J Mol Evol (2009) 68:192–196
DOI 10.1007/s00239-008-9185-2

Global Versus Local Centrality in Evolution of Yeast Protein Network

Alexander E. Vinogradov

Received: 28 April 2008 / Accepted: 12 November 2008 / Published online: 14 January 2009
© Springer Science+Business Media, LLC 2009

Abstract It is shown here that in the yeast protein interaction network the global centrality measure (betweenness) depends on the protein evolutionary age (i.e., on historic contingency) more weakly than the local centrality measure (degree). This phenomenon is not observed in mutational duplication-and-divergence models. The network domains responsible for this difference deal with DNA/RNA information processing, regulation, and cell cycle. A selection vector can operate in these domains, which integrates the network activity and thus compensates for the process of mutational divergence.

Keywords Protein interaction networks · Selection · Mutation-and-divergence · Integration · Systems biology

Introduction

Networks of protein interactions grow in evolution by means of gene duplication and divergence complicated by rewiring of existing interactions. There is much argument over the role of selection in this process (e.g., Pastor-Satorras et al. 2003; Wagner 2003; Hahn et al. 2004; Dosztányi et al. 2006; Beltrao and Serrano 2007; Kim et al. 2007; Stumpf et al. 2007; Wang and Zhang 2007). A contrast between the neutralist and the selectionist standpoints can be sought in the holistic nature of the organism (where a change in one part should be associated with the corresponding adjustments of other parts). For instance, the mutational duplication-and-divergence model can explain the appearance of new genes but does not account for a complementary process—the integration of work of a growing number of diverse genes. Metaphorically, the mutational divergence can be compared with a 'centrifugal force,' whereas the integration, with a compensating 'centripetal force.' Until the emergence of systems biology, there were no approaches to study the latter phenomenon (because integration is a systemic property). As a first trial, it is tempting to look for traces of integrative selection comparing the local and global topology in the biological networks and mutational duplication-and-divergence models.

Here I compare the local and global centrality in the protein interaction network of a unicellular eukaryote (the yeast *Saccharomyces cerevisiae*) and model networks in relation to network growth in evolution. The degree (number of one-step interactions of a given node) was taken as a measure of local centrality, whereas the betweenness (number of shortest paths between any other nodes crossing a given node) was used as a measure of global centrality.

Materials and Methods

The dataset of yeast protein interactions was taken from Batada et al. (2007). This is currently a most complete, high-confidence dataset where each interaction was validated by at least two different experimental methods. The protein evolutionary age was determined using the NCBI phylogenetic tree (Wheeler et al. 2006) and the COG (KOG) orthologous gene groups (Tatusov et al 2003; Koonin et al. 2004) as presented in the STRING database,

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9185-2) contains supplementary material, which is available to authorized users.

A. E. Vinogradov (✉)
Institute of Cytology, Russian Academy of Sciences,
St. Petersburg 194064, Russia
e-mail: aevin@mail.cytspb.rssi.ru

with the addition of NOGs (von Mering et al. 2007). Six evolutionary stages were taken, determined by the following phylogenetic branching: Bacteria, Archaea, Eukaryota, Fungi, Ascomycota, and Saccharomycetes. A yeast protein was regarded as appearing at a corresponding evolutionary stage if it had relatives in the same COG (KOG, NOG) group in the phylogenetic lineages that branched off after this stage and no relatives in the lineages that branched off earlier.

Model networks were constructed using an algorithm of node duplication and divergence similar to one described previously (Pastor-Satorras et al. 2003; Vazquez et al. 2003; Wagner 2003), with modifications. The model starts with two connected nodes. Iteratively, a randomly chosen node is copied with all its interactions, the interactions of each duplicate are randomly removed, and their new interactions are randomly formed. There are two parameters in the model: alpha, the probability of removing of an existing interaction; and beta, the probability of forming a new interaction with any of the existing nodes. Both alpha and beta can be asymmetric for duplicate counterparts (e.g., a lower probability for an older counterpart). In part of the analysis, beta was normalized by dividing it by the number of available nodes in each iteration step. Thus, it was diminishing as the network was growing (because the probability of formation of interaction with a given node may depend on the total number of nodes potentially available for interaction). In another part of the analysis, competition divergence between duplicate counterparts was introduced into the model. For this purpose, the operation of removing and introducing an interaction was applied to only one of the duplicates (drawn randomly for each interaction with probability 0.5). The (reversed) order of appearance in the model was regarded as the evolutionary age of a node in the model network. In another part of the analysis the yeast protein network was rewired (to provide a control). Each interaction was randomly removed and replaced with a purely random one (i.e., two new nodes were drawn randomly from all nonconnected nodes). In the other control test, reshuffling of the evolutionary age of nodes was performed.

