BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 1 2007, pages 14–20
doi:10.1093/bioinformatics/btl558

Phylogenetics

Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes

Daniel Barker, Andrew Meade[1] and Mark Pagel[1,*]

Sir Harold Mitchell Building, School of Biology, University of St Andrews, St Andrews, Fife, KY16 9TH, UK and [1]School of Biological Sciences, University of Reading, Whiteknights, Reading, RG6 6AJ, UK

Received on June 13, 2006; revised on September 28, 2006; accepted on October 30, 2006
Advance Access publication November 7, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Motivation: We compare phylogenetic approaches for inferring functional gene links. The approaches detect independent instances of the correlated gain and loss of pairs of genes from species' genomes. We investigate the effect on results of basing evidence of correlations on two phylogenetic approaches, Dollo parsminony and maximum likelihood (ML). We further examine the effect of constraining the ML model by fixing the rate of gene gain at a low value, rather than estimating it from the data.

Results: We detect correlated evolution among a test set of pairs of yeast (*Saccharomyces cerevisiae*) genes, with a case study of 21 eukaryotic genomes and test data derived from known yeast protein complexes. If the rate at which genes are gained is constrained to be low, ML achieves by far the best results at detecting known functional links. The model then has fewer parameters but it is more realistic by preventing genes from being gained more than once.

Availability: BayesTraits by M. Pagel and A. Meade, and a script to configure and repeatedly launch it by D. Barker and M. Pagel, are available at http://www.evolution.reading.ac.uk

Contact: m.pagel@rdg.ac.uk

Supplementary information: Supplementary Data are available at *Bioinformatics* online.

1 INTRODUCTION

An established computational approach for predicting functional links is the across-species method of phylogenetic profiles (Pellegrini *et al*., 1999). If the genes coding for part of a pathway or structural complex are lost from a species' genome, we might expect that the genes to make the remainder of the proteins involved might also soon be lost, leading to modularity in gain and loss of genes over evolutionary time (Ettema *et al*., 2001). On this assumption, the across-species method of phylogenetic profiles takes a correlated pattern of presence and absence in genes across several genomes as evidence that the products of those genes are functionally linked.

However, species' genomes may have similar gene content for the historical reason of being closely related, rather than as a result of adaptive evolution. When comparing features of different species, a truly phylogenetic approach allows for the historical influence of phylogenetic relationships (Ridley, 1983; Felsenstein, 1985; Harvey and Pagel, 1991). For prediction of functional links among gene products, this may be achieved by seeking not simple correlated presence and absence of genes, but instead considering the effect of the species phylogeny (Vert, 2002; Barker and Pagel, 2005; Zhou *et al*., 2006).

We have shown that seeking correlated gains and losses of genes on a phylogenetic tree of species substantially improves the detection of functionally linked pairs of proteins (Barker and Pagel, 2005), compared to the original across-species method (Pellegrini *et al*., 1999). We here compare the original across-species method (Pellegrini *et al*., 1999) with several phylogenetic methods. Two of the latter are based on Dollo parsimony (Farris, 1977). Two are based on maximum likelihood (ML) with a relatively general model (cf. Barker and Pagel, 2005), and another uses ML but with a constrained model in which the rate of gain of genes is not estimated from the data, but fixed to a low value. The motivation for the latter, novel approach was to model gene content evolution better, by preventing the modelling of multiple gains of the same gene in different parts of the phylogeny. A priori we believe such multiple gains to be extremely rare in eukaryotes, which do not undergo extensive horizontal gene transfer in nature.

We apply each method to a positive and negative test set, based on known protein complexes in yeast. We compare the quality of methods according to sensitivity and specificity. We find that all but one of the phylogenetic methods give higher quality predictions than the across-species method of phylogenetic profiles. Among phylogenetic methods, ML can achieve by far the most reliable results, but only if rates of gene gain are constrained to a low value.

2 METHODS

2.1 Species comparisons and relationships

The methods we investigate require accurate patterns of gene presence and absence across several species. This allows us to form a species-by-proteins matrix, which we refer to as the trait matrix, showing presence ('1') or absence ('0') of each species' ability to code for homologous proteins. The trait matrix was obtained bioinformatically, from species with relatively complete genome sequences. The phylogenetic approaches to seeking correlated gain and loss of genes additionally require a phylogenetic tree showing how the species are related to each other. We obtained the trait matrix and phylogeny as described in the Supplementary material.

To allow validation against the large amount of known data for yeast, we focus our study on fungi. We also include three animals, and a plant (*Arabidopsis thaliana*) as outgroup for the phylogenetic tree, giving 22 species in all (Supplementary material). *A.thaliana* was included

*To whom correspondence should be addressed.

only to provide the root position of the phylogeny. We exclude it from all searches for functional links, owing to its very distant relationship to the other species.

