A Comparison of Three Heuristic Methods for Solving the Parsing Problem for Tandem Repeats

Conference Paper *in* Lecture Notes in Computer Science · August 2012

DOI: 10.1007/978-3-642-31927-3_4

4 authors, including:

Atheer Matroud
The American University of Iraq, Sulaimani
6 PUBLICATIONS **14** CITATIONS
SEE PROFILE

David Bryant
University of Otago
150 PUBLICATIONS **12,204** CITATIONS
SEE PROFILE

Michael D Hendy
University of Otago
144 PUBLICATIONS **5,988** CITATIONS
SEE PROFILE

TANDEM REPEATS PARSING PROBLEM

A. A. MATROUD [1,2,*], M. D. HENDY [3], C. P. TUFFLEY [1], AND D. BRYANT [2,3]

Abstract. In tandem repeats, the placement of the boundaries of the repeating motif is ambiguous when some of the repeating motif extends into the flanking regions. We refer to alternate boundary placements as alternate *parsings* of the tandem repeat. In the case of approximate tandem repeats, the variations in the motif copies allow us to estimate a *Duplication History Tree* (DHT) which describes the possible evolution of the tandem repeat from a single ancestral motif, undergoing processes of motif duplication (and possibly deletion) together with nucleotide substitutions. However different parsings can lead to different DHTs, so we could discriminate among alternate parsings by selecting the parsing which minimises the parsimony score of the resulting DHT. We develop a criterion here which acts as a surrogate for the parsimony score so that the optimal parsing may be selected before the DHT is derived.

1. Introduction

Several studies have proposed different mechanisms for the evolution of tandem repeats Weitzmann et al. (1997), Wells (1996). In the case of approximate tandem

repeats it is possible to infer their duplication history tree (DHT). This is an important step to understand the duplication mechanism that generate them. A number of algorithms to reconstruct DHT have been introduced in the last ten years Bertrand et al. (2008); Lajoie et al. (2007); Rivals (2004); Chauve et al. (2008). A crucial element in the DHT reconstruction process is to identify the repeat pattern boundaries. In the literature, researchers used alignment score to decide on the boundaries of the motif.

Comparing tandem repeats (also called Mapping minisatelites) is the problem of pairwisely aligning and comparing DNA sequences containing tandem repeats. this problem was addressed by Sammeth and Stoye (2006); Berard and Rivals (2003); Behzadi and Steyaert (2003). In order to align two tandem repeats, the repeat boundaries should be determined previously. Different parsing decisions may lead to different conclusions. To date, alignment programs arbitrary decide on the parsing.

Example 1 Consider the following sequence which contains an approximate tandem repeat with periodicity 4:

(1) GACCACGAACGTACGAACGTATTA.

There are 4 possible parsings. For each parsing we define a *mode* motif to be a sequence where the $i-$th nucleotide is a most common nucleotide at the $i-$th site

among the segments. If we set the boundary after GACC we obtain

GACC|ACGA|ACGT|ACGA|ACGT|ATTA,

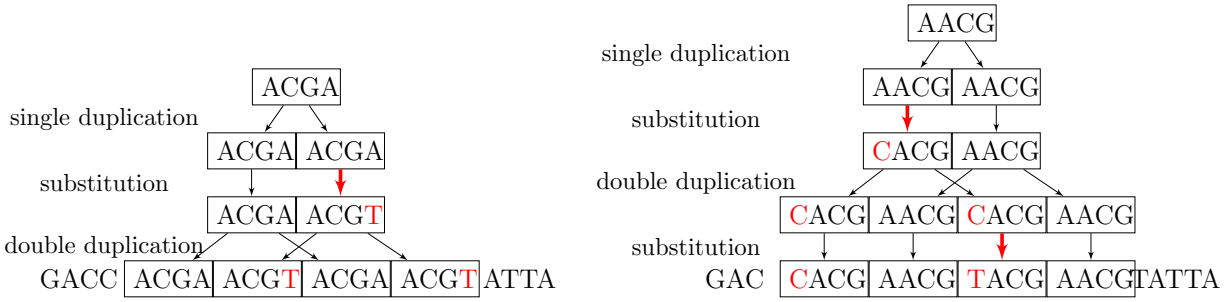
with mode motif ACGA. If we shift the frame one nucleotide to the left we obtain

GAC|CACG|AACG|TACG|AACG|TATTA,

with mode motif AACG. In Figure 1 we see minimal DHTs for both these parsings. Hence, having a criterion that biologically sounds to suggest tandem repeat pattern boundaries is crucial in the analysis process. We refer to the problem of selecting a preferred boundary as the **parsing problem** of tandem repeat pattern. In general there can be n possible parsings, where n is the motif length.

