NTRFINDER: AN ALGORITHM TO FIND NESTED TANDEM REPEATS

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

1. ABSTRACT

1.1. **Motivation:** We introduce the algorithm `NTRFinder` to find a complex repetitive structure in DNA we call a nested tandem repeat (NTR). An NTR is a recurrence of two distinct tandem motifs interspersed with each other. We propose that nested tandem repeats can be used as phylogenetic and population markers.

A major issue is the parsing problem, which is the problem of finding the optimal boundaries of the respective motifs. There has been little discussion about determining the best boundaries of the repeated motifs in tandem repeats. For nested tandem repeats, parsing is a significant issue.

1.2. **Results:** An algorithm for finding Nested Tandem Repeats and a criterion for solving the parsing problem are introduced. We have tested our algorithm on both real and simulated data, and present some real nested tandem repeats of interest. We discuss how one of these may assist in determining the cultivation prehistory of the ancient staple food crop taro (*Colocasia esculenta*).

1.3. **Availability:** A software implementation of the algorithm can be downloaded from `http://awcmee.massey.ac.nz/downloads.htm`.

1.4. **Contact:** a.a.matroud@massey.ac.nz

2. INTRODUCTION

Genomic DNA contains repetition. For example, more than 50% of some genomes consists of repetition. It is suggested that these repeated regions are the result of mutational changes such as duplication, an event that lengthens the DNA by duplicating a certain pattern called a *motif* many times. A tandem repeat is the occurrence of at least two adjacent copies of a motif. Several studies have proposed different mechanisms for the occurrence of tandem repeats [Weitzmann *et al.*, 1997, Wells, 1996].

Recently, researchers have become increasingly interested in tandem repeat regions due to their potential as genetic markers and their importance for the study of evolution. The copy numbers of some tandem repeats usually exhibit polymorphism due to replication slippage and/or unequal crossing-over. This polymorphism is useful in inter-population studies, pedigree analysis, investigating the phylogenic relationships between species, and in evolution studies. In the case of inter-population studies, the number of repeat copies might be the same between individuals within the same population but might differ between individuals of different populations. The case of having different numbers of repeat copies within individuals of the same population is of interest in forensics [Jeffreys *et al.*, 1980].

The biological role of repeat structures is not well understood. Their implication for neurological disorders, and their use to infer evolutionary histories has urged some researchers to develop tools to locate tandem repeats. This has resulted in a number software tools, each of which has its own strengths and limitations. In this paper, we present a new software, NTRFinder, which is designed to find complex repetitive structures in DNA that we call nested tandem repeats. Little attention has been given to finding these structures to date.

2.1. **Background and Definitions.** In this section, definitions of some keywords are presented.

A DNA sequence is a sequence of symbols from the nucleotide alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. We define a DNA *segment* to be a string of contiguous DNA nucleotides and define a *site* to be a component in a segment. For a DNA segment

$$\mathbf{X} = x_1 x_2 \cdots x_n,$$

$x_i \in \Sigma$ is the nucleotide at the $i-$th site.

Copying errors happen in DNA sequences due to different external and internal factors. These changes include substitution, insertion, deletion, duplication, and contraction. We refer to these as *edit operations* as defined below. By giving each operation a weight, one can form a measure of distance for comparing two segments of DNA sequences.

We define the **length** of $\mathbf{X}$ to be the number $l(\mathbf{X}) = n$ of nucleotides in the string.

2.2. **Edit operations.** Let $\mathbf{X}$ be a segment of length n. We define the following edit operations.

- **Substitution:** Let $\theta \in \{\alpha, \beta, \gamma\}$ be a transformation of the Kimura 3ST model [Kimura, 1981]. ($\alpha : \mathtt{A} \leftrightarrow \mathtt{G}, \mathtt{C} \leftrightarrow \mathtt{T}$, $\beta : \mathtt{A} \leftrightarrow \mathtt{C}, \mathtt{G} \leftrightarrow \mathtt{T}$, $\gamma : \mathtt{A} \leftrightarrow \mathtt{T}, \mathtt{C} \leftrightarrow \mathtt{G}$.). The operation $\sigma_i(\theta, \mathbf{X})$ applies the transformation θ to the nucleotide at the $i-$th site of $\mathbf{X}$.
- **k–Deletion:** $\delta_i(k, \mathbf{X})$ removes the k contiguous nucleotides from the i-th to the $(i + k - 1)$-th sites of $\mathbf{X}$ moving the following nucleotides k sites left.
- **k–Insertion:** $\iota_i(z_1 \cdots z_k, \mathbf{X})$ inserts the segment $z_1 \cdots z_k$ of length k at the i-th to $(i + k - 1)$-th sites of $\mathbf{X}$, moving the following nucleotides k steps to the right.
- **Duplication:** $\Delta_i(k, \mathbf{X})$ is the k-insertion $\iota_i(x_{i-k} \cdots x_{i-1}, \mathbf{X})$ of the segment consisting of the k nucleotides of $\mathbf{X}$ immediately preceding site i.

