

Filtering Tandem Repeats in DNA Sequences

Dina Sokol

sokol@sci.brooklyn.cuny.edu

Department of Computer and Information Science, Brooklyn College of the City University of New York

2900 Bedford Avenue, Brooklyn, N.Y. 11210, Phone:(718)951-5000 ext.2065, Fax: (718)951-4842.

Justin Tojeira*

jtojeira@gmail.com

Abstract

A tandem repeat is a sequence of two or more contiguous, approximate copies of a pattern. Tandem repeats occur in the genomes of both eukaryotic and prokaryotic organisms. They are important in numerous fields including disease diagnosis, mapping studies, human identity testing (DNA fingerprinting), sequence homology, and population studies. Although tandem repeats have been used by biologists for many years, there are few tools available for performing an exhaustive search for all tandem repeats in a given sequence. In this paper we describe a software tool that has been implemented as a post-processing stage for a popular tandem repeats program. This new stage allows the program to scale up for use with whole genomic sequences. The program now organizes and filters the data into a meaningful and manageable set. The output is presented as a succinct table of repeats, including several relevant statistics for each repeat.

Keywords: *tandem repeats, Hamming distance, period, filter*

1 Introduction

A *tandem repeat* is a sequence of two or more contiguous, approximate copies of a pattern of nucleotides. Tandem repeats occur in biological sequences with a wide variety. They are important in numerous fields including disease diagnosis, mapping studies, human identity testing, sequence homology, and population studies. Tandem repeats in DNA are also called *satellite DNA*. Satellite DNA is usually classified among satellites (spanning megabases of DNA), minisatellites (repeat units in the range 9-80 bp, spanning 1 kb to 20 kb) and microsatellites (repeat units in the range 1-6 bp, spanning less than 150 bp).

Tandem repeats in the human genome are important as genetic markers and are used in DNA fingerprinting [5]. They are also responsible for a number of inherited diseases involving the central nervous system. For example, in a normal FMR-1 gene, the triplet CGG is tandemly repeated 6 to 54 times, while in patients with Fragile X Syndrome, the pattern occurs more than 200 times. Kennedy disease and Myotonic Dystrophy (MD) are two other diseases that have been associated with triplet repeats [4]. In addition, tandem repeats are used in population studies [12], conservation biology [11], and in conjunction with multiple sequence alignments [1, 6].

In [10], Landau, Schmidt and Sokol presented an algorithm to find all tandem repeats within a sequence. The algorithm has been implemented, and made available on the web [9]. However, it turns out that when this program is run on a biological sequence, the output is so large that it is extremely difficult to analyze. Although each reported repeat satisfies the definition of a tandem repeat, the repeats are fragmented and unorganized, resulting in huge amounts of data, much of which is useless. For example, when the program was run on only the first 20,000 base pairs (bp) of the human chromosome 18, it produced 2,177 repeats.

*This work has been supported in part by the PSC-CUNY Research Award number 67217-00 36.

This problem is an instance of a general phenomenon that occurs when computational tools are used to facilitate research in biology. Whereas, previously, a biologist could spend an entire life studying a single gene, the current available tools can produce so much data, that it is difficult to know where to begin analysis. Thus, it is important that data produced by computational tools is organized in an effective and useful manner.

The goal of this work was to impose an organization on the data produced by the program of [10]. We analyzed the output from several different sequences and noted many important observations. We used these observations to develop a set of criteria by which to reduce the output size. We then implemented these criteria as a post-processing stage for the program. The result is a succinct subset of the reported repeats, which represents all repeats that are found in the sequence. As shown in section 2.4, for many of the test sequences the number of repeats reported was reduced by about 99%.

In this paper we describe the main techniques that we use to filter the repeats found by [10]. Furthermore, we announce a new website¹ for the program. The remainder of this paper is organized as follows. In the rest of this section, we discuss preliminaries and related work. In the following section, we describe the filtering criteria and techniques. In section 3 we present results of our program on actual biological sequences, and in section 4 we conclude with a discussion of the results.

1.1 Preliminaries

In this section we review some known definitions on periodicity, and the relation of periodicity with tandem repeats. Given a string $S = s_1 s_2 \dots s_n$, S is *periodic* if $S = \pi^\ell \pi'$ where $\ell \geq 2$, and π' a (possibly empty) prefix of π . A string S is *cyclic* in string π if it is of the form π^ℓ , $\ell > 1$. In other words, a cyclic string is a periodic string with an integral number of periods. A *primitive* string is a string which is not cyclic in any string.

