BIOINFORMATICS

Vol. 19 no. 12 2003, pages 1497–1504
DOI: 10.1093/bioinformatics/btg191

Greedy method for inferring tandem duplication history

Louxin Zhang[1,*], Bin Ma[2], Lusheng Wang[3] and Ying Xu[4]

[1]Department of Mathematics, National University of Singapore, Singapore,
[2]Department of Computer Science, University of Western Ontario, Canada N6A 5B8,
[3]Department of Computer Science, City University of Hong Kong, Hong Kong and
[4]Department of Computer Science, Peking University, People's Republic of China

Received on November 4, 2002; revised on November 18, 2002; accepted on February 18, 2003

ABSTRACT

Motivation: Genome analysis suggests that tandem duplication is an important mode of evolutionary novelty by permitting one copy of each gene to drift and potentially to acquire a new function. With more and more genomic sequences available, reconstructing duplication history has received extensive attention recently.

Results: An efficient method is presented for inferring the duplication history of tandemly repeated sequences based on the model proposed by Fitch (1977). We validate the method by using simulation results and real data sets of mucin genes, ZNF genes, and olfactory receptors genes. The agreement with conclusions drawn by other biological researchers strongly indicates that our method is efficient and robust.

Availability: The program is available by request.

Contact: matzlx@nus.edu.sg

1 INTRODUCTION

Tandem duplication is a mutational process for DNA molecules in which a short stretch of DNA is transformed into several adjacent copies. Since individual repeat may undergo additional mutations later on, approximate repeats are usually present in a tandem array. There are three main types of tandemly repeated sequences in a genome: (i) large tandem repeats, (ii) minisatellites—repeats of a sequence with 10 to 200 bps, in which each repeat may vary slightly, and (iii) microsatellites—identical repeats of a short sequence. Although tandem duplication is less well understood, unequal recombination is widely thought as the key biological mechanism responsible for it (Ohno, 1970; Fitch, 1977).

Tandem duplication process is a primary mechanism for generating gene clusters on chromosomes. 32% of putative genes on human chromosome 19 (HSA19) are arranged in tandem arrays (Kim *et al.*, 2001). Gene families organized in tandem arrays were also reported in other genomes such as *C. elegans*, *Drosophila*, and *Arabidopsis*. Genome analysis suggests that one copy of a duplicated gene could drift and potentially acquire a new function. Hence, the evolutionary study of a tandem gene family may yield valuable insights into predicting their functions and solving important species-specific questions.

The study of tandemly repeated sequences goes back to more than two decades ago. The duplication model of tandemly repeated sequences was first formulated by Fitch (1977) as a phylogeny constrained by the crossover process. In the same paper, he studied the duplication history of five human hemoglobins and seven tandemly repeats of human apolipoprotein A-I. The duplication history for the human fatal globin was studied by Shen *et al.* (1981).

With more and more genomic sequences available recently, reconstructing the duplication history of the tandemly repeated sequences has received attention again. The ordered nature of tandemly repeated sequences was first considered by Benson and Dong (1999). They presented heuristic reconstruction algorithms for a special case in which the duplicated stretch contains only one basic copy unit. Later, a better algorithm for this special case was given by Jaitly *et al.* (2001). Independently, Tang *et al.* (2001) reproposed the same duplication model as Fitch for studying tandemly repeated sequences. Based on the duplication model, they gave a distance-based method, called the WINDOW method, for inferring the tandem duplication history. Such a method is simple, efficient. But it works under the assumption that sequences evolved in a molecular clock mode. Elemento *et al.* (2002) presented another method that performs an exhaustive search in the space of duplication trees according to the parsimony criterion.

In this paper, we study the problem of systematically inferring the duplication history of tandemly repeated sequences in a different approach. Here, we use the terminology and notation from Tang *et al.* (2001). A

*To whom correspondence should be addressed.

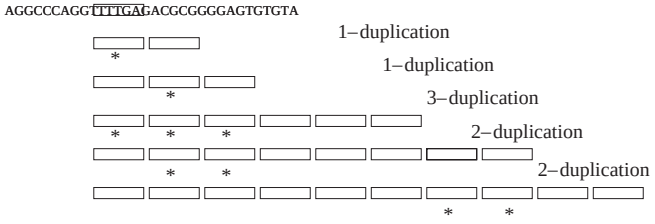

Fig. 1. A duplication process with five tandem duplications. The original repeats are marked with '*'.