Thus for example if $\mathbf{X} =$ `CGGTATCCAGTAGCT`, then

$$\begin{aligned}
\sigma_7(\alpha, \mathbf{X}) &= \mathtt{CGGTAT}\underline{\mathtt{T}}\mathtt{CAGTAGCT},\\
\delta_{11}(\mathbf{X}) &= \mathtt{CGGTATCCAG}-\mathtt{AGCT},\\
\delta_{11}(3, \mathbf{X}) &= \mathtt{CGGTATCCAG}---\mathtt{CT},\\
\iota_4(GAGTTG, \mathbf{X}) &= \mathtt{CGG}\underline{\mathtt{GAGTTG}}\mathtt{TATCCAGTAGCT},\\
\Delta_{12}(6, \mathbf{X}) &= \mathtt{CGGTATCCAGTA}\underline{\mathtt{CCAGTA}}\mathtt{GCT}.
\end{aligned}$$

2.3. **Hamming Distance.** For two segments $\mathbf{X}, \mathbf{Y}$ of length n, we define the Hamming distance

$$d_H(\mathbf{X}, \mathbf{Y}) = \#\{i \in \{1, \cdots, n\} | x_i \neq y_i\}.$$

This is the number of sites where the corresponding nucleotides differ, or equivalently, the minimum number of substitutions required to convert $\mathbf{X}$ to $\mathbf{Y}$. Clearly d_H is a metric.

2.4. **Edit Distances.** Edit distances count the minimum number of edit operations needed to transform one segment into another. It is important to define which edit operations we use as some have little biological relevance.

Given the set of edit operations defined above, each with equal weight, a random sequence $\mathbf{X}$ of length n can be transformed into any other random sequence $\mathbf{Y}$ of length m in two steps:

$$\iota_1(\mathbf{Y}, \delta_1(n, \mathbf{X})) = \mathbf{Y}$$

(the n-deletion of $\mathbf{X}$ followed by the m-insertion of $\mathbf{Y}$). This is not a useful measure of distance for our purposes.

A distance function which reflects some biological relevance is needed, so, for example, assuming the sequences are homologous, the minimum number of edits from a common ancestral sequence reflects the number of mutational events which may have occurred in their evolution. In this case we do not include insertions, except where they are duplications. However we do not usually know the common ancestor of $\mathbf{X}$ and $\mathbf{Y}$. If $\mathbf{Z} = \mathbf{X} \cdot \mathbf{Y}$ is the concatenation of $\mathbf{X}$ and $\mathbf{Y}$, then we can obtain the daughter sequences in two steps, by deleting $\mathbf{X}$ on one line of descent and $\mathbf{Y}$ on the other. We can avoid this by insisting that the length of the common ancestor cannot be greater than the length of its larger daughter. With this restriction it may be difficult to determine the minimum number of steps required to convert a given sequence to any other sequence.

3. Classification of Tandem Repeats

Many classifications of tandem repeat schemas have been introduced in the computational biology literature. We list some which are commonly used.

- **(Exact) Tandem Repeats:** An *exact tandem repeat* (TR) is a sequence comprising two or more contiguous copies $\mathbf{X}\mathbf{X}\cdots\mathbf{X}$ of identical segments $\mathbf{X}$ (referred to as the *motif*).
- $k-$**approximate Tandem Repeats:** A $k-$*approximate tandem repeat* ($k-$TR) is a sequence comprising two or more contiguous copies $\mathbf{X}_1\mathbf{X}_2\cdots\mathbf{X}_n$ of similar segments, where each individual segment $\mathbf{X}_i$ is distance at most k (by some distance measure) from a template segment $\mathbf{X}$.
- **Multiple Length Tandem Repeats:** A multiple length tandem repeat is an approximate tandem repeat where each repeat copy is of the form $\mathbf{X}\,\mathbf{x}^n$, where $n > 1$ and $\mathbf{X} \neq \mathbf{x}$.

Examples:

- **TR:**
  AGG AGG AGG AGG AGG. The motif is AGG.
- $1-$**TR:**
  AGG AG$\underline{\text{C}}$ A$\underline{\text{T}}$G AGG $\underline{\text{C}}$GG. The motif is AGG.
- **MLTR**:
  GACCTTTGG ACGGT ACGGT ACGGT
  GACCTTTGG ACGGT ACGGT ACGGT.
  The motifs are ACGGT and GACCTTTGG.

