Reconstructing the Duplication History of Tandemly Repeated Genes

Olivier Elemento,† Olivier Gascuel,* and Marie-Paule Lefranc†*

*Département d'Informatique Fondamentale et Applications, LIRMM, 161 rue Ada, 34392 Montpellier, France; and †Laboratoire d'Immunogénétique Moléculaire, LIGM, Université Montpellier II, UPR CNRS 1142, IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

We present a novel approach to deal with the problem of reconstructing the duplication history of tandemly repeated genes that are supposed to have arisen from unequal recombination. We first describe the mathematical model of evolution by tandem duplication and introduce duplication histories and duplication trees. We then provide a simple recursive algorithm which determines whether or not a given rooted phylogeny can be a duplication history and another algorithm that simulates the unequal recombination process and searches for the best duplication trees according to the maximum parsimony criterion. We use real data sets of human immunoglobulins and T-cell receptors to validate our methods and algorithms. Identity between most parsimonious duplication trees and most parsimonious phylogenies for the same data, combined with the agreement with additional knowledge about the sequences, such as the presence of polymorphisms, shows strong evidence that our reconstruction procedure provides good insights into the duplication histories of these loci.

Introduction

Tandemly repeated DNA sequences consist of two or more adjacent copies of a stretch of DNA, together forming an array of consecutive repeated sequences. They arise from tandem duplication, in which a sequence of DNA (which may itself contain several repeats) is transformed into two adjacent copies. Because copies are then free to evolve independently and are likely to undergo additional mutation events, they become approximate over time. Tandemly repeated sequences are often termed paralogous sequences because their homology arises via duplication (in contrast with orthologous sequences, where homology arises through speciation). There are three main kinds of tandemly repeated sequences: (1) microsatellites, whose basic motif, generally 2–10 nucleotides long, may be repeated unchanged for up to thousands of times, (2) minisatellites which have core repeating units of 10–100 bases and differ from microsatellites in that each repeat unit may vary slightly in length and base sequence, and (3) larger tandem repeats (from 0.1 to 200 kbp). Besides size, fundamental differences exist between minimicrosatellites and larger repeats; the former do not contain any genes, whereas the latter often do.

The three main distinct mechanisms which generate tandem duplication of DNA stretches are slipped-strand mispairing, gene conversion, and unequal recombination. The latter (also known as unequal crossover) is widely viewed as the predominant biological mechanism responsible for the production of medium to large tandemly repeated sequences. Various examples have been described (Ohno 1970; Smith 1976; Jeffreys and Harris 1981; Collins and Weissman 1984; Gumucio et al. 1988; Ruddle et al. 1994; Honjo and Alt 1995, p. 269). Recombination (Alberts et al. 1995, p. 863) arises during meiosis, just after chromosome duplication, when chromosomes line up in tetrad configuration. At this time they can exchange segments of DNA. In most cases, recombination does not produce repeated segments because chromosomes are well aligned. However, because of the presence of short repeated sequences, unequal pairing of the chromosomes may sometimes occur, and the shift between both chromatids duplicates a fragment of DNA. Because a DNA fragment from one chromosome is transported to another chromosome, unequal recombination also deletes a fragment from one of the two chromosomes. This duplication mechanism is illustrated in figure 1, step 1 (sequences are shortened for the purpose of illustration). Tandemly repeated sequences in turn increase the likelihood of additional tandem duplications (fig. 1, step 2) because they increase the possibilities of mispairing. Block duplication, or simultaneous duplication of several genes in tandem (as shown in fig. 1, step 3), was also found to have occurred in several loci (Lefranc et al. 1986; Corbett et al. 1997; Hordvik et al. 1999).

Gene duplications (in tandem or not) give rise to gene families that are one of the most important evolutionary mechanisms for producing genes with novel functionalities (Ohno 1970; Li 1997, p. 269). Accurate reconstruction of tandem duplication histories is therefore an important issue because it would allow scientists to have a better understanding of the evolution of gene families. More specifically, when applied to immunogenetics data, it should provide valuable insights into the origin and behavior of the immune repertoire.

The first manual reconstruction of the duplication history for a complete locus containing tandemly repeated genes was apparently related in Shen, Slightom, and Smithies (1981) for the human fetal globin. Algorithms for the reconstruction of the ancestral predoubling genome, from a set of chromosomes divided into segments, are presented in El-mabrouk (2000) and applied to the genome duplication that may have occurred in *Saccharomyces cerevisiae*. The emerging field of genome rearrangement describes edit distances (Sankoff and Blanchette 1999) between species as the minimum number of inversions, translocations, duplications, and deletions necessary to transform one genome into another and uses these distances for phylogenetic infer-

Address for correspondence and reprints: Olivier Gascuel, Département d'Informatique Fondamentale et Applications, LIRMM, 161 rue Ada, 34392 Montpellier, France. E-mail: gascuel@lirmm.fr.

Mol. Biol. Evol. 19(3):278–288. 2002

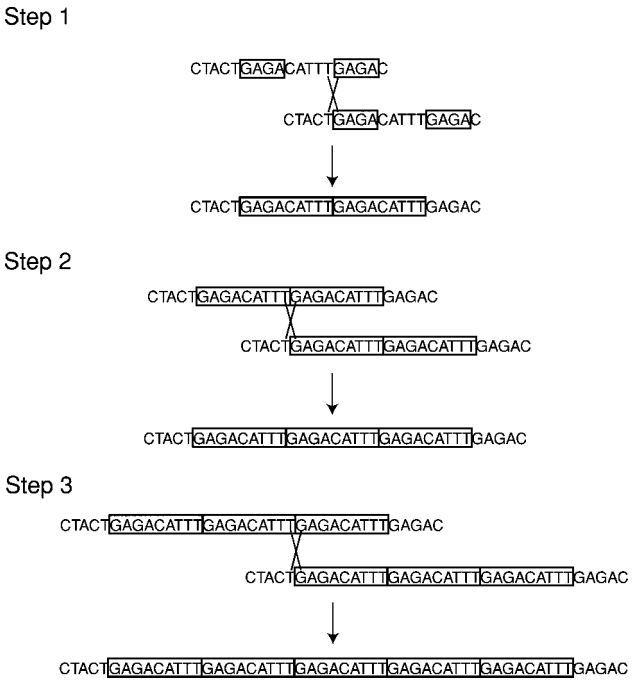

FIG. 1.—Different kinds of duplication events following chromatids misalignment during meiotic crossover. An initial duplication is caused by the presence of small repeated segments (step 1), then additional duplications are favored by the presence of several copies (step 2), and block duplications are possible (step 3).

ence. Closer to our problem, algorithms for phylogenetic analysis of minisatellites were previously presented in (Benson and Dong 1999). However, large repeats produced by unequal recombination have not received much attention to date, and we could not find any program or description of an algorithm for automated reconstruction of duplication histories.

In this paper, we deal with the problem of reconstructing the duplication history of a set of large tandemly repeated genes. We suppose that the main biological mechanism responsible for the generation of tandem repeats is the unequal recombination, and we adopt a single locus approach, i.e., we do not use sequence data from other species or from other loci. We also suppose that our loci have continually expanded via duplications and did not undergo any deletions. Indeed, comparisons between distinct species (Vijverberg and Bachmann 1999) seem to show that positive selection tends to make loci expand, probably in order to generate diversity. However, this hypothesis will be discussed at the end of this article. Finally, we assume that our sequences were not affected by gene conversion events. These assumptions form the basis of the methods and algorithms we present in this paper, and we will see that they are in good agreement with the sequences we study.

