An Algorithm for Finding Tandem Repeats of Unspecified Pattern Size

Gary Benson*

Abstract

A tandem repeat is two or more contiguous, *approximate* copies of a pattern of nucleotides. Tandem repeats occur frequently in the human genome. They have been shown to cause human disease, may play a variety of regulatory and evolutionary roles, and are important laboratory tools. Extensive knowledge about pattern sizes, copy number, mutational history, etc. for tandem repeats has been limited because of the difficulty of detecting them in genomic sequence data. In this paper, we present a new algorithm for finding tandem repeats in DNA sequences without the need to specify either the pattern or pattern size. The algorithm is based on the detection of k-tuple matches. It uses a probabilisitic model of tandem repeats and a collection of statistical criteria based on that model. We demonstrate the algorithm's speed and its ability to detect tandem repeats that have undergone extensive mutational change by analyzing 4 sequences in the 200Kb to 700Kb range.

1 Introduction

DNA sequences are subject to mutational events that transform them over time. One of the less well understood mutational transformations is *tandem duplication* in which a stretch of DNA is duplicated to produce two or more copies, each following the preceding one in a contiguous fashion. For example:

$$...CGG... \longrightarrow ...CGGCGGCGG...$$

(Here the triplet *CGG* has been reproduced twice to form three identical, adjacent copies.) The result of a tandem duplication event is termed a *tandem repeat*. Over time, tandem repeats undergo additional mutations so that typically, only *approximate* tandem copies are present.

*Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, NY 10029-6574. benson@ecology.biomath.mssm.edu; Partially supported by NSF grant CCR-9623532.

RECOMB 98 New York NY USA

Tandem repeats occur frequently, comprising perhaps 10% or more of the human genome. Recently they have been implicated in the causation of several inherited human diseases [23], the so-called trinucleotide repeat diseases, including fragile-X mental retardation [32], Huntington's disease [16], myotonic dystrophy [13], spinal and bulbar muscular atrophy [17], and Friedreich's ataxia [9]. Tandem repeats may play a significant role in gene regulation [19, 14], by interacting with transcription factors or by altering the structure of the *chromatin* [24]. They may promote the evolution of DNA by acting as targets for recombination events [15]. Recently it has been reported that tandem repeats act as protein binding sites [26, 35].

Besides their importance in DNA function and expression, tandem repeats are useful laboratory tools. The number of copies in a tandem repeat is often polymorphic and so is useful in linkage analysis and DNA fingerprinting [10, 33]. Recent studies of allele diversity at tandem repeat loci have provided support for the "Out of Africa" hypothesis of modern human evolution [31, 3].

To date, most of the research on tandem repeats has focused on those with short patterns, probably because such repeats are relatively easy to spot by eye in printed sequences. Repeats with longer patterns are notoriously hard to detect. (Even when the copies are identical. For example see [5] for the 101 bp repeats *undetected* in [15].) Given the importance of known and potential biological roles of tandem repeats and their usefulness for other biological studies, it is essential that *efficient and sensitive* algorithms be developed for detecting unknown tandem repeats in sequence data.

Both exact [18, 30, 4, 29] and heuristic algorithms [22, 7, 27] have been developed for finding tandem repeats. All have critical limitations. For the exact algorithms, the primary limitation is time. With time complexity $O(n^2 \text{ polylog}(n))$, none would be useful for sequences much longer than several thousand bases. (In this paper we report on our analysis of sequences in the 200 kilobase to 700 kilobase range.)

Among the heuristic algorithms, two use methods based on data compression algorithms. One [22] attempts to find "simple sequences," that is, mixtures of fragments that occur elsewhere. Simple sequences may or may not contain tandem repeats and no attempt is made to deduce a pattern. The other algorithm [27] bases the compression on the presence of small preselected patterns (all those of size 1, 2, or 3) and is apparently not readily generalized to longer patterns for which there is an algorithmic need. Both of these

Example 1:

```
*               *               *
A G C T C A C T A G T A C A C A C A C T T A C A C C A G A
C G C T C A C T G G T - - A C A C A C T C A C A C C A G -
T H H H H H H H T H H T T H H H H H H H T H H H H H H H T
```

Example 2:

```
      * *     *     * *                       *
C T A A T G C T A G C A C T A - - A - T G
C T C C T G T T A C A A C T A G T A C T A
H H T T H H T H H T T H H H H T T H T H T
```

Figure 1: Two examples of aligned adjacent sequence from the Human T-cell receptor β chain sequence [28]. The first example is from a tandem repeat with 8 copies. The second is not known to be a tandem repeat.

heuristic methods provide a measure of significance based on the amount of compression. A third heuristic algorithm [7] detects tandem repeats in database scans, but requires that a single pattern size of interest by specified in advance. Thus, to find a range of pattern sizes, the program must be run multiple times, each time with a different pattern size specified.

In this paper, we describe a more flexible algorithm that is not dependent on *apriori* knowledge of the pattern or pattern size. It has already been used as a preprocessor in a new alignment algorithm where tandem duplication augments the standard mutation set of insertion, deletion and substitution [5]. The main features of this algorithm are: 1) it finds tandem repeats without the need to specify the basic pattern size, 2)it detects tandem repeats even when there is a substantial amount of mutational difference between adjacent copies and 3) it finds a smallest consensus unit for the tandem repeat and aligns the tandem repeat with that consensus.

