



SimiTri—visualizing similarity relationships for groups of sequences

John Parkinson* and Mark Blaxter

Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, UK

Received on May 10, 2002; revised on August 14, 2002; accepted on September 17, 2002

ABSTRACT

Global sequence comparisons between large datasets, such as those arising from genome projects, can be problematic to display and analyze. We have developed SimiTri, a Java/Perl-based application, which allows simultaneous display and analysis of relative similarity relationships of the dataset of interest to three different databases. We illustrate its utility in identifying *Caenorhabditis elegans* genes that have distinct patterns of phylogenetic affinity suggestive of horizontal gene transfer. SimiTri is freely downloadable from <http://www.nematodes.org/SimiTri/> and the source code is freely available from the authors.

Contact: john.parkinson@ed.ac.uk

INTRODUCTION

Large scale sequencing initiatives are generating vast amounts of data for a wide variety of organisms, permitting large-scale comparisons between taxa at the molecular level. These analyses may reveal the evolutionary relationships and molecular mechanisms responsible for biological diversity and specialization between organisms. While the majority of genes homologous between genomes might be expected to yield congruent phylogenies, processes such as gene duplication, gene loss, gene conversion, differing selective constraints and lateral gene transfer can result in conflicting trees (Sicheritz-Ponten and Andersson, 2001).

Methods for the analysis of phylogenetic relationships within individual gene families are well developed, but it is more difficult to place such relationships in the context of all gene families. A whole dataset analysis would reveal how representative a particular protein is for a given dataset, and provide insights into the underlying mechanisms responsible for their evolution. While it is clear that only rigorous phylogenetic analysis, including considerations of problems such as paralogy, is effective at defining the evolutionary history underlying a present day pattern of gene similarity, simple pairwise similarity scores can be used as a first estimate of closest neighbour

relationships. These pairwise similarity scores can be derived from a number of algorithms, such as BLAST or FASTA, and have previously been represented as pie charts or Venn diagrams (Blaxter *et al.*, 2002; Wood *et al.*, 2002). Such displays show which group/organism a particular gene is most related to, but are limited in that they must use unitary cut-offs of similarity and thus present one, static view of sequence relationships.

Simultaneous display of relative similarity to a number of datasets can be very revealing, identifying for example, genes that are absent from one or more comparators (suggesting gene loss, or gene birth), or that have divergent similarity relationships compared with the bulk of the dataset (suggesting horizontal gene transfer or convergent evolution). Similarity scores for one dataset compared to two others can be displayed as a simple scatter plot, and individual sequences of interest pinpointed by their position off the main diagonal. For many purposes, this two way comparison is insufficient. For example, in comparing free living and parasitic nematodes (Parkinson *et al.*, 2001) we are interested in both shared (nematode-specific) and unique (potentially parasite specific) novelties that might have promise as drug targets or vaccine candidates. Simultaneous display of the relative similarity of one parasite's proteome to the proteomes of a free-living nematode *Caenorhabditis elegans* (*C.elegans*), a non-nematode and a second parasitic species would aid this. Here we describe a Java/Perl visualization tool that enables such comparisons.

METHODS

Generation of similarity data

For each sequence of interest, a 'similarity profile' was created by searching for sequence similarity against a number of different protein and nucleotide sequence databases using the program BLAST (Altschul *et al.*, 1990, 1997). For each database, the highest BLAST scores (bit score values) in excess of 50 were extracted. In order to maintain consistency in scores between different types of databases (protein and nucleotide), nucleotide versus nucleotide comparisons were performed using the

*To whom correspondence should be addressed.

