BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 9 2012, pages 1270–1271
doi:10.1093/bioinformatics/bts118

Genome analysis

Advance Access publication March 15, 2012

GOHTAM: a website for 'Genomic Origin of Horizontal Transfers, Alignment and Metagenomics'

Sabine Ménigaud†, Ludovic Mallet*†, Géraldine Picord, Cécile Churlaud, Alexandre Borrel and Patrick Deschavanne*

Molécules Thérapeutiques in silico, Institut National de la Santé et de la Recherche Médicale (INSERM) UMR-S 973, Université Paris Diderot, Sorbonne Paris Cité, 35 rue Héléne Brion, 75013, Paris, France

Associate Editor: Martin Bishop

ABSTRACT

Motivation: This website allows the detection of horizontal transfers based on a combination of parametric methods and proposes an origin by researching neighbors in a bank of genomic signatures. This bank is also used to research an origin to DNA fragments from metagenomics studies.

Results: Different services are provided like the possibility of inferring a phylogenetic tree with sequence signatures or comparing two genomes and displaying the rearrangements that happened since their separation.

Availability and implementation: http://gohtam.rpbs.univ-paris-diderot.fr/

Contact: patrick.deschavanne@univ-paris-diderot.fr; ludovic.mallet@jouy.inra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online http://gohtam.rpbs.univ-paris-diderot.fr:8080/Data/bin/GOHTAM_bin.tgz

Received on August 16, 2011; revised on March 2, 2012; accepted on March 6, 2012

1 INTRODUCTION

Horizontal transfers (HTs) are a major force of evolution (Keeling and Palmer, 2008; Ochman *et al.*, 2000) and this website proposes methods for their detection. The genomic signature was demonstrated to be species-specific (Deschavanne *et al.*, 1999; Sandberg *et al.*, 2001) and allows HT detection in terms of tetranucleotide frequencies (Dufraigne *et al.*, 2005). Parametric methods were designed to work only with the information contained in genomic sequences. They rely either on the whole set of genes or on local variations of genomic signature (Dufraigne *et al.*, 2005; Mallet *et al.*, 2010). Recently, a benchmark has determined the most efficient parametric methods in different conditions and has proposed to use a combination of methods to analyze HTs in genomes (Becq *et al.*, 2010). This site provides user-friendly access to such methods as well as some unique features including signature-based phylogeny and potential origin of a set of metagenomics sequences.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

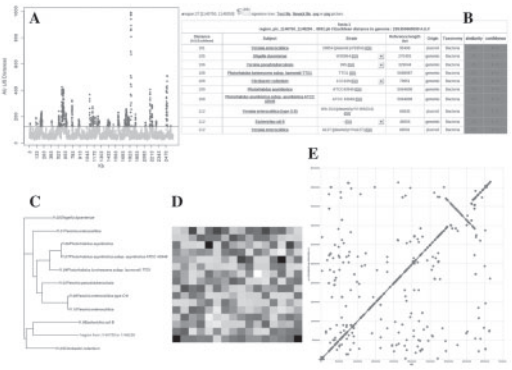

Fig. 1. Some partial screens of the website. (A) Window-based HT detection; (B) table of neighbors; (C) signature-based phylogenetic tree; (D) species signature; and (E) genome alignment.

2 GOHTAM SERVICES

2.1 HT detection

The two methods proposed can be used alone or in combination. The first is a window-based signature method as described in Dufraigne *et al.* (2005), except that the distance used is the Jensen–Shannon divergence, a symmetric version of the Kullback–Leibler divergence (Azad and Lawrence, 2007; Becq *et al.*, 2010). Either sensitivity or specificity can be increased by adjustable classification process (Azad and Lawrence, 2007). A gene-based method is also proposed with the same distance (Becq *et al.*, 2010). Up to now, these methods were never proposed for online genome analysis (Fig. 1A).

2.2 Bank of genomic signatures

A key feature of GOHTAM is the biggest bank of genomic signatures to date. Instead of using only complete genomes (van Passel *et al.*, 2005; Teeling *et al.*, 2004), this bank is based on the whole set of sequences of Genbank (release 188, only sequences <1 kb were discarded) and contains $\sim$248 000 tetranucleotide species signatures. The bank is updated at each major release.