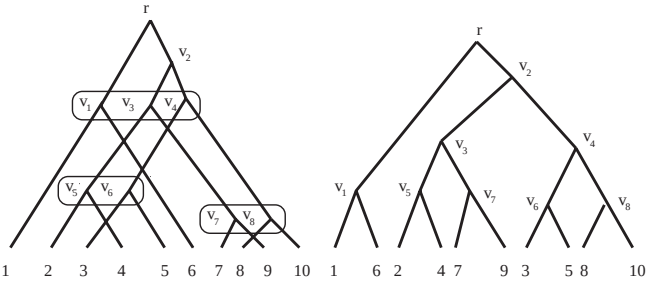

Fig. 2. A duplication model on a sequence family $\{1, 2, \ldots, n\}$ and its associated phylogeny. There are five blocks $[r]$, $[v_2]$, $[v_1, v_3, v_4]$, $[v_5, v_6]$ and $[v_7, v_8]$; blocks with at least two nodes are represented by boxes.

duplication model $\mathcal{M}$ for tandemly repeated sequences is a directed graph that contains *nodes*, *edges* and *blocks* as shown in Figure 2. This model assumes that tandem duplication is the main biological process responsible for generation of the sequences and there are no gene deletions. If only the parent–child relations are considered in a duplication model $\mathcal{M}$, then, the resulting structure is just a rooted binary tree $T_{\mathcal{M}}$ which is unique and called the *associated phylogeny* for $\mathcal{M}$. We first give a linear-time algorithm which, given a phylogeny for a set of sequences, outputs the unique duplication model $\mathcal{M}$ such that $T = T_{\mathcal{M}}$ if it exists. This algorithm provides a basis for developing our algorithm for inferring the duplication model of tandemly repeated sequences in Section 2.3 Finally, we analyze three data sets of tandemly repeated sequences to validate our method.

2 MODEL AND ALGORITHMS

2.1 Duplication Model

In this paper, we study the duplication history of tandemly repeated sequences using the model proposed by Fitch (1977) (and reproposed by Tang *et al.* (2001)). The model captures both evolutionary history and the observed order of sequences on a chromosome. Assume n sequences $\{1, 2, \cdots, n\}$ were formed from a locus only by tandem duplication. Then, the locus had grown from a single copy through a series of tandem duplications as shown in Figure 1. A duplication replaces a stretch of DNA containing several repeats with two identical and adjacent copies of itself. If the stretch contains k repeats, the duplication is called a *k-duplication*.

A *duplication model* $\mathcal{M}$ for tandemly repeated sequences is a directed graph that contains *nodes*, *edges* and *blocks*. For example, the model in Figure 2 describes the duplication history given in Figure 1. A node in $\mathcal{M}$ represents a repeat. A directed edge (u, v) from u to v indicates that v is a child of u. A node s is an *ancestor* of a node t if there is a directed path from s to t. A node that has no outgoing edges is called a *leaf*; it is labeled with a given sequence. A non-leaf node is called an *internal* node; it has a left and right child. The *root*, which has only outgoing edges, represents the original copy at the locus.

A *block* in $\mathcal{M}$ represents a duplication. Each internal node appears in a unique block; no node is an ancestor of another in a block. If the block corresponds to a k-duplication, then it contains k nodes $v_1, v_2, \cdots, v_k$ from left to right. Let $lc(v_i)$ and $rc(v_i)$ be the left and right child of v_i $1 \leqslant i \leqslant k$. Then,

$$lc(v_1), lc(v_2), \cdots, lc(v_k), rc(v_1), rc(v_2), \cdots, rc(v_k)$$

are placed from left to right in the model. Hence, for any i and j, $1 \leqslant i < j \leqslant k$, the edges $(v_i, rc(v_i))$ and $(v_j, lc(v_j))$ cross each other. However, no other edges cross in the model. For simplicity, we will only draw a box for a block representing a k-duplication, $k \geqslant 2$.

Each leaf is labeled with a given sequence. The left-to-right order of leaves in the model is identical to the order of the sequences on a chromosome.

An *ordered phylogenetic tree* for sequences $\{1, 2, \cdots n\}$ is a rooted phylogeny in which its leaves are listed from left to right in the increasing order. Ordered phylogenies form a proper subclass of phylogenies; and each of them is a duplication model in which there are only 1-duplications. Since an ordered phylogeny with n leaves corresponds uniquely to a triangulation of a regular $(n+1)$-polygon, the number of ordered phylogenies with n leaves is just $\binom{2(n-1)}{n-1}/n$, the nth Catalan number. This answers a problem raised in Fitch (1977).

If we consider only the evolutionary relationship defined in a duplication model $\mathcal{M}$, the underlying structure $T_{\mathcal{M}}$ is just a rooted binary tree, called the *associated phylogeny* of $\mathcal{M}$. Clearly, $T_{\mathcal{M}}$ is unique. However, not all phylogenies are associated with duplication models. In fact, Fitch (1977) first observed that only 92 of 945 possible rooted phylogenies with six leaves are associated with duplication models. Counting associated phylogenies is tricky in general.

