Open Access

JSTRING: A Novel Java Tandem Repeats Searcher in Genomic Sequences with an Interactive Graphic Output

Valeria De Fonzo[1], Filippo Aluffi-Pentini[2] and Valerio Parisi[*,1]

[1]*Dipartimento di Medicina Sperimentale, Università di Roma "La Sapienza", Viale Regina Elena 324, 00161 Roma, Italy*

[2]*Dipartimento Metodi e Modelli Matematici, Università di Roma "La Sapienza", Via A. Scarpa 16, 00161 Roma, Italy*

Abstract: The search of Tandem Repeats in a genomic sequence is of growing interest but unfortunately the existing tools have the drawback of producing cumbersome results that require a painstaking interpretation. In order to help users interpret the results at first glance, we implemented JSTRING that shows the results also in a very clear graphic format, thus allowing their immediate survey.

The software is freely available at: http://bioinf.dms.med.uniroma1.it/JSTRING/.

Keywords: Tandem repeat, genome, graphic results, applet, Java application.

INTRODUCTION

The role of Tandem Repeats (TR) in a genomic sequence is of rapidly growing interest; see for example [1] and the references therein.

There exist a number of tools for finding TRs and exhibiting their repeated units (consensus). For some recent examples see [2-5].

A serious comparison between different tools should, of course, be performed by a third party against a sufficient number of carefully selected test problems, carefully tuning many different search parameters; meanwhile we can say that in the case of naïve runs with default parameters on easy problems, all the above programs appear to output basically equivalent results.

However, a major drawback of all such tools is that, when runing against very long sequences, they produce a large amount of cumbersome results, that require a painstaking interpretation.

In order to provide the user with some capability of grasping the TRs at first glance, we decided that the only way to do this is by means of a suitable graphic presentation of the results. And since, as far as we know, no tool with satisfactory graphic features is as yet available, a novel concept for the graphic presentation of the results has been devised.

IMPLEMENTATION

We implemented such a concept in JSTRING, a Java program for TR search, endowed with a rich interactive graphic user interface, that enables a quick grasp and a much easier interpretation of the results, and of course maintains the capability of printing results in a usual text form.

As for the TR search, we implemented a rapid and powerful algorithm based on a dynamic programming procedure, which requires polynomial search time, and exploits some heuristically plausible criteria to reduce the size of the search space; a similar approach is used for the STRING program [3].

In order to select the best alignment between a consensus and a TR, we evaluate the possible alignments by means of the classical simple additive scores: to each pair of aligned bases we assign 10 points for equality, while for inequality we assign a penalty which can be user-selected as mild (-30), medium (-50) or strong (-100). The choice between the above values is suggested by the user preference for more or less perfect TRs. The widely used affine penalty is assigned to each gap: a fixed penalty (-100) for its very existence and a proportionality factor (-10) for the gap length.

Two other tunable parameters are:

- max_gamma: is the maximum detectable length for the repeated unit in a TR; therefore its value should be fixed somehow larger (to allow for a moderate amount of insertions) than the maximum interesting length for the repeated unit.
- score: is the threshold above which the score of a TR is considered interesting and therefore the TR is accepted.

As for the graphic output, the gist of the novel concept is to show at a glance on a suitable Graphic Page – for all the results of a selectable tract of the sequence, or even for the entire sequence – a good deal of useful information about both single TRs and the collective distribution of TRs along the sequence; and this also in the case of a very large number of TRs, and with a great flexibility in the interactive management of the results. The implementation of this concept is briefly described in the following text and in more detail in the User Guide.

*Address correspondence to this author at the Dipartimento di Medicina Sperimentale, Università di Roma "La Sapienza", Viale Regina Elena 324, 00161 Roma, Italy; E-mail: valerio.parisi@uniroma1.it

