BIOINFORMATICS

Vol. 14 no. 3 1998
Pages 285-289

Towards detection of orthologues in sequence databases

Yan P. Yuan, Oliver Eulenstein[1], Martin Vingron[2] and Peer Bork

Biocomputing, European Molecular Biology Laboratory (EMBL), Meyerhofstr. 1, 69012 Heidelberg & Max-Delbrueck-Center, Berlin, [1]Universität Bonn & GMD, St Augustin, Bonn and [2]Deutsche Krebsforschungszentrum (DKFZ), 69120 Heidelberg, Germany

Received on October 17, 1997; revised and accepted on January 20, 1998

Abstract

Motivation: *Numerous homologous sequences from diverse species can be retrieved from databases using programs such as BLAST. However, due to multigene families, evolutionary relationship often cannot be easily determined and proper functional assignment becomes difficult. Thus, discrimination between orthologues and paralogues within BLAST output lists of homologous sequences becomes more and more important.*

Result: *We therefore developed a method that attempts to construct a reconciled tree from a gene tree of selected sequences and its corresponding phylogenetic tree of the species involved (species tree). An interface on the Web is developed to enable users to analyse the BLAST result. BLAST outputs are parsed and, for the selected sequences, multiple alignments are constructed either globally or for local regions. Bootstrapped trees are returned and compared with the expected species tree. In cases of discrepancies, gene duplications are assumed and a reconciled tree is computed. The reconciled tree shows probable orthologues and paralogues as predicted.*

Contact: *yuan@embl-heidelberg.de*

Introduction

The rapidly progressing genome projects provide a huge amount of sequence data which need to be annotated as accurately as possible. One focus is the assignment of biological function to a gene or a gene product, respectively, which is mostly accomplished by computer-aided sequence analysis (Bork *et al*., 1994). Today, this is usually done by searching for homologues in sequence databases and subsequent transfer of functional annotation from the best database match. The heuristic search algorithm BLAST (Altschul *et al*., 1990) is used extensively for this purpose. As >70% of all sequences in current databases have detectable homologues and, for most of them, at least some functional features have been annotated, a more reliable and efficient functional assignment is required. However, current database search techniques are not able to discriminate whether the best hit is an orthologue, i.e. the functionally corresponding counterpart of the query sequence in another species, or only a paralogue, i.e. a homologous member of a multigene family (Fitch, 1970) that shares, in the best case, only some functional features with the query.

To overcome this obvious limitation in functional predictions, we have developed a procedure that tries to decide for a query sequence of a given species whether the function of the database hits can be assigned, i.e. whether those hits are orthologues of the query sequence. This analysis is based on the comparison of a gene tree computed for a set of homologous genes with the phylogenetic tree for the species from which those genes came. Discrepancies between the two trees may be due, among others, to gene duplications. Computation of the so-called reconciled tree displays the duplication events and thus allows the assignment of orthologous and paralogous sequences.

Methods and implementation

We established a pipeline of programs to apply the above approach in the context of a database search. At the time of development, the WU-BLAST2 algorithm for database search (Altschul and Gish, 1996; W.Gish, unpublished; server: http://blast.wusl/.edu/ was the most advanced program, but newer versions can be integrated. We apply this BLAST2 program for the database search as it provides gapped alignments and more sensitivity than the old BLAST version. In addition, several post-processing procedures are implemented to support decision finding:

1.. Sequences above a certain hit threshold are selected for further processing, whereby a graphical digest of the WU-BLAST2 output allows easy manual refinement of the selection.
2. The sequences are multiply aligned with either global or local alignment being performed.
3. A neighbour-joining tree is computed for the selected sequences. Bootstrapping is used to check the robustness of the grouping.
4. A phylogenetic tree for the species under study is extracted from the taxonomy key provided by the NCBI.
5. The gene tree and the species tree are compared, and in the case of discrepancy reconciled trees are constructed (Page,

1994) that suggest possible duplication events. The graphical output of the trees is produced by drawgram/drawtree of the PHYLIP package (Felsenstein, 1989; see also: http://evolution.genetics.washington.edu/phylip.html).

In summary, this semi-automatic procedure allows one to decide which of the sequences identified in the database search are orthologues of the query sequence. The above procedure is integrated as an additional option into a WU-BLAST2 (http://www.bork.embl-heidelberg.de/Blast2/) server.

