. Therefore, a tree cannot aptly represent an evolutionary history; a more appropriate representation of the evolutionary relationships between duplication blocks is a DAG in which the vertices represent duplication blocks and an edge directed from a vertex \( X \) to a vertex \( Y \) indicates that

---

C.L.Kahn et al.

2.8 Given a source string \( X = x_1 \ldots x_n \) and a target string \( Y = y_1 \ldots y_n \) the number \( N^\omega_i(Y) \) of distinct length-\( k \) generators \( \Psi^\omega \) that are for feasible \( Y \) satisfies the following recurrence.

\[
N^\omega_i(Y) = \sum_{j=0}^{\omega} N^\omega_j(Y, i) \quad \text{satisfies the following recurrence.}
\]

\[
N^\omega_0(Y, i) = 1 \quad \text{if } Y = Y_i^1 \quad \text{otherwise,}
\]

\[
N^\omega_1(Y, i) = N^\omega_1(Y_1, i) + \sum_{j=1}^{\omega} \sum_{l=0}^{\omega} N^\omega_l(Y_1^j, i).
\]

Here, the term \( N^\omega_i(Y, j) \) represents the number of feasible generators \( \Psi^\omega \) with length \( k \) given that the character \( y_j \) is generated by a substring of \( X \) starting at \( x_i \).

We compute the restricted partition function \( Q^\omega_i(Y) \) efficiently by first counting the number of feasible generators, namely \( N^\omega_i(Y, j) \) and scoring each generator appropriately by \( \sigma(k, |F|) \).

**Theorem 2.9.** Let \( X = x_1 \ldots x_n \) and \( Y = y_1 \ldots y_n \) be source target string pair and let \( k \) be a positive integer. The restricted partition function \( Q^\omega_i(Y) \) satisfies the following.

\[
Q^\omega_i(Y) = N^\omega_i(Y, j) \cdot \sigma(k, |F|).
\]

The recurrence given in Lemma 2.8 can be computed in time \( O(|Y|^2 |\mu(Y)| |\mu(X)|) \) where \( \mu(Y) \) (resp. \( \mu(X) \)) is the maximum multiplicity of any character that appears in \( Y \) (resp. \( X \)), so computing \( Q^\omega_i(Y) \) takes the same time. We include a proof of correctness in Supplementary Section 1.1.

## 3 ALGORITHM

Here, we formalize the problem of computing a segmental duplication evolutionary history for a set of duplication blocks in the human genome with respect to either a parsimony or likelihood criterion.

### 3.1 Maximum parsimony and maximum likelihood evolutionary histories

The input to our problem is the set of duplication blocks found in the human genome, each represented as a signed string on the alphabet of duplicons. Our goal is to compute a putative duplication history that accounts for the construction of all of the duplication blocks. We assume that the ancestral genome is devoid of segmental duplications. A duplication history is a sequence of duplicate events that first builds up a set of seed duplication blocks by duplicating and aggregating duplicons from their ancestral loci and then successively constructs the remaining duplication blocks by duplicating substrings of previously constructed blocks.

We observed in Kahn and Raphael (2008) strong evidence that many of the duplication blocks identified by Jiang et al. (2007) had been constructed through the duplication and aggregation of substrings of duplicons from several other blocks. Therefore, a tree cannot aptly represent an evolutionary history; a more appropriate representation of the evolutionary relationships between duplication blocks is a DAG in which the vertices represent duplication blocks and an edge directed from a vertex \( X \) to a vertex \( Y \) indicates that...
We analyzed a set of 391 duplication blocks identified by Jiang that maximizes the likelihood:

\[ P(Y) = \sum_{D \in \mathcal{D}} P(D)P(Y,D) \]

where \( P(Y) \) is the probability of observing the data \( Y \), \( P(D) \) is the prior probability of the tree \( D \), and \( P(Y,D) \) is the likelihood of the data under the tree \( D \).

We make two simplifying assumptions. First, we assume that only duplicate events occur and that there are no deletions, inversions, or other types of rearrangements within a duplication block. Second, we assume that a duplication block is not copied and used to make duplicate events occur and that there are no deletions, inversions, or other types of rearrangements.

3.1 Maximum parsimony reconstruction

The maximum parsimony DAG contains 391 nodes and 479 directed edges. For comparison, we then computed the maximum parsimony score of 28 852 and the maximum likelihood score of 390 edges.

