Research article

Stable evolutionary signal in a Yeast protein interaction network

Stefan Wuchty*[1], Albert-Laszlo Barabási[2] and Michael T Ferdig[3]

Address: [1]Northwestern Institute on Complexity, Chambers Hall, Northwestern University, 600 Foster Street, Evanston, IL 60202, USA, [2]Department of Physics, 225 Nieuwland Science Hall, University of Notre Dame, Notre Dame, IN 46556, USA and [3]Department of Biology, 107 Galvin Science Hall, University of Notre Dame, Notre Dame, IN 46556, USA

Email: Stefan Wuchty* - s-wuchty@northwestern.edu; Albert-Laszlo Barabási - alb@nd.edu; Michael T Ferdig - mferdig@nd.edu

* Corresponding author

Abstract

Background: The recently emerged protein interaction network paradigm can provide novel and important insights into the innerworkings of a cell. Yet, the heavy burden of both false positive and false negative protein-protein interaction data casts doubt on the broader usefulness of these interaction sets. Approaches focusing on one-protein-at-a-time have been powerfully employed to demonstrate the high degree of conservation of proteins participating in numerous interactions; here, we expand his 'node' focused paradigm to investigate the relative persistence of 'link' based evolutionary signals in a protein interaction network of *S. cerevisiae* and point out the value of this relatively untapped source of information.

Results: The trend for highly connected proteins to be preferably conserved in evolution is stable, even in the context of tremendous noise in the underlying protein interactions as well as in the assignment of orthology among five higher eukaryotes. We find that local clustering around interactions correlates with preferred evolutionary conservation of the participating proteins; furthermore the correlation between high local clustering and evolutionary conservation is accompanied by a stable elevated degree of coexpression of the interacting proteins. We use this conserved interaction data, combined with *P. falciparum* /Yeast orthologs, as proof-of-principle that high-order network topology can be used comparatively to deduce local network structure in non-model organisms.

Conclusion: High local clustering is a criterion for the reliability of an interaction and coincides with preferred evolutionary conservation and significant coexpression. These strong and stable correlations indicate that evolutionary units go beyond a single protein to include the interactions among them. In particular, the stability of these signals in the face of extreme noise suggests that empirical protein interaction data can be integrated with orthologous clustering around these protein interactions to reliably infer local network structures in non-model organisms.

Background

An ambitious goal of contemporary proteome research is the elucidation of the structure, interactions and functions of the proteins that constitute cells and organisms. During the last few years, large-scale efforts have unraveled the complex web of protein interactions in simple organisms such as *H. pylori* [1], *E. coli* [2] and *S. cerevisiae* [3-7]. Most recently, attention has focused on the first protein interac-
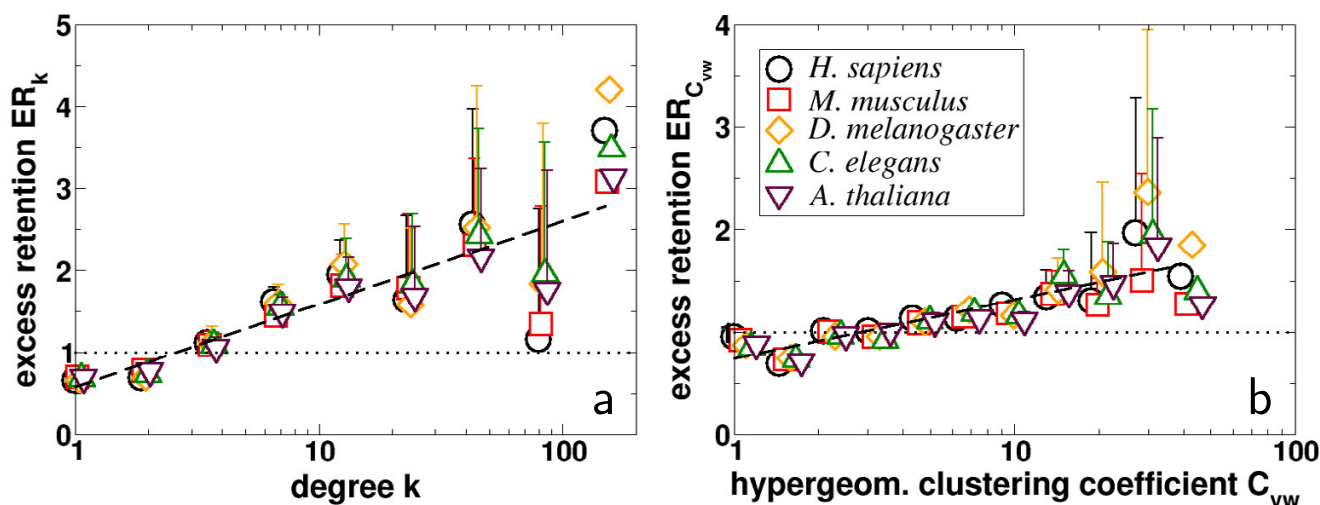
Figure 1
(a) Pooling proteins according to their level of interaction *k* we determined the excess retention *ER* of these grouped protein (pair)s that have orthologs in *H. sapiens, M. musculus, D. melanogaster, C. elegans* and *A. thaliana*. Averaging these *k* depending values of ER_k in bins of logarithmic size we observe clear logarithmic trends. **(b)** Analogously, we pooled interacting protein pairs *v, w* of Yeast according to their hypergeometric clustering coefficient C_{vw} and determined the excess retention *ER* in these groups of protein pairs that both have orthologs in the aforementioned higher eukaryotes. Pooling these C_{vw} dependent values of *ER* in bins of logarithmically increasing size we observe clear logarithmic trends again. In both cases, significant Pearson's and Spearman's rank correlation coefficients [see Additional file 1] support our conclusion that not only highly interacting proteins (as exemplified by a large *k*) are predominately preserved but also interacting protein pairs which are embedded in a highly cohesive neighborhood (as exemplified by a high C_{vw}). Error bars indicate the standard deviations from the mean excess retention in each bin.