A number of ideas incorporated into our algorithm have been utilized in earlier homology detection programs [25, 2], but, ours differs in several ways. First, we are *not* looking for the highest scoring homologous regions, but rather tandem repeats which are often hidden in larger homologous regions or which may fall well below the level of significance required for other programs to report a match. Second, our program is designed more for search in a single sequence than for database scans. Third, the output from our program is designed to give insight into the history of the mutations that could have produced the tandem repeats, thus providing a potentially valuable tool for phylogenetic research [3].

The remainder of this paper is organized as follows. In the Methods section we present a probabilistic model of tandems repeats, a set of criteria that guide the selection process of our algorithm and an algorithm overview. In the Discussion section, we present some examples of newly found tandem repeats. Finally, in the Conclusion we describe directions for future research.

2 Methods

One difficulty in dealing with tandem repeats is accurately defining them. The best definition is that a tandem repeat is a sequence resulting from a tandem duplication event that happened *in the past*. The problem is, we usually cannot know the history of a sequence and so must judge a possible tandem repeat by the sequence appearance today. Our approach has been to define tandem repeats with a probabilistic model and to draw inferences from that model which facilitate detection of the repeats.

2.1 Probabilistic Model of Tandem Repeats

Our model is a Bernoulli process – essentially random coin tossing – which describes *aligned adjacent copies* of a tandemly repeated pattern. Suppose (Fig. 1) that we take two adjacent stretches of nucleotide sequence and align them, one above the other. Below each column of the alignment we write an H if the characters match and a T otherwise. In essence, we convert the alignment into a sequence of heads and tails – a coin toss sequence. Fig. 1 shows two examples drawn from the Human β T cell receptor locus sequence [28]. The first example consists of 2 adjacent copies from a tandem repeat with 8 copies. The copies are only approximate, but most bottom row characters are H. The second example consists of adjacent sequences not known to be part of a tandem repeat. Note the much higher frequency of T in this example.

Our probabilistic model incorporates the idea illustrated above and in fact reverses the process. We choose a target matching probability p_M for aligned characters in *aligned adjacent copies* within a tandem repeat. Then we let $P(H) = p_M$, where $P(H)$ is the probability of heads in our Bernoulli process. Using p_M, we determine a collection of *statistical criteria* that our algorithm uses to find *candidates*. The target p_M serves as a type of upper bound. It describes, on average, the *most mutated* tandem repeats we want to find. A second critical parameter in our model is p_I, the target probability of an indel. The use of both p_m and p_I is described further below.

2.2 k-tuple Matches and Statistical Criteria

The basic premise behind our algorithm is that given two adjacent approximate copies within a tandem repeat, there will be many characters in the first copy that match corresponding characters in the second copy. Our algorithm works by finding the matches at a common distance. For reasons of efficiency, instead of looking for all the matches, we look only for runs of matches, which we call *k-tuple matches*. A k-tuple is merely a window of k consecutive characters from the sequence. Matching k-tuples are two windows with iden-

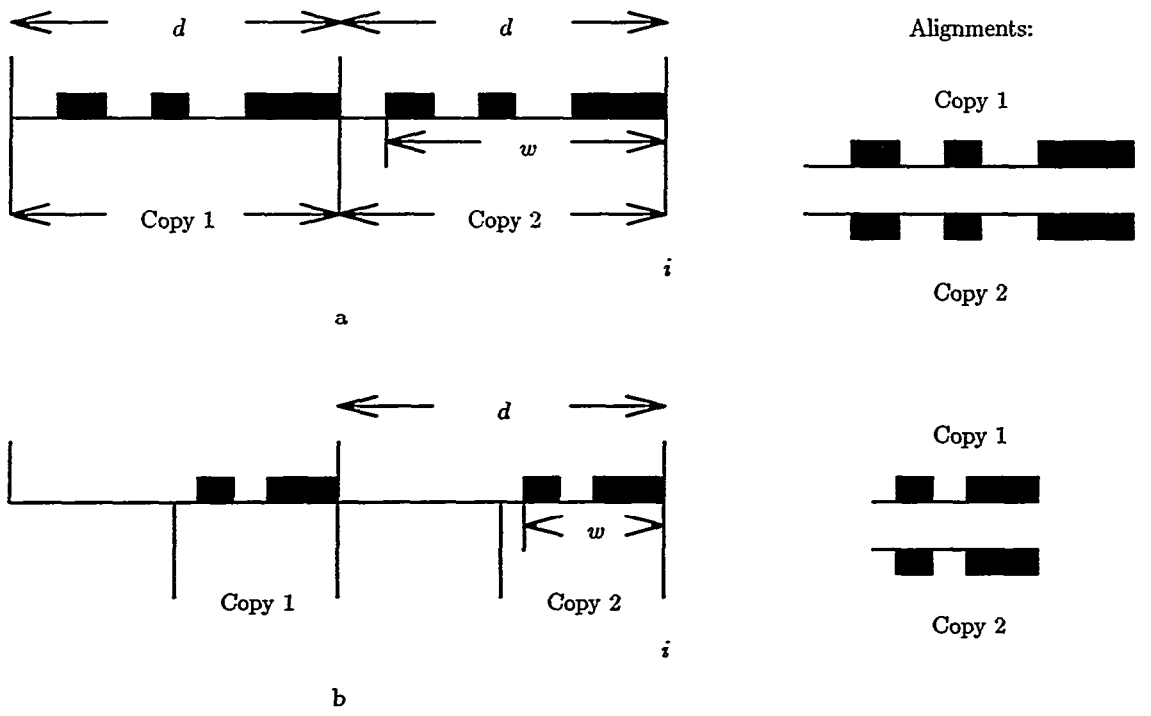

Figure 2: When a sufficient number of matches are detected, we must distinguish between a) a possible tandem repeat (matches spread over the interval of size d) and b) a possible non-tandem, direct repeat (all the matches concentrated on the right). k-tuple or larger matches in the alignment of copies 1 and 2 are indicated by the shaded boxes.

