BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 13 2003, pages 1710–1711
DOI: 10.1093/bioinformatics/btg213

Detecting putative orthologs

D. P. Wall[1,*], H. B. Fraser[2] and A. E. Hirsh[1]

[1]Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA and [2]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA

Received on February 19, 2003; revised on March 5, 2003; accepted on March 10, 2003

ABSTRACT

Summary: We developed an algorithm that improves upon the common procedure of taking reciprocal best blast hits (rbh) in the identification of orthologs. The method—reciprocal smallest distance algorithm (rsd)—relies on global sequence alignment and maximum likelihood estimation of evolutionary distances to detect orthologs between two genomes. rsd finds many putative orthologs missed by rbh because it is less likely than rbh to be misled by the presence of a close paralog.

Availability: A Python program and ReadMe file are freely available from: http://charles.stanford.edu/~dennis/research.html

Contact: dpwall@stanford.edu

When comparing the evolutionary rates of proteins in the absence of a normalizing molecular clock, rate estimates must be based upon comparisons between sequences that are orthologs (sequences that diverged from each other at the species split), and not paralogs (sequences that diverged at another time). Only if all sequence comparisons share the same time of divergence are protein evolutionary distances expected to be proportional to relative evolutionary rates. For example, in Figure 1, the orthologous comparisons would yield evolutionary distances indicative of the relative rates of protein evolution. By contrast, paralogous comparison of A or B with E would yield an evolutionary distance that would badly overestimate the evolutionary rate.

A common procedure for identifying sequence pairs that are putatively orthologous, and therefore admissible for estimation of relative evolutionary rate, is reciprocal best hit (rbh) (Hirsh and Fraser, 2001; Jordan *et al*., 2002). Protein i in genome I is a rbh of protein j in genome J if query of genome J with protein i yields as the top hit protein j, and reciprocal query of genome I with protein j yields as the top hit protein i. However, blast search often returns as the highest scoring hit a protein that is not the nearest phylogenetic neighbor of the query sequence (Koski and Golding, 2001). If the forward blast yields a paralogous best hit but the reciprocal blast recovers an actual ortholog, both pairs will be excluded. Thus,

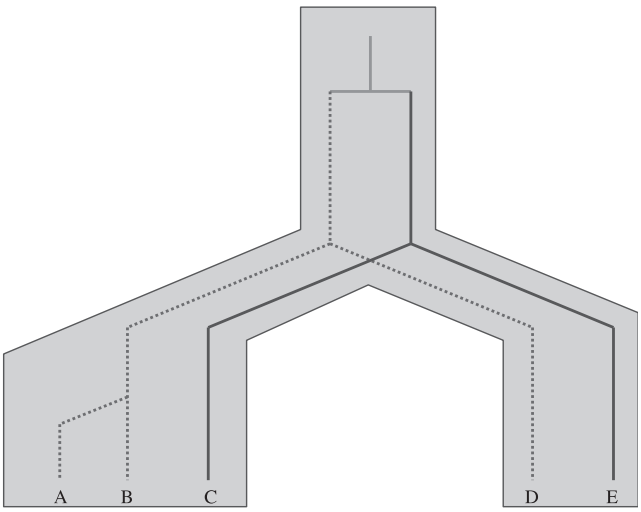

Fig. 1. Identification of probable orthologs between two genomes. The thin lines represent the gene tree, thick lines represent the organismal tree. Admissible orthologous pairs are A or B with D and C with E. Inadmissible paralogous pairs are A or B with E and C with D.

while rbh will rightfully prevent admission of the paralogous pair to the set of proteins for which relative evolutionary rates are estimated, it will also wrongly exclude an authentically orthologous pair from consideration. Here we describe a new algorithm that preserves the safeguard of reciprocal genome queries, but is less vulnerable to exclusion of orthologs due to identification of a paralog in one blast direction.

The method employs blast (Altschul *et al*., 1990) as a first step, starting with a subject genome, J, and a protein query sequence, i, belonging to genome I. A set of hits, H, exceeding a predefined significance threshold (e.g. $E < 10^{-20}$) is obtained. Then, using Clustalw (Thompson *et al*., 2000), each protein sequence in H is aligned separately with the original query sequence i. If the alignable region of the two sequences exceeds a threshold fraction of the alignment's total length (0.8 is our working cutoff), the program PAML (Yang, 2000) is used to obtain a maximum likelihood estimate of the number of amino acid substitutions separating the two protein sequences, given an empirical amino

*To whom correspondence should be addressed.

acid substitution rate matrix (Jones *et al*., 1992). The model under which a maximum likelihood estimate is obtained may include variation in evolutionary rate among protein sites, and for more distant comparisons we have generally assumed a gamma distribution with shape parameter $\alpha = 1.53$ (Nei *et al*., 2001). Of all sequences in H for which an evolutionary distance is estimated, only j, the sequence yielding the shortest distance, is retained. This sequence j is then used for a reciprocal blast against genome I, retrieving a set of high scoring hits, L. If any hit from L is the original query sequence, i, the distance between i and j is retrieved from the set of smallest distances calculated previously. The remaining hits from L are then separately aligned with j and maximum likelihood distance estimates are calculated for these pairs as described above. If the protein sequence from L producing the shortest distance to j is the original query sequence, i, it is assumed that a true orthologous pair has been found and their evolutionary distance is retained.

We tested the algorithm in comparisons between the *Saccharomyces cerevisiae* genome and two other complete genomes: *Candida albicans* and more distantly related *Caenorhabditis elegans*. We compared the results with the set from rbh in each case. In the first, reciprocal smallest distance algorithm (rsd) yielded 2777 unique, putatively orthologous pairs, whereas rbh produced 1824 pairs. All orthologs found by rbh were also found by rsd. In the second, rbh found 526 admissible pairs and rsd found 816. Again, all orthologous pairs found by rbh were also found using rsd.

Orthologous pairs retrieved by rsd but not rbh are presumably cases in which blast returned a paralog as the best hit in at least one direction, missing the ortholog even though it was among the high-scoring hits. In our algorithm, global alignment and evolutionary distance estimation recover the ortholog, revealing that it is in fact the nearest evolutionary neighbor of the query, though not the best blast hit, thereby retrieving more authentically orthologous pairs. An additional improvement over rbh beyond the use of global alignments is that rsd's reliance on maximum likelihood distances obviates the need to convert blast scores or fractional pairwise differences into evolutionary distances.

ACKNOWLEDGEMENTS

Thanks to two referees for their helpful comments. D.P.W. is supported by an NSF postdoctoral fellowship in bioinformatics, H.B.F. by an NSF predoctoral fellowship, A.E.H. and D.P.W. by NIH grants GM 28016 and 28424 to Marcus W. Feldman.

REFERENCES

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.

Hirsh,A.E. and Fraser,H.B. (2001) Protein dispensability and rate of evolution. *Nature*, **411**, 1046–1049.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Computer Application in Biosciences*, **8**, 275–282.

Jordan,I.K., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.*, **12**, 962–968.

Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.*, **52**, 540–542.

Nei,M., Xu,P. and Glazko,G. (2001) Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl Acad. Sci. USA*, **98**, 2497–2502.

Thompson,J.D., Plewniak,F., Thierry,J.C. and Poch,O. (2000) DbClustal: Rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, **28**, 2919–2926.

Yang,Z. (2000) Phylogenetic Analysis by Maximum Likelihood (PAML). University College London, London.