2.2 Validation

Each method allows a range of numbers of predicted functional links by appropriate choice of a score cut-off, with prediction quality tending to decrease and number of predictions tending to increase as the cut-off becomes less stringent. We compare the methods at a range of cut-offs, from extremely liberal to extremely stringent (see Algorithm, below).

Where two proteins are truly functionally linked, we assume that, usually, both proteins will be found within the same cellular component or contribute to the same biological process (in the sense of Gene Ontology Consortium, 2006). They may also form part of the same structural complex. Methods of assessing quality of predicted functional links have relied on, for example, shared SwissProt keywords for the proteins (Pellegrini *et al*., 1999), shared subcellular compartment or functional category (von Mering *et al*., 2002; Lu *et al*., 2003), or shared complex membership (von Mering *et al*., 2002; Barker and Pagel, 2005). These methods of validation are necessarily correlated with each other. For a simple, conservative assessment of prediction quality, we here use data on complex membership.

Our positive test set consists of 9178 pairs of functionally linked proteins, derived from known yeast complexes in the Comprehensive Yeast Genome Database (Güldner *et al*., 2005). Our negative test set consists of 441 217 pairs of proteins that are unlikely to be functionally linked (Supplementary material). We judge quality of predictions for a method as specificity and sensitivity (cf. von Mering *et al*., 2003), where

$$\text{specificity} = (\text{true positives})/(\text{true positives} + \text{false positives}) \quad (1)$$

$$\text{sensitivity} = (\text{true positives})/(\text{true positives} + \text{false negatives}). \quad (2)$$

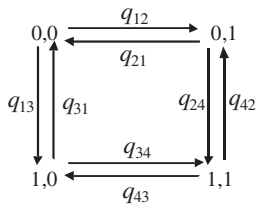
Sensitivity and specificity have a theoretical range of 0 to 1. High specificity suggests the predicted functional links are likely to be correct. High sensitivity suggests the method is able to find a large proportion of functional links so that, when a link is not predicted, it is likely to genuinely not exist. Overall, we desire both high specificity and high sensitivity.

3 ALGORITHM

For each pair of proteins in the positive and negative test sets we sought correlated presence and absence for the across-species method, and correlated gain and loss for the phylogenetic methods. For the across-species method, the first score we used was the number of species that had a matching state for the two proteins, i.e. either both absent (‘0’) or both present (‘1’) (Pellegrini *et al*., 1999). We refer to this method as ‘P99’. Two proteins both present in yeast but with otherwise entirely dissimilar distribution patterns would have a P99 score of 1, and two proteins with an identical cross-species distribution pattern have a score equal to the number of species in the study, here 21. Following the implementation of the across-species method in Barker and Pagel (2005), we also used the Fisher exact test (e.g. Zar, 1996), which we refer to as ‘Fisher’. This provides a *P*-value for the association between binary strings.

For the first two phylogenetic methods, we used Dollo parsimony (Farris, 1977) as an appropriate simple way to reconstruct ancestral distribution patterns of gene presence (Krylov *et al*., 2003; McLysaght *et al*., 2003; Koonin *et al*., 2004; see also Aravind *et al*., 2000). Dollo parsimony maps the trait (here, gene presence or absence) onto the phylogenetic tree with the minimum number of gains and losses, under the constraint that there must be zero or one gains. The number of losses is not constrained, but the overall number of changes (gain + losses) is minimized. The scores we derived from this were, first, the number of branches of the tree on which change occurred in a positively correlated manner (i.e. the number of branches on which both genes were gained together, plus the number of branches on which both were lost together), and second, this value minus the number of branches of the tree on which non-correlated changes occur. We refer to these as ‘Dollo-pos’ and ‘Dollo-overall’, respectively. Dollo-pos seek only positive evidence of correlated evolution. Dollo-overall considers both positive and negative evidence. In the current study we found the Dollo-pos score to range from a minimum of 0 to a maximum of 10, and the Dollo-overall score to range from −18 to 7.

For a more sophisticated phylogenetic approach, we also evaluated ML methods for detecting correlated evolution (Pagel, 1994,1997,1999). A brief summary of these ML methods follows. For a more complete description of the methods in this context, see Barker and Pagel (2005); see also Pagel and Meade (2006) for a Bayesian description of the Pagel (1994) model. Two genes can exhibit four different patterns of presence and absence in each species, with each gene individually either being present (‘1’) or absent (‘0’) from that species’ genome. The diagram in Equation 3 links the four states by arrows with parameters that describe the rates of transition between the two states of one of the genes, while the state of the other is constant.



