BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S31–S37

# *Beyond tandem repeats: complex pattern structures and distant regions of similarity*

*Amy M. Hauth* and Deborah A. Joseph*

*Department of Computer Sciences, University of Wisconsin - Madison, 1210 W. Dayton Ave., Madison, WI 53706, USA*

## ABSTRACT

**Motivation:** Tandem repeats (TRs) are associated with human disease, play a role in evolution and are important in regulatory processes. Despite their importance, locating and characterizing these patterns within anonymous DNA sequences remains a challenge. In part, the difficulty is due to imperfect conservation of patterns and complex pattern structures. We study recognition algorithms for two complex pattern structures: variable length tandem repeats (VLTRs) and multi-period tandem repeats (MPTRs).

**Results:** We extend previous algorithmic research to a class of regular tandem repeats (RegTRs). We formally define RegTRs, as well as two important subclasses: VLTRs and MPTRs. We present algorithms for identification of TRs in these classes. Furthermore, our algorithms identify degenerate VLTRs and MPTRs: repeats containing substitutions, insertions and deletions. To illustrate our work, we present results of our analysis for two difficult regions in cattle and human data which reflect practical occurrences of these subclasses in GenBank sequence data.

In addition, we show the applicability of our algorithmic techniques for identifying Alu sequences, gene clusters and other distant regions of similarity. We illustrate this with an example from yeast chromosome I.

**Availability:** Algorithms can be accessed at http://www.cs.wisc.edu/areas/theory.

**Contact:** kryder@cs.wisc.edu; joseph@cs.wisc.edu

**Keywords:** tandem repeat; satellite; SSR; STR; MPTR; VLTR; DNA; algorithm.

## INTRODUCTION

The goal of our research is to develop algorithms to identify repeated occurrences of a pattern within a genomic sequence. A simple example of recurring patterns is the tandem repeat (TR) which involves consecutive occurrences of a pattern. For example, the sequence TGTGTG is defined by three consecutive occurrences of a TG pattern. TRs can arise when mutational events occur to transform a segment of DNA sequence into two or more copies. Additional mutation allows the individual copies to diverge. The algorithmic challenge is to identify the repeat pattern structure within a sequence without *a priori* knowledge of the composition of the repeat.

Significant effort in recent years has yielded many algorithms for identifying TRs having a simple pattern structure (Review in Benson, 1999; Kurtz *et al.*, 2000, 2001. Also, Adebiyi *et al.*, 2001; Landau *et al.*, 2001; Volfovsky *et al.*, 2001). In this paper, we describe two broad classes of complex pattern structures absent from the literature but present in genomic sequence data. Prior approaches have specific limitations that do not allow identification of TRs in these new classes.

In practice, two issues complicate repeat identification: imperfect conservation of patterns and complex pattern structures. We believe fundamental issues associated with these complications cannot be addressed by a single technique. Thus, our algorithm is a collection of techniques each designed to address a fundamental issue observed in real DNA sequence data. In addition, we extend the algorithm to recognize distant regions of similarity.

## DEFINITIONS

Several classes of tandem repeats can be defined. Simple sequence repeats (SSRs) have patterns which are typically less than six nucleotides in length and are fairly well conserved. Long tandem repeats have long patterns (at least one hundred nucleotides in length) and are often less conserved. Variable length tandem repeats (VLTRs) and multi-periodic tandem repeats (MPTRs) have complex pattern structures.

We assume $S$ is a DNA sequence of length $n$ and over the DNA nucleotide alphabet $\Sigma = \{A,C,G,T\}$. A pattern, $p$, is a finite length word over $S$. A perfect tandem repeat, $T = p_c$, is a simple concatenation of $c$ copies of $p$, where $c \geq 2$ and designates the copy number of the repeat. For example, TGTGTG is a perfect TR with $p =$ TG and $c = 3$.

Many tandem repeats have complex structures defined by two or more patterns. We define and study two classes: VLTRs and MPTRs.

*To whom correspondence should be addressed.

DEFINITION. A simple nested tandem repeat is formed by the nested concatenation of several patterns. Formally, a finite length word, $N$, is a simple nested repeat if $N = ((p_1)_{c_1}(p_2)_{c_2}(p_3)_{c_3}\dots(p_k)_{c_k})_{c_{k+1}}$ where the $p_i$'s are patterns and the $c_i$'s are $\geq 1$. A variable length tandem repeat (VLTR) is a simple nested tandem repeat in which the copy number for some pattern $p_i$ is variable rather than constant.