In the following sections, we describe the mathematical model of evolution by tandem duplication, as induced by the above assumptions. We then present a simple, exhaustive procedure that searches for the best duplication trees, according to the maximum parsimony criterion, when given a set of ordered and aligned DNA sequences. Finally, we analyze two data sets of tandemly

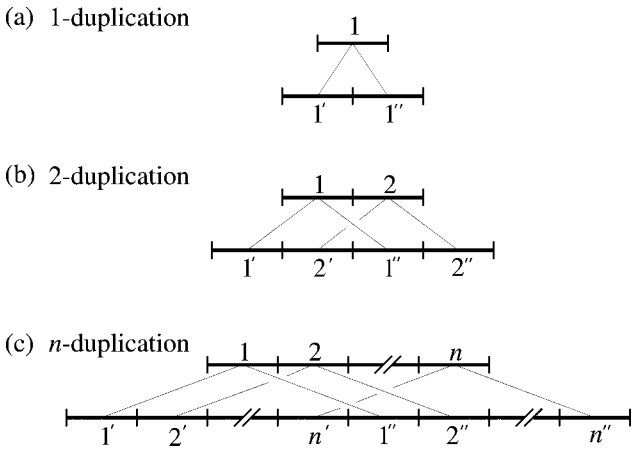

FIG. 2.—*a*, 1-Duplication. *b*, 2-Duplication. *c*, *n*-Duplication.

repeated sequences, the TRGV and the IGLC loci, obtained from the Immunogenetics DataBase IMGT.

Models and Algorithms

Possible Duplication Events

Consider m tandemly repeated sequences originating from the same locus. They form a set of homologous sequences $\{1, 2, \ldots, m\}$ whose elements are ordered according to their position. Adjacency between two repeats i and j is denoted as $i \prec j$. Assuming unequal recombination is the sole mechanism responsible for generating the repeats, the duplication process involves replacing a fragment of DNA with two identical and adjacent copies of itself. When this fragment only contains a single repeat, we say that the duplication event is a 1-duplication. As represented in figure 2*a*, duplication of repeat 1 results in two identical and adjacent repeats which diverge through mutations and become $1'$ and $1''$, with 1 as common ancestor and $1' \prec 1''$. When the duplicated fragment contains 2, 3, or n repeats, we call the duplication event a 2-, 3-, or n-duplication. In figure 2*b*, 2-duplication of repeats $1 \prec 2$ creates two new repeats, identical and immediately adjacent to the initial $1 \prec 2$ segment (i.e., $1 \prec 2 \prec 1 \prec 2$), and after mutation we obtain $1' \prec 2' \prec 1'' \prec 2''$. $1'$ and $1''$ share a common ancestor but are now separated by $2'$ on the locus. When representing this kind of event, branch crossing is necessary to respect leaf ordering. The same holds for n-duplications (fig. 2*c*).

Duplication Histories

The series of consecutive duplication events which has given rise to the m repeated sequences, can then be represented as a rooted tree with labeled and ordered leaves, which we call a duplication history. To be more precise, this tree should be called a ''tandem duplication history,'' but we choose to abbreviate this denomination for conciseness. In a duplication history, internal nodes correspond to duplication events, ordered from top to bottom according to the moment they occurred during the course of evolution. Because a tandem repeat's locus initially contains a single copy, a duplication history is

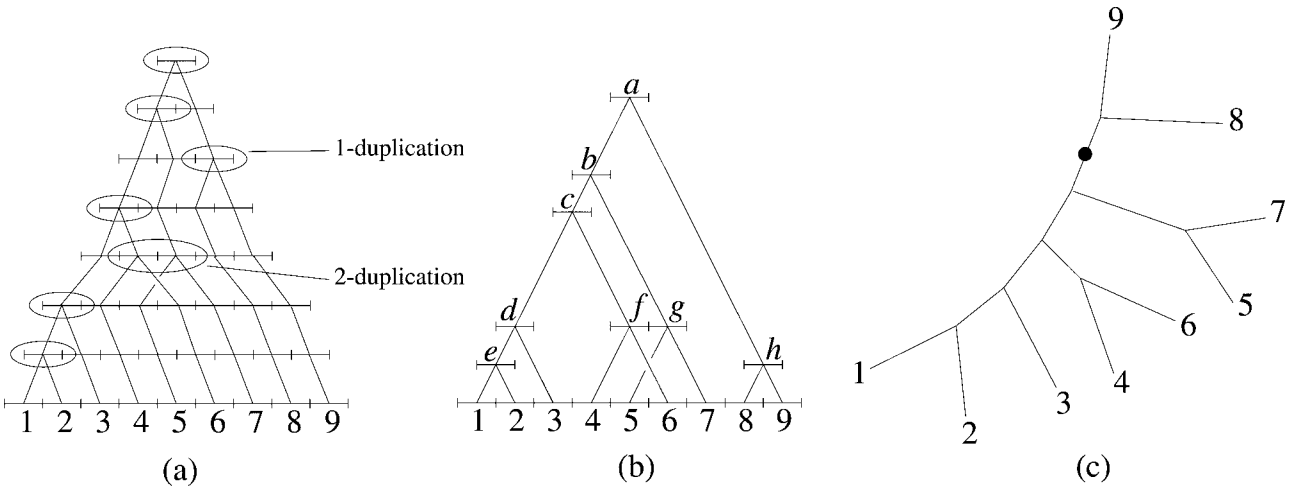

FIG. 3.—*a*, Duplication history. *b*, Partially ordered duplication history. *c*, Duplication tree. The black dot indicates the position of the root in (*b*).

rooted by essence, and the root of a duplication history always lies on the branches linking sequences at both extremities of the locus. A hypothetical duplication history for a locus containing 9 repeats is shown in figure 3*a*.

Partially Ordered Duplication Histories

As in phylogenetic reconstruction, the molecular clock often does not hold with duplicated genes, making the order between the duplication events of two different lineages impossible to recover from the sequences. In this case, we are only able to infer what we call a partially ordered duplication history, i.e., a duplication history in which the duplication events are partially ordered. For example, in figure 3*b*, duplication *c* is after duplication *b*, but the relationship between *c* and *h* is undetermined.

In the initial duplication history, the adjacency relationships between ancestral copies (denoted $i \prec j$, see above) can be clearly identified throughout the evolution of the locus. In a partially ordered duplication history, these relationships are no longer represented. However, not all the adjacency relationships between ancestral copies are possible. For example, in figure 3*b*, $e \prec 3 \prec f \prec g \prec h$ is possible, whereas $e \prec 3 \prec f$ only or $e \prec 3 \prec 4 \prec 5 \prec b$ would never occur. In fact, it can be shown that the possible ancestral combinations of adjacent copies are given by the maximal antichains (Atallah 1999, p. 13) of the partial order on the duplication events.

Duplication Trees

Another consequence of the absence of a molecular clock is that the position of the root cannot generally be recovered from the sequences. Unrooting a partially ordered duplication history creates what we call a duplication tree. As mentioned previously, a more precise denomination would be tandem duplication tree. A duplication tree is an unrooted phylogeny with ordered leaves whose topology is compatible with at least one phylogeny induced from a duplication history (a more formal definition is given below). Figure 3*c* shows a hypothetical duplication tree that is compatible with the duplication histories shown in figures 3*a* and *b*.

