Bioinformatics, 31(13), 2015, 2166–2173
doi: 10.1093/bioinformatics/btv102
Advance Access Publication Date: 24 February 2015
Original Paper

OXFORD

Systems biology

Detection of significant protein coevolution

David Ochoa[1,†], David Juan[2], Alfonso Valencia[2] and Florencio Pazos[1,*]

[1]Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), C/ Darwin 3, 28049 Madrid and [2]Structural Bioinformatics Group, Spanish National Cancer Research Centre (CNIO), C/ Melchor Fernández Almagro 3, 28029 Madrid, Spain

*To whom correspondence should be addressed.

[†]Present address: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

Associate Editor: Janet Kelso

Received on June 18, 2014; revised on January 13, 2015; accepted on February 11, 2015

Abstract

Motivation: The evolution of proteins cannot be fully understood without taking into account the coevolutionary linkages entangling them. From a practical point of view, coevolution between protein families has been used as a way of detecting protein interactions and functional relationships from genomic information. The most common approach to inferring protein coevolution involves the quantification of phylogenetic tree similarity using a family of methodologies termed *mirrortree.* In spite of their success, a fundamental problem of these approaches is the lack of an adequate statistical framework to assess the significance of a given coevolutionary score (tree similarity). As a consequence, a number of *ad hoc* filters and arbitrary thresholds are required in an attempt to obtain a final set of confident coevolutionary signals.

Results: In this work, we developed a method for associating confidence estimators (*P* values) to the tree-similarity scores, using a null model specifically designed for the tree comparison problem. We show how this approach largely improves the quality and coverage (number of pairs that can be evaluated) of the detected coevolution in all the stages of the *mirrortree* workflow, independently of the starting genomic information. This not only leads to a better understanding of protein coevolution and its biological implications, but also to obtain a highly reliable and comprehensive network of predicted interactions, as well as information on the substructure of macromolecular complexes using only genomic information.

Availability and implementation: The software and datasets used in this work are freely available at: http://csbg.cnb.csic.es/pMT/.

Contact: pazos@cnb.csic.es

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

1 Introduction

Coevolution is a widespread phenomenon with important implications at all biological levels. At the molecular level, coevolution plays a fundamental role in key cellular systems, allowing their components to change and evolve while maintaining their interactions (Juan *et al.*, 2013). Quantifying the ongoing coevolution between different molecular features has been proposed as a proxy to predict different types of interactions. Among the most successful approaches are those applied for detecting protein–protein interactions and functional relationships and, more recently, for the accurate prediction of residue contacts in individual proteins (Morcos *et al.*, 2011; Jones *et al.*, 2012).

The evolutionary histories of interacting protein partners are not independent but entangled in many different ways. Coevolution-based approaches benefit from this observation in order to detect protein interactions. As the predictions are only based on the vast amount of genomic and sequence information available nowadays, these methods are suitable for the systematic detection of interactions in a wide range of organisms, such as bacteria (Juan *et al.*, 2008), fungi (Clark *et al.*, 2011) or human (Havugimana *et al.*, 2012).

Protein coevolution is reflected in different genomic and sequence features (Juan *et al.*, 2013). A popular approach for inferring coevolution is to evaluate the similarity of phylogenetic trees. Phylogenetic trees represent the putative evolutionary histories of the corresponding protein families and, consequently, similar protein trees imply similar evolutionary histories. This is the basis of the *mirrortree* and related methods for detecting protein interactions using genomic information [see (Juan *et al.*, 2013) for a recent review].

In its simplest form, *mirrortree*-related approaches represent a phylogenetic tree of a family of orthologs as a distance matrix, and quantify the similarity between two trees as the Pearson's linear correlation between the sets of values of their corresponding matrices (Pazos and Valencia, 2001). In spite of their success in detecting interactions, this approach has a number of 'methodological' and 'biological' problems, which were partially overcome with different variations of the original method. On the methodological side, it is well known that the internal codependencies between the values of distances matrices burden the significance assessment of a given correlation coefficient. The tabulated *P* values, regularly used to assign correlation significances, assume the independence of the vectors' components. Since the distances in the phylogenetic trees cannot freely change to adopt any possible value, strictly speaking, these *P* values are not adequate, in spite of having shown an improvement on the interaction prediction (Juan *et al.*, 2008). Another consequence of the incompleteness of this null model is that these methods required a minimum number of organisms in common between the two trees (generally around 15) in order to take into consideration a given correlation. This additional requisite drastically decreases the coverage of the methods: the number of protein pairs that can be evaluated. Moving into the biological problems, the distances within the matrices are also constrained by the evolutionary characteristics of the trees they come from. For example, the limits in the divergence between homologous sequences constrain the observed distances to certain values. Moreover, tree leaves corresponding to evolutionary close species will present a strong tendency to be close in all phylogenetic trees regardless of the proteins they represent, and vice versa for distant species. As a consequence, all trees of orthologs have a certain level of similarity, between them and with the canonical species tree. This 'background' similarity, reflected in correlation values relatively higher even for non-interacting pairs, has been corrected using external representations of the tree of life (Pazos *et al.*, 2005; Sato *et al.*, 2005) or removing the common signal in a large collection of trees (Juan *et al.*, 2008). Another problem is related to the presence of redundant taxa on the sets of organisms used for constructing the trees. As the sequencing efforts are biased to certain organisms, phylogenetic trees become populated in an unbalanced way. Indeed, previous analyses have shown that using trees based on nonredundant sets of organisms increased the performance of the basic *mirrortree* approach (Herman *et al.*, 2011; Muley and Ranjan, 2012). Some of these problems are alleviated in the recent context-based *mirrortree* variations. These approaches use all pairwise tree similarities of a given proteome to reassess a given coevolutionary signal, largely improving the predictions (Juan *et al.*, 2008). These context-based approaches also help to disentangle direct coevolutionary signals from those due to third proteins (indirect), which is important since the latter are not always related to protein interactions. Indeed, in the case of residue–residue coevolution, recent methods characterized by this capacity to filter indirect coevolutions are able to predict contacts in protein structures with very high reliability when fed with enough sequences (Morcos *et al.*, 2011; Jones *et al.*, 2012).

