Chapter 9

Using Phylogenetic Profiles to Predict Functional Relationships

Matteo Pellegrini

Abstract

Phylogenetic profiling involves the comparison of phylogenetic data across gene families. It is possible to construct phylogenetic trees, or related data structures, for specific gene families using a wide variety of tools and approaches. Phylogenetic profiling involves the comparison of this data to determine which families have correlated or coupled evolution. The underlying assumption is that in certain cases these couplings may allow us to infer that the two families are functionally related: that is their function in the cell is coupled. Although this technique can be applied to noncoding genes, it is more commonly used to assess the function of protein coding genes. Examples of proteins that are functionally related include subunits of protein complexes, or enzymes that perform consecutive steps along biochemical pathways. We hypothesize the deletion of one of the families from a genome would then indirectly affect the function of the other.

Dozens of different implementations of the phylogenetic profiling technique have been developed over the past decade. These range from the first simple approaches that describe phylogenetic profiles as binary vectors to the most complex ones that attempt to model to the coevolution of protein families on a phylogenetic tree. We discuss a set of these implementations and present the software and databases that are available to perform phylogenetic profiling.

Key words: Phylogenetic profiles, Coevolution, Functional associations, Comparative genomics, Coevolving proteins

1. Introduction

The remarkable improvements in sequencing technology that have occurred over the past few decades have made the sequencing of genomes an ever more routine task. To date about 1,000 bacterial genomes have been sequenced along with dozens of eukaryotic ones. Along with the genome sequences themselves, annotation efforts have also progressed so that most new genome sequences are accompanied by detailed descriptions of the positions of the genes encoded within the genome, and the functions of

Jacques van Helden et al. (eds.), *Bacterial Molecular Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 804, DOI 10.1007/978-1-61779-361-5_9,

the proteins and noncoding RNAs that are encoded in the genes. One of the fundamental challenges for computational biologists and bioinformaticians is the inference of interaction networks between these genes that enhance our ability to understand the function of the gene products. To this end, here we discuss the phylogenetic profiling technique, and its uses for probing functional association between proteins.

Sequence homology is the primary tool used to assign a function to a protein. If two proteins have significant similarities between their sequences, then they likely descended from a common ancestor and share a common function. As our experimental knowledge of protein functions increases, this approach allows us to pass functional annotation from a characterized protein to its uncharacterized homologs, and thus annotate an ever-growing number of sequenced proteins. Nonetheless, gaps in our knowledge still remain, which lead to lack of any meaningful functional annotation for many protein families. The function of these orphan proteins may not be studied using homology-based computational approaches, and therefore a different class of nonhomology algorithms must be used in these cases. Among these approaches, phylogenetic profiling is one of the primary tools.

As genomes evolve, genes are deleted or are horizontally transferred from one species to another. The intermingling of genetic material between species, which is particularly common among bacteria, makes the reconstruction of species phylogenetic trees very challenging. At the same time, the abundance of genetic exchanges between organisms offers a unique opportunity to study the coupling of genes within genomes. If at the simplest level we view genomes as bags of genes, and we have access to hundreds of bags, then we can begin to identify pair-wise and higher associations between the elements of the bags. In other words, if genes are often transferred between organisms, we can identify which sets of genes appear to transfer together. Genes that are coupled in this manner are necessarily present or absent within the same organisms, and it is therefore unlikely to find one without the other. The identification of these couplings allows us to infer that the products of the two genes likely function together to achieve a common biological function. They may be subunits of a protein complex, or sequential steps in metabolic pathways. In either case, an organism needs both genes to carry out their function, and having only one of them likely decreases its fitness.

The search for co-occurring protein families across organisms is often referred to as phylogenetic profiling. Here, we review various implementations of phylogenetic profiling, and also discuss databases that use this approach to study protein function. Finally, we discuss extensions of phylogenetic profiling that involve higher order associations between genes.