Let $S = \pi^\ell \pi'$. If π is primitive, then π is called *the period* of S , and $|\pi|$ is the period size of S . Where the context is clear, we use the word “period” to mean both the string π , and $|\pi|$.

A tandem repeat can be viewed as a periodic string, in which the periods of the repeat are not exact. When only mismatches are allowed, the periods all have equal length (except possibly the last period). The definition of a tandem repeat in [10] is built upon the Hamming distance, and it allows up to k mismatches, where k is a given threshold. Mismatches in a repeat (using the definition of [10]) are not counted by counting character replacements. Rather, they are counted by creating a multiple sequence alignment of the periods of a repeat. Each row in the alignment is a period of the repeat, without any inserted gap characters. Column j consists of all characters at position j in each period. A mismatch is counted for every column in the alignment that contains occurrences of two or more distinct characters. Thus, there are k mismatches in a tandem repeat if its multialignment contains k *error columns*, i.e. k columns that are not uniform.

Example: The following repeat has period 11, length 27, and 3 errors.

```

235  AACCCCTACCC 245
      | |      |
246  TAACCCTAACC 256
      | |
257  CAACC      261
```

The program of [10] reports only the k -maximal and k -primitive repeats. A repeat within a sequence is said to be *k-maximal* if it cannot be extended either to the left or to the right with matching characters, and $\leq k$ errors. A repeat is *k-primitive* if it has the smallest period size over all possible periods for a given repeat, or if it has fewer errors than the smallest period size repeat.

¹We have developed a mirror site in the U.S. at <http://www.sci.brooklyn.cuny.edu/~sokol/trepeats/>.

1.2 Related Work

TRDB: A popular software tool for locating tandem repeats is Tandem Repeats Finder, developed by Benson [2]. The program first collects short exact repeats, called k -tuple matches (in practice, 5-7 bp). It then uses a collection of statistical criteria to extend and combine the matches, detecting statistically significant tandem repeats. We have used Tandem Repeats Finder for purposes of evaluation.

MREPS: The `mreps` software [8] is based on the algorithm presented by Kolpakov and Kucherov [7]. Similar to current paper, `mreps` uses a combinatorial algorithm to detect tandem repeats, and then heuristics to filter the found repeats. In their paper, only sample repeats are reported; a table of all repeats is not included. Perhaps this software will be usable as an add-on to `mreps` as well.

SWAN: The tools discussed thus far all detect repeats that are highly conserved. The goal of SWAN [3] is to detect highly divergent repeats. In addition to detecting divergent repeats, SWAN contributes combinatorial formulae for evaluating the statistical significance of a tandem repeat.

2 Methods

The goal of this paper is to reduce and organize a large set of tandem repeats into a meaningful and manageable set of repeats. The following 4-tier approach is taken to produce a succinct table of the significant tandem repeats found in a sequence.

Algorithm Outline

- Step 1: Calculate statistics
- Step 2: Combine like periods
- Step 3: Minimum length filter
- Step 4: Shift filter

We describe these four stages in the following four subsections.

2.1 Calculate Statistics

The output of the program of [10] is given as a listing of the alignment of each found repeat, including the start position, end position, and number of mismatching columns. In order to determine criteria for the selection and combining of repeats, we calculate three scores for each repeat.

Number of Matches: A match is counted when a given character matches the corresponding character in the previous period. Formally, a match is counted when $\sigma = \tau$, where σ is the i th character in period number M and τ is the i th character in period number $M + 1$. We note that given two exact repeats that span the same substring, a smaller period will have more matches since it contains more periods.

Number of Errors: An error is counted for each pair of aligned characters that do not match. We deviate from the definition used in the original program, which counts an error for each non-uniform column (see section 1.1). Simply counting the error columns does not tell much about a repeat, since it gives no indication about its length or number of matches.

Percent Matches: Consider all possible opportunities to generate a match in a sequence. Let $total$ denote this number. Then, $total = matches + errors$. The percent matches equals $\frac{matches}{matches+errors}$.

2.2 Combine like periods

One of the main goals of this project is to consolidate fragmented data into a larger, coherent whole. Thus, the first stage of our post-processor is to combine repeats. Since the program of [10] works with the Hamming distance, each repeat has a fixed period size. In addition, insertions and deletions are not allowed. For these reasons, it only makes sense to combine repeats with the same period size.

