FEBS Letters 584 (2010) 4237–4240

journal homepage: www.FEBSLetters.org

Coevolution is a short-distance force at the protein interaction level and correlates with the modular organization of protein networks

Zhi Liang [a,*,1], Meng Xu [a,1], Maikun Teng [a], Liwen Niu [a], Jiarui Wu [a,b,**]

[a] Hefei National Laboratory for Physical Sciences at Microscale and School of Life Science, University of Science & Technology of China, Hefei, Anhui 230027, China
[b] Key Laboratory of Systems Biology, State Key Laboratory of Molecular Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue-Yang Road, Shanghai 200031, China

ARTICLE INFO

Article history:
Received 19 June 2010
Revised 4 September 2010
Accepted 8 September 2010
Available online 17 September 2010

Edited by Takashi Gojobori

Keywords:
Protein–protein interaction
Protein interaction network
Coevolution
Modularity

ABSTRACT

We investigated what roles coevolution plays in shaping yeast protein interaction network (PIN). We found that the extent of coevolution between two proteins decreases rapidly as their interacting distance on the PIN increases, suggesting coevolutionary constraint is a short-distance force at the molecular level. We also found that protein–protein interactions (PPIs) with strong coevolution tend to be enriched in interconnected clusters, whereas PPIs with weak coevolution are more frequently present at inter-cluster region. The findings indicate the close relationship between coevolution and modular organization of PINs, and may provide insights into evolution and modularity of cellular networks.

1. Introduction

Protein–protein interactions (PPIs) are involved in almost all biological processes and are crucial to cellular functions. Knowledge of the evolution of PPIs is critical to the understanding of the principles on the construction, function and evolution of biological systems. One important theoretical problem, yet still poorly understood, related to PPIs is the coevolution of proteins. As an evolutionary process, coevolution occurs at all biological levels. It is an important force that functions in the evolution of species and their communities [1]. Well characterized examples described by ecologists include the mutualistic relationship between hummingbirds and ornithophilous flowers [2]; the antagonism between predator and prey species, for example, garter snake and rough-skinned newt [3]. Coevolution also takes place at the molecular level as demonstrated by the strongly correlated behavior of the evolutionary histories between proteins and their interacting DNA, RNA and proteins in a variety of scenarios [4–6], for instance, protein ligands and their receptors such as the TGF-β/TGF-β receptor family [7]. Two general hypotheses have been proposed to explain the coevolution of proteins [8,9]. The first states that the similarity of the evolutionary pressure exerted on proteins results in the observed coevolution. The alternative hypothesis argues that it is physical coadaptation between interacting proteins via compensatory changes that causes protein coevolution. Both hypotheses are supported by evidences. Despite recent advances, many problems in the field need further investigation [8,9]. With the progresses in high-throughput biotechnology, a large amount of PPI data has been accumulated, which enables a shift toward the study of protein interaction networks (PINs) from analysis of individual proteins. Here, we address how coevolution interplays with the organization of PINs based on yeast PPI data. We found that coevolutionary constraint is a short-distance force at the PPI level and is highly associated with the modular organization of yeast PIN.

2. Materials and methods

2.1. Protein interaction data

We generated a 'merged yeast PIN' (MYP) by integrating the following high-quality PPI data: (1) the "Y2H-union", "combined-AP/MS" and "LC-multiple" data sets from [10]; (2) the list of PPIs annotated as high confidence in [11]; (3) the MIPS physical interactions from small-scale experiments[12].

Abbreviations: PPI, protein–protein interaction; PIN, protein interaction network; MYP, merged yeast PIN; PCC, Pearson correlation coefficient

* Corresponding author. Fax: +86 551 3600607.
** Corresponding author at: Hefei National Laboratory for Physical Sciences at Microscale and School of Life Science, University of Science & Technology of China, Hefei, Anhui 230027, China.
E-mail addresses: liangzhi@ustc.edu.cn (Z. Liang), wujr@sibs.ac.cn (J. Wu).
[1] These authors contribute equally to the work.