In both the yeast protein network and the model networks, the giant component (i.e., the greatest connected component containing a majority of network nodes) was extracted. The local centrality measure (degree; i.e., the number of one-step interactions of a given node) and the local centrality measure (betweenness; i.e., the number of shortest paths between any other nodes crossing a given node) were calculated for the giant component of each network instance.

Analysis of functional protein (gene) groups reflected in the Gene Ontology (GO) categories (Gene Ontology Consortium 2008) was done as described (Vinogradov and Anatskaya 2007). Briefly, the average value of the parameter under question for a gene group belonging to a given GO category was checked against the average value for the total gene dataset. For each GO category, I collected all its subcategories using the directed acyclic graph of a given GO domain ('biological processes,' 'molecular functions,' or 'cellular components'), and a gene was regarded as belonging to a given category if it was mapped to any of its subcategories in the Entrez Gene (Maglott et al. 2007). Two parameters were tested: (i) the residuals of regression of the log-transformed betweenness on the evolutionary age of a given protein and (ii) the average value of the absolute log-transformed ratios between the evolutionary ages of the interactants in all pairwise interactions of a given protein. In other words, the former parameter emphasizes those proteins (and GO categories) that have higher betweenness for their evolutionary age, whereas the latter parameter estimates the heterogeneity of the evolutionary age of interacted proteins. For evaluation of statistical significance, I did 20,000 random samplings from the total dataset (of a size equal to the size of a tested gene group). After obtaining the two-tailed significance level (p-value), the false discovery rate (q-value) was estimated for correction for multiple comparisons (Storey and Tibshirani 2003).

Centrality and Evolutionary Age

In the yeast protein network the evolutionary age of a node (estimated by the major phylogenetic lineage branching; see Methods) correlated positively with both the local and the global centrality measures beginning from the Archaea (Supplementary Fig. 1). The nonmonotonous region before the Archaea stage suggests a special evolutionary mode, which differs from the further gradual evolution. This special mode is probably related to a hybrid (symbiotic) origin of Eukaryota (e.g., Rivera and Lake 2004; Aravind et al. 2006). As a result, bacterial proteins have lower centrality in the eukaryotic cell compared with their antiquity. In the further analysis only the monotonic region of the gradual network evolution was used (presumably realized by the gene duplication-and-divergence mechanism), which is characterized by a positive correlation between centrality measures and evolutionary age (Archaea-Saccharomycetes).

In the model networks the evolutionary age of a node correlated positively with the network centrality measures under the following conditions: (i) asymmetry of the probability of removing an existing interaction ('alpha')—a lower probability for an older duplicate counterpart; and/or (ii) the probability of formation of a new interaction ('beta') being normalized by division by the number of available nodes (Table 1 and Supplementary Table 1). It should be

Table 1 Statistics for yeast protein network and duplication-and-divergence models[a]

