the "BLAST best bi-directional hit" criterion. Putative orthologs required an *E*-value $\leq 1 \times 10^{-5}$ and an alignment coverage of 70% to be considered for analysis. The set of resulting sequences are then aligned using MUSCLE [214] with the default parameters. The phylogenetic trees are finally created by the neighbor-joining algorithm implemented in ClustalW [217], excluding the gaps for the distance calculations.

3.4.3 Comparing protein interaction predictions

For each phylogenetic tree based on each of the 12 organism sets, a matrix of cophenetic distances is calculated by adding the branch lengths separating the corresponding leaves. These distance matrices served as input for the three co-evolution-based predictors of protein-protein interactions under study.

Mirrortree methodology (MT) (section 3.1.1.1) was applied to the distance matrices obtained using the different selection of organisms. Every pairwise comparison was performed excluding protein pairs with less than 15 organisms in common or a *P*-value $> 1 \times 10^{-5}$. With the matrix of pairwise correlation coefficients fulfilling these thresholds, Profile Correlation (PC) (section 3.1.1.2) was calculated using the same parameters. Finally, the co-evolutionary profiles were compared under the Context-Mirror methodology (CM) (section 3.1.1.3), using a number of different third proteins for partial correlation calculation.

Considering a given subset of organisms, all pairs of proteins in the *E. coli* genome fulfilling the aforementioned requirements were ranked based on the scores of the three methods. We applied ROC analysis (section 3.1.3) to these lists to assess the capacity of the methods to separate positives and negatives on the basis of three different types of interactions (section 3.1.2): binary physical, co-membership in the same macromolecular complex or co-presence in the same metabolic pathway (Figure 3.3). Additionally, we evaluated the ability of the different predictors to recover the maximum number of positives with the best possible accuracy. For the different combinations of datasets, we quantified this tradeoff using the "F-measure" (section 3.1.3). The maximum "F-measure" was used to compare the different predictions assuming that, at this cutoff, the predictor displays its optimal performance.

3.5 Improving the significance of co-evolution detection

In this section, we introduce and evaluate a revisited version of the *P*-values associated to the correlation of distances in *mirrortree*-based approaches. In order to avoid some of the problems affecting the current way of evaluating tree similarity (section 1.7), we designed a new methodology denominated *p-mirrortree*, in which the correlation significance is reassessed by comparing the observed correlation with a null distribution of correlations obtained from the similarities of a large set of permuted phylogenetic trees. Consequently, tree similarities are re-scored using a *P*-value which takes into account the dependencies present in the set of phylogenetic trees under

study. To illustrate the problems associated to the previous versions of *mirrortree* and the improvement obtained by this new method, we compared the original and the new algorithm based on the "historical" sets of organisms available at different time points in the past. Finally, we evaluated the *p-mirrortree* ability to take advantage of the whole matrix of pairwise tree similarities, in a way similar to the PC method (section 3.1.1.2).

3.5.1 *p-mirrortree*

A new methodology denominated p-mirrortree (pMT) to evaluate protein co-evolution was introduced. The key point of this approach is the generation of a null distribution of tree similarities based on the observed distances in a background set of randomized phylogenetic trees.

The background set, which can contain all the phylogenetic trees in an organism or any subset of them, serves as reference to calculate the expected background distribution of tree similarities. Since the correlation coefficients between protein families composed by protein orthologs in many organisms are influenced by the number and characteristics of these organisms, the expected similarity needs to be evaluated in the context of the set of organisms shared by the trees. To build a reliable null model, all the pairwise combinations between phylogenetic trees on the reference set are split into groups based on the number of organisms in common, and a null model is derived for each group independently (Figure 3.4). 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 size groups can be used. For each one of these groups, a iterative process is carried out in order to obtain its corresponding null distribution of tree similarities (Figure 3.4). For each iteration, a pair of trees is randomly sampled with replacement and their distance matrices are retrieved from a pool of pre-calculated matrices of cophenetic distances. A pair of sub-matrices were extracted from the original matrices containing only the distances between sequences belonging to the organisms shared by both trees (Figure 3.4). The resulting sub-matrices are standardized by subtracting from each value their mean and dividing by their standard deviation. Once both matrices are in the same scale, the values corresponding to a given organism are switched between both families. Finally, the distance matrices are de-standardized using their original mean and standard deviation and completed by the distances between organisms present in the original matrices but not shared by the trees. The resulting matrices are returned to the pool in replacement of the original ones and are available for further iterations. Therefore, a single matrix can switch rows/columns multiple times with different matrices. After a number of iterations, the pool contains randomly generated distance matrices but always limited to the distance information available in other trees, reducing the space of possibilities to the ones that have already occurred (Figure 3.4). Finally, for each size group, all posible pairwise correlation coefficients are calculated based on the shuffled matrices, generating a background distribution for that size group.

Once these background distributions are calculated, the significance of a given *mirrortree* correlation coefficient based on a number of organisms in common can be evaluated. This result is

quantified by calculating the probability (*P*-value) of finding a coefficient higher than the observed one in the corresponding background distribution. A low *P*-value indicates a tree similarity much higher than those observed between shuffled trees with similar characteristics and, consequently, is indicative of a meaningful co-evolution.

3.5.2 Generating phylogenetic trees

Using the completely sequenced Eubacteria and Archaea present in the KEGG database (release 59.0 - August 2011) [219], we created phylogenetic trees for all *E. coli* proteins. Prokaryotic protein families with both paralogs and orthologs were retrieved for each protein using KEGG orthology groups. In order to select a single ortholog for each organism, we selected the sequences, which were best ranked against the corresponding *E. Coli* protein on the pre-calculated lists of "BLAST best bi-directional hits" stored in KEGG. The resulting protein families were then aligned by MUSCLE [214] using the default parameters. For each one of the 2,844 MSAs, a phylogenetic tree was created using the neighbor-joining algorithm implemented in TreeBeST [220].