3.1. **Nested Tandem Repeats.** In this section we consider a more complex repetitive structure, Nested Tandem Repeats (NTRs), also referred as *Variable Length Tandem Repeats* by [Hauth and Joseph, 2002]. Let $\mathbf{X}$ and $\mathbf{x}$ be two segments (usually of different lengths) from the alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ such that $d(\mathbf{X}, \mathbf{x}) > k$, where d is the edit distance and k is some threshold value.

Definition 3.1. A **Perfect Nested Tandem Repeat** is a segment of the form

$$\mathbf{x}^{s_0}\prod_{i=1}^{l}(\mathbf{X}\,\mathbf{x}^{s_i}) = \mathbf{x}^{s_0}\mathbf{X}\mathbf{x}^{s_1}\mathbf{X}\cdots\mathbf{x}^{s_{l-1}}\mathbf{X}\mathbf{x}^{s_l},$$

where $l > 2$, $s_i \geq 1$ for $i = 1, \cdots, l-1$, with at least one $s_j \geq 2$. The $\mathbf{x}$ motif is called the *tandem repeat* and the $\mathbf{X}$ motif is the *interspersed repeat*. The concatenations of the tandem repeats $\mathbf{x}^{s_i}$ alone, and of the interspersed motifs $\mathbf{X}$ alone, both form perfect tandem repeats.

Example: $\mathbf{x}$ = ACGGT, $\mathbf{X}$ = GACCTTTGG, $l = 7$, $s_0 = 0$, $s_1 = 3$, $s_2 = 5$, $s_3 = 2$, $s_4 = 4$, $s_5 = 1$, $s_6 = s_7 = 2$

$\mathbf{x}^0\prod_{i=1}^{7}(\mathbf{X}\,\mathbf{x}^{s_i}) = \mathbf{X}\mathbf{x}\mathbf{x}\mathbf{X}\mathbf{x}\mathbf{x}\mathbf{x}\mathbf{x}\mathbf{X}\mathbf{x}\mathbf{x}\mathbf{X}\mathbf{x}\mathbf{x}\mathbf{x}\mathbf{x}\mathbf{X}\mathbf{x}\mathbf{X}\mathbf{x}\mathbf{x}\mathbf{X}\mathbf{x}\mathbf{x} =$

GACCTTTGG ACGGT ACGGT ACGGT
GACCTTTGG ACGGT ACGGT ACGGT ACGGT ACGGT
GACCTTTGG ACGGT ACGGT
GACCTTTGG ACGGT ACGGT ACGGT ACGGT
GACCTTTGG ACGGT
GACCTTTGG ACGGT ACGGT
GACCTTTGG ACGGT ACGGT.

Definition 3.2. $(k_1, k_2)-$**(approximate) Nested Tandem Repeats** A $(k_1, k_2)-$NTR is similar to perfect NTR but with the flexibility that we allow the tandem motifs $\mathbf{x}_i$ to differ from a template motif $\mathbf{x}$ by at most k_1 edit operations, and the interspersed motifs $\mathbf{X}_j$ to differ from a template motif by at most k_2 edit operations.

Examples:

- **NTR**:
  AGG AGG CTCAG AGG CTCAG AGG AGG AGG CTCAG.
  The motifs are AGG, CTCAG.
- $(1, 2)-$**NTR**:
  AG<u>A</u> AGG CT<u>TC</u>G AGG CTCAG A<u>A</u>G.
  The motifs are AGG, CTCAG.

3.2. **Related Work.** Various algorithms have been introduced to find exact tandem repeats. Such algorithms were developed mainly for theoretical purposes, namely to solve the problem of finding squares in strings [Apostolico and Preparata, 1983, Crochemore, 1981, Kolpakov *et al*., 2001, Main and Lorentz, 1984, Stoye and Gusfield, 2002]. These algorithms are not easily adapted to finding the approximate tandem repeats that usually occur in DNA.

A number of algorithms, e.g. [Delgrange and Rivals, 2004, Landau *et al*., 2001] consider motifs differing only by substitutions, using the Hamming distance as a measure of similarity. Others, e.g. [Benson, 1999, Hauth and Joseph, 2002, Domaniç and Preparata, 2007, Sagot and Myers, 1998, Wexler *et al*., 2005], have considered insertions and deletions by considering the edit distance. Most of these algorithms have two phases, a scanning phase that locates possible tandem repeats, and an analysis phase that checks the candidate tandem repeats found during the scanning phase.

Different software use different approaches in the scanning phase. Some of them scan the sequence with a small window looking for frequent occurrences of some patterns then find the distance between these occurrences, e.g. [Benson, 1999] and [Hauth and Joseph, 2002].