4 IMPLEMENTATION

We analyzed a set of 391 duplication blocks identified by Jiang et al. (2007) that were represented as signed strings on an alphabet of \( \mathbb{Z} \). We computed the maximum likelihood evolutionary history (Definition 3.1) for the entire set of blocks (Fig. 2). The DAG exhibited multiple connected components. For comparison, we then computed the maximum likelihood evolutionary histories (Definition 3.2) for several of the subgraphs induced by connected components of the parsimony solution.
Fig. 2. The maximum parsimony DAG for a set of 391 duplication blocks in the human genome. The nodes represent duplication blocks. Edges indicate evolutionary relations; an edge is directed from a node $u$ to a node $v$ if the most-parsimonious duplication scenario includes duplication events that copy substrings of $u$ in the construction of $v$. Jiang et al. (2007) partitioned the duplication blocks into a set of 24 clades (plus one ‘s’ group of duplication blocks found in subtelomeric regions) that we indicate here with 25 colors on nodes. The 3 sets of colored edges represent inheritance networks for 3 conserved subsequences of duplicons. These inheritance networks are almost entirely confined to a single clade each. The green edges represent the inheritance of the duplicon sequence [6968, 6967, 6965, 6963, 6962, 6960] in clade ‘M1’, the red edges represent the inheritance of [7039, 7036, 7037] in clade ‘M2’, and the blue edges represent the inheritance of [9448, 9449] in clade ‘chr16.’
Reconstruction of segmental duplications

Fig. 3. A connected component of the maximum parsimony DAG. Nodes from clade ‘M1’ are red and nodes from clade ‘chr7_2’ are green. Node labels correspond to duplication block IDs. The blue edges represent the inheritance network for non-core duplicon 6970.

Fig. 4. (a) Component comprised entirely of duplication blocks from clade ‘chr16’ in the maximum parsimony DAG. (b) Maximum likelihood DAG for subgraph induced on nodes in (a).

4.3 Maximum likelihood reconstruction

We computed the maximum likelihood DAGs (Definition 3.2) for the sets of duplication blocks appearing within moderately sized connected components of the maximum parsimony DAG in order to compare the two methods. We chose the components comprised of blocks from clades ‘chr16’ and ‘chr10’, respectively (Fig. 2). The maximum likelihood subgraphs for these subproblems are shown in Figures 4b and 5b.

The two DAGs for the ‘chr16’ component in Figure 4 share some characteristics. For example, node 121 is a common ancestor of every other block and block 276 exhibits high out-degree in both solutions. Both solutions are similarly ‘good’ with respect to the parsimony objective: the solution in (a) exhibits an optimal parsimony score of 397, and the one in (b) exhibits a score of 401.
However, the likelihood score for the parsimony solution (a) was nearly zero. One difference that accounts for this discrepancy is the higher average in-degree for blocks in the parsimony solution (2.2) as compared to the likelihood solution (1.3). Also, the parsimony solution exhibits a path with ten edges, whereas the longest path in the likelihood solution has six.

Some of these differences are due to the fact that the parsimony criterion does not penalize edges that do not directly improve the score. For example, block 291 has two parents (276 and 25 ) in the parsimony DAG but only one parent (276) in the likelihood DAG. However, the duplication distance with source 276 ⊙ 25 and target 291 is the same as the duplication distance with source 276 and target 291. Therefore, the edge from 25 to 291 does not improve the parsimony score, underscoring that there are multiple optimal parsimony solutions. In contrast, the likelihood of a target block generally increases as the sum of the lengths of its parent blocks decreases, so the max likelihood DAG will not include edges that do not directly improve the score.

5 DISCUSSION

Our maximum parsimony and maximum likelihood reconstructions show some differences, both from each other and from the analysis of Jiang et al. (2007). In particular, we identify non-core duplicons and subsequences that are arguably as promiscuous within a clade as core duplicons.

There are several directions for future work. From a theoretical perspective, one can incorporate other types of operations into the probabilistic model, such as deletions and inversions which we have described in the parsimony setting (Kahn et al., 2010), as well as single nucleotide mutations. Also, our method could be used to sample over the space of DAGs using a Markov Chain Monte Carlo strategy. From the perspective of applications, a more comprehensive analysis of genes or other elements in the newly identified core duplicons and core subsequences from our reconstruction is warranted, as is a further refinement of the clade annotation by analyzing the clade-induced subgraphs of the DAGs.

Conflict of Interest: none declared.

REFERENCES