EXAMPLE. CATTAGCC TGTGTGTG CATTAGCC TGTG CATTAGCC TGTGTGTGTG

This VLTR sequence is composed of two repeated patterns, CATTAGCC and TG, and can be written as CATTAGCC(TG)$_4$CATTAGCC(TG)$_2$CATTAGCC(TG)$_5$, or as ((CATTAGCC)$_1$(TG)$_*$)$_3$. The later expression explicitly indicates that the copy number for the CATTAGCC pattern is fixed and equal to one, while the copy number for the TG pattern is variable, denoted with a *.

DEFINITION. A multi-period tandem repeat (MPTR) is formed by the nested concatenation of two or more $i$-similar patterns. Consider the following concatenation of 1-similar patterns.

EXAMPLE. CAGTA CAGCA CAATA CAGCA CAGTA CAGCA CAATA CAGCA

This sequence is *almost a* perfect tandem repeat of the form (CAGTA)$_8$. However, in every other copy the T becomes a C and in every fourth copy the G becomes an A. So, this sequence is a simple nested concatenation of three 1-similar patterns, CAGTA, CAGCA and CAATA; a MPTR. The multiple periodicity exhibited is written as follows: a first multiple (1×) of (CA$^G/_A{}^T/_C$A)$_8$, a second multiple (2×) of (CA$^G/_A$TA CA$^G/_A$CA)$_4$ and a fourth multiple (4×) of (CAGTA CAGCA CAATA CAGCA)$_2$ where $^G/_A$ is a position represented by both G and A.

DEFINITION. The class of regular tandem repeats (RegTRs) includes perfect and degenerate TRs, VLTRs and MPTRs.

Existing algorithms handle SSRs and long tandem repeats well. Our algorithm identifies perfect and degenerate repeats having both simple and complex pattern structures including SSRs, MPTRs and VLTRs.

## ALGORITHM

Our identification of tandem repeats involves both locating and characterizing regions. We accomplish this goal by performing three major tasks: (1) isolate a tandem repeat by determining its period and its approximate sequence location, (2) determine the pattern associated with a region period and (3) characterize the region using the pattern (See Figure 1). Our technique is similar to Benson, 1999 in that we analyse $k$-length substrings in a DNA sequence by finding recurring distances between identical substrings. We differ in that a statistical model is not used to locate interesting periods, but rather a simple and accurate filter coupled with techniques to data mine sequence differences.

### Isolate initial TR period and location

We construct a distance array, $D$, parallel to $S$, to record a distance, $d$, between identical occurrences of $k$-length substrings, termed *words*. The $i$th position in $D$, $d_i$, records the distance from the word, $w$, ending in position $i$ in $S$ to the previous occurrence of $w$ at position $j$, $j < i$ such that $w$ does not occur between $i$ and $j$. For example, a word occurs at positions 619 and 1083. A distance of $1083 - 618 = 465$ is recorded at position 1083. The leftmost occurrence of a word in $S$ is recorded in $D$ as a distance of zero. Perfect TRs with a pattern length $\leq k$ correspond to $d \leq k$ and are processed as SSRs in a separate analysis.

Perfect TRs create runs of consecutive, identical $d$ in $D$, termed $d$ runs. Degenerate TRs exhibit $d$ runs but not throughout the entire region. Edit differences (substitutions, insertions and deletions) are reflected in $D$ such that regions with many differences exhibit fewer and shorter $d$ runs. Our algorithm requires at least five identical $d$ occurrences to be present in one or more runs. This requires at minimum one very strong $d$ run of length five, or two moderate $d$ runs of length two and three, or three weak $d$ runs of length two each.

We construct histograms which place the position of identical distances in the same bin or peak. A filtered histogram, $F$, contains only positions associated with $d$ runs. We select peaks in $F$ beginning with the peak containing the most positions. We place proximal positions in each qualifying peak in groups. Each group containing at least five positions is processed.

The distance between identical words determines the period of a region. Previous models identify only simple pattern structures and use the distance associated with a histogram peak as the initial period of a region. Complex pattern structures have several associated periods. Thus, a more thorough period analysis is necessary.