tion maps of complex multicellular organisms such as *C. elegans* [8] and *D. melanogaster* [9]. Although these organisms vary extensively in their complexity, corroborative evidence points to a series of simple organizing principles that characterize all complex protein interaction networks [10]. The most dramatic of these is their scale-free nature [11,12], highlighting a small number of highly connected proteins which secure the integrity and connectivity among modules [13,14] that are discernible, yet topologically overlapping, clusters of densely interconnected protein groups sharing well-defined functions [10,15-18]. A crucial biological corollary of this ubiquitous network organization is the observation that hubs exhibit an elevated propensity to be simultaneously conserved in evolution and are essential for survival [13,19,20]. This role of highly connected proteins is further indicated by a considerable degree of sequence conservation [21-25]. Similarly, cohesively bound modules have been conserved as a whole, suggesting the presence of evolutionary relevant building blocks [26-28]. This hypothesis is further supported by the observation that proteins belonging to a certain module tend to be coexpressed [29] and coregulated [30]. These particular results are utilized for the comparison of protein pathways of various organisms [31], mod-

eling of interactomes [32,33] and prediction of protein functions [34].

These insights have fundamental implications for our understanding of biological processes and potential applications; however the severe error-proneness of methods for the determination of protein interactions casts doubt on the integrity of such datasets. For example, an estimate of the accuracy of protein interactions in *S. cerevisae* uncovered a startling false negative rate of 90%, and a 50% false positive error rate [35].

Despite incoherences in the determination of protein interactions and orthologs, we observe that extensive information remains in the topology of a protein interaction network. In particular, even tremendous experimental noise does not bury the strong evolutionary signal that highly connected nodes in an interaction web of Yeast proteins are preferably conserved in higher eukaryotes. Accounting for interactions between pairs of Yeast proteins, we find that the reliability of an interaction as indicated by a high degree of local clustering around interactions is accompanied by an elevated propensity for the corresponding proteins to be evolutionary conserved.

In addition, we observe that such interactions are preferably coexpressed in both the reference and a target organism, suggesting that conservation occurs not only on the level of individual proteins but also on the level of their interactions. The observation that such link-based evolutionary signals prevail in the topology of an otherwise extremely noisy protein interaction network indicates a novel way to uncover protein interactions in any organism for which orthologs can be identified from sequence data.

Results

As a basis of our considerations we utilized a protein-protein interaction network of *S. cerevisae* from the DIP database [36], providing 3, 833 proteins embedded in 11, 942 interactions. We labeled pairs of proteins as orthologous to each other as of the InParanoid database [37] that relates proteins of *S. cerevisiae* to complete protein sets of various higher eukaryotes, allowing us to utilize 1, 928 Yeast proteins with putative orthologs in *H. sapiens*, 2,073 in *A. thaliana*, 1, 885 in *C. elegans*, 1, 885 in *M. musculus* and 1,631 in *D. melanogaster*.

Evolutionary retention of single proteins

Utilizing these data sets we recently uncovered a correlation between a Yeast proteins level of interaction and its propensity to be evolutionary conserved [20]. Pooling all proteins into groups according to their connectivity k we determine the respective fraction of orthologs in each group. As a null-hypothesis we assume a random distribution of orthologs that is quantified by the fraction of proteins with an ortholog in a target eukaryote and the total number of proteins present in the underlying Yeast protein interaction network. The degree dependent orthologous excess retention, ER_k, defined as the ratio of ortholog fractions in k dependent groups of proteins and fractions of randomly distributed orthologous proteins reflects the dependence of evolutionary protein conservations as function of the proteins connectivity. Logarithmically binning the k-dependent values of ER_k the averages in each bin show a clear and systematic trend toward preferred conservation of proteins that interact on a high level (Fig. 1a). Significant Pearson's and Spearman's rank coefficients support our qualitative observations [see Additional file 1].

### Evolutionary retention of interacting pairs of proteins

While we find that the conservation of single proteins is a function of connectedness we wonder if topology also contains such evolutionary signals on the level of interactions. Because proteins which are placed in cohesive areas (i.e. modules) tend to be evolutionary conserved we wonder if their interactions are conserved too. We utilize a link-based clustering coefficient that reflects the degree of clustering of an interaction's immediate network neighborhood, a topological measure that allows for correlations between local clustering and the actual reliability of observed interactions [38]. Similar to the single protein case, we grouped all interactions according to their hypergeometric clustering coefficient C_{vw} and determined the respective fraction of interacting pairs that are fully conserved as putative orthologs in each bin. In the absence of a correlation between evolutionary conservation and an interactions placement in the network the ratio of the C_{vw}-dependent and random fractions of orthologous protein pairs – defined as the interaction based excess retention $ER_{C_{vw}}$ (see Materials and Methods) – would be unity. Logarithmically binning all interactions according to their local degree of clustering C_{vw} and determining the average excess retention $ER_{C_{vw}}$ in each bin we identify a significant and systematic trend of proteins engaged in highly clustered interactions to be preferably evolutionary conserved [Fig. 1b, see Additional file 1]. These link-based observations are not only consistent with previous node-based results but also allow to suggest that standard single-node measurements of evolutionary conservation can be extended to their neighboring links. This evolutionary corollary indicates that not only single proteins are a target of evolution but also the interactions between conserved proteins.

