TReaDS: Tandem Repeats Discovery Service

Alessio Vecchio
Dept. Ingegneria dell'Informazione
Univ. of Pisa, Pisa, Italy
a.vecchio@ing.unipi.it

M. Elena Renda
IIT-CNR
Pisa, Italy
elena.renda@iit.cnr.it

Marco Pellegrini
IIT-CNR
Pisa, Italy
marco.pellegrini@iit.cnr.it

Abstract. Tandem repeats (TRs) are multiple duplications of substrings in the DNA that occur contiguously, or at a short distance, and may involve some mutations (such as substitutions, insertions, and deletions). The analysis of TRs is an important genetic profiling technique. In fact, TRs can be used, for instance, to detect evolutionary phenomena in populations, to identify the cause of several diseases, and to help in determining parentage. There are several web-based resources or downloadable packages for finding TRs, but such tools rarely give exactly the same result for a given input. Thus, biologists could be interested in a tool that, not only gives them the possibility of querying multiple systems at the same time, but also simplifies the burden of comparing and merging the results. *TReaDS* (Tandem Repeats Discovery Service) is a tandem repeat meta search engine that finds exact, approximate, short and long TRs. *TReaDS* queries several web-based tools and merges their outcome into a single report, providing a global, synthetic, and comparative view of the different results.

Availability. *TReaDS*, the Tandem Repeats Discovery Service, is a web application free and open to all users without login requirement at the following URL: `http://bioalgo.iit.cnr.it/treads`.

Keywords: Genetic Profiling; Genomic Sequences; Exact, Approximate, Short, and Long Tandem Repeats Identification and Extraction.

I. Introduction

Repetitive structures in DNA sequences have attracted the attention of biologists since the discovery of satellite DNA in 1961 [10]. Tandem Repeats (named in several ways, such as microsatellites, minisatellites, Short Tandem Repeats (STR), Variable Number Tandem Repeats (VNTR)) have been studied extensively because of their role in several biological processes. Tandem repeats can be highly polymorphic, thus they are privileged targets in activities such as fingerprinting or tracing the evolution of populations. Since the 1990's, it was recognized that several diseases, disorders and addictive behaviors are linked to specific tandem repeats loci (and sometimes a causal relation can be established) (see e.g. [6] and references therein, [3], and [20]). More recently, TRs have been a target in the study of the evolution of populations (e.g. for humans [9], bacteria [18], protista [8], and multi-species [14]). Also, recent studies have considered the role of TRs within coding regions [15] and their relation to gene functions (see [14] and references therein).

The earlier studies relied on exhaustive enumeration, and were able to detect mostly short tandem repeats with zero or few errors allowed. The scope and depth of the research on tandem repeats have been boosted by the availability, starting from the late 1990's, of efficient non-trivial algorithms for finding TRs, even when mutations in the sequences may introduce non-negligible errors in the repeats. Tandem Repeat Finder [1], ETANDEM [16], Reputer [12], mreps [11], ATRHunter [19], CRISPRFinder [7], and Tread [17] are among the systems available via web interface that are currently operational.

Depending on the underlying algorithmic principle and on the choice of parameters, these systems rarely give exactly the same results for a given input. Comparative studies on the output of several tools, for the case of short tandem repeats with high substitution error rates, are presented in [2] and [13]; both studies report significant differences among the sets of detected TRs. Moreover, in [13] it is highlighted how critical it is the right choice of parameters. Thus, biologists could be interested in a tool giving them the possibility of simultaneously querying multiple systems and getting a global comparative and synthetic view of the results, with the same effort one would exert in using just one of the systems.

In a parallel development, databases of Tandem Repeats have been proposed so to facilitate annotation and cross-reference once new TRs are found (e.g. [6] and [5]). Multiple tools for searching Tandem Repeats have been used in building up the data base VNTRDB described in [4].

II. Methods

TReaDS - Tandem Repeats Discovery Service is a *TRs meta search engine* for querying several web services (each such service is usually based on a different algorithmic principle) for finding exact, approximate, short and long TRs. *TReaDS* allows the user (1) to simultaneously run different algorithms on the same data set, (2) to choose for each one of them different parameters and settings (3) to get back a report downloadable for further, off-line, investigations.

Currently, the publicly available web tools for finding tandem repeats supported by *TReaDS* are: Tandem Repeat Finder (TRF) [1], mreps [11], and Approximate Tandem Repeat Hunter (ATRHunter) [19]. We plan to add more existing systems (and new ones in time as they become available).

At the moment we produce reports in PDF, Excel, RTF and HTML format.