tical contents. The efficiency consideration is important. For sequences in which every character is equally likely, a single character (*e.g.* A) will find a match, on average, every four bases. Most of these matches would not indicate a tandem repeat and just looking at each match would lead to a $O(m^2)$ algorithm, for sequences of length m. This is much too slow for large (100kb to 1Mb range) sequences. On the other hand, with 5-tuples, a match will occur, on average, every 1000 bases and these will often indicate a true repeat. Importantly, looking at all 5-tuple matches is approximately 250 times faster than looking at single character matches.

If the window contents of matching k-tuples are aligned, then, in our coin toss model they appear as a *run of k heads*. Our first three criteria are based on statistical distributions of runs of heads. The distributions depend on p_M, k and the pattern length n. They are: 1) **Sum of Heads** – *the number of heads when we count only those occurring in runs of length k or longer* – used to select candidates with enough matches, 2) **Apparent Size** – *the distance between the first head in the first run of k heads and the last head in the last run of k heads* – used to screen out direct but not tandem repeats, and 3) **Waiting Time** – *the number of coin tosses until the first run of k heads* – used to pick a good window size for detecting patterns of size n.

Our last criteria deals with a statistical distribution for indels as specified by p_I: 4) **Random Walk** – *the maximum displacement of a random walk from the origin* – used to determine the range of distances between matching k-tuples due to insertions and deletions. Below we describe the criteria in more detail.

Sum of Heads. For each pattern size n, we want to look for enough matches to convince us that we have a good candidate. How many matches is enough? Let R_{n,k,p_M} be a *random variable corresponding to the sum of heads*. The distribution of this random variable is well approximated by the normal distribution and we have previously shown that the exact mean and variance of R_{n,k,p_M} can be calculated in constant time [6]. From the normal distribution we determine the largest number, x, such that 95% of the time R_{n,k,p_M} equals or exceeds x. We use x as our **sum-of-heads criterion**. For example, if $p_M = .75$, $k = 5$ and n=100, then 95% of the time R_{n,k,p_M} will be at least 26. Put another way, if a duplicated pattern has length 100 and aligned copies are expected to match in 75 positions, then by counting only matches that fill a window of length 5, we expect to count at least 26 matches 95% of the time.

Apparent Size. Once we have enough matches, we must distinguish a tandem repeat from two copies of a (non-tandem) direct repeat. Non-tandem direct repeats are separated by intervening sequence which is not repeated. The latter should have the matches clustered on the right end of the matching distance, whereas a tandem repeat should have the matches distributed throughout (see Fig. 2.1). We judge this by looking at the *apparent size* w. If w is too small, then we assume the repeat is not a tandem repeat or that we haven't yet seen enough of it to be convinced.

We estimate the distribution of w by simulation. (It is conditional on first meeting the sum-of-heads criterion.) From the distribution, we determine the maximum number y such that 95% of the time w is greater than y. We use y as our **apparent-size-criterion**. For example, if $p_m = .80$, $k = 5$ and $n = 100$, then we expect that 95% of the time the apparent size will be greater than 37 .

Waiting time. Increasing tuple size dramatically decreases

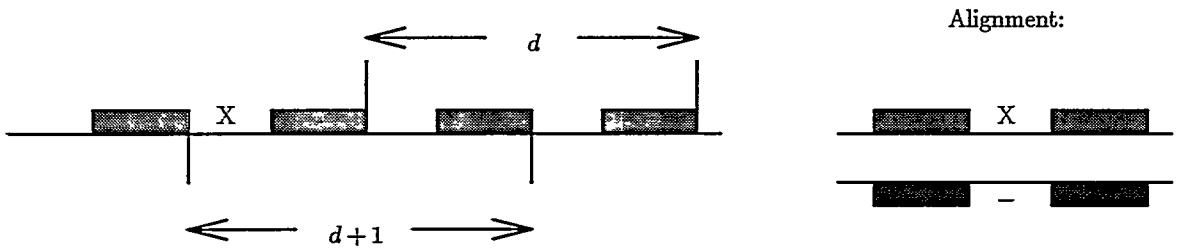

Figure 3: Insertions and deletions change the distance between exact matches. The inserted character X above causes the pair of matching k-tuples to be separated by distance $d+1$ while another pair is separated only by distance d.

the number of random tuple matches (and the algorithm running time). For example, if the nucleotides occur with equal frequency, then increasing the tuple size by one increases the average distance between randomly matching tuples by a factor of four. Thus if $k = 5$, the average distance between random matches is 1kb, but if $k = 7$, the average distance is 16Kb. On the other hand, a large k-tuple match may not occur in a small duplicated pattern. But, decreasing the tuple size increases the probability that the duplicates will contain a k-tuple match. For example, if $p_M = .75$ and $k = 5$ then two copies of a pattern of size 10 will contain a *single* matching 5-tuple only about 54% of the time. But, with $k = 3$, that probability increases to about 91% of the time. Thus if we restrict the algorithm to $k = 5$, duplicated patterns of size 10 will often be missed.