Here we propose a new approach, *P-mirrortree* (pMT), which aims at correcting, in a single shot, all these problems associated with the quantification of tree similarity and the assessment of its significance. pMT generates null distributions of tree similarities (correlation coefficients) obtained from large sets of shuffled phylogenetic trees from which empirical *P* values can be derived.

We show that this approach overcomes previous *mirrortree* versions and produces predictions of high quality and coverage, the latest due to not requiring a threshold of minimum number of organisms in common to evaluate tree similarity. The improvement is particularly high when this approach is coupled to the modern context-based methods. Moreover, pMT is largely insensitive to the characteristics of the set of organisms used to build the trees (number and taxonomic redundancy).

In order to assess these improvements under different scenarios of available genomic information, we perform a retrospective analysis of the predictive power of several *mirrortree*-based approaches when fed with the genomic information available at different time points in the past. Consequently, this work also evaluates, for the first time, how the non-homogeneous exploration of the bacterial taxonomy in terms of sequenced genomes affects the detection of coevolution, and which trends are expected for the future.

2 Methods

We want to highlight some of the aforementioned problems, in particular those associated to the sets of organisms used to generate the trees and to the type of interactions aimed to predict, as well as foresee the impact of these issues as more genomes are sequenced. However, our main aim is to evaluate how the new pMT method can alleviate these problems. For all that, we compared the performances of pMT and previous versions of *mirrortree* (MT) predicting three types of interactions between *Escherichia coli* proteins by using phylogenetic trees constructed with the genomes available at different time points in the past. The different parts of the method are detailed below.

2.1 Mirrortree

The basic *mirrortree* method (Pazos and Valencia, 2001) quantifies the similarity between two phylogenetic trees of orthologs as the Pearson's correlation coefficient between the two corresponding distance matrices. So, for two protein phylogenetic trees *A* and *B* with *n* organisms in common, the *mirrortree* score would be

$$r_{AB} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(dA_{ij} - dA)\cdot(dB_{ij} - dB)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(dA_{ij} - dA)^2}\cdot\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(dB_{ij} - dB)^2}} \quad (1)$$

where dA_{ij} is the distance between organisms *i* and *j* in the tree of family *A*. dB_{ij} is the corresponding distance in tree *B*, and dA and dB are the corresponding average values. Distances are obtained by summing the lengths of the branches separating the two leaves (organisms).

2.2 p-mirrortree

The goal of pMT is to associate a *P* value to a given r_{AB} score by comparing it to a null distribution of scores obtained for a large set of shuffled pairs of trees with a similar number of organisms in common to that of *A* and *B* (*n*). The null distribution of tree similarities of (mostly) noninteracting proteins constructed in this way is

expected to reflect all the problems influencing the observed scores commented in Section 1. Consequently tree similarities deviating from this null distribution are expected to be meaningful.