We combine repeats that overlap, are adjacent, or are at a small distance. The distance is variable, depending upon the length of the repeat. Currently, we are allowing 5% of the length of a repeat between neighboring repeats. Following these rules, the task of combining repeats is fairly straightforward. We sort the repeats by period size, and secondarily by starting point. We begin with the leftmost repeat, and concatenate neighboring repeats. After concatenation, we realign the repeat, and calculate its new statistics (as in section 2.1).

2.3 Minimum length filter

The minimum length filter removes all repeats that are shorter than a prespecified length. The idea is that a repeat that is very short, and was unable to be appended to a different repeat in the combine stage, is statistically insignificant.

The actual filter is a little more involved than a simple length test. The formula used for the minimum length filter is: $matches - errors \geq 10$. In words, we require a repeat to have 10 more matches than the number of errors.

In general, we allow the user to specify the minimum-length parameter; we added this as a new input to the program. We note that it is critical for the minimum length filter to be performed at exactly this point, following the combine, but before the shift filter. It must follow the combine, since a short repeat that can be combined with another repeat may prove to be significant. And, it must precede the shift filter, since the shift filter will drop some repeats in favor of others, and we want to be sure that a short repeat is not chosen to represent a group.

2.4 Shift filter

In the combine stage, neighboring or overlapping repeats with the same period are combined into a single repeat. As explained in section 2.2, it is impossible to combine repeats with different period sizes. Yet, to report all overlapping repeats would explode the size of the output. Hence, the shift filter deals with overlapping repeats that are of different period sizes.

Given two repeats that span exactly the same substring, it is possible to compare the repeats to determine which one is better. The number of matches and/or the percent matches can be used as the criteria for comparison. The difficult situation is when overlapping repeats only partially overlap. For example, a repeat is found spanning locations 1-64, and another spans 25-79. In the case of a shifted repeat, either choice would lose a substring of the input that should be included in the output. The idea of the shift filter is as follows.

We organize the repeats into groups of overlapping repeats. Within each group, we report three representatives of the group: the repeat with the leftmost start (LM), the repeat with the rightmost end (RM), and the “best” repeat (i.e. the one with the most matches). This ensures two important properties. First, all input characters included in a repeat are included in a repeat in the output. Secondly, a high scoring repeat is not missed, since it will be the best in its group.

The groups are constructed by dividing the repeats into intervals, as follows. The leftmost repeat that is found begins the first group. (In case of ties, the longest of the leftmost repeats is chosen.) All repeats that overlap the first repeat in a group are included in the group. The second group begins with the rightmost repeat in the first group. Add to the second group all repeats overlapping its first element. The process continues, until the list is exhausted. If the first element in a group is also the rightmost, the process begins anew, by choosing the leftmost repeat beginning after the endpoint of the last reported repeat.

As an additional filter, we put in a check comparing the best repeat in a group to the LM (resp. RM). If they overlap by more than 90% of the length of the LM (resp. RM), then only the best is reported.

The results of the shift filter were surprisingly good, as will be shown in the following section.

Table 1: The reduction in the number of repeats output is shown, for four complete chromosomes. The sequences were downloaded from <ftp://ftp.ensembl.org/pub/>.

Sequence	Length	Number of Original Repeats	Number of Post Filter Repeats
Chr.1 of <i>S.cerev.</i>	230,254 bp	18,973	267
Chr.12 of <i>Rattus N.</i>	1 st 1Mbp	43,507	471
Chr.Y of <i>Anopheles</i>	27,229 bp	2,325	13
Chr.4 of <i>Drosophila</i>	88,110 bp	6,042	55

3 Results

In Table 1, we contrast the number of repeats found by the program of [10], with the number of repeats reported following our post-processing stage. We begin with chromosome 1 of *S.cerevisiae*, which is cited by [8] as a classic sequence used for testing tandem repeats software. As shown in Table 1, using our post-processor, the number of repeats was reduced by about 99%. Furthermore, the reported repeats represent the significant repeats in the sequence, as shown in the example of Table 2.

In Table 2, we show the output of our program for the first 20,000 bp of human chromosome 18. The definitions of the statistics included in the table are described in section 2.1. There were a total of 26 repeats reported by our program, while in the original output there were 2,177 repeats. We used Tandem Repeats Finder [2] (described in section 1.2) as the evaluation tool. For this sequence, Tandem Repeats Finder reports 10 repeats, all of which have counterparts in our output.