Our period analysis uses word and distance similarities to determine significant periods within a region. We begin with positions from the selected distance peak and their associated words. We find other histogram peaks containing word occurrences identical to words in the selected peak. Twice we extend from a set of words isolated from selected peaks to related histogram peaks. This process determines significant region periods and approximate region location and assigns a base period, $b$, to the region.

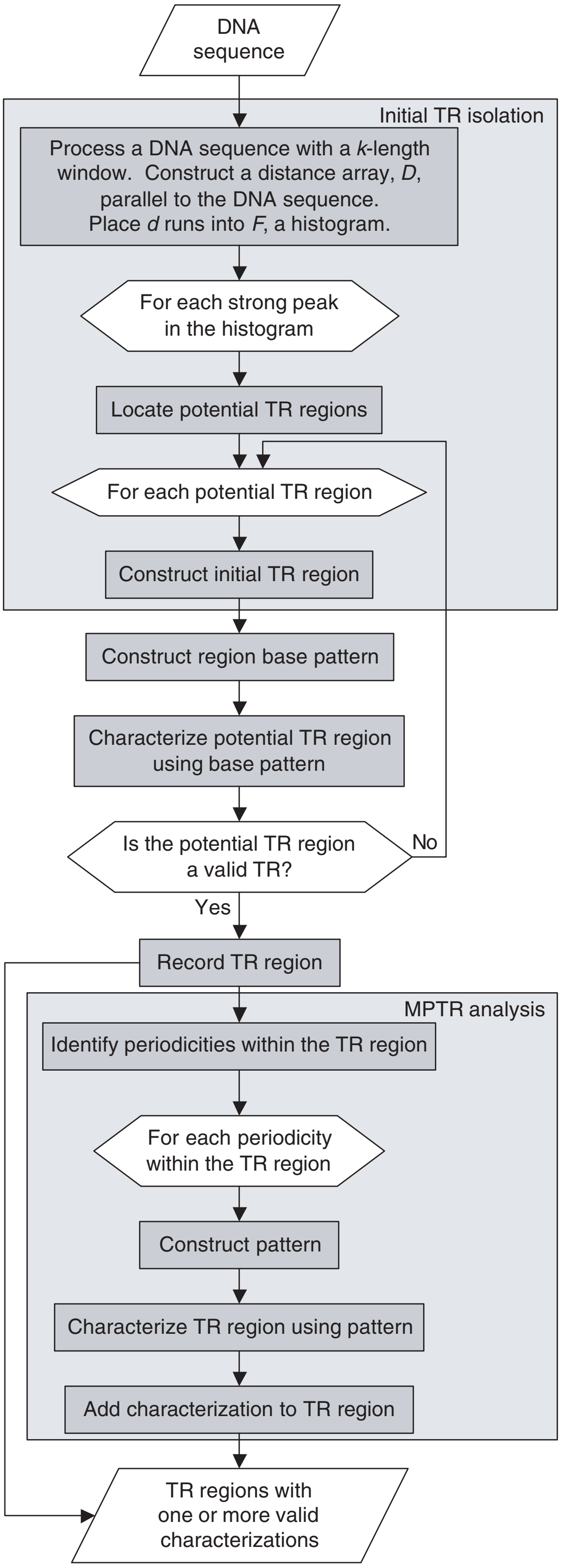

DNA sequence
Initial TR isolation
Process a DNA sequence with a k-length window. Construct a distance array, D, parallel to the DNA sequence. Place d runs into F, a histogram.
For each strong peak in the histogram
Locate potential TR regions
For each potential TR region
Construct initial TR region
Construct region base pattern
Characterize potential TR region using base pattern
Is the potential TR region a valid TR?
No
Yes
Record TR region
MPTR analysis
Identify periodicities within the TR region
For each periodicity within the TR region
Construct pattern
Characterize TR region using pattern
Add characterization to TR region
TR regions with one or more valid characterizations

**Fig. 1.** Basic identification algorithm locates TR regions in a DNA sequence and characterizes the region using one or more patterns.

## Construct region base pattern

Region patterns are constructed in several ways dependent upon the class designation of the region. The general procedure is to select a segment of sequence corresponding to all or a portion of a copy, to align the segment to several copies of the region and to form a pattern using the consensus formed from the alignment. This pattern is designated as the region base pattern.

