BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 6 2007, pages 767–768
doi:10.1093/bioinformatics/btm005

Sequence analysis

SatDNA Analyzer: a computing tool for satellite-DNA evolutionary analysis

Rafael Navajas-Pérez[1,*,§,†], Cristina Rubio-Escudero[2,§], José Luis Aznarte[2,§], Manuel Ruiz Rejón[1] and Manuel A. Garrido-Ramos[1]

[1]Departamento de Genética, Facultad de Ciencias, Universidad de Granada, 18071 and [2]Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Granada, Spain

Received on October 30, 2006; revised on January 8, 2007; accepted on January 11, 2007

Advance Access publication January 22, 2007

Associate Editor: Nikolaus Rajewsky

ABSTRACT

Summary: satDNA Analyzer is a program, implemented in C++, for the analysis of the patterns of variation at each nucleotide position considered independently amongst all units of a given satellite-DNA family when comparing it between a pair of species. The program classifies each site accordingly as monomorphic or polymorphic, discriminates shared from non-shared polymorphisms and classifies each non-shared polymorphism according to the model proposed by Strachan *et al*. in six different stages of transition during the spread of a variant repeat unit toward its fixation. Furthermore, this program implements several other utilities for satellite-DNA analysis evolution such as the design of the average consensus sequences, the average base pair contents, the distribution of variant sites, the transition to transversion ratio and different estimates of intra-specific variation and inter-specific variation. Aprioristic hypotheses on factors influencing the molecular drive process and the rates and biases of concerted evolution can be tested with this program. Additionally, satDNA Analyzer generates an output file containing a sequence alignment without shared polymorphisms to be used for further evolutionary analysis by using different phylogenetic softwares.

Availability: satDNA Analyzer is freely available at http://satdna.sourceforge.net/. SatDNA Analyzer has been designed to operate on Windows, Linux and Mac OS X.

Contact: rnavajas@uga.edu

Satellite-DNA families are comprised of tandem non-coding short repeated sequences distributed through the eukaryotic genomes at heterochromatin basically in centromeric and subtelomeric regions as well as chromosome-specific amplified sequences (Ugarkovic and Plohl, 2002). Repetitive DNA families are influenced by several molecular mechanisms of non-reciprocal exchanges (Ohta and Dover, 1984) that can gradually spread a variant sequence throughout a family within a sexual population in a population genetics process called molecular drive (Dover, 1986) which should explain the evolutionary pattern of repetitive sequences known as concerted evolution. Concerted evolution leads to high levels of family homogeneity for species-diagnostic mutations.

Assuming that molecular drive is a time-dependent process, then the expected stages of transition during the spread of a variant repeat unit toward its fixation can be defined according to the model of Strachan *et al*. (1985). This is a method of partitioning of variation by analyzing the patterns of variation at each nucleotide position considered independently amongst all repeats of a repetitive family when comparing a pair of species. This method classifies the sites in terms of six stages (Classes 1–6) in the spread of variant repeats through the family and the species. In brief, the Class 1 site represents complete homogeneity across all repeat units sampled from a pair of species, whereas Classes 2, 3 and 4 represent intermediate stages in which one of the species shows a polymorphism. The frequency of the new nucleotide variant at the site considered is low in Class 2 and intermediate in Class 3, while Class 4 represents sites in which a mutation has replaced the progenitor base in most members of the repetitive family in the other species. Class 5 represents diagnostic sites in which a new variant is fully homogenized and fixed in all the members of one of the species while the other species retains the progenitor nucleotide. A Class 6 site represents an additional step over stage 5 (new variants appear in some of the members of the repetitive family at a site fully divergent between the two species).

If mutation and spreading were operating at similar rates one would expect a consistently high level of within-species variation for variant repeats representing all four bases at any one position each of which might have spread to varying extents. Empirical observations that the overwhelming number of base positions fall into the six classes proposed by Strachan *et al*. (1985) in most species-pair comparisons for different satellite-DNA families indicate that the rate of production of new sequence variants (mutation) is a slower process than their rate of spread while the general paucity of transition stages indicates also that the replacement is relatively fast (Ugarkovic and Plohl, 2002).

