Genome analysis

AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps

Thomas Derrien, Catherine André, Francis Galibert and Christophe Hitte*
CNRS UMR 6061 Génétique et Développement, Université de Rennes 1, IFR 140, 2 Av du Pr. Léon Bernard, CS 34317, 35043, France

Received on September 20, 2006; revised and accepted on November 29, 2006
Advance Access publication December 4, 2006
Associate Editor: Alfonso Valencia

ABSTRACT
AutoGRAPH is an interactive web server for automatic multi-species comparative genomics analyses based on personal datasets or pre-inserted public datasets. This program automatically identifies conserved segments (CS) and breakpoint regions, assesses the conservation of marker/gene order between organisms, constructs synteny maps for two to three species and generates high-quality, interactive displays facilitating the identification of chromosomal rearrangements. AutoGRAPH can also be used for the integration and comparison of several types of genomic resources (meiotic maps, radiation hybrid maps and genome sequences) for a single species, making AutoGRAPH a versatile tool for comparative genomics analyses.

Availability: http://genoweb.univ-rennes1.fr/tom_dog/AutoGRAPH/
Contact: hitte@univ-rennes1.fr
Supplementary information: A description of the algorithm and additional information are available at http://genoweb.univ-rennes1.fr/tom_dog/AutoGRAPH/Tutorial.php

1 INTRODUCTION
Many large-scale mapping and sequencing projects have been completed in the last 10 years, making it possible to compare the genomes of many species, to study evolutionary changes (Murphy et al., 2005), and to improve genome annotation (Chatterji and Pachter, 2006). Comparative genomics is based on the identification of unique, unambiguous orthologous sequences in different species. These sequences, known as comparative anchors (O’Brien et al., 1993), are used to determine the contingency of anchors between genomes from different species and thus to investigate the correspondence of genomic segments and to define their limits.

Synteny maps can be used to identify conserved segments (CS), corresponding to chromosomal segments containing the same list of markers in all the species studied, and CS ordered (CSO), in which not only are the same markers present in all species, but they are also in the same order (Fig. 1). The genomic regions delimiting CS and/or CSO, the breakpoints, can also be identified by constructing synteny maps. Several tools have been developed for identifying CS and CSO (Pan et al., 2005; Pavesi et al., 2004; Halling-Brown et al., 2004; Clamp et al., 2003; Tesler, 2002). These tools allow accurate comparative genomics analyses, but are often dedicated to a specific genome, are not compatible with the use of large sets of personal data, and are not always available online. We have therefore developed AutoGRAPH, a versatile, online multi-species server for the automatic identification of CS, CSO, breakpoints, marker/gene order conservation, the integration of multiple resources for a given species (Hitte et al., 2005) (Supplementary Figure 2), and the generation of an interactive graphical display of comparative maps. AutoGRAPH frees the user from the tedious task of plotting comparative maps and determining marker order correspondence manually for personal data, enabling the user to focus on interpreting synteny or integrated maps.

2 PRINCIPLES
Using the AutoGRAPH server for constructing synteny maps between species

Public datasets: information for the protein-coding genes (Ensembl v39) of six reference mammalian genomes (human, chimpanzee, cow, dog, mouse and rat) have been inserted into the program database. Orthologous gene pairs with one-to-one relationships (Ensembl v39) are used to construct pairwise or three-way synteny maps. Synteny maps can be built for selected chromosomes or for all the chromosomes of the reference genome. CS, CSO and breakpoints are automatically identified with the fast adjacency function. This function assigns relative integers to the coordinates provided and determines the sequence of integers in the same order for common anchors of the reference and tested genomes (a detailed description is provided in the Supplementary material). Various options are available and can be combined interactively to specify the way in which the synteny map should be constructed. For example, it is possible to specify the nature of the orthology between markers (one-to-one/one-to-zero) and the number of anchors defining CS and CSO. An adjacency penalty value, corresponding to the number of marker that account for an interruption of colinearity can be set by users to specify the marker order conservation criteria.

The program output consists of

- A figure showing gene order relationships by means of connecting lines within each CS and CSO, across two to three species, and identifying the evolutionary breakpoint regions delimiting CS and/or CSO (Fig. 1A). Output figure can be exported in several formats.
- Two modes, making it possible to construct comparative maps with or without one-to-zero relationship anchors, to switch the orientation of each map and to scale images by a factor of 0.5 to 2 (Fig. 1B). When mode 1:0 is selected, genes in the reference genome with no ortholog in the tested genome are placed on the synteny map. Their putative interval location in

*To whom correspondence should be addressed.
the tested genome is inferred from the gene-order conservation rule and is automatically output as a tabulated text file.

- A table listing all CS, CSO and breakpoints, with their size, number of genes, chromosome coordinates and gene density (Fig. 1C).

Personal datasets: two formats of input datasets can be uploaded. GFF format and tabular plain-text columns comprising unique marker IDs, chromosome number and genomic coordinates for each dataset to be studied. Input datasets may contain different sets of markers, with missing or duplicated markers or genes. Personal datasets can be entered or uploaded via web form and example inputs are provided by the web interface.

For comparative studies between species, synteny maps can be built for entire chromosomes, specific parts of chromosome or the whole genome, as specified by the user. Datasets for a single species can also be used. Any type of coordinate-base pairs (bp) kilobases (kb) or megabases (Mb)-mapping data units, centimorgans (cM) or centirays (cR) can be used to indicate location in the genome. For example, (Supplementary Figure 2), the server can integrate high-density genetic/radiation hybrid/genome sequence maps for a single species, facilitating map construction and improving sequence assembly. Various options can be combined to specify comparative analyses. The output displays are the same as for public datasets.

3 SYSTEMS

AutoGRAPH runs on an Apache web server and uses a MySQL relational database. MySQL specific queries and Perl functions are used to apply options and the adjacency function. The web interface was developed in PHP language and the graphical display uses a GD graphic library written in C (http://www.boutell.com/gd/).

ACKNOWLEDGEMENTS

The authors would like to thank Denis Larkin and Simon de Givry for testing AutoGRAPH and providing useful suggestions. The authors also thank the GenOuest Bioinformatics Platform for hosting the web server, the French Centre National de la Recherche Scientifique (CNRS) for supporting this work and the Conseil Regional de Bretagne for supporting T.D. with a fellowship.

Conflict of Interest: none declared.

REFERENCES