The process is illustrated in Figure 1. It starts with a large collection of 'background' phylogenetic trees, those for all proteins in *E.coli* in this case. All the pair-wise combinations between these trees are generated and these tree pairs are split in different groups depending on the number of organisms in common (Fig. 1). The size of the groups is defined in a logarithmic scale to add more sensitivity to the correlation changes in trees sharing a low number of leaves. Depending on the total number of organisms used to model the trees and the computational resources available, a smaller or larger number of groups can be used. For each of these groups, an iterative process is carried out to obtain its corresponding null distribution of tree similarities (Fig. 1). In each iteration, a pair of trees is randomly sampled with replacement and the corresponding distance matrices are retrieved from a pre-calculated pool. The sub-matrices containing only the distances between organisms shared by both trees are extracted (Fig. 1). The values of the sub-matrices are converted to z-scores so as to put them in the same scale. Once both matrices are in the same scale, the distance values corresponding to a given organism are swapped between both families with a given probability. Finally, the distance matrices are put back to their original scales using the original mean and standard deviations, and completed with the distances involving organisms not shared by the trees. This part of the process can be seen as interchanging the branches corresponding to a given organism between both trees (Fig. 1), although everything is done with the distance matrices. The two resulting modified matrices are returned to the pool in replacement of the original ones and are available for further iterations. Therefore, a given matrix can swap organisms (rows/columns) multiple times with different matrices. After a number of iterations, the pool contains randomly modified distance matrices but always limited to the distance information available in other trees (Fig. 1). Finally, for each group, all possible pairwise correlation coefficients are calculated (eq. 1 above) using these shuffled matrices, generating a null distribution of tree similarities for that size group. Once these background distributions are calculated, the significance of a given *mirrortree* correlation coefficient obtained for a pair of (real) trees with a given number of organisms in common can be evaluated by calculating the probability (*P* value) of finding a higher coefficient in the corresponding background distribution (Fig. 1). A low *P* value indicates a tree similarity significantly higher than those observed between shuffled trees with similar characteristics and, consequently, can be interpreted as indicative of a meaningful coevolution.

For this particular work, we used 40 groups (intervals) of number of organisms in common and ran 1000 permutation steps with a branch swapping probability of 0.05.

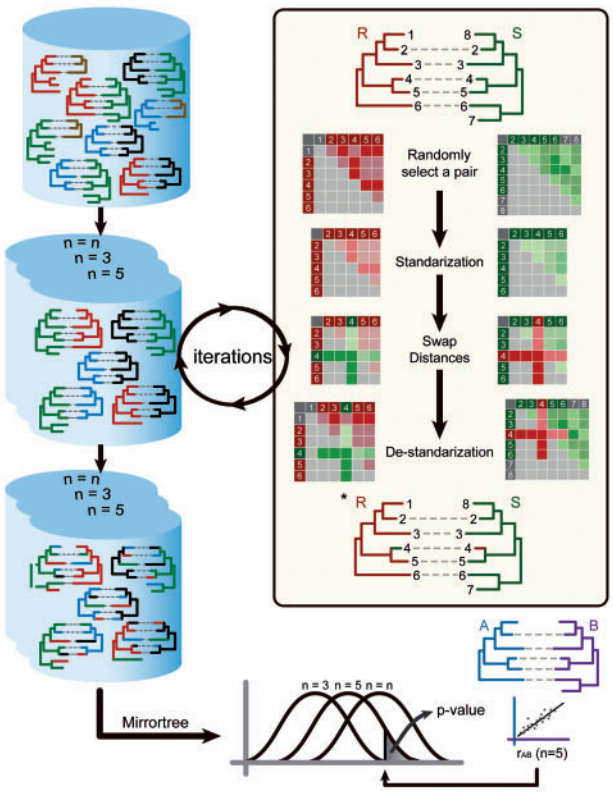

Fig. 1. Overview of the pMT methodology. In the first step, all the possible pairs of phylogenetic trees are split into groups (cylinders) based on the number of organisms in common. For each group, a number of iterations of a distance swapping procedure are run in order to randomize the trees present in the set. In each iteration, a random pair of trees is selected and standardized (bring to the same scale) based on the distances between sequences belonging to the organisms in common. Rows/columns with the distances involving a given organism ('4' in the example) are swapped between the two matrices with a given probability. The resulting matrices are de-standardized to restore their original scales. Both phylogenetic trees are introduced again in the pool of trees for further iterations. The final set of shuffled trees is used to calculate the background distribution of tree similarities. These distributions are used to quantify the statistical significance of an observed tree similarity score

2.3 pMT as input for context-based methods

The *P*-value associated by pMT to each pair of proteins can be used as input for the *mirrortree* context-based methods (Juan *et al.*, 2008). As originally formulated, these methods take as input the whole network or pairwise tree similarities (Pearson's correlations) for a whole proteome. For a given protein, the vector containing all the correlations with the rest of the proteome is called 'co-evolutionary profile'. The 'profile correlation' method (PC) requantifies the coevolution between two proteins as the Pearson's correlation between their corresponding coevolutionary profiles. The 'context-mirror' method (CM) calculates the partial correlation between two profiles in relation to that of a third protein, aiming at discarding non-specific coevolutionary signals shared between many proteins. For this work, the PC and CM methods were applied using the matrix of pairwise *P* values, instead of the original correlation coefficients.

