BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 5 2011, pages 700–706
doi:10.1093/bioinformatics/btq705

Systems biology

Advance Access publication December 17, 2010

A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees

Hung Xuan Ta[1,*], Patrik Koskinen[2] and Liisa Holm[1,2,*]

[1]Institute of Biotechnology and [2]Division of Genetics, Faculty of Biological and Environmental Sciences, PO Box 56, 00014 University of Helsinki, Helsinki, Finland

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Functional linkages implicate pairwise relationships between proteins that work together to implement biological tasks. During evolution, functionally linked proteins are likely to be preserved or eliminated across a range of genomes in a correlated fashion. Based on this hypothesis, phylogenetic profiling-based approaches try to detect pairs of protein families that show similar evolutionary patterns. Traditionally, the evolutionary pattern of a protein is encoded by either a binary profile of presence and absence of this protein across species or an occurrence profile that indicates the distribution of copies of this protein across species.

Results: In our study, we characterize each protein by its enhanced phylogenetic tree, a novel graphical model of the evolution of a protein family with explicitly marked by speciation and duplication events. By topological comparison between enhanced phylogenetic trees, we are able to detect the functionally associated protein pairs. Because the enhanced phylogenetic trees contain more evolutionary information of proteins, our method shows greater performance and discovers functional linkages among proteins more reliably compared with the conventional approaches.

Contact: xuanhung.ta@helsinki.fi; liisa.holm@helsinki.fi

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

Received on September 24, 2010; revised on November 19, 2010; accepted on December 14, 2010

1 INTRODUCTION

A great number of cellular behaviors are mediated by proteins, which always carry out their functions by interacting with each other (Eisenberg *et al.*, 2000). Proteins that participate in a common structural complex or pathway or biological process are defined as functionally linked. Defining functional linkages of proteins is one of the central goals in proteomics that will decipher the molecular mechanisms underlying the biological functions and, then, help to enhance approaches for drug discovery. The complete sequencing of multiple genomes from diverse species provides an excellent opportunity to develop comparative approaches for functional studies in proteomics. Biological interactions and functional linkages of proteins can be inferred via various patterns across many genomes. These include the co-localization of genes on chromosomes (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999), the genetical fusion of two distinct proteins from one organism into a single protein in another organism (Enright *et al.*, 1999; Marcotte *et al.*, 1999), protein domain composition (Sprinzak and Margalit, 2001; Ta and Holm, 2009) and phylogenetic profiles (Dutkowski and Tiuryn, 2009; Marcotte *et al.*, 2000; Pellegrini *et al.*, 1999; Wu *et al.*, 2003).

Phylogeny-based methods for protein functional relationship prediction are broadly applicable due to a sufficiently large number of completely sequenced genomes. These methods are premised on the hypothesis that functionally linked proteins evolve in a correlated fashion, and therefore they have homologs in the same set of organisms (Pellegrini *et al.*, 1999). The original approach uses the simplest form of phylogenetic profile of a protein, a binary string in which each bit indicates the presence or absence of a protein family in a different species (Pellegrini *et al.*, 1999). This basic method, often referred to 'phylogenetic binary profiling', subsequently searches for matching pairs of profiles which differ from each other by less than three bits. In further steps, profiles can be compared by statistical correlation measures, such as co-occurrence probability (Wu *et al.*, 2003), the Pearson correlation coefficient (Glazko and Mushegian, 2004), Fisher's exact test (Barker and Pagel, 2005) or mutual information (Date and Marcotte, 2003; Huynen *et al.*, 2000).

Recently, Ranea *et al.* (2007) introduced phylogenetic occurrence profiling to detect the functionally related protein families in eukaryotic genomes. The phylogenetic occurrence profile of a protein family is a vector in which each element indicates the number of protein members of this family observed in one organism. Unlike prokaryotic genomes, where a high proportion of protein families have about one copy per species, eukaryotic genomes show a large number of multiprotein families having more than one member per species (Ranea *et al.*, 2007). Phylogenetic occurrence profiling, therefore, is able to detect more evolutionary signals that could not be detected by phylogenetic binary profiling.

However, binary and occurrence profiles cannot adequately describe the whole evolutionary history of proteins. Many methods characterize proteins by their reconstructed phylogenetic trees. These approaches, then, compare the phylogenetic trees by utilizing the parsimony principle (Barker *et al.*, 2007), maximum likelihood model (Barker and Pagel, 2005), a tree-kernel model (Vert, 2002) or the correlation between the distances matrices used to build the trees (Pazos and Valencia, 2001).