0014-5793/$36.00
doi:10.1016/j.febslet.2010.09.014

2.2. Sequence data

The sequence data of *Saccharomyces cerevisiae* was downloaded from SGD on Sep 1, 2008. The genome sequences of other 12 fungal species were downloaded from different databases on the same day: Genolevures: *Candida glabrata*, *Debaryomyces hansenii*, *Kluyveromyces lactis*, *Yarrowia lipolytica*; EBI: *Ashbya gossypii*; MIT Broad Institute: *Candida albicans*, *Neurospora crassa*, *Coccidioides posadasii*, *Aspergillus clavatus*, *Sclerotinia sclerotiorum*, *Fusarium graminearum*, *Schizosaccharomyces pombe*. Assignments of orthology were performed using InParanoid algorithm [13].

2.3. Assessment of protein coevolution

We used the mirror-tree approach to quantify coevolution between a pair of proteins [14–16]. Briefly, for a *S. cerevisiae* protein, if its orthologs was present in all the above 12 fungal species, the set of orthologous genes were aligned with CLUSTALW and a distance matrix was computed by using PROTDIST with the Jones–Taylor–Thornton model. For a pair of *S. cerevisiae* proteins having distance matrix available, the Pearson correlation coefficient (PCC) between the corresponding two matrices was determined [14]. The PCCs constituted a symmetric matrix, each row/column of which contained correlation values for a given protein with all of the others. Following Juan et al.'s protocol, the PCC profiles of two *S. cerevisiae* proteins were used to calculate a new correlation coefficient, which was taken as the final coevolutionary score *Re* [15].

2.4. Statistical test for coevolution-distance relationship

To evaluate the significance of the observed coevolution-distance relationship, we measured the deviation of the observation from its random expectation. Two kinds of network null models were used [17]. The first is randomized networks with the same degree distribution as the original network, which were generated by keeping node labels constant while swapping edges randomly but preserving the degree of each node. The second is randomized networks isomorphic to the original network, which was generated by shuffling the node labels while keeping the network topology unchanged. These two null models gave similar results. We computed the shortest path length between every pair of nodes on the randomized networks and generated a random median *Re* score distribution for each distance category. Based on the random distribution constructed by network null models, we used *Z* scores to evaluate the significance. Positive *Z* scores indicate that the observations are more frequent than random expectations, whereas negative *Z* scores indicate the opposite.

2.5. Fragmentation of network

We introduced a quantity *f* to evaluate the degree of fragmentation of a network.

$$f = \sum_{i}^{Ncc} \left(\frac{s_i}{\sum_{j}^{Ncc} s_j} \right)^2 = \sum_{i}^{Ncc} \left(\frac{s_i}{V} \right)^2$$

Ncc is the number of connected components in a network, *si* is the number of nodes in connected component *i*, *V* is the total number of nodes in a network. *f* ranges from $1/V(\approx 0$, *V* is usually pretty large) for a fully disconnected network consisting of only isolated nodes to 1 for a completely connected network (*i.e.* with only one component). For a partially disintegrated network, *f* takes between 0 and 1: the larger the extent of disintegration, the smaller the *f*, vice verse. So, when the edges are gradually removed from an intact network as we did in this analysis, *f* runs from 1 to 0, the trend of which indicates the locations of edges on a network. Since the computation of *f* is very fast, it allows the estimation of an effective interval for the edge attack.

3. Results and discussion

3.1. Coevolutionary constraint is a short-distance force at the molecular level

We first generated a MYP by combining several highly reliable PPI data sets (see Section 2). The resulting MYP contains 14,644 non-self PPIs involving 3466 proteins. To quantify the extent of coevolution between a pair of *S. cerevisiae* proteins, we used the well-established mirror-tree approach to compute a coevolutionary score *Re* (see Section 2). The closer the value of *Re* is to 1, the higher the extent of coevolution. We identified 1365 proteins in the MYP which were *Re* computable, and obtained the *Re* scores for all the 930,930 non-self protein pairs formed by them, among which 5833 pairs involving 1213 proteins corresponded to PPIs in the MYP and was annotated as 'MYP with *Re*' (MYP-R).

