Protein Networks Reveal Detection Bias and Species Consistency When Analysed by Information-Theoretic Methods

Luis P. Fernandes[1], Alessia Annibale[2], Jens Kleinjung[3], Anthony C. C. Coolen[1,2], Franca Fraternali[1]*

1 Randall Division of Cell and Molecular Biophysics, King's College London, London, United Kingdom, **2** Department of Mathematics, King's College London, London, United Kingdom, **3** Division of Mathematical Biology, MRC National Institute for Medical Research, London, United Kingdom

Abstract

We apply our recently developed information-theoretic measures for the characterisation and comparison of protein–protein interaction networks. These measures are used to quantify topological network features *via* macroscopic statistical properties. Network differences are assessed based on these macroscopic properties as opposed to microscopic overlap, homology information or motif occurrences. We present the results of a large–scale analysis of protein–protein interaction networks. Precise null models are used in our analyses, allowing for reliable interpretation of the results. By quantifying the methodological biases of the experimental data, we can define an information threshold above which networks may be deemed to comprise consistent macroscopic topological properties, despite their small microscopic overlaps. Based on this rationale, data from yeast–two–hybrid methods are sufficiently consistent to allow for intra–species comparisons (between different experiments) and inter–species comparisons, while data from affinity–purification mass–spectrometry methods show large differences even within intra–species comparisons.

Citation: Fernandes LP, Annibale A, Kleinjung J, Coolen ACC, Fraternali F (2010) Protein Networks Reveal Detection Bias and Species Consistency When Analysed by Information-Theoretic Methods. PLoS ONE 5(8): e12083. doi:10.1371/journal.pone.0012083

Editor: Kumar Selvarajoo, Keio University, Japan

Received June 3, 2010; **Accepted** July 6, 2010; **Published** August 18, 2010

Copyright: © 2010 Fernandes et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), Medical Research Council (MRC) (grant U.1175.03.001.00001.01) and the Leverhulme Trust (grant F/07040/AL to FF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: franca.fraternali@kcl.ac.uk

Introduction

Comparative genomics has revolutionised the study of biology by shifting its focus from component characterisation to the study of systemic properties. We envisage that in the near future the comparison of interactomes of different species might drive a similar, if not even more powerful transformation [1,2]. Interaction maps (otherwise here named interactomes, referring to the entire set of molecular interactions in the cell) can reveal important mechanistic principles that may guide further progress in the understanding of cellular function, and of dysfunction leading to disease [3–5]. One would expect that the availability of interactome maps for several organisms could give new insights into how biological diversity is embedded in the networks' functionality [6]. In contrast to genomic data, however, the available interactome data are still far from complete and of limited reproducibility [7,8]. One can compare protein-protein interaction network (PPIN) datasets by simply counting the fraction of common interactions, referred to as 'overlap'. However, the overlap values found are typically small, which prohibits a meaningful comparison [9]. Alternative approaches have therefore been proposed. Some focus on identifying conserved 'modules' or recurrent geometrically defined motifs, envisaged to capture biological and functional properties of the underlying networks [10,11] or common functional cores of ancestral origin [12,13]. Others employ alignment strategies where phylogenetic information is derived by the identification of paralogues [14,15]. These studies illustrate the additional information provided by comparative interactomics, beyond comparative genomics, and the benefit of intra-species comparison [16].

However, to progress further in comparative interactomics, a serious problem needs to be resolved. Recent analyses of PPINs sparked a debate about the influence of the experimental method on the quality and biological relevance of the interaction data [17]. Current experimental techniques, such as yeast two-hybrid (Y2H) and co–affinity purification combined with mass spectrometry (AP–MS), sample subsets of the interaction data space [18]. These subsets show very limited overlap [7]. Moreover, AP–MS interaction data are non-binary by nature for any multi-component complex; their conversion to binary pair-interactions is non–trivial and relies on processing protocols that may introduce further biases in the final screening output [17,19]. It is vital that we understand to what extent observed discrepancies between different networks reflect sampling biases of their experimental methods, as opposed to topological features due to biological functionality.

