A cross-comparison of a large dataset of genes

BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 7 2000
Pages 654–655

A cross-comparison of a large dataset of genes

G. Cannarozzi[1], M. T. Hallett[2,*], J. Norberg[2] and X. Zhou[2,3]

[1]Department of Chemistry, University of Florida, Gainesville 32611, USA, [2]Computational Biochemistry Research Group (CBRG), Department of Computer Science, ETH Zentrum, Zürich, Switzerland and [3]Department of Pharmaceutical Chemistry, ETH Irchel, Zürich, Switzerland

Received on November 28, 1999; revised on March 2, 2000; accepted on March 15, 2000

Abstract

Summary: *We make available a large cross-comparison for 16 of the completely sequenced genomes and additional eukaryotic genes. The alignments were performed at the protein level using liberal similarity bounds in order to capture as many significant alignments as possible. This dataset will be updated as new genomes become available.*

Availability: *The dataset is accessible at http://cbrg.inf.ethz.ch/AllAllGeneAlign.html*

Contact: *hallett@inf.ethz.ch*

Many research projects in molecular biology require the alignment of a large set of protein sequences against a large set of protein sequences. These cross-comparisons are important since each alignment provides clues about the evolution, structure and, ultimately, function of these objects [Chervitz *et al.* (1998) Gonnet *et al.* (1992) and Snel *et al.* (1999) provide examples where such cross-comparisons have been used]. With the ever-increasing availability of molecular sequence data—including complete genomes—the size of these sets has become quite large and many bioscience researchers lack the computational power and time needed to perform these computations. Over the past 2 years, the CBRG has computed cross-comparisons numerous times for different research groups on a first-come, first-serve basis. To this end, we make available to the research community a cross-comparison of a set of 83 216 genes where 37 883 of these genes come from 16 of the completely sequenced genomes. The remaining genes come from various eukaryota taxa in order to provide a more complete sampling over the Tree of Life (see Table 1).

The protein and DNA coding sequences for the complete genomes were obtained from GenBank (Benson *et al.*, 1999) and the eukaryota sequences were obtained via SRS at EBI (Etzold *et al.*, 1996). We performed our alignments at the protein level using a standard dynamic programming approach (Smith and Waterman, 1981) in the Darwin system (Gonnet *et al.*, 2000). It is believed that a full dynamic programming approach performs significantly better than a standard BLAST alignment for detecting long-distance homologues (Brenner *et al.*, 1998; Shaper *et al.*, 1996). We used the GCB matrices with an affine gap penalty using the empirical values given in Gonnet *et al.* (1992).

The cross-comparison was carried out in two steps. First, we performed a local alignment of every protein sequence against every other protein sequence with GCB scoring matrices built for PAM distance 250. Any two sequences that had a local alignment that passed a similarity score threshold of 120 were considered significant. Second, all pairs of significant alignments from phase 1 were refined using a (computationally more expensive) Smith–Waterman routine that seeks to find the PAM distance that maximizes the similarity score (Gonnet *et al.*, 1992).

In total, we performed 3.46×10^9 alignments of which 6.33×10^6 were found to be significant at the 120 level (5.87×10^6 were found to be significant at the 150 level). We used these similarity bounds in order to err on the side of *false positives*. In other words, although the dataset contains alignments that do not represent true homology, it should also include most potential long-distance homologues. The purpose of these cross-comparisons is not to directly provide a curated list of gene families but instead to supply researchers with a starting set of alignments that share at least one homologous region. Researchers may then concentrate their analyses on this subset of the data. The protein and its corresponding coding DNA are stored in SGML format. This information is accessible from our server. For a selected set of organisms, it returns the appropriate genes and the cross-comparison.

Via the process support system BioOpera (Alonso *et al.*, 2000), the computation of these large cross-comparisons is now almost entirely automatic. Our intention is to use this software to compute the cross-comparison for each new genome as it becomes available.

*To whom correspondence should be addressed.