The flanking region contains repeat copies that contains high number of mutations. This flanking region misleads in deciding where is the start and the end of the repeated region. In case a tandem repeats contains n copies with the first and the last copies are in the flanking region (we are able to observe only a part of the pattern because of mutations). If these parts of the patterns happen to be the starting part of the last copy and the ending part of the first copy then the alignment score will consider $n - 1$ copies starting at middle of the flanking region (the first copy) and ending in the middle of the flanking region (the last copy).

In Benson and Dong (1999), the authors suggest a method to select a possible boundaries base on his model of duplication which consider dynamic boundaries



(a) (b)

FIGURE 1. The DHTs inferred from the two parsings of Example 1. The parsing (a) has a DHT with two duplications and a single substitution, whereas parsing (b) the DHT has two duplications and two substitutions. (In both cases we see these number of events is minimal for that parsing.) We use the parsimony principle to prefer parsing (a) over parsing (b) as its DHT requires fewer mutational events.

(duplications may occur on different boundaries). In this paper, we present a new criterion to select the pattern boundaries base on the assumption that the boundaries are fixed during the whole duplication process Fitch (1977).

2. TANDEM REPEATS

An *exact tandem repeat* is a string comprising two or more contiguous exact copies of a substring $\mathbf{X}$, called the tandem repeat *motif*.

We obtain an *approximate tandem repeat* by allowing approximate rather than exact copies of the template motif **X**. We will refer to each homologous copy of the motif as a **segment**. When there is variation among the segments it may be possible to infer the duplication tree. If the rate of substitution (per segment) is lower than the rate of duplication, then we should expect some of the segments to be identical. We take a **mode** motif (consensus pattern), as being a segment where at each site, the most common nucleotide at that site among all segments is chosen.

3. Duplication model

Tandem repeats are modeled as a consequence of duplication events interspersed with single nucleotide substitutions, single nucleotide indels and deletions of one or more copies of the motif. A duplication event occurs by introducing an additional copies of the motif from copying of one or more copies of the motif. The number of motifs copied is referred as the size of duplication. In the simplest case these copies remain contiguous and oriented in the same direction in the genome.

Example 1:

Consider the following sequence which contains tandem repeat.

(2) TATGTCATGGTTATGGACATGGTTATGGACACGCTCACGCTTATGGTCAAGGTCACGGTCAATA

which for the parsing displayed, is an approximate tandem repeat with modal motif CATGGT. There are six motif variants, in order as $ababccdef$ where

$$a = \texttt{CATGGT},\ \texttt{b} = \texttt{TATGGA},\ \texttt{c} = \texttt{CACGCT},\ \texttt{d} = \texttt{TATGGT},\ \texttt{e} = \texttt{CAAGGT},\ \texttt{f} = \texttt{CACGGT}.$$

In Figure 2 we see a DHT. On each edge we identify a substitution as a pair $i\theta$, where the substitution $\theta \in \{\alpha, \beta, \gamma\}$ is applied at site $i \in \{1, \cdots, 6\}$. (Here, following Kimura, the substitution types are

$$\alpha = \texttt{A} \leftrightarrow \texttt{G},\ \texttt{C} \leftrightarrow \texttt{T}; \quad \beta = \texttt{A} \leftrightarrow \texttt{T},\ \texttt{G} \leftrightarrow \texttt{C}; \quad \alpha = \texttt{A} \leftrightarrow \texttt{C},\ \texttt{G} \leftrightarrow \texttt{T}.)$$

This tree represents the DHT of Figure 2. Each duplication is identified by the segment to be duplicated enclosed in a rectangle, When the duplicated segment encloses more than one copy of the motif, the descendant motifs alternate as shown. The approximate repeat is fully described by duplication tree T with 5 duplications, the ancestral motif at the root (CATGGT), and the 5 substitutions on the edges of T.