We compared the *Re* score distribution of the 5833 PPIs in MYP-R with that of all the possible 729,245 non-PPI protein pairs involving the same 1213 proteins. These two distributions are significantly different from one another, as measured by the Kolmogorov–Smirnov (KS) test ($P < 10^{-15}$). The former has a significantly higher mean *Re* of 0.72 than the 0.54 of the latter (*t*-test, $P < 10^{-15}$). The result indicates that interacting proteins are closely tied in evolution. We further addressed the question whether coevolution between a pair of proteins was correlated with their distance on the PIN. For every pair of the 1365 proteins having *Re* scores available, we computed the shortest path length between them on the MYP. The result showed that the value of *Re* decreases rapidly when the interacting distance between proteins increases (Fig. 1). To evaluate the significance of this observation, we checked the same relationship measured in randomized networks (see Section 2). It was shown that, on average, when the interacting

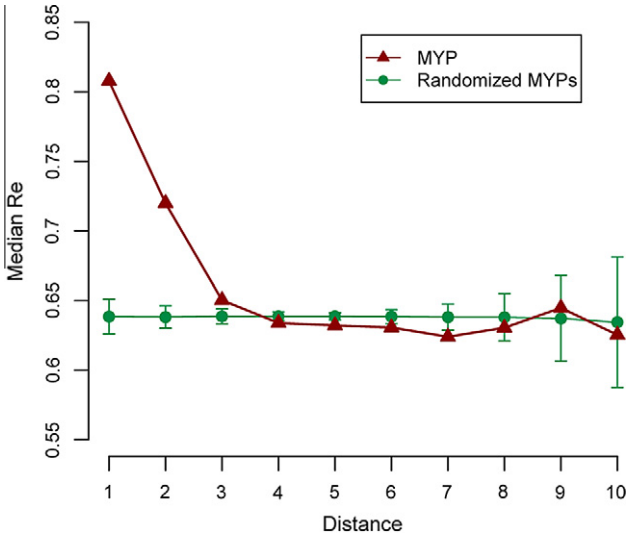

Fig. 1. The correlation between coevolution and interacting distance between a pair of yeast proteins. For a pair of proteins, the length of the shortest path connecting them is defined to be their interacting distance. The extent of coevolution between two proteins decreases rapidly with increase of their distance on the yeast PIN. The red triangles and line show the median *Re*s of different distance categories for the yeast PIN, and the green dots and line indicate the distribution of median *Re*s (mean ± s.d.) of different distance categories for randomized networks (see Section 2 and Supplementary material).

distance between a pair of proteins was equal to or more than about 3, their extent of coevolution could not be distinguished with random protein pairs (Fig. 1 and Supplementary Fig. 1). Taken together, these results suggest that coevolutionary constraint is a short-distance force at the molecular level.

Since interacting proteins on the PIN are more likely to be related in biological functions, these proteins might undergo similar regulation, resulting in similar evolutionary constraint. This argument was supported by the studies that coevolution of interacting proteins could result from similar functional constraint and evolutionary pressure [18,19]. Therefore, the distance-cutoff for coevolution derived from our work may be an average measure of the range that functional constraints can propagate through PPIs in yeast. On the other hand, the short-distance property of coevolutionary constraint could also be explained from a structural scenario, *i.e.* the physical coadaptation between protein structures, which argues that a deleterious mutation in one protein can be compensated by a mutation in its interacting partner in order to maintain the functional interaction [20–22]. Therefore, the observed rapid attenuate of coevolution with the interacting distance implies that the influence of a destabilizing change in one protein can spread at most to its vicinity. In another word, the effects of a genetic perturbation should damp quickly with the interacting distance away from the perturbed node, which might be due to the flexibility of protein structures and reflect the robustness of yeast PIN at the long-time scale of evolution. This observation complements a recent report that the concentration perturbations on the yeast PIN were strongly localized and exponentially decayed with distance [23], which operates at the short time scale of biochemical reactions.

3.2. Coevolution is associated with the modular organization of yeast PIN

If coevolution works in local as described above, we are curious about whether coevolution plays a role in the organization of PINs, particularly in modularity [24]. It is possible that PPIs with different extent of coevolution could distribute homogeneously on PINs. In such case, inter-module PPIs should have the same extent of coevolution as intra-module PPIs have, and coevolution might have nothing to do with modularity. Alternatively, since coevolution works in a short distance, the PPIs with strong coevolution could become a driving force for the modularity in PINs, whereas the PPIs with weak coevolution might disperse between these modules.

