Poster C-17

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

Authors:
Daniel Barker (*School of Biology, University of St Andrews*)
Andrew Meade (*School of Biological Sciences, University of Reading*)
Mark Pagel (*School of Biological Sciences, University of Reading*)

Short Abstract: Correlated gain and loss of genes from species' genomes may be used to predict functional linkage between gene products. We propose and test a novel variation on the phylogenetic maximum likelihood approach for seeking correlation among gain/loss of genes, tailored to eukaryotic gene content evolution. This improves results significantly.

Long Abstract:

One approach to predicting functional linkage is the across-species method of phylogenetic profiles (Pellegrini et al. 1999). If a gene coding for a protein vital to a structural complex or biochemical pathway is lost from a species' genome, the genes involved in the remainder of the pathway or complex may also soon be lost, leading to modularity in change in total gene content over evolutionary time (Aravind et al. 2000, Ettema et al. 2001). The across-species method (Pellegrini et al. 1999) takes a correlated pattern of presence and absence in genes across several genomes as evidence that the products of those genes are functionally linked.

We have improved accuracy of predictions be 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 (Barker & Pagel 2005). This may be modelled within a maximum likelihood (ML) framework, in which a model of correlated change is compared to a model of uncorrelated change by means of a likelihood ratio test (Pagel 1994; Barker & Pagel 2005). As originally applied to gene content (Barker & Pagel 2005), this method uses four parameters for the uncorrelated model, and eight for the correlated model. In the uncorrelated model each gene in the pair is gained and lost at a rate independent of the other, estimated from the species' phylogeny and the gene's distribution pattern across extant species. The correlated model uses eight parameters, representing for two genes (A and B) all possible transitions between the states "gene A present, gene B present", "gene A present, gene B absent", "gene A absent, gene B present" and "gene A absent, gene B absent" (excluding transitions of the state of both genes simultaneously in the interest of simplicity). The uncorrelated model is a special case of the correlated model in which certain parameters are forced to equal each other.

We here reduce the number of parameters of 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. Gain of the same gene by multiple

lineages is rare among eukaryotes, which undergo little horizontal gene transfer between species in nature. The decision as to which low value to use for rate of gene gain is made using an initial training step with a grid search for the optimal rate, judged by sensitivity and specificity according to known test data. The uncorrelated ML model now has two free parameters, representing independent rates of loss of the two genes. The correlated model now has six free parameters, since the rate of gain of each gene in the absence of the other is fixed.

We compare our new method with the ML method of Barker & Pagel (2005), the across-species method of Pellegrini et al. (1999) and an approach seeking correlated gain and loss of genes based on Dollo Parsimony. Dollo parsimony (Farris 1977) 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 have been used for this purpose (Krylov et al. 2003).

Using 21 species of fungi and animals and with gene presence/absence data inferred using Inparanoid (Remm et al. 2001), we test each method on large positive and negative test data derived from known protein complexes (Güldener et al. 2005). 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 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.

References

Aravind, L., Watanabe, H., Lipman, D.J. and Koonin, E.V. 2000. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc Nat Acad Sci U S A 97: 11319-11324.

Barker, D. and Pagel, M. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol 1: 24-31.

Ettema, T., van der Oost, J. and Huynen, M. 2001. Modularity in the gain and loss of genes: Applications for function prediction. Trends Genet 17: 485-487.

Farris, J.S. 1997. Phylogenetic analysis under Dollo's Law. Syst Zool 26: 77-88, 1977.

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.

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.

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: 3745.

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.

Remm, M., Storm, C.E. and Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 14: 1041-1052.