4. Parsing problem

Let $\mathbf{x[i]} = \mathbf{x_i x_{i+1} \ldots x_n x_1 \ldots x_{i-1}}$ be the i^{th} cyclic permutation of motif $\mathbf{x}$, where n is the length of $\mathbf{x}$, and let $\mathbf{S} = \mathbf{s_1} \ldots (\mathbf{x_i x_{i+1} \ldots x_n x_1 \ldots x_{i-1}})^{\mathbf{l_i}} \ldots \mathbf{s_m}$, be a string containing l_i copies of motif $\mathbf{x[i]}$.

We are interested in answering which of the i^{th} cyclic permutation of $\mathbf{x}$ is the best estimation of the ancestral segment parsing.

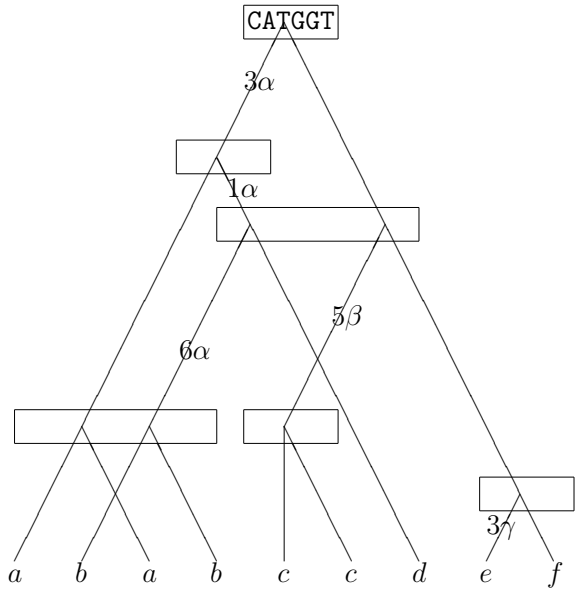


FIGURE 2. *We obtain a maximum parsimony tree of the 9 segments. When we place the root on the edge (arrowed) and order the tips as ababccdef we obtain the duplication tree, descending from the consensus motif a. There are 6 duplications, with the rectangles enclosing the segments being duplicated.*

In example 1, each of the other parsings leads to a duplication tree requiring 6 substitutions and 7 substitutions. Using a parsimonious criterion, we can use a

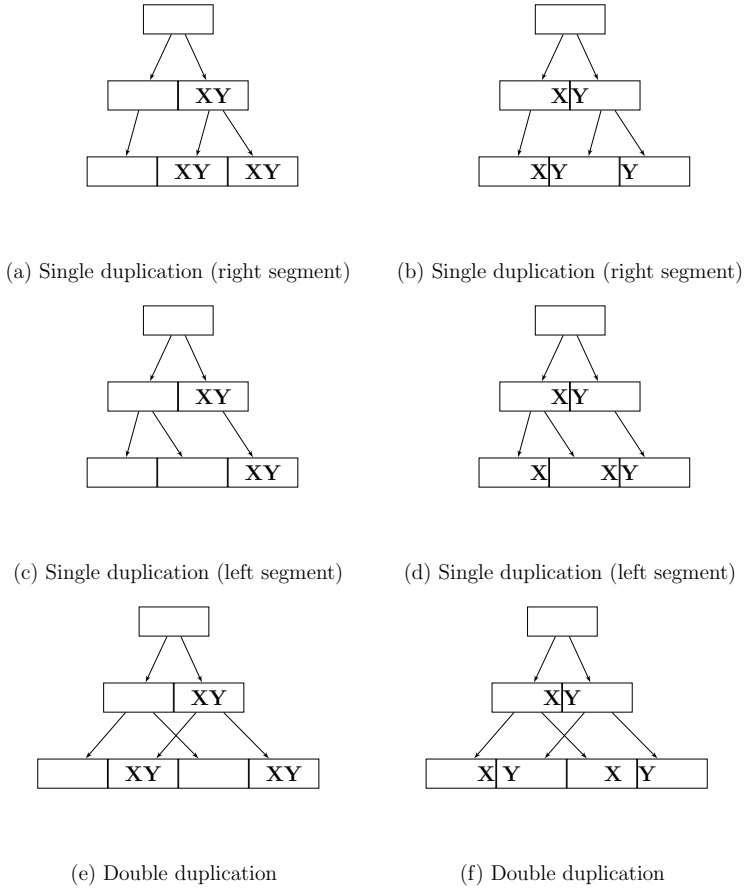
parsimony or information content criterion to discriminate among different parsings of approximate tandem repeats.