2.3 Origin of transferred regions

Each detected region signature is compared with the signatures of the bank and the 10 closest neighbors are displayed with a confidence rating depending on the length of both query and reference sequences and the distance between the two signatures (Fig. 1B).

2.4 Metagenomics

In the case of a metagenomics study, a sequence or a set of sequences (multi-Fasta) is loaded; the signatures of these sequences are compared as above to propose a species of origin.

2.5 Oligonucleotide content

The whole set of tetranucleotides of a sequence represents the signature of a sequence (Deschavanne *et al.*, 1999). This signature of the 256 possible tetranucleotides is under the form of a 16×16 frequency matrix and can be displayed as a signature image (Fig. 1D).

2.6 Phylogenetic tree of sequence signatures

It was shown that the species specificity of genomic signatures could be used to infer phylogenetic trees (Chapus *et al.*, 2005). Loading a multi-Fasta file of sequences leads to build a neighbor-joining phylogenetic tree (Fig. 1C; Felsenstein, 2005).

2.7 Genome alignment

The website uses maximum unique matches (MUMs) to align genomes. All rearrangements superiors to 1 kb between two genomes are graphically displayed with the possibility to choose a region or modify the length of MUMs (Fig. 1E; Delcher *et al.* 1999).

3 IMPLEMENTATION

Except for use of programs like the Phylip package (http://evolution.gs.washington.edu/phylip.html) or Mummer (http://mummer.sourceforge.net/), the original programs are written in Python, Perl or R and available at: http://gohtam.rpbs.univ-paris-diderot.fr:8080/Data/bin/GOHTAM_bin.tgz

An online help is available. Some analyses require time; HT detection lasts ∼6 min and the research for neighbors ∼2 min depending on the server load and the sequence length.

This site provides some unique features in terms of HT detection, origin of HT regions, metagenomics studies as well as for phylogenetic analyses of homologous or non-homologous sequences due to its extended reference database and improves the analyses proposed by other sites of genome analysis (van Passel *et al.*, 2005; Teeling *et al.*, 2004).

Funding: This work was supported by a grant from ANR MIE/TB-Hits 2010.

Conflict of Interest: none declared.

REFERENCES

Azad,R.K. and LawrenceJ.G. (2007) Detecting laterally transferred genes: use of entropic clustering methods and genome position. *Nucleic Acids Res*., **35**, 4629–4639.

Becq,J. *et al.* (2010) 'A benchmark of parametric methods for horizontal transfers detection'. *PLoS ONE,* **5**, e9989.

Chapus,C. *et al.* (2005) 'Exploration of phylogenetic data using a global sequence analysis method' *BMC Evol. Biol.*, **5**, 63–83.

Delcher,A.L. *et al.* (1999) 'Alignment of whole genomes' *Nucleic Acids Res*, **27**, 2369–2376.

Deschavanne,P.J. *et al.* (1999). 'Genomic signature: characterization and classification of species assessed by Chaos Game Representation of sequences'. *Mol. Biol. Evol*., **16**, 1391–1399.

Dufraigne,C. *et al.* (2005) 'Detection and characterization of horizontal transfers in prokaryotes using genomic signature'. *Nucleic Acids Res*., **33**, e6.

Felsenstein,J. (2005) *PHYLIP (Phylogeny Inference Package) Version 3.6. Distributed by the Author*. Department of Genome Sciences, University of Washington, Seattle.

Keeling,P.J. and Palmer, J.D. (2008) 'Horizontal gene transfer in eukaryotic evolution'. *Nat. Rev. Genet*., **9**, 605–618.

Mallet,L.V. *et al*. (2010) 'Whole genome evaluation of horizontal transfers in the pathogenic fungus Aspergillus fumigatus'. *BMC Genomics*, **11**, 171.

Ochman,H. *et al*. (2000) 'Lateral gene transfer and the nature of bacterial innovation'. *Nature*, **405**, 299–304.

Sandberg,R. *et al.* (2001). 'Capturing whole-genome characteristics in short sequences using a naïve bayesian classifier' *Genome Res*., **11**, 1404–1409.

Teeling,H. *et al.* (2004) 'TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences'. *BMC Bioinformatics*, **5**, 163.

van Passel,M.W. *et al.* (2005) 'Deltarho-web, an online tool to assess composition similarity of individual nucleic acid sequences'. *Bioinformatics*, **21**, 3053–3055.