Others fix the distance and look for subsequence matches by scanning the sequence with two windows that are at a fixed distance apart, e.g. [Wexler *et al*., 2005], or three windows of length 1 for example [Boeva *et al*., 2006]. Another algorithm introduced by [Domaniç and Preparata, 2007] looks for the preceding occurrence of substrings of a fixed length.

The only algorithm designed to look for NTRs is that of [Hauth and Joseph, 2002] which searches for tandem motifs of length at most 6 nucleotides.

4. The Algorithm

In this section, we present the algorithm we have developed to find nested tandem repeats in DNA sequences. This algorithm requires preset parameters: k_1, k_2 which bound the edit distances of the target segments from the tandem repeat and interspersed motifs; and size ranges $[\min\ t_{tr}, \max\ t_{tr}]$ and $[\min\ t_{ntr}, \max\ t_{ntr}]$ bounding their lengths.
Search phase: An NTR contains an approximate tandem repeat which occurs with at least two adjacent approximate copies of a motif x, with other approximate copies of this motif in the neighbourhood, but not adjacent. The search is applied to a sequence S.

Consider the following example where S contains the NTR

$$\mathbf{x}\prod_{i=1}^{4}(\mathbf{X}\,\mathbf{x}^{\mathtt{s_i}}) = \mathbf{x}\mathbf{X}\underline{\mathbf{xxx}}\mathbf{X}\mathbf{xxxxxx}\mathbf{X}\mathbf{xx}\mathbf{X},$$

where each $\mathbf{x}$ is within k_1 of the motif GACC, and each $\mathbf{X}$ is within k_2 of TTACGGA.

We begin searching S for a k_1-TR motif x. Several good algorithms for finding $k-$TR have been developed for this [Benson, 1999, Domaniç and Preparata, 2007, Wexler *et al*., 2005]. We have chosen to implement the approach of [Wexler *et al*., 2005]. The sequence is scanned from the left until an approximate tandem repeat with motif $\mathbf{x}$ is reported and verified. In our example this finds the underlined TR: $\mathbf{xxx}$. Then the neighbourhood of this region is searched for further non-adjacent segments which contain approximate $\mathbf{x}$ copies; if none are found then this tandem repeat is ignored, and the scan continues.

In our example we would locate the segments with $\mathbf{x}$ to the left, and $\mathbf{xxxxxx}$, and $\mathbf{xx}$ to the right.

The focus then is to search for the interspersed regions which in this example locates the common (approximate) motif $\mathbf{X}$. If a common interspersed motif is not found, this candidate is abandoned and the TR scan resumed. Otherwise we have an NTR with TR motif $\mathbf{x}$ and interspersed motif $\mathbf{X}$ and we form two lists, one containing the segments which approximate $\mathbf{x}$, the other with the segments approximating $\mathbf{X}$. At this point we align the segments in each list and we construct a consensus $\mathbf{x}$ and $\mathbf{X}$ for each.
Verification phase. The main task of this phase is to align and verify the NTR region. At this point we have the two consensus patterns of the putative NTR . This phase begins by aligning the consensus patterns against the sequence using nested wrap-around dynamic programming (NWDP).

Nested wrap-around dynamic programming is used to align both consensus patterns against the NTR region (Matroud *et al.*, in preparation). The NWDP has time and space complexity $O(n.|\mathbf{x}|.|\mathbf{X}|)$, where n is the length of the NTR region and $|\mathbf{x}|$ and $|\mathbf{X}|$ are the length of the tandem motif and the length of the interspersed motif respectively.

The search phase procedure is described in Figure 1.

4.1. **Results.**

We have implemented the algorithm. Searches on simulated (for testing) and real sequence data have been done. To date our algorithm has found three NTR regions of interest.