*To whom correspondence should be addressed.
†Present address: Plant Genome Mapping Laboratory, Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA.
§The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

However, the rates of homogenization and fixation of sequence variants, i.e. the rates of sequence change, vary for each satellite-DNA family. Levels of sequence identity between repeats would depend on many parameters in each species, such as the rates and biases of transfer between homologous and non-homologous chromosomes, number and distribution of repeat units, physical constraints within the genome, generation time and effective population size as well as selective constraints (Ohta and Dover, 1984).

Patterns of sequence change for satellite-DNA families can be studied by a detailed analysis of the transitional stages in the process of spreading and fixation of sequence variants but, depending on the influence of such many factors influencing the molecular drive process, might it be possible to define some sort of other nucleotide position types when comparing species two to two. Thus, the analysis of shared polymorphisms and of non-shared polymorphisms other than those not included as transitional stages of molecular drive can be highly informative for satellite-DNA evolution analysis (Navajas-Pérez *et al*., 2005; Pons *et al*., 2002; Robles *et al*., 2004). Making aprioristic assumptions one can test different hypotheses on factors influencing the molecular drive process and the rates and biases of concerted evolution.

On these grounds, the partitioning of the variation based in a site by site analysis between all the repeat variants compared between each two species is currently manually performed in an arduous and time-consuming manner. We introduce here satDNA Analyzer, a program for the analysis of the evolutionary patterns of repetitive non-coding sequences. This program has been developed for the analysis of the patterns of variation at each nucleotide position considered independently amongst all units of a repetitive family when comparing a pair of species. The program classifies each site accordingly as monomorphic or polymorphic, discriminate shared from non-shared polymorphisms and classify each non-shared polymorphism within each of the Strachan's stages of transition. Furthermore, this program implements several other utilities for satellite-DNA analysis evolution as the design of the average consensus sequences, the average base pair contents, the distribution of variant sites, the transition to transversion ratio and different estimates of intra-specific variation and inter-specific variation. Available options include the possibility of making these estimations including/excluding gaps and undefined sites as well as, specifically, after including/excluding shared polymorphisms. This latter option is with the aim to avoid the effects of ancestral shared polymorphisms in the estimations of inter-specific divergence, a measure that cannot be accomplished efficiently for satellite-DNA with current available methods (Navajas-Pérez *et al*., 2005).

satDNA Analyzer uses the SEQIO package for reading and writing sequences (Knight, 1996). A first output file is an html document with a site-by-site classification of nucleotide positions and the output of the several options that the program can perform. The second output file is a text document with the alignment without shared polymorphisms in the same format than the input.

ACKNOWLEDGEMENTS

This work was supported by grant CGL2006-00444/BOS awarded to M.A.G-R by the Ministerio de Educación y Ciencia (Spain) and Fondo Europeo de Desarrollo Regional (FEDER, EU). R.N-P was granted by Plan Propio of University of Granada and currently is a Fulbright Postdoctoral Scholar.

Conflict of Interest: none declared.

REFERENCES

Dover,G. (1986) Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. *Trends in Genetics*, **2**, 159–165.

Knight,J. (1996) SEQIO: A C Package for Reading and Writing Sequences. Distributed by the author. Freely available at http://bioweb.pasteur.fr/docs/seqio/seqio.html.

Navajas-Pérez,R. *et al*. (2005) Reduced rates of sequence evolution of Y-linked satellite DNA in *Rumex* (*Polygonaceae*). *J. Mol. Evol*., **60**, 391–399.

Ohta,T. and Dover,G. (1984) The cohesive population genetics of molecular drive. *Genetics*, **108**, 501–521.

Pons,J. *et al*. (2002) Evolutionary dynamics of satellite DNA family PIM357 in species of the genus *Pimelia* (Tenebrionidae, Coleoptera). *Mol. Biol. Evol*., **19**, 1329–1340.

Robles,F. *et al*. (2004) Evolution of ancient satellite DNAs in sturgeon genomes. *Gene*, **338**, 133–142.

Strachan,T. *et al*. (1985) Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*. *EMBO J*., **4**, 1701–1708.

Ugarkovic,D. and Plohl,M. (2002) Variation in satellite DNA profiles—causes and effects. *EMBO J*., **21**, 5955–5959.