Now, we give two basic properties of associated phylogenies. These properties will be used to reconstruct a duplication model from a phylogeny in the next section. Let $\mathcal{M}$ be a duplication model of tandemly repeated sequences $F = \{1, 2, \cdots, n\}$ and $T_{\mathcal{M}}$ its associated phylogeny. Then, for any internal node u in $T_{\mathcal{M}}$, the subtree $T_{\mathcal{M}}(u)$ rooted at u is also associated with a duplication model of the subset of sequences appearing in $T_{\mathcal{M}}(u)$.

Recall that, for an internal node u in the model $\mathcal{M}$, we use $lc(u)$ and $rc(u)$ to denote its left and right child in $\mathcal{M}$ or $T_{\mathcal{M}}$ respectively. Similarly, we use $l^*c(u)$ and $r^*c(u)$ to denote the leftmost and rightmost leaf in the subtree $T_{\mathcal{M}}(u)$ rooted at u respectively. For example, in the model given in Figure 2, $l^*c(v_2) = 2$, and $r^*c(v_2) = 10$. Then we have the following lemma.

LEMMA 1. *For each internal node u in $T_{\mathcal{M}}$, $r^*c(u) > r^*c(lc(u))$ and $l^*c(u) < l^*c(rc(u))$. Equivalently, for an $u \in T_{\mathcal{M}}$, $l^*c(u)$ and $r^*c(u)$ are the smallest and largest labels in the subtree $T_{\mathcal{M}}(u)$.*

2.2 Constructing a duplication model from a phylogeny

A duplication model $\mathcal{M}$ has a unique associated phylogeny $T_{\mathcal{M}}$. However, a phylogeny is not necessarily associated with a model. In this section, we present a linear time algorithm for the following problem: Given a phylogeny T, reconstruct the duplication model $\mathcal{M}$ such that $T = T_{\mathcal{M}}$ if it exists.

A duplication model is said to be *double* if all duplications in it are 1(or 2)-duplication. Here, we first present an algorithm for reconstructing a double duplication model from a phylogeny. Then, we generalize it to arbitrary duplication models. The best known algorithm for this problem takes quadratic time (Tang *et al.*, 2001; Elemento *et al.*, 2002). To represent a duplication model, we only need to list all non-single duplication blocks on the associated phylogeny. Hence, our algorithm outputs a list of non-single duplication blocks if the model exists.

2.2.1 Double duplication models Let T be a phylogeny on sequence family $F = \{1, 2, \cdots, n\}$. We associate a pair (L_v, R_v) of leaf indices with each node v in T as follows. The pair of a leaf labeled with i is (i, i); $(L_v, R_v) = (l^*c(v), r^*c(v))$ for an internal node v. We set $P(T) = \{(L_v, R_v) \mid v \in T\}$. The set $P(T)$ can be computed using dynamic programming in a bottom up fashion: $(L_v, R_v) = (i, i)$ for a leaf v labeled with i; for an internal node v, we compute (L_v, R_v) as $L_v = \min\{L_{lc(v)}, L_{rc(v)}\}$ and $R_v = \min\{R_{lc(v)}, R_{rc(v)}\}$ once $(L_{lc(v)}, R_{lc(v)})$ and $(L_{rc(v)}, R_{rc(v)})$ are known. Since T contains $2n-1$ nodes, the above procedure takes linear time. The set $P(T)$ will play a critical role in reconstructing a duplication model from a phylogeny. (To use $P(T)$ efficiently, we attach each pair (L_v, R_v) at the node v by introducing an extra pointer.)

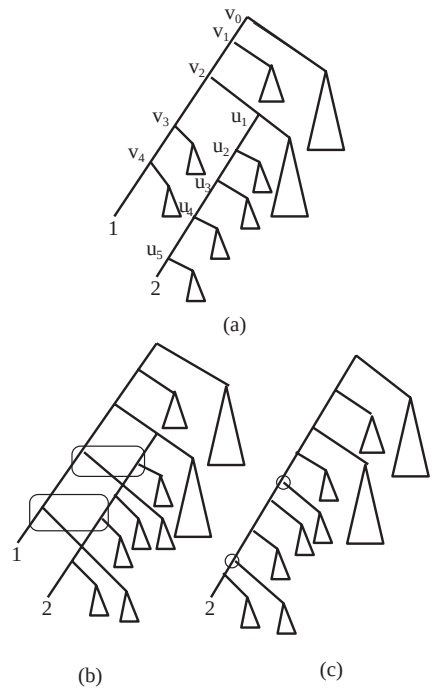

Fig. 3. (a) An example illustrating the position of sequence 2, where $p = 5$, $i = 2$ and $q = 6$. (b) The configuration after duplication blocks $[v_3, u_2]$ and $[v_4, u_4]$ are placed. (c) The tree T' derived after the duplication blocks are eliminated. The 'bad' nodes are marked with circles.