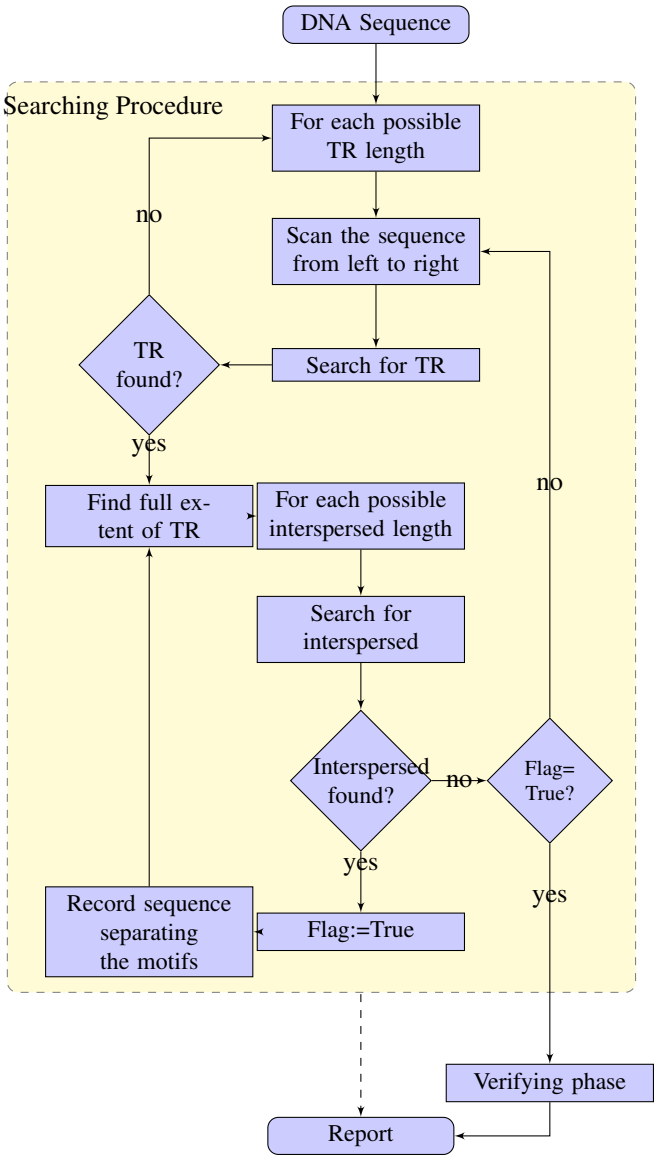

FIGURE 1. The search phase of the `NTRFinder` algorithm.

4.1.1. *Sequence 1: IGS region in Taro (C. esculenta).* This sequence (kindly supplied by A. Clarke, Allan Wilson Centre) is from the IGS region in Taro, and contains 93 approximate copies of the pattern
$\mathbf{x} =$ `TCGCACAGCCG`,
and 11 approximate copies of
$\mathbf{X} =$ `TTCTGGGCAAAACGGCTGGGTGACGTGCTGAACTGGCCAGCTGGTTCG`.

4.1.2. *Sequence 2: human X chromosome.* We found an NTR with consensus TR motif $\mathbf{x} = \texttt{GATA}$ and consensus interspersed motif $\mathbf{X} = \texttt{GAATAAGATGGAT}$. There are 31 approximate copies of $\mathbf{x}$ and and 6 approximate copies of $\mathbf{X}$ at positions 51,830 - 52,031.

4.1.3. *Sequence 3: Reference number AC188994.1.* This sequence is from GenBank, with reference number AC188994.1, from the common Marmoset (*Callithrix jacchus*) chromosome UNK clone CH259-68H12 at positions 145,997 - 149,988. The NTR region is of length 3292 bp. It contains 468 approximate copies of the pattern $\mathbf{x} = \texttt{GATA}$ and 144 approximate copies of the pattern $\mathbf{X} = \texttt{GATGCCA}$.

5. PARSING

If the boundaries of the tandem repeat and interspersed segments are moved a few bp in either direction, the translated segments may still satisfy the definition of an approximate NTR. The *parsing problem* is the problem of finding the optimal frame-shift. The parsing problem occurs in TR also, but is more significant for nested tandem repeats, so we have further developed the work by [Benson, 1999] and [Sammeth and Stoye, 2006] by introducing an additional criterion to decide which boundaries are the most likely. Their solution is to consider the best alignment score to determine the boundaries. However, different parsing may give the same alignment score, so additional criteria are needed. For example consider the following example of an approximate tandem repeat (in bold)

$$\cdots \texttt{GACC}\ \mathbf{ACGA\ ACGT\ ACGA\ ACGT}\ \texttt{ATTA}\ \cdots$$

with the consensus pattern being ACGA. If we shift the boundaries one nucleotide to the left we obtain

$$\cdots \texttt{GGAC}\ \mathbf{CACG\ AACG\ TACG\ AACG}\ \texttt{TATT}\ \cdots,$$

with consensus pattern AACG. Both of these consensus patterns give the same score when they are aligned. However, in the first example there are two variants $\texttt{ACGA}, \texttt{ACGT}$ and in the second there are three variants $\texttt{CACG}, \texttt{AACG}, \texttt{TACG}$. We can select the parsing which minimises the number of variants as a proxy for minimising the number of edit operations required to generate these tandem copies from a single ancestral motif, so in this example we would select the first parsing as our preferred choice of boundaries.

