Discrete Applied Mathematics 194 (2015) 1–22

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

The combinatorics of tandem duplication

L. Penso-Dolfin [a], T. Wu [a], C.D. Greenman [a,b,*]

[a] School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK
[b] The Genome Analysis Center, Norwich Research Park, Norwich, NR4 7UH, UK

ARTICLE INFO

Article history:
Received 3 February 2014
Received in revised form 1 April 2015
Accepted 6 May 2015
Available online 6 June 2015

Keywords:
Combinatorics
Tandem duplication
Posets
Rearrangements
Evolution

ABSTRACT

Tandem duplication is a rearrangement process whereby a segment of DNA is replicated and proximally inserted. A sequence of these events is termed an evolution. Many different configurations can arise from such evolutions, generating some interesting combinatorial properties. Firstly, new DNA connections arising in an evolution can be algebraically represented with a word producing automaton. The number of words arising from n tandem duplications can then be recursively derived. Secondly, many distinct evolutions result in the same sequence of words. With the aid of a bi-colored 2d-tree, a Hasse diagram corresponding to a partially ordered set is constructed, for which the number of linear extensions equates to the number of evolutions generating a given word sequence. Thirdly, we implement some subtree prune and graft operations on this structure to show that the total number of possible evolutions arising from n tandem duplications is $\prod_{k=1}^{n}(4^k-(2k+1))$. The space of structures arising from tandem duplication thus grows at a super-exponential rate with leading order term $\mathcal{O}(4^{\frac{1}{2}n^2})$.

1. Introduction

Tandem duplications occur when a region of DNA is duplicated and inserted adjacent to the original segment, such as portrayed in Fig. 1A.

This biological process has long been known to be implicated in the formation of gene clusters [24,23] and more recently has been implicated in the formation of amplicons in cancer [21,25,26,32]. In both cases Darwinian selection may be acting to increase the number of copies of a target gene. In addition to the biological study of this process, there are a range of algorithmic and mathematical questions that are also of interest. These include identification and alignments of tandem duplications in data [3,2,5,4,20] and the construction of phylogenies describing their evolution [6,9,8]. In [9] this was done in a quite general context, where duplications and losses across multiple genomes were considered. In [8] tree operations were introduced that allowed a full exploration of tandem duplication trees; phylogenetic structures that describe tandem duplication evolution. A survey of algorithmic approaches can be found in [27]. The combinatorial nature of these rearrangement operations leads to some interesting combinatorics. The number of rooted and unrooted tandem duplication trees that arise from the tandem duplication of a loci of interest are explored in [13,30]. The space of permutations arising from a tandem duplication-loss model is characterized in [11,10].

* Corresponding author at: School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK. Tel.: +44 01603 592300; fax: +44 01603 593345.
E-mail address: C.Greenman@uea.ac.uk (C.D. Greenman).

http://dx.doi.org/10.1016/j.dam.2015.05.014
0166-218X/

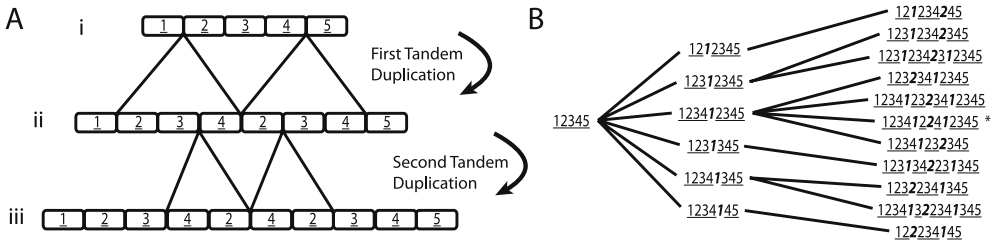

Fig. 1. A Tandem Duplication Process. (A) Three structures (i)–(iii) arising from two tandem duplications on a reference of five regions; 1, 2, 3, 4, 5. (B) Eleven possible evolutions with two tandem duplications. The example in A is highlighted by *. Underlined numbers are segments. Bold italicized numbers ***n*** indicate connections between segments formed in the *n*th tandem duplication.

