International Journal of Computer Applications (0975 - 8887)
Volume 146 - No.9, July 2016

An Improved Method to Identify Exact and Approximate Tandem Repeats in DNA Sequences using Biclustering

Pamela Vinitha Eric
Department of Information Science and Engineering
Rajiv Gandhi Institute of Technology
Bangalore, India

Kusum Rajput
Department of Information Science and Engineering
Rajiv Gandhi Institute of Technology
Bangalore, India

Gopakumar G.
Department of Computer Science and Engineering
National Institute of Technology Calicut
Kerala, India

ABSTRACT

Tandem repeats occur frequently in eukaryotic and prokaryotic genomic sequences. They are associated with several inherited human diseases, DNA fingerprinting, evolution and regulatory processes. In spite of their importance, detection of tandem repeats is still not resolved in the sense that the current existing detection tools do not give the same results for a given input sequence. This is mainly due to the differences in the methods adopted by the search algorithms and the different parameter settings needed when they are executed. This paper proposes an efficient method to identify all exact and approximate tandem repeats within a given DNA sequence and also identifies the presence of any changes brought about by mutation. The method first identifies all potential tandem repeats by clustering using K-means method, followed by biclustering to filter out the actual repeats along with the position of occurrance of approximate tandem repeats. The results obtained by this method are consistent with that of existing methods.

Keywords

Tandem Repeat, DNA Sequence, Micro Satellites, Mini Satellites, Clustering

1. INTRODUCTION

Repeats in DNA sequences consist of short sequences or patterns that recur. These recurrences can be either interspersed or tandem. Interspersed repeats are long, less frequent and far apart whereas tandem repeats are consecutive occurrences of a pattern that are adjacent to each other. These repeated patterns are either exact copies or there may be insertion or deletion or substitution of a base, due to mutations. Accordingly there are exact and approximate tandem repeats. Repeated sequences make up more than 50% of the human genome [1].

Tandem repeats are further classified into microsatellites, minisatellites and satellite DNA [2] based on the period of the repeat. Repeats consisting of less than 1 to 6 nucleotides are referred to as microsatellites. These are more frequently found and are mostly exact repeats. Minisatellites are those that have a period size greater than 6 nucleotides and satellite DNA have greater than 100 nucleotides. It is observed that minisatellites and satellite DNA are mostly approximate repeats. Tandem repeats are involved in various regulatory mechanisms like protein binding [3], they affect the chromatin structure and are also involved in heat-shock inducible expression mechanism [4]. Tandem repeats are related to many human diseases caused by genetic disorders such as Huntingtons disease [5], fragile-X mental retardation [6], myotonic dystrophy [7], spinal and bulbar muscular atrophy [8] and Friedreichs ataxia [9]. These repeats also help in determining the parentage and individual inheritance traits [10]. Certain changes in the frequency of these repeats can cause cancer [11].

Detection of tandem repeats is very important but there are few methods that do this efficiently. Imperfect conservation of patterns and the complex structure makes it difficult to identify tandem repeats. Moreover the output obtained for a given input varies from method to method. Most tandem repeat detectors produce different, often nonoverlapping inferences, reflecting characteristics of the underlying algorithms rather than the true distribution of Tandem Repeats (TRs) in genomic sequences [12].

The proposed method uses K-means algorithm [13,14] for detecting potential tandem repeats and biclustering is used to identify exact and approximate tandem repeats. Clustering groups a given set of patterns into disjoint clusters, such that the patterns within a group are more similar to each other than the patterns in different groups. Euclidean distance is used to find distance between data points and centroids.

The biclustering method was introduced by Cheng and Church [15] and is based on mean squared residue score. Biclustering finds subgroups of rows and columns which are as similar as possible to each other and as different as possible to the rest [16].

This paper is organized as follows. Section 2 reviews various tandem repeat identification methods, section 3 describes the proposed

method, section 4 specifies implementation details and the results, which is followed by conclusion in section 5.

