

LETTER

On Counting Tandem Duplication Trees

Jialiang Yang and Louxin Zhang

Department of Mathematics, National University of Singapore, Singapore

Introduction

Large genomes are full of repeated DNA sequences. It was estimated that over half of the human DNA consists of repeated sequences (Baltimore 2001; Eichler 2001; Leem et al. 2002). Tandem duplication is one of the important evolutionary mechanisms for producing repeated DNA sequences, in which the copies that may or may not contain genes are adjacent along the genome. Fitch (1977) first observed that tandem duplication histories are much more constrained than speciation histories and proposed to model them assuming that unequal crossover is the biological mechanism from which they originate. The corresponding trees are now called *tandem duplication trees*, the term *tandem* sometimes being omitted for the sake of conciseness. With more and more genomic sequences becoming known, inferring tandem duplication history has again redrawn researchers' attention (Benson and Dong 1999; Tang, Waterman, and Yooseph 2002; Elemento, Gascuel, and Lefranc 2002; Zhang et al. 2003).

The aim of this article is, first, to present a simple recurrence formula for the number of rooted duplication trees based on the recurrence formula proved in Gascuel et al. (2003). We also give a simple non-counting proof of the fact that the number of rooted duplication trees for n segments is exactly twice the number of unrooted duplication trees for n segments. Notice that this fact was proved based on a counting argument in Gascuel et al. (2003).

Duplication Tree Model

Assume n sequence segments $\{1, 2, \dots, n\}$ were formed from a locus by tandem duplication. Then, assume that the locus had grown from a single copy through a series of tandem duplications. Each duplication replaced a stretch of DNA sequences 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 rooted *duplication tree* \mathcal{M} for tandemly repeated segments $\{1, 2, \dots, n\}$ is a rooted binary tree that contains blocks as shown in figure 1. A node in \mathcal{M} represents a repeat. Obviously, the root represents the original copy at the locus and leaves the given segments.

A *block* in \mathcal{M} represents a duplication event. Each non-leaf node appears in a unique block; no node is an

ancestor of another in a block. If the block corresponds to a k -duplication, it has k nodes u_1, u_2, \dots, u_k listed from left to right. Assume $lc(u_i)$ and $rc(u_i)$ are the left and right children of u_i , $1 \leq i \leq k$. Then, in the model \mathcal{M} ,

$$lc(u_1), lc(u_2), \dots, lc(u_k), rc(u_1), rc(u_2), \dots, rc(u_k)$$

are placed from left to right. Hence, for any i and j , $1 \leq i \leq j \leq k$, the directed edges $(u_i, rc(u_i))$ and $(u_j, lc(u_j))$ cross each other. But no other edges cross in the model. For simplicity, we will only draw blocks corresponding to multi-duplication events that contain more than one internal node.

The leaves representing given segments are placed from left to right in the order of the segments appearing on the chromosome. Here we assume that such an order is the increasing order.

Counting Rooted Duplication Trees

An r -duplication event in a rooted duplication tree is *visible* if none of the $2r$ copied segments produced by the event have been duplicated subsequently. It is easy to see that the children of the nodes contained in a visible duplication block are leaves. For example, there are five visible duplications in the rooted duplication tree in figure 1: $[q]$, $[n]$, $[j, k]$, $[o]$, $[p]$. For a visible r -duplication event, if there are i given segments remaining to the right of the $2r$ copies produced by the event, we refer it as a (i, r) -duplication event. In figure 1, the visible 2-duplication $[j, k]$ is a $(7, 2)$ -duplication.

We use r_n to denote the number of all the rooted duplication trees for n segments. For $n \geq 2$ and $0 \leq i \leq n - 2$, let $p(n, i)$ denote the number of all the rooted duplication trees for n segments in which the leftmost visible duplication is an (i, r) -duplication for some r , $1 \leq r \leq (n - i)/2$. Then,

Lemma 1 (Gascuel et al. 2003) For any $n \geq 2$ and $1 \leq i \leq n - 4$,

$$r_n = \sum_{i=0}^{n-2} p(n, i), \quad n \geq 2, \quad (1)$$

$$p(n, i) = p(n - 1, i + 1) + p(n, i - 1). \quad (2)$$

By applying Lemma 1, we obtain the following simple recurrence formula for r_n .

THEOREM 1 For any $n \geq 2$,

$$r_n = \begin{cases} 1 & \text{if } n = 2, \\ \sum_{k=1}^{\lfloor (n+1)/3 \rfloor} (-1)^{k+1} \binom{n+1-2k}{k} r_{n-k} & \text{if } n \geq 3. \end{cases}$$

Proof. Obviously, $r_2 = 1$. Now, we only consider the case $n \geq 3$. By definition,

Key words: molecular phylogeny, tandem duplication history, duplication tree model.