In turn, rooting a duplication tree may or may not produce a valid partially ordered duplication history because the mathematical model of duplication allows the root of a duplication history to be located somewhere on the path linking the most distant genes, but not everywhere. For example, in figure 4 there are three potential root locations (a, b, and c), and only two of them (b and c) lead to valid partially ordered duplication histories.

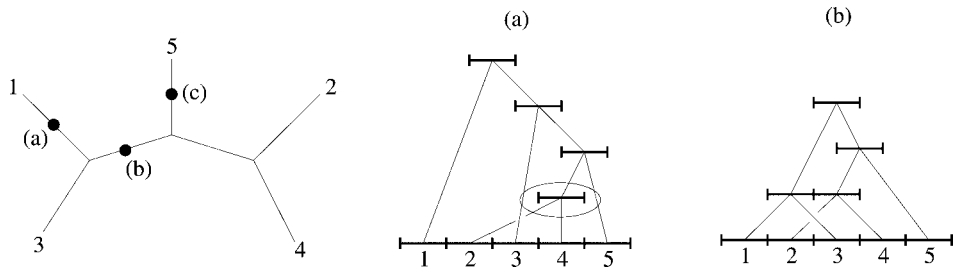

FIG. 4.—Not all potential root locations lead to valid partially ordered duplication histories. History (a), which represents the above tree rooted at position (a) is not valid because it contains an event that is not a tandem duplication: 2 and 4 are not adjacent but separated by 3. However, rooting the tree at position (b) leads to a valid partially ordered duplication history, and the same holds for position (c) (not shown).

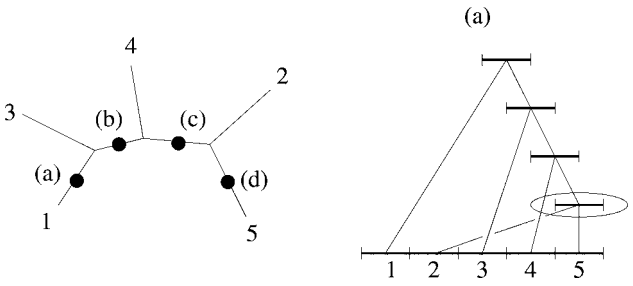

FIG. 5.—Not all phylogenies with ordered leaves are duplication trees. The above unrooted phylogeny can be rooted at locations (a), (b), (c), or (d), but none of them lead to a valid partially ordered duplication history. For example, when rooting at position (a), the obtained partially ordered duplication history contains a duplication event in which the duplicated genes 2 and 5 are not adjacent. The same holds for (b), (c), and (d) (not shown).

By extension, we say that given an order on its leaves, an unrooted phylogeny is a duplication tree if, among all possible root locations, at least one leads to a valid partially ordered duplication history. For 2, 3, and 4 leaves, it is easily shown that every phylogeny is also a duplication tree. However, when considering 5 or more leaves, not all phylogenies can be duplication trees for every ordering of the leaves (fig. 5).

Determining Whether a Phylogeny is a Duplication Tree

To determine whether a given rooted phylogeny is also a partially ordered duplication history, we devised a simple algorithm, called Possible Duplication History (PDH). This algorithm takes as input a rooted phylogeny and an order on its leaves, detects the patterns corresponding to duplication events, and outputs whether or not it is a partially ordered duplication history.

The PDH algorithm also allows us to determine whether a phylogeny can be a duplication tree; we simply run it on each root location along the path linking the leaves associated with the two most distant sequences on the locus. If at least one of these roots yields a valid partially ordered duplication history, the unrooted phylogeny is a possible duplication tree. The PDH algorithm thus provides us with a mathematical characterization of both the partially ordered duplication history and the duplication tree objects.

We define the following notation: (1) T is a rooted phylogeny with ordered leaves and (2) a cherry (i, u, j) is a pair of leaves (i and j) separated by a single node u in T; we call $C(T)$ the set of cherries of T. The PDH algorithm is a recursive procedure, which progressively reduces T by agglomerating the cherries that belong to recognized duplication events. It merges cherries until T has been reduced to its root, meaning that it constitutes a valid partially ordered duplication history, or it cannot go further, in which case T cannot be a partially ordered duplication history. It must be noted that the order in which cherries are agglomerated is not important; a cherry belongs to at most one duplication event. The PDH algorithm is given in figure 6.

PossibleDuplicationHistory(Tree T, Order O, Root R)
if $(T = R)$ **then**
  return true
else if $(\exists\ (i_1, u_1, j_1), (i_2, u_2, j_2), ..., (i_k, u_k, j_k) \in C(T)$ and $i_1 \prec i_2 \prec ... \prec i_k \prec j_1 \prec j_2 \prec ... \prec j_k \subset O)$ **then**
  remove $i_1, j_1, i_2, j_2, ..., i_k, j_k$ in T
  replace $i_1 \prec i_2 \prec ... \prec i_k \prec j_1 \prec j_2 \prec ... \prec j_k$ by $u_1 \prec u_2 \prec ... \prec u_k$ in O
  return(PossibleDuplicationHistory(T, O, R))
else
  return false
end if

FIG. 6.—The PDH algorithm (possible duplication history).

Number of Duplication Histories, Partially Ordered Duplication Histories and Duplication Trees

In this section, we compute (or estimate) the number of duplication histories, partially ordered duplication histories, and duplication trees for a given number of repeats *n*. As we shall see in the next sections, counting these objects allows various reconstruction procedures to be compared, and it also gives us sound arguments supporting our tandem duplication model.

A locus containing n repeats can be obtained from any of $(n - 1)$ 1-duplications from a locus containing $(n - 1)$ repeats or from any of $(n - 3)$ 2-duplications from a locus containing $(n - 2)$ repeats, etc. Therefore, the number $DH(n)$ of possible duplication histories for n repeats is given by the following recursive formula:

$$DH(n) = (n - 1)DH(n - 1) + (n - 3)DH(n - 2) + \cdots + (n - 2\alpha + 1)DH(n - \alpha) \ldots,$$

where the last term is equal to $DH(n/2)$ if n is even and otherwise to $2DH((n + 1)/2)$. Moreover, we have $DH(1) = 1$.

To count the number of partially ordered duplication histories $PODH(n)$ for n repeats, we used the PDH algorithm. For relatively small n (i.e., for $n \leq 10$), we generated every possible rooted phylogeny and applied the PDH algorithm to obtain the total number of partially ordered duplication histories. For $n > 10$, this procedure takes too much time, and the number of partially ordered duplication histories has to be estimated. This estimation is given by $\widehat{PODH}(n) = \widehat{P_{PODH}}(n) \times N_{\text{rooted}}(n)$, where $P_{PODH}(n)$ is the proportion of partially ordered duplication histories among the set of all possible rooted phylogenies, and $N_{\text{rooted}}(n)$ is the total number of rooted phylogenies, that is $(2n - 3)\Pi_{i=3}^{n}(2i - 5)$ (Cavalli-Sforza and Edwards 1967). To estimate $P_{PODH}(n)$, we constructed with the uniform distribution a large number of rooted phylogenies and fed them into our PDH algorithm.