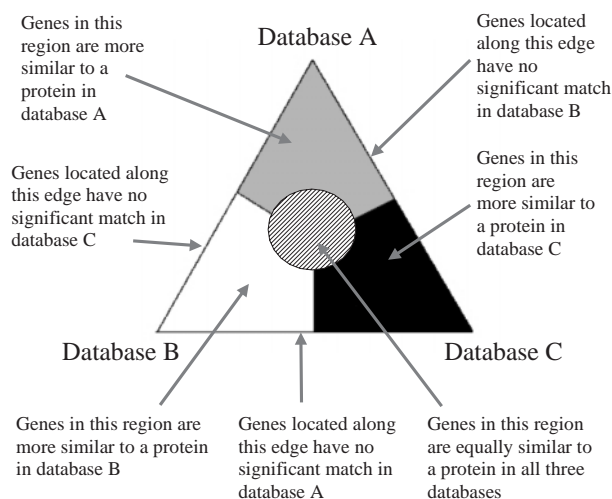


Fig. 1. Schematic indicating how the relative similarity relationship to three datasets for a particular sequence can be shown as a point in triangular phase space.

TBLASTX program.

The profile for each sequence consists of a row of numbers indicating its similarity to sequences from the chosen databases. This file is then used as input for the viewer tool, SimiTri (see below). Here we have utilized BLAST to generate similarity information, however it should be noted that any other tool or measure of similarity could be used. Perl scripts are used to parse the BLAST output to extract, for each query sequence, the relevant score, and output it to a profile file.

Visualization of data—SimiTri

With a two-dimensional display it is possible to identify the similarity relationships of a group of proteins to three distinct datasets using triangular phase space (see Figure 1). For a particular gene, its position in this phase space can be calculated from the BLAST scores obtained for three selected databases. In effect, each score represents the length of the vector from the centre of the triangle to each of its nodes, and the position of a gene within this phase space thus indicates its relative BLAST score to each of the three chosen datasets. Genes which have significant similarity to only two datasets will be mapped to the edge joining the two datasets. Genes with significant similarity scores to only one, or no, comparator datasets may be tabulated separately.

SimiTri is available as a standalone package for use with local datasets. It has been installed as a web front end to our nematode datasets (<http://www.nematodes.org/nematodeESTs/nembase.html>).

The standalone package consists of two main applications. The first, written in Perl, takes as input a space-

delimited file containing a list of sequence identifiers associated with the scores for three user selected databases and calculates the phase space coordinates for each sequence from this data. The second program consists of a Java applet that can be launched using an appropriate appletviewer tool (a standard web page is included in the package) and displays the results of the query in a user interactive application. The java applet consists of a large window displaying the location of the selected genes in the triangular phase space. Each gene is represented as a square tile, coloured by the highest similarity score found in comparisons to the three datasets. The user can zoom in and out, and move around the triangular phase space. The user can determine the window of similarity scores displayed, using interactive controls.

In addition to the standalone package, we have included SimiTri as part of the services offered for analyzing our nematode sequence datasets. In brief, for our nematode datasets, the user may select any 3 of 17 different comparator databases containing the precomputed BLAST scores.

RESULTS AND DISCUSSION

Comparison of *C.elegans* proteins with yeast, fly and human proteins

Figure 2 shows an example of SimiTri analysis of the predicted *C.elegans* proteome (as found in WORMPEP, release 69, http://www.sanger.ac.uk/Projects/C_elegans/wormpep/; The *C.elegans* Sequencing Consortium, 1998). For each protein, a BLASTP search was performed against 3 different organism-specific protein databases; *Saccharomyces cerevisiae* (31,897 proteins) and *Drosophila melanogaster* (22,677 proteins) (downloaded on 5/12/01 using the SRS facility at the European Bioinformatics Institute; <http://srs.ebi.ac.uk>) and *Homo sapiens* (181,288 proteins, including 152,582 predicted by genscan; downloaded on 5/12/01 from Ensembl <http://www.ensembl.org>). It should be noted that the unexpectedly large numbers of proteins obtained for *S.cerevisiae* and *D.melanogaster*, arise from the presence of alternative splice variants, isoforms and duplicate sequences in GenBank. The highest BLAST score was obtained for each protein/dataset comparison and used to calculate SimiTri coordinates. Of the 20,311 predicted *C.elegans* proteins, 8938 had significant BLAST similarity to more than one protein database and are shown in Figure 2. The majority of proteins are located in the lower half of the triangle, suggesting that they are more closely related to *D.melanogaster* and *H.sapiens* than to *S.cerevisiae*, as would be expected. The relatively equal spread of proteins from left to right indicates that overall, *C.elegans* proteins are equally related to *H.sapiens* and *D.melanogaster* proteins. Interestingly, the proteins