The best possible copy is selected from the initial segment. Best is based on two criteria: copy length closest to $b$ and occurrences of $d = b$ in $D$ either at positions in the copy or at positions referring to the copy.

Current VLTR analysis finds nested TRs with short patterns. SSRs with identical patterns across multiple region copies are removed by adjusting distances within the region. The remaining sequence is analysed as a region having a simple pattern structure and a best copy is selected. A pattern is constructed by appending the SSR pattern to the best copy. For example, the SSRs have a TG pattern which is appended to CATTAGCC to form the pattern CATTAGCC(TG)$_*$.

## Characterize using region pattern

One or more patterns characterize a region by aligning each pattern to the region. We use wraparound dynamic programming (WDP; Fischetti *et al.*, 1992; Myers and Miller, 1989) modified to handle regular expressions (RegWDP) in the context of tandem repeat identification. WDP allows wraparound of scores from the rightmost column to the leftmost column of the ‘pattern’ in a 2D alignment. RegWDP allows wraparound of scores from any column to any other column dependent upon the regular expression of the pattern. Concatenation joins patterns together. Unions separate alternative patterns (see MPTR analysis). Closures indicate TRs and surround TR patterns. Our closure implementation allows partial matching of a pattern, unlike true regular expressions, and results from the biological reality of partial copies at the edges of a TR region.

We display the final alignment as a series of copies, each copy representing a pattern occurrence in $S$. We stack the copies such that a column through the copies corresponds to the same pattern position. Furthermore, we form a consensus pattern by tallying nucleotide and gap occurrences in each column. These are depicted as ambiguous nucleotide characters, as appropriate.

MPTR analysis is performed on all regions following alignment based pattern characterization. We extract a column in the alignment and treat it as an ordered sequence. We convert the sequence into a distance array, $C$, using $k = 3, 4, 5$ and 6, and identify significant periods in $C$. We analyse each column independently. Significant region periods coupled with significant column periods determine significant patterns across copies. If a column contains a three-nucleotide pattern then a new pattern is assembled composed of three occurrences of the consensus pattern formed from the base pattern alignment, modified to reflect the cross copy, three-nucleotide pattern. In addition, our analysis indicates period irregularities occur in which complete copies are lacking or extras are present relative to the pattern. To ensure proper alignment, our approach generalizes this to allow any modified consensus pattern to match to the next copy by forming a union between the modified consensus patterns. Thus, a new pattern which we believe is CAGTACAGCA is actually represented as CAGTA|CAGCA where '|' is read as 'or'.

### Identify SSRs

SSRs are identified during construction of $D$. Every $d \leq k$ indicates a SSR. We know this since the region length $\geq (k + d) \geq 2d$, i.e., at least two copies of the pattern. This approach is reasonable since we use $k = 6$ and SSRs are typically defined as TRs with a pattern of at most six nucleotides. Our algorithm identifies single $d$ and $d$ runs as perfect SSRs and checks for degenerate SSRs using pattern characterization procedures.

### Identify regions of similarity

Regions of similarity are pairs of non-adjacent regions with strong similarity. which are identified during construction of $D$. Potential TR regions, which fail to become a TR due to fewer than two copies of a pattern within the region, are examined. A region of similarity satisfies these criteria: (1) period $\geq$60, (2) region length $\geq$20 and either (3) the region contains $\geq 3d$ runs of the period or (4) the region contains $\geq$ 30 occurrences of the period. This analysis is preliminary with a wealth of potential extensions.

### Fundamental TR issues

Our early analysis explored the type of TRs present in annotated sequence data. Sequence data contains a wide assortment of TRs: TRs with short patterns, several hundred long patterns and perfect, poor and erratic pattern conservation, as well as TRs embedded within TRs and TRs with higher order patterns across multiple copies of a base pattern. We explored TR identification by existing algorithms. Early algorithms concentrated on short patterns and eventually were scaled up to long patterns. Also, initial algorithms identified perfectly conserved patterns with eventual extension to degenerate patterns. Now the pattern structure model, $T = p_c$, $c \geq 2$, requires modification to include complex pattern structures. Here we discuss several fundamental TR issues: the period, the conservation model and the role of data similarities and differences.