Since the property in Lemma 1 can easily be verified using $P(T)$ given T, we may assume that T satisfies that property in the rest of this section. (Otherwise, T cannot be associated with a duplication model.) Hence, the leftmost and rightmost leaves in T are 1 and n respectively. Where does 2 locate? Let

$$v_0 = r, v_1, v_2, \cdots, v_{p-1}, v_p = 1 \tag{1}$$

be the path from the root r to leaf 1, the *leftmost path* in T. Since $l^*c(v) < l^*c(rc(v))$ for any $v \in T$ (Lemma 1), 2 can only appear at the leftmost leaf of a subtree rooted at some node $rc(v_i)$, $0 \leqslant i \leqslant p-1$ (see Fig. 3a). Let

$$u_1 = rc(v_i), u_2, \cdots, u_{q-1}, u_q = 2 \tag{2}$$

be the path from $rc(v_i)$ to leaf 2, where $q \geqslant p - i$ as we will see later.

Suppose that T is associated with a double duplication model $\mathcal{M}$. Then, we have the following lemma.

LEMMA 2. *$\mathcal{M}$ must contain $p - i - 1$ double duplications*

$$[v_{i+1}, u_{j_1}], [v_{i+2}, u_{j_2}], \cdots, [v_{p-1}, u_{j_{p-i-1}}],$$

where $1 \leqslant j_1 < j_2 < \cdots < j_{p-i-1} \leqslant q-1$. *Hence,* $q \geqslant p-i$.

PROOF. If v_{i+k} does not belong to a double duplication block in $\mathcal{M}$, the leaf labeled with 2 cannot be placed before the leftmost leaf in the subtree rooted at $rc(v_{i+k})$, contradicting the fact that 2 is right next to 1 in $\mathcal{M}$. Hence, v_{i+k} must appear in a double duplication block for each k, $1 \leqslant k \leqslant p-i-1$. This finishes the proof. □

Notice that $R_{u_1} > R_{u_2} > \cdots > R_{u_{q-1}} > R_{u_q}$ and $R_{v_{i+1}} > R_{v_{i+2}} > \cdots > R_{v_{p-1}}$. By merging these two sequences in $p-i+q \leqslant 2q$ comparisons, we are able to determine all the u_{j_k}'s, $1 \leqslant k \leqslant p-i-1$, by just scanning the resulting sequences. This is because $R_{v_{i+k}}$ appears between $R_{u_{j_k}}$ and $R_{u_{j_k+1}}$ in the resulting sequences. (To locate u_{j_k}'s efficiently, we need not only a pointer from a parent node to its child, but also a pointer from a child to its parent.) Hence, the $p-i-1$ double duplication blocks can be determined by using at most $2q$ comparisons, where q is the length of the path given in (2).

Assume that all u_{j_k}'s have been determined for all $k = 1, 2, \cdots, p-i-1$. After all the duplication blocks $[v_{i+k}, u_{j_k}]$ are placed on T, the leaf 2 should be right next to the leaf 1 (see Fig. 3b) if T is associated with a model. We derive a rooted binary tree T'' from the subtree $T(u_1)$ by inserting a new node v'_k in the edge (u_{j_k}, u_{j_k+1}) for each $1 \leqslant k \leqslant p-i-1$, and assigning the subtree $T(rc(v_{i+k}))$ rooted at $rc(v_{i+k})$ as the right subtree of v'_k. Notice that the left child of v'_k is u_{j_k+1} in T'' now. Then, we form a new phylogeny T' from T by replacing subtree $T(v_i)$ with T'' as illustrated in Figure 3c. To distinguish the inserted nodes v'_k, $1 \leqslant k \leqslant p-i-1$, from the original ones in T, we mark these inserted ones as 'bad' nodes. It is not difficult to see that the leftmost leaf is 2 in T' and the number of nodes in T' is 1 less than those of T. Most importantly, we have the following straightforward fact: T is associated with the double duplication model $\mathcal{M}$ only if T' is associated with the model $\mathcal{M}'$ that is the restriction of $\mathcal{M}$ on the nodes of T', in which the 'bad' nodes will not be in any k-duplication blocks, $k \geqslant 2$.

ALGORITHM FOR RECONSTRUCTING A DDM

Input: A phylogeny T with root r on a sequence family $\{1, 2, \cdots, n\}$.
Output: The DDM derived from T if it exists.

$DS = \phi$; / DS keeps double duplication blocks */*
Determine duplication blocks described in Lemma 2 that place the leaf 2 right next to the leaf 1;
if such a series of blocks do not exist, then T is not associated with a DDM and exit the procedure.

Add all the duplication blocks obtained into DS;
Construct T' from T as described after Lemma 2, and recursively determine if T' is associated to a DDM or not and update DS accordingly.

Output duplication blocks that are stored in DS.

This implies the above recursive algorithm for reconstructing a double duplication model (DDM) from a phylogeny. Since we can charge the number of comparisons taken in different recursive steps to disjoint left paths in the input tree T, the whole algorithm takes at most $2 \times 2n$ comparisons for determining all the duplication blocks. Therefore, the above algorithm takes linear time for reconstructing a double duplication model from a phylogeny T if it exists.

