Tandem repeats finder: a program to analyze DNA sequences

Gary Benson*

Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, NY 10029-6574, USA

Received September 10, 1998; Revised and Accepted November 12, 1998

ABSTRACT

A tandem repeat in DNA is two or more contiguous, *approximate* copies of a pattern of nucleotides. Tandem repeats have been shown to cause human disease, may play a variety of regulatory and evolutionary roles and are important laboratory and analytic tools. Extensive knowledge about pattern size, copy number, mutational history, etc. for tandem repeats has been limited by the inability to easily detect them in genomic sequence data. In this paper, we present a new algorithm for finding tandem repeats which works without the need to specify either the pattern or pattern size. We model tandem repeats by percent identity and frequency of indels between adjacent pattern copies and use statistically based recognition criteria. We demonstrate the algorithm's speed and its ability to detect tandem repeats that have undergone extensive mutational change by analyzing four sequences: the human frataxin gene, the human β T cell receptor locus sequence and two yeast chromosomes. These sequences range in size from 3 kb up to 700 kb. A World Wide Web server interface at c3.biomath.mssm.edu/trf.html has been established for automated use of the program.

INTRODUCTION

DNA molecules are subject to a variety of mutational events. One of the less well understood is *tandem duplication* in which a stretch of DNA, which we call the *pattern*, is converted into two or more *copies*, each following the preceding one in a contiguous fashion. For example we could have

. . . TCGGA . . . → . . . TCGGCGGCGGA . . .

in which the single occurrence of triplet CGG has been transformed into three identical, adjacent copies. The result of a tandem duplication event is termed a *tandem repeat*. Over time, individual copies within a tandem repeat may undergo additional, uncoordinated mutations so that typically, only *approximate* tandem copies are present.

Tandem repeats are presumed to occur frequently in genomic sequences, comprising perhaps 10% or more of the human genome. But, accurate characterization of the properties of tandem repeats has been limited by the inability to easily detect them. In recent years, the discovery of the trinucleotide repeat diseases has piqued interest in tandem repeats. These diseases, including fragile-X mental retardation (1), Huntington's disease (2), myotonic dystrophy (3), spinal and bulbar muscular atrophy (4) and Friedreich's ataxia (5), are the result of a dramatic increase in the number of copies of a trinucleotide pattern. In afflicted individuals, the copy number has been amplified from the normal range of tens of copies to hundreds or thousands, resulting in the disease. It has been suggested that the repeats themselves produce unusual physical structures in the DNA, causing polymerase slippage and the resulting amplification (6,7).

A more salubrious potential role for tandem repeats is gene regulation, in which the repeats may interact with transcription factors, alter the structure of the chromatin or act as protein binding sites (8–12). Tandem repeats have an apparent function in development of immune system cells. Breakpoints for immunoglobulin heavy chain switch recombination occur within tandem repeats preceding the heavy chain constant region genes (13). Because the number of copies in any specific tandem repeat is often polymorphic in the population, tandem repeats have proven useful in linkage analysis and DNA fingerprinting (14,15). Recent studies of allele diversity at tandem repeat loci have provided support for the 'Out of Africa' hypothesis of modern human evolution (16,17).

To date, much of the research on tandem repeats has focused on those with short patterns (2–5 nt), presumably because such repeats are relatively easy to spot by eye in printed sequences. Repeats with long patterns (sometimes called variable number of tandem repeats or VNTRs) are notoriously harder to detect [even when the copies are identical, for example see Benson (18) for the 101 bp repeats *undetected* in Hellman *et al*. (19), a paper on the role of tandem repeats as hot spots for recombination]. Given the importance of known and potential biological roles for tandem repeats and their usefulness in other biological studies, it seemed essential to us to develop an *efficient* and *sensitive* algorithm for detecting these repeats so that they may receive further study.

A number of algorithms already exist which either directly or indirectly detect tandem repeats. All suffer from significant limitations. One group of algorithms is based on computing alignment matrices (20–22). Their primary limitation is excessive running time. The best algorithm in this group (22) has time complexity $O[n^2 \text{polylog}(n)]$ for a sequence of length n and would not be useful for sequences much longer than several thousand

*Tel: +1 212 241 5777; Fax: +1 212 860 4630; Email: benson@ecology.biomath.mssm.edu bases. (In this paper we report on our analysis of sequences up to 700 kb in length.)

Another group of algorithms finds tandem repeats indirectly using methods from the field of data compression. An algorithm by Milosavljevic and Jurka (23) detects ‘simple sequences’, i.e. mixtures of fragments that occur elsewhere. Simple sequences may or may not contain tandem repeats and this algorithm makes no attempt to deduce a repeated pattern. An algorithm by Rivals *et al*. (24) bases the compression on the presence of small preselected patterns (all those of size 1–3) and is not readily generalized to longer patterns for which there is an algorithmic need. To their credit, both of these methods provide a measure of statistical significance based on the amount of compression.

Another collection of algorithms aim more directly at finding tandem repeats. Of these, one exact algorithm (25) is limited by its definition of approximate patterns, requiring that two copies differ either by *k* or fewer substitutions (Hamming distance) or by *k* or fewer substitutions and indels (unit cost edit distance). Besides treating substitutions and indels as equals, the requirement for a fixed number of differences rather than a *percentage* difference is unsatisfactory. Any fixed number of differences suitable for small patterns (say five differences for patterns of size 20) would be unreasonably restrictive for larger patterns (five differences for patterns of size 100). Conversely, any fixed number for large patterns would allow too much variability in small patterns. A heuristic algorithm by Karlin *et al*. (26) is similarly hampered by the use of matching blocks separated by error blocks of fixed size. The remaining two algorithms in this group require input from the user which limits their usefulness. An earlier heuristic algorithm by Benson (27) finds tandem repeats only if they have a pattern size which is specified in advance. An exact algorithm by Myers and Sagot (28) (limited to patterns with size of at most 40 bases) requires that the approximate pattern size and a range for the number of copies be specified.

The algorithm (29) presented in this paper is designed to overcome many of the aforementioned limitations: (i) it uses the method of *k-tuple matching* to avoid the need for full scale alignment matrix computations; (ii) it requires no *a priori* knowledge of the pattern, pattern size or number of copies; (iii) there are no restrictions on the size of the repeats that can be detected; (iv) it uses percentage differences between adjacent copies and treats substitutions and indels separately; (v) it determines a consensus pattern for the smallest repetitive unit in the tandem repeat. The program 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 (18).