Perturbation analysis

To demonstrate this gain of evolutionary information, we simulated the impact of extremely high false negatives rates of protein interactions by removing up to 70% of experimentally determined links between randomly selected protein pairs. Additionally, to address the effects of false positives, we randomly distributed up to 70% more interactions than were previously identified in the original Yeast network.

Because there are no significant differences in the distributions of organism-specific excess retention in Fig. 1, we examine orthologs of *C. elegans* as a representative comparative set for these analyzes. After generating 1, 000 different realizations to each case of incomplete false data, we determine the excess retention ER of proteins and their interactions that have orthologs in *C. elegans* according to their degree k and hypergeometric clustering coefficient C_{vw}. Determining the average excess retention in bins of increasing logarithmic size we find that the relationship between excess retention, level of interaction and local clustering is widely unaltered (single proteins: Figs. 2a,b; protein interactions: Figs. 2c,d). Addressing the statistical significance of these results we determined correlation
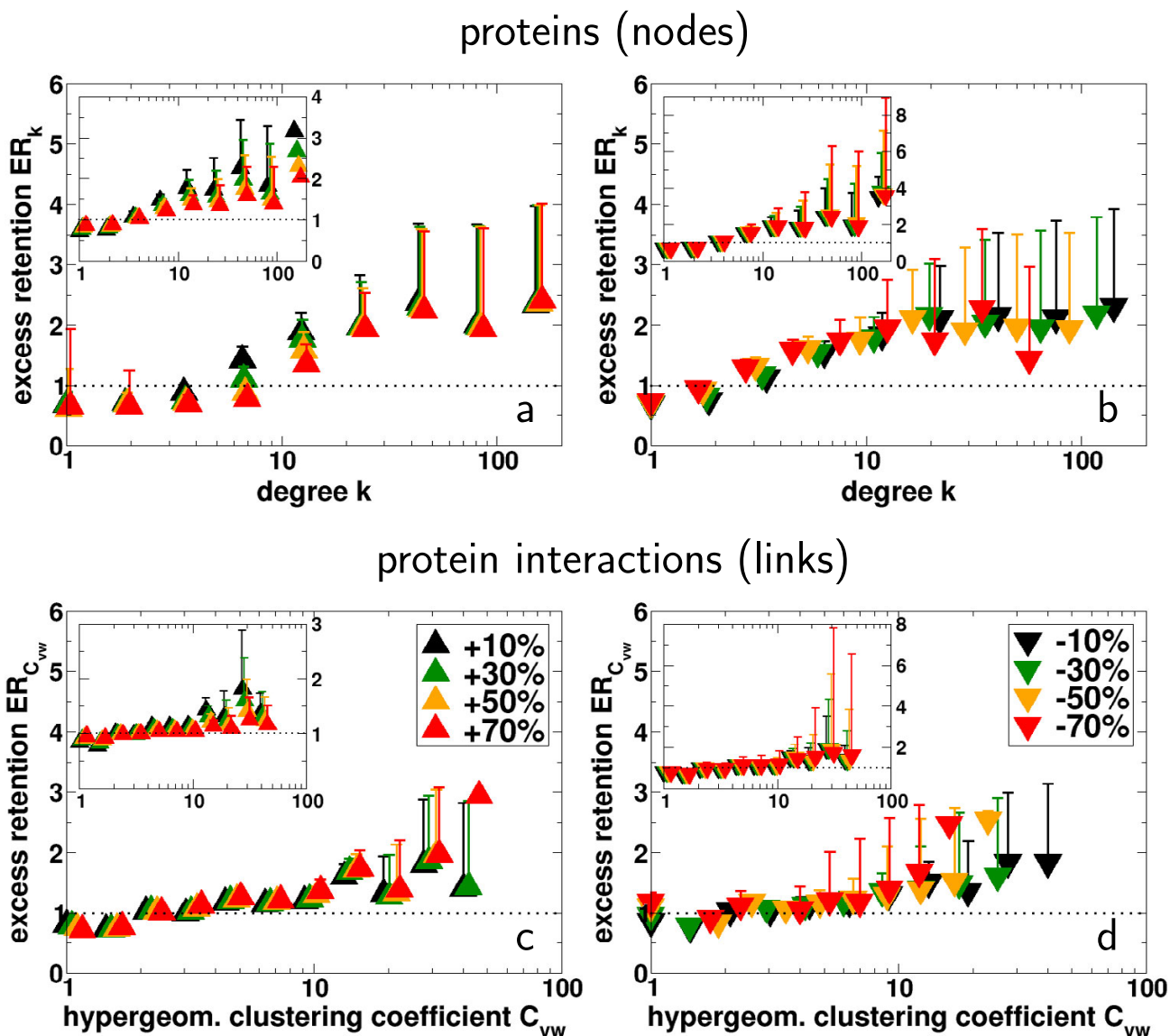
Figure 2
(a) In order to assess the impact of severely inconsistent protein interaction data, we removed 10–70% of interactions between randomly selected protein pairs, mimicking false negatives. **(b)** Simulating the effects of false positives, we randomly added 10–70% more interactions than originally present in the network. In each case, we averaged the degree dependent excess retention of interacting proteins ER_k that have orthologs in *C. elegans* over 1, 000 different samples. Analogously, we assessed the consequences of false negative orthologs by eliminating 10–70% of the proteins present in the set of worm orthologs ((a) inset). Mimicking the presence of false positive orthologs we labeled 10–70% more proteins as orthologs in worm that were originally present ((b) inset). Analogously, we test the robustness of the trend that highly clustered interactions (as exemplified by the hypergeometric clustering coefficient C_{vw}) are indeed predominantly conserved (as exemplified by a link based excess retention $ER_{C_{vw}}$) toward the presence of false negative **(c)** and false positive **(d)** interactions and orthologs (insets). In each case, we observe that the initial (empirically derived) ascending trend prevails, results which are further supported by strong and significant Pearson's and Spearman's rank correlation coefficients and Kolmogorov-Smirnov scores [see Additional file 1]. Error bars indicate the standard deviations from the mean excess retention in each bin.
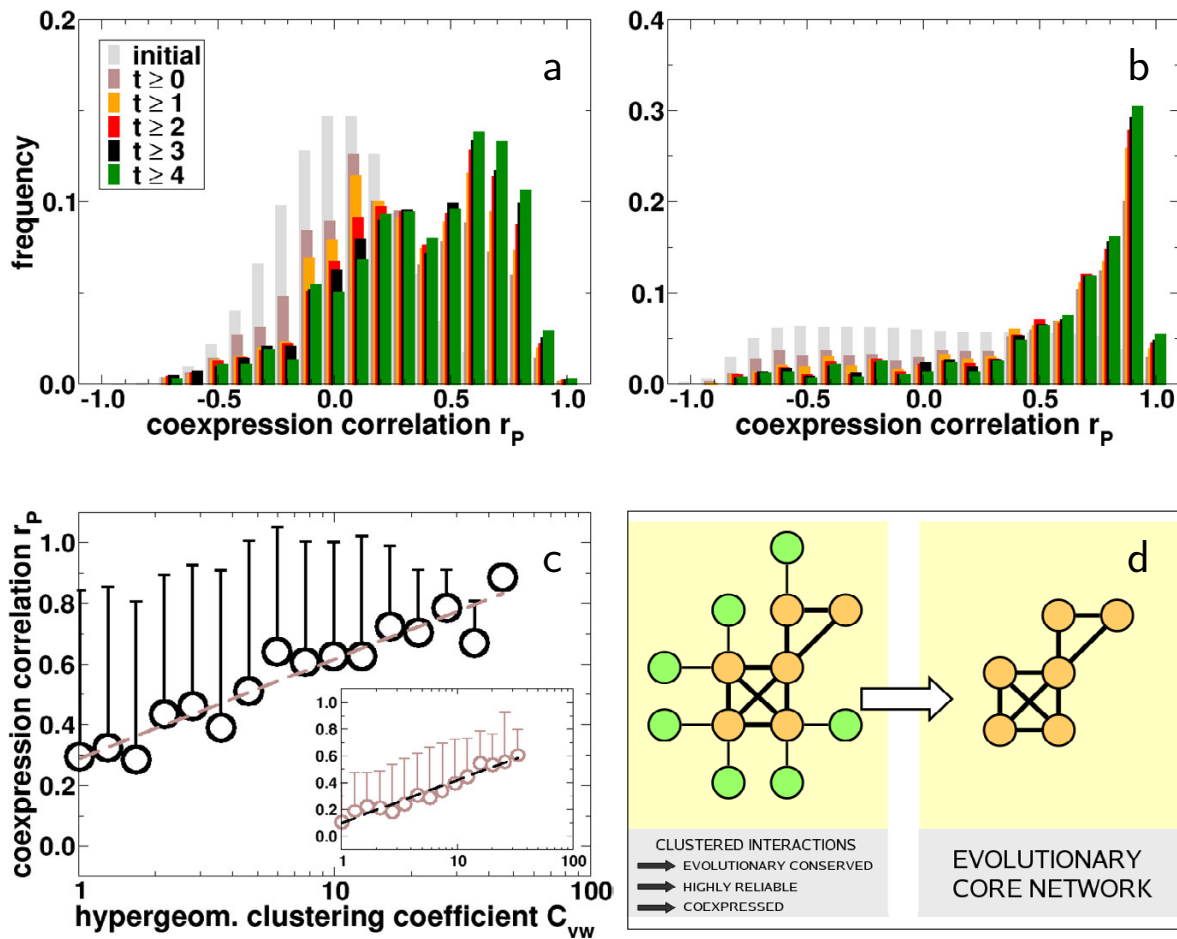
Figure 3