(3)

If two genes are gained and lost independently of one another then the rates of change between the presence and absence of one gene will not depend upon whether the other is present or absent. For example, if the rate of gain of the second gene does not depend upon the state of the first, then $q_{12} = q_{34}$. More generally, the model of independent (uncorrelated) evolution implies that $q_{13} = q_{24}$, $q_{42} = q_{31}$, $q_{43} = q_{21}$ and $q_{12} = q_{34}$, and therefore requires a maximum of four parameters. The most general model of dependent (correlated) evolution does not imply these restrictions, and uses a maximum of eight parameters to describe the data.

The dependent model will improve on the independent model when the distribution of the genes across the species of the phylogeny implies that some of the pairs of transition rates constrained in the independent model to be equal to each other, in fact differ.

The method is formally described by a rate matrix $\mathbf{Q}$:

$$\mathbf{Q}_{I,D} = \begin{array}{c} \\ 0,0 \\ 0,1 \\ 1,0 \\ 1,1 \end{array} \begin{array}{c} \begin{array}{cccc} 0,0 & 0,1 & 1,0 & 1,1 \end{array} \\ \begin{bmatrix} - & q_{12} & q_{13} & 0 \\ q_{21} & - & 0 & q_{24} \\ q_{31} & 0 & - & q_{34} \\ 0 & q_{42} & q_{43} & - \end{bmatrix} \end{array}, \quad (4)$$

where we use the $\mathbf{Q}_{I,D}$ notation to indicate that the matrix can be configured to either the independent or dependent model depending upon whether some pairs of transition rates are constrained to be equivalent. The main diagonal elements are defined as minus the sum of the other rate coefficients in the row of the matrix, such that each row sums to zero. The values of all dual transitions, or cases in

which the states for both genes change simultaneously, are set to zero in the matrix in Equation 4. These can be modeled appropriately as two separate transitions. In the independent model, for convenience we may represent the 'gain' parameters as α_1 and α_2 and the 'loss' parameters as β_1 and β_2, where $\alpha_1 = q_{13} = q_{24}$, $\alpha_2 = q_{12} = q_{34}$, $\beta_1 = q_{42} = q_{31}$ and $\beta_2 = q_{43} = q_{21}$.

In contrast to Dollo parsimony, the ML approach as we implement it accounts for branch length, regarding a change as more probable on a long branch of the phylogeny than on a short one. Hence uncorrelated change on a short branch is regarded as weaker negative evidence than uncorrelated change on a long branch. Another advantage of ML is that the likelihood of each model (independent and dependent) is summed over all possible ancestral state reconstructions on the tree (Felsenstein, 1981; Pagel, 1994), whereas Dollo parsimony forces a single state at each internal node, even if the chosen state is only weakly supported.

In its full unconstrained form illustrated above, the ML model allows a gene to arise more than once on the tree. We refer to this model as 'ML-unconstrained'. In reality we expect that most correlated evolution takes the form of coincident losses of genes because gaining the same gene twice is improbable, especially among eukaryotes. On the other hand, given that homology is assessed using pairwise sequence similarity (Supplementary material), protein domain rearrangements or small amounts of convergent evolution could make two initially dissimilar sequences become more similar, and thereby appear as orthologues (Barker and Pagel, 2005). To slightly reduce the effect of allowing multiple gains in the model, one may fix the state of the root of the 21-species phylogeny at '1' for any gene found on both sides of the major bifurcation between animals and fungi. This causes the model to favour losses for those pairs, and is the approach used by Barker and Pagel (2005). We refer to it as 'ML-root'.

We also now evaluate a different ML model, modified by our prior knowledge of mechanisms of eukaryotic gene content evolution. We refer to this as 'ML-constrained'. It is the same as ML-unconstrained except that the initial gain parameters are not estimated by ML, but fixed a priori at a single low value r, where $q_{12} = q_{13} = r$ for the dependent model and $\alpha_1 = \alpha_2 = r$ for the independent model. This imposes a more Dollo-like constraint in that the algorithm will now tend to reconstruct multiple gene losses rather than multiple gains, with the strength of the tendency depending on the value of r. As an objective way to set the value of r, for any study we propose that analyses of known test data are run with a range of different values of r, and the specificity and sensitivity of each of these are calculated. This allows choice of r based on its demonstrated sensitivity and specificity. On the basis of initial tests to discover a range of values of r likely to include the optimum (data not shown), in the current study we investigate values for r in the range $0.1 \leq r \leq 6.0$, specifically $r = 0.1$, 0.2, 0.4, 0.8, 1.5, 3.0 and 6.0. To investigate the effect of errors in setting r, we record sensitivity and specificity at values of r deviating from the optimum.

r and all rates in $\mathbf{Q}_{I,D}$ have, as units, the reciprocal of the units of branch length in the species phylogeny. One value of r is unlikely to perform well across studies. The units of branch length in the phylogenetic tree will affect choice of r. Tree topology, relative branch lengths and the actual rate of gain for the genes in the study will also cause variation. We provide a script, *bms_runner*, which assists the user in discovering an optimal r for a given phylogeny and user-supplied training data. This script shows sensitivity and specificity at various score (likelihood ratio) cut-offs for each of a range of values of r. From this, a value of r and a cut-off may be chosen, which give sensitivity and specificity considered appropriate by the user.

