TReaDS: Tandem Repeats Discovery Service

Renda M E (1), Vecchio A (2), Pellegrini M (1)
(1) Istituto di Informatica e Telematica del CNR, Pisa, Italy
(2) Dipartimento di Ingegneria dell'Informazione, Univ. di Pisa, Pisa, Italy

Motivation

Tandem Repeats (TRs) are multiple duplications of substrings in the DNA that occur contiguously and may involve mutations, such as substitutions, insertions, and deletions. Not only TRs are privileged targets in activities such as fingerprinting or tracing the evolution of populations, but they have also been linked to several diseases, disorders and addictive behaviors. Even though the depth of the research on TRs has been boosted by the availability of efficient non-trivial algorithms for finding TRs (even when mutations occur with non-negligible probability), comparative studies report significant differences among the sets of TRs that can be detected by using different tools and show how critical it is the choice of the input parameters. Thus, biologists could highly benefit from a tool that gives 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. Here we present TReaDS, the Tandem Repeats Discovery Service, which allows the user to: simultaneously run different algorithms on the same data set; manually choose for each algorithm a different parameter settings, or express her/his request in a simple and concise way (exact or approximate, short or long TRs), delegating to TReaDS the burden of choosing the right choice of parameters for all the systems; and get back a report that can be also downloaded for further, off-line, investigations. To the best of our knowledge TReaDS is the first meta search engine for tandem repeats and there is no similar and comparable system freely available.

Methods

TReaDS is a Java-based web application with the proper structure of a meta search engine. In particular, a pool of Servlets takes care of handling the users' and collects the results generated by the queried systems. The publicly available tools for finding TRs currently supported by TReaDS are: ATRHunter [5], mreps [3], TandemSwan [2], TRF [1], and TRStalker [4] (an algorithm developed by our team aimed at finding long fuzzy TRs under weighted edit distance). On the client side the only requirement is a standard web browser. The main page of TReaDS is essentially composed of four sections: Algorithms, Parameter Settings, Report, and Sequence: - Algorithms section: the user can choose any combination of the supported systems;- Parameter Settings section: the user can chose two ways to set the parameters for the chosen systems: the simple mode, where it is possible to specify the kind of TRs to look for, by setting the minimum

and maximum motif length, the minimum number of repetitions, and the maximum percentages of allowed substitutions, insertions and deletions; or the advanced mode, where the user can run each system with manually selected parameters, if she wants a fine-grained control over the settings;- Report section: the user can set a certain number of parameters on the final report creation, like the format among the available ones (HTML, Excel, PDF, RTF);- Sequence section: to submit a (FASTA or plain text genomic) sequence as a file or pasted in a given text area, and chose if the whole sequence or just a part of it must be analyzed. TReaDS merges the results received by the queried services and produces a final report with the JasperReports publicly available libraries. The report contains detailed information on the submitted sequence, and on the results returned by each queried system. Furthermore, TReaDS clusters the results of all algorithms giving a global view of them both in a textual and a graphical way. The user can provide a valid email address to receive the results via email.

Results

We run TReaDS, with different parameter settings, over the Saccharomyces Cerevisiae chromosome I left arm sequence (GenBank accession number U12980, 103, 681 bp long). The results show that the number of TRs found by the algorithms is subject to large variations. It is also worth noting that the sets of TRs found by each queried system can be scarcely overlapped. Furthermore, TReaDS makes clear the inclusion relationship between TRs even when they are produced by the same algorithm. For example, mreps often returns a list of results where shortest TRs are included in longer TRs. These results clearly show how TReaDS can simplify the search for TRs by using and combining the power of different techniques. Furthermore, merging and comparing the outcome of different search tools on the same data could be useful for gaining higher confidence that all the relevant TRs have been found. Through TReaDS we are currently investigating the relation between fuzzy TRs (TRs with high rate of divergence) and the well known tri-nucleotide disorders (as the ones responsible for the Huntington disease).

Availability

http://bioalgo.iit.cnr.it/treads

References

[1] Benson G: Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Research 1999, 27(2): 573-580.

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

[3] Kolpakov R, Bana G, Kucherov G: mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Research 2003, 31(13): 3672-3678.

[4] Pellegrini M, Renda M E, Vecchio A: TRStalker: an efficient heuristic for finding fuzzy tandem repeats. Bioinformatics 2010, 26(12): i358-i366.

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

Contact email
elena.renda@iit.cnr.it