Evolutionary divergence, convergence and horizontal transfer events in multiprotein families are more complex than can be described by conventional phylogenetic profiles. In this work, we introduce a novel graphical model of the evolution of proteins, which enhances the information content of phylogenetic profiles

*To whom correspondence should be addressed.

by accounting for the reconstructed proteomes of ancestral species and synchronous gene duplication events. While orthologs evolved by vertical descent (speciation) from a single protein in the last common ancestor of the compared genomes, paralogs evolved by gene duplication. In this work, one-to-one correspondences between orthologs are detected using the reciprocal best hits criterion, and duplications are identified using the InParanoid criterion (O'Brien *et al.*, 2005). This so-called enhanced phylogenetic tree (EPT) explicitly traces all the descendants of proteins from the last universal common ancestor (LUCA) down to the extant species. This tree is different from the tree reconciliation (Zmasek and Eddy, 2002) in that protein data are directly fitter on the species tree, and speciation or duplication events are inferred using an exclusion principle to ensure strict one-to-one correspondences between proteins derived by speciation. The EPTs, subsequently, are topologically compared to find the evolutionarily correlated proteins families. Two proteins of a target organism belonging to two correlated families are likely to be functionally linked.

In this article, we benchmark the EPT method by using positive and negative reference sets of functional linkages in *Homo sapiens* and *Saccharomyces cerevisiae*. The biological features of the predicted functional linkages are examined by studying GO annotations in the biological process, cellular component and molecular function branches of Gene Ontology. Our novel method shows a significant improvement compared with conventional phylogenetic profiling methods and makes predictions that are complementary to other prediction methods.

2 METHODS

2.1 Benchmark datasets

Benchmark sets including positive and negative reference sets (PRS and NRS, respectively) are needed to calculate the receiver operating characteristic (ROC), which represents the performance of a classification method. By true positive (TP), we mean an interaction or linkage that is predicted by our method and present in the PRS. Analogously, a true negative (TN) is a pair of proteins predicted by our method not to interact that is present in the NRS. A false positive (FP) is a protein pair in the NRS that is predicted to interact by our method while a false negative (FN) is a protein pair in the PRS predicted not to interact by our method. In the plot of the ROC curve, the x-axis represents false positive rate (FPR) or 1−specificity, that is FP/(TN + FP), and the y-axis represents true positive rate (TPR) or sensitivity, TP/(TP + FN), as the threshold gradually varies.

To evaluate the performance of our methods at predicting protein functional linkages, we choose the pairs of proteins existing within the same complex as referent positives and the random pairs of proteins from different complexes, which are not connected in the interaction graph, as referent negatives. For yeast datasets, 5888 pairs of co-complex proteins and 5888 randomly chosen pairs were derived from the manually curated catalog in MIPS (Mewes *et al.*, 2004).

The benchmark datasets of human functional linkages are derived from the CORUM database, a resource of manually annotated protein complexes from mammalian organisms (Ruepp *et al.*, 2010). After filtering out co-complex protein pairs discovered by high-throughput experimental methods, we obtained 26 813 pairs of protein for the human PRS. The human NRS contains 26 813 pairs of proteins chosen randomly from different complexes.

2.2 Protein families

Phylogenetic profiling requires that proteins from different organisms are classified into families, which can be compared across species. We clustered the proteins in the proteomes of 572 complete genomes (560 prokaryotic and 12 eukaryotic organisms) based on their Basic Local Alignment Search Tool (BLAST) similarity. This study considered similarities above a bitscore of 50, which effectively filters FP hits from the sequence comparison (Remm *et al.*, 2001). We tested both lower and higher thresholds, which yielded worse performance in the benchmark. The clusters have an internal hierarchical structure that guarantees strict one-to-one correspondences between proteins from different species in each subfamily. The clustering algorithm is a hierarchical modification of the InParanoid algorithm (Remm *et al.*, 2001). The inputs are the species tree and the protein sequences of a set of organism. The algorithm outputs proteins classified into families with gene duplication events mapped to taxa in the species tree. The details of the clustering algorithm are shown in Figure 1.

There were 2 188 588 proteins in the proteomes. Of total, 254 418 singletons (about 11.6%) had no similarity to other proteins. The other proteins were clustered into 91 428 protein families (EPTs) having more than one member. The average number of leaf proteins in an EPT was 21.2 approximately. The distribution of EPT size is scale free, i.e. it follows a power law.

2.3 Enhanced phylogenetic trees

The strength of functional linkage between two protein families is assessed by comparing the topologies of the corresponding EPTs.

Figure 2 shows two examples of EPT, tree A and tree B. In tree A, four round-dot lines represent the proteomes of four organisms. Tree A (B) has four (two) subtrees f_1, f_2, f_3, f_4 (f_1', f_2'), forming in three (two) layers. Each subtree, representing a subfamily, is composed of the identical leaves, internal node and solid edges (speciation). They are separated by dashed edges (duplication). The corresponding occurrence and binary profiles of the family are shown below. Intuitively, the EPT contains more evolutionary signals than conventional profiling. The NCBI taxonomy tree is used in constructing EPTs. In a subtree, the dashed edges (empty nodes) indicate the edges (nodes) of the NCBI taxonomy tree that do not exist in the subtree. In tree A, the first layer contains the subtree f_1, the second layer contains f_2 and f_3, and the third layer contains f_4.