The region period is the single most crucial aspect of TR analysis. Patterns can vary due to edit difference, but the period should remain roughly constant. Period variation does occur due to insertions, deletions and complex pattern structures. Thus any analysis must take these factors into consideration. Our approach begins with a single region period which is expanded to a collection of periods based on recurrence of both words and distances between identical words.

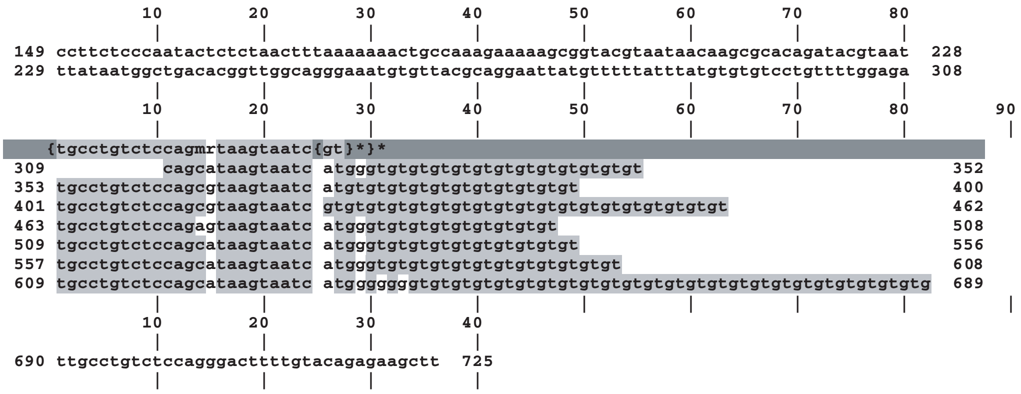
The conservation model in use by an algorithm dictates how a base pattern can vary within a region. Our sequence data analysis revealed a challenging TR class in which a poorly conserved copy is surrounded by moderate to well conserved copies. Models which focus on fixed differences from a pattern (e.g., $k$-difference approaches) have difficulty bridging across poorly conserved copies. Our approach focuses on matches and allows significant local pattern divergence when poor conservation is offset by good conservation.

Sequence similarities reveal the 'big picture' structure of a region while differences contain the nuances which characterize a particular region. Our identification of both VLTRs and MPTRs utilizes difference or noise mining which is described by a four step process: throw out data, analyse the remainder, acquire new information and incorporate information into the region's working model. For VLTRs, we throw out the SSRs, analyse the remaining sequence by adjusting region distances, acquire period and pattern information for the adjusted region and construct a new pattern containing a nested TR. For MPTRs, we iteratively throw out all but one column, analyse the columns for cross copy periods, acquire period and pattern information for multi-copy patterns and construct a new pattern composed of several modified base patterns.

## DISCUSSION

We develop and implement a collection of algorithmic techniques for identifying simple and complex repeat pattern structures in DNA sequences. Three examples illustrate the success of this project. The first is a bovine sequence containing a highly conserved VLTR. The second is a human sequence containing a difficult MPTR region which illustrates the need to detect multiple periodicities. The third is an analysis of yeast chromosome I in which we identify highly similar regions 12 to 14 kilobasepair (kbp) apart. The difficulties presented in these examples underscore fundamental issues encountered with repetitive structure identification beyond simple TR pattern structures.

CASE EXAMPLE I. A bovine sequence contains a fairly simple VLTR region (Kashi *et al.*, 1990, GenBank Locus: BOVTGN). Each copy of the VLTR contains a nested TR with a TG pattern which varies from 12 to 23 copies.



```
            10        20        30        40        50        60        70        80
             |         |         |         |         |         |         |         |
149 ccttctcccaatactctctaactttaaaaaaactgccaaagaaaaagcggtacgtaataacaagcgcacagatacgtaat  228
229 ttataatggctgacacggttggcagggaaatgtgttacgcaggaattatgtttttatttatgtgtgtcctgttttggaga  308
             |         |         |         |         |         |         |         |
            10        20        30        40        50        60        70        80        90
             |         |         |         |         |         |         |         |         |
   {tgcctgtctccagmrtaagtaatc{gt}*}*
309           cagcataagtaatc atgggtgtgtgtgtgtgtgtgtgtgtgtgt                              352
353 tgcctgtctccagcgtaagtaatc atgtgtgtgtgtgtgtgtgtgtgt                                    400
401 tgcctgtctccagcgtaagtaatc gtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt                      462
463 tgcctgtctccagagtaagtaatc atgggtgtgtgtgtgtgtgtgt                                      508
509 tgcctgtctccagcataagtaatc atgggtgtgtgtgtgtgtgtgtgt                                    556
557 tgcctgtctccagcataagtaatc atgggtgtgtgtgtgtgtgtgtgtgtgt                                608
609 tgcctgtctccagcataagtaatc atgggggggtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtg   689
             |         |         |         |         |         |         |         |         |
            10        20        30        40
             |         |         |         |
690 ttgcctgtctccagggacttttgtacagagaagctt  725
             |         |         |         |
```