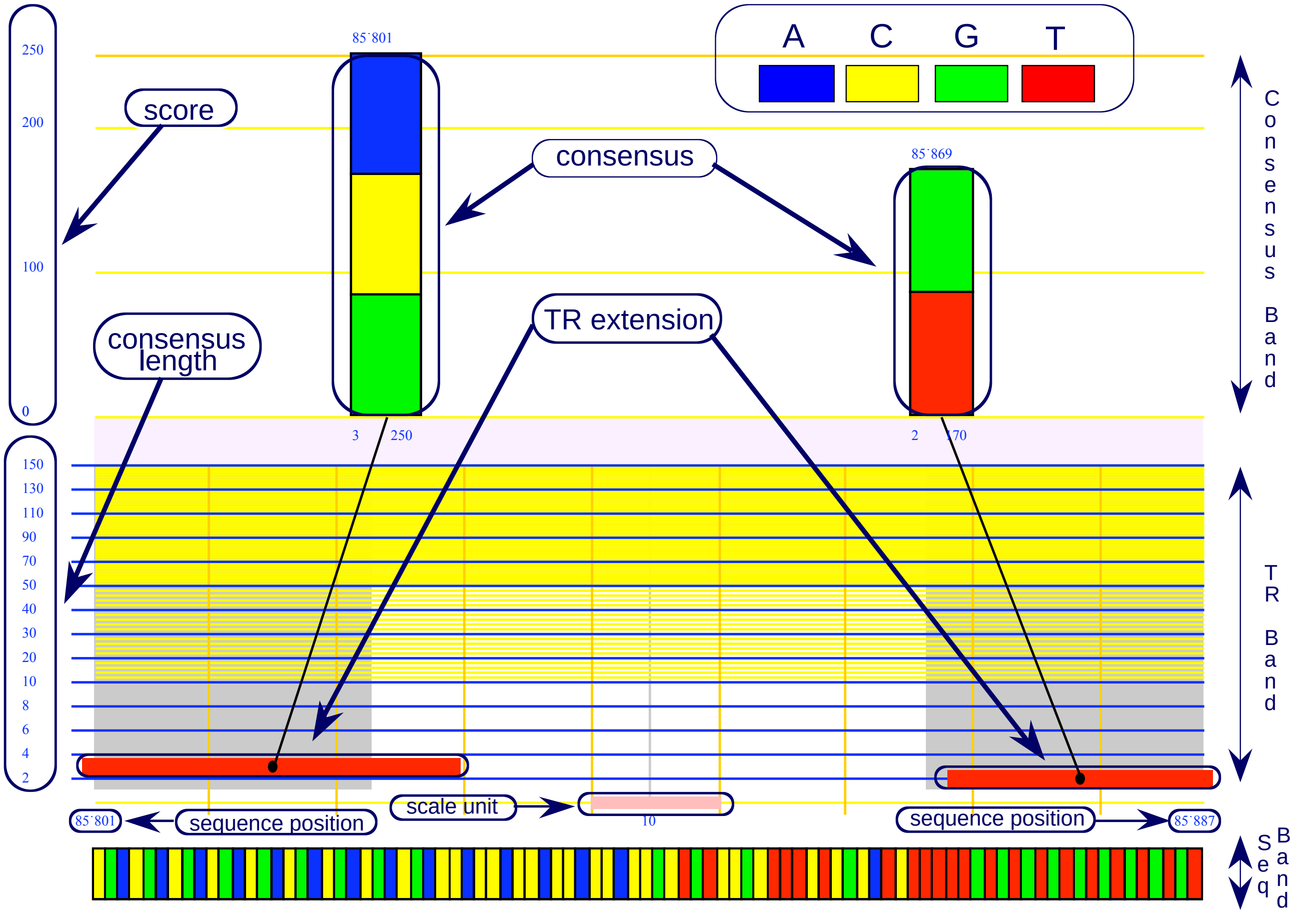

Fig. (1). A commented snapshot of the output for a small tract of the considered sequence (Leishmania major, see text) - between positions 85801 and 85887 - with two TRs, having a consensus respectively of three and two bp.

As good examples of "first sight" survey of the results we quote firstly the easy detection of the presence or absence of TRs in a given range of consensus lengths and in a given range of positions along the sequence; and secondly the possibility to see at once the base composition of a number of TRs.

The Graphic Page is made of three horizontal bands, i.e., from below, the Sequence Band, the TR Band, and the Consensus Band (Fig. **1**).

The Sequence Band displays (whenever graphically possible) only the selected tract of the input sequence without the results, but with the colour-coded bases giving already an idea of the prevailing base types.

The TR Band displays, in a cartesian plane, all the TRs as horizontal segments, whose location represents both the positions along the sequence (as abscissas) and the consensus length (as ordinate).

The Consensus Band shows, for the best results in the displayed tract, their score, and the (colour-coded) nucleotide sequence of the consensus.

In more detail, in the cartesian plane of the TR Band, each result is represented by a horizontal red segment with a central black dot: the abscissas of the extremities of the segment represents their position along the sequence, so that the length of the segment represents the length of the TR (although, in order to enhance the visibility, the segments undergo a small graphic enlargement).

The scale of the abscissas is a linear scale, simply indicated, below the TR Band, by the positions along the sequence of the selected tract end points and by a representation of the scale unit length, while the scale of the ordinates is an approximately logarithmic piecewise-linear scale.