We balance the two consequences of tuple size by employing different tuple sizes to detect different pattern sizes. From our coin toss model, we use the distribution of the *waiting time* to determine appropriate tuple sizes. The exact distribution of waiting times is given by a simple recursive formula [1]. We require that each pattern size exhibit a minimum of $k+1$ matches for the sum-of-heads criterion when using a tuple of size k. In reality, selecting an efficient tuple size is only a problem for patterns of length 30 or less.

Random Walk. Indels change the distance between matching characters in adjacent copies (see Fig. 2.2). How much change should we expect? As a first approximation, we model indels as single nucleotide events occuring with probability p_I in the aligned copies. We assign equal probability for insertion or deletion. That is, a distance d between matching k-tuples changes either to $d+1$ or $d-1$ with probability 1/2. Viewed in this way, we can treat the distance change caused by the indels as the problem of repeated reflections in a symmetric random walk [11]. We are interested in the distribution of *the maximum displacement of the random walk from the origin.* Let RW_n be the random variable denoting the position furthest from the origin in a random walk of n steps. It can be shown that, 95% of the time, RW_n ranges approximately between $2.3\sqrt{n}$ and $-2.3\sqrt{n}$. Since n is not fixed, but is itself a random variable with expectation $E(n) = p_I d$ and since $E(RW_n) = 0$, we have that, 95% of the time, the maximum displacement from the origin ranges between $2.3\sqrt{p_I d}$ and $-2.3\sqrt{p_I d}$. We use $r = 2.3\sqrt{p_I d}$ as our **random-walk-criterion**. For example if $p_I = .2$ and $d = 100$ the neighborhood of distances predicted by the model is 100 ± 10.3.

2.3 Algorithm Overview

Our algorithm finds tandem repeats by observing a collection of matching k-tuples at a common distance in a common region of the DNA sequence. The basic features are illustrated in Fig. 3. (For the following description, we assume that we use only one tuple size.) Let the sequence S have length n. We select a small integer k, for example $k = 5$, and then maintain a list representing all 4^k-tuples (strings of length k) in Σ^k where $\Sigma = \{A, C, G, T\}$. This collection of strings constitutes our *probes*. Next, we slide a window of size k across the sequence, and determine the probe at each position i in S. For each probe p, we maintain a history list H_p of the positions at which it occurs. Since there is one probe per position, these lists are easily maintained within a single array of size n.

Once a position i has been added to H_p, we scan H_p for *all* earlier occurrences of p. For each earlier occurrence, say at j, we calculate the distance $d = i - j$ between the indices i and j. For every distance d, we maintain a Distance list D_d. This list stores the positions and total of all detected matches *in a sliding window of size d.*

We update D_d by adding the position i to the list of positions and increasing the total by the number of new matches contributed by the tuple match at i and j. We discard matches detected before position $j+1$ and reduce the total by the corresponding number of matches. After D_d has been updated, the criteria for a candidate tandem repeat are tested. Since insertions and deletions change the distance, we use our **random-walk-criterion** r to select a neighborhood of distances $D_{d \pm h}$, for $h = 0, 1, \ldots, r$ to include in the test. The remaining criteria ask 1) does this collection of distance lists contain enough matches (**sum-of-heads criterion**) and 2) are the matches spread out enough (**apparent-size criterion**)? If the answers are both yes, a candidate pattern is drawn from the sequence and aligned with surrounding sequence using wraparound dynamic programming (WDP) [21, 12]. Finally, if at least two copies of the pattern are found in the alignment, the tandem repeat is reported in the output.

3 Results

We have analyzed 4 genomic sequences with our program: yeast chromosomes 1 [8], 6 [34] and 8 [20] and the human β T cell receptor locus sequence [28]. In the analysis, we have looked for all pattern sizes between 10 and 500 bases.