More generally our criterion to decide the preferred parsing is the *length of the minimum spanning tree* connecting the variants, with edges weighted by edit operations. In the next section a detailed explanation of our approach to solve the problem is illustrated.

5.1. **Relationship Graph of Repeat Variants.** To obtain the *minimum spanning tree* connecting the repeat variants we construct their *relationship graph*. The set of variants are hypothesised to be homologous (to have evolved from a single ancestral segment) with an ancestral history of duplication, substitution and deletion. If we knew this history, we could illustrate it by a tree T, rooted at the common ancestor, where each vertex represents a variant (historical or contemporary), and each directed edge represents a set of edit operations transforming an ancestor to its descendant. We use a parsimony principle for finding a tree T connecting the variants, where the total number of edit operations to transform the connected variants is minimal.

When the variants are closely connected we consider the $1-$cluster graph $G_1 = (V, E)$ [Hendy *et al*., 1980], where V is the set of contemporary variants and the set of edges $E = \{(u, v) | u, v \in V; d(u, v) = 1\}$ connects each pair of variants which differ by a single edit operation. However G_1 will not usually be a tree. The graph G_1 may contain circuits, for example, the four variants a, d, q, z in Figure 2. This may be the consequence

of a parallel mutation, which can happen when the density of substitutions is high and the segments are short.

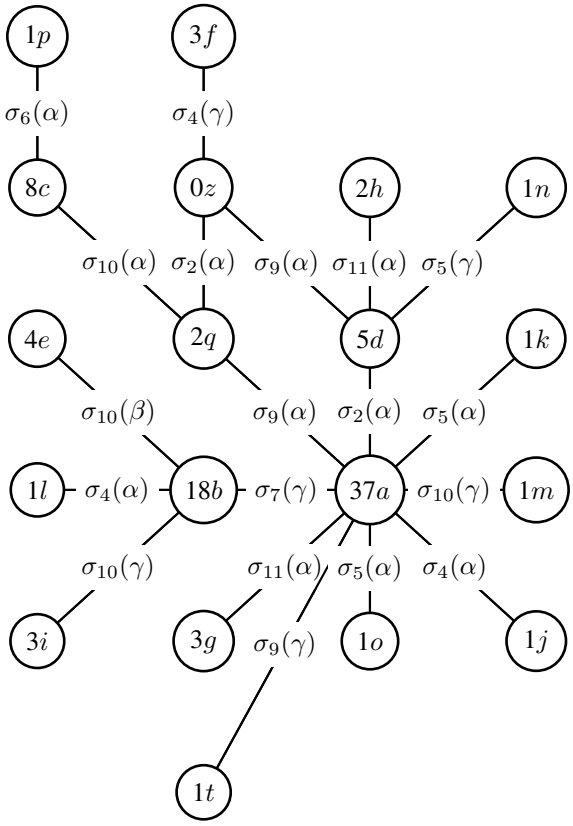

FIGURE 2. The relationship graph of the taro variants with the parsing being CGTCGCACAGC which is the variant *a*. Inside each node is the frequency of each variant followed by the variant, for example, *a* occurs 37 times in the NTR region. The arrows contain the edit operation needed to change one variant to another, for example, a γ operation at site 10 changes variant *a* to variant *m*. The variant *z* has frequency 0 because it does not appear in the NTR region.

The set V may not contain all ancestral variants, as some may have been lost by deletion, and hence G_1 might not be connected. This may occur if the substitution rate is higher than the duplication rate. Connectivity can be achieved by adding edges (u, v) with $d(u, v) > 1$, where u, v are in different components of G_1 and *Steiner points* (new vertices which could represent ancestral variants) when adjacent edges share some common edit operations, to reduce the total number of edit operations across the edges.

We will refer to a connected graph G which connects all the variants, and which may include additional hypothetical variants, and in which every edge represents a set of edit operations, as a *relationship graph*. For any such relationship graph we identify a *spanning tree* T of G to be any subgraph of G whose vertex set includes all the observed variants. A *minimal spanning tree* is a spanning tree where the number of edit operations summed over