2.2.2 Arbitrary duplication models Now, we generalize the above algorithm into arbitrary duplication models. Again, we assume the leftmost paths leading to leaf 1 and leaf 2 in T are given in (1) and (2) respectively. Then, we have the following observation:

> Assume a phylogeny T is associated with a duplication model $\mathcal{M}$. Then, there exist $p-i-1$ double duplication blocks $[v_{i+k}, u_{j_k}]$ $(1 \leqslant k \leqslant p-i-1)$ such that, after these duplications are placed in T, the leaf 2 is right next to the leaf 1. But, these double duplication blocks may not be in $\mathcal{M}$.

We construct a new phylogeny T' from T in the same way as described in Section 2.2.1. Recall that there are two types of nodes on the leftmost path of T'. Some nodes are original ones in the input tree T; some are inserted due to duplication blocks we have examined so far. To extend the existing duplication blocks to larger ones, we associate a flag to each original node on the leftmost path of T', which indicates whether the node is in an existing duplication block or not.

Let x be an original node on the leftmost path P of T' appearing in a duplication block $[x_1, x_2, \cdots, x_t, x]$ of size $t+1$ so far, then, there are t inserted nodes x'_i right below x on the path P, which correspond to x_i for $i \leqslant t$. To determine whether $[x_1, x_2, \cdots, x_t, x]$ can be extended to a large duplication block in the model with which the original tree T is associated, we need to consider x and all the x'_is $(1 \leqslant i \leqslant t)$ simultaneously. For this purpose, we introduce the concept of *hyper-double (duplication) blocks*. We say that x and y form a hyper-double block $[x, y]$ in T' if the following three conditions hold: (i) x is a node in some non-single duplication block that we have obtained so far; (ii) x and y are not an ancestor of each other; (iii) the block $[x_1, x_2, \cdots, x_t, x]$ can be extended to a block $[x_1, x_2, \cdots, x_t, x, y]$ of size $t+2$ in the original tree T. Hence, when we place a hyper-double block $[x, y]$ in the current tree T', the edge $(y, l(y))$ crosses not only the edge $(x, r(x))$, but also the edges $(x'_i, r(x'_i))$, $1 \leqslant i \leqslant t$. With these notations in mind, we have that a phylogeny T is associated with a model if and only if (i) there exist $p-i-1$ double duplication blocks $[v_{i+k}, u_{j_k}]$ $(1 \leqslant k \leqslant p-i-1)$ in T such that, after these duplication blocks are placed in T, leaf 2 is right next to leaf 1, and (ii) T'

constructed above is associated to 'a duplication model' with introducing hyper-double duplication blocks. Hence, the algorithm in Section 2.2.1 can be generalized to an arbitrary duplication model.

To make the algorithm run in linear time, we refine the algorithm in two aspects. First, we assign a pair (R'_x, R''_x) of indices to a node x on the leftmost path of T' in each recursive step: if x is in a duplication block $[x_1, x_2, \cdots, x_t, x]$ in the current stage, we set $R'_x = R_{x_1}$ and $R''_x = R_x$, which are defined in Section 2.2.1. Since $R'_x < R_{x_i} < R''_x$ for $2 \leqslant i \leqslant t$, only R'_x and R''_x will be examined for determining if x is in a hyper-double block in next step. Secondly, if the duplication block $[x_1, x_2, \cdots, x_t, x]$ is extended into a larger hyper-double block $[x_1, x_2, \cdots, x_t, x, y]$ in a step, the binary tree T' for next step is constructed by inserting the right subtrees of x_i's and x into the edge between y and its left child $lc(y)$. To do these insertions, we need to point the left child of x_1 to $l(y)$, and then point the left child of y to x. In this way, we are able to insert all the subtrees in only two pointer operations.

Now, we illustrate the algorithm using the phylogeny given in Figure 4. To arrange 2 next to 1 in step 1, we obtain 3 hyper-double duplications $[v_1, v_2]$, $[v_3, v_5]$ and $[v_8, v_6]$. After this step, leaves $1, 2, 3, 4$ have been arranged in increasing order. To arrange 5 next to 4 in step 2, we obtain a hyper-double block $[v_5, v_4]$ and hence we extend the double duplication $[v_3, v_5]$ to $[v_3, v_5, v_4]$; finally, to arrange 6 next to 5 in step 3, we obtain a hyper-double block $[v_4, v_7]$ and further extend the duplication $[v_3, v_5, v_4]$ to $[v_3, v_5, v_4, v_7]$. After step three, all sequences have been arranged in the increasing order and the algorithm terminates successfully by outputting three duplications $[v_1, v_2]$, $[v_3, v_5, v_4, v_7]$ and $[v_8, v_6]$.