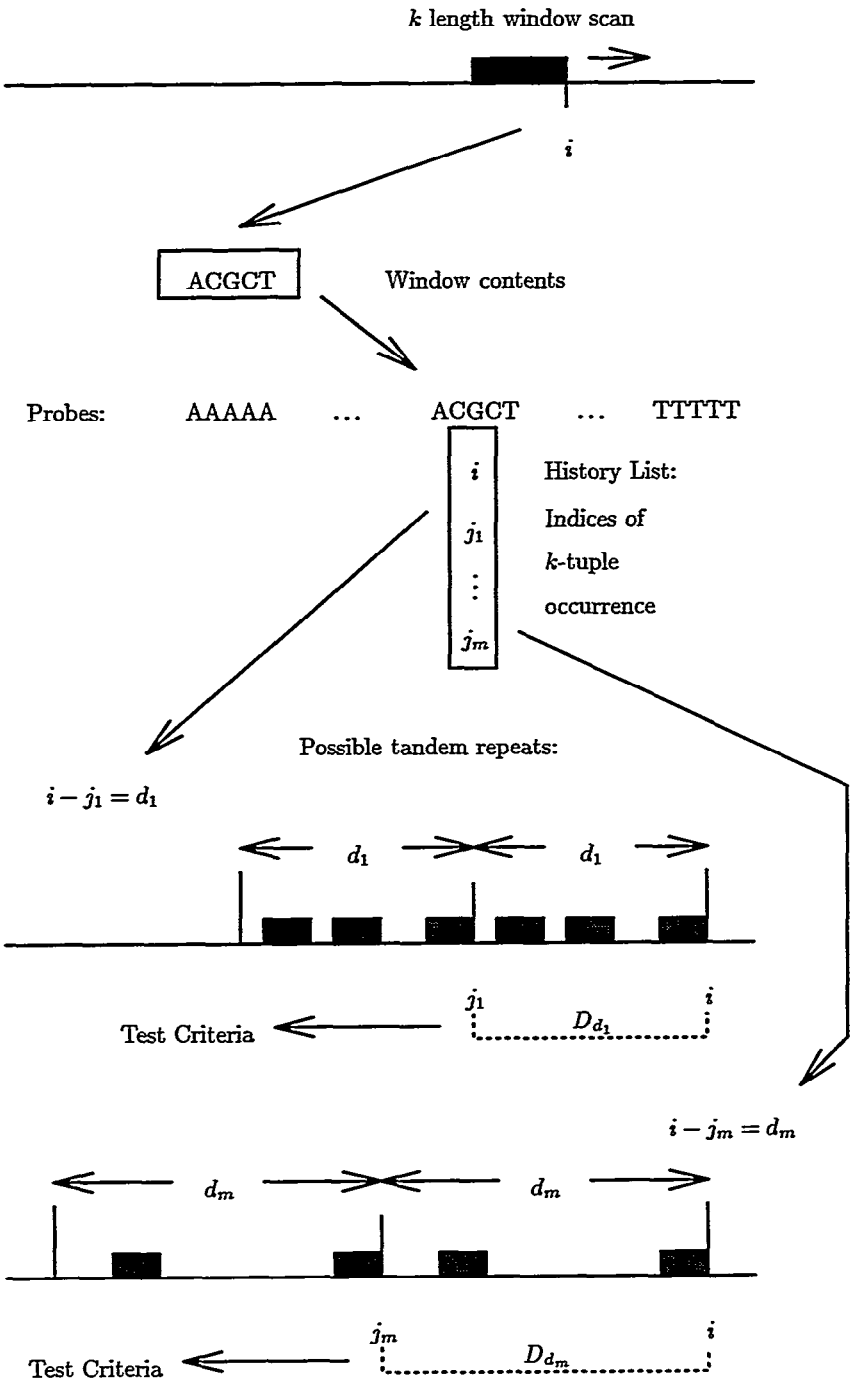

Figure 4: Tandem repeats are detected by scanning the sequence with a small window and then determining the distance between exact matches.

We performed two searches on each sequence, one using parameter values $p_M = .75$ and $p_I = .20$ and the second using values $p_M = .80$ and $p_I = .10$. While the two searches proceed at much different speeds, the detected repeats are nearly identical. Table 2 lists sample running times of the program on the analyzed sequences. Tables 5, 4, 5, and 6 list the tandem repeats found. Due to space limitations, we discuss briefly only the Human T-cell receptor and yeast chromosome 1 sequences below.[1]

Human β T cell receptor locus sequence. (See Table 6.) Several large pattern tandem repeats were found, including those with patterns sizes in the range 39 to 60. Two in particular, the 60 base pattern with approximately 16 copies and the 39 base pattern with 7.7 copies display substantial amounts of mutation between adjacent copies, illustrating the power of our program.

Yeast Chromosome 1. (See table 5.) Several interesting aspects of chromosome 1 can be observed in this data. Most of the tandem repeats found have small pattern sizes (14 - 28 bases). Only two large pattern repeats were found. One has pattern size 48 with 7.7 copies. The other has pattern size 135 and is part of the yeast flocculation gene *FLO1*. The gene itself occurs on the right arm and contains 18 copies of the 135 base pattern. The other occurrence of the 135 base pattern is on the left arm and appears to be part of a duplicated fragment of the *FLO1* gene. (A similar fragment occurs on chromosome 8.) Patterns of mutation in the tandem copies within the *FLO1* gene suggest that the copies were produced by tandem duplication events. *This may be the first example of a large protein which has evolved in this way.*

The *FLO1* gene and its homolog occur on opposite ends of the chromosome adjacent to a repeated structure that has been designated W'. Interestingly, within the *FLO1* gene and its homolog, there are clusters of 3 other tandem repeats with sizes 27, 21 and 15. An additional cluster occurs to the left of the *FLO1* homolog, indicating that part of the *FLO1* gene is repeated a *third* time. We designate these clusters from the left end of the chromosome, Clusters 1, 2, and 3. (See Table 3.)

Within Clusters 1 is a fourth tandem repeat with pattern size 48. This same pattern occurs in the other two clusters, but was not detected because in each of those clusters there are only 1.7 copies of the pattern. Additionally, within Cluster 1 there is a fragment of less than one copy of the 135 base pattern. Significantly, the number of copies of every one of the pattern sizes varies among the three clusters, implying that duplication or excision (deletion of copies) events have occurred since the time when the separate clusters were incorporated into the chromosome.

[1]The complete sequences of the yeast chromosomes were obtained via ftp from ftp.ebi.ac.uk directory pub/databases/yeast in files chri_230209.ascii, chrvi_270148.ascii and chrviii_562638.ascii. The human T-cell receptor sequence was obtained from GenBank. All indexing in this paper is relative to the sequences in these files. Corresponding data file accession numbers for these sequences are: yeast chromosome 1: U12980, L20125, L05146, L22015, L28920; yeast chromosome 6: D50617; yeast chromosome 8: U11583, U11582, U11581, U10555, U10400, U10399, U00062, U00061, U10556, U00060, U00059, U10398, U10397, U00027, U00028, U00030, U00029; Human T-cell receptor: L36092.