3.5.3 Year-based selection of reference organisms

For each year, from 1995 to 2010, we created two different sets of reference organisms, "redundant" and "non-redundant". The "redundant" list of organisms contains the fully-sequenced Eubacteria and Archaea included in the KEGG database [219] in the corresponding year. A "non-redundant" set was obtained from it by removing the evolutionary close organisms. In order to evaluate which organisms are redundant, pairwise identities between ortholog sequences were calculated using the aforementioned MSAs (section 3.5.2). If two proteomes have more than 70% of the orthologous 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 the one with the lowest. The total number of organisms present in both datasets is shown in Figure 3.5. To assure a minimum number of 15 organisms in the "redundant" and "non-redundant" datasets, we focused on the period 2000–2010 for further analysis.

3.5.4 Year-based distance matrices

For each phylogenetic tree generated based on the available genomes in 2011 (section 3.5.2), a matrix of cophenetic distances was calculated by summing the lengths of the branches separating each pair of proteins. Depending on the size of the protein family, the size of these squared matrices range from 3 rows/columns to the total number of organisms in the corresponding reference set. The distance matrixes for each year were constructed by taking the rows/columns corresponding to the organisms in the redundant and non-redundat sets for that year. For comparative purposes, neither MSAs no trees are recalculated, so distances between a particular pair of sequences in a given protein family remain constant independently of the set of organisms used as reference. As a result, for each protein in the *E. coli* genome, year-based distance matrixes were obtained, based on the redundant or non-redundant sets of organisms available at that time.

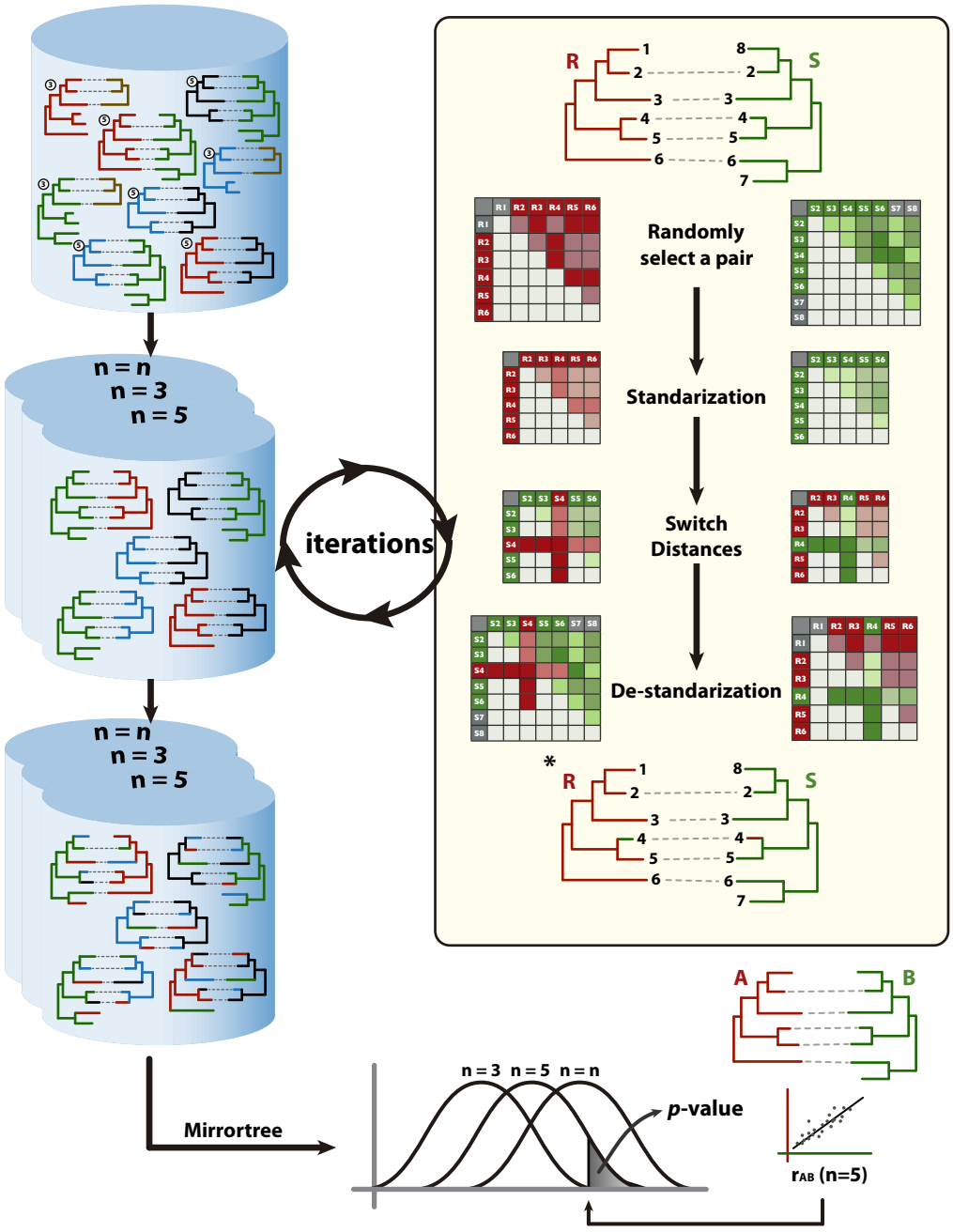

Figure 3.4: **Generating p-mirrortree null distributions**. In the first step all the pairwise combinations between phylogenetic trees are split into groups based on the number of organisms in common - red bubbles. For each group of pairs of trees a number of iterations of a distance swapping procedure are run in order to randomize the trees present in the set. In each iteration, a random pair of trees is selected and standardized based on the distances between sequences belonging to the organisms in common. Rows/columns with the distances belonging to the same organism are swapped between matrixes with a given probability. The resulting matrix 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. (*) The trees with the swapped branches are shown to illustrate the rationale of the approach, but all the process is applied to the distance matrixes only.