2.4 Subtree similarity

The topological similarity of two subtrees is computed guided by the NCBI taxonomy tree, meaning that the comparison is based on the taxonomy identifiers of nodes in two subtrees (see Fig. 3). Let T_i^A be a part of subtree A that roots from node i and $C(i,A)$ denote the set of children of node i in subtree A. The similarity of two subtrees A and B rooting from node i, which is denoted by $S(T_i^A, T_i^B)$, can be calculated as follows

$$S(T_i^A, T_i^B) = \alpha + \sum_{k \in C(i,A) \cap C(i,B)} S(T_k^A, T_k^B) - \beta |(C(i,A) - C(i,B)) \cap (C(i,B) - C(i,A))| \quad (1)$$

where $X - Y$ denotes the set of objects that belong to X and not to Y, $X \cap Y$ denotes the set of objects that belong to both set X and set Y and $|X|$ means the number of elements in set X. In Equation (1), $\alpha(\beta)$ is the reward (penalty) coefficient. From Equation (1), we have $S(T_i^A, T_j^B) = -\beta$ if $j \neq i$ or one of two trees is null. In words, each common node yields a reward α and each deletion of a subtree costs a penalty β. In the current implementation, we require that two subtrees have the same root node identifiers, that is, only subtrees that were created by gene duplications at the same taxonomic node contribute to the EPT score (the tree in the first layer is always rooted at LUCA).

2.5 Measuring the similarity of enhanced phylogenetic trees

Two EPTs are compared layer-by-layer (see an example in Fig. 2). Let T_{ijk}^A be subtree k in the i-th layer of EPT A whose root taxonomy identifier is j.

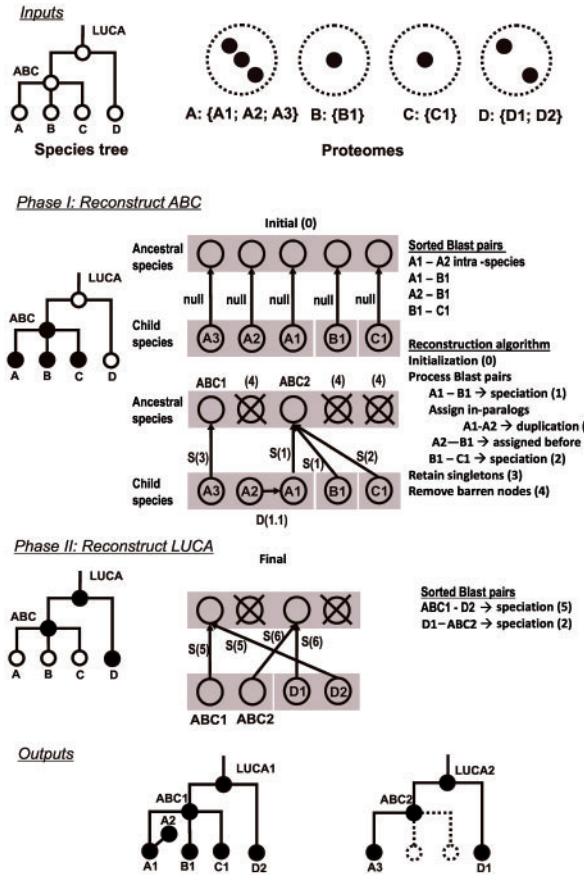

Fig. 1. Schematic example to show how the family definition algorithm works. *Phase I*: the proteomes of ancestral species are reconstructed hierarchically based on the species tree. Here, species ABC is reconstructed from child species A, B and C. Proteins are clustered based on all versus all Blast hits between the child species. The list of Blast hits is sorted by similarity score in decreasing order. Initially, each child protein has an image in the ancestor but the actual parent is yet unassigned. Child proteins that derive from the same ancestral protein are identified based on Blast hits with the highest similarity score. Only one protein from each child species may be assigned to the same ancestral protein. Recent duplications (generating in-paralogs within a child species) have higher similarity scores than BLAST hits to a sister species. Singleton proteins have no similarity to sister species; they are assumed to be present in the ancestor. Finally, the algorithm removes image nodes from the ancestor which are without progeny. *Phase II*: here, the root node (LUCA) is reconstructed from ABC and D. The similarity scores from reconstructed proteins are estimated as the maximum of the similarities of the cluster members; for example, ABC2-D1 is the maximum of A1-D1, B1-D1 and C1-D1. The final assignments after the reconstruction are shown. Following the edges from the leaf proteins to reconstructed LUCA proteins gives the protein families and lineages shown in *Outputs*.