5. Heuristic methods to estimate the parsing point

Consider large and long tandem repeats, obtaining the maximum parsimony duplication history tree of the motif copies can be computationally expensive, and some cases the maximum parsimony tree cannot be expressed as a duplication tree [reference to be added]. It may be preferable to avoid these constructions when comparing different parsings.

We describe below an easily determined measures which we can use as surrogates for the comparisons. These rules are intended as a guide to discriminate between alternate parsings. It will often be the case that there remain several alternate equally good parsings, in which case either external information may offer some guide, or an arbitrary choice is required.

Pair of substitutions. In this method we identify all neighboring mutations that are less than the repeat length apart. In Figure 3, all duplication that can occur on two adjacent segments are listed. 3(a,c,e) show the results of different duplication when the pair of substitutions **XY** are not separated by the boundaries. 3(b,d,f) present all duplication event on two adjacent segment and the pair **XY** is separated by the boundaries of the two segments.



(a) Single duplication (right segment)

(b) Single duplication (right segment)

(c) Single duplication (left segment)

(d) Single duplication (left segment)

(e) Double duplication

(f) Double duplication

FIGURE 3. *All possible duplication events on two adjacent segments, in (a) a single duplication on the right segment results in doubling the pair* **XY**, *where the same duplication in (b) does not change the number of occurrences of* **XY**. *In (c) and (d) a single duplication on the right segment results in no changes on the pair* **XY**. *In (d) and (e) a double duplication has the same impact on both case.*

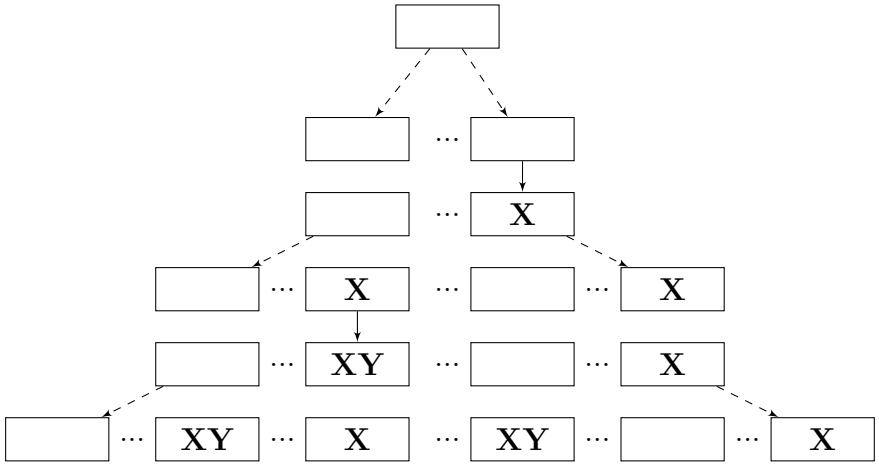


FIGURE 4. A duplication tree, the dash line represent a series of duplication events.

For example,

```
tatGCATGGTtATGGaCATGGTtATGGaCAcGcTCAcGcTtATGGTCAaGGTCAcGGTCAata
```

The neighbor pairs are `t`, `a` at sites $(11, 16), (23, 28)$, and `c`, `c` at $(31, 33), (37, 39)$. This suggests the natural segment boundary should not separate them, leaving the most favoured parsing of segments of length 6 starting at site 5. This gives rise to the same parsing as before.