2. Structure of Phylogenetic Profiles

The first implementation of phylogenetic profiles consisted of binary vectors that captured the presence or absence of homologs of a reference protein across organisms (1). To construct these profiles, first a reference genome was selected (e.g., *E. coli*). Each *E. coli* protein was then aligned to the proteome of each fully sequenced bacteria using BLAST (2). If a hit to one of the proteins in another organism is identified by a significant BLAST threshold (e.g., $E < 1e\text{-}6$), then a 1 is inserted in the corresponding position of that organism in the phylogenetic profile. If no significant hit is found, then a zero is inserted. Using this approach, the entire length of the phylogenetic profile vector is populated. A schematic representation of the construction of phylogenetic profiles is shown in Fig. 1.

When using reference genomes to construct phylogenetic profiles, one generates a separate phylogenetic profile matrix for each organism. However, instead of computing the homologs between a reference genome and other organisms, it is also possible to use orthologous protein families to populate a general phylogenetic profile that does not require a reference genome. The definitions of orthologous protein families vary across implementations, but one example is that provided by the clusters of orthologous groups (COGs) database (3, 4).

Representations of phylogenetic profiles other than binary vectors have also been developed. For example, in one case,

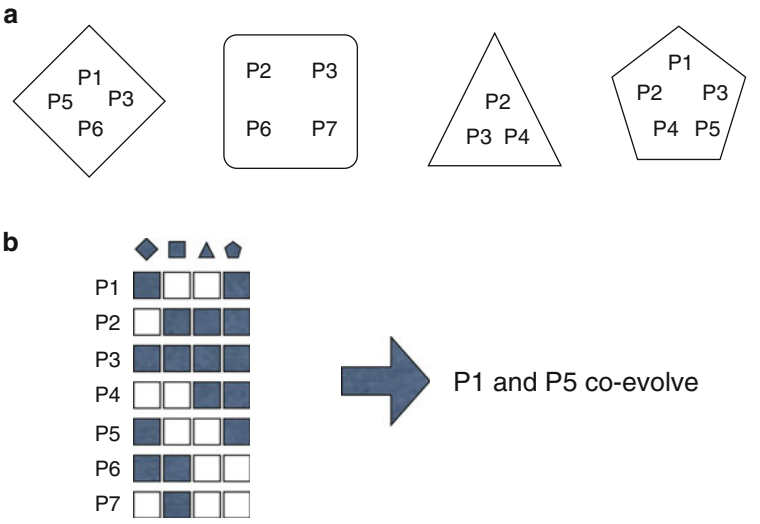

Fig. 1. Schematic representation of the phylogenetic profile method. In panel (**a**), we see four sample genomes with their respective proteins. Panel (**b**) shows phylogenetic profiles that capture the presence or absence of these protein families across genomes. We note that protein family one and five have identical profiles, and are thus coevolving.

instead of a binary value, the logarithm of the BLAST expectation score is used (5). Another approach used matrices instead of vectors to represent each profile (6). The entries of the matrix are the evolutionary distances between orthologous proteins, and the matrix has dimension $N \times N$, where N is the number of organisms being considered, and one would thus generate $N \times N \times P$ profiles for a genomes, where P is the number of protein families. These distance matrices may be compared by computing an internal product that measures the similarity of the corresponding cells across two matrices.

3. Metrics for Comparisons Between Phylogenetic Profiles

Once the phylogenetic profiles have been constructed, the next step consists of systematically comparing all pairs of profiles to detect protein families that have coupled evolutionary patterns. The first approach developed to compare profiles used the Hamming distance, or simply the number of positions where two binary profiles have different values (1). However, this metric does not provide a statistical framework for evaluating the likelihood of observing a specific Hamming distances. To this end, subsequent approaches used the hypergeometric distribution to estimate the probability of observing a certain number, or greater matches between two profiles (7). This metric was further refined in other work to account for the different proteome sizes of the organisms that are used to populate the phylogenetic profiles. A weighted hypergeometric distribution P value was developed to address this limitation (8).