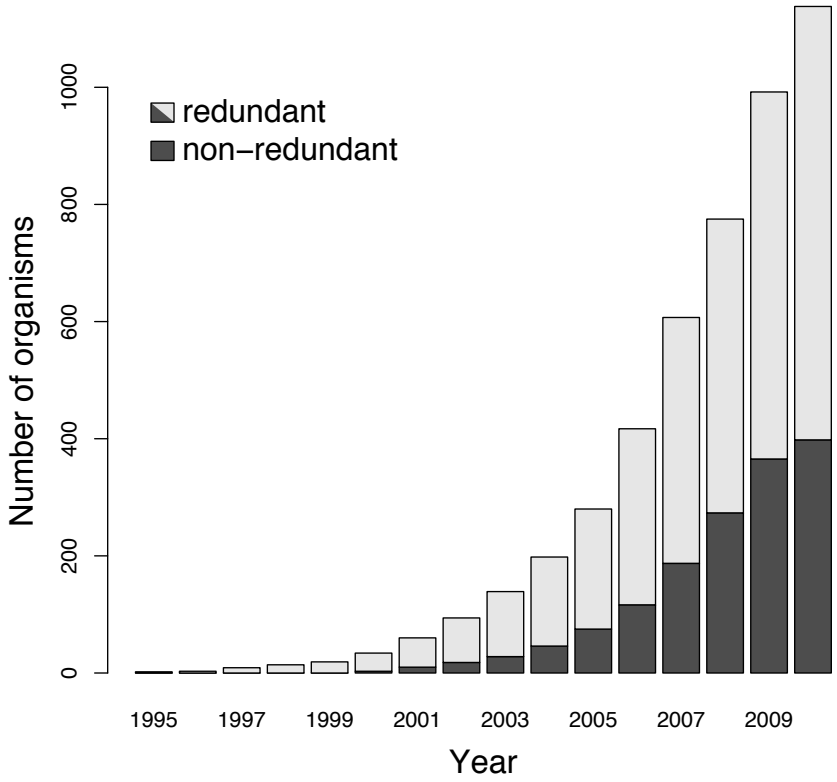

Figure 3.5: **Number of completely sequenced Eubacteria and Archaea available in KEGG for each year in the period 1995–2010.** In dark grey, we represent the number of organisms present in the "non-redundant" set, within the total number of "redundant" organisms represented by the whole bar.

3.5.5 Comparative performance analysis

The original *mirrortree* (section 3.1.1.1) and *p-mirrortree* (section 3.5.1) were applied to each of the "historical" sets of distance matrices (section 3.5.4). The matrices contained the distances between the protein sequences belonging to the available organisms over the years, considering two different redundancy criteria. For the *p-mirrortree*, we executed the algorithm creating a maximum of 40 intervals of number of organisms in common, and ran 1.000 permutation steps with a branch-switch probability of 0.05 (section 3.5.1). Both methods produced lists of putative interacting pairs ranked by their corresponding scores, correlation coefficient or *P*-value, respectively. Those pairs whose proteins are present in the same KO group were excluded to avoid artifacts caused by extremely similar trees.

In order to compare the performances within the same year (redundant/non-redundant and *mirrortree/p-mirrortree*), only those protein pairs present in the four results lists were considered, limiting the evaluation to the discriminant capacity of the scores. The resulting ranked lists were evaluated in the context of ROC analysis (section 3.1.3) using the three types of reference interaction sets previously described (section 3.1.2). Complementary evaluations were performed using only the pairs of trees with at least 15 and 30 organisms in common.

3.5.6 Context-based *p-mirrortree*

The same way the genome-wide matrix of pairwise *mirrortree* correlation coefficients is used by the PC method (section 3.1.1.2) to improve the prediction of interactions, we evaluated how a similar approach works with this new score (*p-mirrortree P*-value). Using the organisms dataset of 2010, we applied the PC method to the matrix of pairwise *mirrortree* scores, and a similar approach to the matrix with the new *P*-values (Figure 3.6). We evaluated both predictions based on the "Complexes" gold standard (section 3.1.2) using ROC analysis (section 3.1.3).

Additionally we introduced a method named Hierarchical Co-evolutionary Analysis (HCA) to explore the "co-evolutionary hierarchy" defined by the *P*-values. The distance between a pair of *p-mirrortree* coevolutionary profiles was defined as 1 minus the PC score previously defined. Using these distances, different clustering algorithms were applied. These include Ward's minimum variance, complete linkage, neighbor-joining and UPGMA. This generates a hierarchy of co-evolutionary relationships between *E. coli* proteins (Figure 3.6). The accuracies of the top-scoring results were calculated and compared with those obtained from PC using either correlation coefficients or *P*-values as input. Additionally, the biological meaning of some co-evolutionary groups showing up in this clustering is evaluated.