2.3 Inferring duplication models from sequence data

Let $F = \{s_1, s_2, \cdots, s_n\}$ be a set of tandem sequences of a fixed length l over an alphabet A and let $\mathcal{M}$ be a duplication model of F, in which each node v is assigned with a sequence $s(v) \in A^l$, where $s(v)$ is in F if v is a leaf. We measure $\mathcal{M}$ using the following cost $C(\mathcal{M})$:

$$C(\mathcal{M}) = \sum_{(u,v)\in E(\mathcal{M})} dist(s(u), s(v)),$$

where $E(\mathcal{M})$ denotes the set of directed edges in $\mathcal{M}$ and $dist(,)$ is a distance function.

By the parsimony principle, inferring a duplication model on F is to find a duplication model of minimum cost. Such an inference problem is NP-hard as proved in Jaitly *et al.* (2001). Therefore, there is unlikely a polynomial time algorithm for it. Here, we introduce an efficient heuristic for inferring a duplication model from sequence data using the algorithm in Section 2.2.2 and the *nearest neighbor interchange* (NNI) operation for tree transformation. Every edge, e, in an unrooted tree partitions the set of leafs. We may consider alternative hypothesis associated with e by applying an NNI on e. This rearrangement generates alternate bipartition for e while leaving all the other parts of the tree unchanged (see Swofford *et al.*, 1996, for more information).

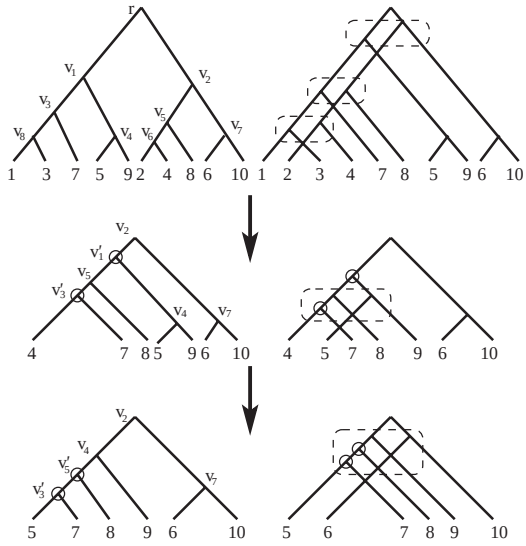

Fig. 4. Step-by-step demonstration of reconstructing a duplication model from a phylogeny. Here, we use a box drawn in dashed lines to denote a hyper-double block; we mark the inserted nodes with circles.

Our method is analogous to the Feng–Doolittle algorithm for multiple alignments. First, an unrooted guide tree is constructed for the given set of sequences using the parsimony method or neighbor-joining (see Swofford *et al.*, 1996). Then, the NNI's are applied to the guide tree T for searching a duplication model that is closest to T in terms of NNI distance. In the kth stage, we examine all the trees that are k NNI operations away from the guide tree T. During the search process, we use the algorithm for determining whether the resulting tree is associated with a duplication model or not by rooting the resulting tree on an edge on the path from 1 to n in each stage. Since a phylogeny with n leaves can always be transformed into another by applying at most $n \log_2 n + O(n)$ NNI's (Li *et al.*, 1996), our algorithm will terminate quickly.

3 RESULTS

We now apply our method to three data sets: tandem repeats in the human mucin gene MUC5B, zinc-finger (ZNF) genes, and olfactory receptors (OR). These sequences are strongly believed to be generated by tandem

duplication. Our aim is to provide an accurate and reliable history of duplication events that has given rise to the observed repeats.

3.1 Mucin gene MUC5B

Mucus consists mainly of mucin proteins, which are heterogeneous, highly glycosylated and produced from epithelial cells. Four mucin genes MUC6, MUC2, MUC5AC and MUC5B have been identified within a 400 kb genomic segment in human 11p15.5 (Pigny *et al.*, 1996). All mucin genes contain a central part of tandem repeats. In the central part of MUC5B, a single large exon of about 11 K bp (Desseyn *et al.*, 1997), three Cyc-subdomains are followed by four repeats. Each repeat contains an R-subdomain (11 or 17 tandem repeats of a motif of 29 amino acids), an R-End subdomain, and a Cys-subdomain. Another R-subdomain of 23 tandem repeats of 29 amino acids follows the fourth repeat. This suggests that the central part of MUC5B arose through tandem duplications.

The amino acid sequences for the five R-subdomains were taken from (Desseyn *et al.*, 2000) and aligned using an alignment program at Michigan Tech (Huang, 1994). We chose the following parameters—substitution matrix: Blosum62, mismatch score: -15, gap open penalty 10, gap extension penalty 5. The alignment was used to produce a guide tree by using the parsimony method (Felsentein, 1995; Lim and Zhang, 1999), which is rooted on the edge that partitions the set of leafs into $\{RI, RII, RIV\}$ and $\{RIII, RV\}$. Then we applied our algorithm to this guide tree. The resulting tandem duplication history for the central part of MUC5B agrees with the evolutionary history described verbally by Desseyn *et al.* (2000) in their original papers.