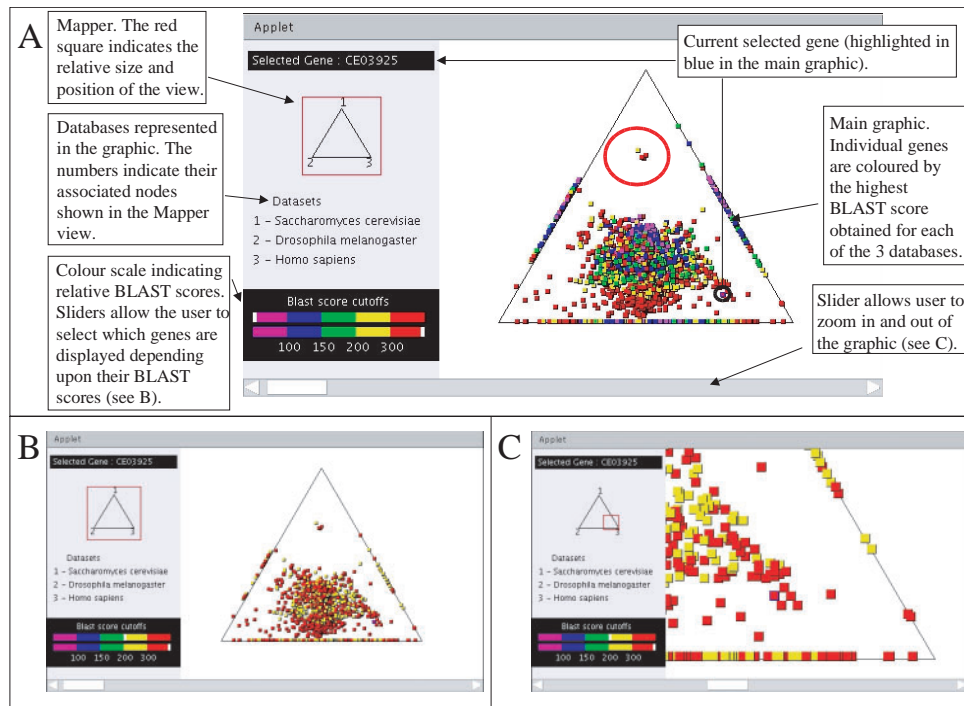


Fig. 2. SimiTri representation of the *C.elegans* predicted proteome (Wormpep 69) showing their BLAST score relationships with *S.cerevisiae*, *D.melanogaster* and *H.sapiens* protein databases. (a) Entire dataset. (b) Dataset showing genes which have a BLAST score >200 to one or more of the three databases. (c) Enlarged portion of dataset B.

aligned along the edge joining nodes 1 and 2 and those aligned along the edge joining nodes 1 and 3 represent those proteins that may have been lost from the *H.sapiens* (1–2) and *D.melanogaster* (1–3) genomes respectively. Proteins along the edge joining nodes 2 and 3 have either arisen in the metazoan lineage, or possibly, have been lost from yeast.