Period Size	Cluster		
	1	2	3
27	2.2	2.2	3.3
21	3.2	4.3	3.0
48	7.7	1.7	1.7
15	9.1	9.1	5.5
135	0.7	13	18

Table 1: Varying copy numbers in the three similar tandem repeat clusters found in yeast chromosome 1. Clusters are numbered from left arm to right arm along the chromosome. For locations, see Table 5 in the appendix.

4 Conclusion

In this paper, we have presented a new algorithm for finding tandem repeats in DNA sequences without the need to specify either the pattern or pattern size. The algorithm is based on the detection of k-tuple matches. It uses a probabilisitic model of tandem repeats and a collection of statistical criteria based on that model. We have demonstrated the speed of the algorithm and its ability to detect tandem repeats that have undergone extensive mutational change by analyzing 4 sequences in the 200Kb to 700Kb range. Several avenues for future research are raised by this work, including methods to estimate the probability of a tandem repeat occurring at random and algorithms to determine plausible mutational histories for tandem repeats.

Statistical Issues. We have yet to develop a good probability measure for the tandem repeats found by our algorithm. For now, we use simulation to determine common high alignment scores when running the program on randomly generated sequences with the same nucleotide frequency as the analyzed sequence. The repeats reported in the appendix exceed those scores. For small patterns, we could use the estimate of significance from [7], but those estimates are too high in this application because they apply to tandem repeats of one pattern size only, rather than for the range of sizes considered here.

One of the difficulties of getting a reliable statistical estimate is the local variation in nucleotide frequency. Some parts of a sequence consist almost entirely of two bases and others consist of three or four. A useful statistical measure would have to account for the fact that with fewer bases it is more likely for an apparent tandem repeat to occur by chance. Thus it may not be suitable to use a measure derived from a random sequence if the frequencies of the nucleotides in any particular genomic sequence locally vary more than would those in a randomly generated sequence. Another difficulty is that significance based on alignment score takes no account of copy number. If for example, a pattern of length 16 appears twice, is that more or less significant than a pattern of length 8 appearing 4 times?

Mutational History. Analyzing the mutational history of tandem repeats involves using the pattern of mutations among adjacent copies to describe the interwoven progression of substitutions, indels and duplication/excision events in such a way as to minimize the number of identical mutations that arise independently. For example, if 3 out of

Sequence	Length (bases)	Running Times	
		$P_M = .75$ $P_I = .20$	$P_M = .80$ $P_I = .10$
Yeast Chromosome 1	230,209	1 min 19 sec	7 sec
Yeast Chromosome 8	562,638	2 min 36 sec	13 sec
Human β T cell receptor locus sequence	684,973	3 min 34 sec	20 sec

Table 2: Running times of program on selected sequences using a Silicon Graphics O2 RS10000.

10 copies show an $A \rightarrow T$ mutation in the same position, it is more likely that the mutation arose only once and that the original mutation was duplicated than that it arose independently 2 or 3 times.

5 Acknowledgement

The author would like to thank Xiaoping Su for his help in analyzing the sum-of-heads criterion, Astrid Jervis for her help in simulating many of the statistical measures and examining the program output, and Lan Dong for her help with some of the programming and examining the output. Thanks also to Mike Waterman, Richard Arratia and Rolf Backofen for helpful discussions.

References

[1] S. Aki, H. Kuboki, and K. Hirano. On discrete distributions of order k. *Ann. Inst. Statist. Math.*, 36:431–440, 1984.

[2] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[3] J. Armour, T. Anttinen, C. May, E. Vega, A. Sajantila, J. Kidd, K. Kidd, J. Bertranpetit, S. Pääbo, and A. Jeffreys. Minisatellite diversity supports a recent African origin for modern humans. *Nature Genetics*, 13:154–160, 1996.

[4] G. Benson. A space efficient algorithm for finding the best non-overlapping alignment score. In M. Crochemore and D. Gusfield, editors, *Proc. 5th annual Symp. on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, volume 807, pages 1–14. Springer-Verlag, 1994.

[5] G. Benson. Sequence alignment with tandem duplication. *J. Computational Biology*, 4:351–367, 1997.

[6] G. Benson and X. Su. On the distribution of k tuple matches for sequence homology: a constant time exact calculation of the variance. *J. Computational Biology* in press.

[7] G. Benson and M. Waterman. A method for fast database search for all k-nucleotide repeats. *Nucleic Acids Research*, 22:4828–4836, 1994.

[8] H. Bussey, D. B. Kaback, W. Zhong, D. T. Vo, M. W. Clark, N. Fortin, J. HallB. F. Ouellette, T. Keng, A. B. Bartoon, S. Yuping, C. J. Davies, and R. K. Storms. The nucleotide sequence of Chromosome I from Saccharomyces Cerevisiae. *Proc Natl Acad Sci U S A*, 92:3809–3813, 1995.

[9] V. Campuzano, L. Montermini, M.D. Molto, L. Pianese, and M. Cossee. Friedreich's ataxia: Autosomal recessive disease caused by an intronic gaa triplet repeat expansion. *Science*, 271:1423–1427, 1996.

[10] A. Edwards, H. Hammond, L. Jin, C. Caskey, and R. Chakraborty. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, 12:241–253, 1992.

[11] W. Feller. *An introduction to probability theory and its applications*, volume I. John Wiley & Sons, 3rd edition, 1968.

[12] V. Fischetti, G. Landau, J. Schmidt, and P. Sellers. Identifying periodic occurrences of a template with applications to a protein structure. In A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, editors, *Proc. 3rd annual Symp. on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, volume 644, pages 111–120. Springer-Verlag, 1992.