When nonbinary phylogenetic profiles are constructed, then other metrics have to be used. Various studies have used mutual information to measure the similarities between profile vectors (5).

4. Accounting for the Phylogenetic Tree of Organisms

Since the goal of phylogenetic profiling is the identification of protein families that have coevolved, it is important to account for the underlying phylogeny of the organisms used to construct the profiles. The tree of organisms may be used to infer what loss or acquisition events explain the profile of a specific protein family. When comparing two profiles, it is then possible to use parsimony to estimate the smallest number of differences that explain the evolution of two protein families (9). This approach allows us to separate pairs of profiles that have identical scores when non-tree-based metrics, such as the ones described in the previous section, are used.

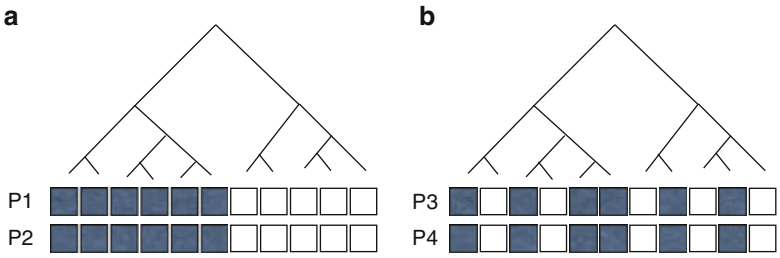

Fig. 2. Tree-based analysis of phylogenetic profiles allows us to identify the most significant coevolutionary events. In panel (**a**), we see two protein families that co-occur across genomes, but are only found in one branch of the species tree, thus we can explain the pattern with a single loss event at the right branch of the first bifurcation, or a single gain in the left branch of the first bifurcation. In contrast, panel (**b**) shows a more interesting co-occurrence pattern that arose due to multiple coordinated loss events across the species surveyed. We conclude that P3 and P4 are more likely coevolving, and hence functionally related, than P1 and P2.

The simple example shown in Fig. 2 illustrates the advantage of considering the underlying tree of organisms when comparing two polygenetic profiles. In panel A, we see two profiles that are identical, but can be explained by the acquisition of the protein in one branch of the tree. In contrast, in panel B we see two proteins that have repeatedly been lost and acquired in multiple branches of the tree. Our hypothesis is that the latter pair is far more likely to be functionally coupled than the former, since their evolutionary pattern requires more coordinated loss events.

Pagel et al. provided a formal model for the estimation of the likelihood of coevolution of two protein families on a tree of organisms (10, 11). They explicitly estimate the coupling parameters of two protein families and are thus able to derive an estimate of their coevolution. However, this formalism requires significant computational resources compared to the simpler metrics.

Recently, we introduced an intermediate approach that partially accounts for the tree of organisms, but is able to do so more efficiently than the Pagel et al. approach (12). This method does not consider the full tree of organisms, but only the ordering of organisms within the tree. Dynamic programming may be used to estimate the optimal ordering or organisms on a tree, by minimizing the distance between adjacent organisms represented by the leaves of the tree (13). Armed with this information, it is possible to implement a simple probabilistic model for the likelihood of observing runs of matching ones in two binary profiles. This allows us to separate pairs of profiles that have matches in only one branch of the tree, from those that are co-occurring in diverse branches of the tree. It also allows us to partially compensate for the fact that certain branches are more populated than others in cases, where a protein family is conserved across the entire branch,

since both large and small branches would contribute a single run of occurrences.

The simplest approach to partially account for the tree of organisms when comparing phylogenetic profiles is to prune the tree so that each group of related organisms is represented by a single individual (14). Of course, there is some ambiguity associated with the definition of a related group of organisms, as this must be based on a somewhat arbitrary level of taxonomic similarity. However, once this parameter has been established, this method reduces the biases in the analysis that may arise from the over-representation of one group with respect to another, but does not explicitly correct the remaining biases that are addressed by the two previous approaches.