Of particular interest are the four proteins shown within the outlined red circle. These proteins are more similar to a *S.cerevisiae* protein than to either a fly or human protein (see Figure 3). One, CE29804 (F48E8.3) is similar to a range of formate dehydrogenase flavoenzymes from a wide range of organisms but absent from metazoans except *C.elegans*. Phylogenetic analysis groups CE29804 with a clade of yeast and protozoan sequences (Figure 3a). Related genes were also identified in EST datasets from the nematodes *Strongyloides ratti* (GenBank accession BI742141) and *Ascaris suum* (GenBank accession BF050077) but not elsewhere. The three other proteins identified; CE20628 (D2063.1), CE12214 (K12G11.3) and CE12212 (K12G11.4), are all very closely related (Figure 3b). K12G11.3 and K12G11.4 are neighbouring genes (tail-to-tail conformation) on chromosome V, while D2063.1 also located on chromosome V, is about 7 MB distant. The three genes are most closely related to each

other, and all identify a family of fungal, eubacterial and plant alcohol dehydrogenases. BLAST analysis reveals no metazoan similarity within the top 100 high-scoring segment pairs. A single EST from the nematode *Zeldia punctata* (GenBank accession AW783790) encodes a protein most closely related to these *C.elegans* sequences. Thus these four genes, identified by SimiTri, do indeed have atypical phylogenetic profiles, and are candidates for genes either present in an ancestral metazoan, and lost from all those examined to date except *C.elegans* and some related nematodes, or acquired, probably from fungal donors, by horizontal transfer to an ancestral nematode.

BLAST similarity scores for a population of sequences reflects accepted phylogeny for a range of parasitic nematodes

As indicated above, BLAST or other similarity scores derived from pairwise comparison may not be an accurate measure of phylogenetic proximity for some sequences (Mushegian *et al.*, 1998; Xie and Ding, 2000; Koski and Golding, 2001). For completed genomes, systems have been developed which automatically derive neighbour joining trees for all protein families from a given set of organisms (Sicheritz-Ponten and Andersson, 2001;

