35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

A Software Tool for the Selection of Tandem Repeats for MLVA Analysis

Giuseppe D'Avenio, Mauro Grigioni, *Senior Member, IEEE*, and Roberta Creti

***Abstract*— DNA fingerprinting is a reliable tool for identifying, tracing and characterizing bacterial clonal population structure. A recent technique is given by the Multiple Loci VNTR Analysis (MLVA), where VNTR denotes Variable Number of Tandem Repeats, that meets the need for fast and reliable typing methods by combining the polymorphic nature of tandem repeats (TR) and the use of Polymerase Chain Reaction (PCR) methodology.**

The key point in the MLVA technique is the selection of a set of TRs which have a sufficient variability (polymorphism) across strains, in order to allow easy strain typing.

In this work, we present a program which analyses a set of N genomes and outputs the list of shared TRs and associated information. The program compares the TRs for each possible genome pair, and finds the sets of TRs that are shared by at least M genomes. The subsequent determination of "virtual amplicons" enables the user to consider the actual polymorphism exhibited by the different strains with regards to each given TR, which is the critical parameter for the experimental strain typing.

I. Introduction

DNA fingerprinting is a reliable tool for identifying, tracing and characterizing bacterial clonal population structure. Different typing techniques have been developed (e.g. MLST, PFGE) which differ in discriminative power, reproducibility and ease of interpretation.

Exploiting the fact that repetitive units of variable size can be found on multiple loci on the chromosome of different strains of a given bacterial species, an alternative approach has been recently proposed: the Multiple Loci VNTR Analysis (MLVA), where VNTR denotes Variable Number of Tandem Repeats, that meets the need for fast and reliable typing methods by combining the polymorphic nature of tandem repeats (TR) and the use of Polymerase Chain Reaction (PCR) methodology. Tandem repeats are found in DNA when a pattern of two or more nucleotides is repeated and the repetitions are adjacent to each other. Examining the sequence nucleotides (nt) flanking such repetitive patterns, suitable primers can be designed, and TR-containing amplicons can be obtained by PCR. Then, the size of the amplicons can be measured, and used to derive the copy number of each TR.

The key point in the MLVA technique is the selection of a set of TRs which have a sufficient variability (polymorphism) across strains, in order to allow easy strain typing.

The recent surge of the number of available sequenced bacterial genomes suggests that an automated search for suitable TRs across strains of a bacterial species can be carried out, sparing the time otherwise required for manually sorting the potential candidates for MLVA. In this work, we present a program which analyses a set of N genomes and outputs the list of shared TRs and associated information.

II. Materials and Methods

The program is organized in successive steps, detailed in the following.

Step 1 – collection of TR loci shared by different strains.

At first, the software tool inputs each genome sequence of interest to the Tandem Repeats Finder (TRF) software [1], with a suitable choice of the minimum repeat size and the conservation between the tandem repeats (in the following, we set minimum repeat size=10 bp and conservation>80%, i.e., a less restrictive search option than that proposed in [2]). The TRF program has detection and analysis components. The detection component uses a set of statistically based criteria to find candidate TRs. The analysis component attempts to produce an alignment for each candidate and if successful gathers a number of statistics about the alignment (percent identity, percent indels). Also other softwares for MLVA technique (e.g., [3]) leverage on the TRF, which is the most popular software for finding TRs in single genomes.

The output of the TRF is sifted through by a parser written in the Matlab programming language, and a list of TRs and relative attributes is created. These attributes are consensus pattern, copy number, genomic position of the TR loci, number of indels… If available, the annotations for each genome are used to tag each TR as intergenic or intragenic. As reported in [4], there has been an interest in finding TRs located in the intergenic regions of the genome, due to the complex role played by small intergenic repeat sequences in molecular and functional aspects of the bacterial cell. The software gives the possibility of checking the inter-/intragenic distribution of TRs for a particular set of bacterial strains.

G. D'Avenio is with the Department of Technology and Health, Italian National Institute of Health (ISS), Rome, Italy (phone: ++39-06-49902855; fax: ++39-06-49903096; e-mail: davenio@iss.it).

M. Grigioni is with the Department of Technology and Health, Italian National Institute of Health (ISS), Rome, Italy (e-mail: grigioni@iss.it).

R. Creti is with the Department of Infectious, Parasitic And Immune-Mediated Diseases, Italian National Institute of Health (ISS), Rome, Italy (e-mail: roberta.creti@iss.it).

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

After the phase of data collection relative to single genomes, the program starts a systematic comparison between genomes: each pattern (s1), considered as a the basic element of a candidate TR, in the list relative to the k-th genome, is juxtaposed to each pattern (s2) in the list relative to the h-th genome, forming the string s1s2, as shown in Fig. 1.