**Fig. 2.** This bovine sequence (Genbank Locus: BOVTGN) contains a VLTR. The nested TR, (TG)$_*$, has a variable number of copies.

Less obvious, a well conserved 22-nucleotide subsequence occurs between each TG based repeat.

Figure 2 is a display generated by our web-based software characterizing this (and other) sequence(s). A regular expression represents the pattern (depicted as a gray bar across the top of the alignment). Region copies aligned to the pattern are stacked such that vertical columns through the copies correlate to the same pattern position. Columns of well conserved bases are coloured.

Earlier approaches do not properly characterize the complete region. For instance, Tandem Repeat Finder (Benson, 1999) locates each nested TG based TR, as well as adjacent copies in which the number of TG copies is held constant. The actual structure of the nested repeat is not determined.

CASE EXAMPLE II. A human sequence contains a fairly difficult MPTR region (Boan *et al.*, 1997, GenBank Locus: HSVDJSAT). On first examination, it is quite clear that the sequence contains a TR with a 9- or 10-nucleotide period that is highly conserved (Figure 3a, conserved columns are coloured red). Upon closer examination, the 9- and 10-nucleotide copies frequently alternate to form a period of 19 nucleotides which we refer to as the 2× pattern (Figure 3b). Furthermore, frequent alternations of the 2× pattern form a 38-nucleotide period, which we refer to as the 4× pattern (Figure 3c). Regular gapping occurs in both the 2× and the 4× pattern alignments. The 11× pattern captures the remaining structure (Figure 3d and 3e).

Our algorithm detects important multiples at 2×, 4×, 5×, 6× and 11×. Three patterns, CTGGGAGAGG, CTGGGAGAG, and CTGGGATTG, primarily compose the region. The 10-nucleotide pattern primarily occurs every other copy. The two 9-nucleotide patterns primarily occur every fourth copy. The 2× and 4× alignments capture these pattern structures. Their significance is visible by observing additional column colouring, a reflection of increased conservation. The 5× and 6× alignments offer alternative alignments to the 4× alignment and are echoes of the 11× pattern alignment.

The algorithm detects a period for an 11× pattern; a 104-nucleotide TR. The 11× alignment produces a better alignment, but still considerable gapping occurs (Figure 3d). However, a slightly different pattern produces a very good alignment in which very little gapping occurs (Figure 3e).

