Simplified Models of Evolution Lead to Improved Prediction of Functional Linkage from Correlated Gain and Loss of Genes among Eukaryotes

D. Barker[1]*, A. Meade[2], M. Pagel[2]

[1]Sir Harold Mitchell Building, School of Biology, University of St Andrews, St Andrews, Fife, KY16 9TH, UK. [2]AMS Building, School of Biological Sciences, University of Reading, Whiteknights, Reading, RG6 6AJ, UK. *Corresponding author, email db60@st-andrews.ac.uk.

We have improved the accuracy of predictions of functional linkage from whole genomic gene content [1] by seeking not correlated presence and absence of genes, but rather correlated gains and losses of those genes on branches of a phylogenetic tree of species [2]. This may be modelled within a maximum likelihood (ML) framework [2,3]. As originally applied to gene content [2], this method involves estimation of rates of gene gain as well as rates of gene loss.

We here simplify the ML models, by not estimating the initial rates of gain of genes but fixing them to constant, low values. The motivation for this novel approach is to better model gene content evolution, by preventing the modelling of multiple gains of the same gene in different parts of the phylogeny. A suitable low rate of gene gain is estimated using an initial training step with a grid search for the optimal rate, judged by specificity and sensitivity of predictions according to known test data.

We compare our new method with the ML method of [2], the across-species method [1] and a method seeking correlated gain and loss of genes based on Dollo Parsimony. Dollo parsimony [4] provides a rapid, non-statistical method of reconstructing ancestral states on a phylogenetic tree. It allows zero or one gains of a trait, but any number of losses subject to the constraint that the total number of changes on the phylogenetic tree is minimized. The implied assumptions seem suitable for reconstructing gene content evolution in eukaryotes and it has been used for this purpose [5].

Using 21 species of fungi and animals, we test each method on large positive and negative test data derived from known protein complexes [6]. We compare the quality of methods according to sensitivity and specificity. We find that all three phylogenetic methods (using ML models or Dollo parsimony) give higher quality predictions than the across-species method. The ML approach with rates of gene gain constrained to a low value gives by far the best results. The specific value used for rates of gain is not critical, and values of half or double the optimum for the study are found to give results of almost equal quality. Thus the novel version of the ML phylogenetic method, using simpler ML models of gene gain/loss, will be able to give high quality predictions of functional linkage even for studies of species for which little training data is available.

[1] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A, 96, 4285-4288.
[2] Barker, D. and Pagel, M. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol 1: 24-31.
[3] Pagel, M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. Proc R Soc Lond B Biol Sci 255: 37-45.
[4] Farris, J.S. 1977. Phylogenetic analysis under Dollo's Law. Syst Zool 26: 77-88.
[5] Krylov, D.M., Wolf, Y.I., Rogozin, I.B. and Koonin, E.V. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res 13: 2229-2235.
[6] Güldener, U., Münsterkötter, M., Kastenmüller, G., Strack, N., van Helden, J. et al. 2005. CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Res 33: D364-D368.