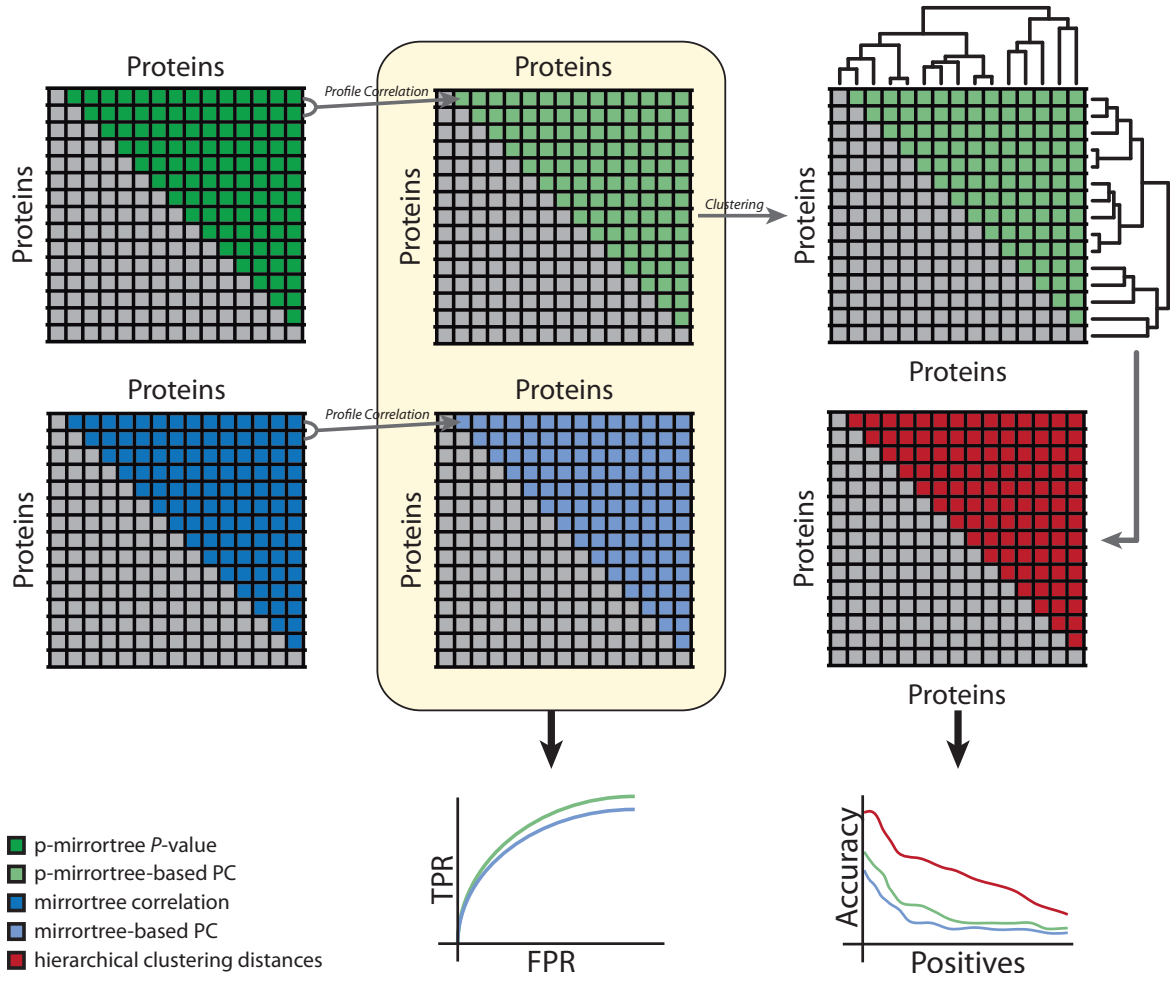

Figure 3.6: **Context methods based on mirrortree and p-mirrortree results**. Profile correlation were calculated using both genome-wide MT correlation coefficients (dark blue) and pMT *P*-values (dark green). The results of both methodologies (light blue and light green, respectively) were evaluated using the "Complexes" set (section 3.1.2) as gold standard. The True Positive Rates (TPR) and False Positive Rates (FPR) were drawn for the different possible thresholds in the context of ROC analysis (section 3.1.3). Moreover, a hierarchical clustering was applied to pMT PC results and the cophenetic distances of the resulting clustering were used as a new scoring schema. The performance of this score predicting interactions in the "Complexes" set (section 3.1.2) was evaluated within the top ranked results and compared with that of PC applied over MT correlations or pMT *P*-values.

subunit (DPO3E_ECOLI). If we attend to the results of the co-evolution analysis using both sets of organisms, we can appreciate significant differences. Whereas PC using “level 9” is able to produce significant scores for 306 pairs involving the alpha subunit, this number is considerably reduced to 128 when the “nearest 2” set is introduced. Moreover, the PC score calculated with the full set of organisms as reference (“level 9”) is 0.73, dropping to 0.57 when using only enterobacterias (“nearest 2”). As a direct consequence of the correlation drop, the proportion of false positives grows, negatively affecting the AUC, which decreases from 0.70 in “level 9” to 0.11 in “nearest 2”. We observe exactly the opposite behavior in the “recent” interactions where the “nearest 2” dataset performs better predicting interacting partners (Table 4.2).

4.4 Improving the detection of significant co-evolution

4.4.1 More insight on *p-mirrortree* null distributions

Previous versions of *mirrotree* either disregard the *P*-values associated to the correlation score or use tabulated ones, calculated analytically or derived from random sets of numbers not fulfilling the properties of tree-based distances. In order to get insight into the variations introduced by pMT and their *P*-values specifically derived for the genomic tree comparison problem, we compared the null distributions obtained by both approaches.

The pMT method introduces an additional step in the prediction of protein interactions: the calculation of the null distribution of tree similarities generated as a consequence of the permutation of a “background” set of trees. This process, described in section 3.5.1, creates a number of distributions of expected tree similarities for different intervals of number of organisms in common between the pairs of trees. Some of these distributions for the genomes available at different years are compared with the corresponding distributions of correlations between sets of random numbers (those tabulated and used in previous versions of mirrortree)(Figure 4.7). As expected, the average correlation coefficient of random numbers is always 0. Moreover, as soon as the sets of numbers being correlated are larger, the probability of obtaining “extreme correlations”, either positive or negative, decreases. However, some of these general observations are not extended to the correlation coefficients calculated using distance matrixes of permuted trees. As described in different studies, phylogenetic trees tend to share a background similarity and, consequently, the distribution of correlations is always shifted to higher values. Besides, the expected correlation distributions highly depend on the number of organisms shared by a pair of trees. Figure 4.7 shows that pairs of trees sharing a small set of organisms present a wider range of correlation coefficients, whereas pairs of trees sharing many orthologs present correlations in a narrower range. Contrary to the random numbers, these distributions are not centered in the same value, so the null distribution of correlation coefficients varies depending on the number of organisms in common. Therefore, the probability of obtaining a given correlation coefficient varies depending on the distribution of organisms used to generate the trees.