5. Assessing the Functional Relationships of Coevolving Protein Families

Our hypothesis is that coevolving protein families likely share a related function. A variety of methods have been developed to test this theory. The simplest is to use existing pathway annotation, such as that provided by the Gene Ontology Consortium (15). This annotation allows us to group together proteins that act within the same biological process. Examples of biological processes are the cell cycle, specific metabolic pathways or large protein complexes, such as the flagella. In order to demonstrate that coevolving pairs of protein are functionally related, we simply demonstrate that these pairs are more often members of the same biological process than random pairs. If pairs of proteins within the same process are considered true positives, and those in different processes true negatives, then using a receiver operator characteristic (ROC) curve allows us to measure the enrichment for true positives versus false positives in a ranked list. These types of analyses also enable the comparison of different approaches for the identification of phylogenetic profile pairs, as the methods that yields the largest area under the ROC curve performs best (12, 16).

Beyond the demonstration that phylogenetic profiling allows one to identify functionally related proteins, it is also possible to use this approach to assign a function to a previously uncharacterized protein family. That is, if nothing is known about the function of protein family A, but using phylogenetic profiling one identifies that it is coevolving with functionally characterized protein family B, then it is likely that the function of A is similar to that of B. A recent review demonstrates that over the years a variety of examples of this “guilt by association” technique have been used to predict and then verify the functions of previously uncharacterized protein families (17).

6. Higher-Order Relationships Among Phylogenetic Profiles

The analysis discussed so far has been restricted to the identification of pair-wise relationships between phylogenetic profiles. However, it is also possible that higher order relationships may be identified. Here, we present two examples of approaches that have been used to identify relationships between triplets and larger groups of phylogenetic profiles.

The first approach searches for logic combinations of pairs of triplets that match a third profile (4). For example, protein family C may be present across organisms only when both protein families A and B are present (see Fig. 3). Thus, C is not correlated with the presence of A or B, but is only correlated with the combined presence of families A and B. Bowers et al. developed an implementation of this approach and were able to uncover a large number of examples of logic triplets. They analyzed all eight possible logic relationships and developed a metric for scoring the significance of logic triplets.

A second method that uses higher-order relationships between phylogenetic profiles was developed to identify protein complexes that are duplicated within genomes (18). In this approach, phylogenetic profiles are first compared to identify significant pair-wise relationships. This leads to the creation of a $P \times P$ binary matrix, where P is the number of profiles, whose entries are one if the corresponding (i,j) pair is found to be significantly related and zero otherwise. This matrix is then clustered so that groups of profiles that are significantly related form clusters along the matrix diagonal. These clusters might correspond to components of pathways or protein complexes.

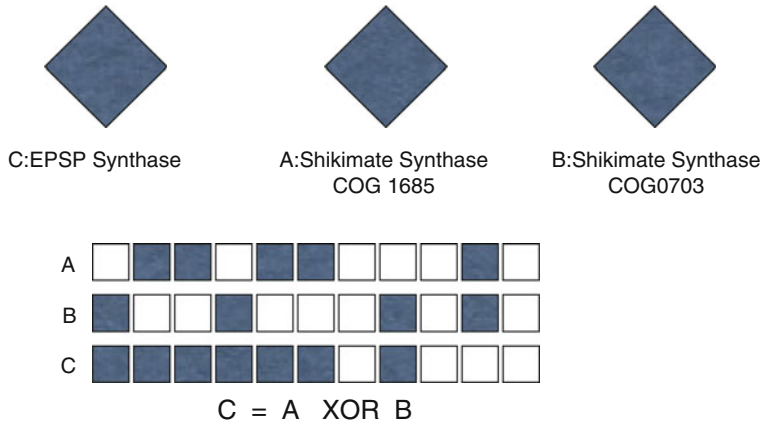

Fig. 3. Logic analysis of phylogenetic profiles. Protein family C (EPSP synthase) is only present in genomes that contain either protein families A (COG1685 Shikimate Synthase) or B (COG0703 Shikimate Synthase), but not both (exclusive or, XOR).