Let t_i^A be the set of the root taxonomy identifiers of the subtrees in the i-th layer of EPT A. The number of layers of EPT A is denoted by ℓ_A. The similarity of two EPTs A and B, with an assumption that $\ell_A > \ell_B$, is calculated as follow

$$S_{AB} = \sum_{i=1}^{\ell_B} w_i \left(\sum_{j \in t_i^A \cap t_i^B} \left[\frac{1}{n_{ij}^A n_{ij}^B} \sum_{k=1}^{n_{ij}^A} \sum_{l=1}^{n_{ij}^B} S(T_{ijk}^A, T_{ijl}^B) \right] \right)$$

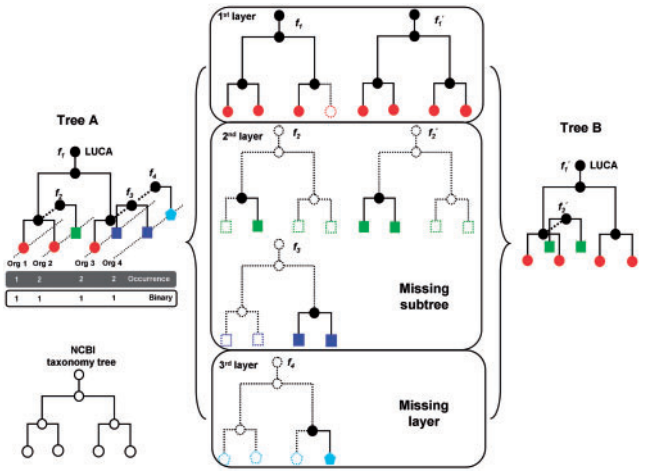

Fig. 2. Two examples of EPTs and their decompositions. Trees A and B are compared layer-by-layer. The missing layer in tree B is penalized [corresponding to the third term of Equation (2)]. Subtrees (f_1, f_1') in the first layer and (f_2, f_2') in the second layer are compared [corresponding to the first term of Equation (2)]. The missing subtree in the second layer of tree B is penalized [corresponding to the second term of Equation (2)].

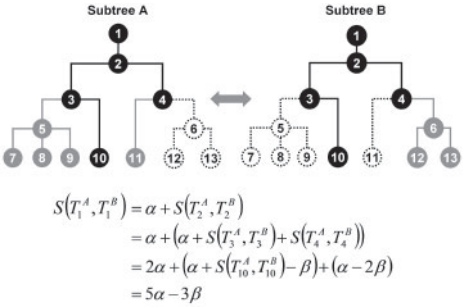

Fig. 3. The topological comparison of two trees is based on the taxonomy identifiers of the nodes in two trees. The numbers labeling the nodes in the trees are taxonomy identifiers. The similarity is recursively calculated by Equation 1. In this example, there are five common nodes (black) and three subtree deletions (dashed subtrees).

$$-\beta \left[\sum_{j \in (t_i^A - t_i^B)} n_{ij}^A + \sum_{j \in (t_i^B - t_i^A)} n_{ij}^B \right] \Bigg) + \sum_{i=\ell_B+1}^{\ell_A} w_i (-\beta n_i^A) \qquad (2)$$

where $S(T_1, T_2)$ notifies the similarity between two subtrees T_1, T_2 which is computed by Equation (1). In Equation (2, $n_{ij}^A = \left| \{T_{ijk}^A\}_{\forall k} \right|$ denotes the number of subtrees whose root taxonomy identifier is j in the i-th layer of EPT A and $n_i^A = \sum_{j \in t_i^A} n_{ij}^A$ is the number of all the subtrees in the i-th layer of EPT A. Therefore, the first term compares subtrees having the same root taxonomy identifiers, the second term is the penalty for subtrees in one layer of an EPT and not in the corresponding layer of the other EPT and the third term is the penalty for missing layers. In this study, we set $w_i = 1/i$, meaning that the layers are assigned weights depending on their lineage. Finally, the

scores are normalized as follows

$$\overline{S_{AB}} = \frac{\left(S_{AB} - \min\{S_{AB}\}_{A,B}\right)}{\left(\max\{S_{AB}\}_{A,B} - \min\{S_{AB}\}_{A,B}\right)}. \quad (3)$$

2.6 Predicting functional linkages of proteins

A protein is characterized by the EPT of the family to which it belongs. Two similar EPTs mean correlated patterns of evolution and, by implication, a functional linkage. We construct EPTs for all the families that contain more than one member and at least one of the members is a protein of the target organism. Two proteins whose EPTs have a higher similarity score than a given threshold are classified as functionally linked or physically interacting proteins.

2.7 Calculating GO semantic similarity

To investigate the biological features of predicted functionally linked protein pairs, we quantify the functional similarity for these pairs by computing the semantic similarity (SS) score (Wang *et al.*, 2007) of the GO terms with which these proteins are annotated. The semantic score takes into account both the level of hierarchy of GO terms and also their relations with their ancestor terms. All predicted protein pairs are assigned semantic similarity scores for three GO branches: molecular function (MF); biological process (BP) and cellular component (CC). Semantic scores in this study were calculated using GOSemSim package in Bioconductor (Gentleman *et al.*, 2004).

Functional linkages are predicted among all the proteins in yeast and human proteomes by all the competing methods (binary profiles, occurrence profiles and EPT). The mean GO semantic score is calculated over all the protein pairs in the top predictions to present the biological features of the predicted pairs. The method that shows higher mean GO-SS at the same number of top predictions has stronger biological implications.

