BIOINFORMATICS APPLICATIONS NOTE Vol. 15 no. 5 1999
Pages 424–425

JaDis: computing distances between nucleic acid sequences

Isabelle Gonçalves[1], Marc Robinson[2], Guy Perrière[1] and Dominique Mouchiroud[1]

[1]Laboratoire de Biométrie, Génétique et Biologie des Populations, UMR CNRS n° 5558, Université Claude Bernard, Lyon 1, 43, bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex and [2]Laboratoire de Biologie Moléculaire et Cellulaire, École Normale Supérieure de Lyon, 6, allée d'Italie, 69364 Lyon Cedex 07, France

Received on November 26, 1998 revised on January 15, 1999; accepted on January 28, 1999

Abstract

Summary: *JaDis is a Java application for computing evolutionary distances between nucleic acid sequences and G+C base frequencies. It allows specific comparison of coding sequences, of non-coding sequences or of a non-coding sequence with coding sequences.*
Availability: *http://pbil.univ-lyon1.fr/software/jadis.html*
Contact: *{perriere, goncalve}@biomserv.univ-lyon1.fr*

The estimation of distances between sequences is a major concern of comparative molecular studies aimed at understanding genome dynamic evolution, related either to the structure of genetic information or to the way this information is expressed. Such distances are also useful for the comparison of this evolution between different genomes. Several software packages allow the computation of such distances, such as DNADIST from the PHYLIP package (Felsenstein, 1993), PHYLO_WIN (Galtier *et al.*, 1996) or DISTREE (Schafer and Schoniger, 1997). However, these programs are mostly dedicated to phylogeny. This is why we have developed JaDis, which is aimed at molecular evolution studies rather than at phylogeny. This Java application, with a graphical user interface (GUI), allows easy computation of distances between aligned nucleic sequences and G+C frequencies (Figure 1). JaDis can compute distances between several coding sequences, or several non-coding sequences (like introns or flanking regions), but also between a non-coding sequence and several coding sequences. The last option is notably most useful in comparing a pseudogene with functional genes. JaDis results can be saved into text files that can be used directly by spreadsheet or statistical software like EXCEL™ or STATVIEW™ in order to analyse and visualize the results.

JaDis should be able to run on any computer for which a Java 1.1.x runtime is available. It has been tested on MacOS 8.0 and Solaris 2.5 operating systems, respectively, with the MRJ SDK 2.0.1 and JDK 1.1.3 Java Virtual Machines. The

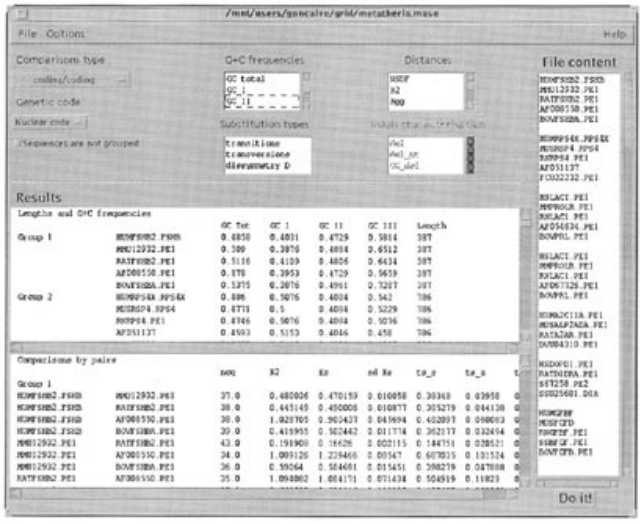

Fig. 1. Main window of JaDis interface. In this example, the results of a comparison between aligned coding sequences are displayed.

Java language was chosen in order to allow portability of the GUI on any kind of platform.