2. RELATED WORK

The different approaches to tandem repeat identification can be broadly classified as combinatorial approach or statistic or heuristic approach. Combinatorial methods employ exhaustive searching techniques to identify subsequences which are then compared with adjacent subsequences to identify tandem repeats. Statistic or heuristic methods uses small windows to detect possible repeats and then uses them to search for longer tandem repeats. Once tandem repeats are detected they are filtered so that biologically significant repeats are extracted. TRF[17], Mreps[20], Sputnik[25], STAR[26], RepeatMasker[23], BWtrs[19], Treks[22] and MCMC[21] are a few tools that are popularly used to detect tandem repeats. Tandem repeat finder (TRF) adopts the statistic based approach and models tandem repeats as percent identity and frequency of indels between adjacent pattern copies, uses statistically based recognition criteria[17]. TRF reports a higher number of overlapping repeats (35.1%) since it allows up to three overlapping repeats to be reported [18].

BWtrs, based on Burrows-Wheeler Transform, searches for the exact occurrences of tandem repetitions in DNA sequences [19]. Mreps uses the Hamming distance model to find all approximate tandem arrays, redundancies are then eliminated and statistically insignificant repeats are filtered out [20]. Mreps has no limitation on the size of the repeated pattern, the major drawback is that it can handle only repeats with mismatches and not indels. Markov chain Monte Carlo (MCMC) is a full probabilistic model which uses the sequence motif model to detect the enriched dispersed pattern in multiple sequences [21]. T-REKS, designed for protein sequences, used for ab initio identification of tandem repeats, clusters lengths between identical short strings by using K-means algorithm[22]. RepeatMasker is based on a local alignment strategy and uses a list of pre-selected common motifs, stored in a reference database called RepBase, to scan a query for these sequences. Even though this is highly suitable for selective motif searches, it is not an effective substitute for more comprehensive search [23,24].

There is a need for a method to identify exact and approximate tandem repeats accurately, where mismatches and indels are accommodated. The proposed method is able to identify all tandem repeats and in case of approximate tandem repeats the amount of similarity can be chosen based on the requirement so that appropriate filtering can be done.

3. METHOD

This section describes the proposed method which is a heuristic based approach. Let S be a DNA sequence of length n over the alphabet $\sum = \{A, C, G, T\}$ and p be a substring of finite length in S. A perfect or exact tandem repeat T consists of a concatenation of two or more identical copies of p. The length of p is referred to as the period and the number of copies of p that is concatenated is the exponent. In the tandem repeat ACGACGACGACGACGAC, p is ACG, its period is 3 and the exponent is 17/3=5.6. The approximate tandem repeat, ACGAGGACGTACGAGAC has 5.6 inexact copies of the substring ACG, with a total of three mismatches at positions, 5 where G is substituted in place of C, 10 where a T is inserted and 15 where C has been deleted. If the substring p consists of only one symbol drawn from $\sum$ then we have a homo repeat. The substring AAAAAAAAAAAAAA consists of A concatenated 14 times, thus this is a homo repeat of length 14, p is A and period is 1. Our goal is to identify all exact and approximate tandem repeats with at most k mismatches, where k is a predetermined threshold.

The proposed method is similar to [22] in that it first filters out all homo repeats, chooses short strings SS, finds the distance between two identical SS and uses K-means algorithm for the initial identification of potential tandem repeats. The method is based on the notion that the distance between the repeats of a given symbol in a tandem repeat region gives the period of that tandem repeat.

Initially strings of 4 symbols are chosen from the given DNA sequence as SS and compared with the sequence. When a match is found the distance between these two substrings is taken as the short string length (SL). This procedure continues until all SLs of that SS are detected, this yields a set of all short string lengths (SSL) of that SS. Each SL can be a candidate for a tandem repeat. The SSL is now divided into K clusters using K-means clustering. From each cluster the substrings whose lengths are equal or close to the most frequent short string length of a given cluster are selected as these are potential tandem repeats. In case the cluster has several SLs that occur the same number of times, the shortest length is chosen. The SLs so chosen are referred to as FSL. The short strings selected in this manner are potential tandem repeats. Biclustering of these short strings will identify actual tandem repeats.

A data matrix is constructed from the substrings obtained after K-means clustering. Each a_{ij} of the data matrix is the ratio of the number of occurrence of a symbol in a given position within all repeats to the total number of repeats of a given FSL, Fig-1 shows the construction of data matrix. The Cheng and Church biclustering algorithm [15] is then applied on this data matrix to filter out the actual tandem repeats.

