Available online at www.sciencedirect.com
ScienceDirect
Biochimie 90 (2008) 609–614
BIOCHIMIE
www.elsevier.com/locate/biochi

Research paper

Automatic identification of large collections of protein-coding or rRNA sequences

Anne-Muriel Arigon*, Guy Perrière, Manolo Gouy

Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon; université Lyon 1, CNRS, UMR 5558, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France

Received 3 July 2007; accepted 24 August 2007
Available online 2 September 2007

Abstract

The number of available genomic sequences is growing very fast, due to the development of massive sequencing techniques. Sequence identification is needed and contributes to the assessment of gene and species evolutionary relationships. Automated bioinformatics tools are thus necessary to carry out these identification operations in an accurate and fast way. We developed HoSeqI (Homologous Sequence Identification), a software environment allowing this kind of automated sequence identification using homologous gene family databases. HoSeqI is accessible through a Web interface (http://pbil.univ-lyon1.fr/software/HoSeqI/) allowing to identify one or several sequences and to visualize resulting alignments and phylogenetic trees. We also implemented another application, MultiHoSeqI, to quickly add a large set of sequences to a family database in order to identify them, to update the database, or to help automatic genome annotation. Lately, we developed an application, ChiSeqI (Chimeric Sequence Identification), to automate the processes of identification of bacterial 16S ribosomal RNA sequences and of detection of chimeric sequences.

Keywords: Automatic identification; Similarity; Alignment; Phylogeny; Chimera

1. Introduction

Identification is used in many fields, such as microbiology, medicine and environment. Sequence identification consists in the attribution of an unknown taxonomic unit to a taxonomic group of a pre-established classification. Thus, to identify a new taxon or a new sequence, it is necessary to find its nearest known taxon. In the medical field, methods of identification are used to detect and recognize micro-organisms implied in pathologies, which thus helps choosing the most suitable treatment. Identification can also be used in the agro-alimentary field as tools for food traceability. In other contexts such as identification of species or taxons from environmental organism molecular markers, the confrontation of a new sequence with a database, or sequence database update, the assignment of a new sequence to a collection is necessary. The number of available biological sequences increasing considerably with the development of massive sequencing techniques, it is necessary to rapidly classify these sequences into existing databases.

According to analyzed data, the approach used for identification differs and several tools exist. Identification tools often vary with the type of sequences and thus with the sequence databases for which they were developed. Several tools exist to make sequence identification and most of them are domain specific or data specific. For instance, some applications allow bacterial identification as BIBI (Bioinformatic Bacterial Identification) [1], PhyID/CD [2] or MicroSeq (Microbial identification System); others are specialized for the medical domain as RIDOM (Ribosomal Differentiation Of Medical Organisms) [3] or for the identification of Ribosomal RNA sequences as the RDP classifier (Ribosomal Database Project) [4], or TaxI [5] based on DNA barcodes.

* Corresponding author. Tel.: +33 4 26 23 44 75; fax: +33 4 72 43 13 88.
E-mail address: arigon@biomserv.univ-lyon1.fr (A.-M. Arigon).

0300-9084/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.biochi.2007.08.006

We are interested in the homologous gene family databases HOVERGEN and HOGENOM [6] developed in our group. In these databases, homologous sequences are clustered into families, i.e., sequences of the same family share a common ancestor. Sequence alignments and phylogenetic trees for each family are also stored in these databases. Thus, these databases can be used for different purposes, among which phylogenetic analyses, and they allow the study of sequence evolutionary relationships. In order to build these family databases, several complex automated procedures are needed (similarity search, gene clustering, multiple alignment and tree computations). With the very fast growth of biological data, gene family database updates are time-consuming and tedious. Moreover, the addition of a single sequence to a given family from these databases can have many repercussions on the topology of the associated phylogenetic tree; these changes may be located near the introduced sequence, but they may also be located in deep nodes. In such case, the phylogenetic information brought by the whole family should be taken into account. Also, as HOVERGEN and HOGENOM contain large families, with several thousand sequences, powerful algorithms are required in order to manage large amount of sequences. Available identification tools, such as those presented previously, are developed to treat specific data and cannot be used effectively with large family databases. Thus, it is necessary to develop methods and bioinformatics tools (i) to carry out identification processes in a precise and rapid way, and (ii) to quickly add sequences to these databases without integrally updating them.