To distinguish these two hypotheses, we investigated the possible location distribution of PPIs with different extents of coevolution by comparing their edge betweenness, which measures the number of shortest paths between all node pairs that traverse a given edge [25]. Edges between highly interconnected clusters were proposed to have high edge betweenness [25,26]. Therefore, if the first model is true, we should expect no coevolution differences between PPIs with high and low edge betweenness, whereas for the second model, we should expect stronger coevolution for PPIs with low edge betweenness than that for PPIs with high edge betweenness. We defined the top 10% of PPIs with the highest betweenness in the MYP as bottlenecks and the rest as non-bottlenecks. It was found that the bottleneck PPIs had a significantly lower mean *Re* of 0.60 than the 0.73 of the non-bottleneck PPIs (Fig. 2A; *t*-test $P < 10^{-12}$), which supports the second model and suggests that the PPIs with low extent of coevolution show their tendency to be outside of densely interconnected clusters, whereas the PPIs with high extent of coevolution are highly enriched inside these clusters. We repeated the analysis with different parameter settings for the bottlenecks and non-bottlenecks, and the similar results were obtained (see Supplementary Table 3).

To further confirm our above observations, we applied a decomposing approach in network analysis [27]. We removed PPIs from the largest connected component (including 1089 proteins and 5716 PPIs) of MYP-R in an ascendant, descendant and random order based on their *Re* scores, respectively. At each round of attack, we computed a score *f* which quantifies the fragmentation of network topology (see Section 2). Distinct effects were observed for different removing experiments (Fig. 2B and Supplementary Fig. 2). Ascendant attack has the most deleterious effect on the network integrity, descendant attack has the least and random attack locates between. These results suggest that PPIs with high extent of coevolution tend to be enriched in the interconnected clusters and thus their removals have small effect on the network integrity, whereas PPIs with low extent of coevolution have the tendency to be located between those clusters and thus their removals break the network quickly. These results are in agreement with the above edge betweenness analysis. The results indicate that the short-distance property of coevolution is associated with the modular organization of yeast PIN. All results were robust to different data sets

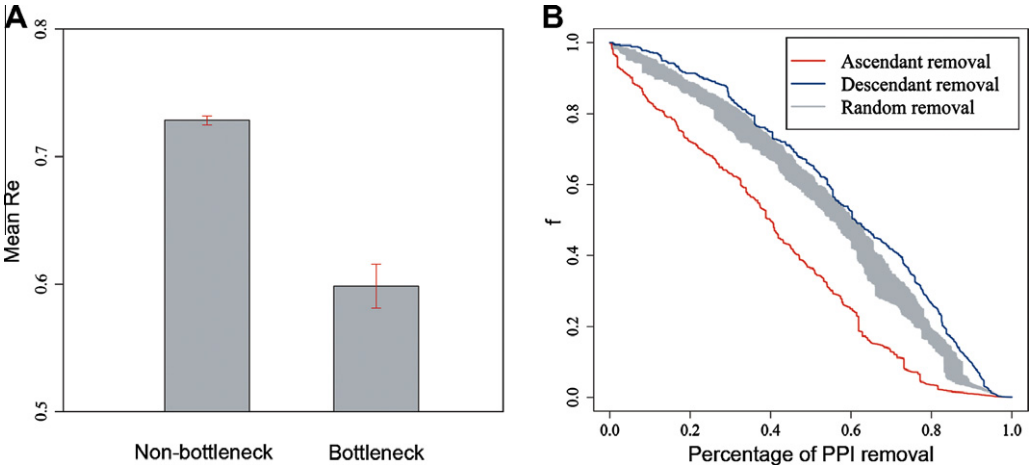

Fig. 2. PPIs with different extent of coevolution have different topological distributions on the yeast PIN. Weakly coevolving PPIs are central to the global organization of network topology, as strongly coevolving PPIs are to the local. (A) The mean *Re* scores (±s.e.) for bottleneck and non-bottleneck PPIs. The distributions of *Re* scores for these two types of PPIs are significantly different from one another (KS test, $P < 10-13$) and the bottleneck PPIs has a significantly lower mean *Re* value than that of the non-bottleneck PPIs (*t*-test, $P < 10^{-12}$). (B) The effect of PPI removal on the network integrity. Lines in red and blue depict the disintegration of network due to PPI attack in ascendant and descendant order based on *Re* scores, respectively. Gray area shows the effective interval of random attacks.

and parameter selections in our computation (see Supplementary material).

Acknowledgements

China Postdoctoral Science Foundation (20070420731); National Natural Science Foundation of China (30900270, 3821065, 30571066), Ministry of Science and Technology of China (2006CB503900, 2006CB901200, 2006CB806500) and Chinese Academy of Sciences (KSCX1-YW-02, INFO-115-D01-2009, KSCX2-YW-R-127).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2010.09.014.