SS = AAGG FSL = 12

A	A	G	G	C	T	C	G	A	G	T	T
5/5	5/5	5/5	5/5	5/5	5/5	4/5	4/5	5/5	4/4	5/5	5/5

Fig-1 – Potential tandem repeats and the data matrix constructed from it.

A bicluster is a subset of rows and a subset of columns with high similarity. The similarity score or the mean squared residue (H) is used to measure the coherence of the rows and columns in a single bicluster. Given the data matrix A = (X;Y); a bicluster is defined as a uniform submatrix (I;J) having a low mean squared residue score. The mean squared residue score which checks against threshold value (δ) specified by the user is given as

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$$

Where

a_{ij} is an element of the data matrix representing the probability of occurrence of a given base at a given position in the i[th] potential tandem repeat,

a_{iJ} is the mean of row i, given by the equation

$$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij} \quad ,$$

a_{Ij} is the mean of column j which is given by the equation

$a_{Ij} = \frac{1}{|I|}\sum_{i\in I} a_{ij}$ and

a_{IJ} is the overall mean and is derived as

$a_{IJ} = \frac{1}{|I||J|}\sum_{i\in Ij\in J} a_{ij} = \frac{1}{|I|}\sum_{i\in I} a_{iJ} = \frac{1}{|J|}\sum_{j\in J} a_{Ij}$

When the value of H(I,J) is greater than the threshold, deletion takes place and if value is lesser than threshold a bicluster is found. Its a greedy method. Initially the bicluster contains all rows and columns. At each iteration, the values of a_{Ij}, a_{iJ}, a_{IJ} and H(I,J) are computed for reuse and a row or column that gives the maximum decrease of H is deleted. This process terminates when there is no further decrease in H. Now this bicluster is masked and the process continues. Each bicluster so identified corresponds to a tandem repeat.

Algorithm TR(S)

(1) Filter out homo repeats
(2) Repeat
(3) Choose the short string ($SS \geq 4$) and determine the various short string lengths(SSL). If an SL corresponds to a tandem repeat already detected then discard that SL.
(4) Choose K the number of clusters.
(5) Kmeans(K, SSL)
(6) For each cluster $k \leftarrow 1$ to K do
  (a) Bicluster(k)
  (b) Identify tandem repeat.
  (c) Mask this bicluster and continue to step 6.a
(7) Until there are no short strings available.

Algorithm Kmeans(K, SSL)

(1) Select K SLs at random as initial cluster centroids.
(2) Repeat
  (a) For each SL ϵ SSL
    i. Find distance from SL to each centroid
    ii. Assign SL to the nearest cluster
  (b) End for
  (c) Re-compute the cluster centroid by taking the mean of the elements of each cluster
(3) Until convergence

Algorithm Bicluster(k)

(1) For each FSL in the cluster extract the corresponding substring from the sequence and construct the data matrix for the cluster k.
(2) If no possible rows or no change in the data matrix then exit.
(3) Obtain $\alpha \geq 1$, a parameter for multiple node deletion and $\delta \geq 0$, the maximum acceptable mean square residue score.
(4) Compute a_{iJ} for all i ∈ I, a_{Ij} for all j ∈ J , a_{IJ} and H(I, J).
(5) While$H(I,J) > \delta$ do
  (a) Remove the rows i∈ I with

$\frac{1}{|J|}\sum_{j\in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 > \alpha$ H(I,J)

  (b) Recompute a_{Ij},a_{IJ} and H(I,J).
  (c) Remove the columns j ∈ J with

$\frac{1}{|I|}\sum_{i\in I}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 > \alpha$ H(I,J)

  (d) If no rows or columns have been removed then switch to single node deletion.