Repeat variants. Tandem repeat copies slightly differ from each other by point mutation (substitution or deletion of a single character) resulting in *variants* of the ancestral copy. The number of distinct variants depends on the rate of mutation. High rate of mutation results in high number of variants.

The pair of mutations **XY** in Figure 3 is a result of two mutation happened in the past, namely **X** happen first then **Y** occurs sometime later. As a result, at the present time, three variants are obseved; a variant with no mutations; we may also observe another variant with only **X** mutation; and a third variant with **XY** pair of mutations. By setting the boundaries between the pair of mutation **XY**, this might results in observing four variants containing one of the following four mutations {**X**, **YX**, **Y**, variant with no mutation}.

However, base on the model of duplication we consider in Figure 3, we may observe a variant containing only mutation **Y** only if this mutation happened parallel to the mutation **XY** or as a result of another mutation on **X** that change it back to its original character.

length of the variants minimum spanning tree. In this section we consider calculating the minimum spanning tree as a method to distinguish between different parsing. The minimum spanning tree of the variants is calculated for each parsing and the parsing which gives the minimum length is chosen.

6. Results

We have implemented the three methods in section 5 and run them on synthetic tandem repeat. We have tested 100 tandem repeats where the size of the patterns are 40 bp. The number of repeated copies in each tandem repeats are around 100.

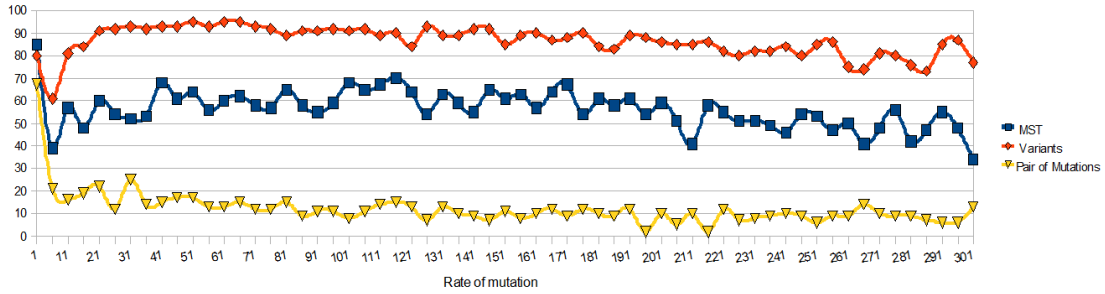


FIGURE 5. A Comparison between the three methods. The x-axis represent the rate of mutation. This graph shows the number of times the true parsing is one of the suggested parsing points.

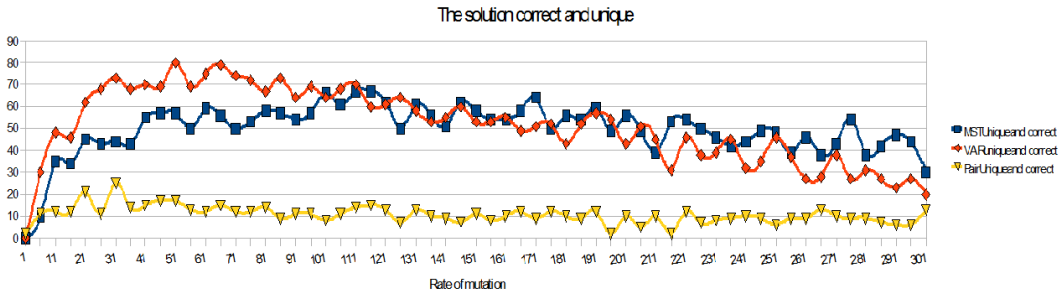


FIGURE 6. A Comparison between the three methods. The x-axis represent the rate of mutation. This graph shows the number of times the suggested parsing point is correct and unique.

A comparison between the three methods above is shown in Figure 5. Each method might suggest more than one parsing point. In Figure 5 the correct parsing point is one of the suggested points.

Figure 6 shows the number of times each method is suggesting only one correct parsing parsing point.