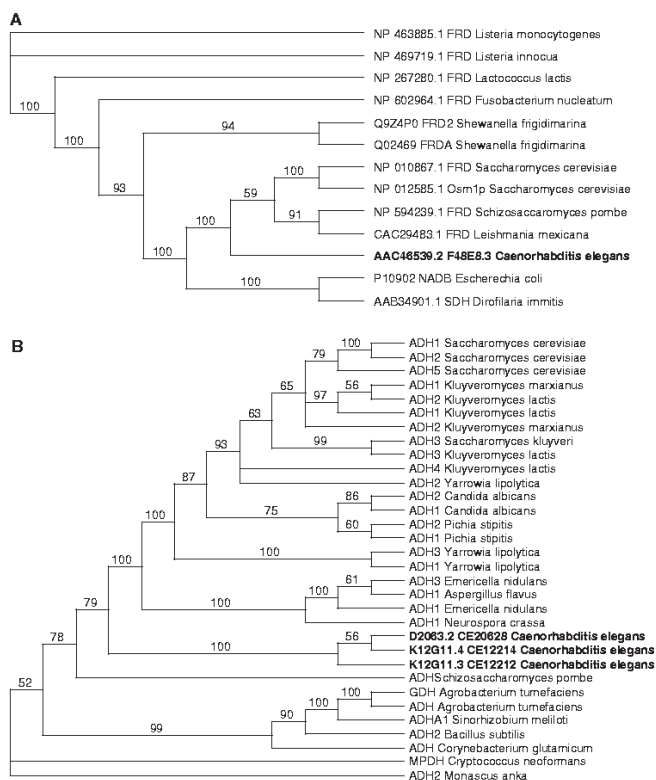


Fig. 3. Possible lateral gene transfers in *C.elegans*. (a) F48E8.3 CE29804 is similar to a range of formate dehydrogenase flavoenzymes from a wide range of organisms. The 100 proteins with highest PSI-BLAST scores were selected and aligned using CLUSTALX, and the resulting alignment subjected to maximum parsimony phylogenetic analysis using PAUP*v4.10 (Swofford, 2000) with 1000 bootstrap replicates. From this global analysis the twelve most closely related sequences were analyzed in detail. F48E8.3 is found within a clade of eukaryote formate dehydrogenases, with representatives from *S.cerevisiae*, *Schizosaccharomyces pombe* and the kinetoplastid protozoan *Leishmania major*. There are no other closely related metazoan sequences. A succinimide dehydrogenase from the parasitic nematode *Dirofilaria immitis* is related to this clade, but is itself more similar to NADB from *Escherichia coli*. Sequences related to F48E8.3 were identified in EST datasets from the nematodes *Strongyloides ratti* (BI742141), and *Ascaris suum* (BF050077) (data not shown). (b) D2063.1 CE20628, K12G11.3 CE12212, and K12G11.4 CE12214. K12G11.3 and K12G11.4 are neighbouring genes (tail-to-tail conformation) on chromosome V, while D2063.1 is also on chromosome V, but over 5 Mb away. The three genes are most closely related to each other, and all identify a family of fungal and eubacterial Zinc-containing alcohol dehydrogenases (Glaser *et al.*, 1995). There are no other metazoan alcohol dehydrogenases of this class. A single EST from the nematode *Zeldia punctata* (AW783790) encodes a protein closely related to these *C.elegans* sequences.

Daubin *et al.*, 2002). Resulting tree topologies are then analyzed either by consensus; (Sicheritz-Ponten and Andersson, 2001), or by parsimony analysis of reconstructed

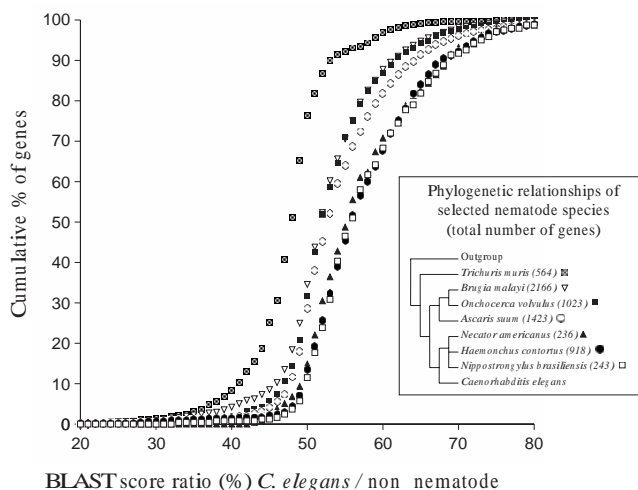


Fig. 4. Graph showing the cumulative percentage of genes for 6 different species of nematode plotted as a function of the percentage ratio of BLAST scores obtained from comparisons to the *C.elegans* and non-nematode protein databases. Genes which are more similar to a *C.elegans* protein will have a ratio of >50% whilst those which are more similar to a non-nematode protein will have a ratio of <50%. Only genes with a significant BLAST score (>50) to both databases were used.

nodes in each tree (Daubin *et al.*, 2002), to both define the 'majority consensus' for organismal phylogeny and highlight protein families that yield trees incongruent with this. Principal coordinates analysis has also been used to identify individual protein phylogenies that are in conflict with the bulk of a dataset (Daubin *et al.*, 2002; Matte-Tailliez *et al.*, 2002).

SimiTri was developed to provide a quick and robust method to examine similarity relationships and identify sequences demonstrating atypical behaviour. It is worth noting that SimiTri is not limited to the use of sequence alignment scores, as any measures of similarity such as relative expression levels or other sequence properties could be used for comparative purposes. For the purposes of this paper we have used BLAST as it provides a fast and efficient heuristic.