Here we used the number of organisms in common between a pair of phylogenetic trees in a given set of organisms as a proxy for the set of organisms under comparison. Since the *mirrortree*-based approaches are usually focused on predicting protein interactions in a reference organism, it is expected that the number of organisms in common reflects the conservation of these families along the tree of life. However, this number is highly dependent on the total number of organisms used to generate the phylogenetic trees. Indeed, we need to be aware that a given correlation coefficient calculated with 10 organisms in common has a completely different meaning if the phylogenetic trees are generated using the 38 "non-redundant" organisms available in 2001 or the 335 of 2010 (Figure 4.7). This condition, previously ignored by the tabulated *P*-values, is addressed by pMT, which independently constructs a null distribution for each number of possible organisms in common between the pairs of phylogenetic trees.

4.4.2 Historical assesment of p-mirrortree predictions

We compared the performance of p-mirrortree (pMT) with that of the original mirrortree (MT) in predicting different types of interactions using the set of organisms available in the period 2000–2010, as well as non-redundant versions of these sets (section 3.5).

The performance of MT and pMT predicting two different types of physical interactions are shown in Figure 4.8. The clearest observation is that both methods based on similarities of phylogenetic trees are able to capture part of the co-evolutionary signal related to physical protein interactions. The general trends observed suggest that protein interaction predictions have benefited from the increase in the number of sequenced genomes during the last decade. Indeed, there is no clear evidence suggesting these trends have reach a plateau, so further improvement can be expected over the next few years. Similarly to previous studies, interactions defined as co-membership in the same macromolecular complex present the highest AUCs, followed by binary physical interactions. Interactions based on co-membership in the same metabolic pathway present poor and constant AUCs (Figure S8), suggesting that co-evolution may not be a generalized process between the proteins of the same pathways.

The performance of pMT when predicting physical interactions using the organisms available during the last 10 years is higher than that of MT ($\sim$0.10–0.15 increase of AUC) (Figure 4.8). This improvement is obtained 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.

Previous results suggested that MT predictions are negatively affected by redundant taxa (section 4.3). To obtain further insight into this, we compared MT and pMT performances using the aforementioned year-based lists of organisms against a subset of the same organisms from which redundancy was removed. In Figure 4.8, we observe that MT performances benefit from the usage of the "non-redundant" sets, supporting previous evidences. Indeed, the gap in AUC between "redundant" and "non-redundant" sets becomes bigger as the taxonomical redundancy increases over the years. On the other hand, pMT presents a higher robustness to redundancy. Although "non-redundant" sets achieve slightly better performances, this improvement seems to be constant