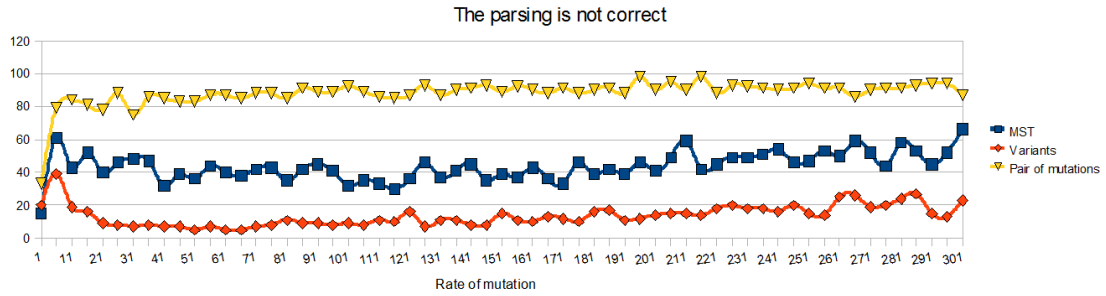


FIGURE 7. A Comparison between the three methods. The x-axis represent the rate of mutation. This graph shows the number of times the suggested parsing point is not the right parsing

7. DISCUSSION

REFERENCES

Behshad Behzadi and Jean-Marc Steyaert. An improved algorithm for generalized comparison of minisatellites. In Ricardo Baeza-Yates, Edgar Chvez, and Maxime Crochemore, editors, *Combinatorial Pattern Matching*, volume 2676 of *Lecture Notes in Computer Science*, pages 32–41. Springer Berlin / Heidelberg, 2003. ISBN 978-3-540-40311-1.

Gary Benson and Lan Dong. Reconstructing the duplication history of a tandem repeat. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 44–53. AAAI Press, 1999. ISBN 1-57735-083-9. URL `http://portal.acm.org/citation.cfm?id=645634.660817`.

S. Berard and E. Rivals. Comparison of minisatelites. *Journal of computational biology*, 10(3-4):357–372, 2003.

Denis Bertrand, Mathieu Lajoie, and Nadia El-Mabrouk. Inferring ancestral gene orders for a family of tandemly arrayed genes. *Journal of Computational Biology*, 15(8):1063–1077, 2008.

Cedric Chauve, Jean-Philippe Doyon, and Nadia El-Mabrouk. Gene family evolution by duplication, speciation, and loss. *Journal of Computational Biology*, 15(8):1043–1062, 2008.

Walter M. Fitch. Phylogenies constrained by the crossover process as illustrated by human hemoglobins and a thirteen-cycle, eleven-amino-acid repeat in human apolipoprotein a-i. *Genetics*, 86(3):623–644, 1977. URL `http://www.genetics.org/content/86/3/623.abstract`.

Mathieu Lajoie, Denis Bertrand, Nadia El-Mabrouk, and Olivier Gascuel. Duplication and inversion history of a tandemly repeated genes family. *Journal of Computational Biology*, 14(4):462–478, 2007.

Eric Rivals. A survey on algorithmic aspects of tandem repeats evolution. *Int. J. Found. Comput. Sci.*, 15(2):225–257, 2004.

Michael Sammeth and Jens Stoye. Comparing tandem repeats with duplications and excisions of variable degree. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3:395–407, 2006. ISSN 1545-5963. doi: http://doi.ieeecomputersociety.org/10.1109/TCBB.2006.46.

M. N. Weitzmann, K. J. Woodford, and K. Usdin. DNA Secondary Structures and the Evolution of Hypervariable Tandem Arrays. *J. Biol. Chem.*, 272(14):9517–9523, 1997. doi: 10.1074/jbc.272.14.9517. URL `http://www.jbc.org/cgi/content/abstract/272/14/9517`.

R. D. Wells. Molecular Basis of Genetic Instability of Triplet Repeats. *J. Biol. Chem.*, 271(6):2875–2878, 1996. doi: 10.1074/jbc.271.6.2875. URL `http://www.jbc.org`.

[1] Institute of Fundamental Sciences, Massey University, Private Bag 11 222, Palmerston North 4442, New Zealand., [2] Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Private Bag 11 222, Palmerston North 4442, New Zealand., [3]Department of Mathematics and Statistics, University of Otago, PO Box 56, Dunedin 9054, New Zealand., * Corresponding author