Algorithm Single-node deletion

(1) Find r the row $i \in I$ with the largest

$d(i) = \frac{1}{|J|}\sum_{j\in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$

(2) Find c the column $j \in J$ with the largest

$d(j) = \frac{1}{|I|}\sum_{i\in I}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$

(3) If $d(r) > d(c)$ then remove r from I else remove c from J
(4) Compute a_{ij} ∀ i, a_{Ij} ∀ j, a_{IJ} and H (I,J)
(5) Add the columns $j \notin J$ with

$\frac{1}{|I|}\sum_{i\in I}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \leq H(I,J)$

(6) Recompute a_{iJ}, a_{IJ} and H(I,J)
(7) Add the rows $i \notin I$ with

$\frac{1}{|J|}\sum_{j\in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \leq H(I,J)$

(8) For each row $i \notin I$ add its inverse if

$\frac{1}{|J|}\sum_{j\in J}(-a_{ij} + a_{iJ} - a_{Ij} + a_{IJ})^2 \leq H(I,J)$

(9) If I and J have not changed then return this bicluster .

4. IMPLEMENTATION

The method was tested on data sets in FASTA format. K-means clustering of the DNA sequence created several clusters and the basic pattern of these clusters so formed are compared. Experimental results show that as the length of DNA sequence increases the probability of occurrence of tandem repeats also increases. It is seen that the results obtained by the method is more accurate if the number of clusters are increased when the length of the DNA sequence increases. It was observed that a value of 10 for K was good enough for short sequences, while for long sequences a value of 20 gave optimum results. Fixing the value of K as 20 is ideal in cases where varying K may become cumbersome.

	No. of tandem repeats detected	
DNA sequence	Treks (sim= 0.8)	Proposed method
HUMDYSTROP	31	35
HUMGHCSA	63	58
HUMHBB31	104	96
HUMHDABCD	65	70
HUMHPRTB	69	74

Table-1 – Comparison of proposed method with TReks.

Once the clusters are formed, biclustering filters out the actual tandem repeats. While biclustering the value of δ is assumed to be 100 as it is a termination condition and the value of H(I,J) is observed to be above 100 normally for the substrings being biclustered. The amount of similarity required between the substrings that form a tandem repeat is decided by the parameter α. By varying α from 1 to 5 we can obtain tandem repeats with different similarity levels, where the similarity between substrings decreases with increase in α. It is observed that the tandem repeats obtained are very dissimilar (false positives) when the value of α is above 5. Also values

of $\alpha \leq 4$ ensured detection of true tandem repeats with potential biological meaning, hence the default setting of α is taken as 4.

The method was tested for several DNA test data like HUMHBB31, HUMDYSTROP, etc.. The number of tandem repeats detected by the proposed method for these benchmark sequences was found to be at par with TReks. Table-1 shows the comparison of performance of the two methods. Both the methods filter overlapping repeats. The output obtained for HUMDYSTROP is given in Fig-2.