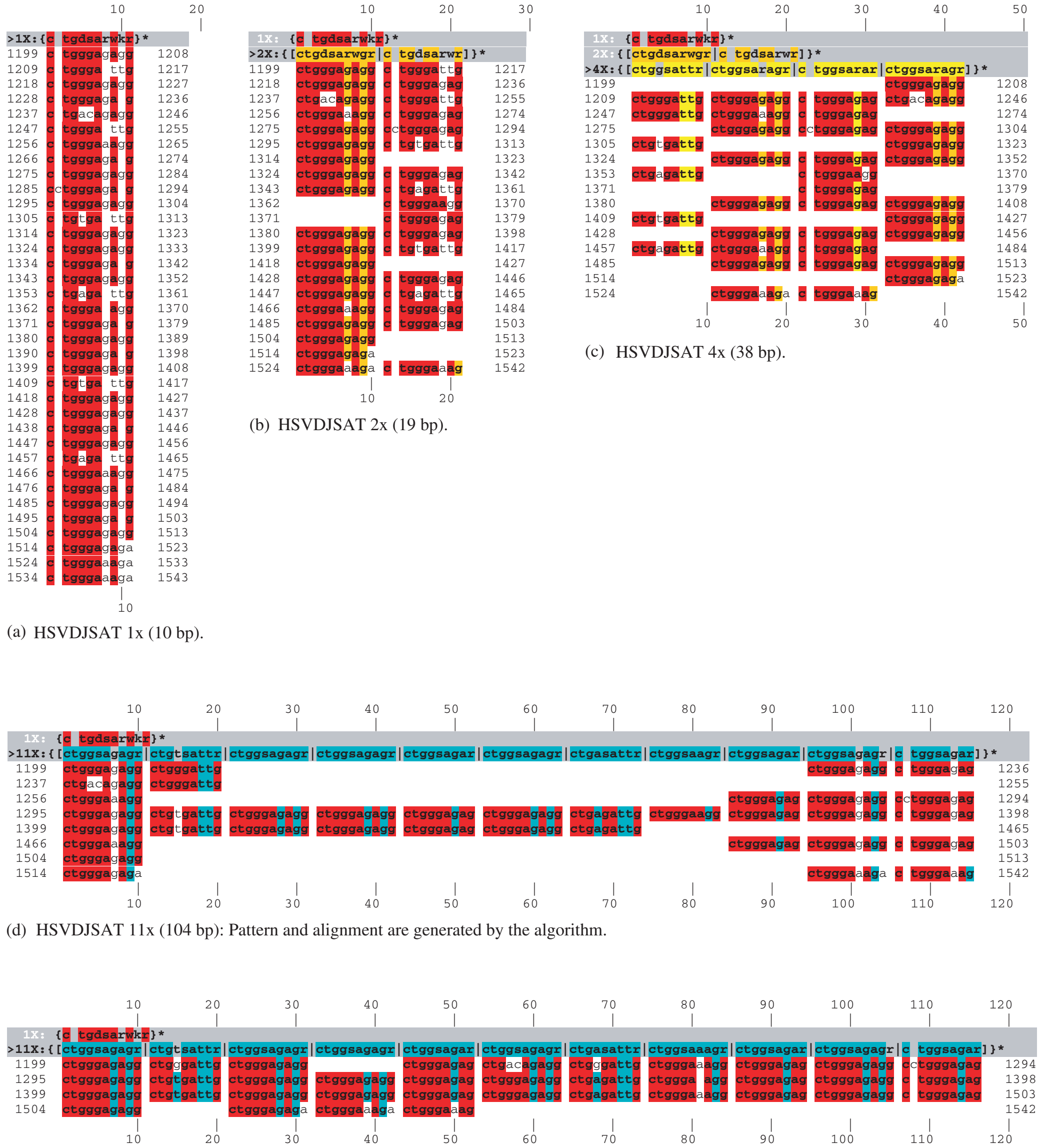
We believe that this sequence is a good illustration of the success of our techniques for characterizing MPTR regions by isolating base patterns, detecting patterns across copies of the base pattern and aligning and displaying the pattern structures in a easily understood manner.

CASE EXAMPLE III. A yeast sequence illustrates identification of distant regions of similarity (Bussey *et al.*, 1995, GenBank Locus: NC_001133; *Saccharomyces cerevisiae* chromosome I). Our preliminary analysis uses words of length six ($k = 6$) and reveals six distinct regions of similarity. We discuss three.

The first two are two FLO genes located at opposite ends of this chromosome; FLO9 at positions 24 to 28 kbp and FLO1 at positions 204 to 208 kbp. Each has a shadow located 10 to 14 kbp distal to its own location. Benson (1999) observed several similar TRs associated with the FLO1 gene, the FLO9 gene and FLO9 shadow. Our TR analysis identifies the same TR clusters. In addition, we identify regions of similarity which link each gene with its shadow. Analysis of the FLO9 and its shadow is shown in Figure 4 (left).

The third is a gene and pseudogene cluster bookmarked by LTRs 5.6 kbp apart (Figure 4 right). The cluster has many regions of similarity, many of which are linked to the Prm9p gene.

Analysis using larger windows spans greater distances. Using a window of $k = 12$, we identify a region of



(a) HSVDJSAT 1x (10 bp).

(b) HSVDJSAT 2x (19 bp).

(c) HSVDJSAT 4x (38 bp).

(d) HSVDJSAT 11x (104 bp): Pattern and alignment are generated by the algorithm.

(e) HSVDJSAT 11x (104 bp): Pattern contains an A inserted at position 80. The alignment shown is aligned by hand.