We adopted the same approach to obtain the number of duplication trees $DT(n)$. For $n \leq 10$, we generated every possible unrooted tree and obtained $DT(n)$ using the PDH algorithm. For $n > 10$, we estimated the number of duplication trees using $\widehat{DT}(n) = \widehat{P_{DT}}(n) \times N_{\text{unrooted}}(n)$, where $P_{DT}(n)$ is the proportion of dupli-

Table 1
Number of Partially Ordered Duplication Histories and Duplication Trees

n	$PODH(n)$	$P_{PODH(n)}$	$\sigma_{PODH(n)}$	$\frac{DH(n)}{PODH(n)}$	$\frac{\sigma_{DH(n)}}{PODH(n)}$	$DT(n)$	$P_{DT(n)}$	$\sigma_{DT(n)}$	$\frac{PODH(n)}{DT(n)}$	$\frac{\sigma_{PODH(n)}}{DT(n)}$
3.....	22	0.209	—	1.455	—	11	0.733	—	2	—
6.....	92	0.097	—	1.989	—	46	0.438	—	2	—
7.....	420	0.040	—	2.952	—	210	0.222	—	2	—
8.....	2,042	0.015	—	4.749	—	1021	0.098	—	2	—
9.....	10,404	5.13e-3	—	8.249	—	5202	0.038	—	2	—
10.....	54,954	1.59e-3	—	15.41	—	27,477	0.014	—	2	—
11.....	2.95e5	4.51e-4	0.047	31.23	0.047	1.46e5	4.43e-3	0.047	1.934	0.067
12.....	1.63e6	1.18e-4	0.029	67.45	0.029	8.64e5	1.32e-3	0.027	1.879	0.040
13.....	9.52e6	3.01e-5	0.057	148.8	0.058	5.02e6	3.34e-4	0.055	2.072	0.080
14.....	5.22e7	6.60e-6	0.087	377.6	0.089	3.35e7	9.35e-5	0.032	1.764	0.093

NOTE.—n is the number of leaves. $PODH$ (n) is the number of partially ordered duplication histories. $P_{PODH(n)}$ is the proportion of partially ordered duplication histories among rooted phylogenies. $DH(n)/PODH(n)$ is the ratio between the number of duplication histories and the number of partially ordered duplication histories; it reflects the gain obtained by the algorithmic refinement of DTEXPLORE. $DT(n)$ is the number of duplication tress. $P_{DT(n)}$ represents the proportion of duplication trees among unrooted phylogenies. $PODH(n)/DT(n)$ is the ratio between the number partially ordered duplication histories and the number of duplication trees; it reflects the gain that could be obtained by further refining DTEXPLORE. These numbers are exact for $5 \leq n \leq 10$, and estimated for $11 \leq n \leq 14$. $\sigma_{PODH(n)}$, $\sigma_{DH(n)/PODH(n)}$, $\sigma_{DT(n)}$, $\sigma_{PODH(n)/DT(n)}$ are the ratios between the standard deviations of the estimators and the estimates.

cation trees among the set of all possible unrooted phylogenies, and $N_{\text{unrooted}}(n)$ is the total number of unrooted phylogenies, i.e., $\Pi_{i=3}^{n}(2i-5)$.

To construct rooted and unrooted trees with the uniform distribution, we used the classical addition scheme described in Gascuel (2000). Estimations were computed for $11 \leq n \leq 14$ because duplication histories and duplication trees become too scarce for higher values of n (e.g., only 77 rooted phylogenies out of 5×10^7 were found to be duplication histories when $n = 15$). This procedure provided us with good estimates for $n \leq 14$. For example, in the $n = 10$ experiments, where the exact numbers can be calculated, the relative deviation between $PODH(n)$ and $\widehat{PODH}(n)$ is equal to 0.8%, whereas between $DT(n)$ and $\widehat{DT}(n)$ it is equal to 1.1%. Moreover, the standard deviations of the estimators (approximated using the Gaussian distribution assumption) are relatively low (i.e., less than 10% of the estimates) for $11 \leq n \leq 14$ and for all estimated quantities.

Table 1 provides the results of this study. It clearly shows that $PODH(n)$ and $DT(n)$ are much smaller than $N_{\text{rooted}}(n)$ and $N_{\text{unrooted}}(n)$, respectively. Linear regression on the log values suggests that the ratio between $N_{\text{rooted}}(n)$ and $PODH(n)$, the ratio between $N_{\text{unrooted}}(n)$ and $DT(n)$, and the ratio between $DH(n)$ and $PODH(n)$ are at least of exponential order in n. Finally, it appears that the ratio between $DT(n)$ and $PODH(n)$ is strictly equal to two for exact values and remains close to two with estimated values. The constancy of this ratio is surprising and indicates that better analysis of these combinatorial objects represents an interesting direction for further research.

Reconstructing the Optimal Duplication Trees

We now deal with the process of reconstructing duplication trees from a set of aligned, tandemly repeated genes, using the duplication model we have just defined. If we assume that every kind of duplication is equiprobable and that each lineage evolves independently, the classical optimality criteria from phylogenetic analysis (maximum parsimony, least squares, minimum evolution, and maximum likelihood) are relevant in the scope of duplication trees. In this study, we adopt a maximum parsimony approach. We apply this criterion in the usual way, i.e., we use the topology of our duplication trees and label the leaves with our sequences. The labeling process simply consists of matching the ordered set of sequences with the ordered set of leaves belonging to the duplication tree. The algorithm that computes the maximum parsimony of a given tree topology is computationally efficient and easy to implement (Fitch 1971). Moreover, parsimony is commonly acknowledged (Swofford et al. 1996) as being a good criterion when dealing with sequences sharing more than 75% of homology, which is the case with our sequences. However, other traditional criteria could be used as well.

There are two ways to reconstruct duplication trees under the parsimony criterion. The most straightforward method is to apply a phylogenetic reconstruction program, such as DNAPENNY from the PHYLIP package, to our aligned sequences and check using PDH whether the output tree(s) is a duplication tree or not. This method can be efficient in practice, but it is not guaranteed to find a duplication tree because DNAPENNY does not restrict its search to duplication trees. The other method uses an algorithm that only searches the space of duplication trees. Although we cannot explore the space of all possible phylogenies for n sequences in reasonable time, it becomes possible to explore the more restricted space of all duplication trees, for n relatively large. A simple approach is then to generate all possible duplication trees with the desired number of leaves and select the best ones. For this purpose, we devised the DTEXPLORE algorithm, which performs a depth-first exploration of the solution space through a simulation of the unequal recombination process, as represented in figure 7. This algorithm can be summarized as follows: we start with a tree consisting of two leaves and one root and one or several leaves are duplicated, as implied by