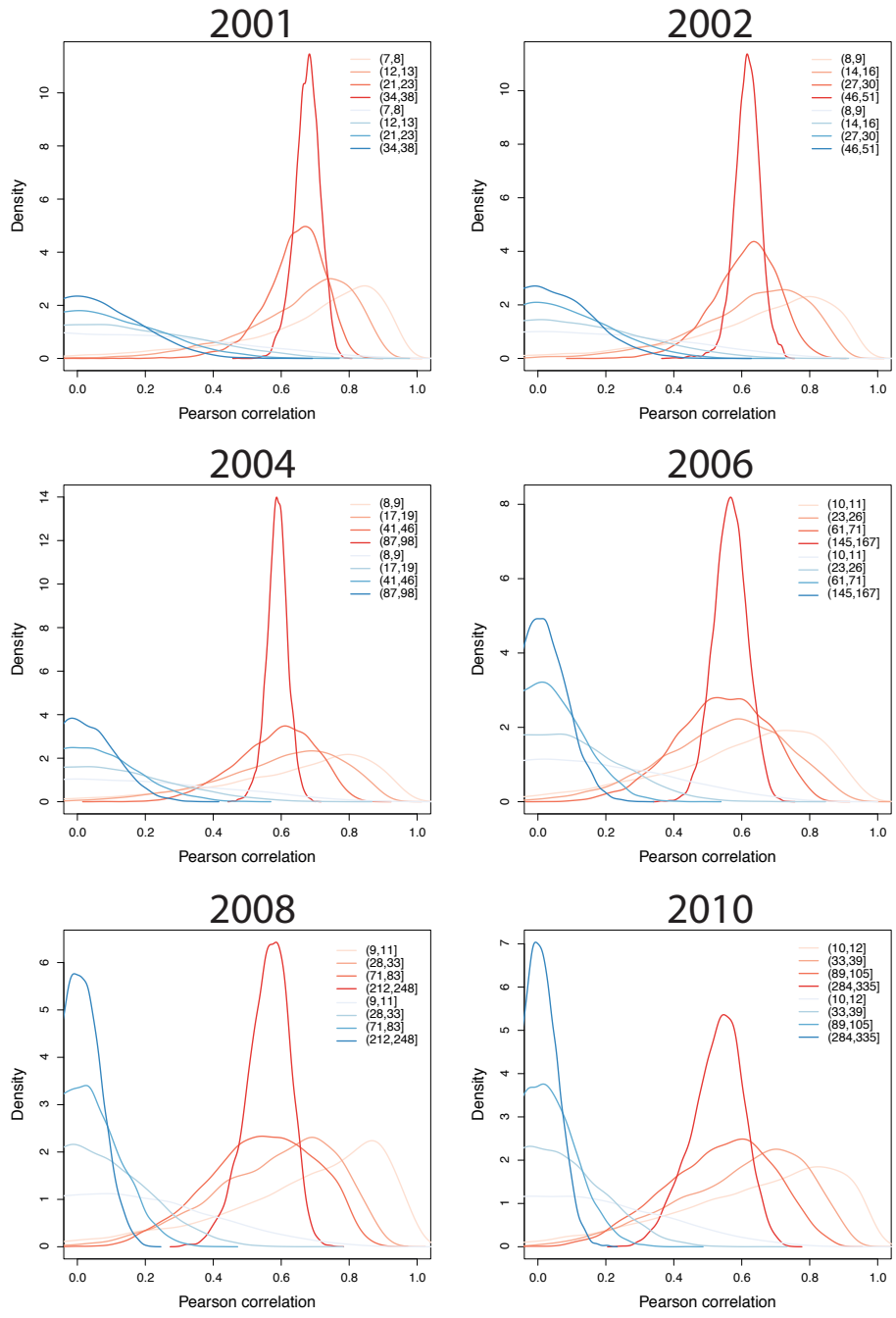

Figure 4.7: **Density functions for the distribution of expected correlation coefficients in sets of random pairs of numbers and sets of distances extracted from pairs of permuted phylogenetic trees**. The genomes available at different time points in the past were used as reference to generate shuffled 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).

over the years, indicating that it could be independent of the growing redundancy and most likely related with the methodological setup.

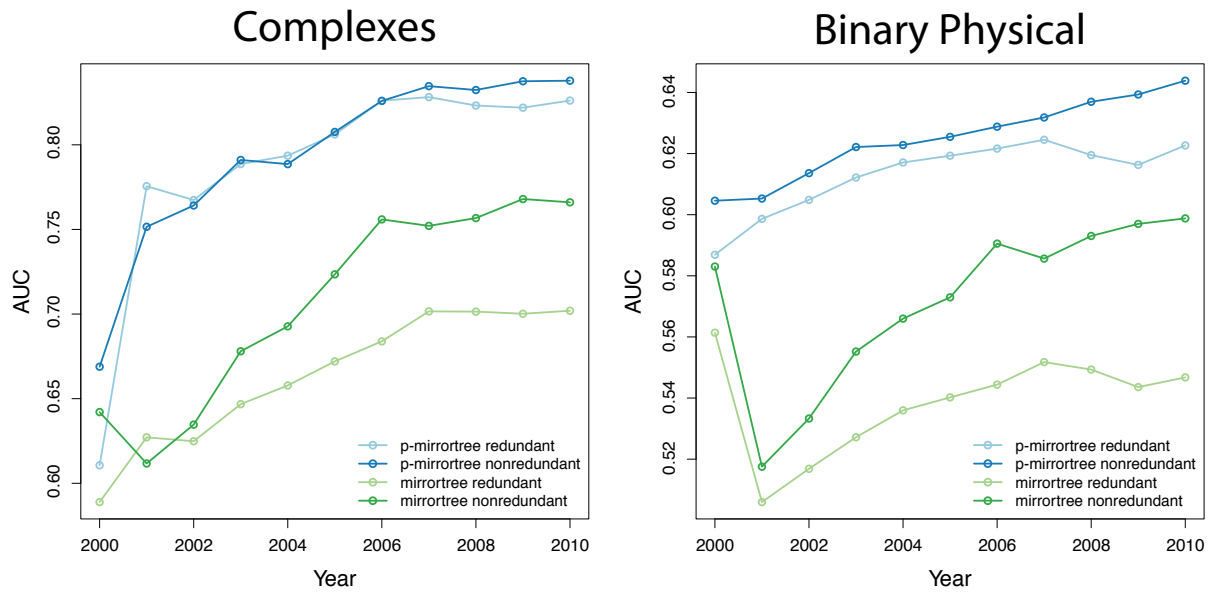

Figure 4.8: **Performance of the MT and pMT methods when predicting physical interactions using different sets of organisms based on the fully-sequenced genomes available in the period 2000–2010 (with and without redundancy)**. The performances were evaluated in terms of AUC using two gold standard datasets of protein interactions: binary physical and co-membership in the same macromolecular complex.

As a consequence of the incomplete *P*-values previously used in MT implementations, some workarounds have been proposed in order to improve global performance. The most common approach is to ignore those protein pairs below a given number of organisms in common. We benchmarked this workaround using MT and pMT to predict proteins in the same macromolecular complex. The same "historical" sets of organisms were evaluated excluding those pairs with less or equal than 15 and 30 organisms (Figure 4.9).

Although MT predictions show better AUCs than those obtained using all the predictions (section 4.3), the performance starts to drop drastically at a certain number of sequenced organisms for both "redundant" and "non-redundant" sets. The maximum performance is different depending on the number of minimum organisms and also with the set of organisms used to construct the phylogenetic trees. For example, using all the sequenced organisms available, the optimal performance of MT was reached in 2003 using tree pairs with more than 15 organisms in common and in 2006 when pairs with more than 30 organisms in common were considered. Disregarding the obvious loss in prediction coverage, as more organisms are used to generate the trees, pairs with larger number of organisms in common need to be excluded in order to get the optimal performance. Even on those maximum values of AUC, MT never outperforms the more stable pMT.

Certain limitations arise as a consequence of all of the above. Firstly, the difficulties to assign the proper threshold of minimum organisms in common necessary to consider a pair of trees constructed with a given set of organisms limits the applicability of MT methodology. As it has

been pointed out, this threshold varies depending on unknown factors related with the size and redundancy of the set of organisms, and thus it remains difficult to define as a single recipe. Secondly, even if the adequate threshold is detected, predicting less protein pairs have a strong negative impact since it can bias the pairs towards proteins conserved in large number of organisms. The drawbacks of excluding protein pairs are not limited to the performance evaluation. Having less pairs evaluated is also critical in those methods, such as PC or CM, that benefit from the whole coevolutionary network to predict single interactions.