In information-theoretic terms overlap is not a good measure of the similarity between two sampled networks, just as the size in bits of a file does not give its true information content. It is therefore natural to explore the potential of information-theoretic measures for comparing interaction networks. These require a systematic characterisation of network topologies, which is a general prerequisite in network science [20], and formulations in terms

of network sample probabilities, based on macroscopic topological features. One is thus led to study the relationship between structured random graph ensembles and real biological signalling networks. The rationale is that PPIN data should be regarded as noisy samples of a true underlying network, and that the family of such samples is best described and studied statistically as a structured random graph ensemble with controlled macroscopic topological features. If the control parameters of the random graph ensemble can be derived from sufficiently accurate and complete network data, it is in principle possible to calculate (asymptotically) explicit formulae for entropies and complexities, and for information-theoretic distances between network families.

In recent years there have been efforts to define and generate random graphs whose topological features can be controlled and tailored to experimentally observed networks. In Perez-Vicente and Coolen [21] a parameterised random graph ensemble was defined where graphs have not only a prescribed degree distribution but also prescribed degree correlations. We have recently been able to show [22] that this graph ensemble (described in [21]) can be tailored asymptotically to the generation of graphs with any prescribed degree distribution and any prescribed degree correlation function. Moreover, for this ensemble the combinatorial problem of calculating the network complexities and information-theoretic distances between network families can be solved analytically. The result is a novel, practical and precise mathematical framework, that allows for the large-scale analysis and unbiased comparison of PPINs from different species and measured with different techniques. Here we apply this formalism to an extensive range of PPINs, and show that it provides a quantitative window on interactome data. The topological network distance is applied here to cluster network data and to estimate intra-species similarity for differently detected interaction data and inter-species distances within and between experimental methods. In particular, the presence of methodological data biases and the topological similarity between networks with small microscopic overlap can be detected clearly and at low computational cost.

Results

The PPIN data taken from literature

In the table of Fig. 1 we give a comprehensive table of all the PPIN data that are used in the analysis presented in this paper, colour coded according to the experimental method that was used. The table lists for each PPIN dataset various simple quantitative characteristics, such as the number of nodes (NP), interactions (NI), protein coding genes (PCG) and the average (AD) and maximum (k_{max}) degree k, which is defined as the number of interaction partners of a node.

Macroscopic characterisation: degree statistics and degree–degree correlations

We characterise each PPIN by its degree distribution $p(k)$ and its normalised degree–degree correlations (DDCs) $\Pi(k, k')$ (the latter quantity gives the likelihood that two proteins with degrees k and k' interact, relative to what would be found in random networks with the same degree distribution but uncorrelated degrees). The precise definitions are given in the Methods section. A value $\Pi(k, k')>1$ indicates that protein pairs with degrees (k, k') interact more than what one would expect on the basis of their degree values, whereas if $\Pi(k, k')<1$ they interact less than expected; either case would signal topological information beyond that encoded in the degree statistics alone. We applied a weak Gaussian smoothening to these functions, to prevent probabilities from being strictly zero. The resulting numerical differences in the macroscopic quantities are irrelevant for the presented data. Various quantities have been proposed in the past for characterising the structure of networks. One reason for choosing the macroscopic features $p(k)$ and $\Pi(k, k')$ is that many of the previously proposed quantities are either similar or equivalent to (or expressible in terms of) $p(k)$ and $\Pi(k, k')$. Examples are degree sequences [23], degree distributions [24], degree correlations [25], and assortativity [26]. Some authors, however, used measures that are qualitatively different, such as clustering coefficients [27] and so-called community structures [28].

Before embarking upon an information-theoretic analysis of our PPIN datasets, based on the macroscopic topological features captured by $p(k)$ and $\Pi(k, k')$, we first verify that for these datasets the function $\Pi(k, k')$ actually contains topological information, *i.e.* deviates significantly from the value one. It would also be useful to know how these topological features may have evolved; one would expect that closely related species should also have PPINs with more similar topological features.