A number of ideas incorporated into this new algorithm have been utilized in earlier homology detection programs (30,31), yet the goals and methods differ. Instead of looking for highest scoring homologous regions, the algorithm looks for 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. The detection criteria are based on a stochastic model of tandem repeats specified by percent identity and frequency of insertions and deletions, rather than some minimal alignment score. Finally, the program aligns repeat copies against a consensus sequence, revealing patterns of common mutations. These patterns yield insight into the history of duplications that produced the tandem repeat, thus providing a potentially valuable tool for phylogenetic research.

```
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
```

Figure 1. Two adjacent copies from a tandem repeat in the human β T cell receptor locus sequence (37). H indicates a match, T indicates a mismatch, insertion or deletion.

The remainder of this paper is organized as follows. In Methods we present a probabilistic model of tandem repeats, an algorithm overview and the set of criteria that guide the recognition process. In the Discussion we present our analysis of the frataxin (Friedreich’s ataxia) gene sequence, the human β T cell receptor locus and two yeast chromosomes. Finally, in the Conclusion we describe directions for future research.

METHODS

Probabilistic model of tandem repeats

We model *alignment* of two tandem copies of a pattern of length *n* by a sequence of *n*-independent Bernoulli trials (coin tosses). The probability of success, *P* (*heads*), which we also call p_M or *matching probability*, represents the *average* percent identity between the copies. Each head in the Bernoulli sequence is interpreted as a match between aligned nucleotides. Each tail is a mismatch, insertion or deletion. A second probability, p_I or *indel probability*, specifies the average percentage of insertions and deletions between the copies. Figure 1 illustrates the underlying idea for the model.

While Figure 1 is an interpretation of a particular alignment as a Bernoulli sequence, we are more generally interested in the distribution of Bernoulli sequences and the properties of alignments that they represent when dealing with a specific pair (p_M, p_I), for example ($p_M = 0.80$, $p_I = 0.10$). Note that these *conservation parameters* serve as a type of extremal bound, i.e. as a quantitative description of the *most divergent copies* we hope to detect.

Program outline

Our program has *detection* and *analysis* components. The detection component uses a set of statistically based criteria to find *candidate* tandem repeats. The analysis component attempts to produce an alignment for each candidate and if successful gathers a number of statistics about the alignment (percent identity, percent indels) and the nucleotide sequence (composition, entropy measure).

Detection component. We assume that adjacent copies of any pattern will contain some matching characters in corresponding positions. Just how many matches and how the distance between those matches should vary depend on the fixed values of p_M and p_I. In the next section, we develop the statistical criteria to answer these questions. Here, we describe how the matches are detected.

The algorithm looks for matching nucleotides separated by a common distance *d*, which is not specified in advance. For reasons of efficiency it looks for *runs of k matches*, which we call *k-tuple matches*. A *k-tuple* is a window of *k* consecutive characters from the nucleotide sequence. Matching *k*-tuples are two windows with identical contents and if aligned in the Bernoulli model would produce a run of *k* heads. Because we limit ourselves to *k*-tuple matches, we will *not* detect all matching characters. For example, if $k = 6$ and two windows contain

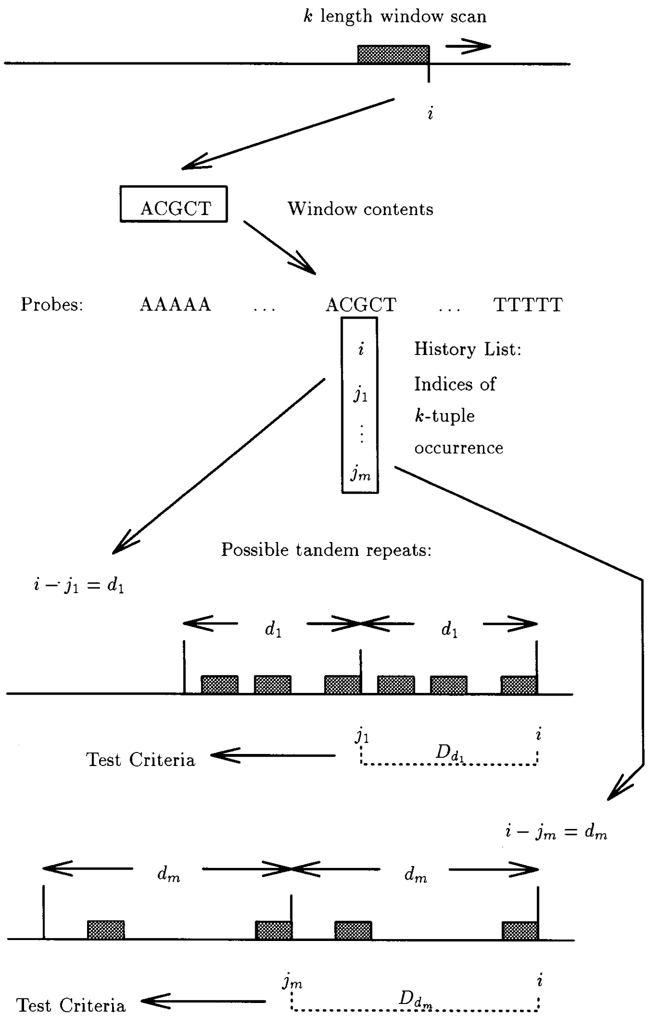


Figure 2. Tandem repeats are detected by scanning the sequence with a small window, determining the distance between exact matches and testing the statistical criteria.

TC<u>A</u>TGT and TC<u>T</u>TGT we will *not* know that there are 5 matching characters because the window contents are not identical. Put in terms of the Bernoulli model, the aligned windows would be represented by the sequence HHTHHH, which is *not* a run of 6 heads.

The basic operation of the detection component is illustrated in Figure 2. Let *S* be a nucleotide sequence. We select a small integer *k* for the tuple or window size (*k* = 5 for example) and keep a list of all possible *k* length strings (there are 4^k for the DNA alphabet A,C,G,T) which we call the *probes*. By sliding the window across the sequence, we 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 *p* occurs.

When a position *i* is added to H_p, we scan H_p for *all* earlier occurrences of *p*. Let one earlier occurrence be at *j*. Since *i* and *j* are the indices of matching *k*-tuples, the distance $d = i - j$ is a *possible pattern size* for a tandem repeat. For the criteria tests, we need information about other *k*-tuple matches at the same distance *d* where the leading tuple occurs in the sequence between *j* and *i*. A *distance list* D_d stores this information. It can be thought of as a sliding window of length *d* which keeps track of the positions of matches and their total.