We also attained a tandem duplication hypothesis of the third repeat segment RIII. It is composed of 17 tandem repeats (RIII-1 to RIII-17) of an irregular motif of 29 amino acids rich in Ser and Thr. We obtained a guide parsimony tree from a multiple alignment of these repeats (see Fig.5a). Then, we derived a duplication model using 3 local NNI arrangements in the guide tree (see Fig. 5b). This model has 126 substitutions. This hypothesis is also consistent with the conclusion drawn in Desseyn *et al.* (2000).

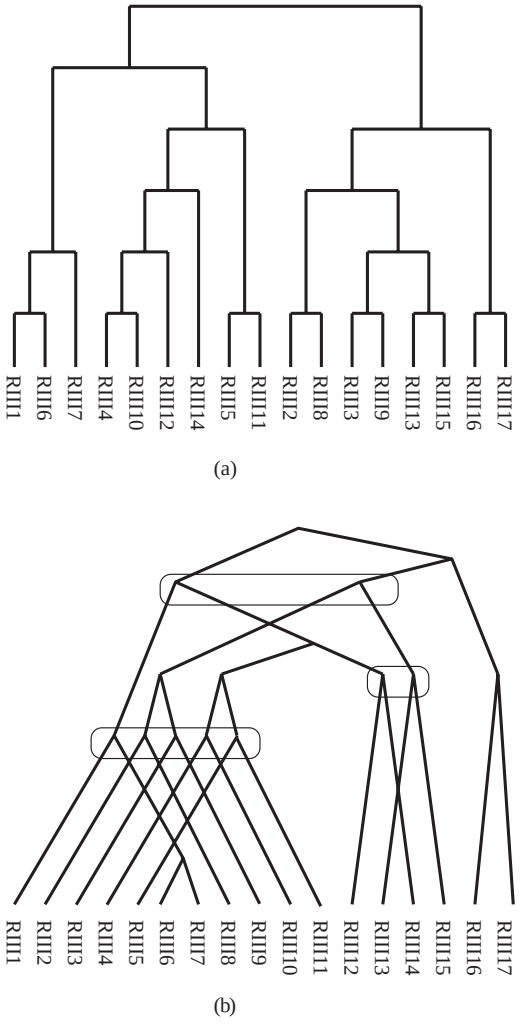

Fig. 5. The duplication history of the third repeat RIII in the MUC5B central part.

3.2 ZNF and OLF gene families

Human genome contains roughly 700 *Krüppel*-type (C2H2) zinc-finger (ZNF) genes, which encode putative transcription factors, and 900 olfactory receptors (OR), which encode proteins that recognize distinct types of olfactants and that function in different regions of the olfactory epithelium. Because of their importance in species-specific aspects of biology, Dehal *et al.* (2001) analyzed the evolutionary relationship between C2H2 ZNF genes and OR genes in human chromosome 19 (HSA19). 262 C2H2 ZNF genes identified on the HSA19 were grouped into 11 different locations (Z1 to Z11); 49 OR genes were grouped into four location clusters (OLF1 to OLF4) (see Tables C and G in the Supplementary material of Dehal *et al.* (2001)). *In situ* tandem duplication events are likely to have given rise to these features.

Because of their locations, we choose gene clusters ZNF45 and OLF3 and obtained their gene-duplication history using our algorithm. The gene family ZNF45 contains 16 genes and OLF3 14 genes in HSA19q13.2. The alignment of each gene family was used to produce a matrix of pairwise distances between the family members and a guide tree by using the Neighbor-Joining method (Felsentein, 1995; Lim and Zhang, 1999). Then, we applied our algorithm to the guide tree. Our algorithm

performed better than the WINDOW method on the gene family ZNF45 according to the parsimony principle. Our model for ZNF45 contains two double duplications and has 4690 substitutions while the one in Tang *et al.* (2001) has 4800 substitution, which were constructed from the same alignment data.

4 DISCUSSION

We have presented an efficient algorithm for inferring a duplication model from sequence data. Such an algorithm performed well on three test data sets—MUC5B, gene families ZNF45 and OLF3. The obtained duplication models indicates that single-copy duplication occurs more often than multiple-copy duplications (Tang *et al.*, 2001; Elemento *et al.*, 2002). If this is indeed the case, then, the following reconstructing algorithm is more efficient:

> **Greedy Search:** We root the given guide tree in the middle edge on the path from leaf 1 to leaf n using the molecular clock hypothesis. For each $k = 2, 3, \cdots, n-1$, we repeatedly arrange k next to $k-1$ using duplications as described in last two subsections. If k cannot be placed next to $k-1$ by introducing duplications, we just move k to the branch having $k-1$ as an end so that these two leaves become siblings.