To test how BLAST scores reflect the phylogenetic relationships for the nematode taxa examined, we investigated the pattern of relative BLAST scores for each of seven parasitic nematode species compared to *C.elegans* and non-nematode proteins. The sequences used in this study were generated as part of the Edinburgh parasitic nematode EST project (Parkinson *et al.*, 2001). In brief, ESTs were clustered on the basis of sequence similarity and used to generate consensus nucleotide sequences using an in-house program (Parkinson *et al.*, 2002). These consensus sequences were then used to generate 'similarity profiles' (see Figure 4). The shape of each species'

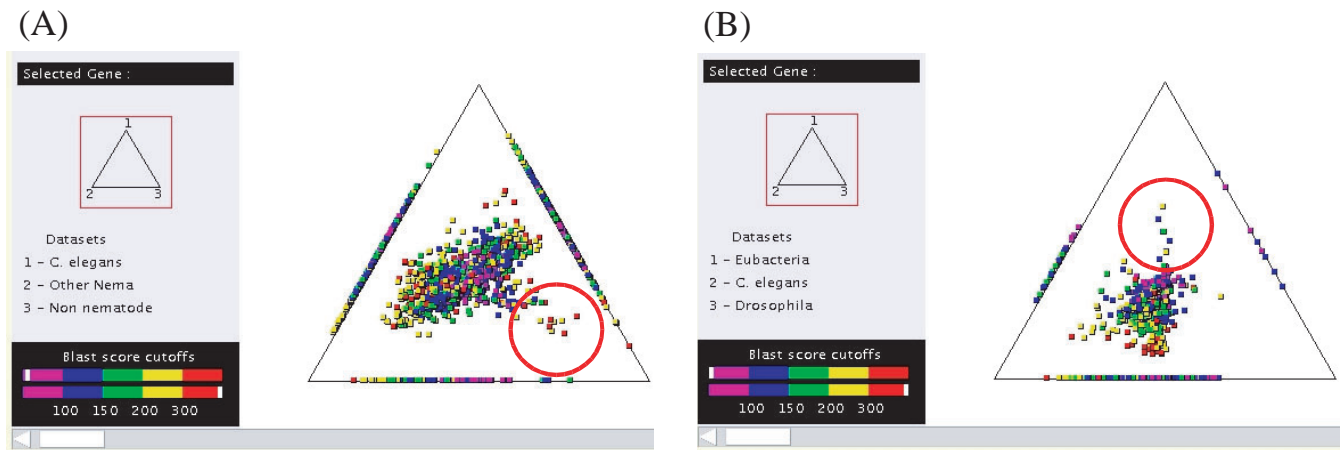


Fig. 5. SimiTri representation of predicted *Brugia malayi* genes. (a) *B. malayi* compared to *C. elegans*, other nematodes and non-nematode proteins. (b) *B. malayi* compared to eubacteria, *D. melanogaster* and *C. elegans*.

curve correlates well with the predicted phylogenetic relationships obtained from small subunit ribosomal RNA analysis (Blaxter *et al.*, 1998). For example, the curves for *Haemonchus contortus* and *Necator americanus*, both Clade V nematodes relatively closely related to *C. elegans*, are located to the right of the graph compared with the other curves, whilst the curve representing the *Trichuris muris* dataset, a Clade I nematode, most distantly related to *C. elegans*, is located to the left of the graph.

SimiTri identifies genes with unexpected patterns of phylogenetic affinity

Whilst the results above indicate that the use of BLAST similarity for a population of sequences gives very good agreement with accepted phylogenies, there are sequences that do not fit with the standard accepted phylogeny. These include genes that are evolving with a different dynamic from the majority of the genes, perhaps due to functional constraints and genes that may have originated by horizontal gene transfer from another distantly related species.

Figure 5 shows two SimiTri profiles of predicted genes from the filarial nematode *Brugia malayi* (Blaxter *et al.*, 2002) compared to *C. elegans* proteins, nematode proteins excluding *C. elegans*, and non-nematode proteins (Figure 5a) or to *C. elegans* proteins, *D. melanogaster* proteins and eubacterial proteins (Figure 5b). Outlined in red in Figure 5a is a region that identifies genes that are more similar to the non-nematode dataset than to the nematode datasets. These proteins may include genes that have evolved to interact with the host (Maizels *et al.*, 2001), and host contaminants of the original cDNA libraries used. In Figure 5b, there are a number of genes that appear to be more related to eubacteria than to the other two datasets (again outlined in red). Since *Brugia malayi*

is known to have an alphaproteobacterial endosymbiont (Bandi *et al.*, 1998), these genes are candidate endosymbiont transcripts, and indeed map to the endosymbiont genome (unpublished data).