Network	Alpha[b]	Beta[c]	No. nodes in giant component	Avg degree	Avg betweenness	Spearman correlation with evolutionary age[d]		
						Degree	Betweenness	Rank(degree) minus rank(betweenness)
Yeast network	—	—	2228	5.57	0.0021	0.208	0.141	0.099
Model[e]	0.3	0.3n	2213 ± 421	5.24 ± 0.41	0.0020 ± 0.0005	0.189 ± 0.026	0.198 ± 0.022	−0.045 ± 0.012 ($p < 0.05$)
Model[e]	0.3	0.5n	2362 ± 543	6.23 ± 0.83	0.0016 ± 0.0007	0.198 ± 0.024	0.205 ± 0.027	−0.041 ± 0.016 ($p < 0.05$)
Model[e]	0.5	0.9n	2631 ± 482	4.93 ± 0.67	0.0018 ± 0.0006	0.228 ± 0.031	0.223 ± 0.026	0.018 ± 0.019 ($p > 0.6$)
Model[e]	0.3/0.7	0.002	2498 ± 496	6.18 ± 0.71	0.0017 ± 0.0007	0.205 ± 0.029	0.201 ± 0.021	0.001 ± 0.023 ($p > 0.9$)

[a] Some typical models are shown (many more examples are given in Supplementary Table 1)

[b] Probability of interaction being removed. Where two values of alpha are listed, the alpha was asymmetric for older and younger duplicate counterparts

[c] Probability of interaction appearing. Where the beta value is followed by "*n*," it was normalized by the number of available nodes

[d] $p < 0.0001$, if not indicated otherwise

[e] Ten simulations were done for each model (varying the number of iterations); means ± confidence intervals

noted that the asymmetric alpha suggests a stronger conservation selection for the older duplicate counterpart, which is reasonable from the selectionist standpoint (because the younger counterpart turns out in novel genomic surroundings) but has no ground in the neutralist case. Furthermore, in the case of an asymmetric alpha it is reasonable to make beta similarly asymmetric (i.e., the older counterpart has a lower probability both to break the old interactions and to form new ones). However, asymmetry of beta reduced the positive correlation between evolutionary age and centrality measures. Therefore, the case with a normalized beta seems more plausible. In this case we only assume that the probability of formation of interaction with a given node is diminishing as the number of available nodes is increasing. (In fact, this assumption is also not strictly neutral. Under neutrality, if a protein mutationally acquires the ability to interact with some other protein, it does not matter how many other proteins also acquire the ability to interact with this protein.)

In the yeast protein network the global centrality (betweenness) increases more slowly than the local one (degree) with an increase in evolutionary age, which is not observed in the model networks (Table 1 and Supplementary Table 1). In other words, the yeast protein duplicates tend to form interactions with nodes of high global rather than local centrality (and thus reduce the correlation between global centrality and evolutionary age), which can be an indication of integrative selection. When the most ancient part of the model networks was removed from the analysis (similarly as was done with the bacterial proteins in the yeast network), the results were qualitatively similar (Supplementary Table 1). The effect holds in the case of normalized beta and/or asymmetric alpha (Table 1 and Supplementary Table 1). It should be emphasized that the model networks were tested over a wide range of network parameter values (number of nodes, degree, betweenness) that extends far on both sides from the parameter values of the yeast network (Supplementary Table 1). However, over this wide range of parameters, model networks do not show the stronger correlation of evolutionary age with degree (than with betweenness) that is characteristic for the yeast network. In Table 1 the model networks with parameters closest to those of the yeast network are listed (with 10 simulations for each model).

When another simple selectionist assumption was introduced into the model—the competition divergence between duplicate counterparts—the results did not change qualitatively (Supplementary Table 1). When the preferential attachment proportional to a node degree, which is frequently used to explain real-network properties (e.g., D'Souza et al. 2007; Davids and Zhang 2008), was added, the results were also similar (Supplementary Table 1). (Noteworthily, the preferential attachment is difficult to explain from the neutralist standpoint).

Robustness of the Revealed Effect

The revealed difference between the correlation coefficient of evolutionary age with betweenness and that of evolutionary age with degree in yeast network reaches about half the value of betweenness correlation coefficient (Table 1). It should be emphasized that this difference is very robust. When the yeast network was randomly rewired with the increasing probability of replacing an interaction, the effect was gradually diminishing but it held up to $p = \sim 0.5$ (Fig. 1). The other testing approach, reshuffling of evolutionary age of the nodes, gave similar results (Supplementary Fig. 2).