List D_d is updated every time a match at distance *d* is detected. Position *i* of the match is stored on the list and the total is increased. The right end of the window is set to *i* and matches that occurred before $j = i - d$ are dropped from the list and subtracted from the total. Lists for other nearby distances are also updated at this time (Random Walk Distribution in the next section), but only to reset their right ends to *i* and remove matches that have been passed by the advancing windows. Information in the updated distance lists is used for the sum of heads and apparent size criteria tests as described in the next section. If both tests are passed, the program moves on to the analysis component.

Statistical criteria. The statistical criteria are based on runs of heads in Bernoulli sequences, corresponding to matches detected with the *k*-tuples and stored in the distance lists. The criteria are based on four distributions which depend upon: (i) the pattern length, *d*; (ii) the matching probability, p_M; (iii) the indel probability, p_I; (iv) the tuple size, *k*. For each distribution, we either calculate it with a formula or estimate it using simulation. Then, we select a cut-off value that serves as our criterion. Below we describe the distributions and criteria in more detail.

Sum of heads distribution. This distribution indicates how many matches are required. Let the random variable $R_{d,k,pM}$ = the total number of heads in head runs of length *k* or longer in an iid Bernoulli sequence of length *d* with success probability p_M. The distribution of $R_{d,k,pM}$ is well approximated by the normal distribution and we have previously shown that its exact mean and variance can be calculated in constant time (32). For the *sum of heads criterion*, we use the normal distribution to determine the largest number, *x*, such that 95% of the time $R_{d,k,pM} \geq x$. For example, if $p_M = 0.75$, $k = 5$ and $d = 100$, then the criterion is 26. Put another way, if a 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.

Random walk distribution. This distribution describes how distances between matches may vary due to indels. Because indels change the distance between matching *k*-tuples (Fig. 3), there will be situations where the pattern has size *d*, yet the distance between matching *k*-tuples is $d \pm 1$, $d \pm 2$, etc. In order to test the sum of heads criterion, we count the matches in $D_{d \pm \Delta d}$, for $\Delta d = 0, 1, ..., \Delta d_{max}$ for some Δd_{max}. In our model, indels are single nucleotide events occurring with probability p_I. Insertions and deletions are considered equally likely and we treat the distance change as a problem of random walks. Let the random variable $W_{d,pI}$ = the maximum displacement from the origin of a one-dimensional random walk with expected number of steps equal to $p_I \cdot d$. It can be shown (33) that 95% of the time, $W_{d,pI}$ ranges between $\pm 2.3\sqrt{p_I \cdot d}$. We set $\Delta d_{max} = \lfloor 2.3\sqrt{p_I \cdot d} \rfloor$. For ⌋example if $p_I = 0.1$ and $d = 100$, then $\Delta d_{max} = 7$.

Apparent size distribution. This distribution is used to distinguish between tandem repeats and non-tandem direct repeats (Fig. 4). For tandem repeats, the leading tuples in matching *k*-tuples will be distributed throughout the interval from *j* to *i*, whereas for non-tandem repeats, they should be concentrated on the right side of the interval near *i*. Let the random variable $S_{d,k,pM}$ = the distance between the first and last run of *k* heads in an iid

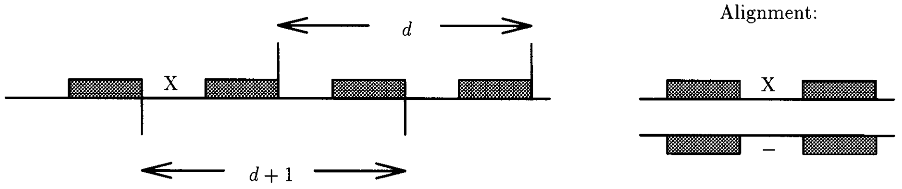


Figure 3. Insertions and deletions change the distance between exact matches. The inserted character X causes one pair of matching *k*-tuples to be separated by distance *d* + 1 while another pair is separated only by distance *d*.

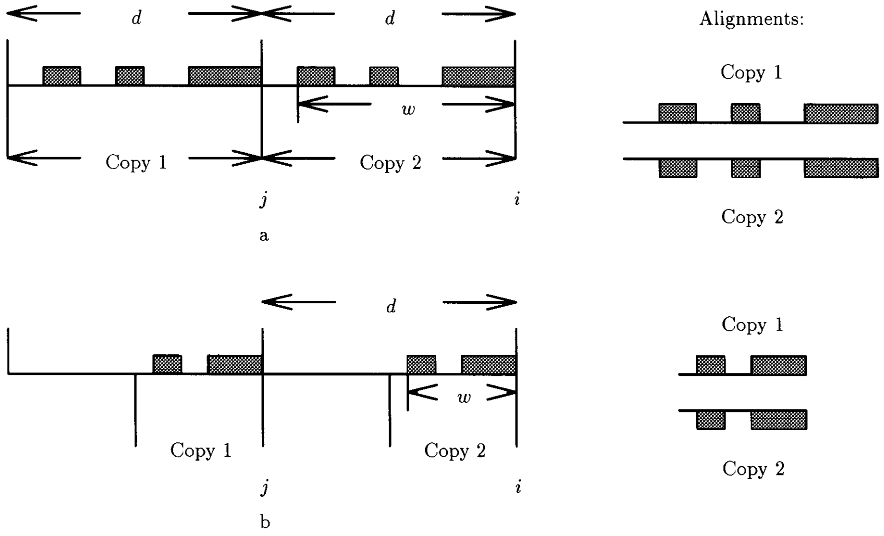


Figure 4. We must distinguish between (**a**) a tandem repeat (leading tuples in *k*-tuple matches spread over the interval between *i* and *j*) and (**b**) a non-tandem, direct repeat (leading tuples concentrated on the right). Matching *k*-tuples are indicated by the shaded boxes. *w* is the distance between the first and last leading tuple.

Bernoulli sequence of length *d* with success probability p_M. $S_{d,k,pM}$ is the apparent size of the repeat when using *k*-tuples to find the matches and will usually be shorter than the pattern size *d*. We estimate the distribution of $S_{d,k,pM}$ by simulation because we make it conditional on first meeting the sum of heads criterion. For given *d*, *k* and p_M, random Bernoulli sequences are generated using p_M. For every sequence that meets or exceeds the sum of heads criteria, the distance between the first and last run of heads of length *k* or larger is recorded. From the distribution, we determine the maximum number *y* such that 95% of the time $S_{d,k,pM} > y$. We use *y* as our *apparent size criterion*. For example, if $p_M = 0.75$, $k = 5$ and $d = 100$, then the criterion is 56. In order to test the apparent size criterion, we compute the distance between the first and last tuple on list D_d. If the distance between the tuples is smaller than the criterion, we assume the repeat is not tandem or that we have not yet seen enough of it to be convinced.