5. CONCLUSION

The method successfully identified all approximate and exact tandem repeats of different lengths. An advantage of the method is that it is able to identify the position of occurrence of the repeat. Forming a consensus sequence will give additional information regarding the presence of any edit operations brought about by mutations along with the position and type of edit operation involved. This data can be useful for studies on diseases caused by mutations, finding phylogenetic relations, compression, etc,.

```
HUMDYSTROP:
Number of homorepeats - 3
Residue: T   Length: 21 Positions: 5779-5799
Residue: A   Length: 19 Positions: 7442 - 7460
Residue: A   Length: 19 Positions: 17179-17197

Number of repeats of 8 residues - 10
GACTTCTTGACTTCTCAAGGTCAT       Length: 24 Positions: 4479 to 4502
TAAAATAGTAAAATAATATATTCT       Length: 24 Positions: 5739 to 5762
TTTCTTTTTTTCTTTTTATAATTT       Length: 24 Positions: 6177 to 6200
GGATTAAAGGATTAATGTGTTTGA       Length: 24 Positions: 15249 to 15272
TTCTATCTTTCTATCCATCCAAGC       Length: 24 Positions: 18056 to 18079
CCCTCCCCCCCTCCCCCCACCCCA       Length: 24 Positions: 23188 to 23211
TGAAGAAATGAAAATAATGAGAAA       Length: 24 Positions: 26551 to 26574
AATATGATAATATGATACATTATA       Length: 24 Positions: 29583 to 29606
GTTTTTGTGTTTTTAAGTGTTAGG       Length: 24 Positions: 32056 to 32079
AAAAGTACAAAAAGTAAATTATTT       Length: 24 Positions: 32712 to 32735

Number of repeats of 9 residues - 10
CTCTCCCTCCCTCTCCTGCCTCTCTTCTTGCTCCTTC Length: 37 Positions: 2220 to 2256
CTATATAAACTATAAAGACAGAAGATATA  Length: 29 Positions:  7620 to 7648
TCTTGCTTCATCTTGCCATCTTGGATA    Length: 27 Positions:  12589 to 12615
TCATACAGTCATACATATCACCTAAA     Length: 26 Positions:  15928 to 15953
GAAGAAATTGAAGAAACAGTAGCAGCC    Length: 27 Positions:  17796 to 17822
AACTTTTCAACTTTCCCTCTTATACC     Length: 26 Positions:  21387 to 21412
TTAGGTCTTTTAGGTCCAAGAATACAA    Length: 27 Positions:  21729 to 21755
TTGTTCCCTTTGTGCTCTTTGTAATAA    Length: 27 Positions:  28037 to 28063
CCACTAAGAACCCACTTCAACCACTGCCA  Length: 29 Positions:  30504 to 30532
CAAAGGAAATGCAGACATGCAAGAAA     Length: 26 Positions:  34296 to 34321

Number of repeats of 10 residues - 2
ATTTTATCAAATTTAACAAAATATTTTATATC       Length: 32 Positions:  4085 to 4116
TCACTATGATCACTATGGAGTCACATTAAA         Length: 30 Positions:  30193 to 30222

Number of repeats of 11 residues - 5
GATGATAAACTAAGATGATAATAC       Length: 24 Positions: 6959 to 6982
GTTTACAGGGAGTGTTTAAAGTGG       Length: 24 Positions: 8402 to 8425
ATGTACATTCATGTGCTTCATGTTACTGAA Length: 30 Positions: 20065 to 20094
CTTATTATCCTGGCTTATTATCTA       Length: 24 Positions: 33445 to 33468
TCATCTAGTGCTTTCATAGTCCTT       Length: 24 Positions: 37834 to 37857

Number of repeats of 12 residues - 3
AGACACTTCCTCAGACACTTCCTT       Length: 24 Positions: 1508 to 1531
TTCTTTGTGTGTCTTCTGTGTCCAC      Length: 26 Positions: 23003 to 23028
TTAGAAAGAAACTTTTAGAAAGAAATTTATAAACTGAA Length: 38 Positions: 26252 to 26289

Number of repeats of 14 residues - 1
CATATTCTGACTTTCATATTCTGGGTTC   Length: 28 Positions: 21354 to 21381

Number of repeats of 18 residues - 1
TTCAGGGACATGTGACTATTCAGGGGAAACTA       Length: 32 Positions: 23620 to 23651

Number of repeats of 19 residues - 1
TTCCTCTTGTTATTTTCTATTTCCTCTGTTTCTATTTGT Length: 39 Positions: 21136 to
21174
```

Fig 2 – Output obtained for HUMDYSTROP

6. REFERENCES

[1] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R. "Initial sequencing and analysis of the human genome". Nature. 2001 Feb 15;409(6822):860-921.

[2] Richard, Guy-Franck, Alix Kerrest, and Bernard Dujon. "Comparative genomics and molecular dynamics of DNA repeats in eukaryotes". Microbiology and Molecular Biology Reviews 72.4 (2008): 686-727..

[3] Richards, R. I., Holman, K., Yu, S., Sutherland, G. R. (1993). "Fragile X syndrome unstable element, p (CCG) n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins". Human molecular genetics, 2(9), 1429-1435

[4] Leibovitch, Boris A., Quinn Lu, Lawrence R. Benjamin, Yingyun Liu, David S. Gilmour, and Sarah CR Elgin. "GAGA factor and the TFIID complex collaborate in generating an open chromatin structure at the Drosophila melanogaster hsp26 promoter". Molecular and cellular biology 22, no. 17 (2002): 6148-6157.

[5] MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, Barnes G, Taylor SA, James M, Groot N, MacFarlane H. "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes". Cell. 1993 Mar 26;72(6):971-83..