E-mail: matzlx@nus.edu.sg.

Mol. Biol. Evol. 21(6):1160–1163. 2004

DOI:10.1093/molbev/msh115

Advance Access publication March 10, 2004

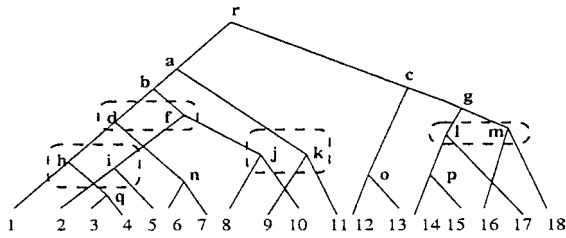


FIG. 1.—A rooted duplication tree \mathcal{M} . Multi-duplication blocks are $[d, f]$, $[h, i]$, $[j, k]$, and $[l, m]$.

$$p(n, n - 2) = p(n, n - 3) = p(n, n - 4) = r_{n-1}, \quad n \geq 3 \tag{3}$$

as demonstrated in Gascuel et al. (2003). This can be generalized into

$$p(n, n - k) = \sum_{i=1}^{\lfloor (k+1)/3 \rfloor} (-1)^{i+1} \binom{k - 2i}{i - 1} r_{n-i} \tag{4}$$

for any k from 2 to n .

Now, we show equation (4) by induction on k . When $k = 2, 3$, equation (4) becomes equation (3) and hence is true. Assume it is true for $k \leq j$. For $k = j + 1 > 3$, by equation (2),

$$\begin{aligned} p(n, n - (j + 1)) &= p(n, n - j) - p(n - 1, n - j + 1) \\ &= p(n, n - j) - p(n - 1, (n - 1) - (j - 2)) \\ &= \sum_{i=1}^{\lfloor (j+1)/3 \rfloor} (-1)^{i+1} \binom{j - 2i}{i - 1} r_{n-i} \\ &\quad - \sum_{i=1}^{\lfloor (j-1)/3 \rfloor} (-1)^{i+1} \binom{j - 2 - 2i}{i - 1} r_{n-1-i} \\ &= \sum_{i=1}^{\lfloor (j+1)/3 \rfloor} (-1)^{i+1} \binom{j - 2i}{i - 1} r_{n-i} \\ &\quad + \sum_{i'=2}^{\lfloor (j+2)/3 \rfloor} (-1)^{i'+1} \binom{j - 2i'}{i' - 2} r_{n-i'} \\ &= r_{n-1} + \sum_{i=2}^{\lfloor (j+2)/3 \rfloor} (-1)^{i+1} \left(\binom{j - 2i}{i - 1} + \binom{j - 2i}{i - 2} \right) r_{n-i} \\ &= \sum_{i=1}^{\lfloor (j+2)/3 \rfloor} (-1)^{i+1} \binom{j + 1 - 2i}{i - 1} r_{n-i} \end{aligned}$$

where we assume $\binom{a-1}{a} = 0$ to derive the 5th equality and use the formula $\binom{b}{a} + \binom{b}{a-1} = \binom{b+1}{a}$ for two integers a, b such that $1 \leq a, a - 1 \leq b$. This concludes the induction proof and hence equation (4) holds for any k from 2 to n .

Now, by equation (1),

$$\begin{aligned} r_n &= \sum_{j=2}^n p(n, n - j) \\ &= \sum_{j=2}^n \sum_{i=1}^{\lfloor (j+1)/3 \rfloor} (-1)^{i+1} \binom{j - 2i}{i - 1} r_{n-i} \\ &= \sum_{k=1}^{\lfloor (n+1)/3 \rfloor} (-1)^{k+1} \left(\sum_{3k-1}^n \binom{j - 2k}{k - 1} \right) r_{n-k} \\ &= \sum_{k=1}^{\lfloor (n+1)/3 \rfloor} (-1)^{k+1} \binom{n + 1 - 2k}{k} r_{n-k} \end{aligned}$$

where we use the formula $\sum_{i=a}^b \binom{i}{a} = \binom{b+1}{a+1}$ for two integers $0 \leq a \leq b$. This finishes the proof.