The Consensus Band is a rectangular area where a consensus relative to a given result is represented as a vertical rectangle whose height is the score of the result and whose horizontal position is arbitrary (i.e. suggested only by graphic convenience), with a straight line connecting the consensus with the corresponding segment in the TR Band. Within the rectangle, the nucleotide bases of the consensus are represented by coloured slices, and are in top-down order.

Whenever possible, the numerical values of three important quantities are displayed in the Consensus Band: the TR initial position is displayed above the rectangle, while the consensus length and the score of the result are displayed below the rectangle (we note that these three quantities can be obtained also in analog form).

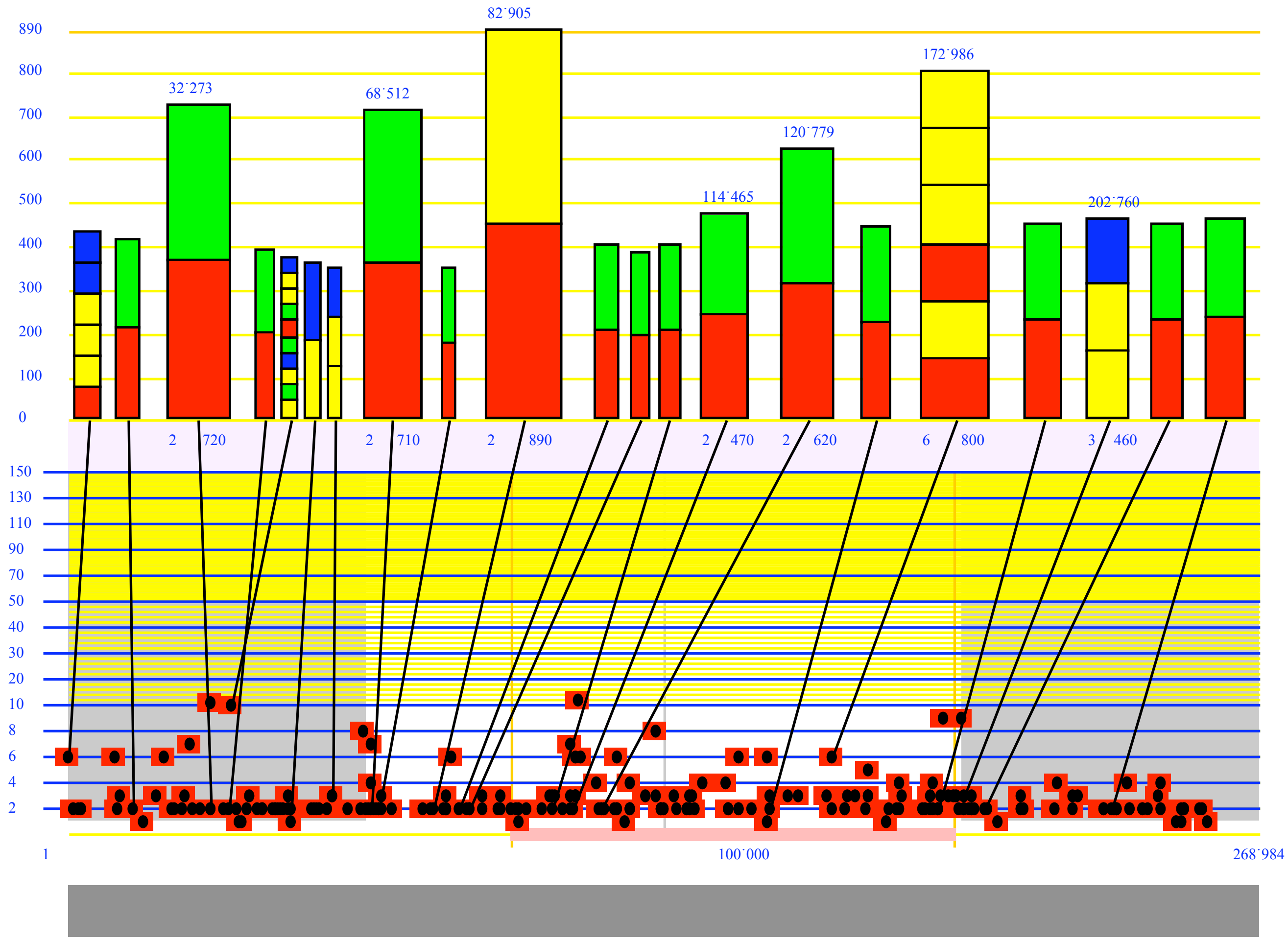

Fig. (2). A snapshot of the output for the entire considered sequence (268984 bp). The TR Band shows all the 238 TRs (possibly indistinguishably superposed), while the Consensus Band graphically displays the main features for only the best 21 TRs (with added numerical values for only the best 7).

