Predicting functional gene-links from phylogenetic-statistical analyses of whole genomes

Daniel Barker and Mark Pagel
School of Animal and Microbial Sciences, University of Reading, Whiteknights, P.O. Box 228, Reading, Berkshire, RG6 6AJ, UK
m.pagel@reading.ac.uk (MP)

Abstract

We describe a novel application of computational phylogenetic approaches to predict functional linkage among proteins, using proteomes derived from whole genome sequence data. The methods detect independent instances of the correlated gain or loss of pairs of genes on branches of a phylogenetic tree, on the assumption that functionally linked genes are often gained and lost at approximately the same time during evolution. According to this view, several correlated gain and/or loss events between a pair of genes suggests the gene products are functionally linked. We implement this approach using Dollo parsimony and maximum likelihood (ML) to seek correlated evolution among 21 eukaryotic species. We compare these approaches to each other and to the existing method of phylogenetic profiles, which seeks an across-species correlation but does not explicitly incorporate a phylogenetic tree. We assess all methods according to a positive test set of functionally linked protein pairs based on the MIPS catalogue of yeast protein complexes, and a negative test set of random protein pairs. Both Dollo parsimony and ML are able to achieve far greater specificity than the existing method of phylogenetic profiles. We show that ML is by far the best approach, provided that an appropriate model is used. Best results are obtained if the rate of gain of genes is fixed at a low value, to prevent modeling of multiple gains. With such a model, proteins with strong ML evidence of correlated evolution among eukaryotes are almost certainly functionally linked.

1. Introduction

A popular computational approach for predicting functional links is the method of phylogenetic profiles [1]. This predicts functional linkage on the basis that genes with correlated patterns of presence and absence across several species' genomes are likely to be functionally linked.

We show that results are improved if one seeks not correlated presence and absence of genes, but rather correlated gains and losses of genes on a phylogenetic tree. We implement this phylogenetic approach using Dollo parsimony [2], and ML approaches to detect correlated gain and loss [3] both with a relatively general ML model [4] and with a restricted model tailored to evolution of eukaryotic gene content. We show that not only the framework (phylogenetic profiles or correlated gain/loss), but the model within that framework has a profound effect on the quality of results.

2. Materials and methods

We obtained a species-by-proteins matrix showing presence ("1") or absence ("0") of orthologues of each yeast (*Saccharomyces cerevisiae*) gene in 20 other species using BLASTP [5] with Inparanoid [6]. We built a sequence-based phylogenetic tree using a concatenated multiple alignment of 19 single-copy genes found to be present across the 21 species.

For the method of phylogenetic profiles, the prediction score for candidate pairs of genes was the number of species which had a matching state for the two proteins [1].

For Dollo parsimony the scores we used were, firstly, the number of branches of the tree on which ability to code the two proteins changed in a positively correlated manner, and secondly, this value minus the number of branches of the tree on which non-correlated changes occur. We refer to these as "Dollo 1" and "Dollo 2", respectively.

We also evaluated ML approaches that compare a model of correlated evolution with a model of uncorrelated evolution of binary traits on a phylogeny [3]. Here, traits represent presence and absence of genes. We allowed any number of gains of the genes on the tree [4] ("ML 1"), and also restricted parameters repre-

senting initial gains of genes to an arbitrary low value of 0.1 ("ML 2"). In both cases the score is the likelihood ratio statistic for the model comparison.

Across the range of score cut-offs for each method we calculated specificity and sensitivity, using a negative test set of 441,217 pairs of proteins and a positive test set of test set of 9,178 pairs of proteins based on the MIPS Comprehensive Yeast Genome Database complex catalogue [7].

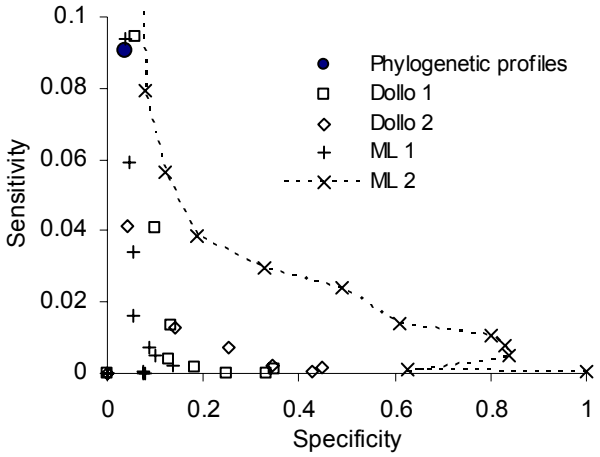

Figure 1. Performance of the five methods used for predicting functional links.

3. Results and discussion

The quality of results for each method are illustrated in Figure 1. Score cut-offs achieving sensitivity greater than 0.1 are not shown because they have poor specificity, for any of the methods.

The method of phylogenetic profiles predicts with poor discrimination, achieving a maximum specificity of only 0.04 (at sensitivity = 0.09). Only the point representing the maximum cut-off, of perfectly correlated distribution patterns, appears in Figure 1. Other cut-offs give greater sensitivity, but with even worse specificity.

The Dollo 1 method is more accurate, achieving specificity of 0.35 (at sensitivity 0.0009). The Dollo 2 method achieves still higher specifity peaking at at 0.45 (at an improved sensitivity of 0.001). However, at less strict cut-offs it is inferior to Dollo 1, giving lower specificity for a given sensitivity.

The ML 1 method achieves greater specificity than the method of phylogenetic profiles, peaking at 0.14 (at sensitivity = 0.002). But ML 1 performs more poorly than the Dollo 1 and Dollo 2 methods. This is due to a poor match between gene content evolution and the ML 1 model. ML 1 specifies no limit on the number of independent gains a gene may have, even though we do not believe multiple gains of genes are likely among eukaryotes. The ML 2 model is more realistic because, to an extent, it represents this prior belief. ML 2 results are greatly superior to results of all other methods used. ML 2 even achieves specificity of 1, at sensitivity = 0.0007. Various less extreme cut-offs will be useful in practice due to their higher sensitivity, for example a sensitivity of 0.01 at specificity = 0.8 (Figure 1).

4. Acknowledgements

We acknowledge the financial support of the Biotechnology and Biological Sciences Research Council, UK (Grant G19848 to MP) and the University of Reading.

5. References

[1] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg and T.O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles", *Proceedings of the National Academy of Sciences of the USA* **96**: 4285–4288, 1999.

[2] J.S. Farris, "Phylogenetic analysis under Dollo's Law", *Systematic Zoology* **26**: 77–88, 1977.

[3] M. Pagel, "Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters", *Proceedings of the Royal Society of London Series B Biological Sciences* **255**: 37–45, 1994.

[4] D. Barker and M. Pagel, "Predicting functional gene links from phylogenetic-statistical analyses of whole genomes", *PLoS Computational Biology* **1**, in press, 2005.

[5] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Research* **25**: 3389–3402, 1997.

[6] M. Remm, C.E. Storm and E.L. Sonnhammer, "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons", *Journal of Molecular Biology* **14**: 1041–1052, 2001.

[7] U. Güldener, M. Münsterkötter, G. Kastenmüller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. García-Martínez, J. E. Pérez-Ortín et al., "CYGD: the Comprehensive Yeast Genome Database", *Nucleic Acids Research* **33** (database issue): D364–D368, 2005.