(a) Accounting for the expression coefficients of all protein pairs in Yeast, we observe a bell shaped frequency distribution of the expression correlation coefficient r_P, peaking around 0.0 (initial). Focusing on interacting pairs of Yeast proteins that both have an ortholog in *P. falciparum* and score above a certain threshold t of the respective hypergeometric clustering coefficient C_{vw} we observe shifted frequency distributions of the expression correlation coefficients r_P. This observation is further indicated by significant Students t-test scores [see Additional file 1]. **(b)** Assuming that those interactions constituted by Yeast proteins with an ortholog are conserved as well in *P. falciparum*, we determined frequency distributions of expression correlation coefficients r_P in this organism. Similarly to (a), we find significant shifts toward enforced coexpression if we focus on conserved interactions in Plasmodium that were embedded in an increasingly cohesive neighborhood in Yeast. Comparing our results to a background distribution of coexpression correlations of all protein pairs in *P. falciparum* Students t-test scores indicate the significance of our results [see Additional file 1]. **(c)** The significant shifts toward elevated levels of coexpression allow us to assume that there exists a pronounced correlation between the local cohesiveness of an interaction and the tendency that the involved proteins are coexpressed. Logarithmically binning data points according to their hypergeometric clustering coefficient C_{vw} in Yeast, we determined the mean expression correlation r_P in each bin, allowing us to observe a positive and significant trend that interacting Yeast proteins having an ortholog in *P. falciparum* are increasingly coexpressed if they are placed in a cohesive neighborhood (inset, Pearson's $r = 0.38$, $P = 8.8 \times 10^{-44}$, Spearman's rank $\rho = 0.41$, $P = 2.8 \times 10^{-47}$). Similarly, we observe qualitatively the same trend considering interactions in *P. falciparum* that have been inferred from Yeast protein interactions which have an ortholog in *P. falciparum* (Pearson's $r = 0.25$, $P = 1.9 \times 10^{-27}$, Spearman's rank $\rho = 0.27$, $P = 2.1 \times 10^{-30}$). Error bars indicate the standard deviation from the mean coexpression coefficient in each bin. **(d)** Concluding, we observe a perturbation persistent coincidence of (i) coexpression of interacting proteins, (ii) an enhanced clustering of their immediate neighborhood and (iii) their elevated tendency to be evolutionary conserved (yellow circles) in *S. cerevisiae*. Since high clustering around a certain protein interaction coincide well with an elevated reliability the integration of knowledge about the local clustering of an interacting pair of conserved proteins and their tendency to be coexpressed can be used to infer evolutionary core protein interactions in other organisms for which orthologs can be identified.