Table 1. Genomes and other sequences

Taxa	Number of genes	Kingdom	Total(%)
Archaeoglobus fulgidus	2407	Archaea	2.9
Methanococcus jannaschii	1715	Archaea	2.1
Methanobacterium thermoautotrophicum	1869	Archaea	2.3
Pyrococcus horikoshii	2064	Archaea	2.5
			9.8
Aquifex aeolicus	1522	Eubacteria	1.8
Borrelia burgdorferi	850	Eubacteria	1.0
Bacillus subtilis	4100	Eubacteria	4.9
Escherichia coli K12	4289	Eubacteria	5.2
Haemophilus influenzae Rd	1709	Eubacteria	2.1
Helicobacter pylori	1553	Eubacteria	1.9
Mycoplasma genitalium	480	Eubacteria	0.6
M.pneumoniae	677	Eubacteria	0.8
Mycobacterium tuberculosis	3918	Eubacteria	4.7
Synechocystis sp.	3169	Eubacteria	3.8
Treponema pallidum	1031	Eubacteria	1.2
			28.0
Saccharomyces cerevisiae	6530	eukaryota	7.8
Homo sapiens	10740	eukaryota	12.9
Rodents	5664	eukaryota	6.8
Other mammals	6234	eukaryota	7.5
Other invertebrates	16014	eukaryota	19.2
Other vertebrates	6681	eukaryota	8.0
			62.2
Total	83216		100

A.fulgidus, Klenk,H.-P. *et al.* (1997) *Nature*, **390**, 364–370; *A.aeolicus*, Deckert,G. *et al.* (1998) *Nature*, **392**, 353–358; *B.burgdorferi*, Fraser,C.M. *et al.* (1997) *Nature*, **390**, 580–586; *B.subtilis*, Kunst,F. *et al.* (1997) *Nature*, **390**, 249–256; *E.coli*, Blattner,F.R. *et al.* (1997) *Science*, **277**, 1453–1462; *H.influenzae*, Fleischmann,R.D. *et al.* (1995) *Science*, **269**, 496–512; *H.pylori*, Tomb,J.-F. *et al.* (1997) *Nature*, **388**, 539–547; *M.genitalium*, Fraser,C.M. *et al.* (1995) *Science*, **270**, 397–403; *M.jannaschii*, Bult,C.J. *et al.* (1996) *Science*, **273**, 1058–1073; *M.pneumoniae*, Himmelreich,R. *et al.* (1996) *Nucleic Acids Res*., **24**, 4420–4449; *M.thermoautotrophicum*, Smith,D.R. *et al.* (1997) *J. Bacteriol*., **179**, 7135–7155; *M.tuberculosis*, Cole,S.T. *et al.* (1998) *Nature*, **393**, 537–544; *P.horikoshii*, Kawarabayasi,Y.*et al.* (1998) *DNA Res*., **5**, 55–76; *Synechocystis* sp., Kaneko,T. *et al.* (1996) *DNA Res*., **3**, 109–136; *T.pallidum*, Fraser,C.M. *et al.* (1998) *Science*, **281**,375–388; *S.cerevisiae*, Goffeau,A. *et al.* (1997) *Nature (Suppl.)*, **387**, 5–105.

Acknowledgments

J.N. acknowledges a postdoctoral fellowship from The Swedish Foundation for International Cooperation in Research and Higher Education.

References

Alonso,G., Bausch,W., Hallett,M., Kahn,A. and Pautasso,C. (2000) BioOpera:managing large-scale computations in bioinformatics. Submitted to *ISMB '00*.

Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) GenBank. *Nucleic Acids Res.*, **27**, 12–17.

Brenner,S.E., Chothia,C. and Hubbard,T.J. P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA*, **95**, 6073–6078.

Chervitz,S.A., Aravind,L., Sherlock,G., Ball,C.A., Koonin,E.V., Dwight,S.S., Harris,M.A., Dolinski,K., Mohr,S., Smith,T., Weng,S., Cherry,J.M. and Botstein,D. (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.

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

Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1445.

Gonnet,G.H, Hallett,M.T., Korostensky,C. and Bernardin,L. (2000) Darwin v. 2.0. An interpreted computer language for the biosciences. *Bioinformatics* (accepted).

Shpaer,E.G., Robinson,M., Yee,D., Candlin,J.D., Mines,R. and Hunkapiller,T. (1996) Sensitivity and selectivity in protein similarity searches: a comparison of Smith-Waterman in hardware to BLAST and FASTA. *Genomics*, **38**, 179–191.

Snel,B., Bork,P. and Muynen,M.A. (1999) Genome phylogeny based on gene content. *Nat. Genet.*, **21**, 108–110.

Smith,T. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.