Waiting time distribution. This distribution is used to pick tuple sizes. Tuple size has a significant inverse effect on the running time of the program because increasing tuple size causes an exponential decrease in the expected number of tuple matches. If the nucleotides occur with equal frequency, then increasing the tuple size by Δk increases the average distance between randomly matching tuples by a factor of $4^{\Delta k}$. If $k = 5$, the average distance between random matches is ~1 kb, but if $k = 7$, the average distance is ~16 kb. Thus, by using a larger tuple size, we keep the history lists short. On the other hand, increasing the tuple size decreases the chance of noticing approximate copies because they may not contain a long, unbroken run of matches. Let the random variable $T_{k,pM}$ = the number of iid Bernoulli trials with success probability p_M until the first occurrence of a run of *k* successes. $T_{k,pM}$ follows the *geometric distribution of order k*. If we let $p = p_M$ and $q = 1 - p$ then the exact probability $P(T_{k,pM} = x)$ for $x \geq 0$ is given by the recursive formula (34)

$$P(T_{k,pM} = x) = \begin{cases} 0 & \text{for } x < k \\ p^k & \text{for } x = k \\ qp^k\,[1 - \sum_{i=0}^{x-k-1} P(T_{k,pM} = i)] & \text{for } x > k \end{cases}$$

For example, if $p_M = 0.75$ and $k = 5$ then we need at least 31 trials (coin tosses) to have a 95% chance of seeing a run of 5 heads. For patterns smaller than 31 characters, we need to use a smaller *k*-tuple. The waiting time distribution allows us to balance the running time and sensitivity of our algorithm by picking a set of tuple sizes, *each applying to a different range of pattern sizes*. The program processes the sequence once, simultaneously checking these different tuple sizes. We require that the smallest pattern for tuple size *k* have a sum of heads criterion of at least *k* + 1. Table 1 shows the range of tuple sizes and the corresponding pattern sizes currently used by the program.

Analysis component. If the information in the distance list passes the criteria tests, a candidate pattern consisting of positions *j* + 1 . . . *i* is selected from the nucleotide sequence and aligned with the surrounding sequence using wraparound dynamic programming (WDP) (35,36). If at least two copies of the pattern are aligned with the sequence, the tandem repeat is reported. Several implementation details of the analysis component are described below.

Table 1. Tuple sizes and the range of pattern sizes each is used to detect

Tuple	Pattern Sizes	
Sizes	p_M = .75	p_M = .8
3	1 – 29	–
4	30 – 43	1 – 29
5	44 – 159	30 – 159
7	160 – 500	160 – 500

Multiple reporting of repeat at different pattern sizes. When a single tandem repeat contains many copies, several pattern sizes are possible. For example, if the basic pattern size is 26, then the repeat may be reported at sizes 26, 52, 78, etc. We limit this redundancy in the output to, at most, three pattern sizes. Note that we do not automatically limit the output to the smallest period size because a much better alignment may come from a larger size (for example Table 5, indices 410172–410459).

Narrow band alignment. Alignments are the program's most time intensive calculations. To decrease running time, we limit WDP calculations to a narrow diagonal band in the alignment matrix for patterns larger than 20 characters. In accordance with the random walk results, the band radius is $\Delta d_{\max}$. The band is periodically recentered around a run of matches in the current best alignment.

Consensus pattern and period size. An initial candidate pattern *P* is drawn from the sequence, but this is usually not the best pattern to align with the tandem repeat. To improve the alignment, we determine a *consensus pattern* by majority rule from the alignment of the copies with *P*. The consensus is used to realign the sequence and this final alignment is reported in the output. Period size is defined as the most common matching distance between corresponding characters in the alignment and may not be identical to consensus size.

Program usage and output

Input to the program consists of a sequence file and the following parameters: (i) alignment weights for match, mismatch and indels; (ii) p_M and p_I; (iii) a minimum size for patterns to report; (iv) a minimum alignment score to report. We have developed a web based interface for the program. Using an HTML form at c3.biomath.mssm.edu/trf.html , the user provides an input DNA sequence file. Defaults can be used for the remaining parameters. After program execution, two files are returned. The first is a summary table describing the location and statistical properties of the tandem repeats found. The second contains the alignment of each repeat with its consensus sequence. The files are linked so that selecting an entry from the table opens a second browser window which contains the proper alignment. The summary table includes the following information: (i) indices of the repeat in the sequence; (ii) period size; (iii) number of copies aligned with the consensus pattern; (iv) size of the consensus pattern (may differ from the period size); (v) percent of matches between adjacent copies overall; (vi) percent of indels between adjacent copies overall; (vii) alignment score; (viii) percent composition for each of the four nucleotides; (ix) entropy measure based on percent composition.

RESULTS

To demonstrate the capabilities of our program, we used it to analyze four sequences, the human frataxin gene sequence (Friedreich's ataxia) (5), the human β T cell receptor locus sequence (37) and two yeast chromosomes (I and VIII). [The frataxin gene sequence and the human β T cell receptor sequences were obtained from GenBank. The yeast chromosomes sequences were obtained via ftp from ftp.ebi.ac.uk directory pub/databases/yeast in files chri_230209.ascii and chrviii_562638.ascii. Indexing in this paper is relative to the sequences in these files. Data file accession numbers for these sequences are: frataxin gene promoter and intron 1, U43748; human T cell receptor, L36092; yeast chromosome 1, U12980, L20125, L05146, L22015, L28920; yeast chromosome 8, U11583, U11582, U11581, U10555, U10400, U10399, U00062, U00061, U10556, U00060, U00059, U10398, U10397, U00027, U00028, U00030, U00029.] In our analysis, we searched for all pattern sizes between 1 and 500 bases (the implementation's current upper size limit, to be extended in subsequent versions). We used one of two sets of alignment parameters (match, mismatch, gap), either (+2,–7,–7) or (+2,–5,–7). Only those repeats scoring at least 50 with these parameters are reported. Occasionally, the same repeat is reported at different pattern sizes. We have omitted these redundancies.

We performed two searches on each sequence, using different conservation parameter values, (p_M = 0.75, p_I = 0.20) and (p_M = 0.80, p_I = 0.10). While the first search is slower than the second, the detected repeats are nearly identical. Table 2 shows running times of the program and Tables 4–7 list the tandem repeats found.