coefficients (Pearson's r and Spearman's rank ρ) and performed Kolmogorov-Smirnov tests, recovering significant similarities between the original and perturbed distributions [see Additional file 1].

Moreover, to represent missed orthologs, we randomly eliminated up to 70% from the set of Yeast proteins that have an ortholog in *C. elegans*. In turn, we randomly labeled up to 70% more proteins as orthologs in *C. elegans* than were previously present in the initial set. Sampling 1, 000 different realizations each, we calculated the excess retention according to the proteins degree k and local clustering around each interaction C_{vw}. Logarithmically binning the results thus obtained we averaged the excess retention of orthologous proteins in each bin, allowing us to find that the introduction of noise on the level of orthologs determination does not alter our initial observations (single proteins: insets Figs. 2a,b; protein interactions: insets Figs. 2c,d). Significantly similar correlation coefficients and Kolmogorov-Smirnov scores [see Additional file 1] support our conclusions.

Clustering, coexpression and evolutionary conservation

The observation that highly clustered links between evolutionarily conserved proteins are reliable and stable toward severe perturbation enhances our expectation that an elevated degree of coexpression of interacting proteins will retain this relationship as well. In particular, a strong coexpression signal of the orthologs of proteins that embrace the interactions in question would strongly indicate the actual presence of the interaction in a reference and target organism. As a test case, we extend our investigations to the malaria parasite *Plasmodium falciparum*, a single celled organism that has 895 putative orthologous proteins with Yeast. Elucidating those Yeast interactions between yeast proteins conserved in *P. falciparum*, we find a web of 3, 071 interactions among 659 proteins in *P. falciparum*. To evaluate the quality of these inferred interactions we utilized a comprehensive set of *P. falciparum* specific coexpression data [39] to calculate Pearson's correlation coefficients r_P for the inferred protein interactions. In the same way, we utilized an extensive set of Yeast coexpression data [40] to investigate the coexpression tendency of those interacting Yeast proteins that served as the template for the inferred interactions in Plasmodium. In both cases, we use the Yeast specific C_{vw} values as an approximate measure of an interaction's reliability. Focusing on interactions that score above increasing thresholds of $C_{vw'}$ we observe a strong shift toward coexpression of the considered protein pairs (Fig. 3a,b). The difference between the individual coexpression patterns is further indicated by significant Students t-test scores when the C_{vw} dependent distributions are compared to a background distribution of coexpression coefficients of all protein pairs in the considered organism [see

Additional file 1]. The significant shifts toward elevated levels of coexpression identify a pronounced correlation between the local cohesiveness of an interaction and the tendency that the involved proteins are coexpressed. The determination of mean coexpression coefficients r_P of interactions that have been logarithmically grouped according to their C_{vw} allows us to find a statistically significant trend toward elevated levels of coexpression of conserved yeast interactions that are placed in highly clustered neighborhood (inset, Fig. 3c). Although we adopt measurements of the local cohesiveness around links C_{vw} from Yeast, we find that the corresponding interactions in Plasmodium exhibit a similar trend (Fig. 3c). Remarkably, the latter distribution exhibits better and statistically more significant correlation coefficients than its template in Yeast [see Additional file 1].

In the same way we investigated the stability of the interactions propensity to be evolutionary conserved, we checked for the robustness of the obtained correlation between local clustering and coexpression. Mimicking the presence of false positive/negative links we randomly eliminated/added up to 70 % of interactions in the Yeast interaction network. Recalculating the hypergeometric clustering coefficient for each of 1,000 runs, we grouped all interacting pairs of Yeast proteins with an ortholog in *P. falciparum* according to C_{vw} in bins of logarithmically increasing size. Averaging over the respective coexpression correlation coefficient r_P of all Yeast interactions in each bin, we observe that the initial ascending trend prevails [see Fig. 1ab of Additional file 1]. Assuming that all interactions between proteins that have an ortholog in Plasmodium were conserved we repeated this procedure by superimposing and averaging over the respective coexpression correlation values of Plasmodium. Similar to the Yeast specific case, we observe the same qualitative trends [see Fig. 2ab of Additional file 1]. Significant correlation coefficients and Kolmogorov-Smirnov scores support our observations of our findings. In the same way, we simulated the presence of false positive /negative orthologs by eliminating/adding up to 70% of orthologs in *P. falciparum*. Averaging over 1,000 runs each, we determined the average coexpression coefficient in each bin utilizing Plasmodium and Yeast specific coexpression data. In each case, we find that the original trend of high local clustering around interactions coincides with an increased propensity to be coexpressed strongly prevails [see Figs. 1, 2bc of Additional file 1], observations that are supported by significant correlation coefficients and Kolmogorov-Smirnov scores [see Additional file 1].