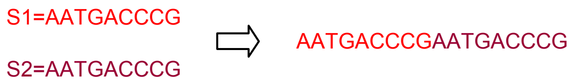

Figure 1 - Fragments from each genome sequence of a given pair are juxtaposed, to be input to the TRF

More precisely, the comparison is not of the type one-against-all, since this would entail an analysis time of O(N^2), N being the average number of TR in each genome. Since the positions of the TRs are very similar to each other, as far as different strains are concerned, we chose to compare s1 to s2, where $s2 \in S$, and S is a list of 21 consecutive TRs in the 2nd genome, with the central TR of the list being updated at each iteration of the procedure, in order to maintain a suitable search range in the 2nd genome. Thus, the complexity of the algorithm is of the order O(21*N)=O(N) When the same TR is considered in both lists, since s1 and s2 are the basic part of a repetitive sequence, the string s1s2 will represent a TR in itself, with a copy number=2, and the indices of the related TRs for both strains will be recorded. The program checks that s1 and s2 are closely related. The result of the comparison between strains is saved as a file for further analysis.

The next step is the construction of a subset of the TR list proposed by the TRF analysis on the entire genomes: since we are interested only in the TRs which are shared by different strains, only the TRs which can be found also in at least one other genome are retained. More precisely, the user can specify the minimum number M (<=N) of genomes that must share the same TR. The series of TRs is written in separate files, one for each TR, with all the relevant information for identification of its occurrence in each species. Each file contains a list of parameters, namely, the position of the TR locus in each genome, the consensus pattern,…

It must be recalled that the TRs typically output by a single-genome analysis are on the order of 250 (with Minscore =20 and minimum length for the base pattern=10 nt), which would lead to 250^2=62,500 comparisons between the TRs of each pair of genomes. Of course, this procedure should be repeated for every genome pair, being highly time-consuming.

Actually, less than such number of comparisons is performed by the program, which uses positional information in each genome to limit the search to a zone encompassing the position of the TR locus to be compared. Then, if the k-th TR in genome A is to be compared to the TRs in genome B, the program considers the TRs comprised between the (k-10)-th and the (k+10)-th element in the list of all TRs relative to genome B.

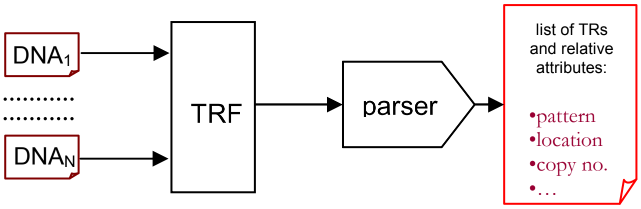

Figure 2 – The first step of the procedure yields a list of putative TRs, that constitute a potential link from each genome to the others.

Step 2 – Detection of "virtual amplicons"

For each TR identified by the previous step, the program verifies that the putative local polymorphism (in terms of number of repetitive elements), as given by the TRF's statistical approach, is actual, and not induced by, e.g., spurious variations of nucleotidic bases: in fact, especially for shorter lenghts of the base elements, a repetition can be detected by the TRF as having an inaccurate value of the copy number, according to the parameters' choice (in particular, to their restrictivity). In order to cope with such spurious matches, for each TR identified by the previous step the program checks these matches as follows. Starting from the position *x* in the genome (in nt) of the first TR, the program identifies a short nt sequence, or "virtual primer" (20 nt long) before and after the position *x*.

More precisely, the TR possessing the largest copy number, across the available strain sequences, is used the calculate the virtual primer downstream of the putative TR, owing to the fact that it has the largest distance from the identificative position of the TR itself., thus avoiding the selection of a downstream primer too close to the position *x*. Such a downstream virtual primer is then searched for, considering all of the TRs that had been identified, in step 1, as associated to the reference TR. The same operation is performed for the upstream virtual primer.

Thereafter, a couple of virtual primers is assigned to every strain sequence, and the length is calculated for the "virtual amplicon" potentially amplified by such a couple of primers. This allows to evaluate the effective polymorphism that an experimental trial would highlight. Thus, it is a more robust measure than that involving only the copy number data directly derived by the TRF, since it is not limited by statistical uncertainty.

For instance, let l_1, l_2,.. l_N the amplicon lengths associated to the TRs (TR_1, TR_2, .. TR_N), found in N different genomes, the polymorphism can be evaluated by means of the difference

$$\Delta = \max_i \{l_i\} - \min_i \{l_i\} \tag{1}$$

where $l_i \in \{l_1, l_2, \ldots, l_N\}$.

Alternatively, the relative variation can be calculated as follows:

$$r = \Delta / L, \tag{2}$$

L being the consensus pattern length (representative of the average repetitive element forming the TRs, across genomes). In both (1) and (2) the value of the parameter is not affected by the length of the virtual primers chosen as previously said. The length of the virtual amplicons, that, is the nucleotide distance between the virtual primers, was checked by using the BLASTN program. [4].