Search for homologous sequences and multiple alignment

The multiprocessor WU-BLAST2 program currently appears to be the best compromise between speed and sensitivity of database searches. It provides gapped alignments and good statistics that can be used for automatic large-scale analysis. For these reasons, we use it to scan for homologues of a given query sequence in sequence databases. We have developed a WWW server for it in which we provide by default several useful post-processing steps such as a hyperlinked graphical display of the matched regions with respect to both query and database sequence. It facilitates the detection of 'twilight zone' matches (i.e. matches with a *P* value >0.001) and the selection of database hits for further analysis. HTML-based output (Figure 1) allows the retrieval of functional information from various hyperlinked databases utilizing SRS (Sequence Retrieval System; Etzold *et al.*, 1996).

As long as the selected database sequences are reasonably related to each other, a subsequent global alignment performed by CLUSTALW (Higgins *et al.*, 1996) should reveal most of the required information. However, there are many cases where only domains are homologous or where the sequences are so divergent that automatic multiple alignment procedures have difficulties. Therefore, we implemented a second option that only considers local alignment of motifs that are common to all selected sequences. Although the signal may become weaker due to the loss of some segments, the remaining regions are the most conserved ones and misalignment is avoided. The output is the multiple alignment and the set of evolutionary distances among aligned sequences.

Construction of the gene tree

The phylogenetic gene tree is constructed by using routines of the widely distributed computer program CLUSTALW. The CLUSTALW program uses the neighbour-joining method (Saito and Nei, 1987) to construct the phylogenetic tree, but does not provide a root for the tree. As our procedure requires a rooted tree, two rules are used: (i) choose the longest branch as the tree root; (ii) use a sequence as an outgroup from an organism that has to be outside of the rest of the genes (Figure 2a). A bootstrapped tree of the genes is also computed. However, for the time being, bootstrapping results are given only as an interpretation aid and are not automatically inserted into the tree graphical presentation.