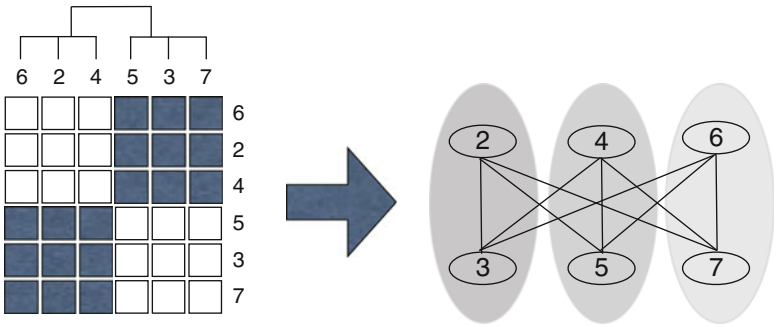

Fig. 4. Phylogenetic profiles may be used to identify duplicated complexes. *On the left*, we see a matrix of significant pair-wise relationships between the phylogenetic profiles of the two corresponding proteins (*black indicates significance and white lack of significance*). These proteins are clustered and ordered in the same manner on the *x* and *y* axes. A block of significant relationships occurs off the matrix diagonal (*upper left to lower right*). This pattern is due to coevolutionary relationships between the subunits of the three complexes shown on the right. The 2 subunits of the blue complex coevolves with the 3, 5, and 7 subunits, which are all homologous to each other. The 4 and 6 subunits show identical co-occurrence patterns. By using additional criteria, such as the presence within an operon, it is possible to separate the three complexes from each other.

Alternatively, one can also cluster the nonbinary version of the profile similarity matrix, whose entries represent the distance between two profiles.

Using this approach, Li et al. also identified off-diagonal clusters. These arise when protein complexes are duplicated within the genome. Consider, for example, a complex that is formed of two subunits (2 and 3), and two other homologous complexes which are formed of two subunits (4 and 5, 6 and 7), as illustrated in Fig. 4. We assume that subunit 2 is homologous to 3, 5, and 7. In this case, the phylogenetic profile of protein 2 would be found to coevolve with proteins 3, 5, and 7. Similarly, protein 3 coevolves with 2, 4, and 6. This pattern of relationships leads to an off-diagonal cluster that contains the pair-wise interactions between the six subunits. Separating these complexes using additional information, such as co-occurrence within operons, allows one to identify the correct association between the subunits. When applied to the genome of *R. palustris*, Li et al. used this approach to identify a large off-diagonal cluster that arose from the duplication of a nitrogenase complex.

7. Available Software

Several databases are available that utilize phylogenetic profiles for functional annotation or network reconstruction. The most widely used is the STRING database (19). This database is constructed using COGs and implements a continuous version of

phylogenetic profiles, whose pair-wise similarity is evaluated using mutual information. The underlying phylogenetic tree of organisms is accounted for by collapsing into a single node those taxa in which the presence or absence of a specific gene pair is in agreement in all the species.

Another implementation of phylogenetic profiles is found in the Prolinks database (20). Here, orthology is determined using a simple BLAST cut-off criterion, and a reference organism must be specified. The phylogenetic profiles are encoded as binary vectors and their similarity is determined using the hypergeometric distribution probability. In this implementation, the phylogenetic relationships between organisms are not accounted for, and each genome is treated as an independent random variable (and therefore some closely related genomes can lead to a bias in the analysis). However, future implementations will also include the "runs" correction discussed above to partially compensate for this effect. The PLEX (21) and ViSANT (22) databases implement phylogenetic profiling in a very similar manner.