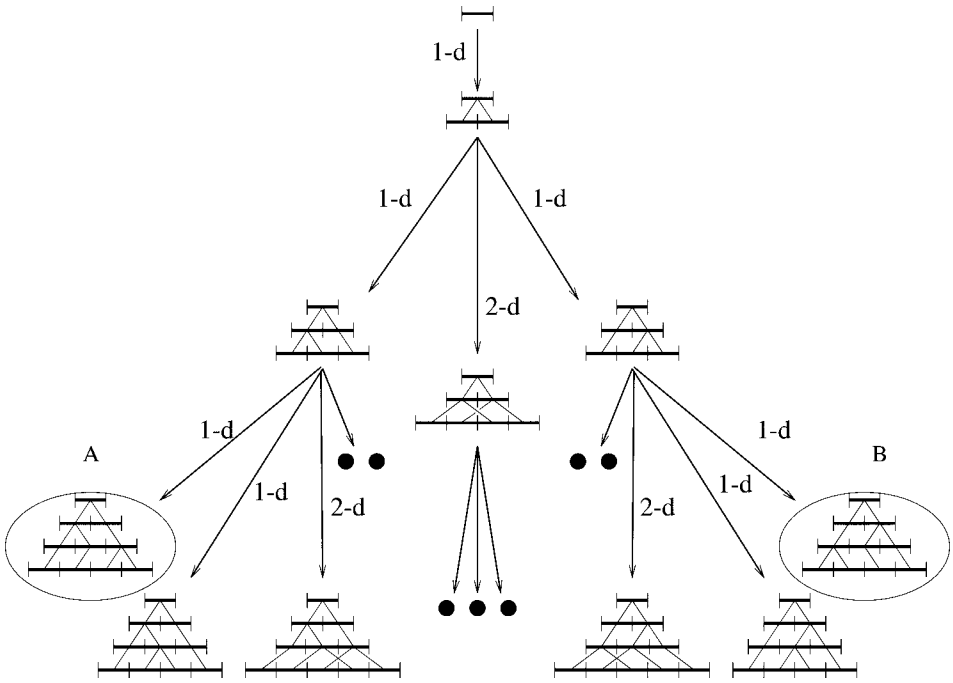

FIG. 7.—Partial view of the search space of duplication histories. For example, the locus associated with duplication history A contains four repeats and is generated by three 1-duplications. In duplication history A, the left duplication occurs before the right duplication, whereas in duplication B, the left duplication occurs after the right duplication. Because these two duplications occur in two separate lineages, duplication histories A and B correspond to the same partially ordered duplication history.

our duplication model (fig. 2). When the desired number of leaves is reached, leaves are labeled according to their order so as to associate them with gene sequences, and the parsimony value of the resulting tree is computed. The search algorithm then backtracks, and alternative duplications are tried until the search space has been completely explored. Finally, the most parsimonious trees are outputted. The DTEXPLORE algorithm is shown in figure 8.

```
procedure DTEXPLORE(Tree T, Integer pmin)
if |leaves(T)| = n then
  p ← parsimony-value-of(T)
  if p < pmin then
    L = ∅
    store T in L
    pmin = p
  else if pmin = p then
    store T in L
  end if
else
  for i = 1 to |leaves(T)| do
    for j = i to min(i + n − |leaves(T)| − 1, |leaves(T)|) do
      T′ ← T with duplication of sequences i to j
      DTEXPLORE(T′, pmin)
    end for
  end for
end if
```

FIG. 8.—The exhaustive search algorithm, n is the number of sequences being studied, L is the list of most parsimonious duplication trees found so far, and p_{min} is the parsimony value of the trees in L (with $p_{min} = \infty$ initially).

Algorithmic Refinement to DTEXPLORE

Because it generates every possible duplication history, this procedure also generates every possible duplication tree. However, this procedure usually generates the same partially ordered duplication history several times (and therefore the same duplication tree) because several duplication histories can be compatible with a single partially ordered duplication history. For example, in figure 7, duplication histories A and B correspond to the same partially ordered duplication history. In order to avoid evaluating identical duplication trees many times, we add a memorization feature to our procedure. We encode the partially ordered duplication history that corresponds to a given duplication history into a character string, according to a nonambiguous coding scheme, and store this encoded history within a data structure called a prefix tree (Aho, Hopcroft, and Ullman 1974, p. 346). Every time a new duplication history is generated, it is first encoded into a character string, and a lexicographic search is performed within the prefix tree to check whether it has already been generated or not. If it has, we stop exploring and evaluating the current subspace and explore alternative histories. Otherwise, we add the encoded history to the prefix tree and continue the search. Because it takes approximately 100 ms to compute the parsimony value of a duplication tree that has nine leaves associated with 1,300-bp-long sequences and less than 1 ms to encode a duplication history and search it within our prefix tree, this algorithmic sophistication led to enormous speed improvements. For example, only 10,404 duplication trees are evaluated for

nine repeats when turning on the memorization feature, whereas 85,820 are evaluated without memorization. The gain in CPU time obtained using this refinement is then approximately equal to the ratio between the number of duplication histories and the number of partially ordered duplication histories. For nine repeats, this ratio is approximately equal to eight, but it has an exponential increase in n (see previously and table 1).

However, the same duplication tree can still be generated several times (the number of duplication trees for nine repeats is 5,202) because several partially ordered duplication histories (corresponding to every possible root position) are sometimes compatible with a single duplication tree. Designing a more sophisticated memorization system to eliminate redundant partially ordered duplication histories represents an interesting direction for further research. However, according to our previous results, the ratio between the number of partially ordered duplication histories and the number of duplication trees seems to remain close (or be identical) to two, and so this possible refinement would at the most divide the CPU times by a factor of two.

Results

Materials, Methods, and Software

We now apply the model and the algorithms we described to two data sets of tandemly repeated genes from the human TRGV locus and the human IGLC locus. Because these sequences fall in the large tandem repeats category and each of them contains a single gene, they are strongly suspected to have been generated by unequal recombination. Our aim is to verify the hypothesis that these sequences were produced by this duplication mechanism, and once verified, to provide a complete, accurate, and reliable duplication tree for these loci. We also strive, if possible, to root this duplication tree so as to provide the most thorough understanding of the history of duplication events that has given rise to the observed repeats. These data sets come from the ImMunoGeTics database IMGT (Ruiz et al. 2000) available at http://imgt.cines.fr. Sequences extracted from the database were first aligned with CLUSTALW (Thompson, Higgins, and Gibson 1994). Positions with gaps were then excluded from the analysis because parsimony cannot deal comfortably with them.

We then applied DTEXPLORE to the multiple alignments so as to obtain the most parsimonious duplication trees explaining these sequences. With the same input, DNAPENNY, from the PHYLIP software package, was then used to compute the most parsimonious trees among all possible phylogenies, without the duplication tree restriction, using a branch-and-bound strategy (Hendy and Penny 1982). Because DNAPENNY implements Fitch parsimony and shares the traditional phylogenetic assumptions we used (such as independent evolution of lineages), the most parsimonious duplication trees and the most parsimonious phylogenies should be identical, provided our preliminary hypotheses are respected. Thus, identity (when it occurs) between the trees produced using both methods supports the hypothesis that the repeats were generated by an unequal recombination process, provided the probability for a phylogeny to be a duplication tree is small enough (which depends on the number of repeats, as we saw previously).

We used the bootstrap procedure (Felsenstein 1985) to assess the reliability of our duplication trees and the robustness of the tendency of DNAPENNY to find the same trees as DTEXPLORE. As in traditional phylogenetic analysis, this involved creating pseudosamples by randomly sampling with replacement characters from the initial multiple alignment. Each time we generated a new pseudosample, we searched for the optimal duplication trees using DTEXPLORE and for the most parsimonious phylogenies using DNAPENNY. Once every pseudosample had been analyzed, we computed the bootstrap proportion of every branch in the initial duplication tree (or phylogeny). We also computed the proportion of pseudosamples where the duplication trees found by DTEXPLORE were identical, or close, to the most parsimonious phylogenies. Both these indicators gave measures of the repeatability and tolerance of the results with respect to the sampling noise. Pseudosamples for the bootstrap procedure were generated using SEQBOOT from the PHYLIP package (Felsenstein 1989), and 1,000 pseudosamples were generated for each data set. Bootstrap computations were distributed among several computers using the Parallel Virtual Machine (PVM) library (Sunderam 1990).