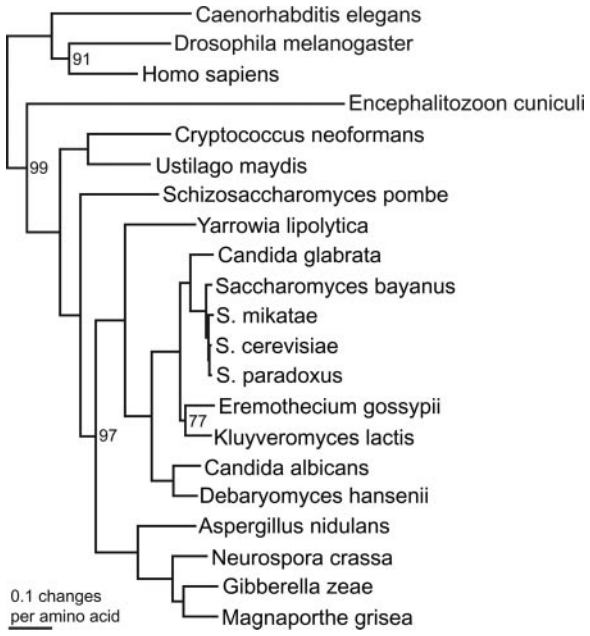


Fig. 1. Reconstructed phylogeny of the 21 ingroup species, from a concatenation of unambiguously alignable regions of 19 single-copy proteins (Supplementary Table S1). Bootstrap support is 100% for all nodes except where shown.

With ML-unconstrained and ML-root, the independent model has four parameters and the dependent model has eight. With ML-constrained, the independent model has two parameters and the dependent model has six. With all the ML approaches, the strength of evidence of correlated evolution is expressed as a likelihood ratio statistic *LR* (Cox, 1962; Goldman, 1993), where

$$LR = -2\ln(H_0) - \ln(H_1). \quad (5)$$

H_0 is here the likelihood of the model of independent evolution and H_1 is the likelihood of the model of dependent evolution, with both at their ML values. *LR* is zero if the dependent model does not improve at all on the independent model, and rises with increasing evidence of correlated evolution. In the current study *LR* ranged from 0 to 25.90 for ML-unconstrained, from 0 to 25.94 for ML-root, and from 0 to 37.94 for ML-constrained with a value of r empirically determined as appropriate for the current study.

ML models were fitted using the program BayesTraits, launched repeatedly by means of the *bms_runner* Perl script. Details of the implementation are given in the Supplementary material.

Negatively correlated distribution patterns are unlikely to indicate functional linkage and there is an argument for excluding those (Barker and Pagel, 2005). This separate pre-processing step is likely to be most useful for Fisher and ML approaches, and has no effect for the P99 approach. To assess the impact of negative correlations on results, we calculated Pearson correlations between distribution

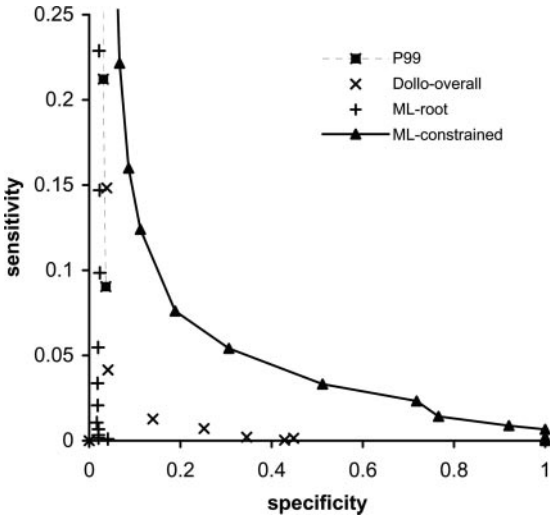


Fig. 2. Comparison of specificity and sensitivity for four of the seven methods investigated. The graph focuses on cut-offs giving sensitivities up to 0.25. Higher sensitivities are only possible at very low specificity, for any of the methods. The Fisher, Dollo-pos and ML-unconstrained methods gave results broadly similar to P99, Dollo-overall and ML-root, respectively, and are omitted for clarity. For details of all methods see Supplementary Table S3. For ML-constrained, the rate of gene gain, r, was set to 0.8.