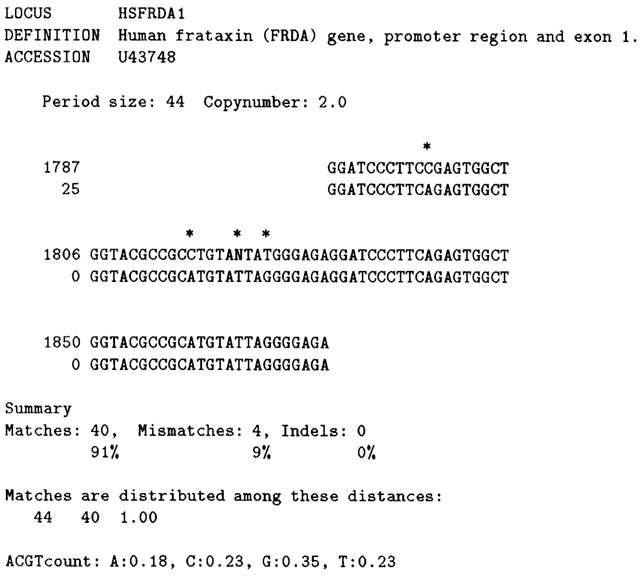
Table 2. Running times of program on selected sequences using a Silicon Graphics O2 RS10000

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

Time grows linearly with sequence length. With conservation parameter values (p_M = 0.75, p_I = 0.20) running time is ~10 times slower than with values (p_M = 0.80, p_I = 0.10) although the detected repeats are nearly identical. Alignment weights also affect running time. The most liberal weights tested increase the times shown here by ~50%.

Human frataxin gene (Friedreich's ataxia), intron 1

Friedreich's ataxia is one of the triplet repeat diseases (5). It is caused by copy number expansion of the triplet GAA in the first intron of the frataxin gene. Table 4 lists the repeats found in the sequence. Besides the triplet repeat, our program found two others which were apparently unknown, a 44 bp pattern and a 14 bp pattern. Figure 5 shows the program's alignment of the 44 bp repeat.



```
LOCUS       HSFRDA1
DEFINITION  Human frataxin (FRDA) gene, promoter region and exon 1.
ACCESSION   U43748

    Period size: 44  Copynumber: 2.0

                                           *
    1787                          GGATCCCTTCCGAGTGGCT
      25                          GGATCCCTTCAGAGTGGCT

                  *    *  *
    1806 GGTACGCCGCCTGTANTATGGGAGAGGATCCCTTCAGAGTGGCT
       0 GGTACGCCGCATGTATTAGGGGAGAGGATCCCTTCAGAGTGGCT


    1850 GGTACGCCGCATGTATTAGGGGAGA
       0 GGTACGCCGCATGTATTAGGGGAGA

Summary
Matches: 40,  Mismatches: 4, Indels: 0
         91%               9%         0%

Matches are distributed among these distances:
   44   40  1.00

ACGTcount: A:0.18, C:0.23, G:0.35, T:0.23
```

Figure 5. The program's alignment of the 44 bp repeat from the frataxin gene intron 1 (Friedreich's ataxia). This repeat was apparently unknown. The actual sequence is on the top; the consensus sequence is on the bottom. Each pair of lines represents one period. Position of the beginning of the repeat is relative to the detected pattern when the criteria were met and is therefore arbitrary. Symbol * indicates a mismatch. Summary refers to matches, mismatches and indels between adjacent copies in the sequence, not between the sequence and the consensus pattern.

Table 3. Varying copy numbers in the four similar tandem repeat clusters found in yeast chromosomes 1 and 8

Period Size	Cluster			
	C_1	C_2	C_3	C_4
27	2.2	2.2	3.3	1.2
21	3.2	4.3	3.0	3.8
48	7.7	1.7	1.7	1.7
15	9.1	9.1	5.5	6.3
135	0.7	13	17.9	7.8

See text and Tables 6 and 7 for cluster locations.

Table 4. Tandem repeats detected in the human frataxin gene intron 1 sequence

Human Frataxin Gene (Friedreich's ataxia) Intron 1											
Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0–2)
822–854	14	2.4	14	89	0	57	6	48	42	0	1.28
1787–1874	44	2.0	44	90	0	140	18	22	35	22	1.95
2183–2211	3	9.7	3	100	0	58	68	0	31	0	0.89

Human β T cell receptor locus sequence

This sequence (37) contains a family of immune recognition coding elements, the T cell receptor variable, diversity, joining and constant gene segments. It was selected for its size and because many tandem repeats within the sequence had already been identified. Table 5 lists the new repeats we found. Of the 83 repeats that we found, 38 were previously annotated and most of those were for patterns of size 5 or smaller. We missed 6 annotated repeats: 4 dinucleotide repeats and 1 tetranucleotide repeat (alignment scores were below our cut-off) and 1 repeat with period size 10 567 bases (beyond the current implementation's pattern upper size limit). Of the 45 unannotated repeats, 13 have short patterns (2–6 bp) and may be polymorphic and thus useful for linkage analysis. Six unannotated repeats have large pattern sizes (116, 65, 52, 49, 34 and 30 bp). The 116 base pattern is also reported at size 39 with a lower scoring alignment. The annotated 60 base pattern repeat (indices 12596–13266) is indicative of the program's ability to find repeats with substantial amounts of mutation between adjacent copies (74% matching characters and 7% indels overall).