When many consensuses are shown in the Consensus Band, the width of a rectangle gives a rough indication of its relative importance: the width is smaller if the score is smaller.

We note that while all results for the considered tract are displayed in the TR Band (although possibly partly or completely superposed), only a limited number of consensuses are displayed in the Consensus Band.

Of course, some features are shown only if graphic space allows.

Two other important features are that the Graphic Page are the following: it is interactive, allowing to repeatedly select a tract of the sequence in order to examine the results, and it is sensitive, in the sense that many choices for displaying the results can be performed by simply clicking in various areas of the frame. The interactive features of the sensitive Graphic Page (including zooming, moving the selected tract along the sequence and querying for a detailed description of some results) and all the possible commands (pushbuttons, keystroke shortcuts) are explained in detail in the User Guide.

As an example, we considered the complete sequence of the chromosome 1 of Leishmania major [GenBank:NC_001905].

With all the default values for parameters and dimensions, except for score=150 (chosen in order to get many TRs) and cmaxwin=2000 (as requested by a runtime warning), we found 238 TRs.

Fig. (**2**) is a snapshot of the Graphic Page, with the whole sequence results. A number of interesting results can be seen at a glance. For example, from the TR Band, one can see that there are no repeated units longer than 13 in the whole sequence, nor repeated units longer than 4 in the last part of the sequence (roughly above location 205.000). Moreover, one can see from the TR Band that many repeated units are dinucleotides, while from the Consensus Band (containing the highest-score consensus), one can see that most of such consensuses are GT dinucleotides.

CONCLUSIONS

The algorithm is implemented in the Java 1.1 release, for the sake of compatibility with all current operating systems and browsers.

A detailed User Guide is at: http://bioinf.dms.med.uniroma1.it/JSTRING/UG.

The byte-code JSTRING.class can be used both as an applet and as an application and is freely available at: http://bioinf.dms.med.uniroma1.it/JSTRING/. It can be immediately used, by any browser, as an applet (although with the constraints imposed by the sandbox philosophy).

JSTRING can be also used as a Java-interpreted application (without the sandbox constraints), using a Java Virtual Machine, with the advantage of the full functionalities.

JSTRING has been tested on a number of operating systems (for example some versions of Microsoft Windows and of Mac OS) with a number of sequences of various length, always obtaining valid results in a short time. For example, to give a rough estimation of the execution time, using an input file of length greater than 10 MB with a sequence length greater than 4 Mbp, a complete run required (depending on the used values for the parameters) from five to twenty minutes.

Besides the advantage of the graphic interactivity, others stem from providing the user with a Java byte-code. The program can be used locally (without burdening the central computer and the transmission lines as with some web pages, an asset still greater when using a graphic presentation), with no need of a risky compilation (as when the source is given) and no need to install the appropriate version of the executable (and to produce and maintain a version for each environment); moreover it is much more portable, reliable, and robust than compiled executable.

REFERENCES

[1] V. Parisi, V. De Fonzo, and F. Aluffi-Pentini, "A Survey of Dynamical Genetics," *Riv. Biol.*, Vol. 97, No. 2, pp. 223-253, May-August 2004.

[2] G. Benson, "Tandem repeats finder: a program to analyse DNA sequences," *Nucleic Acids Res.,* Vol. 27, No. 2, pp. 573-580, 1999.

[3] V. Parisi, V. De Fonzo, and F. Aluffi-Pentini, "STRING: finding tandem repeats in DNA sequences," *Bioinformatics*, Vol. 19, No. 14, pp. 1733-1738, September 2003.

[4] R. Kolpakov, G. Bana, and G. Kucherov, "mreps: Efficient and flexible detection of tandem repeats in DNA," *Nucleic Acids Res.*, Vol. 31, No. 13, pp. 3672-3678, July 2003.

[5] Y. Wexler, Z. Yakhini, Y. Kashi, and D. Geiger, "Finding Approximate Tandem Repeats in Genomic Sequences," in *Proceedings of the Eighth Annual International Conference on Computational Molecular Biology*, 2004, San Diego, California, USA, March 27-31, edited by P. E. Bourne, D. Gusfield, New York, ACM Press, pp. 223-232, 2004.

Received: February 25, 2007 Revised: April 2, 2008 Accepted: April 4, 2008