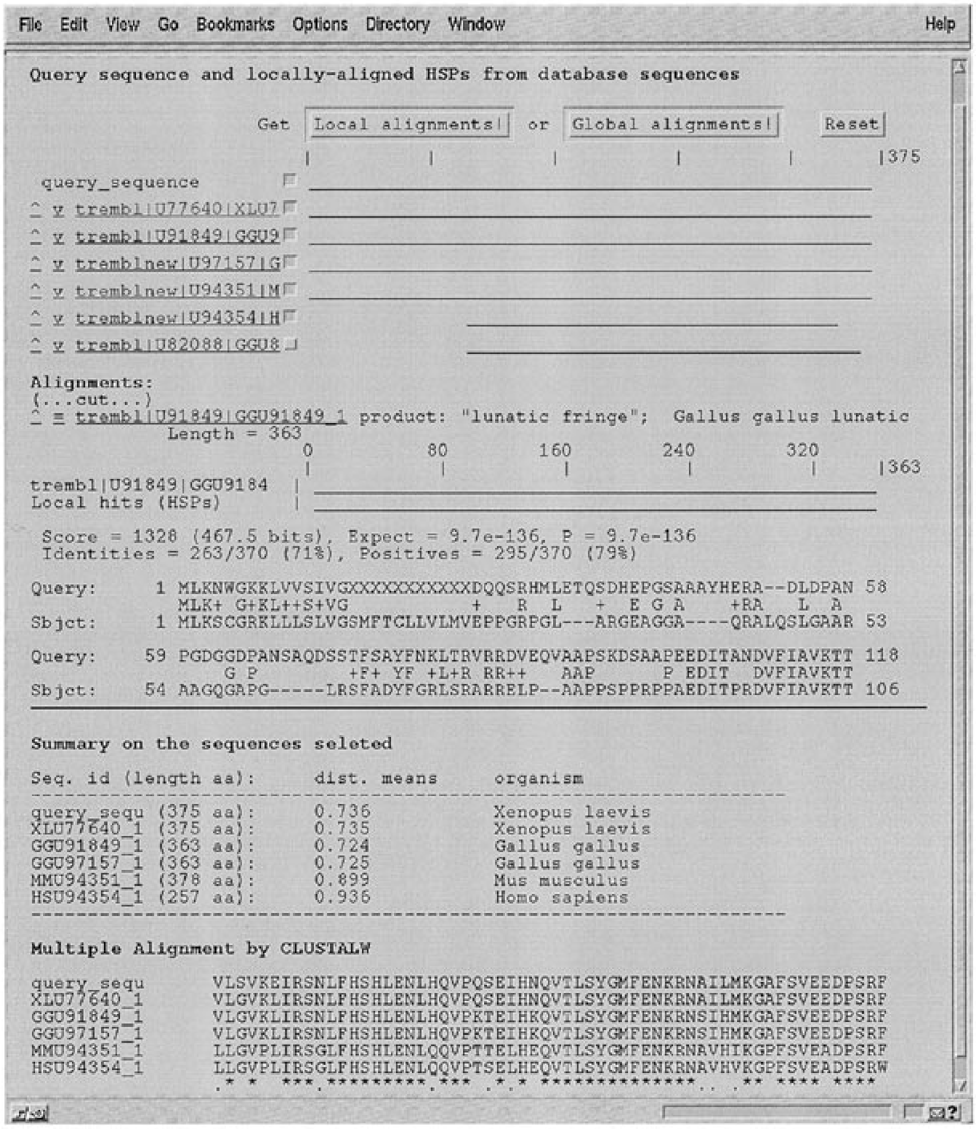
Construction of the species tree

A phylogenetic tree that represents our current knowledge on the evolution of species we call a species tree. The taxonomy of species reflects this knowledge, although there are numerous branches that are not bifurcating. Based on this taxonomic tree and using the newest information from molecular evolutionary analysis and other sources, NCBI has been developing a taxonomic species tree (D.Leipe, V.Soussov, http://www.ncbi.nlm.gov/Taxonomy). This taxonomic species tree has recently been adapted by EMBL's databases. We use this information to build a species tree for the species from which genes were selected in the prior step. For fast access to the species tree, a species database is created where all the internal nodes starting from an organism all the way to the root are easily retrievable. This allows the relevant part of the phylogenetic tree of the species involved in the analysis to be built quickly (Figure 2b). In our approach, the taxonomic species tree is built up first, then the corresponding genes are sorted accordingly to construct the species tree.

Construction of reconciled phylogenetic trees

The phylogenetic gene tree and the corresponding species tree either agree or disagree with each other. In the case of agreement, the sequences involved may be considered to be orthologues and a functional assignment of the query based on the database hits would be justified. On the other hand, a disagreement might be caused by the uncertainty in relationships among the sequences and thus by the limits of the methods used. To check this, we provide the bootstrapping option. This minimizes the risk of making false deductions by mistaking small reconstruction errors for real disagreement between trees. When real disagreement is suspected, the conflict between species and gene tree is resolved by postulating gene duplication events. For this purpose, a computer program has been developed that integrates both gene tree and species tree into the so-called reconciled tree (Page, 1994). In this reconciled tree, distinguishing between orthologues and paralogues becomes straightforward (Figure 2c). All the trees (species tree, gene tree and finally reconciled tree) are given in PHYLIP's newick format which can be displayed by the graphical program treetool. Additionally, each of them is displayed in graphics produced by drawing programs of PHYLIP.

Another example for our approach is given in Figure 3. Kawamura and Yokoyama (1995) analysed rhodopsin-like opsin genes in lizards and predicted that the opsin gene RH2Ac (OPSB_ANOCA) of the species *Anolis carolinensis* and the gecko blue opsin gene OPSB_GECGE of the species *Gecko gecko* should be derived as paralogues from duplicate



```
Query sequence and locally-aligned HSPs from database sequences

                          |          |          |          |         |       |375
 query_sequence
^ v trembl|U77640|XLU7
^ v trembl|U91849|GGU9
^ v tremblnew|U97157|G
^ v tremblnew|U94351|M
^ v tremblnew|U94354|H
^ v trembl|U82088|GGU8

Alignments:
(...cut...)
^ = trembl|U91849|GGU91849_1 product: "lunatic fringe";  Gallus gallus lunatic
            Length = 363
                          0          80        160        240        320
                          |          |          |          |          |      |363
trembl|U91849|GGU9184   |
Local hits (HSPs)       |

 Score = 1328 (467.5 bits), Expect = 9.7e-136, P = 9.7e-136
 Identities = 263/370 (71%), Positives = 295/370 (79%)

Query:      1 MLKNWGKKLVVSIVGXXXXXXXXXXXDQQSRHMLETQSDHEPGSAAAYHERA--DLDPAN 58
              MLK+ G+KL++S+VG           +   R  L   +  E G A    +RA   L  A
Sbjct:      1 MLKSCGRKLLLSLVGSMFTCLLVLMVEPPGRPGL---ARGEAGGA----QRALQSLGAAR 53

Query:     59 PGDGGDPANSAQDSSTFSAYFNKLTRVRRDVEQVAAPSKDSAAPEEDITANDVFIAVKTT 118
                  G P        +F+ YF +L+R RR++   AAP      P EDIT  DVFIAVKTT
Sbjct:     54 AAGQGAPG-----LRSFADYFGRLSRARRELP--AAPPSPPRPPAEDITPRDVFIAVKTT 106

Summary on the sequences seleted

Seq. id (length aa):     dist. means     organism
------------------------------------------------------------------
query_sequ (375 aa):     0.736           Xenopus laevis
XLU77640_1 (375 aa):     0.735           Xenopus laevis
GGU91849_1 (363 aa):     0.724           Gallus gallus
GGU97157_1 (363 aa):     0.725           Gallus gallus
MMU94351_1 (378 aa):     0.899           Mus musculus
HSU94354_1 (257 aa):     0.936           Homo sapiens
------------------------------------------------------------------

Multiple Alignment by CLUSTALW

query_sequ      VLSVKEIRSNLFHSHLENLHQVPQSEIHNQVTLSYGMFENKRNAILMKGAFSVEEDPSRF
XLU77640_1      VLGVKLIRSNLFHSHLENLHQVPQSEIHNQVTLSYGMFENKRNAILMKGAFSVEEDPSRF
GGU91849_1      VLGVKLIRSNLFHSHLENLHQVPKTEIHKQVTLSYGMFENKRNSIHMKGAFSVEEDPSRF
GGU97157_1      VLGVKLIRSNLFHSHLENLHQVPKTEIHKQVTLSYGMFENKRNSIHMKGAFSVEEDPSRF
MMU94351_1      LLGVPLIRSGLFHSHLENLQQVPTTELHEQVTLSYGMFENKRNAVHIKGPFSVEADPSRF
HSU94354_1      LLGVPLIRSGLFHSHLENLQQVPTSELHEQVTLSYGMFENKRNAVHVKGPFSVEADPSRW
                .* *  ***.*********.*** .*.* **************.. .** **** ****
```