The recurrence formula in Theorem 1 allows us to find a closed formula for computing r_n . Let $X = (r_n, r_{n-1}, \dots, r_3, r_2)^T$. Then, by the recurrence formula,

$$AX = (0, 0, \dots, 0, 1)^T$$

where $A = (a_{ij})_{(n-1) \times (n-1)}$ is defined as

$$a_{ij} = \begin{cases} (-1)^{j-i} \binom{n+2+i-2j}{j-i} & \text{if } i \leq j \leq (n+2+2i)/3; \\ 0 & \text{otherwise.} \end{cases}$$

Since A is an upper triangular matrix having 1s along the diagonal, its determinant is 1. Hence, the fact that only the last entry '1' is non-zero in the right-hand vector implies that r_n is the co-factor of the row $n - 1$ and column 1 in A . For example, we have

$$r_6 = \begin{vmatrix} -\binom{5}{1} & \binom{3}{2} & 0 & 0 \\ 1 & -\binom{4}{1} & \binom{2}{2} & 0 \\ 0 & 1 & -\binom{3}{1} & 0 \\ 0 & 0 & 1 & -\binom{2}{1} \end{vmatrix}$$

Unrooted Duplication Trees

An unrooted duplication tree is a tree derived from a rooted duplication tree by removal of the root. Let \mathcal{U} be an unrooted duplication tree for n segments $\{1, 2, \dots, n\}$. By definition, at least one rooted duplication tree can be obtained from \mathcal{U} by rooting it at some edge in the path from 1 to n . We say that \mathcal{U} can be rooted at an edge e if a rooted duplication tree can be formed by rooting it at e . Let S_i denote the set of unrooted duplication trees that can be rooted at exactly i edges, $i \geq 1$. Then, the following facts are true:

- (1) If the unrooted duplication tree \mathcal{U} can be rooted at two edges e and e' , then it can be rooted at any edge between e and e' in the path from segment 1 to segment n .
- (2) Let $\mathcal{U} \in S_k$ for some $k \geq 3$. Assume the edges in the path from 1 to n in \mathcal{U} are

$$e_1, e_2, \dots, e_m, \quad m \leq n$$

where $e_i = (u_{i-1}, u_i)$, $u_0 = 1$, and $u_m = n$. If \mathcal{U} can only be rooted at the edges e_j ($i \leq j \leq i'$) where $i' - i + 1 = k$. Then, u_{i-1} must be contained in a multi-duplication block and so is $u_{i'}$.

Assume $\mathcal{U} \in S_k$, $k \geq 3$ and i and i' are as in (2). Let T_j denote the subtree of \mathcal{U} rooted at a child of u_j that is off the path from 1 to n . We also let $\mathcal{U}_{j(j+1)}$ denote the unrooted duplication tree obtained from \mathcal{U} by interchanging T_j and T_{j+1} as illustrated in figure 2. Then, we attain $i' - i - 2$ unrooted trees $\mathcal{U}_{(i+1)(i+2)}$, $\mathcal{U}_{(i+2)(i+3)}$, \dots , $\mathcal{U}_{(i'-2)(i'-1)}$ from \mathcal{U} . It is easy to see that, for each j from $i + 1$ to $i' - 2$, $\mathcal{U}_{j(j+1)}$ can be rooted uniquely at the edge e_{j+1} .

Conversely, if \mathcal{U} can be rooted uniquely at some edge $e_{i+1} = (u_i, u_{i+1})$ in the path from 1 to n , then, u_i and u_{i+1} must be contained in a double duplication block (and hence the right subtree of u_i and the left subtree of u_{i+1} are swapped). Thus, $\mathcal{U}_{i(i+1)}$ obtained by interchanging T_i and T_{i+1} can be rooted at three edges e_i, e_{i+1}, e_{i+2} . This implies that $\mathcal{U}_{i(i+1)}$ is an unrooted duplication tree that can be rooted in at least three ways.

Therefore, the mapping from \mathcal{U} to $C_{1\mathcal{U}} = \{\mathcal{U}_{j(j+1)} \mid i + 1 \leq j \leq i' - 2\}$ is one-to-one from unrooted duplication trees $\mathcal{U} \in S_k$ ($k \geq 3$) to the $(k - 2)$ -subsets of S_1 . Recall that r_n denotes the number of rooted duplication trees for n segments. Letting u_n be the number of unrooted duplication trees, we obtain