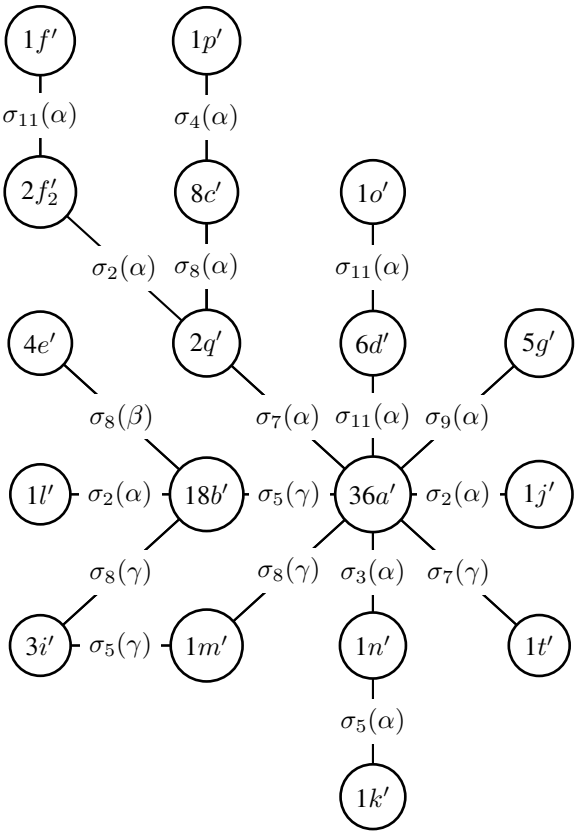

FIGURE 3. The relationship graph of the taro variants obtained by shifting the boundaries two nucleotides to the right. The variant *a'* here is TCGCACAGCCG. The number of variants is 18, one less than the previous parsing in Figure 2.

all edges is minimal. We define the *length* $l(G)$ of G to be the number of edit operations in any minimal spanning tree, and seek to find a *tight relationship graph* for which $l(G)$ is minimal across all possible relationships graphs. In general we expect this to be an NP-hard problem.

When the edit operations are only single nucleotide substitutions, a minimal spanning tree is a maximum parsimony tree of V, which can be found using standard parsimony tree methods. Then every spanning tree will be minimal. (When the variants are not too sparse we can use $2-$clusters as in [Hendy *et al*., 1980], to efficiently find a tight relationship graph.)

Example:

Consider the tandem repeats in the IGS region in Taro. The length of the repeat pattern is 11 so there are 11 potential parsings to consider. In Figure 2 there are 19 variants (including one Steiner point) and in Figure 3 where the boundaries are moved two steps right there are only 18 variants. The minimum spanning tree of Figure 3 of length 17 gives a preferred parsing to that of Figure 2 which will have length 18.

6. DISCUSSION

In reviewing the literature, little work exists on complex repetitive structures such as nested tandem repeats. The problem of finding nested tandem repeats is addressed in this study. In particular, a biologically relevant criterion for the parsing problem for tandem repeats had not been addressed. The parsing solution that we propose is biologically significant as it considers the evolutionary history of duplication events.

The motivation for our study has been the potential use of NTRs as a marker for genetic studies of populations and of species. We have done some analysis on the nested tandem repeat in the intergenic spacer region in *C. esculenta*, noting some variation in the NTRs derived from domesticated varieties sourced from New Zealand, Australia and Japan (further varieties are currently being analysed). By considering some edit operations such as deletion, mutation, and duplication we can align both nested tandem repeat regions of the domesticated varieties and calculate the score of the alignment, which can be considered as a measure of distance between both sequences. In particular they appear to share some common inferred histories of the development of the NTRs from a simpler structure of two motifs. These edit operations appear to be occurring on a 1,000 year timescale, so this analysis offers the potential to date the prehistory of the early agriculture of this ancient staple food crop.

7. CONCLUSION

The nested tandem repeat structure is a complex structure that requires further analysis and study. The number of copy variants in the NTR region and the relations between these copies might suggest a tandem repeat generation mechanism. In this paper, we have introduced a new algorithm to find nested tandem repeats. The first phase of the algorithm has $O(n(\max n_{tr})(\max n_{ntr}))$ time complexity, while the second phase (the alignment) needs $O(n(\max n_{tr})(\max n_{ntr}))$ space and time, where n is the length of the NTR region, $\max n_{tr}$ and $\max n_{tr}$ are the maximum tandem motif and interspersed motif length respectively.

ACKNOWLEDGEMENT

Andrew Clarke, Peter Matthews (for providing data and useful background about Taro), Hussain Matawa (for assisting in the development of the program interface).
Funding: This project was funded by the Allan Wilson Centre for Molecular Ecology and Evolution.

REFERENCES

[Apostolico and Preparata, 1983] Apostolico, A., Preparata, F. P., (1983) Optimal Off-Line Detection of Repetitions in a String, *Theor. Comput. Sci.*, **22**, 297-315.

[Benson, 1999] Benson G., (1999) Tandem repeats finder: a program to analyze DNA sequences, *Nucl. Acids Res.*,**27, 2**, 573-580.

[Boeva *et al*., 2006] Boeva, V. and Regnier, M. and Papatsenko, D., Makeev, V., (2006) Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression, *Bioinformatics*, **22, 6**, 676-684.