As a complementary analysis, we used a Bayesian approach, implemented in BAMBE (Larget and Simon 1999), to get a quantitative assessment of the support of the unequal recombination hypothesis. Given a set of nucleotide sequences and a model of substitution, BAMBE computes the posterior probabilities of a large sample of phylogenies. We then apply the PDH algorithm to find duplication trees among these phylogenies and sum their posterior probabilities to obtain the posterior probability of our duplication model. F84 (Felsenstein and Churchill 1996) was used as a substitution model. For each data set, we ran BAMBE five times with a different random initialization.

Once our duplication trees were constructed, we rooted them using both the outgroup method and the molecular clock hypothesis on functional genes. In the first method, we selected appropriate outgroup sequences, i.e., homologous sequences from other species, with the constraint that the minimum distance between our initial sequences and the outgroup sequences had to be larger than the maximum distance within our initial sequences. We then constructed a global tree containing both our sequences and the outgroup sequences. Because our initial sequences and the outgroup sequences were relatively divergent, we computed a distance matrix from our data using the F84 model of substitution, and used BIONJ (Gascuel 1997) to construct the global tree. Using a distance approach allows us to limit the long-branch attraction phenomenon, to which parsimony methods are very sensitive (Felsenstein 1978). Also because the duplicated genes share the same environment, it is reasonable to think that at least the functional genes follow a form of molecular clock in which the total mu-

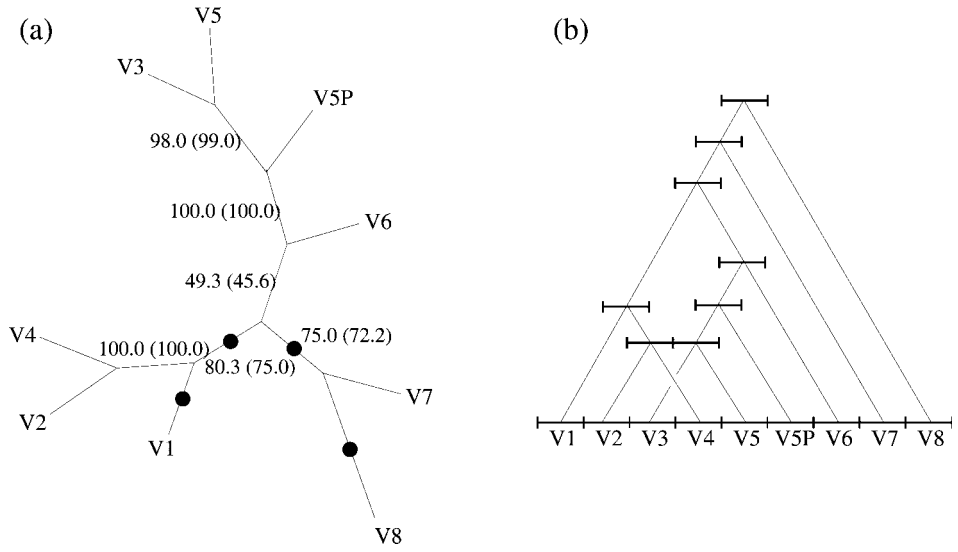

FIG. 9.—The duplication tree of the TRGV locus. Tree (*a*) is the most parsimonious duplication tree found by DTEXPLORE. The same tree was found by DNAPENNY. Bootstrap values for each internal branch of the duplication and DNAPENNY trees (between brackets) are also shown. This tree can be rooted at four branches, shown with the black dots. When rooted with the outgroup or with the molecular clock–based method (both locate the root on the branch leading to V8), it can be represented as the partially ordered duplication history (*b*).

tation rate from root to leaves is relatively constant. To root a duplication tree using this hypothesis, we constructed a tree from the initial data using BIONJ and used the branch lengths between functional genes to locate the minimum variance point.

The TRGV Locus

Our first data set stems from the human TRGV locus (Lefranc, Forster, and Rabbitts 1986; Lefranc et al. 1986), which corresponds to the variable region of the gamma T-cell receptor. It contains nine repeated genes; eight of them are 4.5–5 kb long, and the last one is slightly shorter (3 kb). These nine genes are named V1, V2, V3, V4, V5, V5P, V6, V7, and V8. Three of them are pseudogenes (V5P, V6, and V7), and one is an ORF (V1). The whole TRG locus was recently fully sequenced (accession number AF057177). We kept the DNA stretch starting 500 nucleotides upstream and finishing 500 nucleotides downstream of the coding sequences. After multiple alignment and gap removal, our sequences were 1,318 bp long.

The TRGV locus was shown to be polymorphic in French, Lebanese, Tunisian, Black-African, and Chinese populations (Ghanem et al. 1989). The most striking polymorphism is the simultaneous absence of the V4 and V5 sequences in some individuals from these populations. Another polymorphism stems from the insertion of an additional copy called V3P between V3 and V4. Unfortunately, V3P has not been sequenced so far.

DTEXPLORE evaluated all possible duplication trees with nine repeats and finally came up with the single most parsimonious one shown in figure 9*a*. DNAPENNY came up with the same tree topology as DTEXPLORE. Considering that only 3.5% of phylogenies are also duplication trees for nine leaves, the identity between the most parsimonious phylogeny and the most parsimonious duplication tree strongly supports our assumptions concerning the biological model by which repeats are generated. The bootstrap analysis for this data set showed that most of the internal branches of the duplication tree are strongly supported. Similar bootstrap proportions (using the same pseudosamples) were found to support the branches of the most parsimonious phylogeny. The bootstrap analysis also showed that DNAPENNY came up with the same trees as those found by the DTEXPLORE for 86.8% of the pseudosamples and with nearly identical trees (at most one branch of difference) for 92.7% of them. In addition, the mean BAMBE results on this data set indicated that our duplication model is supported by a posterior probability of 0.977.

We then rooted the duplication tree using seven sheep TRGV sequences (accession numbers Z12998, Z12999, Z13000, Z13001, Z13002, Z13005, and Z13006). Using the F84 model of substitution, the minimum distance between our initial sequences and these sequences is 0.3821, whereas the maximum distance within the TRGV sequences is 0.2148. Therefore, the sheep sequences can be safely used in the outgroup rooting procedure. The tree constructed with BIONJ did not modify the topology of our initial tree and located the root on the branch leading to repeat V8. This branch belongs to the set of potential root locations (the duplication tree contains 15 branches, and only 4 of them are possible). Then we constructed another BIONJ tree from our initial sequences (this tree was also identical to the most parsimonious duplication tree), and the tree distances between functional genes (V1, V2, V3, V4, V5, and V8) were used to compute the minimum variance point. It appeared to be located on the same branch as that indicated by the previous method. Therefore, both the molecular clock–based and outgroup-based rooting procedures produced the same partially ordered duplication history (shown in fig. 9*b*).

This partially ordered duplication history clearly indicates that segments V2-V3 and V4-V5 arose from a recent 2-duplication event and therefore respects the polymorphism that occurs in the TRGV locus. Although we cannot rule out that the missing segments could be explained by a deletion event, this strongly suggests that the 2-duplication simply did not occur in some populations. This agreement between our duplication tree and the polymorphism data provides further support for both our assumptions concerning the biological mechanism that produces the repeats and our reconstruction method.