TReaDS is a web application completely developed with Java-based technologies. In particular, a pool of Servlets takes care of handling the users' request (file upload, parameter settings, search), and collects the results generated by the queried systems. *TReaDS* merges the results obtained and generates the final report with the support of the JasperReports publicly available libraries[1]. On the client side, there is no special requirement: just a standard browser and a viewer (for the selected format). *TReaDS* has been tested under Internet Explorer, Firefox, and Opera web browsers.

III. Input/Output of *TReaDS*

TReaDS has the proper structure of a meta search engine, with options for changing the parameters set of each queried algorithm, and for choosing the output format. *TReaDS*' main page is essentially composed of four sections.

- **Algorithm section**. Here it is possible to choose any combination of the supported systems, and, for each of them, to change the standard input parameter setting. Take the TRF system as an example. After clicking on *Change Parameters* in the main page, the TRF's setting page is showed. In order to explain the specific meaning of any of the system parameters, an help page will pop up on user request. Note that the systems supported by *TReaDS* can be also downloaded, and that there can be interface differences between the web-based and the downloadable versions, especially in terms of the number of parameters to be set. *TReaDS* is interfaced with the version of these tools available on-line.
- **Parameter setting**. In *TReaDS* the user has the possibility of changing the parameter setting for each of the selected tool. A default parameter setting is provided, to be checked before starting the system.
- **Sequence section**. Here it is possible to submit a sequence as a file, or to paste the input sequence in a given text area. It is also provided a default sequence to load for trying the engine. *TReaDS* takes in input either a FASTA or plain text genomic sequence. The user has the option of including/omitting (all or part of) the input sequence in the final report. *TReaDS* pre-processes the input sequence in order to remove spaces, numbers, CR and LF characters. Then, if the sequence is in fasta format, *TReaDS* checks that all characters belong to the set of allowed characters. Before querying the systems, *TReaDS* verifies if the sequence is acceptable according to the systems policies (ATRHunter accepts sequences containing only ACGT, mreps accepts sequences containing ACGTN where the percentage of N's is less than 5%, TRF accepts everything). If the sequence is not acceptable for a given system, that system will not be queried. The summary subsection of the report gives feedback on the motivation for de-selecting a tool.

[1] http://www.jasperforge.org/

Algorithm	Consensus	Start	End	Length	Repetitions
ATRHunter	CAATGTGCTTAC	1742	1814	12	6.0
mreps	TGCTTACCAGAT	1747	1820	12	6.08
TRF	TGCTTACCAGAT	1747	1820	12	6.1

TABLE I
EXAMPLE OF CLUSTERED TANDEM REPEATS: THE CLUSTER STARTING POSITION IS 1742, AND THE ENDING POSITION IS 1820. THE INPUT SEQUENCE IS THE STRAND S288C OF THE GENOME OF S. CEREVISIAE (270K BPS).

- **Format section**. Here it is possible to chose the final report format. The currently supported formats are PDF, Excel, RTF and HTML. Moreover, it is possible to set the length of the *flanking sequence* to include in the final report.

When the search button is clicked, *TReaDS* queries the selected systems with the input sequence and the chosen parameters. The output of the systems are collected and a report is produced, containing: information on the sequence submitted, the performance of each algorithm, and the list of the TRs found by each algorithm. Furthermore, *TReaDS* merges the results of all algorithms to give a global view of them. In fact, for all the TRs found, *TReaDS* identifies TRs with some overlapping. Formally, the overlap relationship among TRs induces a graph, and a cluster is a connected component of this graph. Graphically, a cluster covers a contiguous segment of the input sequence without gaps (Figure 1). In Table I it is shown an example of three partially overlapping TRs where each TR is found by a different algorithm. The cluster is identified by the segment with starting position 1742 (the lowest starting position) and ending position 1820 (the highest ending position).

In particular, *TReaDS* reports contain the following sections.

- **Header page**, with the execution date and time.
- **Sequence report**, containing information on the sequence submitted, such as its length, the A, C, G, T distribution, and the sequence itself (of part of it), if requested.
- **Summary report** of the systems queried: the algorithm name, the number of TRs found, whether the connection has been successful, the response time, and a chart showing a system comparison simply based on the numbers of TRs found. If the connection has not been successful, the type of error encountered is reported: *Malformed URL* in case of wrong system URL, *Connection error* in case of no connection is available, *Invalid input sequence/parameters*, in case the parameters are invalid or the input sequence is not accepted.
- **Algorithm reports**. There is one report for each queried system, composed of two sub-reports: the *parameters sub-report*, reporting the input parameters as set by the user, and the *result sub-report* with the list of the TRs found, reporting, for each TR, the initial position, the length, the number of repetitions, and the TR consensus. Each system might provide additional detailed information on each single TR found (e.g. a graphical display of

the alignments), which is not reported by *TReaDS*.