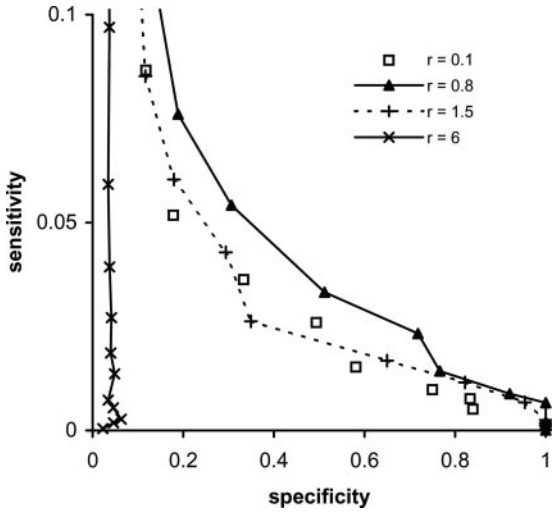


Fig. 3. Sensitivity and specificity for the ML approach with the rate of gene gain (r) fixed at various values, for various cut-offs. Cut-offs with sensitivities greater than 0.1 have relatively poor specificity for all r are omitted. $r = 0.8$ is appoximately optimal for the current study, although reduced amounts of training data led to choice of r in the range 0.8–1.5 (Supplementary Table S5). Some results have been omitted for clarity. For full results, see Supplementary Table S4.

patterns. In the interests of a straightforward comparison of underlying methods, without pre-or post-processing, we did not exclude such pairs from predictions.

4 RESULTS

4.1 Phylogeny

The ML phylogenetic tree of species is shown in Figure 1. Bootstrap support for most branches is high though there are two areas of ambiguity, including the relationship among the three groups of animals represented by *Homo*, *Drosophila* and *Caenorhabditis* (Aguinaldo *et al*., 1997; Wolf *et al*., 2004). Resolution of such long-standing ambiguities is not our current purpose. We have run our phylogenetic searches for correlated gain and loss over the single ML phylogenetic tree, rather than over a sample of trees. Any effect of ambiguity is limited to proteins whose distribution varies among the ambiguously placed species. Incorrect placement of species in the tree can only reduce the quality of phylogenetic predictions and cause us to judge the phylogenetic methods more harshly.

4.2 Predicted links and prediction quality

Even for those methods which could predict 'functional links' on the basis of negatively correlated cross-species distribution patterns of genes, such cases are few or absent among the more stringent score cut-offs (Supplementary Table S2). Inclusion of negatively correlated pairs does not affect our assessment of the relative quality of methods.

The sensitivities and specificities achieved with each method, according to the positive and negative test sets, are shown in Figure 2 and Supplementary Table S3. As expected, each method gives a range of results, depending on the score cut-off used. Within a given method, stricter cut-offs tend to give fewer predictions, lower sensitivity and higher specificity (Fig. 2; Supplementary Tables S2 and S3). Due to incompleteness of the reference set, values for specificity and sensitivity are approximate. However, they allow us to assess the relative merits of the methods, for some of which these statistics differ widely.

The ML-constrained method is clearly superior to any of the other methods used, giving higher specificity for a given sensitivity, and even achieving the theoretical maximum specificity of 1 at the most stringent cut-offs (though with low sensitivities, of 0.0066 or less). At more moderate, but still stringent cut-offs, the approach produces relatively high numbers of relatively high quality predictions (Fig. 2; Supplementary Tables S2 and S3).

The ML-constrained approach relies on esimtation of an appropriate rate of gene gain r, on the basis of known positive and negative data. The approach is robust to the exact value of r. We assessed sensitivity and specificity at seven values for r, ranging from 0.1 to 6. Of these, $r = 0.8$ gave best results in the current study. This generally led to moderate fitted rates of loss, with the overall median of mean (β_1, β_2) being 0.78 ($n = 450\,395$, interquartile range = 1.64). Results are still superior to those of any other method even when r deviates considerably from this optimum (Fig. 3; Supplementary Table S4). A good estimate of r is possible even with relatively small amounts of training data (Supplementary Table S5), making the approach suitable for non-model organisms where little known data exists for training.