2.4 Generation of phylogenetic trees

Using the completely sequenced Eubacteria and Archaea genomes available in KEGG (release 59.0, August 2011) (Kanehisa *et al.*, 2004), we created phylogenetic trees for all *E.coli* proteins. Prokaryotic protein families, including both paralogs and orthologs, were retrieved for each protein directly from the orthology groups in KEGG (KO groups). In order to select a single ortholog for each organism, we took the sequence best ranked against the original *E.coli* protein on the precalculated lists of 'BLAST best bi-directional hits' stored in KEGG. These sets of orthologs were aligned with MUSCLE (Edgar, 2004) using default parameters. For each of the resulting multiple sequence alignments (MSAs), a phylogenetic tree was created using the neighbor-joining algorithm implemented in TreeBeST (Edgar, 2004), again running this program with default parameters. This produced a final set of 2844 phylogenetic trees.

The corresponding matrices of cophenetic distances were generated by summing the length of the branches separating each pair of organisms in these trees, as explained above.

2.5 Historical sets of reference organisms and trees

For each year, from 1995 to 2010, we created two sets of reference organisms, 'redundant' and 'nonredundant'. The 'redundant' set contains all the fully sequenced prokaryotic organisms deposited in KEGG in a particular year. The 'nonredundant' set was obtained from it by removing the evolutionary close organisms. In order to do that, for a given pair of organisms the pair-wise identities between their orthologous sequences were calculated from the aforementioned MSAs. If the two proteomes have more than 70% of the orthologs with 95% or more sequence identity, one of them is excluded. We ran this iterative process starting from the organism with the highest sequence identity with *E.coli* to that with the lowest. The number of organisms for each year within both datasets is shown in Supplementary Figure S1. We discarded the period 1995–1999 for the forthcoming analysis due to the low number of organisms available at these years.

For each distance matrix obtained with the genomes available in 2011, 'historical' versions were derived by keeping only the organisms (rows and columns) available in a particular year. This is done for both the 'redundant' and 'nonredundant' year-based sets. Neither the original MSAs nor the trees are recalculated, they are only 'trimmed' leaving only the sequences available at a given year. So, for each *E.coli* protein, we ended up with 22 distance matrices that try to reflect what a user would had obtained for each of the 11 years using redundant and non-redundant versions of the set of genomes available each year. These year-based sets of matrices can be used as input for the *mirrortree, tol-mirrortree* and pMT methods so as to 'simulate' a genome-wide prediction of interactions using the genomic information available at a particular year.

2.6 Tol-mirrrotree

In order to compare pMT with a *mirrortree* variant which explicitly corrects the background similarity due to speciation, we applied *the tol-mirrortree* (tol-MT) method to the same datasets. tol-MT (Pazos *et al.*, 2005) corrects the distance matrices of both proteins (dA and dB in Equation (1)) with the overall phylogenetic distances between species in an attempt to correct the background tree similarity due to speciation.

In order to generate the species tree, the 16S rRNA genes for the species within our dataset were retrieved from KEGG (K01977 orthologs group) and aligned with MAFFT (Katoh and Standley, 2013) (default options). When more than one 16S rRNA gene is present in a given species, that with the highest identity (calculated from the alignment above) with *E.coli* b0201 is chosen as the putative ortholog. The set of genes is not realigned after this filtering. A phylogenetic tree is generated from this alignment with FasttTree (Price *et al.*, 2010) (options '-nt' and '-gtr'). Distances between species are extracted from this tree as described previously for the protein trees.

The protein family whose tree is most similar to the species tree (highest correlation) and contains 90% or more of the species is taken as the 'molecular clock' to calculate the ratio between nucleic acid and protein distances. This ratio is used for rescaling the 16S rRNA distances before subtracting them from each protein's distances (Pazos *et al.*, 2005), and is calculated as the average of the rRNA distances over that of the protein distances for that molecular clock family. This is done for each historical dataset independently to get closer to the real scenarios where not all proteins/species are available at a given time point.

2.7 Performance evaluation

For a given set of distance matrices, the three methods produce a list of protein pairs sorted by their corresponding scores which aim to quantify the coevolution between the two proteins of the pair (raw Pearson's correlation for MT and tol-MT, and *P* value for pMT). Pairs involving homologous proteins (those within the same KO group of KEGG) are excluded for this evaluation since they can eventually point to the same orthologs and consequently have identical trees in spite of not being interacting, which would produce artifacts in the evaluation. For comparative purposes, only the pairs with predictions in the four sets (MT/tol-MT versus pMT and redundant versus nonredundant) are evaluated. Each pair in these lists can be labeled as 'positive' (interacting) or 'negative' (noninteracting) according to different interaction criteria. In this work we used three different independent gold standard datasets containing different types of physical and funtional interactions for the model organism *E.coli*:

- *Binary physical:* direct binary physical interactions obtained from MPIDB (Goll *et al.*, 2008). These interactions were manually curated from the literature or imported from other databases. This version of the database contains 2103 binary interactions between 1538 different *E.coli* proteins.
- *Complexes:* physical interactions inferred by copresence in the same macromolecular complex. These physical interactions may be direct or not (i.e. two proteins in the same complex but not 'touching' each other). The protein complexes are experimentally determined and extracted from the EcoCyc databases (Keseler *et al.*, 2005). The set includes 1354 pairs between 591 proteins.
- *Pathways:* functional interactions inferred as co-presence in the same metabolic pathway as defined in EcoCyc. This dataset comprises 4491 pairs between 719 proteins.