3 RESULTS AND DISCUSSION

3.1 Evaluation of performance of the EPT method at predicting functional linkages

Benchmark datasets are tested by the enhanced phylogenetic tree method. We compare our method with the conventional methods using binary profile with Pearson correlation (Glazko and Mushegian, 2004) (called bin_ps) and occurrence profile (Ranea *et al.*, 2007) with Pearson correlation (called occ_ps) or Euclidean distance (called occ_ed) (see Supplementary Material). The binary and occurrence profiles are converted from the corresponding EPTs as described in Figure 2A. Figure 4 shows the assessment of the EPT method for different test sets including human (see Fig. 4A) and yeast (see Fig. 4B) PRS and NRS of functional linkages. It is clear that the EPT method can capture part of the co-evolutionary signal of protein functional linkages. In particular, by using a threshold of 0.245 (0.546), the EPT method discovers about 20% (27%) functional linkages in human (yeast) datasets with a low level of FPR (around 5%) (see Fig. 4 and Supplementary Fig. S2). For the top 1000 predictions, we obtained a precision of more than 90% both in the human and in the yeast datasets (see Insets in Fig. 4).

The EPT method shows a drastic improvement compared with traditional approaches using binary or occurrence profiles. In the top 1000 predictions of human datasets, the EPT method shows a precision of more than 90% while others show precisions less than 70%. For yeast datasets, the precision of the EPT method is about 90% while those of other methods are about 60%. The advance of the EPT method seems to be a direct consequence of using the EPT as an evolutionary pattern. It is clear that the similarity between two

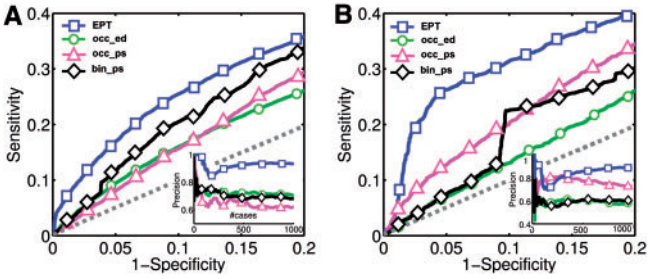

Fig. 4. ROC curves of the EPT, occ_ed, occ_ps and bin_ps methods for (**A**) human and (**B**) yeast datasets. The dashed gray diagonal lines are corresponding to random predictions. The ROC curves with full scales are presented in Supplementary Figure S1. The insets of (A) and (B) are the precisions of the methods at different number of top predictions (cases with highest scores) for human and yeast datasets, respectively. The precision is the ratio of the number of true positives over the number of top predictions, TP/(TP + FP).

EPTs reflects the co-evolution of the pair of proteins belonging to these EPTs.

Clearly, the EPT method helps to reduce FPs by other methods using binary and occurrence profiles (see Supplementary Fig. S3). In the top 1000 human predictions by the bin_ps method, there are 343 FPs (>34%). The number of FPs in the top 1000 predictions by the occ_ed and occ_ps methods are 289 (≈29%) and 378 (≈38%), respectively. Nearly the FPs by the traditional methods are out of the top 1000 predictions by the EPT method, meaning that they are likely identified as TNs by the EPT methods. For yeast datasets, most of the FPs in the top 1000 predictions by the traditional methods are also identified as TNs by the EPT method. Namely, the EPT method ranks 404 among 418 FPs (≈97%) by the bin_ps method out of the top 1000 predictions. Our method also identified 420 /425 FPs (259/274 FPs) by the occ_ed (occ_ps) method as TNs.

Two examples of potential FPs by the traditional methods that are identified as TNs by the EPT methods are presented in Figure 5A for human datasets and Figure 5B for yeast datasets. The upper and lower panels of Figure 5A, respectively, show the EPTs, occurrence and binary profiles of negative elongation factor A protein (Q9H3P2, NELFA_HUMAN) and syntaxin-12 protein (Q86Y82, STX12_HUMAN), which belong to different complexes (Ruepp *et al.*, 2010). They seem to participate in different biological processes, implement different molecular functions and locate in different cellular components (Ashburner *et al.*, 2000). iHOP shows no evidence for functional associations or physical interactions between these two proteins (Hoffmann and Valencia, 2004). NELFA_HUMAN and STX12_HUMAN share the identical binary profiles and very similar occurrence profiles, making them to be in the top predictions by traditional methods. However, their EPTs are topologically different. There are about 55% of all the protein pairs that have higher EPT scores than this pair does, meaning that NELFA_HUMAN and STX12_HUMAN are identified as unlikely to be functionally linked by the EPT method. Figure 5B shows a protein pair, N-terminal acetyltransferase A complex subunit protein (P12945, NAT1_YEAST) and Alpha-soluble NSF attachment protein (P32602, SEC17_YEAST), in yeast that is a TN by the EPT method but a FP by traditional methods using binary and occurrence profiles. These above examples indicate that