The across-species methods, which do not use a phylogenetic tree, predict with poor discrimination. P99 achieves a maximum specificity of only 0.036 (at sensitivity = 0.090). Fisher achieves a maximum specificity of only 0.032 (at sensitivity = 0.18). Both these across-species approaches predict large numbers of functional links, giving high sensitivity (Fig. 2; Supplementary Table S3). But crucially, they fail to distinguish false positives from true positives sufficiently, even at the most stringent cut-offs available. Specificity

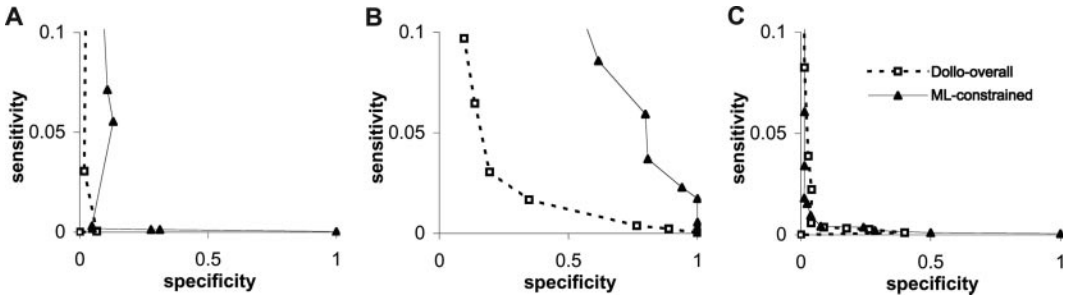


Fig. 4. Comparison of specificity and sensitivity for Dollo-overall (white squares) and ML-constrained (black triangles), stratified by rate of gene loss among the pairs compared. (**A**) Low rate of gene loss, (**B**) moderate rate of loss, (**C**) high rate of gene loss. 'Low' is arbitrarily defined to include the 163 949 pairs in the test data where mean $(\beta_1, \beta_2) < 0.5$, 'moderate' to include the 148 926 pairs where $0.5 \leq$ mean $(\beta_1, \beta_2) < 1.5$, and 'high' to include the 137 520 pairs where mean $(\beta_1, \beta_2) \geq 1.5$. Values of β_1 and β_2 were estimated by the ML-constrained method with $r = 0.8$.

for these across-species methods is so low in the current study that we do not regard the predicted functional links as useful.

The Dollo-pos method is more accurate than the across-species methods. At moderate to strict cut-offs, it tends to predict fewer functional links, but they are far more likely to be correct (Supplementary Table S3). In a sense it 'takes up where the across-species method leaves off', allowing specificity of up to 0.25 (at sensitivity = 0.00011). The proportion of true positives tends to increase with the number of correlated events. But the suggestion of Barker and Pagel (2005), that at least two or three correlated events of gain or loss almost certainly indicate functional linkage, does not seem to be a general rule.

The Dollo-overall method achieves still higher specificity, with less sacrifice of sensitivity. Its specificity peaks at 0.45 (at sensitivity 0.0014). However, at less strict cut-offs it is inferior to Dollo-pos, giving lower specificity for a given sensitivity.

The ML-unconstrained method performs worst out of the methods examined, peaking at specificity = 0.021 (at sensitivity = 0.17). The ML-root method achieves greater specificity than ML-unconstrained or the across-species methods, peaking at specificity = 0.041 (at sensitivity = 0.00098) (Fig. 2; Supplementary Table S3). However, ML-root performs more poorly than the Dollo and ML-constrained methods. The poor performance of ML-unconstrained and ML-root is a result of the poor match between their models, which specify no limit on the number of independent gains a gene may have, and the reality of eukaryotic gene content evolution, in which multiple gains of the same gene in different species are very unusual. Horizontal gene transfer has contributed to the yeast genome, but only rarely (Hall *et al.*, 2005), and convergent sequence evolution may be unlikely on a large scale. We used ML-unconstrained and ML-root not out of a desire to model these rare situations, but because trait matrix could in fact give the impression of multiple gains through discretizing sequence similarity to a binary measure of 'presence' versus 'absence' (see 'Algorithm', above). It appears such problems are not severe in practice, and, for the current broad study, the trait matrix can be taken at face value. The minor adjustment in ML-root (fixing of the root's ancestral state to '1' where appropriate) is inadequate to compensate for an unconstrained rate of gain of genes.

To further characterize the relative performance of Dollo-overall and ML-constrained, we stratify the quality of results by rate of gene loss (Fig. 4). Each method performs best with a 'moderate' rate of gene loss. Within any of the three rate categories ('low', 'moderate' or 'high'), ML-constrained tends to give superior results to Dollo-overall. ML-constrained can achieve a specificity of 1 in each category. Dollo-overall only achieves a specificity of 1 in the 'moderate' category, but even here performs considerably worse than ML-constrained at most score cut-offs (Fig. 4). It is clear that rate of gene loss affects quality of predictions for both methods, but ML-constrained gives superior results to Dollo-overall within any rate category. This is perhaps expected, given the body of evidence suggesting ML methods in phylogeny reconstruction give superior results to parsimony methods in most circumstances (e.g. see Felsenstein, 2004).

The trait matrix and full results for the ML-constrained method are given in the Supplementary material.