These datasets describe only 'positive' cases. For each of them, the corresponding negative set was constructed by generating all possible pairs between the proteins reported in the set, excluding those pairs already reported as interacting. The sorted lists of pairs generated by each method, once the pairs are labeled as 'positive' or 'negative' following the aforementioned criteria, can be subject to different analysis aimed at assessing the capacity of the methods' scores to separate positives from negatives. In this case, we performed ROC analysis (Fawcett, 2006), and calculated the *F*-measure and the accuracy of the top-N pairs.

3 Results

3.1 Specific versus general NULL model

Previous versions of *mirrortree* [i.e. (Pazos and Valencia, 2001)] either disregard the *P* values associated to the correlation scores or use those calculated analytically, or derived from random sets of numbers, that do not fulfill the properties of tree-based distances ('tabulated *P* values'). In order to get insight into the differences between these and the pMT *P* values, specifically derived for the genomic tree comparison problem, and better understand the problems the former were producing, we compared the null distributions obtained by both approaches.

In Figure 2A and Supplementary Figure S2 some of these pMT distributions used to extract *P* values obtained for the genomes available at different years are compared with the equivalent

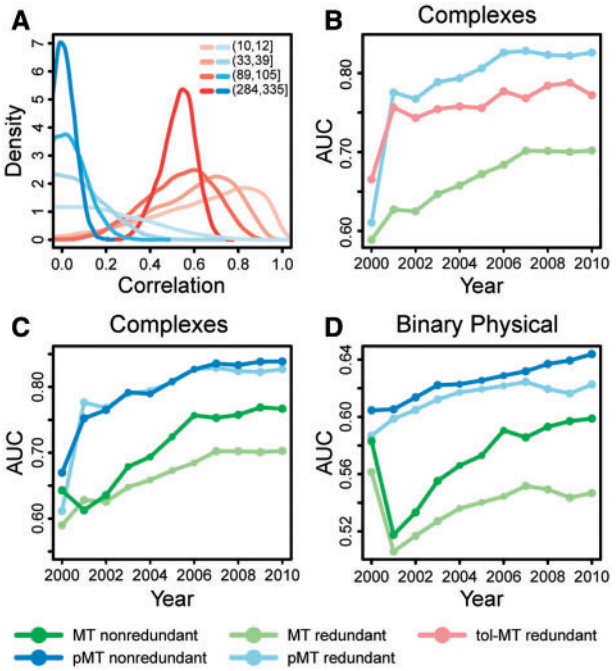

Fig. 2. Prediction performances. (**A**) Density functions for the distribution of correlation coefficients in sets of random pairs of numbers and sets of distances extracted from pairs of permuted phylogenetic trees. The genomes available in 2010 were used as reference to generate shuffled trees for *E.coli* proteins and the corresponding distributions of tree similarities (red) were calculated for the pairs of trees sharing different numbers of organisms in common (between brackets). Those distributions were compared with equivalent ones generated from random sets of numbers in the same size intervals (blue). Equivalent plots for the remaining years are available in the Supplementary Figure S2. (**B**) Performance of the MT, tol-MT, and pMT methods when predicting co-membership to the same macromolecular complexes using the fully-sequenced genomes available in the period 2000–2010. The performance was evaluated in terms of AUC. (**C**, **D**) Effect of the organism redundancy on the MT and pMT performances predicting comembership to the same macromolecular complexes and binary physical interactions

distributions of correlations between sets of random numbers. The *P* values used in previous versions of *mirrortree* would indeed be obtained from these latest distributions (bluish in the figures). It is clear that these two types of distributions (correlations between sets of random numbers—blue- and between trees—red-) are largely different. As expected, the average correlation coefficient between sets of random numbers is always 0. Moreover, as the size of the sets increases, the probability of obtaining 'extreme correlations', either positive or negative, decreases, leading to slightly narrower distributions. These general observations are not extended to the correlation coefficients calculated using distance matrices of permuted trees. As previously described, phylogenetic trees tend to share a background similarity and, consequently, the distributions of correlations are always shifted to high values and not centered in 0. Moreover, the correlation value at which these distributions are centered depends on the number of points being correlated (proportional to the number of organisms in common between both trees) and, in general, gets lower as trees with more organisms in common are being compared. This means that higher correlation values are expected by chance for smaller trees. Not only the average value but also the shape of the distributions varies largely with the number of organisms. Pairs of trees with a small set of organisms in common present a very wide range of correlation coefficients, whereas pairs of trees sharing many orthologs tend to present a narrower range of correlations (Fig. 2A and Supplementary Fig. S2).