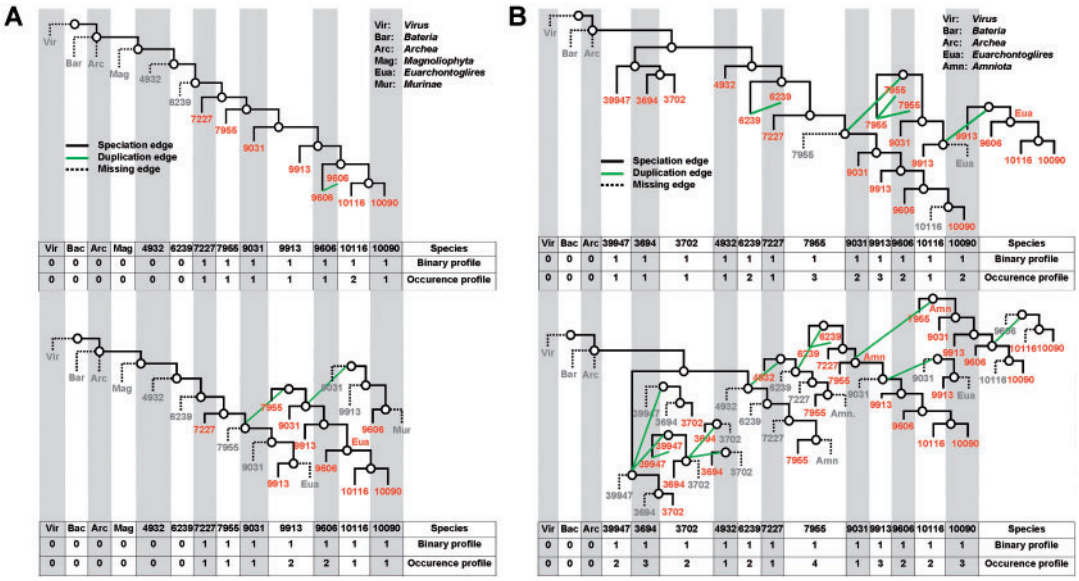

Fig. 5. Examples of potential FPs by the traditional methods that are TNs by the EPT method. The EPTs, the corresponding occurrence and binary profiles of protein Q9H3P2 (upper) and Q86Y82 (below) in human (**A**) and P12945 and P32602 in yeast (**B**). The gray (red) numbers are the taxonomy identifiers of missing (present) proteins. The gray characters are the taxonomy names of the roots of the missing subtrees. The speciation, duplication and missing edges are represented by solid black, red and dashed edges.

the use of EPTs provides a significant increase of the precision and sensitivity for co-evolution signal detection.

3.2 Biological feature of predicted protein functional linkages

The relationship between the prediction methods and GO attributes (Ashburner *et al.*, 2000) is shown in Figure 6. For BP and CC, the average GO semantic score (GO-SS) of the top predictions by the EPT method is higher than that of the top predictions by other methods (see Fig. 6B and C for yeast and Fig. 6E and F for human), indicating that the EPT method is able to find functional linkages among proteins that participate in the same biological processes or locate in the same cellular components. Figure 6A and D show that the linkages predicted by the EPT method exist between proteins that do not perform the same molecular functions. This is as expected because typically a set of different biochemical activities (molecular functions) are required to implement biological processes.

3.3 Examples of predicted functional linkages—linking cancer to DNA repair complexes

Supplementary Table S1 and S2 show several pairs of proteins that participate in the same biological process in yeast and human. These pairs have very high EPT scores so they are predicted as likely functional linkages by the EPT methods. A number of these predicted linkages are confirmed by experimental assays. The others can be considered as 'novel' linkages, which might be candidates for further experimental validation.

An interesting example of yeast functional linkages predicted by the EPT method is the linkage between the large (alpha) and small (beta) subunits of yeast phenylalanyl-tRNA synthetase (Genes FRS1 and FRS2, Uniprot accession numbers P15625 and P15624 , respectively). The association among these two proteins has been found by many experimental assays using Affinity Capture-MS methods (Collins *et al.*, 2007; Gavin *et al.*, 2002, 2006; Krogan *et al.*, 2006; Snitkin *et al.*, 2006). Among 6 931 774 yeast scored protein pairs, the EPT method ranks the pair of P15625 and P15624 as the 13th, indicating that this pair is likely identified as a TP by the EPT method. However, this pair is out of top 100 000 predictions by any conventional methods using either occurrence or binary profiles (top 7% for occ_ed, top 32% for occ_ps and top 50% for bin_ps), meaning that this protein pair is potentially identified as a FN.

As an example of human predicted functional linkages by the EPT method, we predict several interaction partners for DNA-directed polymerase theta (POLQ, Uniprot accession number Q59EE4).

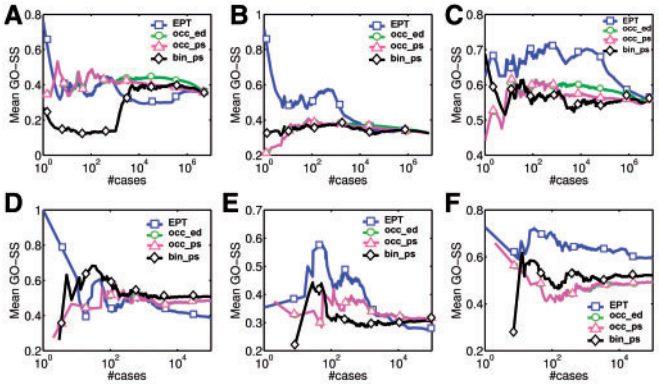

Fig. 6. GO-SS for molecular function (**A** and **D**), biological process (**B**, and **E**) and cellular component (**C**, and **F**) of yeast and human prediction, respectively. The *x* axes represent the number of top predictions (cases with highest prediction scores) by the EPT method, occ_ed, occ_ps and the bin_ps and the *y* axes represent the average GO-SS of the top predictions.