References

[1] Thompson, J.N. (1994) The Coevolutionary Process, University of Chicago Press, Chicago.
[2] Brown, J.H. and Kodric-Brown, A. (1979) Convergence, competition, and mimicry in a temperate community of hummingbird-pollinated flowers. Ecology 60 (5), 1022–1035.
[3] Dawkins, R. and Krebs, J.R. (1979) Arms races between and within species. Proc. R. Soc. Lond. B 205, 489–511.
[4] Metzenberg, S., Joblet, C., Verspieren, P. and Agabian, N. (1993) Ribosomal protein L25 from *Trypanosoma brucei*: phylogeny and molecular co-evolution of an rRNA-binding protein and its rRNA binding site. Nucleic Acids Res. 21, 4936–4940.
[5] Athanikar, J.N. and Osborne, T.F. (1998) Specificity in cholesterol regulation of gene expression by coevolution of sterol regulatory DNA element and its binding protein. Proc. Natl. Acad. Sci. USA 95, 4935–4940.
[6] Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D. and Cohen, F.E. (2000) Coevolution of proteins with their interaction partners. J. Mol. Biol. 299, 283–293.
[7] Goh, C.S. and Cohen, F.E. (2002) Coevolutionary analysis reveals insights into protein–protein interactions. J. Mol. Biol. 324, 177–192.
[8] Juan, D., Pazos, F. and Valencia, A. (2008) Coevolution and coadaptation in protein networks. FEBS Lett. 582, 1225–1230.
[9] Lovell, S.C. and Robertson, D.L. (2010) An integrated view of molecular coevolution in protein–protein interactions. Mol. Biol. Evol., Advance Access doi:10.1093/molbev/msq144.
[10] Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.F., Dricot, A., Vazquez, A., Murray, R.R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.S., Motyl, A., Hudson, M.E., Park, J., Xin, X., Cusick, M.E., Moore, T., Boone, C., Snyder, M., Roth, F.P., Barabási, A.L., Tavernier, J., Hill, D.E. and Vidal, M. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322, 104–110.
[11] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417, 399–403.
[12] Mewes, H.W., Frishman, D., Mayer, K.F., Münsterkötter, M., Noubibou, O., Pagel, P., Rattei, T., Oesterheld, M., Ruepp, A. and Stümpflen, V. (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res. 34, D169–D172.
[13] Brien, K., Remm, M. and Sonnhammer, E. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 33, D476–D480.
[14] Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng. 14, 609–614.
[15] Juan, D., Pazos, F. and Valencia, A. (2007) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc. Natl. Acad. Sci. USA 105, 934–939.
[16] Tillier, E.R.M. and Charlebois, R.L. (2009) The human protein coevolution network. Genome Res. 19, 1861–1871.
[17] Maslov, S. and Sneppen, K. (2002) Specificity and stability in topology of protein networks. Science 296, 910–913.
[18] Fraser, H.B., Hirsh, A.E., Wall, D.P. and Eisen, M.B. (2004) Coevolution of gene expression among interacting proteins. Proc. Natl. Acad. Sci. USA 101, 9033–9038.
[19] Hakes, L., Lovell, S., Oliver, S.G. and Robertson, D.L. (2007) Specificity in protein interactions and its relationship with sequence diversity and coevolution. Proc. Natl. Acad. Sci. USA 104, 7999–8004.
[20] Choi, S.S., Li, W. and Lahn, B.T. (2005) Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis. Nat. Genet. 37, 1367–1371.
[21] Mintseris, J. and Weng, Z. (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. Proc. Natl. Acad. Sci. USA 102, 10930–10935.
[22] Ferrer-Costa, C., Orozco, M. and Cruz, X. (2007) Characterization of compensated mutations in terms of structural and physicochemical properties. J. Mol. Biol. 365, 249–256.
[23] Maslov, S. and Ispolatov, I. (2007) Propagation of large concentration changes in reversible protein-binding networks. Proc. Natl. Acad. Sci. USA 104, 13655–13660.
[24] Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) From molecular to modular cell biology. Nature 402, C47–C52.
[25] Girvan, M. and Newman, M. (2002) Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826.
[26] Yu, H., Kim, P.M., Sprecher, E., Trifonov, V. and Gerstein, M. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput. Biol. 3, e59.
[27] Albert, R., Jeong, H. and Barabási, A. (2000) Error and attack tolerance of complex networks. Nature 406, 378–382.