4.1 Simulation tests

We test the Greedy search method on the simulated data sets generated using the following procedure. First, a random tree is generated using DNATree (Kuhner and Felsenstein, 1994). Then, a random duplication model $\mathcal{M}$ is obtained from the tree by randomly placing multiple duplication blocks. Branch lengths are then assigned on $\mathcal{M}$ so that the model satisfies the molecular clock assumption as shown in Figure 1.

To test our method, we generated a multiple sequence alignment from the model $\mathcal{M}$ by Seq-Gen (Rambaut and Grassly, 1997) and obtained a guide tree using the NJ method. Then, we obtained a duplication hypothesis $\mathcal{M}'$ by applying the Greedy search method on this guide tree.

We run our method against various data sets generated by the above procedure using different parameter settings: mutation rate is from 0.05 to 0.50 (increments of 0.05), sequence number from 10 to 60 (increments of 5), sequence length from 100 to 800 (increments of 200). For each combination of parameter values, 50 data sets were generated. We measure the performance of our method by counting the number of duplication events our method recovers. The results indicate that its performance improves as sequence length increases or sequence number decreases. We also noticed that in our experiments, mutation rate shows no great impact on performance.

4.2 Inversion and deletions

It seems that inversion or deletion cannot straightforwardly be incorporated into the tandem duplication model we have studied here. If we allow gene deletions in the model, every phylogeny may be associated with a model. As a future research topic, we shall study how to include gene inversion and deletion into the duplication model.

ACKNOWLEDGEMENTS

The authors would like to thank one referee for suggestions in revising Section 2.2. L.Zhang thanks M.Steel for useful discussions and O.Gascuel for bringing Fitch (1977) and Elemento *et al.* (2002) to our attention after this work was done. M.Bin thanks C.Liang for Desseyn *et al.* (2000) and thank M.Tang for sharing their data. L.Zhang is financially supported by BMRC01/1/21/19/140. B.Ma is supported by NSERC RGP0238748. L.Wang and Y.Xu are supported by a RGC grant CityU/1087/00E.

REFERENCES

Benson,G. and Dong,L. (1999) Reconstructing the duplication history of a tandem repeat. In *Proceedings of ISMB'99*. pp. 44–53.

Dehal,P. *et al.* (2001) Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution. *Science*, **293**, 104–111.

Desseyn,J.L. *et al.* (1997) Human mucin gene MUC5B, the 10.7-kb large central exon encodes various alternate subdomains resulting in a super-repeat. Structural evidence for a 11p15.5 gene family. *J. Bio. Chem.*, **272**, 3168–3178.

Desseyn,J.L. *et al.* (2000) Evolution of the large secreted gel-forming mucins. *Mol. Biol. Evol.*, **17**, 1175–1184.

Elemento,O., Gascuel,O. and Lefranc,M.-P. (2002) Reconstructing the duplication history of tandemly repeated genes. *Mol. Biol. Evol.*, **19**, 278–288.

Fitch,W. (1977) Phylogenies constrained by cross-over process as illustrated by human hemoglobins in a thirteen cycle, eleven amino-acid repeat in human apolipoprotein A-I. *Genetics*, **86**, 623–644.

Felsentein,J. (1995) *PHYLIP, version 3.57c*. Department of Genetics, University of Washington, Seattle.

Huang,X. (1994) On global sequence alignment. *Comput. Appl. Biosci.*, **10**, 227–235.

Jaitly,D. *et al.* (2001) Methods for reconstructing the history of tandem repeats and their application to the human genome. *J. Comp.. Sci. Sys.*, in press.

Kim,J. *et al.* (2001) Homology-driven assembly of a sequence-ready mouse BAC contig map spanning regions related to the 46-Mb gene-rich euchromatic segments of human chromosome 19. *Genomics*, **74**, 129–141.

Kuhner,M.K. and Felsenstein,J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, **11**, 459–468. (Erratum 12: 525 1995).

Li,M., Tromp,J. and Zhang,L. (1996) On the nearest neighbor

interchange distance between evolutionary trees. *J. Theor. Biol.*, **182**, 463–467.

Lim,A. and Zhang,L. (1999) WebPHYLIP: a web interface to PHYLIP. *Bioinformatics*, **12**, 1068–1069.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer, New York.

Pigny,P. *et al.* (1996) Human mucin genes assigned to 11P15.5: identification and organization of a cluster of genes. *Genomics*, **8**, 340–352.

Shen,S., Slightom,J. and Smithies,O. (1981) A history of the human fetal globin gene duplication. *Cell*, **26**, 191–203.

Rambaut,A. and Grassly,N.C. (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.

Swofford,D.L. *et al.* (1996) Phylogeny inference. In Hillis,D.M. *et al.*, (eds), *Molecular Systematic*. Sinauer Associate, MA, pp. 407–514.

Tang,M., Waterman,M. and Yooseph,S. (2001) Zinc finger gene clusters and tandem gene duplication. In *Proceedings of RECOMB'01*. pp. 297–304.