Discussion & conclusion

Extending a previous study indicating that highly interacting proteins are predominantly conserved in evolution we generalize the concept that evolutionary signals are car-

ried by the topology of the underlying protein interaction network. In particular, a protein's propensity to be conserved while interacting with a high number of partners – a node-based evolutionary signal – has a link based counterpart, as indicated by the propensity of interacting proteins to be evolutionary conserved with increasing local clustering around the interaction in question. Although the obtained correlations are significant, the alarmingly high error rates in the determination of protein interactions cast doubt on the obtained results.

By focusing on perturbation events on node and interaction levels, we observe that extreme error rates of both protein interactions and orthologs do not ablate the evolutionary signal carried by the network structure. The introduction of noise at the node, by simulation of inconsistent determination of orthologs, does not override the preference of highly connected nodes to be evolutionary conserved; as theoretically predicted, random perturbations will rarely affect a hub in a scale-free network [41]. The low probability that a hub is hit by a random perturbation event also explains that interacting proteins that are placed in a highly clustered environment retain their evolutionary signal. Indeed, the definition of the hypergeometric clustering coefficient assures a high score for interacting proteins that share a lot of their interaction partners.

On an interaction level, we observe that the massive insertion/deletion of links does not obliterate the local structure of networks as indicated by the stable preference of highly connected proteins and protein pairs that are embedded in a well clustered neighborhood to be evolutionary conserved. In particular, we conclude that insertion/deletion of random links on average impact sparsely connected parts of the networks much more than densely connected ones; indeed, loss of information in highly clustered neighborhoods and highly connected hubs would require massive, targeted deletion/insertion of links to obliterate their local structure. Therefore, the observation that links which are placed in a highly clustered neighborhood are highly reliable [38] is nested in our observation that highly clustered neighborhoods compensate severe random perturbations much better than sparsely connected ones.

While our results allow us to conclude that degree alone is a robust indicator for a proteins propensity to be evolutionary conserved, the inherent topological robustness of locally clustered links emphasizes the emergent role of cohesive areas [26] as mediators of evolutionary information. In the simplest case, we confirmed that not only single proteins are a potential target of evolution but interaction among them can be potentially conserved as well. As a strong indicator that an interaction indeed has

been conserved, the correlation between high local clustering and evolutionary conservation is accompanied by a stable elevated degree of coexpression of the interacting proteins in both a model and target organism. Superimposing the extreme error rates simulating the incoherent determination of orthologs and interactions as well we see that trends in both the model and target organism prevail, strongly indicating that evolution also happens on the level of interactions and putative bundles of interactions.

Although we utilized very noisy and inconsistent data of protein interactions and putative orthologs, we see that high connectivity and high clustering on average harbor significantly more evolutionary relevant information that sparsely connected and clustered areas. The coincidence of (i) high local clustering around highly reliable interactions of proteins, (ii) their propensity to be evolutionary conserved, (iii) their tendency to be coexpressed even in the face of tremendous experimental noise sketches a hypothetical framework to infer an evolutionary core of single protein-protein interactions by elucidating interacting proteins of a reference organism that have orthologs in the targeted organism. The quality of an interaction is assessed by calculating the corresponding hypergeometric clustering coefficient. Choosing the highest scoring – thus most reliable – ortholog interaction allows the selection of a core interaction network in the targeted organism. Unlike our case, where evolutionary relationships between proteins were approximated by similarity searches, the quality of predicted interactions will be enhanced by utilizing more sophisticated methods (such as tree-base methods) which allow a more reliable assignment of orthology. Finally, the cross-validation with high resolution coexpression data can refine specific protein-protein interaction subnetworks, allowing for checks of the actual presence of a proposed interaction. Ultimately, such a framework would allow a first insight into evolutionary conserved parts in interactomes of organism for which no interaction data currently exists.

Methods
Protein interactions
As a source of protein interactions we chose the DIP database [36] which provides a set of manually curated protein-protein interactions in the organism *S. cerevisiae*. The current version contains 3, 833 proteins involved in 11, 942 interactions derived from combined, non-overlapping data which are mostly obtained from the high-throughput application of the two-hybrid method.

Assignment of orthology
Orthologs are genes in different species that originate from a single gene in the last common ancestor of these species. Such genes often have retained identical biological roles in present day organisms, indicated by a high

degree of sequence homology. Unfortunately, orthology analysis between organisms is often difficult and error prone because of large numbers of paralogs within protein families. As a source of reliable and robust information about orthologous relationships between proteins in different species we utilized the InParanoid database [37,42] which provides putative orthologous sequence information for *S. cerevisiae* and numerous other organisms. The algorithm for assigning orthologous relationships is based on pairwise similarity scores which are by default calculated with the BLASTP program. Best pairwise hits between the proteomes of two species are seeds – labeled as the main ortholog groups – of orthologous protein sequence clusters. In a further step, other sequences are added to this group if they are closely homologous to one of the main orthologs, members of orthologous groups which are called in-paralogs. In a final quality checking step, confidence values for each ortholog and in-paralog is determined allowing the detection of putative orthologous relationships that has been only reliably possible by multiple alignments and phylogenetic trees previously [37]. In our study, we considered the main ortholog pairs of each orthologous group as sequences that are putatively orthologous to each other allowing us to obtain 1, 928 Yeast proteins with orthologs in *H. sapiens*, 2,073 in *A. thaliana*, 1, 885 in *C. elegans*, 1, 885 in *M. musculus*, 1,631 in *D. melanogaster* and 895 in *P. falciparum*.