It should be noted that in addition to whole genome datasets, SimiTri can also be applied to subsets of genes such as specific gene families or those displaying distinct expression profiles. For example, SimiTri could be used to view comparisons of gene families (e.g. protein kinases) and help identify recent, within genome duplications compared to duplications of more ancient origin. As mentioned previously, SimiTri is not limited to the use of BLAST similarity scores. Other indices such as FASTA or Smith–Waterman scores, relative expression profiles derived from e.g. microarray experiments and even simple measures such as sequence length could also be used for visual comparative purposes.

SimiTri is available as part of NEMBASE—a nematode specific sequence resource (<http://www.nematodes.org>). Groups of putative genes from NEMBASE may be selected either on the basis of whole organism, by stage specificity or by BLAST annotation.

REFERENCES

- Altschul,S.F., Gish,W., Miller,W., Myers,M.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Bandi,C., Anderson,T.J.C., Genchi,C. and Blaxter,M. (1998) Phylogeny of *Wolbachia* in filarial nematodes. *Proc. R. Soc. Lond. B.*, **265**, 2407–2413.

- Blaxter,M.L., DeLey,P., Garey,J.R., Liu,L.X., Scheldeman,P., Vierstraeten,J.R., Mackey,L.Y., Dorris,M., Frisse,L.M., Vida,J.T. and Thomas,W.K. (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.
- Blaxter,M.L., Daub,J., Guiliano,D., Parkinson,J. and Whitton,C. (2002) The *Brugia malayi* genome project: expressed sequence tags and gene discovery. *Trans. R. Soc. Trop. Med. Hyg.*, **96**, 7–17.
- Daubin,V., Gouy,M. and Perriere,G. (2002) A phylogenetic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.*, **12**, 1080–1090.
- Glasner,J.D., Kocher,T.D. and Collins,J.J. (1995) *Caenorhabditis elegans* contains genes encoding two new members of the Zn-containing alcohol dehydrogenase family. *J. Mol. Evol.*, **41**, 46–53.
- Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbour. *J. Mol. Evol.*, **52**, 540–542.
- Maizels,R.M., Blaxter,M.L. and Scott,A. (2001) Immunological genomics of *Brugia malayi*: filarial genes implicated in immune evasion and protective immunity. *Parasite Immunol.*, **23**, 327–344.
- Matte-Tailliez,O., Brochier,C., Forterre,P. and Philippe,H. (2002) Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.*, **19**, 631–639.
- Mushegian,A.R., Garey,J.R., Martin,J. and Liu,L.X. (1998) Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res.*, **8**, 590–598.
- Parkinson,J., Whitton,C., Guiliano,D., Daub,J. and Blaxter,M.L. (2001) 200 000 nematode expressed sequence tags on the net. *Trends Parasit.*, **17**, 394–396.
- Parkinson,J., Guiliano,D. and Blaxter,M.L. (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinf.*, **3**, 31.
- Sicheritz-Ponten,T. and Andersson,S.G.E. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res.*, **29**, 545–552.
- Swofford,D.L. (2000) *PAUP** (*Phylogenetic Analysis using Parsimony and Other Methods*), Version 4.10, Sinauer Associates, Sunderland, MA.
- The *C.elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C.elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
- Wood,V., William,R.G., Rajandream,M.A., Lyne,M., Stewart,A. *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.
- Xie,T. and Ding,D. (2000) Investigating 42 candidate orthologous protein groups by molecular evolutionary analysis on genome scale. *Gene*, **261**, 305–310.