3.2 pMT versus mirrortree and tol-MT

We compared the performance of pMT with that of the original *mirrortree* (MT) and one of its variants, tol-MT, in predicting different types of interactions using the set of organisms available in the period 2000–2010, as well as their nonredundant versions (see Section 2 for details). The historical datasets allow not only to compare these approaches and assess how the characteristics of the set of organisms (in terms of redundancy, etc.) affect their performances, but also to foresee how these methods will work in the future with the current trend of genome sequencing efforts.

The performances of these methods when predicting two different types of physical interactions are shown in Figure 2 (panels B–D). The ROC analysis confirms the ability of all these methods based on phylogenetic tree similarity to capture part of the coevolutionary signal related to protein interactions. The general trends observed suggest that protein interaction predictions benefit from the increase in the number of sequenced genomes. Indeed, this trend has not reach a plateau, so further improvement can be expected over the following years. Predicted interactions defined as those belonging to the same macromolecular complex present the highest performances, followed by binary physical interactions. Functional interactions based on copresence in same metabolic pathway present poor and constant AUCs (Supplementary Fig. S3), suggesting that coevolution may not be a generalized process between the proteins of the same pathways, although the artificial and broad definition of 'pathway' might also be affecting these results.

The performance of pMT when predicting physical interactions using the organisms available during the explored 10 years is higher than that of MT (0.10–0.15 increase of AUC) (Fig. 2). The performance of tol-MT lies in the middle between the original MT and the new pMT. The performance of pMT is more stable over time and has a much lower dependence on the number of available genomes. MT performances improve when using 'nonredundant' sets, confirming previous observations (Herman *et al.*, 2011). Indeed, the performance gap between the 'redundant' and 'nonredundant' sets becomes larger as taxonomical redundancy increases over time (Fig. 2 and Supplementary Fig. S1). Interestingly, pMT is much more robust when dealing with redundancy, reflected in a very small difference between the redundant and non-redundant sets which, additionally, remains constant over time.

Supplementary Figure S7 shows the dependence of two commonly used performance figures ('positive predictive value' –PPV- and F-measure) of the *P* value cutoff taken. It can be seen, for example, that a *P* value cut of 0.005 renders the best equilibrium of positive/negative recovery, as quantified by the F-measure.

A common shortcut in previous *mirrortree* and related approaches is to ignore the protein pairs with less than a given number of organisms in common, generally 15. We evaluated the historical MT and pMT performances with this limitation (Supplementary Figure S4). pMT performance is only slightly improved by this additional constraint, indicating that this method is able to efficiently deal with the comparison of trees with a low number of organisms in common. As expected, the original MT method takes advantage of this constraint. Nevertheless, the performance starts to drop drastically at a certain number of sequenced organisms for both 'redundant' (around 2003) and 'nonredundant' (2006) sets, indicating that the threshold of a minimum number of organisms, which so far has been usually fixed at 15, would have to be adapted to the set of

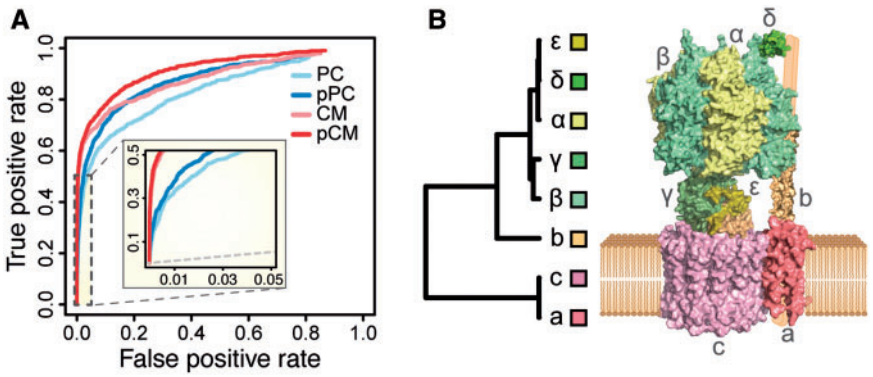

Fig 3. Results of pMT coupled with context-based methods. (**A**) ROC plot comparing the performance of the context methods (PC and CM) when fed with the original *mirrortree* raw score (Pearson's correlation) and with pMT *p* values (pPC and pCM). The inset contains a zoomed view of the ROC region corresponding to the top scores of the methods. The AUC values for these ROC curves are: PC (0.83), CM (0.85), pPC (0.85), and pCM (0.89). (**B**) Coevolutionary analysis of the eight subunits of the *E.coli* ATP synthase. The hierarchical clustering of the pPC profiles of the subunits is calculated using Ward's minimum variance algorithm. The three-dimensional representation of the *E.coli* ATP synthase was composed based on the structures available (PDB ids: 2A7U, 1C17, 1E79, 1L2P)

available organisms so as to obtain the optimal performance. As more organisms are available, this threshold needs to be more restrictive (higher).