Three formats of input files are recognized: FASTA, CLUSTAL and MASE. The advantage provided by CLUSTAL and MASE formats is that they allow the user to have several different groups of aligned sequences in the same file. JaDis may work on the whole input file or on a subset of the data. Before all computations, except those that are related to insertions and deletions (indels), a global gap removal is carried out.

JaDis allows computation of the distances and general features of the analysed sequences. The array of distances available for computation on coding sequences includes measures of divergence SDF (Silent Difference Frequency) and NSDF (NonSilent Difference Frequency) (Mouchiroud and Gautier, 1990), corrected distances Ka and Ks (Li, 1993), and

their weighted sum Kas (Robinson *et al*., 1997). The frequency of silent differences in position III of codons (SDF) and the percentage of non-silent differences (NSDF) are simple parameters with a clear biological meaning. In Kas, information from all sites, as well as the specificity of codon structure, is taken into account. JaDis is the first published software which computes these three distances (SDF, NSDF and Kas). For non-coding sequences and quartet positions of coding sequences, the two-parameter distance of Kimura (1980), and a distance which takes GC content bias into account (Galtier and Gouy, 1995), are implemented. For non-coding sequences, the NCDF (Non-Coding Difference Frequency) and the Jukes and Cantor (1969) distances are also available, as well as a distance which takes into account substitution rate variations among different sites (Tamura and Nei, 1993). This last distance is also computable on codon positions I and II of coding sequences. In addition, the source of JaDis allows easy programming of your own pairwise distances for coding or non-coding sequences. Finally, all distances can be saved in PHYLIP format for use by phylogenetic software.

The general features of the analysed sequences computed by JaDis include various G+C frequencies, transition/transversion frequencies and an index of substitution dissymmetry (Mouchiroud and Gautier, 1990). This index has been specially designed to handle codon usage changes between coding sequences linked to the variation in G+C III level.

A particular option of JaDis is adapted to the study of pseudogenes, but is suitable for any kind of comparison between a non-coding sequence and one or several coding sequences. The non-coding sequence must be at the first position in each group, and will be compared with the first following coding sequence. The other coding sequences in the group are used to orient changes. On the non-coding sequence, JaDis thus computes the number and the size of deletion and insertion events, as well as the number of deleted and inserted nucleotides per site, the G+C frequency of indels and the substitution bias (i.e. the relative difference between the frequency of substitutions from AT to GC and from GC to AT). In this specific option of JaDis, the computable distances are the NCDF and the two-parameter distance of Kimura.

Acknowledgements

Some functions were adapted from Nicolas Galtier's PHYLO_WIN, to whom we are especially grateful. We also wish to thank Hubert Charles, Manolo Gouy, Sandrine Hughes, Loïc Ponger, Bruno Spataro and Nicolas Tourasse.

References

Felsenstein,J. (1993) *PHYLIP: Phylogeny Inference Package, Version 3.5*. University of Washington, Seattle, WA.

Galtier,N. and Gouy,M. (1995) Inferring phylogenies from DNA sequences of unequal base compositions. *Proc. Natl Acad. Sci. USA*, **92**, 11317–11321.

Galtier,N., Gouy,M. and Gautier,C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Applic. Biosci*., **12**, 543–548.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–132.

Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol*., **16**, 111–120.

Li,W.-H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol*., **36**, 96–99.

Mouchiroud,D. and Gautier,C. (1990) Codon usage changes and sequence dissimilarity between human and rat. *J. Mol. Evol*., **31**, 81–91.

Robinson,M., Catzeflis,F., Briolay,J. and Mouchiroud,D. (1997) Molecular phylogeny of rodents, with special emphasis on murids: evidence from nuclear gene LCAT. *Mol. Phyl. Evol*., **8**, 423–434.

Schafer,J. and Schoniger,M. (1997) DISTREE: a tool for estimating genetic distances between aligned DNA sequences. *Comput. Applic. Biosci*., **13**, 445–451.

Tamura,K. and Nei,M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol*., **10**, 512–526.