Table 5. Tandem repeats detected in human β T cell receptor locus sequence

Human β T cell receptor locus sequence – 684,973 bp											
Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0–2)
4627–4657	15	2.1	14	94	5	53	41	25	6	25	1.79
*12596–13266	60	11.2	58	74	7	495	28	29	10	30	1.91
21863–21905	14	2.8	16	86	13	63	44	13	6	34	1.72
48825–48928	52	2.0	52	94	0	181	42	9	20	27	1.83
84876–84913	17	2.2	17	95	0	67	26	13	10	50	1.73
92444–92488	5	8.8	5	85	4	54	0	17	0	82	0.68
99921–99961	2	20.5	2	100	0	82	0	0	51	48	1.00
111227–111251	12	2.1	12	100	0	50	24	44	8	24	1.80
121522–121659	65	2.1	65	98	0	267	8	41	18	31	1.81
124841–124887	2	23.0	2	86	4	67	6	0	46	46	1.28
134820–134856	2	18.5	2	91	0	56	48	2	2	45	1.30
189754–189806	2	26.5	2	92	0	88	0	0	45	54	0.99
193027–193059	6	5.5	6	92	0	57	0	0	18	81	0.68
216002–216043	19	2.2	19	95	0	75	38	16	23	21	1.93
283516–283544	12	2.4	12	100	0	58	17	17	41	24	1.90
344711–344810	49	2.0	49	100	0	200	33	26	21	20	1.97
376278–376322	22	2.0	22	95	0	81	24	17	15	42	1.88
409705–409783	6	13.5	6	90	5	81	0	67	5	27	1.12
†410172–410459	39	7.5	37	79	9	290	11	37	33	17	1.86
410172–410459	116	2.5	116	86	1	391	11	37	33	17	1.86
431581–431607	12	2.2	12	100	0	54	40	7	0	51	1.30
442182–442208	2	13.5	2	100	0	54	0	0	48	51	1.00
444936–444983	2	24.0	2	91	0	78	47	47	4	0	1.21
455160–455204	18	2.6	18	89	6	67	53	0	24	22	1.46
465203–465246	2	22.0	2	95	0	79	50	47	0	2	1.13
470590–470626	4	9.2	4	100	0	74	75	0	0	24	0.80
512990–513015	5	5.2	5	100	0	52	61	0	38	0	0.96
539683–539720	2	19.0	2	100	0	76	50	0	0	50	1.00
568123–568164	13	3.1	13	86	6	57	52	4	2	40	1.35
612431–612470	19	2.1	19	90	0	62	32	10	22	35	1.87
614497–614565	34	2.0	34	100	0	138	20	28	17	33	1.95
638632–638674	12	3.7	12	87	3	52	46	23	0	30	1.52
684213–684417	30	7.0	30	90	7	332	37	25	9	27	1.87

Not shown are all mononucleotide repeats and those repeats already annotated in the GenBank entries (accession nos L36092, U66059, U66060 and U66061) except for the 60 bp repeat marked with symbol *. Symbol † indicates a pattern which is included even though a longer pattern has a better scoring alignment.

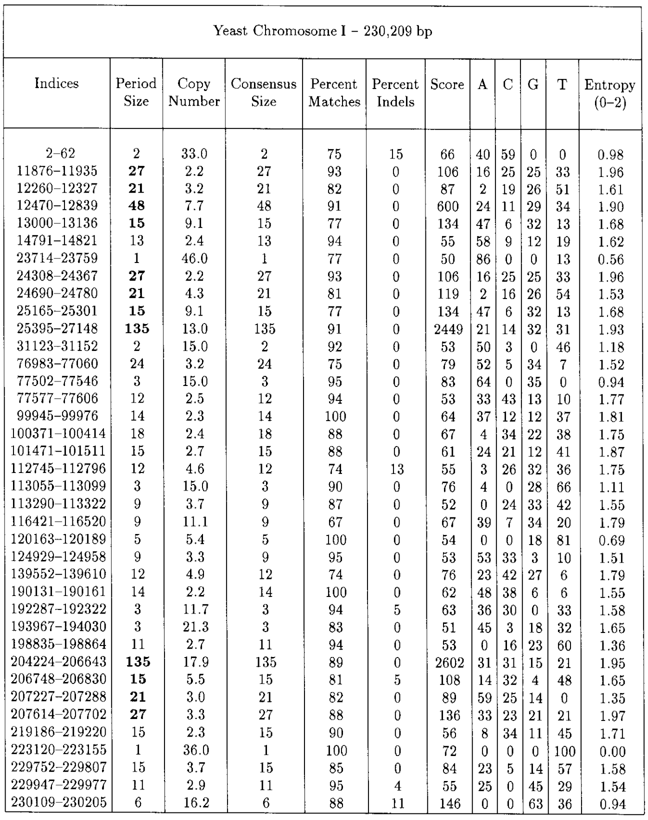
Yeast chromosomes

Tables 6 and 7 list the tandem repeats found for the yeast sequences. Of special interest are the clusters of tandem repeats which show up repeatedly at the ends of the chromosomes, suggesting recent swapping of the ends. Chromosome 8, in particular, has two different clusters on its right end.

The (27, 21, 48, 15, 135) cluster

The *FLO1* gene and its paralogous pseudogenes in chromosomes 1 and 8 contain a cluster of 5 tandem repeats with pattern sizes 27,

Table 6. Tandem repeats detected in yeast chromosome 1



Yeast Chromosome I – 230,209 bp

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0–2)
2–62	2	33.0	2	75	15	66	40	59	0	0	0.98
11876–11935	**27**	2.2	27	93	0	106	16	25	25	33	1.96
12260–12327	**21**	3.2	21	82	0	87	2	19	26	51	1.61
12470–12839	**48**	7.7	48	91	0	600	24	11	29	34	1.90
13000–13136	**15**	9.1	15	77	0	134	47	6	32	13	1.68
14791–14821	13	2.4	13	94	0	55	58	9	12	19	1.62
23714–23759	1	46.0	1	77	0	50	86	0	0	13	0.56
24308–24367	**27**	2.2	27	93	0	106	16	25	25	33	1.96
24690–24780	**21**	4.3	21	81	0	119	2	16	26	54	1.53
25165–25301	**15**	9.1	15	77	0	134	47	6	32	13	1.68
25395–27148	**135**	13.0	135	91	0	2449	21	14	32	31	1.93
31123–31152	2	15.0	2	92	0	53	50	3	0	46	1.18
76983–77060	24	3.2	24	75	0	79	52	5	34	7	1.52
77502–77546	3	15.0	3	95	0	83	64	0	35	0	0.94
77577–77606	12	2.5	12	94	0	53	33	43	13	10	1.77
99945–99976	14	2.3	14	100	0	64	37	12	12	37	1.81
100371–100414	18	2.4	18	88	0	67	4	34	22	38	1.75
101471–101511	15	2.7	15	88	0	61	24	21	12	41	1.87
112745–112796	12	4.6	12	74	13	55	3	26	32	36	1.75
113055–113099	3	15.0	3	90	0	76	4	0	28	66	1.11
113290–113322	9	3.7	9	87	0	52	0	24	33	42	1.55
116421–116520	9	11.1	9	67	0	67	39	7	34	20	1.79
120163–120189	5	5.4	5	100	0	54	0	0	18	81	0.69
124929–124958	9	3.3	9	95	0	53	53	33	3	10	1.51
139552–139610	12	4.9	12	74	0	76	23	42	27	6	1.79
190131–190161	14	2.2	14	100	0	62	48	38	6	6	1.55
192287–192322	3	11.7	3	94	5	63	36	30	0	33	1.58
193967–194030	3	21.3	3	83	0	51	45	3	18	32	1.65
198835–198864	11	2.7	11	94	0	53	0	16	23	60	1.36
204224–206643	**135**	17.9	135	89	0	2602	31	31	15	21	1.95
206748–206830	**15**	5.5	15	81	5	108	14	32	4	48	1.65
207227–207288	**21**	3.0	21	82	0	89	59	25	14	0	1.35
207614–207702	**27**	3.3	27	88	0	136	33	23	21	21	1.97
219186–219220	15	2.3	15	90	0	56	8	34	11	45	1.71
223120–223155	1	36.0	1	100	0	72	0	0	0	100	0.00
229752–229807	15	3.7	15	85	0	84	23	5	14	57	1.58
229947–229977	11	2.9	11	95	4	55	25	0	45	29	1.54
230109–230205	6	16.2	6	88	11	146	0	0	63	36	0.94