The IGLC Locus

Our second data set stems from the IGLC locus (Hieter et al. 1981; Dariavach, Lefranc, and Lefranc 1987; Vasicek and Leder 1990), which corresponds to the constant region of the light chain of the Ig structure. It contains seven tandemly repeated genes (C1, C2, C3, C4, C5, C6, and C7), whose accession numbers are, respectively (J00252, J00253, J00254, J03009, J03010, J03011, and X51175), of which three are pseudogenes (C4, C5, and C6). Because the whole locus has not yet been entirely sequenced, we used the V-REGIONS (in the IMGT standardized notation) of these sequences to construct our multiple alignment. After alignment and removal of positions with gaps, each DNA sequence was 285 bp long.

DTEXPLORE evaluated all possible duplication trees with seven repeats and finally returned the single most parsimonious duplication tree shown in figure 10*a*. With the same data, DNAPENNY came up with four equally parsimonious phylogenies. One of them is identical to the duplication tree we found during the exhaustive search, whereas the remaining ones (figs. 10*b–d*) are not duplication trees. Because the IGLC locus only contains seven genes and DNAPENNY found four equally parsimonious phylogenies, the probability of one of these phylogenies also being a duplication tree is relatively high (approximately equal to 0.6). Therefore, these results are in agreement with our assumptions, but they cannot be considered as a strong support of our approach. The bootstrap analysis indicated that duplications C2-C3 and C4-C5 are strongly supported, whereas other internal branches are less supported (fig. 10*a*). Similar bootstrap proportions were found for the most parsimonious phylogenies. The bootstrap procedure also showed that for 60.3% of the pseudosamples, the phylogenies produced by DNAPENNY include the duplication trees found during the exhaustive search, and for 85.7% of the pseudosamples, every inferred duplication tree has at most one branch of difference with one of the most parsimonious phylogenies. Finally, the mean BAMBE results indicated that our duplication model is supported by a posterior probability of 0.958 on this data set.

We then used six murine IGLC sequences (accession numbers J00587, J00595, J00585, X58416, M16554, and M16628) to root our duplication tree. Using the F84 model of substitution, the minimum distance between our initial sequences and these murine sequences is 0.2938, whereas the maximum distance within the IGLC sequences is 0.2495. This indicates that the murine sequences can safely be used in the outgroup rooting procedure. The global tree produced by BIONJ locates the root on the branch leading to pseudogene C4 and thus produces a rooted tree which does not correspond to a valid partially ordered duplication history. This may stem from the shortness of the sequences or from the relatively high level of divergence of pseudogene C4 (it only shares between 72% to 78% of identity with the other six human IGLC genes, whereas the mean identity rate between IGLC genes is approximately 86%). The long-branch attraction phenomenon (Felsenstein 1978) may then cause the outgroup sequences to insert on this long branch. Another BIONJ tree was constructed from the initial gene sequences (once again, this tree was identical to the most parsimonious duplication tree), and the tree distances between the functional genes (C1, C2, C3, and C7) were used to compute the minimum variance point. It located the root on one of the allowed branches of the duplication tree (the duplication tree contains 11 branches, and only two of them are possible locations for the root), thus producing the partially ordered duplication history represented in figure 10*e*. Because there is no consensus between the rooting results, we cannot propose a partially ordered duplica-

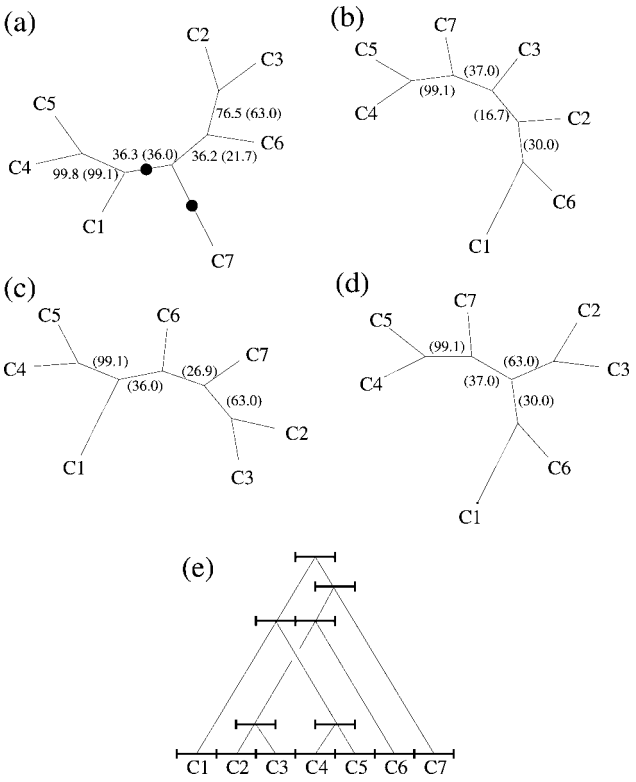

FIG. 10.—The duplication tree of the IGLC locus. Tree (*a*) is the most parsimonious duplication tree found by DTEXPLORE. DNAPENNY found the same tree and three others (*b–d*), which are not duplication trees. Bootstrap values for each internal branch of the duplication and DNAPENNY trees (between brackets) are also shown. Duplication tree (*a*) can be rooted at only two branches, shown with the black dots. When rooted by its minimum variance point, it can be represented as the partially ordered duplication history (*e*).

tion history that has a high level of certainty, as we did with the TRGV locus. However, because the partially ordered duplication history rooted by the molecular clock–based approach completely agrees with our model of duplication trees, we tend to think that it still constitutes a good candidate for explaining some parts of the duplication history of the IGLC locus.

Discussion

The duplication trees found by our exhaustive search procedure always correspond to the most parsimonious phylogenies computed on the same data. Moreover, our results are robust to sampling noise simulated by the bootstrap procedure and clearly agree with polymorphism data revealed by other biological studies. Finally, the Bayesian analysis yields very high posterior probabilities for our assumptions. These results provide further evidence that unequal recombination is a predominant mechanism in tandem repeats production. Although we need to test more data to draw firm and general conclusions, our duplication trees also suggest that single-copy duplications are predominant over multiple-copies duplications.

A duplication history which only contains 1-duplications is analogous to a binary search tree, which is a classical object in computer science. It is easily shown that if we delete a rooted subtree from a binary search tree, the resulting tree is still a binary search tree. Therefore, if a deletion occurs during the evolution of a set of tandemly repeated genes which only undergo 1-duplications, the duplication history model is still valid. This means that our duplication model is (relatively) robust to deletions, provided the duplication history only (mostly) contains 1-duplication events. However, this important property does not always hold when the duplication history contains n-duplications with $n > 1$, and more work would be needed to characterize the effects of deletions on our current duplication model.

Currently, we have two different ways to reconstruct duplication trees. The first one combines existing programs, such as DNAPENNY or DNAPARS, with our PDH algorithm to select the possible duplication trees among the optimal phylogenies. Its drawback is that it is not guaranteed to find a duplication tree because there may be some cases where none of the optimal phylogenies are duplication trees. In the second one, we use DTEXPLORE to perform an exhaustive search of the duplication tree space; this is a guaranteed but nonefficient way to find the optimal duplication trees. We need to increase the speed of this reconstruction procedure, so as to be able to tackle larger loci containing higher numbers of repeats. A refined solution would be to include the duplication tree constraint into the search procedure and to use optimal (e.g., branch-and-bound) or heuristic techniques to explore the restricted solution space.