2. Two applications adapted to homologous gene family databases: HoSeqI and MultiHoSeqI

We have developed an application, HoSeqI (Homologous Sequence Identification), and another derived from the first, MultiHoSeqI. HoSeqI [7] is a Web application (http://pbil.univ-lyon1.fr/software/HoSeqI/) that allows to automatically identify sequences in large gene family databases. The identification process of an unknown sequence into these databases consists in (i) finding the homologous gene family to which this sequence belongs, using similarity search, (ii) aligning the analyzed sequence with the whole family and (iii) reconstructing the phylogenetic tree of the family including the new sequence. HoSeqI proposes an interface allowing the user to submit his/her sequence and to choose the database. When the family of the studied sequence is determined, the user can obtain information on the selected family and choose between several multiple alignment and tree building programs. Then it is possible to visualize the resulting alignment and phylogenetic tree (Fig. 1).

The developments carried out for HoSeqI were also used to create another application—MultiHoSeqI—allowing to quickly add several thousand sequences to a homologous gene family database. MultiHoSeqI corresponds to a generalization of HoSeqI to *n* sequences. For each one of these *n* sequences, the application identifies the family to which it belongs, then for each family containing at least an added sequence, alignments are computed and phylogenetic trees are built including the new sequences. MultiHoSeqI is publicly usable on the HoSeqI Web interface with a restriction on the number of sequences: the user can choose between the identification of only one sequence or of several sequences. If the ‘‘several sequences’’ option is selected, the user submits his/her sequences, chooses the database, the multiple alignment and tree building programs. The process is then started on the server and the results are sent to the user by e-mail.

3. Use of MultiHoSeqI with sequences of bacterial genus *Frankia*

MultiHoSeqI has been used to add genes from several collections of protein sequences to the databases developed by the PBIL (Pôle BioInformatique Lyonnais): putative protein sequences from metagenomes and from completely sequenced bacterial genomes. In collaboration with Philippe Normand (Laboratory of Soil Microbial Ecology, University of Lyon), Vincent Daubin and Simon Penel (Laboratory of Biometry and Evolutionary Biology, University of Lyon), this application was used to add predicted protein sequences from two completely sequenced genomes of the bacterial genus *Frankia* into the HOGENOM database in order to study the evolution of these genomes and to detect possible horizontal gene transfers to *Frankia*.

The bacterial genus *Frankia* belongs to the class Actinobacteria. Among Actinobacteria, there are in particular genus *Mycobacterium* (agents of tuberculosis and leprosy) and genus *Streptomyces* (soil bacteria at the origin of many antibiotics). Twelve species of *Frankia* are recognized today. These bacteria fix nitrogen and convert atmospheric N_2 gas into ammonia, this in symbiosis with a large spectrum of dicot plants, called actinorhizal. These plants, with their symbiotic bacteria, are collectively responsible for approximately 15% of the biologically fixed nitrogen in the world. This association presents a major ecological interest and many actinorhizal plants are used by the pharmaceutical industry because of their large production of phenolic molecules of various activities (e.g., antimicrobial, antioxidant, antiviral, anti-inflammatory drugs, antispasmodic, antitumor). So it is interesting to search information about genes involved in symbiosis and to study how symbiosis evolved.

Three strains of *Frankia* were available when we made this study: HFPCcI3 (CcI3), EAN1pec (EAN) and ACN14a (ACN). The two first strains were sequenced in the DOE Joint Genome Institute in collaboration with D. Benson (University of Connecticut) and L. Tisa (University of New Hampshire). The third strain was sequenced in the Génoscope in collaboration with P. Normand. Genomes of these strains are circular and their size vary between 5.38 millions of base pairs (Mb) for CcI3, 7.50 Mb for ACN and 9.08 Mb for EAN.

MultiHoSeqI was used to add sequences of genomes of the strains CcI3 (4557 sequences) and EAN (7976 sequences) to a local version of the database HOGENOM (based on the version 2 of October 2004) containing sequences of the strain ACN of *Frankia*. Among the 12,533 sequences that were analyzed,