With objective parameters such as these, the user can rationally select TRs that will presumably give a satisfying experimental result.

For instance, in terms of relative variation *r*, high enough values can be looked for. The threshold for the relative variation is dependent on the specifications of the experimental technique used to detect the amplicon length. For not particularly sensitive equipments, a threshold of 2/3 can be high enough, in view of the necessity to have a sufficient signal-to-noise ratio (SNR), during the MLVA analysis. Of course, the choice of the most appropriate TRs for strain typing must be experimentally validated.

Evaluation of TRs in selected genomes, for MLVA Analysis

The program was used to analyze the TR distribution in available genomes of *Streptococcus pyogenes* (11 strains) and *Streptococcus agalactiae* (3 strains). For the former, the following sequences, downloaded from the NCBI website, were used:

- S py M1 GAS
- S py MGAS10394
- S py MGAS315
- S py SSI-1
- S py MGAS8232
- S py MGAS10270
- S py MGAS10750
- S py MGAS2096
- S py MGAS5005
- S py MGAS6180
- S py MGAS9429

For *S. agalactiae*, we analyzed the following:

- S. agalactiae A909
- S. agalactiae 2603
- S. agalactiae NEM316

III. Results

Figure 3 reports graphically the relationships between TRs shared by the different S. pyogenes genomes. A link in the graph denotes a common TR, belonging to both ends (i.e., genomes) of the link. It is evident how the selection of the appropriate TRs for MLVA analysis leverages on automatic evaluation of the TRs themselves, such as that enabled by the software tool herein presented, especially for a large number of sequenced genomes.

Recalling that M denotes the minimum number of genomes that share the same TR, with the choice M=2, 20 groups of shared TRs were found for *S. pyogenes* and 31 for *S. agalactiae*. Predictably, this number was found to decrease for increasing values of M.

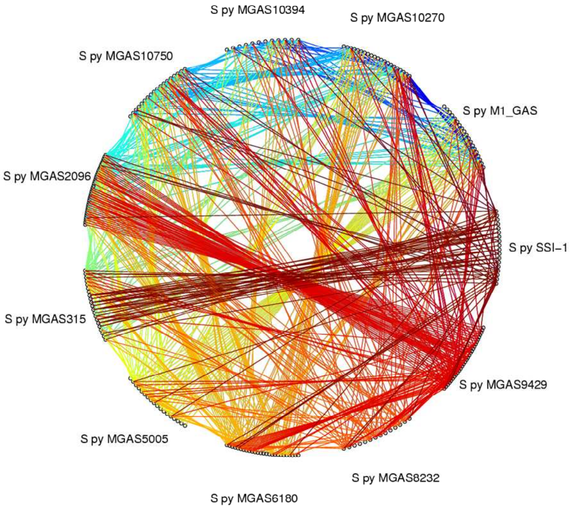

Fig 3 - An intricate web connects the TRs shared by different strains (N=11, S. pyogenes). A link in the graph denotes a common TR, belonging to both ends (i.e., genomes) of the link.

Figure 4 reports the relative variation (2) for the TRs found by the program (strain: *S. pyogenes* MGAS315). A total of 219 TRs were identified, shared by at least 2 of the 11 genomes considered. Some of these have a relatively large value of *r*, and are the natural candidates for the experimental validation of TR selection for MLVA analysis.

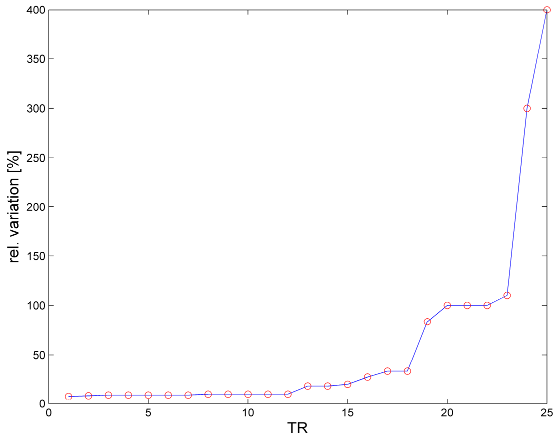

Fig 4 –TRs in S. pyogenes MGAS315, plotted after having been ordered in increasing of relative variation *r*.

In its current form, the software requires an input about the sequence to be considered as reference. Actually, the TRs belonging to multiple genomes must all have a common reference, and the most straightforward choice is to number them according to the list of TRs found in a a given strain. Figs. 5 and 6 show the different TRs associated to different reference strains (no restrictions was enforced about the relative variation of amplicon length).

In the selection of the reference strain, it is convenient to select to strain with the largest number of TRs in single-genome search.