We also tried to reconstruct the duplication history of the 11 repeats of the UbiA polyubiquitin locus (Graham, Jones, and Candidio 1989) in *Caenorhabditis elegans*. Unfortunately, the five most parsimonious duplication trees (184 parsimony steps) found by DTEXPLORE were different from the unique most parsimonious phylogeny found by DNAPENNY (178 parsimony steps). This indicates that our model of evolution by tandem duplication needs to be refined in some cases by introducing other mechanisms such as deletions or gene conversions, for example.

Acknowledgments

We thank David Bryant, Andy McKenzie, Eric Rivals, two anonymous referees, and Mike Hendy for their helpful comments on the preliminary versions of the manuscript.

LITERATURE CITED

AHO, A., J. HOPCROFT, and J. ULLMAN. 1974. The design and analysis of computer algorithms. Addison Wesley, Reading, Mass.

ALBERTS, B., D. BRAY, J. LEWIS, M. RAFF, K. KOBERTS, and J. WATSON. 1995. Molecular biology of the cell. 3rd edition. Garland Publishing Inc., New York.

ATALLAH, M., ed. 1999. Algorithms and theory of computation handbook. CRC Press LLC, Boca Raton, Fla.

BEUSON, G., and L. DONG. 1999. Reconstructing the duplication history of a tandem repeat. Pp. 44–53 *in* Proceedings of the Intelligent Systems in Molecular Biology ISMB'99.

CAVALLI-SFORZA, L., and A. EDWARDS. 1967. Phylogenetic analysis: models and estimation procedure. Evolution **21**: 550–570.

COLLINS, F., and S. WEISSMAN. 1984. The molecular genetics of human hemoglobin. Prog. Nucleic Acids Res. Mol. Biol. **31**:315–462.

CORBETT, S., I. TOMLINSON, E. SONNHAMMER, D. BUCK, and G. WINTER. 1997. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, ''minor'' D segments or D-D recombination. J. Mol. Biol. **270**:587–597.

DARIAVACH, P., G. LEFRANC, and M. LEFRANC. 1987. Human immunoglobulin C lambda 6 gene encodes the Kern–Oz-lambda chain and C lambda 4 and C lambda 5 are pseudogenes. Proc. Natl. Acad. Sci. USA **84**:9074–9078.

EL-MABROUK, N. 2000. Duplication, rearrangement and reconciliation. Pp. 537–550 *in* D. SANKOFF and J. H. NADEAU, eds. Comparative genomics. Kluwer Academics Publishers, Dordrecht.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**: 401–410.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

———. 1989. PHYLIP—Phylogeny Inference Package. Cladistics **5**:164–166.

FELSENSTEIN, J., and G. CHURCHILL. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

FITCH, W. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Syst. Zool. **20**: 406–416.

GASCUEL, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

———. 2000. Evidence of a relationship between algorithmic scheme and shape of inferred trees. Pp. 157–168 *in* W. Gaul,

O. Opitz, and M. Schader, eds. Data analysis. Scientific modeling and practical applications, Springer-Verlag, Berlin.

GHANEM, N., C. BURESI, J. MOISAN, M. BENSMANA, P. CHUCHANA, S. HUCK, G. LEFRANC, and M. LEFRANC. 1989. Deletion, insertion, and restriction site polymorphism of the T-cell receptor gamma variable locus in French, Lebanese, Tunisian, and Black African populations. Immunogenetics **30**:350–360.

GRAHAM, R., D. JONES, and E. CANDIDIO. 1989. UbiA, the major polyubiquitin locus in *Caenorhabditis elegans*, has unusual structural features and is constitutively expressed. Mol. Cell. Biol. **9**:268–277.

GUMUCIO, D., K. WIEBAUER, R. CALDWELL, L. SAMUELSON, and M. MEISLER. 1988. Concerted evolution of human amylase genes. Mol. Cell. Biol. **8**:1197–1205.

HENDY, M., and D. PENNY. 1982. Branch and bound algorithms to determine minimal evolutionary trees. Math. Biosci. **59**:277–290.

HIETER, P., G. HOLLIS, S. KORSMEYER, T. WALDMANN, and P. LEDER. 1981. Clustered arrangement of immunoglobulin lambda constant region genes in man. Nature **294**(5841): 536–540.

HONJO, T., and F. ALT, eds. 1995. Immunoglobulin genes. Academic Press, London.

HORDVIK, I., J. THEVARAJAN, I. SAMDAL, N. BASTANI, and B. KROSSOY. 1999. Molecular cloning and phylogenetic analysis of the Atlantic salmon immunoglobulin D gene. Scand. J. Immunol. **2**(50):202–210.

JEFFREYS, A., and S. HARRIS. 1981. Processes of gene duplication. Nature **296**:9–10.

LARGET, B., and D. SIMON. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. **16**:750–759.

LEFRANC, M., A. FORSTER, R. BAER, M. STINSON, and T. RABBITTS. 1986. Diversity and rearrangement of the human T-cell rearranging genes: nine germ-line variable genes belonging to two subgroups. Cell **45**:237–246.

LEFRANC, M., A. FORSTER, and T. RABBITTS. 1986. Rearrangement of two distinct T-cell gamma-chain-variable-region genes in human DNA. Nature **319**:420–422.

LI, W. 1997. Molecular evolution. Sinauer Inc., Sunderland, Mass.

OHNO, S. 1970. Evolution by gene duplication. Springer-Verlag, New York.

RUDDLE, F., J. BARTELS, K. BENTLEY, C. KAPPEN, M. MURTHA, and J. PENDLETON. 1994. Evolution of Hox genes. Annu. Rev. Genet. **28**:423–442.

RUIZ, M., V. GIUDICELLI, C. GINESTOUX et al. (12 co-authors). 2000. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. **28**:219–221.

SANKOFF, D., and M. BLANCHETTE. 1999. Phylogenetic invariants for genome rearrangements. J. Comput. Biol. **3**: 431–445.

SHEN, S., J. SLIGHTOM, and O. SMITHIES. 1981. A history of the human fetal globin gene duplication. Cell **26**:191–203.

SMITH, G. 1976. Evolution of repeated DNA sequences by unequal crossover. Science **191**:528–535.

SUNDERAM, V. 1990. PVM: a framework for parallel distributed computing. Concurrency: Pract. Experience **2**:315–339.

SWOFFORD, D., P. OLSEN, P. WADDELL, and D. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MALLE, eds. Molecular systematics. Sinauer Associates, Sunderland, Mass.

THOMPSON, J., D. HIGGINS, and T. GIBSON. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**(22):4673–4680.

VASICEK, T., and P. LEDER. 1990. Structure and expression of the human immunoglobulin lambda genes. J. Exp. Med. **172**:609–620.

VIJVERBERG, K., and K. BACHMANN. 1999. Molecular evolution of a tandemly repeated trnF(GAA) gene in the chloroplast genomes of Microseris (Asteraceae) and the use of structural mutations in phylogenetic analyses. Mol. Biol. Evol. **16**:1329–1340.

MIKE HENDY, reviewing editor

Accepted November 13, 2001