- **Clustered tandem report** (Figure 1). It contains a list of *clusters*, where each cluster is formed by a set of tandem repeats covering a segment of the input sequence without gaps. For each cluster it is reported the starting and ending positions of the covered segment, the list of tandem repeats it contains, and for each TR in the cluster its starting position, ending position, length, the number of repetitions, the algorithm(s) responsible for it, and the consensus of the TR. It is also reported the flanking sequence as requested by the user in the main search page. It is possible to view in graphical form, for each cluster, the correspondent segment of the input sequence, where lines of different color and shape are associated with the TRs found by each tool (Figure 2).

IV. Conclusion

As whole-genome TRs studies are being pursued it is important to be able to facilitate the use of several TRs search engines available. Merging and comparing the outcome of several search tools on the same data could be useful for gaining higher confidence that all the relevant TRs present in the data have been found. For this reason, we developed *TReaDS*, a Tandem Repeats Discovery Service that allows to harness the power of several Web-based Tandem repeats finding servers with a minimum effort.

References

[1] G. Benson. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580, 1999. http://tandem.bu.edu/trf/trf.html.

[2] V. Boeva, M. Regnier, D. Papatsenko, and V. Makeev. Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics*, 22(6):676–684, 2006.

[3] S. Bouffler, A. Silver, and R. Cox. The role of DNA repeats and associated secondary structures in genomic instability and neoplasia. *Bioessays*, 15(6):409–12, 1993.

[4] C.-H. Chang, Y.-C. Chang, A. Underwood, C.-S. Chiou, and C.-Y. Kao. VNTRDB: a bacterial variable number tandem repeat locus database. *Nucleic Acids Research*, 35(suppl 1):D416–421, 2007.

[5] F. Denoeud and G. Vergnaud. Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource. *BMC Bioinformatics*, 5(1):1–12, 2004.

[6] Y. Gelfand, A. Rodriguez, and G. Benson. TRDB - the tandem repeats database. *Nucleic Acids Research*, 35(Database-Issue):80–87, 2007.

[7] I. Grissa, G. Vergnaud, and C. Pourcel. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, 35(Web Server Issue), 2007.

[8] M. Imwong, D. Sudimack, S. Pukrittayakamee, L. Osorio, J. M. Carlton, N. P. J. Day, N. J. White, and T. J. C. Anderson. Microsatellite Variation, Repeat Array Length, and Population History of Plasmodium vivax. *Molecular Biology and Evolution*, 23(5):1016–1018, 2006.

[9] Y. D. D. Kelkar, S. Tyekucheva, F. Chiaromonte, and K. D. D. Makova. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research*, 18:30–38, November 2008.

[10] S. Kit. Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *Journal of Molecular Biology*, 3:711–716, 1961.

[11] R. Kolpakov, G. Bana, and G. Kucherov. mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research*, 31(13):3672–3678, 2003. http://bioinfo.lifl.fr/mreps/.

[12] S. Kurtz, J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, 29(22):4633–42, 2001.

[13] S. Leclercq, E. Rivals, and P. Jarne. Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics*, 125(8), April 2007.

[14] M. Legendre, N. Pochet, T. Pak, and K. J. Verstrepen. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research*, 17(12):1787–1796, 2007.

[15] C. O'Dushlaine, R. Edwards, S. Park, and D. Shields. Tandem repeat copy-number variation in protein-coding regions of human genes. *Genome Biology*, 6(8):R69, 2005.

[16] P. Rice, I. Longden, and A. Bleasby. EMBOSS: The european molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, June 2000.

[17] D. Sokol, G. Benson, and J. Tojeira. Tandem repeats over the edit distance. *Bioinformatics*, 23(2):e30–35, 2007.

[18] A. Vogler, C. Keys, Y. Nemoto, R. Colman, Z. Jay, and P. Keim. Effect of repeat copy number on variable-number tandem repeat mutations in escherichia coli O157:H7. *Journal of Bacteriology*, 188(12):4253–63, June 2006.

[19] Y. Wexler, Z. Yakhini, Y. Kashi, and D. Geiger. Finding approximate tandem repeats in genomic sequences. *Journal of Computational Biology*, 12(7):928–942, 2005.