In Fig. 2 we show the normalised DDCs for the bacterial species in our dataset collection in heat map representation, with a colour scale ranging from black ($\Pi(k, k')$ close to zero) to white ($\Pi(k, k')$ very large). Since $\Pi(k, k')$ is a symmetric function, the plots are always symmetric around the main diagonal. The figure reveals that generally the normalised DDCs deviate significantly from those of random networks with the same degree statistics, where one would have found $\Pi(k, k')=1$ throughout (modulo small fluctuations). Apparently there is significant topological information contained in the degree correlations, and this is seen to give rise to quite diverse patterns for the different bacterial species. Some species (*e.g. Synechocystis*) appear to exhibit normalised DDCs mostly higher than the random level, some (*e.g. C. jejuni*) exhibit normalised DDCs that are mostly lower, whereas for *e.g. T. pallidum* one observes strong deviations from the random level in either direction. The most closely related bacterial species in our datasets are *H. pylori* and *C. jejuni*, which both belong to the *Campylobacterales* genus, yet this is not reflected in their DDC patterns. On the contrary, the *H. pylori* network exhibits only minor DDC deviations from the random level, unlike *C. jejuni*. Similarly, comparison of the networks of *C. jejuni*, *T. pallidum* and *H. pylori*, which all belong to the *Proteobacteria* phylum family (comprising the majority of gram-negative bacteria), does not reveal any conserved pattern.

Even more strikingly and worryingly, consistent DDC fingerprints are not even observed for plots that refer to datasets of the same species. In Fig. 3 we show the normalised DDCs for yeast, which has been the focus of most of the large-scale PPIN determinations so far. The plots in Fig. 3, displayed in order of experimental determination and date, do not suggest conservation of the macroscopic topological PPIN features.

A hint at a possible explanation emerges if one compares only plots that refer to the same experimental technique. The DDC patterns then appear more similar, differing mostly in terms of the strengths of the deviations from the random level, which increase roughly with the time of publication of the PPIN dataset. Compare for example *S. cerevisiae* II (core) to *S. cerevisiae* XII (both obtained *via* Y2H), and *S. cerevisiae* VIII to *S. cerevisiae* X (both obtained *via* AP–MS). The interactions reported on the *S. cerevisiae* X dataset were in fact derived from the raw purification data of two AP-MS datasets (*S. cerevisiae* VIII and *S. cerevisiae* IX), but these data were processed using a different scoring and clustering protocol. A strong positive correlation is observed along the diagonal for this latter dataset, indicating an enhancement of interactions between nodes of similar degree. In general, the AP-MS datasets show

Species	NP	PCG	NI	<k>	kmax	DM	Ref
C.elegans	2528	20176	3864	2.96	99	Y2H	Simonis *et al.* 2008
C.jejuni	1324	1736	11796	17.5	207	Y2H	Parrish *et al.* 2007
H.pylori	724	1587	1403	3.87	55	Y2H	Rain *et al.* 2001
H.sapiens I	1499	21370	2530	3.37	125	Y2H	Rual *et al.* 2005
H.sapiens II	1655	21370	3076	3.71	95	Y2H	Stelzl *et al.* 2005
M.loti	1803	7343	3094	3.43	401	Y2H	Shimoda *et al.* 2008
P.falciparum	1267	5385	2709	4.17	51	Y2H	Lacount *et al.* 2005
S.cerevisiae I	991	6532	948	1.82	24	Y2H	Uetz *et al.* 2000
S.cerevisiae II	787	6532	806	1.91	55	Y2H	Ito *et al.* 2001(core)
S.cerevisiae III	3241	6532	4367	2.69	279	Y2H	Ito *et al.* 2001
S.cerevisiae XII	1544	6532	1809	2.34	86	Y2H	Yu *et al.* 2008
Synechocystis	1903	3725	3100	3.25	51	Y2H	Sato *et al.* 2007
T.pallidum	724	1039	3627	10.01	285	Y2H	Titz *et al.* 2008
E.coli	2457	4246	8664	7.05	641	AP-MS	Arifuzzaman *et al.* 2006
H.sapiens III	2268	21370	6433	5.67	314	AP-MS	Ewing *et al.* 2007
S.cerevisiae IV	1576	6532	3617	4.58	62	AP-MS	Ho *et al.* 2002
S.cerevisiae VI	1358	6532	3220	4.73	53	AP-MS	Gavin *et al.* 2002
S.cerevisiae VIII	2551	6532	21394	16.77	955	AP-MS	Gavin *et al.* 2006
S.cerevisiae IX	2708	6532	7121	5.25	141	AP-MS	Krogan *et al.* 2006
S.cerevisiae X	1630	6532	9089	11.15	127	AP-MS	Collins *et al.* 2007
S.cerevisiae XI	1078	6532	2770	4.7	58	PCA	Tarassov *et al.* 2008
D.melanogaster	7286	14141	25102	6.85	176	DD	Stark *et al.* 2006
H.sapiens IV	9306	21370	35021	7.52	247	DD	Prasad *et al.* 2009
S.cerevisiae V	2617	6532	11855	9.05	118	DI	Von Mering *et al.* 2002
S.cerevisiae VII	1379	6532	2493	3.61	32	DI	Han *et al.* 2004