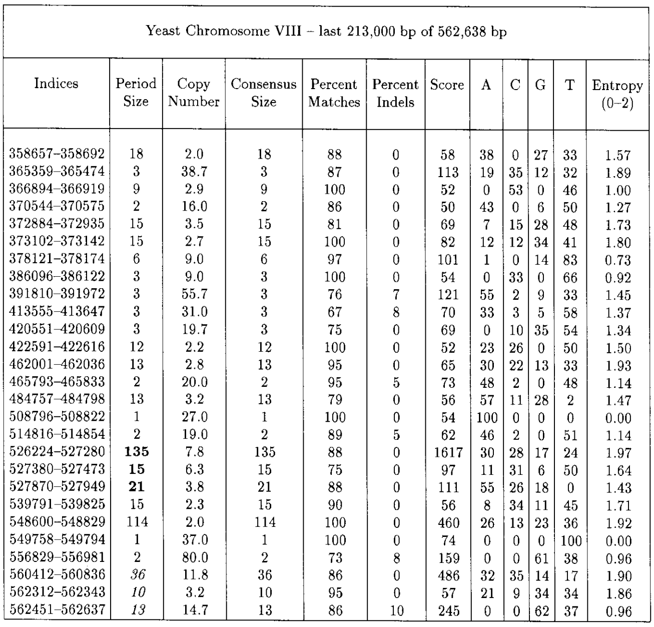
Period sizes in bold indicate similar clusters found at the ends of chromosomes 1 and 8. From the top, these are clusters C_1, C_2 and C_3.

21, 48, 15 and 135. We designate these clusters C_1 and C_2 (adjacent on the left end of chromosome 1), C_3 (right end, opposite strand) and C_4 (right end of chromosome 8). The 27, 48 and 135 base patterns are not reported in every cluster in Tables 6 and 7. Subsequent analysis of the surrounding sequences, however, revealed that every pattern is present but not necessarily as two or more copies (Table 3). For each pattern size, the number of copies varies among the four clusters. More specifically, no cluster is identical in its copy number to any other cluster. This implies that duplication or excision events (deletion of copies) have occurred since the separate clusters were incorporated into the chromosomes. The sequences around these clusters also reveal close homology. For example, C_3 and C_4 are nearly identical over 18 000 bases and C_2 and C_3 display homology over 15 000 bases.

The (13, 10, 36) cluster

A cluster of 3 tandem repeats with pattern sizes 13, 10 and 36 bases appears on both ends of chromosome 8 (Table 7). The 36 bp pattern also appears on the left end (low index numbers) of chromosome 6 (not shown). For the 36 bp pattern, each occurrence has a different copy number. The 10 and 13 bp patterns are identical in their occurrences. Surrounding sequences comprising 4200 bases are nearly identical for these three clusters.

Table 7. Tandem repeats detected in yeast chromosome 8 (only the latter half of the sequence is shown)



Yeast Chromosome VIII – last 213,000 bp of 562,638 bp

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0–2)
358657–358692	18	2.0	18	88	0	58	38	0	27	33	1.57
365359–365474	3	38.7	3	87	0	113	19	35	12	32	1.89
366894–366919	9	2.9	9	100	0	52	0	53	0	46	1.00
370544–370575	2	16.0	2	86	0	50	43	0	6	50	1.27
372884–372935	15	3.5	15	81	0	69	7	15	28	48	1.73
373102–373142	15	2.7	15	100	0	82	12	12	34	41	1.80
378121–378174	6	9.0	6	97	0	101	1	0	14	83	0.73
386096–386122	3	9.0	3	100	0	54	0	33	0	66	0.92
391810–391972	3	55.7	3	76	7	121	55	2	9	33	1.45
413555–413647	3	31.0	3	67	8	70	33	3	5	58	1.37
420551–420609	3	19.7	3	75	0	69	0	10	35	54	1.34
422591–422616	12	2.2	12	100	0	52	23	26	0	50	1.50
462001–462036	13	2.8	13	95	0	65	30	22	13	33	1.93
465793–465833	2	20.0	2	95	5	73	48	2	0	48	1.14
484757–484798	13	3.2	13	79	0	56	57	11	28	2	1.47
508796–508822	1	27.0	1	100	0	54	100	0	0	0	0.00
514816–514854	2	19.0	2	89	5	62	46	2	0	51	1.14
526224–527280	**135**	7.8	135	88	0	1617	30	28	17	24	1.97
527380–527473	**15**	6.3	15	75	0	97	11	31	6	50	1.64
527870–527949	**21**	3.8	21	88	0	111	55	26	18	0	1.43
539791–539825	15	2.3	15	90	0	56	8	34	11	45	1.71
548600–548829	114	2.0	114	100	0	460	26	13	23	36	1.92
549758–549794	1	37.0	1	100	0	74	0	0	0	100	0.00
556829–556981	2	80.0	2	73	8	159	0	0	61	38	0.96
560412–560836	*36*	11.8	36	86	0	486	32	35	14	17	1.90
562312–562343	*10*	3.2	10	95	0	57	21	9	34	34	1.86
562451–562637	*13*	14.7	13	86	10	245	0	0	62	37	0.96

Period sizes in bold indicate one of four similar clusters found at the ends of chromosomes 1 and 8. Cluster C_4 is shown. Period sizes in italics indicate one of three similar clusters found at both ends of chromosome 8 and one end of chromosome 6.

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 and utility of the algorithm by analyzing four sequences ranging in size up to 700 kb. Several avenues for future research are raised by this work, including methods to estimate statistical significance for tandem repeats and algorithms to determine plausible mutational histories.

Statistical issues

We have yet to develop a good statistical significance measure for tandem repeats. For now, we use a cut-off alignment score based on simulations with random sequences. Difficulties include the local variation in nucleotide content in real sequences, which is decidedly non-random, and the problem of accounting for copy number as well as total repeat length. Estimates of significance developed in Benson and Waterman (27) are too high in this application because they apply to tandem repeats of one pattern size only, rather than the range of sizes considered here.