The correlation coefficients are low but highly significant (as well as the difference between them). Probably,

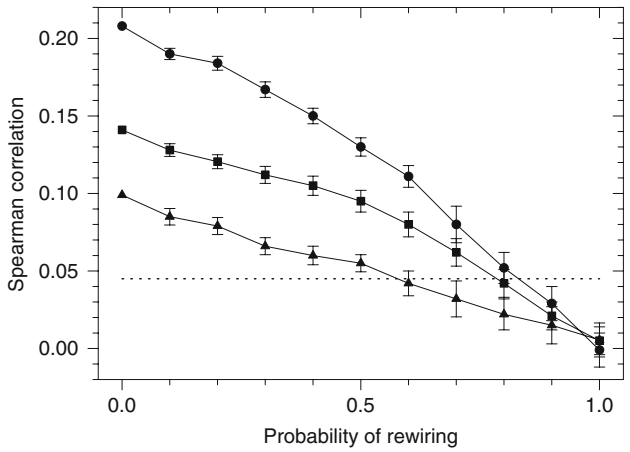

Fig. 1 Spearman rank correlation coefficients between evolutionary age and centrality measures in the yeast protein network with different levels of random rewiring of network edges. Circles, degree; squares,- betweenness; triangles, rank(degree) minus rank(betweenness); dotted line, significance level for $p = 0.05$. Means ± confidence intervals; 10 simulations were done for each 'probability of rewiring' point above zero

they are the vestiges of the evolutionary growth of the network from the historic center. Noteworthily, the correlation coefficient between evolutionary age and degree in the yeast network is close to the maximum, which is obtained in the model networks with a normalized beta. A stronger correlation (albeit still not higher for degree than for betweenness) was observed in model networks only at a greatly asymmetric alpha (e.g., three- or fourfold) without the corresponding asymmetry of beta (Supplementary Table 1), which is not realistic. It should also be mentioned that the correlation coefficients should have a larger error in the yeast network compared to the model networks (because in the model networks the evolutionary age is determined without error and discreteness).

Gene Ontology Categories

Because the revealed effect is a systemic property, its functional correspondence should be sought on the above-genic level. The analysis of overrepresented Gene Ontology categories showed a rather consistent effect for all GO branches ('biological processes,' 'molecular functions,' or 'cellular components'). Proteins whose betweenness is relatively high for their evolutionary age belong to categories devoted to DNA/RNA processing, cell cycle, reproduction, transcription, chromosome organization, and cytoskeleton (the latter category can mostly be related to spindle formation during the cell cycle) (Table 2 and Supplementary Tables 2A and B). These are network domains that reduce the correlation between global centrality and evolutionary age. Similarly, the significant discrepancy in evolutionary age of interactants is observed in proteins belonging to regulation, transcription, chromatin remodeling, chromosome organization, signal transduction, transcription regulator activity, and DNA binding categories (Table 3 and Supplementary Tables 3A and B). Generally, the 'deviant' network domains deal with information processing, regulation, and cell cycle. A selection process can operate in these domains which

Table 2 Gene Ontology 'biological processes' overrepresented for the residuals of regression of the log-transformed betweenness on the evolutionary age of a given protein

Category ID	Category name	Contrast[a]	No. of genes in category	q-value[b]
GO:0006139	Nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process	0.736	624	0
GO:0000003	Reproduction	2.068	117	0
GO:0006259	DNA metabolic process	1.268	227	0.005
GO:0007266	Rho protein signal transduction	4.993	10	0.005
GO:0019953	Sexual reproduction	2.425	50	0.015
GO:0007010	Cytoskeleton organization and biogenesis	1.644	119	0.015
GO:0016070	RNA metabolic process	0.743	489	0.018
GO:0051276	Chromosome organization and biogenesis	0.984	291	0.023
GO:0000278	Mitotic cell cycle	1.378	156	0.025
GO:0007049	Cell cycle	1.100	238	0.030
GO:0006366	Transcription from RNA polymerase II promoter	1.125	208	0.032
GO:0016573	Histone acetylation	3.066	29	0.032
GO:0051704	Multiorganism process	2.281	53	0.037
GO:0032505	Reproduction of a single-celled organism	1.918	67	0.045
GO:0022402	Cell cycle process	1.087	206	0.048