[20] R. Wooster, A.-M. Cleton-Jansen, N. Collins, R. Mangion, J. Cornelis, C. Cooper, B. Gusterson, B. Ponder, A. von Deimling, O. Wiestler, C. Cornelisse, P. Devilee, and M. Stratton. Instability of short tandem repeats (microsatellites) in human cancers. *Nature Genetics*, 6(2):152–156, 1994.

Cluster number: 0 start: 891 end: 1279

Start	End	Length	Repetitions	Algorithm	Consensus
891	999	36	3.0	ATR	CGTTGGTGCTGGCAGTGGTAGTAGCATTAGTCCTGA
891	1279	36	10.8	TRF	CGTTGGTACTGGCAGTGGTAGTAGCATTAGTCCTGA
1007	1079	36	2.0	ATR	CTTTCAGTGGTAGTAGCACTAGTCCTGACGTTGATG
1103	1199	24	4.0	ATR	CTGG-AGTTGGTAGTCGCATTGGTA
1106	1178	36	2.0	ATR	G-AGTTGGTAGTCGCATTGGTACTGGCATTGGTAGTC
1109	1163	24	2.25	MREPS	TTGGTAGTCGCATTGGTACTGGCA
1142	1175	12	2.75	MREPS	GCATTGGTACTG
1238	1262	12	2.0	ATR	GCATTAGTACTG

Cluster number: 1 start: 4031 end: 4068

Start	End	Length	Repetitions	Algorithm	Consensus
4031	4068	17	2.2	TRF	AAGAAAACTCACTAAGAT

Cluster number: 2 start: 4266 end: 4288

Start	End	Length	Repetitions	Algorithm	Consensus
4266	4288	11	2.0	ATR	CCTACGCTTAG

Cluster number: 3 start: 4683 end: 5009

Start	End	Length	Repetitions	Algorithm	Consensus
4683	4701	8	2.25	MREPS	ACCCACAC
4683	4819	6	22.7	TRF	ACCCAC

Fig. 1. An example of the report showing the clustered tandem repeats found by all queried systems.

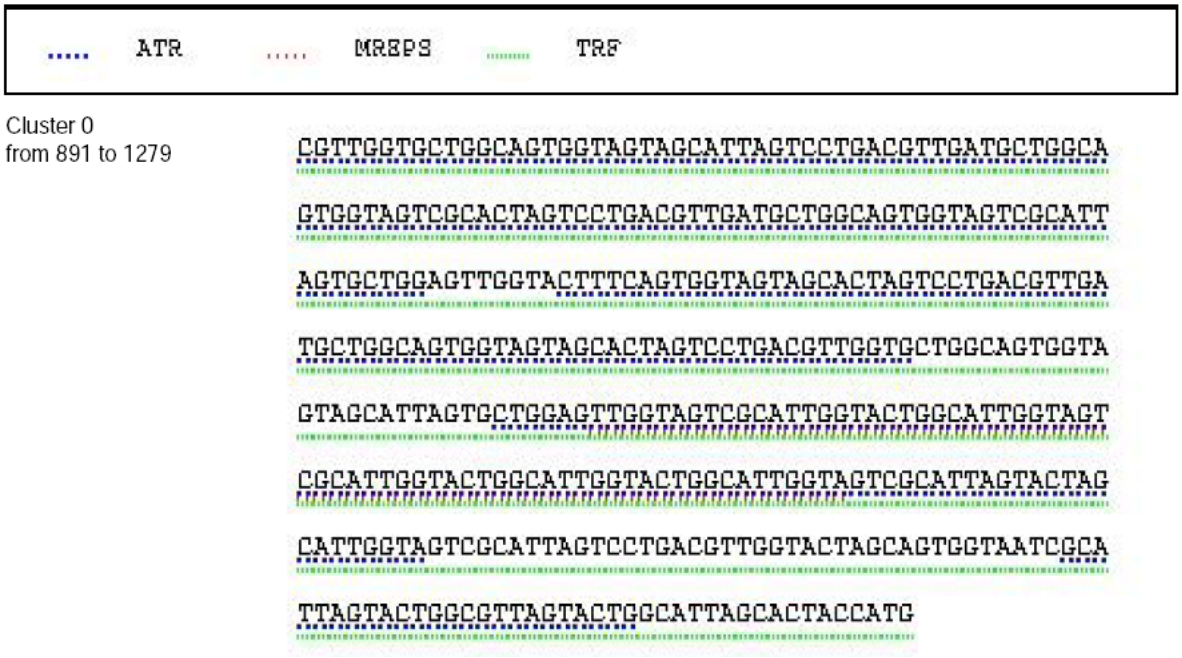

Fig. 2. The clustered tandem repeats in graphical form. TRs found by different algorithms are underlined with different color/symbol combinations.