The predicted interaction partners of POLQ include proliferating cell nuclear antigen (PCNA, P12004), replication factor C (RFC4, P35249) and structural maintenance of chromosomes protein 1A (SMC1A, Q14683). Among more than 10^9 scored pairs of human proteins, the pair POLQ-PCNA is ranked 403rd. These two proteins are both mapped to DNA replication process. There has not been any direct evidence by experimental assays on the relationship between proliferating cell nuclear antigen and polymerase (DNA directed) theta protein. However, there exist associations between PCNA and other DNA polymerases, namely DNA polymerase kappa protein (Haracska *et al.*, 2002; Maga *et al.*, 2002; Shimazaki *et al.*, 2002), polymerase (DNA directed) delta protein Ducoux *et al.* (2001); Liu *et al.* (2003); Ohta *et al.* (2002); Pohler *et al.* (2005) and DNA polymerase eta protein (POLH, Q9Y253). POLH is specifically involved in DNA repair. DNA polymerase kappa and delta require activator 1 (alias RFC1-5) in addition to PCNA. The association between POLQ and PCNA is a very promising target for experimental validation. The second of POLQ's predicted interaction partners, RFC4, is part of activator complex 1 (composed of RFC1-5), which is known to bind to PCNA (Cai *et al.*, 1997; Ohta *et al.*, 2002). The third predicted interaction partner, SMC1A, is linked to disease. SMC1A is a central component of the cohesin complex, which apparently forms a large proteinaceous ring within which sister chromatids can be trapped after DNA replication. The cohesin complex interacts with a number of other proteins, including breast cancer associated BRCA1. Defects in SMC1A are the cause of Cornelia de Lange syndrome type 2 (CDLS2)[MIM:300590]; also known as Cornelia de Lange syndrome X-linked. CDLS is a clinically heterogeneous developmental disorder associated with malformations affecting multiple body parts. Mutated Cornelia de Lange cell lines display genomic instability and sensitivity to ionizing radiation and interstrand cross-linking agents (Apweiler *et al.*, 2004). Now that is an interesting piece of information, because POLQ has been implicated in the repair of interstrand cross-links (Apweiler *et al.*, 2004). So we hypothesize that SMC1A mutations manifest a radiation sensitive phenotype because of a disrupted physical association of the cohesin complex with POLQ.

4 CONCLUSIONS

All the phylogeny-based methods for predicting functional protein linkages are based on the common observation of the similarity between evolutionary patterns of interacting proteins. Therefore, the performance of these methods strongly depends on how the evolutionary patterns of proteins are described.

In prokaryotes, protein families typically contain one copy per species (Ranea *et al.*, 2007). This facilitates the prediction of functional linkages of protein using binary profiles. However, binary phylogenetic profiles are not evolutionarily informative enough for protein families in eukaryotic genomes, many of which contain more protein homologs per species. Occurrence profiles partly overcome the problem with multiprotein families in eukaryotes by counting the number of protein copies per species. Enhanced phylogenetic trees, which are graphical models of the evolution of proteins by speciation and duplication events, add richer information about the evolutionary history of proteins and help to overcome the problems which the conventional profiles encounter.

The EPT algorithm creates groups of proteins using the reciprocal best BLAST criterion that is commonly used to detect orthologs. The orthodox way to define orthologs is a subject of hot debate (Koonin, 2001; Ouzounis, 1999); we note that any protein family classification can be used for phylogenetic profiling and perfect recognition of orthologs is not necessary for our prediction purposes.

The current EPT method produces a simple clustering of proteins based on BLAST scores. The method can be improved in a number of ways to take into account the complexity of evolutionary relations including divergence, convergence, domain recombination and horizontal gene transfer events. For example, the choice of the descendant protein is somewhat arbitrary as BLAST scores do not reflect the complexity of these relations. We are going to address this problem in the future by using synteny information for complete genomes to maximize the number of syntenic protein pairs within layers of an EPT. Protein sequence profiles [e.g. HMMER (Eddy, 1998)] would be better representatives of the ancestral sequences than a single representative as in the current version. The present version of the method only models gene loss whereas inferring *de novo* creation or horizontal gene transfer requires post-processing of the trees. The EPT algorithm was designed so that singletons, as any other protein, are propagated to ancestral species for possible merging with clusters coming from other lineages. We define the ancestral taxon of an EPT as the smallest taxonomic node that encompasses all non-zero instances of the leaf proteins. In the future, we will test setting nodes to zero above the ancestral taxon, as well as modelling possible horizontal gene transfer events (which generate more than one origins of extant protein groups in the EPT).