Fig. 1. An HTML-linked output of the BLAST2 server as currently implemented at the EMBL in Heidelberg (http://www.bork.embl-heidelberg.de/Blast2e/) and displayed using Netscape. Some graphical features have been adapted from Worley *et al.* (1995). The example shows members of the diverse fringe family of glycosyltransferases (Yuan *et al.*, 1997). The query is a frog sequence of the fringe family with a nearly 100% hit to frog 'lunatic fringe' (see alignment part). Apart from the query, five other sequences have been selected (upper part). In the example, a global alignment has been chosen and the distance matrix as well as multiple alignment are displayed (lower part). Finally, gene tree, species tree and reconciled tree are computed and given in a form that can be read by tree-drawing tools.

ancestral genes. Our approach using their data could fully confirm this result.

Discussion

The functional assignment of a query sequence is carried out extensively today by using the database search and transfer of the function of the best hit sequence out of the database. It turns out that the BLAST algorithm represents the best compromise between search efficiency and sensitivity. There exist some post-processing procedures either available on diverse WWW sites (like BEAUTY's server; Worley *et al.*, 1995) or available for stand-alone runs on different computer platforms (like the Visual BLAST of Durand *et al.*, 1997). The greatest weakness in all these approaches is that users cannot or only inconveniently retrieve all the hit sequences with their full length and all their biological information such as their taxonomic classifications. This makes the further

