

Auteurs :

Sébastien Leclercq, Philippe Jarne & Eric Rivals

Titre :

Detecting Microsatellites within genomes : No exact solution

Abstract :

Microsatellites are short tandem repeats (period of 1 to 6 pb) that are present in the genomes of all living organisms. For some species, they account for a significant DNA proportion, with approximately 3% of the *Homo sapiens* genome for example (*International HGSC, 2001*). Some of these elements have a remarkable hypermutability, with an average mutation rate of the order of 0.001 for the human (*Goldstein & Schlötterer, 1999*), which is primarily caused by insertion or deletion of one or more repeats. These length variations are the consequence of a specific molecular mechanism named DNA slippage, which is not well understood yet (*Goldstein & Schlötterer, 1999*).

Microsatellites are extensively used as molecular markers since many years, but the question of their evolution started to be studied a dozen of years ago. One technique is to compare length theoretical distributions (generated from mutation models) to real distributions. The latter are obtained from known microsatellite loci (*Jarne et al 1998, Rolfsmeier et al 2000, Dettmann et Taylor 2004*), or by extraction from genomic sequences (from Genbank) either with personal algorithms (*Kruglyak et al. 1998, Dieringer et Schlötterer 2003, Sainudiin et al. 2004*) or with dedicated softwares based on advanced algorithmic notions.

More than a dozen of these algorithms were published since 1997, without counting dedicated databases. It is possible to group them into 4 major classes :

- methods of alignment against a consensus sequence
- combinatorial algorithms of repeat identification
- heuristic approaches based on statistical criterias
- methods based on the compression capacity of repeated sequences.

We propose here to expose some of these algorithms, and to compare major differences. Four softwares were chosen, each representing one of the above classes : RepeatMasker (<http://repeatmasker.org>) for the sequence alignment, Mreps (*Kolpakov et al 2003*) for the combinatorial method, TRF (*Benson 1999*) for the statistical method and finally STAR (*Delgrange & Rivals 2004*) for the compression method.

Each software have specific parameters, constraints and output formats, that impose to normalize datas before doing inter-algorithm comparisons. These comparisons are based on 4 microsatellite features : their length, their perfection degree (i.e. the percentage of mutation), the repeat length and the chromosomal position.

First observations show that, on the scope of a single algorithm, parameter choice can have a significant influence on detected microsatellite distributions. For example, TRF detection number can vary by a factor 20 simply by changing the minimum score parameter. To take these variations into account in the inter-algorithm comparison, we chose the STAR distribution as a reference (STAR does not take parameter), and we calibrated the parameters for each other algorithm to obtain a distribution the closest to this reference.

Results for inter-algorithm comparison on the human X chromosome show a significant detection divergence. TRF and Mreps detect much more tandem repeats than STAR and RepeatMasker, and particularly for small lengths. On the other hand, Star and TRF are more stringent for highly degraded microsatellites. This study highlights the fact that the way the microsatellites are detected can change biological models fitted on, and finally lead to mistaken interpretations.

References :

International Human Genome Sequencing Consortium, 2001, *Initial sequencing and analysis of the Human Genome*, **Nature** [vol. 409, p 860]

Goldstein D. B. & Schlötterer C., 1999, *Microsatellites Evolution and Applications*, **Oxford University Press**

Jarne P., David P. & Viard D., 1998, *Microsatellites, Transposable Elements and the X Chromosome*, **Mol. Biol. Evol.** [vol. 15(1), p 28]

Rolfsmeier M. L., Dixon M. J. & Lahue R. S., 2000, *Mismatch repair blocks expansions of interrupted trinucleotide repeats in Yeast*, **Mol. Cell** [vol 6, p 1501]

Dettman J. R., Taylor J. W., 2004, *Mutation and Evolution of microsat in Neurospora*, **Genetics** [vol 168, p 1231]

Kruglyak S. et al., 1998, *Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations*, **Proc. Natl. Sci. USA** [vol 95, p 10774]

Dieringer D. & Schlötterer C., 2003, *Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequence of nine species*, **Genome Research** [vol 13, p 2242]

Sainudiin R. et al., 2004, *Microsatellite mutation models : insights from a comparison of humans and chimpanzees*, **Genetics** [vol 168, p 383]

Smit A. F. A., Hubley R. & Green P., *RepeatMasker at <http://repeatmasker.org>*

Delgrange O. & Rivals E., 2004, *STAR : an algorithm to search to tandem approximate repeats*, **Bioinformatics** [vol 20(16), p 2812]

Benson G., 1999, *tandem repeats finder, a program to analyse dna sequences*, **Nucl. Acid. Res.** [vol 27(2), p 573]

Kolpakov R. et al , 2003, *Mreps - efficient and flexible detection of tandem repeats in DNA*, **Nucl. Acid. Res.** [vol 31(13), p 3672]