Yeast-two-Hybrid | Affinity Purification-Mass Spectrometry | Protein Complementation Assay | Database Datasets | Data Integration

Figure 1. Table of all 25 experimental PPIN datasets analysed in this study, corresponding to 11 different species. These included nine eukaryotic organisms (*Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Plasmodium falciparum* and *Saccharomyces cerevisiae*), and six bacterial species (*Campylobacter jejuni*, *Escherichia coli*, *Helicobacter pylori*, *Mesorhizobium loti*, *Synechocystis* and *Treponema pallidum*). Abbreviations stand for: NP, Number of Proteins; NI, Number of Interactions; PCG, Number of Protein Coding Genes; AD, Average Degree; k_{max}), Maximum Degree; Ref, References. Most datasets were derived from high–throughput experiments detected by either Y2H [29,46–55] or AP-MS [56–62]; we also included a recent PCA dataset [63]. In addition we analysed a series of consolidated datasets that include both high-throughput experiments and literature-mined small-scale studies [64,65]. The Ito et al. (2001) [47] dataset was divided into two sets: a high confidence set defined as the 'core' set and a low confidence set, as suggested by the authors. The Collins et al. (2007) [58] dataset consists of the raw purifications identified by the Krogan et al. (2006) [59] and Gavin et al. (2002) [56] studies, but re-analysed by a different scoring and clustering algorithm. Lastly, for completeness we have also included two commonly used yeast datasets: the Dong et al. (2004) [66] network, which is a consolidated dataset referred to in the literature as the 'Filtered Yeast Interactome' (comprising experimentally determined and *in silico* predicted interactions), and the von Mering et al. (2002) [67] dataset, which has been assembled from two catalogues of yeast protein complexes (the MIPS catalogue and the Yeast Protein Database catalogue).
doi:10.1371/journal.pone.0012083.g001

stronger DDC patterns than the Y2H datasets (this we also observed for the *H. sapiens* datasets) although the regions where the main deviations from the random level occur are quite different.

Assortativity

The assortativity of a network is a quantity that measures in a single number the extent to which the degrees of connected nodes are correlated with each other; it can be expressed in terms of $p(k)$ and $\Pi(k, k')$ *via* a simple formula.

Positive assortativity indicates positive correlation between the degrees of connected nodes (implying that nodes prefer to interact with other nodes of similar degree) while the contrary is true for negative assortativity (here high degree nodes prefer to interact with low degree nodes). We can therefore view and use the assortativity as a single parameter that summarises part of the full information provided in our DDC plots. To assess the relevance of any observed topological feature in a network, it must be compared to its frequency of observation in appropriate null models. These are benchmark networks, generated randomly and with uniform probabilities from the set of *all* networks that share specified features with the network under study. In this paper we choose as null models random networks that share with our biological PPINs the degree distribution $p(k)$. Many properties of these null models can be calculated analytically if the number of nodes is sufficiently large; for instance, lacking further topological structure, our null models would have $\Pi(k, k')=1$ for all (k, k'), and zero assortativity. In this section we compare our observations for each dataset to null models that have been generated *via* numerical simulations (by careful re-shuffling of the network under study; see the Methods section for a detailed description of the randomisation algorithm), to capture also finite size effects.