Mutational history

Analyzing the mutational history of tandem repeats requires utilizing the pattern of mutations among adjacent copies to describe the interwoven progression of substitutions, indels and duplication/excision events leading from a single copy of the pattern to the present day sequence. Such histories can suggest how the boundaries and size of the duplication unit vary and may reveal details about the duplication mechanism.

ACKNOWLEDGEMENTS

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 programing and examining the output. Thanks also to Mike Waterman, Richard Arratia and Rolf Backofen for helpful discussions. This work was partially supported by NSF grant CCR-9623532.

REFERENCES

1 Verkerk,A., Pieretti,M., Sutcliffe,J. Fu,Y., Kuhl,D., Pizzuti,A., Reiner,O., Richards,S., Victoria,M., Zhang,F., Eussen,B., van Ommen,G., Blonden,A., Riggins,G., Chastain,J., Kunst,C., Galjaard,H., Caskey,C., Nelson,D., Oostra,B. and Warren,S. (1991) *Cell*, **65**, 905–914.

2 Huntington’s Disease Collaborative Research Group. (1993) *Cell*, **72**, 971–983.

3 Fu,Y.-H., Pizzuti,A., Fenwick,J., King,R.G.Jr. and Rajnarayan,S., Dunne,P.W., Dubel,J., Nasser,G.A., Ashizawa,T., DeJong,P., Wieringa,B. Korneluk,R., Perryman,M.B., Epstein,H.F. and Caskey,C.T. (1992) *Science*, **255**, 1256–1258.

4 La Spada,A., Wilson,E., Lubahn,D., Harding,A. and Fischbeck,K. (1991) *Nature*, **352**, 77–79.

5 Campuzano,V., Montermini,L., Molto,M.D., Pianese,L. and Cossee,M. (1996) *Science*, **271**, 1423–1427.

6 Wells,R. (1996) *J. Biol. Chem*., **271**, 2875–2878.

7 Weitzmann,M., Woodford,K. and Usdin,K. (1997) *J. Biol. Chem*., **272**, 9517–9523.

8 Hamada,H., Seidman,M., Howard,B. and Gorman,C. (1984) *Mol. Cell. Biol*., **4**, 2622–2630.

9 Pardue,M., Lowenhaupt,K., Rich,A. and Nordheim,A. (1987) *EMBO J*., **6**, 1781–1789.

10 Yee,H., Wong,A., van den Sande,J. and Rattner,J. (1991) *Nucleic Acids Res*., **19**, 949–953.

11 Richards,R., Holman,K., Yu,S. and Southerland,G. (1993) *Hum. Mol. Genet*., **2**, 1429–1435.

12 Lu,Q., Wallrath,L., Granok,H. and Elgin,S. (1993) *Mol. Cell. Biol*., **13**, 2802–2814.

13 Du,J., Zhu,Y., Shanmugam,A. and Kenter,A. (1997) *Nucleic Acids Res*., **25**, 3066–3073.

14 Edwards,A., Hammond,H., Jin,L., Caskey,C. and Chakraborty,R. (1992) *Genomics*, **12**, 241–253.

15 Weber,J. and May,P. (1989) *Am. J. Hum. Genet*., **44**, 388–396.

16 Tishkoff,S.A., Dietzsch,E., Speed,W., Pakstis,A.J. and Kidd,J.R. (1996) *Science*, **271**, 1380–1387.

17 Armour,J., Anttinen,T., May,C., Vega,E., Sajantila,A., Kidd,J., Kidd,K., Bertranpetit,J., Pääbo,S. and Jeffreys,A. (1996) *Nature Genet*., **13**, 154–160.

18 Benson,G. (1997) *J. Comput. Biol*., **4**, 351–367.

19 Hellman,L., Steen,M., Sundvall,M. and Pettersson,U. (1988) *Gene*, **68**, 93–100.

20 Kannan,S.K. and Myers,E.W. (1996) *SIAM J. Comput*., **25**, 648–662.

21 Benson,G. (1995) *Theor. Comput. Sci*., **145**, 357–369.

22 Schmidt,J.P. (1998) *SIAM J. Comput*., **27**, 972–992.

23 Milosavljevi**c**,A. and Jurka,J. (1993) *CABIOS*, **9**, 407–411.

24 Rivals,E., Delgrange,O., Delahaye,J.-P., Dauchet,M., Delorme,M.-O., Hénaut,A. and Ollivier,E. *CABIOS*, **13**, 131–136, 1997.

25 Landau,G. and Schmidt,J. (1993) In Apostolico,A., Crochemore,M., Galil,Z. and Manber,U (eds), *Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching*, Lecture Notes in Computer Science, Vol. 648. Springer-Verlag, Berlin, pp. 120–133.

26 Karlin,S., Morris,M., Ghandour,G. and Leung,M.-Y. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 841–845.

27 Benson,G. and Waterman,M. (1994) *Nucleic Acids Res*., **22**, 4828–4836.

28 Sagot,M. and Myers,E. (1998) In Istrail,S., Pevzner,P. and Waterman,M. (eds), *Proceedings of the Second Annual International Conference on Computational Molecular Biology.* ACM Press, NY, pp. 234–242.

29 Benson,G. (1998) In Istrail,S., Pevzner,P. and Waterman,M. (eds), *Proceedings of the Second Annual International Conference on Computational Molecular Biology.* ACM Press, NY, pp. 20–29.

30 Pearson,W. and Lipman,D. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

31 Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) *J. Mol. Biol*., **215**, 403–410.

32 , Benson,G. and Su,X. (1998) *J. Comput. Biol*., **5**, 87–100., Benson,G. and Su,X. (1998) *J. Comput. Biol*., **5**, 87–100.

33 Feller,W. (1968) *An Introduction to Probability Theory and its Applications*, 3rd Edn, Vol. I. John Wiley & Sons, New York, NY.

34 Aki,S., Kuboki,H. and Hirano,K. (1984) *Ann. Inst. Statist. Math*., **36**, 431–440.

35 Miller,W. and Myers,E. (1989) *Bull. Math. Biol*., **51**, 5–37.

36 Fischetti,V., Landau,G., Schmidt,J. and Sellers,P. (1992) In Apostolico,A., Crochemore,M., Galil,Z. and Manber,U (eds), *Proceedings of the 3rd Annual Symposium on Combinatorial Pattern Matching*, Lecture Notes in Computer Science, Vol. 644. Springer-Verlag, Berlin, pp. 111–120.

37 Rowan,L., Koop,B. and Hood,L. (1996) *Science*, **272**, 1755–1768.