8. Strength and Pitfalls

As we have discussed, the phylogenetic profile approach treats genomes as bags of genes, and does not consider the order of genes within the genome. This is clearly an oversimplification, and a variety of methods have been developed to search for pairs or larger groups of genes that maintain their proximity across varied organisms (23). Because transcriptional units in bacteria are operons and not individual genes, the evolutionarily conserved sets of genes typically correspond to operons, or fragments of operons. In fact, the conservation of proximity of groups of genes is a strong indication that the group is a part of a bacterial operon.

Many of the databases we have discussed above incorporate not only phylogenetic profiles in their analyses, but also some form of gene proximity conservation measure. Since the pairwise predictions produced by these two approaches do not always overlap, they may be combined using Bayesian techniques to arrive at more accurate coevolutionary networks.

9. Perspectives

We have already seen that as the number of organisms in phylogenetic profiles grew from tens to hundreds, the ability of this technique to identify biologically interesting coevolution events has dramatically increased. We expect that these trends will continue as the number of genomes continues to increase. However,

along with these improvements the challenge will be the development of efficient implementations of the approach that can handle the ever-growing number of organisms with fully sequenced genomes. The all versus all alignments of proteins across thousands of genomes is a very computationally intensive task. Furthermore, all the approaches that incorporate the tree of organisms when computing the significance of two profiles are also computationally demanding. Over the next few years, these methodologies will need to be optimized to handle the thousands of bacterial genomes that will be sequenced. These increases in efficiency will undoubtedly be accompanied by ever more useful functional predictions.

Acknowledgments

The author wishes to acknowledge the UCLA-DOE Institute for Genomics and proteomics for support.

References

1. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A, 96:4285–4288.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25:3389–3402.
3. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 4:41.
4. Bowers PM, Cokus SJ, Eisenberg D, Yeates TO. (2004) Use of logic relationships to decipher protein network organization. Science, 306:2246–2249.
5. Date SV, Marcotte EM. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnol, 21:1055–1062.
6. Pazos F, Valencia A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng, 14:609–614.
7. Wu J, Kasif S, DeLisi C. (2003) Identification of functional links between genes using phylogenetic profiles. Bioinformatics, 19: 1524–1530.
8. Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM. (2006) Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics, 7:177.
9. Liberles DA. (2001) Evaluation of methods for determination of a reconstructed history of gene sequence evolution. Mol Biol Evol, 18:2040–2047.
10. Barker D, Pagel M. (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol, 1:e3.
11. Barker D, Meade A, Pagel M. (2007) Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. Bioinformatics, 23:14–20.
12. Cokus S, Mizutani S, Pellegrini M. (2007) An improved method for identifying functionally linked proteins using phylogenetic profiles. BMC Bioinformatics, 8(Suppl 4):S7.
13. Bar-Joseph Z, Gifford DK, Jaakkola TS. (2001) Fast optimal leaf ordering for hierarchical clustering. Bioinformatics, 17(Suppl 1): S22–S29.
14. Sun J, Li Y, Zhao Z. (2007) Phylogenetic profiles for the prediction of protein-protein interactions: how to select reference

organisms? Biochem Biophys Res Commun, 353:985–991.

15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25:25–29.
16. Jothi R, Przytycka TM, Aravind L. (2007) Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. BMC Bioinformatics, 8:173.
17. Kensche PR, van Noort V, Dutilh BE, Huynen MA. (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface, 5:151–170.
18. Li H, Pellegrini M, Eisenberg D. (2005) Detection of parallel functional modules by comparative analysis of genome sequences. Nat Biotechnol, 23:253–260.
19. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. (2009) STRING 8 – a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res, 37: D412–D416.
20. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol, 5:R35.
21. Date SV, Marcotte EM. (2005) Protein function prediction using the Protein Link EXplorer (PLEX). Bioinformatics, 21:2558–2559.
22. Hu Z, Hung JH, Wang Y, Chang YC, Huang CL, Huyck M, DeLisi C. (2009) VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res, 37:W115–W121.
23. Dandekar T, Snel B, Huynen M, Bork P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci, 23:324–328.