Like other phylogeny-based methods, enhanced phylogenetic tree method is limited at inferring physical interactions of proteins but is promising at predicting functionally linked proteins. Functionally linked proteins were benchmarked against GO semantic similarity. Unlike domain-based approaches where domains, e.g. binding domains, are believed to perform the same molecular function but may occur in proteins that participate in very different biological processes, the thinking behind such a per-protein model as the EPT is that orthologs will be performing the same biological process across species. There are tools available which combine evidence—often weak and always noisy—from many different types of experimental and computational data to make integrated predictions of functional linkages (e.g. String, FunCoup). Examination of our top predictions indicated that they were often complementary to these existing tools, but had strong evidence from the co-expression data. We believe that the EPT method would be a valuable, orthogonal addition to such integrative tools. Our method significantly surpasses conventional methods in prediction performance and potentially discovers more reliable functional linkages.

ACKNOWLEDGEMENTS

We thank the members of Bioinformatics group for discussion.

Funding: Marie Curie grant (MRTN-CT-2005-019475).

Conflict of Interest: none declared.

REFERENCES

Apweiler,R. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet,*, **25**, 25–29.

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

Barker,D. *et al.* (2007) Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics*, **23**, 14–20.

Cai,J. *et al.* (1997) A complex consisting of human replication factor C p40, p37, and p36 subunits is a DNA-dependent ATPase and an intermediate in the assembly of the holoenzyme. *J. Biol. Chem.*, **272**, 18974–18981.

Collins,S.R. *et al.* (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics*, **6**, 439–450.

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

Date,S.V. and Marcotte,E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol*., **21**, 1055–1062.

Ducoux,M. *et al.* (2001) Mediation of proliferating cell nuclear antigen (PCNA)-dependent DNA replication through a conserved p21Cip1-like PCNA-binding motif present in the third subunit of human DNA polymerase. *J. Biol. Chem.*, **276**, 49258–49266.

Dutkowski,J. and Tiuryn,J. (2009) Phylogeny-guided interaction mapping in seven eukaryotes. *BMC Bioinformatics*, **10**, 393.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eisenberg,D. *et al.* (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.

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

Gavin,A.-C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gavin,A.-C. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

Gentleman,R. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Glazko,G. and Mushegian,A. (2004) Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol.*, **5**, R32.

Haracska,L. *et al.* (2002) Stimulation of DNA synthesis activity of human DNA polymerase kappa by PCNA. *Mol. Cell. Biol.*, **22**, 784–791.

Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664.

Huynen,M. *et al.* (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.

Koonin,E. (2001) An apology for orthologs - or brave new memes. *Genome Biol.*, **2**, comment1005.

Krogan,N.J. *et al.* (2006) Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**, 637–643.

Liu,L. *et al.* (2003) Identification of a novel protein, PDIP38, that interacts with the p50 subunit of DNA polymerase and proliferating cell nuclear antigen. *J. Biol. Chem.*, **278**, 10041–10047.

Maga,G. *et al.* (2002) Human DNA polymerase functionally and physically interacts with proliferating cell nuclear antigen in normal and translesion DNA synthesis. *J. Biol. Chem.*, **277**, 48434–48440.

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

Marcotte,E.M. *et al.* (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **97**, 12115–12120.

Mewes,H.W. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.

O'Brien,K.P. *et al.* (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.

Ohta,S. *et al.* (2002) A proteomics approach to identify proliferating cell nuclear antigen (PCNA)-binding proteins in human cell lysates: identification of the human CHL12/RFCs2-5 complex as a novel PCNA-binding protein. *J. Biol. Chem.*, **277**, 40362–40367.

Ouzounis,C. (1999) Orthology: another terminology muddle. *Trends Genet.*, **15**, 445.

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

Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.*, **14**, 609–614.

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

Pohler,J.R. *et al.* (2005) An in vivo analysis of the localisation and interactions of human p66 dna polymerase delta subunit. *BMC Mol. Biol.*, **6**, 17.

Ranea,J.A.G. *et al.* (2007) Predicting protein function with hierarchical phylogenetic profiles: the gene3d phylo-tuner method applied to eukaryotic genomes. *PLoS Comput. Biol.*, **3**, e237.

Remm,M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

Ruepp,A. *et al.* (2010) CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.*, **38**, D497–D501.

Shimazaki,N. *et al.* (2002) Over-expression of human DNA polymerase lambda in E. coli and characterization of the recombinant enzyme. *Genes Cells*, **7**, 639–651.

Snitkin,E. *et al.* (2006) Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics*, **7**, 420.

Sprinzak,E. and Margalit,H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.*, **311**, 681–692.

Ta,H.X. and Holm,L. (2009) Evaluation of different domain-based methods in protein interaction prediction. *Biochem. Biophys. Res. Commun.*, **390**, 357–362.

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

Wang,J.Z. *et al.* (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.

Wu,J. *et al.* (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, **19**, 1524–1530.

Zmasek,C. and Eddy,S. (2002) Rio: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.