4 Discussion

We have presented a software tool that organizes and filters the tandem repeats found in a sequence. With the use of our tool, it is now possible to apply the tandem repeats program of [10] to process entire chromosomes. This has been previously impossible, due to the large amounts of fragmented data that was reported.

We also believe that the ideas presented in this paper will be useful for other tandem repeats programs. Specifically, we are currently developing a more general program, which will allow for insertions and deletions within the periods of a repeat. Our hope is that this filter will prove useful for the more general definition as well.

Table 2: The 26 repeats reported from the first 20,000 bp of chromosome 18 of the Homo Sapiens. The original program reported 2,177 repeats in this sequence.

Start	End	Length	Period	Copies	Errors	Matches	Percent
1	653	653	6	108.8	84	563	0.87
929	1101	173	29	6.0	4	140	0.97
1048	1147	100	35	2.9	2	63	0.97
1147	1164	18	6	3.0	1	11	0.92
2152	2182	31	13	2.4	4	14	0.78
4251	4278	28	4	7.0	7	17	0.71
5017	5084	68	32	2.1	4	32	0.89
5336	5361	26	12	2.2	2	12	0.86
5573	5616	44	11	4.0	11	22	0.67
6261	6273	13	1	13.0	0	12	1.00
8575	8606	32	13	2.5	4	15	0.79
10301	10335	35	6	5.8	8	21	0.72
10325	10358	34	16	2.1	4	14	0.78
11549	11563	15	1	15.0	0	14	1.00
11783	11836	54	4	13.5	13	37	0.74
11804	11838	35	17	2.1	4	14	0.78
13119	13223	105	4	26.2	20	81	0.80
13155	13224	70	8	8.8	15	47	0.76
13565	13610	46	11	4.2	12	23	0.66
13570	13610	41	14	2.9	8	19	0.70
14695	14717	23	7	3.3	2	14	0.88
16433	16446	14	3	4.7	0	11	1.00
16563	16578	16	1	16.0	0	15	1.00
17295	17305	11	1	11.0	0	10	1.00
19135	19170	36	17	2.1	4	15	0.79
19518	19576	59	10	5.9	18	31	0.63

References

- [1] G. Benson. Sequence alignment with tandem duplication. *J. Comp. Biology*, 4:351–367, 1997.
- [2] G. Benson. Tandem repeats finder – a program to analyze DNA sequences. *Nucleic Acids Research*, 27:573–580, 1999.
- [3] V. Boeva, V. Makeev, and M. Régnier. SWAN: searching for highly divergent tandem repeats in DNA sequences and statistical significance. In *JOBIM'04*. IEEE Computer Society, 2004. In Proceedings JOBIM'04, Montréal.
- [4] C. T. Caskey et al. An unstable triplet repeat in a gene related to Myotonic Dystrophy. *Science*, 255:1256–1258, 1992.
- [5] A. J. Jeffreys. DNA typing: approaches and applications. *Journal of the Forensic Science Society* 33, pages 204–211, 1993.
- [6] H. Kitada, K. Tono, M. Yamamoto, T. Mitamura, A. Ohuchi, T. Ohyanagi, and N. Matsushima. Multiple alignment of biological sequences containing tandem repeats. *Genome Informatics*, 7:276–277, 1996.
- [7] R. Kolpakov and G. Kucherov. Finding approximate repetitions under hamming distance. In *9-th European Symposium on Algorithms (ESA), Lecture Notes in Computer Science*, volume 2161, pages 170–181, 2001.
- [8] R. Kolpakov and G. Kucherov. mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research*, 31:3672–3678, 2003. <http://www.loria.fr/mreps/>.
- [9] G.M. Landau. A Library for Computational Biology Programs. Available at: <http://cswb.cs.haifa.ac.il/library/> and <http://www.sci.brooklyn.cuny.edu/~sokol/trepeats/>.
- [10] G.M. Landau, J.P. Schmidt, and D. Sokol. An algorithm for approximate tandem repeats. *Journal of Computational Biology*, 8:1–18, 2001.
- [11] G. Spong and L. Hellborg. A near-extinction event in lynx: do microsatellite data tell the tale? *Conservation Ecology*, 6(1):15, 2002. <http://www.consecol.org/vol6/iss/art15/1>.
- [12] M.W. Uform and R.K. Wayne. Microsatellites and their application to population genetic studies. *Current Opinion in Genetics and Development*, 3:939–943, 1993.