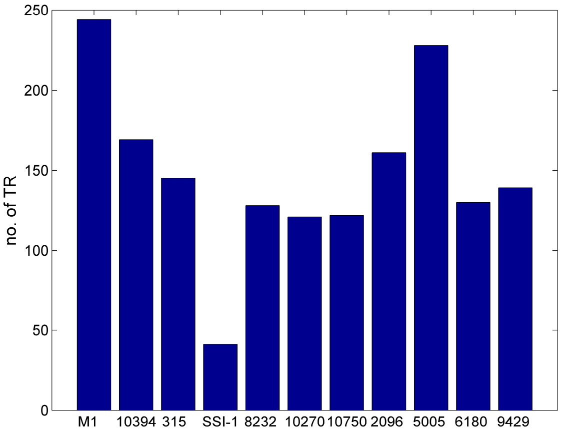

Fig. 5 – TRs found by the program, for the 11 genomes analyzed, considering S. pyogenes M1 as the reference strain.

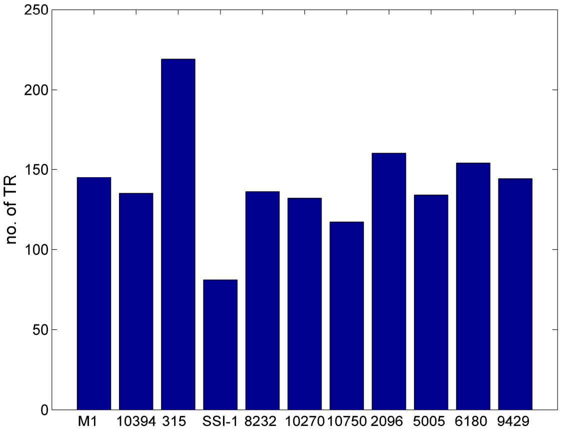

Fig. 6 – TRs found by the program, for the 11 genomes analyzed, considering S. pyogenes MGAS315 as the reference strain.

The software is meant to offer a multiple-genome comparison tool, in order to allow the experimenter to search for the most promising candidates for MLVA genotyping. Other tools have been previously presented, for this objective: an online comparison tool has been introduced in [3], addressing the selection of TRs across multiple genomes, using TRF-derived quantities such as unit length, copy number, total length, %GC, GC bias, %matches. The software reported in [3], nevertheless, does not offer figures of merit such as the relative variation r (Eq. 2) hereby introduced. Moreover, it does not address the problem of actual polymorphism of TRs.

Instead, in order to rule out inaccurate differences in TR lengths estimation, which can arise from the statistical nature of the analysis provided by the TRF software, we considered also the primers associated to each TR, introducing the concept of “virtual amplicon”, which renders the search for candidate TRs more amenable to experimental validation. In the approach hereby presented, the search is systematically carried out for all of the TRs output by the TRF tool, which satisfy a certain search criterion (minimum repeat size and conservation, i.e., %matches), the only limitation being the restriction to fairly close regions in different genetic sequences, in order to avoid excessive computation times.

The validation of the method has been done using a miniaturiazed electrophoresis platform able to size and quantify PCR fragments [6-7]. The results confirmed the predicted polymorphism of the TR loci.

IV. Conclusion

The presented program, which requires only the availability of the DNA sequences of interest, is able to provide the experimenter with all the relevant information for rationally choosing the TR loci for strain typing. Preliminary investigations proved that the proposed approach is capable of yielding useful data for MLVA typing of bacterial species.

References

[1] G. Benson, “Tandem repeats finder: a program to analyze DNA sequences.”, *Nucleic Acids Res*. 1999; 27:573-580

[2] J. Top, L.M. Schouls, M.J. Bonten, R.J. Willems, “Multiple-locus variable-number tandem repeat analysis, a novel typing scheme to study the genetic relatedness and epidemiology of Enterococcus faecium isolates”, J Clin Microbiol. 2004 Oct;42(10):4503-11.

[3] F. Denœud and G. Vergnaud, “Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource”, *BMC Bioinformatics* 2004 Jan 12;5:4.

[4] Delihas N. “Impact of small repeat sequences on bacterial genome evolution“ *Genome Biol Evol.* 2011;3:959-73.

[5] http://blast.ncbi.nlm.nih.gov/Blast.cgi

[6] Pittiglio V, Ciammaruconi A, D'Avenio G, Gherardi G, Pourcel C, Creti R. “A MLVA assay for genotyping of Streptococcus pyogenes.” MEEGID X Conference, Nov. 3-5, 2010, Amsterdam.

[7] Pittiglio V, Ciammaruconi A, D'Avenio G, Imperi M, Baldassarri L, Gherardi G, Visaggio D, Lista F, Pourcel C, Creti R. “A novel genotyping scheme based on MLVA for Streptococcus pyogenes”. In: XVIII Lancefield International Symposium. Abstract book. Palermo, Italy 4-8 Settembre 2011.