[13] Y.-H. Fu, A. Pizzuti, J. Fenwick, R.G.Jr.and King, S. Rajnarayan, P.W. Dunne, J. Dubel, G.A. Nasser, T. Ashizawa, P. DeJong, B. Wieringa, R. Korneluk, M.B. Perryman, H.F. Epstein, and C.T. Caskey. An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science*, 255:1256–1258, 1992.

[14] H. Hamada, M. Seidman B. Howard, and C. Gorman. Enhanced gene expression by the poly(dT-dG) poly(dC-dA) sequence. *Molecular and Cellular Biology*, 4:2622–2630, 1984.

[15] L. Hellman, M. Steen, M. Sundvall, and U. Pettersson. A rapidly evolving region in the immunoglobulin heavy chain loci of rat and mouse: postulated role of $(dC\text{-}dA)_n$ $(dG\text{-}dT)_n$ sequences. *Gene*, 68:93–100, 1988.

[16] Huntington's disease collaborative research group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72:971–983, 1993.

[17] A. La Spada, E. Wilson, D. Lubahn, A. Harding, and K. Fischbeck. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, 352:77–79, 1991.

[18] G. Landau and J. Schmidt. An algorithm for approximate tandem repeats. In *Proc. 4th Annual Symp. on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, volume 648, pages 120–133. Springer-Verlag, 1993.

[19] Q. Lu, L. Wallrath, H. Granok, and S. Elgin. $(CT)_n$ $(GA)_n$ repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the *Drosophila hsp26* gene. *Molecular and Cellular Biology*, 13:2802–2814, 1993.

[20] Johnston M., Andrews S., Brinkman R., Cooper J., Ding H., Dover J., Du Z., Favello A., Fulton L., Gattung S., Geisel C., Kirsten J., Kucaba T., Hillier L., Jier M., Johnston L., Langston Y. Latreille P., Louis E.J., Macri C., Mardis E., Menezes S., Mouser L., Nhan M., Rifkin L., Riles L., St.Peter H., Trevaskis E. Vaughan K., Vignati D., Wilcox L., Wohldman P., Waterston R. Wilson R., , and Vaudin M. Complete nucleotide sequence of *s. cerevisiae* Chromosome VIII. *Science*, 265:2077–2082, 1994.

[21] W. Miller and E. Myers. Approximate matching of regular expressions. *Bulletin of Mathematical Biology*, 51:5–37, 1989.

[22] A. Milosavljević and J. Jurka. Discovering simple DNA sequences by the algorithmic significance method. *CABIOS*, 9:407–411, 1993.

[23] S. Panzer, D. P. Kuhl, and C. Caskey. Unstable triplet repeat sequences: A source of cancer mutations? *Stem Cells*, 13:146–157, 1995. See also http://uranus.gmu.edu:443/TriNuc/chart.html.

[24] M. Pardue, K. Lowenhaupt, A. Rich, and A. Nordheim. $(dC\text{-}dA)_n$ $(dG\text{-}dT)_n$ sequences have evolutionarily conserved chromosomal locations in *Drosophila* with implications for roles in chromosome structure and function. *The EMBO Journal*, 6:1781–1789, 1987.

[25] W. Pearson and D. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acac. Sci. USA*, 85:2444–2448, 1988.

[26] R. Richards, K. Holman, S. Yu, and G. Southerland. Fragile X syndrome unstable element, $p(CCG)_n$, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. *Hum. Mol. Genet.*, 2:1429–1435, 1993.

[27] E. Rival, O. Delgrange, J.-P. Delahaye, M. Dauchet, M.-O. Delorme, A. Hénaut, and E. Ollivier. Detection of significant patterns by compression algorithms: the case of approximate tandem repeats in DNA sequences. *CABIOS*, 13:131–136, 1997.

[28] L. Rowan, B. Koop, and L. Hood. The complete 685-kilobase DNA sequence of the Human β T cell receptor locus. *Science*, 272:1755–1768, 1996.

[29] J.P. Schmidt. All highest scoring paths in weighted grid graphs and its application to finding all approximate repeats in strings. In *Third Israel Symposium on Theory of Computing and Systems*, pages 67–77. IEEE Computer Society Press, 1995.

[30] S.K.Kannan and E.W.Myers. An algorithm for locating regions of maximum alignment score. In *Proc. 4th Annual Symp. on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, volume 648, pages 74–86. Springer-Verlag, 1993.

[31] S.A. Tishkoff, E. Dietzsch, W. Speed, A.J. Pakstis, and J.R. Kidd. Global patterns of linkage disequilibrium at the cd4 locus and modern human origins. *Science*, 271:1380–1387, 1996.

[32] A. Verkerk, M. Pieretti, J. Sutcliffe, Y. Fu, D. Kuhl, A. Pizzuti, O. Reiner, S. Richards, M. Victoria, F. Zhang, B. Eussen, G. van Ommen, A. Blonden, G. Riggins, J. Chastain, C. Kunst, H. Galjaard, C. Caskey, D. Nelson, B. Oostra, and S. Warren. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, 65:905–914, 1991.

[33] J. Weber and P. May. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.*, 44:388–396, 1989.

[34] Murakami Y., Naitou M., Hagiwara H., Shibata T.and Ozawa M., Sasanuma S.I., Sasanuma M., Tsuchiya Y., Soeda E., Yokoyama K., Yamazaki M., Tashiro H., and Eki T. Analysis of the nucleotide sequence of Chromosome VI from Saccharomyces cerevisiae. *Nature Genet.*, 10:281–268, 1995.