These methods make a range of assumptions regarding the information that is available and the process that takes place. In particular, there are two issues that relate to the problem we consider.

Firstly, the genomic sequence information that is analyzed differs. In [9], the signed gene orders of several genomes are compared and explanatory phylogenetic evolutions derived. In [13,30] a single copy of a loci is analyzed, and all the possible different evolutions that can take place counted. In the problem we consider, we also start with a single region of known (reference) sequence, and investigate the number of different possible evolutions that arise. Our approach differs from [13,30] with regard to the second issue.

This relates to the assumption that breakpoints can be reused. A breakpoint in this context can mean the gap between two contiguous loci, such as a pair of genes in a gene cluster, which can cover a wide region and be implicated in more than one duplication event with reasonable probability, or it can mean the precise end points of the duplicated region, which are less likely to be implicated on more than one occasion (for larger scale tandem duplications at least). Modern sequencing (paired-end) data can resolve breakpoints to the base pair level and reveal tandem duplications to great precision, such as with cancer data [21]. In such cases, when a tandem duplication occurs, two breakpoints are implicated in a presumably random process. The chance that precisely the same nucleotide positions are subsequently implicated in another TD is likely to be small and assuming unique breakpoint use is reasonable in these circumstances. The questions considered in this work are restricted to the case of unique breakpoint use. We now outline the main problem we consider.

In Fig. 1A we start with five contiguous segments, labeled 1, 2, 3, 4 and 5. This is the original configuration and is termed the *reference*. The four reference positions between the segments represent *breakpoint* sites that demarcate where duplicated regions of tandem duplications may start or finish. We then have an initial tandem duplication, copying region 234 and inserting a new copy next to the first, to give sequence 12341**1**2345. Here we have used (not underlined, bold symbol) **1** to indicate our first *connection* between two segments not seen in the reference; the right side of segment 4 is connected to the left side of segment 2, as seen in Fig. 1Aii. Note also that the left hand end of the duplicated region 234 implicates the breakpoint between segments 1 and 2, the right hand end implicates the breakpoint between segments 4 and 5. We have thus used two of the four breakpoints available. Next we have the second tandem duplication, copying region 42 to finally give 12341**2**2**4**1**2**2345. We now have another connection, labeled **2**, between the right side of segment 2 and the left of segment 4, as seen in Fig. 1Aiii. Note that we now have two copies of the connection labeled **1**, which was also duplicated. The left hand end of the duplicated region represented by subword 4**1**2 implicates the reference position between 3 and 4, the right hand end implicates that between 2 and 3. We have thus implicated all four breakpoints between the five reference segments exactly once; unique breakpoint use.

In Fig. 1B we see all 11 different ways that two tandem duplications can act on five segments with unique breakpoint reuse. Note that N tandem duplications will implicate $2N$ breakpoints and so $2N + 1$ segments. We are then primarily interested in solving the following problem.

Problem 1.1. Count the number of different ways that an initial string of $2N + 1$ segments can evolve under N tandem duplications, using each of the N breakpoints once.

To solve this involves a better understanding of the connections we have labeled. If we ignore all the labels representing segments, we get simpler sequences to consider. For example, the sequence [12345 → 12341**1**2345 → 12341**2**2**4**1**2**2345] becomes the simpler sequence [$\epsilon \to$ **1** → **121**], where ϵ denotes the empty word. Although this representation is simpler, it is not unique—five of the eleven cases in Fig. 1B contain this sequence of connections. However, we will need to consider these sequences in more detail to solve Problem 1.1.

We then attack the problem as follows. Firstly, we formalize the representations by segments and connections given above. We then explore the size of the space of word sequences involving connection symbols. Each such word sequence will be seen to correspond to many different structures formed by tandem duplications. Thus, thirdly, we consider how to count the distinct cases that all correspond to a single sequence of words containing connection symbols. This involves counting linear extensions of a suitable partially ordered set (poset). Fourthly, we combine these two pieces of information and provide an explicit formula to answer Problem 1.1. Concluding remarks complete the paper.