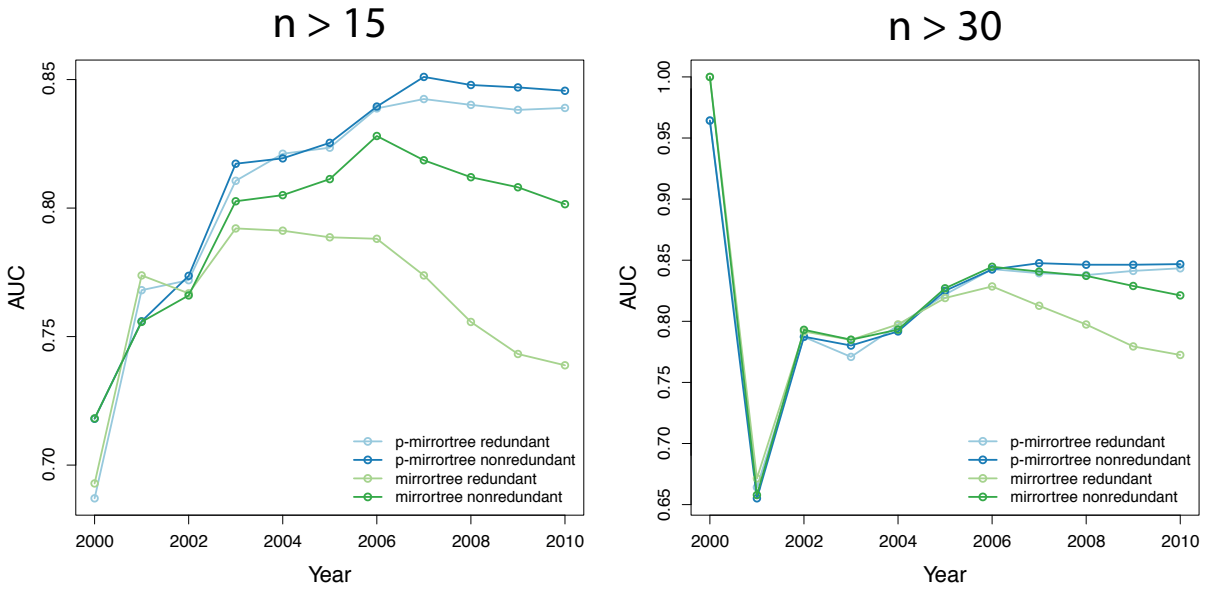

Figure 4.9: **Performance of the MT and pMT methods when predicting proteins in the same macromolecular complex using different sets of organisms based on the fully-sequenced genomes available in the period 2000–2010 (with and without redundancy)**. Their performance was evaluated in terms of AUC. Those pairs of trees with less or equal than 15 or 30 organisms in common were excluded for evaluation.

4.4.3 Context-based *p-mirrortree*

A novel generation of methods succeeded in using the whole matrix of *mirrortree* pairwise correlation coefficients to predict single interactions (PC [163], section 3.1.1.2; and CM, section 3.1.1.3). In order to evaluate the applicability of *p-mirrortree* results to context-based methodologies, we benchmarked PC based on MT correlation coefficients and based on *P*-values. In Figure 4.10, we observe that the PC calculated using *P*-values outperforms the original PC implementation which used correlation coefficients. Considering that the accuracy of pMT are higher than those of MT (section 4.4.2), the improvement in PC predictions can simply arise as a consequence of this.

One of the benefits of these context-based methodologies is the significant reduction in the number of false positives. A new approach, named Hierarchical Co-evolutionary Analysis (HCA), was introduced in order to reassess the similarity between a pair of co-evolutionary profiles. In this approach, a hierarchal clustering was applied over the whole list of co-evolutionary profiles

(section 4.4.3). The purpose of this was to re-score the candidate pairs based on the cophenetic distances in the resulting clustering. By using this score, the enrichment in positive interactions among the first ranked pairs is much larger than using the PC method based on either correlation coefficients or *P*-values (Figure 4.10). Different algorithms for hierarchical clustering showed similar results (Figure S9).