[a] Difference between the average value for a category and the average value for the total dataset

[b] False discovery rate

Table 3 Gene Ontology 'biological processes' overrepresented for the average value of the absolute log-transformed ratio between the evolutionary ages of a given protein and its interactants in all pairwise interactions

Category ID	Category name	Contrast[a]	No. of genes in category	*q*-value[b]
GO:0006350	Transcription	0.020	295	0.006
GO:0006139	Nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process	0.011	579	0.011
GO:0006461	Protein complex assembly	0.057	37	0.023
GO:0065007	Biological regulation	0.016	335	0.031
GO:0006338	Chromatin remodeling	0.035	90	0.032
GO:0019222	Regulation of metabolic process	0.020	220	0.036
GO:0007001	Chromosome organization and biogenesis (sensu Eukaryota)	0.018	263	0.036
GO:0006351	Transcription, DNA-dependent	0.017	275	0.036
GO:0032774	RNA biosynthetic process	0.017	280	0.036
GO:0051276	Chromosome organization and biogenesis	0.017	267	0.036
GO:0050789	Regulation of biological process	0.015	309	0.036
GO:0007165	Signal transduction	0.040	56	0.049

[a] Difference between the average value for a category and the average value for the total dataset

[b] False discovery rate

integrates the network activity and thus compensates for the process of mutational divergence.

Acknowledgments I thank two anonymous reviewers for helpful comments. This work was supported by the Russian Foundation for Basic Research (RFBR) and by the Programme of the Russian Academy of Sciences 'Molecular and Cellular Biology' (MCB RAS).

References

Aravind L, Iyer LM, Koonin EV (2006) Comparative genomics and structural biology of the molecular innovations of eukaryotes. Curr Opin Struct Biol 16:409–419

Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hurst LD, Tyers M (2007) Still stratus not altocumulus: further evidence against the date/party hub distinction. PLoS Biol 5:e154

Beltrao P, Serrano L (2007) Specificity and evolvability in eukaryotic protein interaction networks. PLoS Comput Biol 3:e25

Davids W, Zhang Z (2008) The impact of horizontal gene transfer in shaping operons and protein interaction networks—direct evidence of preferential attachment. BMC Evol Biol 8:23

Dosztányi Z, Chen J, Dunker AK, Simon I, Tompa P (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. J Proteome Res 5:2985–2995

D'Souza RM, Borgs C, Chayes JT, Berger N, Kleinberg RD (2007) Emergence of tempered preferential attachment from optimization. Proc Natl Acad Sci USA 104:6112–6117

Gene Ontology Consortium (2008) The Gene Ontology project in 2008. Nucleic Acids Res 36:D440–D444

Hahn MW, Conant GC, Wagner A (2004) Molecular evolution in large genetic networks: does connectivity equal constraint? J Mol Evol 58:203–211

Kim PM, Korbel JO, Gerstein MB (2007) Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. Proc Natl Acad Sci USA 104:20274–20279

Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:R7

Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 35:D26–D31

Pastor-Satorras R, Smith B, Sole RV (2003) Evolving protein interaction networks through gene duplication. J Theor Biol 222:199–210

Rivera MC, Lake JA (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431:152–155

Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. Proc Natl Acad Sci USA 100:9440–9445

Stumpf MP, Kelly WP, Thorne T, Wiuf C (2007) Evolution at the system level: the natural history of protein interaction networks. Trends Ecol Evol 22:366–373

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

Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Modeling of protein interaction networks. ComPlexUs 1:38–44

Vinogradov AE, Anatskaya OV (2007) Organismal complexity, cell differentiation and gene expression: human over mouse. Nucleic Acids Res 35:6350–6356

von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P (2007) STRING 7—recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35:D358–D362

Wagner A (2003) How the global structure of protein interaction networks evolves. Proc Biol Sci 270:457–466

Wang Z, Zhang J (2007) In search of the biological significance of modular structures in protein networks. PLoS Comput Biol 3:e107

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, Kapustin Y, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E (2006) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 34:D173–D180