[35] H. Yee, A. Wong, J. van den Sande, and J. Rattner. Identification of novel single stranded $d(TC)_n$ binding proteins in several mamalian species. *Nucleic Acids Res.*, 19:949–953, 1991.

Appendix – Output for the tested sequences

The tables below list the tandem repeats found in the analyzed sequences. The information presented is 1) beginning and ending indices of the repeats, 2) the consensus sequence size, 3) the number of copies aligned against the consensus, 4) the most common distance between matching characters which can be interpreted as the true pattern size, 5) the percentage of matches and 6) of indels when comparing all adjacent pattern copies, and 7) the alignment score which was used to screen out randomly occurring candidates.

Yeast Chromosome 1						
Indices	Consensus Size	Copy Number	Period Size	Percent Matches	Percent Indels	Score
11876–11935	27	2.2	**27**	93	0	106
12260–12327	21	3.2	**21**	82	0	87
12470–12839	48	7.7	**48**	91	0	600
13000–13136	15	9.1	**15**	72	9	119
14791–14821	13	2.4	**13**	94	0	55
24308–24367	27	2.2	**27**	93	0	106
24690–24780	21	4.3	**21**	79	5	121
25165–25301	15	9.1	**15**	72	9	119
25395–27148	134	13.0	**135**	91	2	2414
99945–99976	14	2.3	**14**	100	0	64
100371–100414	18	2.4	**18**	88	0	67
101471–101511	15	2.7	**15**	88	0	61
190128–190161	14	2.5	**14**	95	4	63
198835–198864	11	2.7	**11**	94	0	53
204224–206643	135	17.9	**135**	88	3	2690
206748–206830	15	5.5	**15**	84	8	114
207227–207288	21	3.0	**21**	82	0	89
207614–207702	27	3.3	**27**	88	0	136
229752–229807	15	3.7	**15**	85	0	84
229947–229987	11	3.8	**11**	87	9	58

Table 3: Tandem repeats detected in Yeast Chromosome 1. Period sizes in boldface denote the three repeated clusters denoted from the top Clusters 1, 2 and 3.

Yeast Chromosome 8						
Indices	Consensus Size	Copy Number	Period Size	Percent Matches	Percent Indels	Score
5–35	13	2.4	13	100	0	62
144–175	10	3.2	10	95	0	57
1651–1937	36	8.0	36	87	2	330
33591–33703	18	6.3	18	69	24	107
49307–49566	57	4.3	60	65	11	153
58094–58123	15	2.0	15	100	0	60
242207–242238	13	2.5	13	94	0	57
282712–282820	27	4.0	27	76	12	116
316465–316497	15	2.2	15	100	0	66
373102–373142	15	2.7	15	100	0	82
413602–413633	13	2.5	13	94	0	57
462001–462036	13	2.8	13	95	0	65
526224–527280	135	7.8	135	88	0	1619
527380–527473	15	6.3	15	76	4	99
527862–527952	21	4.3	21	86	5	116
548600–548829	114	2.0	114	100	0	460
560412–560836	36	11.8	36	83	5	432
562312–562343	10	3.2	10	95	0	57
562451–562637	13	14.7	13	86	10	273

Table 4: Tandem repeats detected in Yeast Chromosome 8.

Yeast Chromosome 6						
Indices	Consensus Size	Copy Number	Period Size	Percent Matches	Percent Indels	Score
892–1265	36	10.4	36	82	4	394
16275–16309	17	2.1	17	100	0	70
92376–92567	30	6.4	30	81	4	208
115125–115192	30	2.3	30	97	0	129
168312–168361	25	2.0	25	100	0	100
178016–178298	141	2.0	141	99	0	559
186748–186778	12	2.6	12	94	0	55
202601–202636	18	2.0	18	94	0	65
226600–226634	10	3.6	10	92	3	58
227333–227366	12	2.8	12	90	0	54
270096–270148	13	4.0	13	95	2	92

Table 5: Tandem repeats detected in Yeast Chromosome 6.

Human β T cell receptor locus sequence						
Indices	Consensus Size	Copy Number	Period Size	Percent Matches	Percent Indels	Score
8764–8794	15	2.1	15	100	0	62
12586–13535	59	15.8	60	71	9	745
21863–21905	16	2.8	14	86	13	69
48825–48928	52	2.0	52	94	0	187
84876–84913	17	2.2	17	95	0	69
86699–86743	21	2.1	21	88	8	71
121522–121659	65	2.1	65	98	0	269
133305–133357	21	2.5	21	84	6	73
140508–140697	20	9.6	20	83	7	276
149566–149694	40	3.2	40	93	0	230
154085–154125	20	2.0	20	95	0	75
178765–178803	18	2.1	19	90	4	64
193559–193619	27	2.3	27	88	0	94
197973–198134	28	5.7	28	87	5	230
202684–202755	32	2.2	32	78	9	92
216002–216043	19	2.2	19	95	0	77
255371–255434	30	2.2	29	91	8	118
344711–344810	49	2.0	49	100	0	200
376274–376322	22	2.2	22	92	3	84
403499–403549	24	2.1	24	85	0	74
410172–410470	38	7.7	39	79	9	356
516196–516237	18	2.3	18	87	4	63
614493–614565	34	2.1	34	97	0	139
653054–653277	29	8.0	29	81	9	307
684213–684417	30	7.0	30	90	7	352

Table 6: Tandem repeats detected in Human β T cell receptor locus sequence.