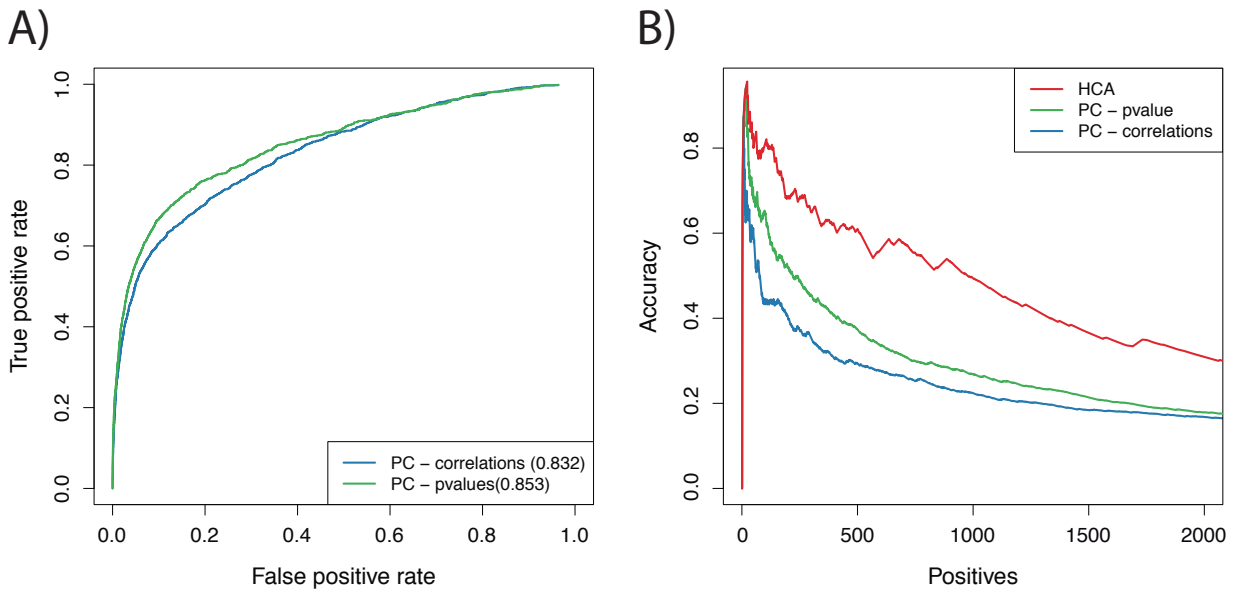

Figure 4.10: **Performances of context-based approaches to predict protein interactions in the "Complexes" gold standard using mirrortree and p-mirrortree scores. A)** ROC curves for PC based on *mirrortree* correlation coefficients - blue - and *p-mirrortree* *P*-values - red. The AUCs of both curves are shown within brackets. **B)** Accuracy *vs.* number of positives in three protein interaction predictions based on PC scores using the matrixes of pairwise correlations - blue - and *P*-values - red -, and cophenetic distances when applying a hierarchal clustering based on Ward's minimum variance over the whole list of co-evolutionary profiles - red.

Apart from its higher performance, an additional advantage to this analysis is that the clustering describes the "co-evolutionary relationships" between a set of proteins in a hierarchical representation, which might be used to infer additional information on the substructure and functioning of the interactome. For example, if we analyze the hierarchical coevolutionary relationships of the *E. coli* ATP synthase (Figure 4.11), we observe how the different members of the complex form clusters, represented as a tree, ranging from the most similar pairs of coevolutionary profiles to the cluster including all the proteins. At each intermediate cut of the tree we can split the proteins into different groups, based on the distances between their coevolutionary profiles. In the case of the ATP synthase, if we cut the tree into three different clusters, we observe a cluster containing the "a" and "c" subunits, a second cluster formed uniquely by the subunit "b" and a third cluster containing the 5 different members of the F_1 particle. These results are compatible with the three-dimensional model of the ATP synthase, in which the "a" and "c" subunits are embedded in the membrane to create the proton pore, the F_1 particle is the cytosolic machinery in charge of the ADP phosphorilation and the subunit "b" connects both sub-complexes (Figure 4.11). Consequently, this coevolutionary analysis generated some clues on the architecture of the macromolecular complex,

only using information at sequence level.

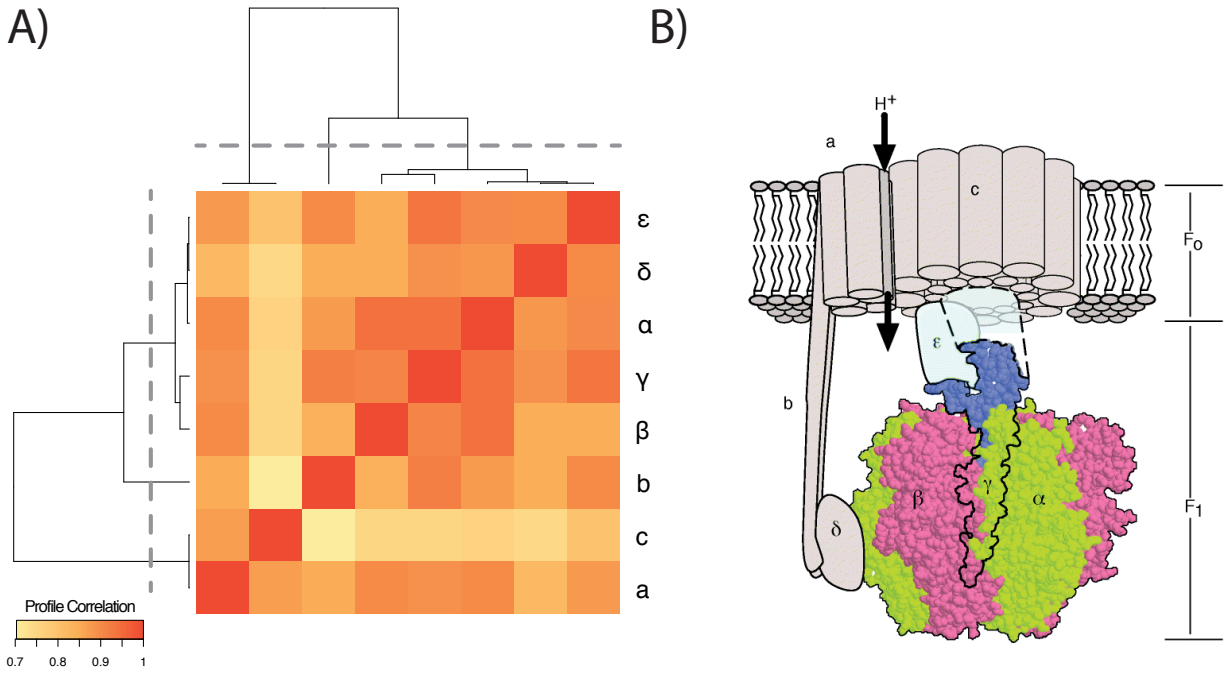

Figure 4.11: **Hierarchical Coevolutionary Analysis (HCA) of the 8 subunits of the *E. coli* ATP synthase**. **A)** Heat map representing the pairwise PC between the coevolutionary profiles of the ATP synthase subunits. The hierarchical clustering is calculated using the Ward's minimum variance algorithm over the pairwise PC results. The clustering is therefore based on the full matrix of pairwise similarities and not only on the similarities shown in the heat map. **B)** Three-dimensional representation of the E.coli ATP synthase based on the model of Wang and Oster [223].