[Crochemore, 1981] Crochemore M., (1981) An Optimal Algorithm for Computing the Repetitions in a Word, *Inf. Process. Lett.*, **12, 5**, 244-250.

[Delgrange and Rivals, 2004] Delgrange, O., Rivals, E., (2004) STAR: an algorithm to Search for Tandem Approximate Repeats, *Journal of Comp. Bio.*, **20, 16**, 2812-2820.

[Domaniç and Preparata, 2007] Domaniç, N. O., Preparata, F. P., (2007) A Novel Approach to the Detection of Genomic Approximate Tandem Repeats in the Levenshtein Metric, *Journal of Comp. Bio.*, **14, 7**, 873-891.

[Hauth and Joseph, 2002] Hauth, A. M., Joseph, D., (2002) Beyond tandem repeats: complex pattern structures and distant regions of similarity, *ISMB.*, 31-37.

[Hendy *et al*., 1980] Hendy, M., Foulds, L. R., Penny, D., (1980) Proving phylogenetic trees minimal with $l-$clustering and set partitioning, *Math. Biosc.*, **51, 1-2**, 71-88.

[Jeffreys *et al*., 1980] Jeffreys, A.J. and Wilson, V. and Thein, S.L. (1980) Individual-specific fingerprints of human DNA., *Nature.*, **51**, 71-88.

[Kimura, 1981] Kimura, M., (1981) Estimation of evolutionary sequences between homologous nucleotide sequences, *Proc. Nat. Acad Sci. USA*, **78**, 454-458.

[Kolpakov *et al*., 2001] Kolpakov, R., Kucherov, G., Logiciel, T. G., (2001) Finding approximate repetitions under Hamming distance, *Theor. Comput. Sci.*, **22, 6**, 170-181.

[Landau *et al*., 2001] Landau, G. M. and Schmidt, J. P. and Sokol, D., (2001) An Algorithm for Approximate Tandem Repeats, *Journal of Comp. Bio.*, **8, 1**, 1-18.

[Macdonald *et al*., 1993] Macdonald, M. E. *et al*., (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group, *Cell.*, **72**, 971-983.

[Main and Lorentz, 1984] Main, M. G., Lorentz, R. J.. (1984) An $O(n\ log\ n)$ Algorithm for Finding All Repetitions in a String, *J. Algorithms*, **5, 3**, 422-432.

[Matthews, 2004] Matthews, P. J., (2004) Genetic diversity in taro, and the preservation of culinary knowledge. *Ethnobotany Res. Appl.* **2**, 55-71.

[Matroud *et al*., 2010] Matroud, A. A., Hendy, M. D., Tuffley, C. P., (2010) Nested Wrap-Around Dynamic Programming: An Algorithm to Align Nested Tandem Repeat. In prep.

[Sagot and Myers, 1998] Sagot, M. F., Myers, E. W., (1998) Identifying Satellites and Periodic Repetitions in Biological Sequences, *Journal of Comp. Bio.*, **5, 3**, 539-554.

[Sammeth and Stoye, 2006] Sammeth, M., Stoye, M.,(2006) Comparing Tandem Repeats with Duplications and Excisions of Variable Degree, *IEEE/ACM Trans. on Comp. Bio. and Bioinf.*, **3**, 395-407.

[Stoye and Gusfield, 2002] Stoye, J., Gusfield, D. (2002) Simple and flexible detection of contiguous repeats using a suffix tree, *Theor. Comput. Sci.*, **270, 1-2**, 843-856.

[Weitzmann *et al*., 1997] Weitzmann, M. N. and Woodford, K. J. and Usdin, K. (1997) DNA Secondary Structures and the Evolution of Hypervariable Tandem Arrays, *J. Biol. Chem.*, **272, 14**, 9517-9523.

[Wells, 1996] Wells, R. A. (1996) Molecular Basis of Genetic Instability of Triplet Repeats, *J. Biol. Chem.*, **271, 6**, 2875-2878.

[Wexler *et al*., 2005] Wexler, Y. and Yakhini, Z. and Kashi, Y.,Geiger, D., (2005) Finding Approximate Tandem Repeats in Genomic Sequences, *Journal of Comp. Bio.*, **12, 7**, 928-942.

[1]ALLAN WILSON CENTRE FOR MOLECULAR ECOLOGY AND EVOLUTION, MASSEY UNIVERSITY, PRIVATE BAG 11 222 PALMERSTON NORTH, NEW ZEALAND., [2] INSTITUTE OF FUNDAMENTAL SCIENCES , MASSEY UNIVERSITY, PRIVATE BAG 11 222 PALMERSTON NORTH, NEW ZEALAND.