5 DISCUSSION AND CONCLUSION

5.1 Quality of results

The current analysis demonstrates the importance of using a truly phylogenetic approach when predicting functional linkage among proteins from the cross-species pattern of gene presence and absence. All of the phylogenetic methods except ML-unconstrained achieved higher specificity than the across-species approach. Among phylogenetic methods, two things were found to be crucial. First, the phylogenetic model must approximate biological reality reasonably closely. In the current study, this means the rate of gain of genes must be constrained to be low, even though this makes the model less accurate from a purely numerical, descriptive point of view. Second, a ML model is capable of greater accuracy and sensitivity than a Dollo parsimony-based approach. An appropriate model within a ML framework gives by far the best results. This appropriately weights gains and losses of genes, and allows correct modelling of the effect of different branch lengths in the phylogeny.

Of the methods examined, the relatively unsuccessful ones were incapable of achieving high specificity at any score cut-off. For the relatively successful methods, including the best-performing method ML-constrained, high specificity was only achieved at low sensitivity. In other words, where accurate predictions of functional linkage are required, only few predictions are possible. This is partly a limitation of the methods. For example, none of the ML or Dollo methods make any predictions for proteins present across all species in the study. (In constrast, the P99 method predicts

that all such proteins are functionally linked to each other.) However, the number and the quality of predictions are expected to increase as further genomes are included in the study (cf. Sun *et al*., 2005).

The number of species we have used reflects the state of fungal genome sequencing and annotation at the start of the current work, but is still sufficient to predict 298 functional links out of the 450 395 pairs of proteins we examined, at a specificity of 0.72 (Supplementary Tables S2 and S3). Such rates of prediction and accuracy are already at a level where one may begin to draw global inferences about the nature of molecular and biochemical organization within the cell.

5.2 Extensions

To allow validation with large amounts of reliable known data, we here confine ourselves to protein-coding genes, though in principle our approach would work with all genes including RNA genes, and indeed any genomic elements including promoters, enhancers, introns, UTR motifs, conserved DNA of unknown function and protein domains and motifs (cf. Pagel *et al*., 2004), either per-class or, perhaps most interestingly, together in the same study. This could allow functional annotation of some conserved regions of noncoding DNA that are increasingly being revealed by comparative genomics (Bejerano *et al*., 2004; Sabarinadh *et al*., 2004).

Our validation of phylogeny, particularly appropriate ML models of trait evolution in analysis of gene presence and absence suggests that phylogeny should also be investigated within several other computational comparative genomics methods for predicting protein function and functional linkage, e.g. gene fusion (Enright *et al*., 1999; Marcotte *et al*., 1999a), predicted operons (Dandekar *et al*., 1998; Overbeek *et al*., 1999), bidirectionally transcribed gene pairs (Korbel *et al*., 2004) and negatively correlated cross-species distribution patterns (Morett *et al*., 2003). None of these methods models change on a phylogenetic tree, but all could be modified to do so. Incorporation of the species phylogeny when predicting physical interaction partners by means of correlated sequence evolution has been found to improve results (Pazos *et al*., 2005; see also Akmaev *et al*., 2000), demonstrating the value of phylogeny extends beyond our current application.

Bayesian Markov chain Monte Carlo (MCMC) approaches, in which we fit the model of trait evolution to more than one tree (Pagel *et al*., 2004; Pagel and Meade, 2005,2006) may improve quality of results compared to ML, especially where there is ambiguity in the phylogeny. However in the current case study, we found comparatively little uncertainty in the tree topology (Fig. 1). Because of this we have pursued our investigation of models within an ML framework. Since Bayesian-MCMC trait reconstruction uses the same likelihood models as ML, our conclusions concerning appropriate models will generalize to both approaches.

Any bioinformatic or laboratory method is limited to finding only those functional links, which match the assumptions of the method. The most complete and accurate results will be obtained using several different methods together (Marcotte *et al*., 1999b; Hishigaki *et al*., 2001; von Mering *et al*., 2002,2003), for example as input to a Bayesian network (Jansen *et al*., 2003; Lin *et al*., 2004) or logistic regression (Lin *et al*., 2004), but it remains important that the quality of the inputs to the combining algorithm be as high as possible. Phylogeny can clearly assist in this area.

ACKNOWLEDGEMENTS

The authors thank Valerie Wood, Mukund Unavane and Riccardo Percudani for discussion and advice, and two anonymous referees for their helpful comments. Computer clusters used for analyses were maintained by the School of Systems Engineering at the University of Reading and by the School of Mathematics and Statistics at the University of St Andrews. The authors acknowledge the financial support of the Biotechnology and Biological Sciences Research Council, UK (Grant G19848 to M.P.) and a Research Councils UK Academic Fellowship to D.B. Preliminary results were presented as a poster at the Computational Systems Bioinformatics Conference, Stanford, August 2005, as a poster at the European Fission Yeast Meeting, Hinxton, March 2006, and as a poster and an oral presentation at the International Conference on Intelligent Systems for Molecular Biology, Fortaleza, August 2006.