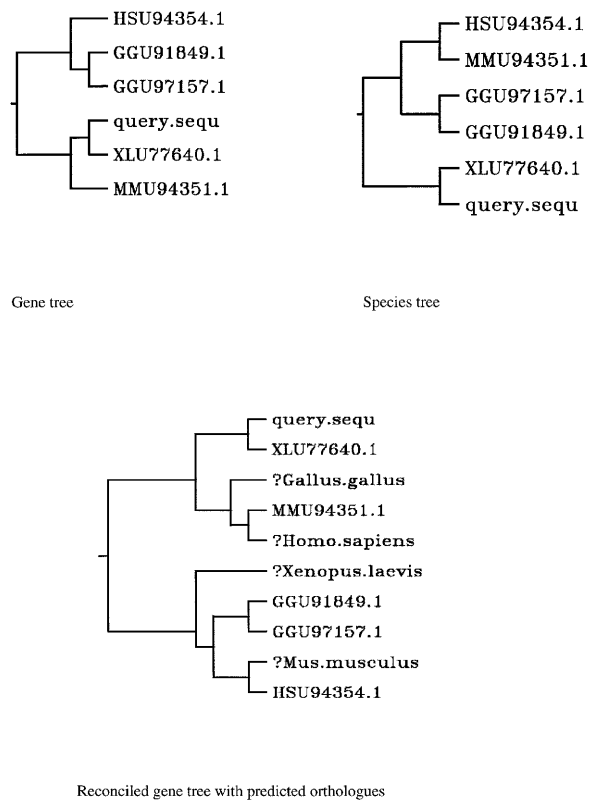


Fig. 2. Phylogenetic trees as derived from the procedure (these trees are displayed graphically and also given in a digital form that can be used by various tree-drawing procedures on different platforms). (**a**) Gene tree as computed by CLUSTALW (Higgins *et al.*, 1996). The identifiers from a non-redundant protein database at EMBL are given. (**b**) Tree as taken from the GENBANK taxonomy according to the species included. The species of the query sequence has to be entered into the Web interface; species corresponding to the matching database sequences are retrieved and appended. (**c**) Reconciled tree that solves discrepancies in (a) and (b). A root is implied by assuming the minimal number of gene duplication events. Duplications after the divergence of species (see, for example, the two chicken sequences) are considered as being independent events that do not need to be considered in the computation. As a minimal number of duplications is assumed, the chicken and the human sequence have been considered as being part of one branch and the root has been placed accordingly. A '?' sign precedes the species name if a gene in this species is assumed to have to exist. Accordingly, four genes (either extinct or not yet observed) are assumed here and a gene duplication event leads to two groups of orthologues.

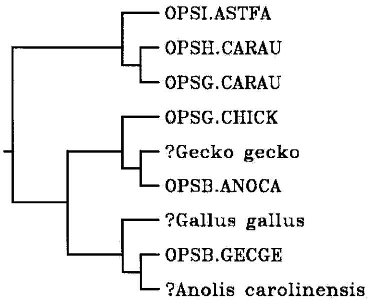


Fig. 3. Reconciled tree of the opsin RH2 group (drawn by the PHYLIP's drawgram). The same relevant protein sequences are chosen as indicated in the paper of Kawamura and Yokoyama (1995). The proteins OBSB_GECGE and OPSB_ANOCA are paralogues as they might be derived from a common ancestral sequence. On other side, OBSB_ANOCA could be considered as an orthologue to the chicken's green opsin protein OPSG_CHICK.

analysis of the evolutionary relationship between the sequences difficult. Furthermore, it is hard to distinguish between orthologues and paralogues among the homologous sequences. In order to avoid this weakness and provide a new approach in this direction, we developed this method and give it as a service on the WWW site (http://www.bork.embl-heidelberg.de/Blast2e/). Using this approach, users can easily retrieve all the sequences in BLAST output and all the phylogenetic relationship between the sequences. A table on all pairwise distances of the involved sequences and bootstrapping values about the gene tree branches are given to enable users to judge the differences between the sequences and the quality of the gene tree. The global alignment option includes the full length of sequences which is useful as only the full length can provide the whole information about the sequences involved and the gene tree can be constructed much more accurately. In some cases, focus on local matched regions taken out of BLAST search results proves to be reasonable as the sequences involved only share some functional similar parts (like multi-domain proteins). The reconciled tree provides the easiest way to decide on paralogues or orthologues among the homologous sequences. Nevertheless, this approach is limited if all species involved belong to the same genus, i.e. the species background cannot provide further resolving information for the genes under consideration. In these cases, the reconciled concept will fail. Another limitation of this approach is that only bifurcating trees are supported by the reconciling program, i.e. a root has to be chosen and only up to two genes from a species can be analysed by this reconciling program. Further developments are possible to overcome these limitations and close the gap between database search, gene family analysis and final functional prediction for the family members.

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (DFG).

References

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol*., **266**, 460–480.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Bork,P., Ouzounis,C. and Sander,C. (1994) From genome sequences to protein function. *Curr. Opin. Struct. Biol*., **4**, 393–403.

Durand,P., Canard,L. and Mornon,J.P. (1997) Visual BLAST and Visual FASTA: graphic workbenches for interactive analysis of full BLAST and FASTA outputs under Microsoft Windows 95/NT. *Comput. Applic. Biosci*., **13**, 407–413.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*., **266**, 114–128.

Felsenstein,J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

Fitch,W. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool*., **19**, 99–113.

Higgins,D.G., Thompson,J.D. and Gibson,T.J. (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*., **266**, 382–402.

Kawamura,S. and Yokoyama,S. (1995) Paralogous origin of the rhodopsinlike opsin gene in lizards. *J. Mol. Evol*., **40**, 594–600.

Page,R.D.M. (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol*., **43**, 58–77.

Page,R.D.M. and Charleston,M.A. (1997) From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol*., **7**, 231–240.

Saito,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*., **4**, 406–425.

Worley,K.C., Wiese,B.A. and Smith,R.F. (1995) BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res*., **5**,173–184.

Yuan,Y.P., Schultz,J., Mlodzik,M. and Bork,P. (1997) Secreted Fringe-like signaling molecules might be glycosyltransferases. *Cell*, **88**, 9–11.