[6] Verkerk, Annemiske JMH, Maura Pieretti, James S. Sutcliffe, Ying-Hui Fu, Derek PA Kuhl, Antonio Pizzuti, Orly Reiner et al. "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, no. 5 (1991): 905-914.

[7] Fu, Y. H., A. Pizzuti, R1G1 Fenwick, J. King, S. Rajnarayan, P. W. Dunne, J. Dubel, G A. Nasser, T. Ashizawa, and P. De Jong. "An unstable triplet repeat in a gene related to myotonic muscular dystrophy". science 255, no. 5049 (1992): 1256-1258.

[8] La Spada, Albert R., Elizabeth M. Wilson, Dennis B. Lubahn, A. E. Harding, and Kenneth H. Fischbeck. "Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy". Nature 352, no. 6330 (1991): 77-79.

[9] Campuzano V, Montermini L, Molt MD, Pianese L, Cosse M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, Zara F. "Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion". Science. 1996 Mar 8;271(5254):1423-7.

[10] Manasatienkij, Chairat, and Chatchawin Ra-Ngabpai. "Clinical application of forensic DNA analysis: a literature review". Journal of the Medical Association of Thailand= Chotmaihet thangphaet 95.10 (2012): 1357-1363.

[11] Xia X, Rui R, Quan S, Zhong R, Zou L, Lou J, Lu X, Ke J, Zhang T, Zhang Y, Liu L. "MNS16A tandem repeats minisatellite of human telomerase gene and cancer risk: a meta-analysis". PloS one. 2013 Aug 22;8(8):e73367.

[12] Schaper, Elke, Andrey V. Kajava, Alain Hauser, and Maria Anisimova. "Repeat or not repeat?statistical validation of tandem repeat prediction in genomic sequences". Nucleic acids research 40, no. 20 (2012): 10005-10017.

[13] Elmasri, Ramez. "Fundamentals of database systems". Pearson Education India, 2008.

[14] Cormen, Thomas H., Charles E. Leiserson, and Ronald L. Rivest. C. Stein "Introduction to Algorithms". MIT Press 5.3 (2001): 55.

[15] Cheng, Yizong, and George M. Church. "Biclustering of expression data". Ismb. Vol. 8. 2000.

[16] Kaiser, Sebastian, and Friedrich Leisch. "A toolbox for bicluster analysis in R". (2008).

[17] Benson, Gary. "Tandem repeats finder: a program to analyze DNA sequences". Nucleic acids research 27.2 (1999): 573.

[18] Lim, Kian Guan, Chee Keong Kwoh, Li Yang Hsu, and Adrianto Wirawan. "Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance". Briefings in bioinformatics 14, no. 1 (2013): 67-81.

[19] Pokrzywa, Rafal, and Andrzej Polanski. "BWtrs: a tool for searching for tandem repeats in DNA sequences based on the BurrowsWheeler transform". Genomics 96.5 (2010): 316-321.

[20] Kolpakov, Roman, Ghizlane Bana, and Gregory Kucherov. "mreps: efficient and flexible detection of tandem repeats in DNA". Nucleic acids research 31.13 (2003): 3672-3678.

[21] Li, Qiwei, Xiaodan Fan, and Tong Liang. "An MCMC algorithm for detecting short adjacent repeats shared by multiple sequences". Bioinformatics 27.13 (2011): 1772-1779.

[22] Jorda, Julien, and Andrey V. Kajava. "T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm". Bioinformatics 25.20 (2009): 2632-2638.

[23] Smit, Arian FA, Robert Hubley, and P. Green. "RepeatMasker." Published on the web at http://www. repeatmasker. org (1996)..

[24] Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J." Repbase Update, a database of eukaryotic repetitive elements". Cytogenetic and genome research. 2005 Jul 21;110(1-4):462-7.

[25] Abajian, Chris. "Sputnik: DNA microsatellite repeat search utility". Program available at: http://epressoftware. com/pages/sputnik. jsp (1994).

[26] Delgrange, Olivier, and Eric Rivals. "STAR: an algorithm to search for tandem approximate repeats". Bioinformatics 20.16 (2004): 2812-2820.