Hypergeometric clustering coefficient

Recently, a network topology based approach uncovered a remarkable correlation between enhanced quality of protein interactions and the degree of clustering of their immediate network neighborhood [38]. Considering a protein-protein interaction network with N nodes, we define the hypergeometric clustering coefficient as

$$C_{vw} = -\log \sum_{i=|N(v)\cap N(w)|}^{\min(|N(v)|,|N(w)|)} \frac{\binom{|N(v)|}{i}\binom{N-|N(v)|}{|N(w)|-i}}{\binom{N}{|N(w)|}} \quad (1)$$

where $N(x)$ represents the neighborhood of a vertex x. Given fixed neighborhood sizes $N(v)$ and $N(w)$ of proteins v and w, the hypergeometric clustering coefficient increases with elevated overlap between the protein's neighborhoods. Provided that the neighborhoods are independent, the summation can be interpreted as a p value, reflecting the probability of obtaining a number of mutual neighbors between proteins v and w at or above the observed number by chance.

Orthologous excess retention

According to their hypergeometric clustering coefficient C_{vw} of the interactions they are involved in, we grouped all

interactions in groups of same C_{vw} that have been rounded to integers. For each group of $N_{C_{vw}}$ proteins, the fraction of interacting pairs of proteins that both have an ortholog in an other organism is defined as $e_{C_{vw},o} = n_{C_{vw},o} / N_{C_{vw}}$. In the absence of a correlation between evolutionary conservation of interacting protein pairs and their position in the network, $e_{C_{vw},o}$ has the general C_{vw}-independent value $e_o = n_o/N$, where n_o is the total number of interactions between Yeast proteins that have an ortholog, and N is the total number of Yeast protein interactions in the underlying network. Thus, we define the clustering-dependent excess retention of such proteins as $ER_{C_{vw},o} = e_{C_{vw},o} / e_o$ which has the C_{vw}-independent value $ER_{C_{vw},o} = 1$ for a random distribution of orthologous proteins [20]. Basically, we applied the same framework for single proteins, by grouping them according to their degree k. For each group of N_k proteins, the fraction of proteins that also have an ortholog is defined as $e_{k,o} = n_{k,o}/N_k$. Analogously, the node based excess retention ER_k is defined as $ER_k = e_{k,o}/E_k$, where E_k is the ratio of all proteins with an ortholog in the whole network.

Coexpression data

To evaluate the quality of these inferred interactions we utilized a comprehensive set of Plasmodium specific [39] and Yeast specific [40] coexpression data. In each dataset, we utilized the expression profiles to determine the respective Pearson's correlation coefficient r_P for each interacting pair of proteins.

Logarithmic binning

To guarantee balanced sampling of our distributions we generally use logarithmic binning of the respective x-axis, a procedure for curve estimation that corrects for the skewed nature of the scale-free distribution.

On a logarithmic scale, we define the bin size $\Delta = \frac{1}{N}\log\left(\frac{b}{a}\right)$, where N corresponds to the selected number of bins. Values a and b refer to the minimal and maximal value of data points on the x-axis, $b = \max_i(x_i)$ and $a = \min_i(x_i)$. Thus, $n_i = \log\left(\frac{x_i}{a}\right)/\Delta$, $n_i \in [0, N-1]$ reflects the number of the bin we assign a data point with a x_i coordinate. Representing the n_ith bin on the x-axis, we place x_{n_i} at the end of each bin using $x_{n_i} = ae^{\Delta(n_i+1)}$.

The advantage of logarithmic binning is an elevated degree of noise reduction which is dependent on the bin size [41,43]. Although this procedure causes a loss of accuracy, we still uncover the buried trends to a satisfying extent applying our statistical methods on the binned data.

Additional material

Additional File 1

contains detailed statistical measurements of presented correlations. Furthermore, we show results of the perturbation analysis of coexpression coefficients as briefly addressed in the main paper.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-6-8-S1.pdf]