$$\begin{aligned} r_n &= \sum_{j=1}^{n-2} j|S_j| \\ &= \sum_{j=3}^{n-2} j|S_j| + |S_1| + 2|S_2| \\ &= \sum_{j=3}^{n-2} \sum_{\mathcal{U} \in S_j} j + \sum_{j=3}^{n-2} \sum_{\mathcal{U} \in S_j} |C_{1\mathcal{U}}| + 2|S_2| \\ &= \sum_{j=3}^{n-2} \sum_{\mathcal{U} \in S_j} (j + |C_{1\mathcal{U}}|) + 2|S_2| \\ &= \sum_{j=3}^{n-2} \sum_{\mathcal{U} \in S_j} (j + j - 2) + 2|S_2| \\ &= 2 \sum_{j=3}^{n-2} \sum_{\mathcal{U} \in S_j} (1 + |C_{1\mathcal{U}}|) + 2|S_2| \\ &= 2 \sum_{j=1}^{n-2} |S_j| \\ &= 2u_n \end{aligned}$$

where we use the fact $|C_{1\mathcal{U}}| = j - 2$ for $\mathcal{U} \in S_j, j \geq 3$.

In conclusion, the following two facts hold for duplication trees:

1. The number r_n of rooted duplication trees for n segments is twice the number of unrooted duplication trees (Gascuel et al. 2003); it satisfies the following recurrence relation:

$$r_n = \sum_{k=1}^{\lfloor (n+1)/3 \rfloor} (-1)^{k+1} \binom{n+1-2k}{k} r_{n-k}.$$

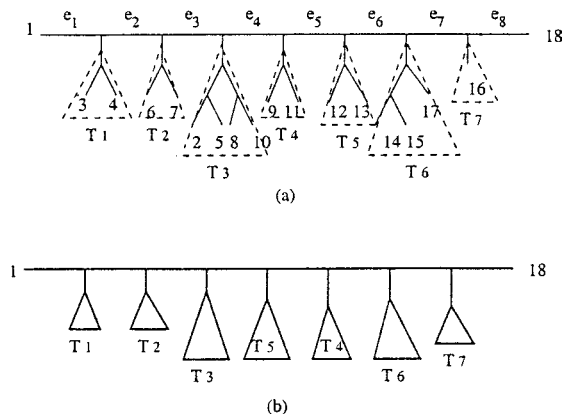


FIG. 2.—(a) An unrooted duplication tree \mathcal{U} . It can be rooted at 5 edges e_3, e_4, e_5, e_6, e_7 . The rooted duplication tree derived from \mathcal{U} by rooting it at e_5 is given in figure 1. (b) An unrooted duplication tree \mathcal{U}_{45} obtained from \mathcal{U} by interchanging subtrees T_4 and T_5 . \mathcal{U}_{45} can only be rooted at e_5 .

2. A rooted duplication tree for segments $\{1, 2, \dots, n\}$ is *ordered* if it contains only 1-duplications. Such a duplication tree is just a rooted binary tree having its leaves listed from left to right in the increasing order. The number of rooted, ordered duplication trees for n segments is the same as the number of unrooted, ordered duplication trees for $n + 1$ segments (Elemento and Gascuel 2003); it is $\binom{2(n-1)}{n-1}/n$ (Zhang, Ma, and Wang 2003; Elemento and Gascuel 2003).

Acknowledgments

L. Zhang thanks Mike Steel for drawing attention to the counting problem studied here, and for further discussions. Thanks, too, to Olivier Gascuel for useful comments on the first version of this paper. The work is partially supported by Singapore BioMedical Research Council Research grant BMRC01/1/21/19/140.

Literature Cited

Baltimore, D. 2001. Our genome unveiled. *Nature* **409**:814–816.
 Eichler, E. E. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**:661–669.
 Elemento, O., and O. Gascuel. 2003. An exact and polynomial distance-based algorithm to reconstruct single copy tandem duplication trees. Pp. 96–108 in *Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching (CPM'03)*, Morelia, Mexico. Lecture Notes in Computer Science, vol. 2676, Springer-Verlag.
 Elemento, O., O. Gascuel, and M.-P. Lefranc. 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.
 Gascuel, O., M. D. Hendy, A. Jean-Marie, and R. McLachlan. 2003. The combinatorics of tandem duplication trees. *Syst. Biol.* **52**:110–118.

- Leem, S.-H., J. A. Londoño-Vallejo, J.-H. Kim et al. 2002. The human telomerase gene: complete genomic sequence and analysis of tandem repeat polymorphisms in intronic regions. *Oncogene* **21**:769–777.
- Tang, M., M. Waterman, and S. Yooseph. 2002. Zinc finger gene clusters and tandem gene duplication. *J. Comput. Biol.* **9**: 429–446.
- Zhang, L., B. Ma, L. Wang, and Y. Xu. 2003. Greedy method for inferring tandem duplication history. *Bioinformatics* **19**:1497–1504.

Peter Lockhart, Associate Editor

Accepted February 25, 2004