The main problem of imposing a threshold of minimum number of organisms in common is the tremendous loss in coverage (number of pairs that can be evaluated) (Supplementary Fig. S4). For example, using the 198 organisms available in 2004, from the 215 026 pairs that can be evaluated without imposing any threshold, only 122 518 can be calculated when more than 15 organisms are required (43% loss in coverage).

3.3 Context pMT

We evaluate the improvement that the use of these pMT *P* values produces in *mirrotree*-based methods by analyzing their effect on the results of the more advanced implementations, i.e. 'context methods' (PC and CM, (Juan *et al.*, 2008)). When these methods are applied to the 2011 set of KEGG genomes (see Section 2) both PC and CM fed with pMT scores (pPC and pCM) perform better than their standard implementations with raw correlation values (Fig. 3A). This is evident in both the global discriminative capacity evaluated by the ROC analysis (Fig. 3A) and in the accuracy of the top-N pairs (Supplementary Fig. S5).

3.4 Detailed evolutionary analysis of macromolecular complexes

To illustrate the capacity of these methods not only to detect interacting pairs but also to obtain detailed information on macromolecular complexes, we show the results of the analysis of the coevolutionary relationships between the members of the *E.coli* ATP synthase obtained with these new approaches. The tree in Figure 3B shows a hierarchical clustering of the pairwise PC scores (based on *P* values –pPC-) for the eight members of this membrane macromolecular complex. This tree highlights the hierarchical coevolutionary relationships between these proteins. If we examine the clearest partition of the tree (that rendering three clusters), we can see a cluster containing the 'a' and 'c' subunits, a second cluster formed by the subunit 'b' alone, and a third cluster containing the five different members of the F1 particle (greek letters). These results are in agreement with the three-dimensional model of the ATP synthase, in which the 'a' and 'c' subunits are embedded in the membrane forming the proton pore, the F1 particle is the cytosolic machinery in charge of the ADP phosphorylation, and the subunit 'b' connects both sub-complexes (Fig. 3B). Consequently, this coevolutionary analysis generates clues on the architecture of the macromolecular complex, using only sequence information.

3.5 *E.coli* coevolutionary *network*

The whole coevolutionary network for *E.coli* obtained with this new approach is available as Supplementary File 1. This file can be interactively inspected with Cytoscape v.3 (www.cytoscape.org). An interactive representation of the same network is available online at http://csbg.cnb.csic.es/colievolution.

This network contains the coevolutionary relationships between *E.coli* proteins obtained with CM fed with pMT *P* values (pCM, level 10 of coevolutionary specificity (Juan *et al.*, 2008)) and a partial correlation cutoff of 0.56. The provided network contains additional information to contrast the predictions, such as detailed information on the proteins and different interaction evidences for the pairs. The clusters of proteins within this coevolutionary network are colored according with the results of an enrichment analysis of GeneOntology terms (Harris *et al.*, 2004). Inspecting these functional features in the network, it becomes evident that this new coevolution-based approach can produce a reliable genome-wide network of interactions and functional relationships and, consequently, provide insight into the underlying biological processes, all using only sequence information.

4 Discussion

Coevolution takes place at all biological levels (species, proteins, amino-acids, . . .) (Juan *et al.*, 2013). This evolutionary linkage between the biological entities is crucial for maintaining relationships and interactions while allowing the two partners to evolve and change. In this sense, coevolution is fundamental for evolutionary innovation.

At the molecular level, protein coevolution has provided a wealth of information about different systems. Recent advances on the detection of residue contacts have joined the detection of protein interactions as some of the most popular applications based on coevolution. For a recent review see (Juan *et al.*, 2013).

The maturity of coevolution-based approaches, together with the increase in the genomic information that feeds them, has led in recent years to the quotidian application of these approaches to many protein families of interest, in many cases directing or

combined with experimental approaches. See for example (Edgar *et al.*, 2012; Havugimana *et al.*, 2012; Sandler *et al.*, 2013; Zamir *et al.*, 2012), and (Ochoa and Pazos, 2014) for a review.

Nevertheless, the observed protein–protein coevolution, quantified as the similarity of the corresponding phylogenetic trees, is influenced by many factors complicating the disentanglement of those more directly related to the interaction.

The results presented here demonstrate that the prediction of protein interactions at different levels (pairs, complexes, and whole networks) is clearly improved when the statistical confidence of the pairwise tree similarity is evaluated based on a background distribution of tree similarities. Assessing the significance of a given tree similarity using this distribution of expected correlations corrects, in a natural way, many of the confounding factors that affect the performance of the original MT, including the background similarity due to the underlying speciation process, the redundancy of the set of organisms used for building the trees and the different range of organisms in which the two proteins are present. Previous attempts to correct these factors implied workarounds, pre- and post-filtering, arbitrary thresholds and other ad-hoc heuristic approaches. pMT intrinsically corrects all these factors producing an estimation of the likelihood of the co-evolution under a solid statistical framework. This is especially important taking into account that the genomic information used as input by these methods will change constantly due to the stream of newly sequencing genomes. Whereas heuristic approaches would have to change and adapt constantly to these new data, our results show how pMT is robust to this change in the input genomic information.