Conflict of Interest: none declared.

REFERENCES

Aguinaldo,A.M. *et al*. (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, **387**, 489–493.

Akmaev,V.R. *et al*. (2000) Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics*, **16**, 501–512.

Aravind,L. *et al*. (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl Acad. Sci. USA*, **97**, 11319–11324.

Barker,D. and Pagel,M. (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput. Biol*., **1**, e3.

Bejerano,G. *et al*. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.

Cox,D.R. (1962) Further results on tests of separate families of hypotheses. *J. R Stat. Soc. Series B (Methodological)*, **24**, 406–424.

Dandekar,T. *et al*. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci*., **23**, 324–328.

Enright,A.J. *et al*. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Ettema,T. *et al*. (2001) Modularity in the gain and loss of genes: applications for function prediction. *Trends Genet*., **17**, 485–487.

Farris,J.S. (1977) Phylogenetic analysis under Dollo's Law. *Syst. Zoology*, **26**, 77–88.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*., **17**, 368–376.

Felsenstein,J. (1985) Phylogenies and the comparative method. *Am. Nat*., **125**, 1–15.

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*., **34**, D322–D326.

Goldman,N. (1993) Statistical tests of models of DNA substitution. *J. Mol. Evol*., **36**, 182–198.

Güldener,U. *et al*. (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res*., **33**, D364–D368.

Hall,C. *et al*. (2005) Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot. Cell*, **4**, 1102–1115.

Harvey,P.H. and Pagel,M.D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.

Hishigaki,H. *et al*. (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, **18**, 523–531.

Jansen,R. *et al*. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, **302**, 449–453.

Koonin,E.V. *et al*. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol*., **5**, R7.

Korbel,J.O. *et al*. (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol*., **22**, 911–917.

Krylov,D.M. *et al*. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res*., **13**, 2229–2235.

Lin,N. *et al*. (2004) Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, **5**, 154.

Lu,L. *et al*. (2003) Multimeric threading-based prediction of protein–protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res*., **13**, 1146–1154.

Marcotte,E.M. *et al*. (1999a) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

Marcotte,E.M. *et al*. (1999b) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

McLysaght,A. *et al*. (2003) Extensive gene gain associated with adaptive evolution of poxviruses. *Proc. Natl Acad. Sci. USA*, **26**, 15655–15660.

von Mering,C. *et al*. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.

von Mering,C. *et al*. (2003) Genome evolution reveals biochemical networks and functional modules. *Proc. Natl Acad. Sci. USA*, **100**, 15428–15433.

Morett,E. *et al*. (2003) Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat. Biotechnol*., **21**, 790–795.

Overbeek,R. *et al*. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Pagel,M. (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R Soc. Lon. B Biol. Sci*., **255**, 37–45.

Pagel,M. (1997) Inferring evolutionary processes from phylogenies. *Zool. Scr*., **26**, 331–348.

Pagel,M. (1999) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol*., **48**, 612–622.

Pagel,M. *et al*. (2004) Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol*., **53**, 673–684.

Pagel,M. and Meade,A. (2005) Bayesian estimation of correlated evolution across cultures: a case study of marriage systems and wealth transfer at marriage. In Mace,R., Holden,C.J. and Shennan,S. (eds), *The Evolution of Cultural Diversity: A Phylogenetic Approach*. University College London Press, London, pp. 235–256.

Pagel,M. and Meade,A. (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat*., **167**, 808–825.

Pagel,P. *et al*. (2004) A domain interaction map based on phylogenetic profiling. *J. Mol. Biol*., **344**, 1331–1346.

Pazos,F. *et al*. (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol*., **352**, 1002–1015.

Pellegrini,M. *et al*. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

Ridley,M. (1983) *The Explanation of Organic Diversity: The Comparative Method and Adaptations for Mating*. Clarendon Press of Oxford University Press, Oxford.

Sabarinadh,C. *et al*. (2004) Extreme conservation of noncoding DNA near HoxD complex of vertebrates. *BMC Genomics*, **5**, 75.

Sun,J. *et al*. (2005) Refined phylogenetic profiles method for predicting protein–protein interactions. *Bioinformatics*, **21**, 3409–3415.

Vert,J.-P. (2002) A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, **18**, S276–S284.

Wolf,Y.I. *et al*. (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis, *Genome Res*., **14**, 29–36.

Zar,J.H. (1996) *Biostatistical Analysis*. 3rd edn. Prentice Hall, Upper Saddle River, NJ.

Zhou,Y. *et al*. (2006) Inferring functional linkages between proteins from evolutionary scenarios. *J. Mol. Biol*., **359**, 1150–1159.