**Fig. 3.** This human sequence (GenBank locus: HSVDJSAT) contains a MPTR. (a)-(d) depict alignments generated by the algorithm. (e) shows an alignment edited by-hand which uses a pattern similar to the pattern in (d); an A is inserted at position 80.

similarity between a Pau7 gene and an open reading frame located 174 kbp apart (not shown). Also, the larger window increased the identified similarity between the FLO genes and their respective shadows but was not able to link the FLO genes to one another.

This preliminary analysis is intriguing and by no means complete. Though simple, it is powerful. Analysis using larger windows can span greater distances, however, it decreases the sensitivity of the algorithm. We believe extensions to our technique will enable more complete identification while retaining the sensitivity inherent with small window sizes.

## CONCLUSION

In this paper, we characterize complex repeat pattern structures and present identification algorithms for MPTRs and VLTRs which contain SSRs. In addition, distant regions of similarity are identified using our techniques.

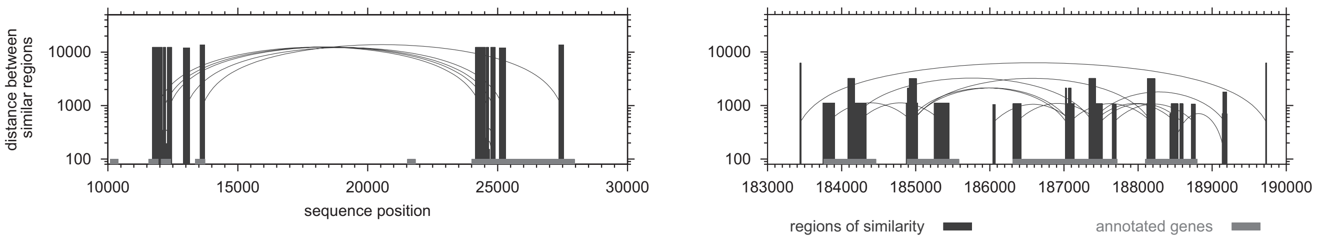


**Fig. 4.** Yeast Chromosome I Analysis. At left, the FLO9 gene (24 to 28 kbp) is linked to its shadow (11 to 14 kbp) by regions of similarity. At right, the Prm9p gene (about 187 kbp) is linked to several pseudogenes. The gene cluster is bookmarked by LTRs about 6 kbp apart. A region of similarity is two similar sequence regions and is denoted by two spikes linked by an arc. For convenience, the height of the arc and the two spikes reflects the distance between the similar regions. The width of a spike denotes the range of sequence positions associated with one half of a region of similarity.

## REFERENCES

Adebiyi,E.F. *et al.* (2001) An efficient algorithm for finding short approximate non-tandem repeats. *Bioinformatics*, **17**, S5–S12.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Boan,F. *et al.* (1997) Molecular characterization of a new human minisatellite that is able to form single-stranded loops *in vitro* and is recognized by nuclear proteins. *FEBS Lett.*, **418**, 251–257.

Bussey,H. *et al.* (1995) The nucleotide sequence of chromosome I from *Saccaromyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **92**, 3809–3813.

Fischetti,V. *et al.* (1992) Identifying Periodic Occurrences of a Template with Applications to Protein Structure. In Apostolico,A., Crochemore,M. *et al.*, (eds), *Proceedings of the Third Annual Symposium on Combinatorial Pattern Matching*, Lecture Notes in Computer Science, Springer, Berlin, **644**, pp. 111–120.

Kashi,Y. *et al.* (1990) $(\mathrm{TG})_n$ uncovers a sex-specific hybridization pattern in cattle. *Genomics*, **7**, 31–36.

Kurtz,S. *et al.* (2000) Computation and visualization of degenerate repeats in complete genome. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 228–238.

Kurtz,S. *et al.* (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, **29**, 4633–4642.

Landau,G.M. *et al.* (2001) An algorithm for approximate tandem repeats. *J. Comput. Biol.*, **8**, 1–18.

Myers,E.W. and Miller,W. (1989) Approximate matching of regular expressions. *Bull. Math. Biol.*, **51**, 5–37.

Volfovsky,N. *et al.* (2001) A clustering method for repeat analysis in DNA sequences. *Genome Biol.*, **2**, research0027.1–research0027.11.