The pMT method does not require an artificial threshold on the minimum number of organisms to evaluate a given pair of trees. Leaving apart the difficulty in deciding such a threshold (which depends on the characteristics of the dataset, as we show), not requiring it has two main advantages, both related with the increase in coverage (number of pairs that can be evaluated). On one hand, the results are not biased to 'central' proteins: these proteins involved in core cellular processes are present in many organisms and consequently pairs involving them would more probably pass that threshold of minimum number of organisms. On the other hand, that increase in coverage is crucial to the context-based methods, which use the whole network of pair-wise tree similarities as input. With pMT this network is much more populated and that, together with the intrinsic better performance of pMT at the pair level, makes these context-based approaches render better results when coupled with pMT.

This new method is not only better for detecting interacting pairs, but can be used to get insight into the substructure and functioning of macromolecular complexes (as illustrated for the ATPase), as well as to obtain a highly reliable network of protein interactions at a genomic scale (as exemplified by the *E.coli* network generated here). In the latter case, the pMT method presented here is crucial since previous approaches were either highly reliable but presented a low coverage, or the other way around. pMT has a high performance, specially when coupled with context based methods, and can be applied to pairs of trees never explored before (e.g. those with a small number of species in common).

These improvements and advantages come at no cost in terms of applicability, since no additional restrictions are required to run pMT, apart from the contextual information necessary to generate the null distributions ('background' set of trees). As presented in this work, this can be seen as a drawback since trees for the whole proteome of interest are required to be used as background. On one hand, this is common to previous context-based approaches. On the other hand, it remains to be explored whether other more restricted sets of trees can eventually be used as background (e.g. trees for the membrane proteins or for those in a given biological process) with better results, as they could serve as a better background for representing the characteristics of the system of interest.

As a side result, the historical perspective of our analysis allows to foresee a continuous increase in the performance of co-evolution based approaches as more genomes are sequenced, highlighting the value of the ongoing genome sequencing projects.

There is a plethora of approaches based on the original *mirrortree* method (see (Juan *et al.*, 2013) for a review). The results presented here indicate that the pMT approach, which touches the base of the methodology itself, could improve all of them. That was the case for the two context-based *mirrortree* variations evaluated here. Although all the quantification of performances presented are based on the method's ability to detect interactions, we are confident that this approach goes beyond the 'practical' applicability in interaction prediction, and will serve to better understand the complex phenomenon of protein coevolution as well.

Acknowledgement

We sincerely thank the members of the Computational Systems Biology Group (CNB-CSIC) for interesting discussions.

Funding

Spanish Ministry for Science and Innovation (BIO2010-22109 and BIO2012-40205). Fellowship from the Basque Country (D.O.).

Conflict of Interest: none declared.

References

Clark,G. *et al.* (2011) Using coevolution to predict protein-protein interactions. *Methods Mol. Biol.*, **781**, 237–756.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Edgar,R.S. *et al.* (2012) Peroxiredoxins are conserved markers of circadian rhythms, *Nature*, **485**, 459–464.

Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recogn. Lett.*, **27**, 861–874.

Goll,J. *et al.* (2008) MPIDB: the microbial protein interaction database, *Bioinformatics*, **24**, 1743–1744.

Harris,M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

Havugimana,P.C. *et al.* (2012) A census of human soluble protein complexes. *Cell*, **150**, 1068–1081.

Herman,D. *et al.* (2011) Selection of organisms for the co-evolution-based study of protein interactions. *BMC Bioinformatics*, **12**, 363.

Jones,D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Juan,D. *et al.* (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. USA*, **105**, 934–939.

Juan,D. *et al.* (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.

Kanehisa,M. *et al.* (2004) The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, **32**, D277–D280.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Keseler,I.M. *et al.* (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D334–D337.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families, *Proc. Natl. Acad. Sci. USA*, **108**, E1293–E1301.

Muley,V.Y. and Ranjan,A. (2012) Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction. *PLoS One*, **7**, e42057.

Ochoa,D. and Pazos,F. (2014) Practical aspects of protein co-evolution. *Front Cell Dev. Biol.*, **2**, 14.

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.

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

Price,M.N. *et al.* (2010) FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Sandler,I. *et al.* (2013) Protein co-evolution: how do we combine bioinformatics and experimental approaches? *Mol. Biosyst.*, **9**, 175–181.

Sato,T. *et al.* (2005) The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, **21**, 3482–3489.

Zamir,L. *et al.* (2012) Tight coevolution of proliferating cell nuclear antigen (PCNA)-partner interaction networks in fungi leads to interspecies network incompatibility. *Proc. Natl. Acad. Sci. USA*, **109**, E406–E414.