References

1. Rain JC, Selig L, DeReuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schächter V, Chemama Y, Labigne A, Legrain P: **The protein-protein interaction map of Helicobacter pylori.** *Nature* 2001, **409:**211-215.
2. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A: **Interaction network conatining conserved and essential protein complexes in Escherichia coli.** *Nature* 2005, **433:**531-537.
3. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y: **Towards a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins.** *Proc Nat Acad Sci USA* 2000, **97(3):**1143-1147.
4. Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast.** *Nature Biotechn* 2000, **18:**1257-1261.
5. Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshorn D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg J: **A comprehensive analysis of protein-protein interactions of Saccharomyces cerevisiae.** *Nature* 2000, **403:**623-627.
6. Gavin A, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick J, Michon AM, Cruciat CM, Remor M, Böfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley R, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415:**141-147.
7. Ho Y, Gruhler A, Heilbut A, Bader G, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutillier K, coauthors: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.** *Nature* 2002, **415:**180-183.
8. Walhout A, Sordella R, Lu X, Hartley J, Temple G, Brasch M, Thierry-Mieg N, Vidal M: **Protein interaction mapping in C. elegans using proteins involved in vulval development.** *Science* 2000, **287:**116-122.
9. Giot L, Bader J, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao Y, Ooi C, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath
    K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon C, Finley R Jr, White K, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets R, McKenna M, Chant J, Rothberg J: **A Protein Interaction Map of Drosophila melanogaster.** *Science* 2004, **302:**1727-1736.
10. Barabaś A, Oltvai Z: **Network Biology: Understanding the Cell's Functional Organization.** *Nature Rev Gen* 2004:101-113.
11. Barabási A, Albert R: **Emergence of Scaling in Random Networks.** *Science* 1999, **286:**509-512.
12. Albert R, Barabási AL: **Statistical mechanics of complex networks.** *Rev Mod Phys* 2002, **74:**47.
13. Jeong H, Mason S, Barabási AL, Oltvai Z: **Lethality and centrality in protein networks.** *Nature* 2001, **411:**41-42.
14. Han J, Bertin N, Hao T, Goldberg DS, Berriz G, Zhang L, Dupuy D, Walhout A, Cusick M, Roth F, Vidal M: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network.** *Nature* 2004, **430:**88-93.
15. Rives A, Galitski T: **Modular organisation of cellular networks.** *Proc Natl Acad Sci USA* 2003, **100:**1128-1133.
16. Spirin V, Mirny L: **Protein complexes and functional modules in molecular networks.** *Proc Natl Acad Sci USA* 2003, **100:**12123-12128.
17. Wuchty S, Almaas E: **Peeling the Yeast Interaction Network.** *Proteomics* 2005, **5:**444-449.
18. Snel B, Bork P, Huynen M: **The identification of functional modules from genomic association of genes.** *Proc Natl Acad Sci USA* 2002, **99:**5890-5895.
19. Wuchty S: **Interaction and Domain Networks of Yeast.** *Proteomics* 2002, **2:**1715-1723.
20. Wuchty S: **Topology and Evolution in Yeast Interaction Networks.** *Genome Res* 2004, **14:**1310-1314.
21. Fraser H, Hirsh A, Steinmetz L, Scharfe C, Feldman M: **Evolutionary Rate in the Protein Interaction Network.** *Science* 2002, **296:**750-752.
22. Fraser H, Wall D, Hirsh A: **A simple dependence between protein evolution rate and the number of protein-protein interactions.** *BMC Evol Biol* 2003, **3(11):**.
23. Jordan I, Wolf Y, Koonin E: **No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly.** *BMC Evol Biol* 2003, **3:**.
24. Jordan I, Wolf Y, Koonin E: **Correction: No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly.** *BMC Evol Biol* 2003, **3(5):**.
25. Williams E, Hurst L: **The evolution of linked genes evolve at similar rates.** *Nature* 2000, **407:**900-902.
26. Wuchty S, Oltvai Z, Barabási AL: **Evolutionary conservation of motif constituents within the yeast protein interaction network.** *Nature Genetics* 2003, **35:**176-179.
27. Vespignani A: **Evolution thinks modular.** *Nature Gen* 2003, **35:**118-119.
28. von Mering C, Zdobnov E, Tsoka S, Ciccarelli F, JB Pereira-Leal CO, Bork P: **Genome evolution reveals biochemical networks and functional modules.** *Proc Natl Acad Sci USA* 2003, **100:**15428-15433.
29. Ge H, Liu Z, Church G, Vidal M: **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae.** *Nature Genetics* 2001, **29:**482-486.
30. Babu M, Luscombe N, Aravind L, Gerstein M, Teichmann S: **Structure and evolution of transcriptional regulatory networks.** *Curr Opin Struct Biol* 2004, **14:**283-291.
31. Sharan R, Suthram S, Kelley R, Kuhn T, McCuine S, Uetz P, Sittler T, Karp R, Ideker T: **Conserved patterns of protein interaction in multiple species.** *Proc Natl Acad Scie USA* 2005, **102:**1974-1979.
32. Vidal M: **Interactome modelling.** *FEBS Lett* 2005, **579:**1834-1838.
33. Bork P, Jensen L, von Mering C, Ramani A, Lee I, Marcotte E: **Protein interaction networks from yeast to human.** *Curr Opin Struct Biol* 2004, **14:**292-299.
34. Vazquez A, Flammini A, Maritan A, Vespignani A: **Modeling of Protein Interaction Networks.** *ComPlexUs* 2003, **1(38):**38-44.
35. Von Mering C, Krause R, Snel B, Cornell M, Oliver S, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2003, **417:**399-403.

36. Xenarios I, Salwinski L, Duan X, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.** *Nucl Acids Res* 2002, **30:**303-305.
37. Remm M, Storm C, Sonnhammer E: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314:**1041-1052.
38. Goldberg D, Roth F: **Assessing experimentally derived interactions in a small world.** *Proc Natl Acad Sci USA* 2003, **100:**4372-4376.
39. Bozdech Z, Llinas M, Pulliam B, Wong E, Zhu J, DeRisi J: **The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum.** *PLoS Biology* 2003, **1:**1-16.
40. Eisen M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Nat Acad Sci* 1998, **95:**14863-14868.
41. Albert R, Jeong H, Barabási A: **Error and attack tolerance of complex networks.** *Nature* 2000, **406:**378-382.
42. O'Brien K, Remm M, Sonnhammer E: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucl Acids Res* 2005, **33:**D476-D480.
43. Goldstein M, Morris S, Yen G: **Fitting to the Power-Law Distribution.** 2004 [http://arxiv.org/abs/cond-mat/0402322].