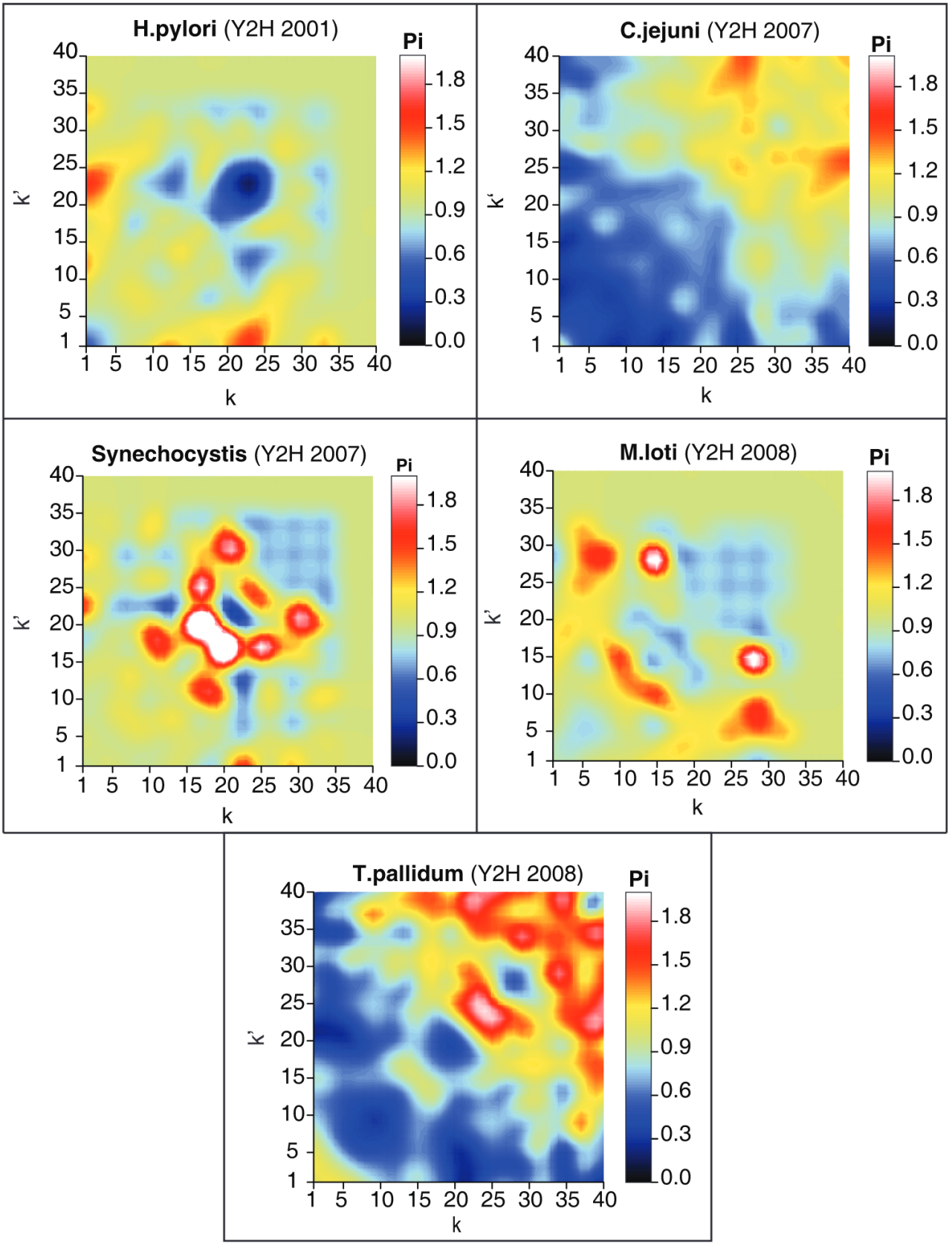

H.pylori (Y2H 2001)
C.jejuni (Y2H 2007)
Synechocystis (Y2H 2007)
M.loti (Y2H 2008)
T.pallidum (Y2H 2008)
Pi
1.8
1.5
1.2
0.9
0.6
0.3
0.0
k'
k
1 5 10 15 20 25 30 35 40

Figure 2. Heat map representations of the normalised DDC function $\Pi(k, k')$ of bacterial PPINs. For each combination of k and k', $\Pi(k, k')$ gives the likelihood that two proteins with degrees k and k' interact, relative to what would be found in appropriate null models (random networks with the same degree distribution but uncorrelated degrees). The degree axes k and k' were truncated to the value $k_{\max} = 40$. White regions indicate strongly enhanced propensity for protein-protein interaction (values of $\Pi(k, k')$ larger than expected on the basis of degree information alone) while dark regions indicate reduced protein-protein interaction (values of $\Pi(k, k')$ smaller than expected on the basis of the degree distribution alone).
doi:10.1371/journal.pone.0012083.g002

In Fig. 4 we plot in black the assortativities of our PPIN datasets (Original), together with those of their randomisations (null models) in green (Reshuffled). Most sets are seen to have slightly negative assortativity values, indicating a weak preference for interactions between nodes with different degrees. The main deviants from this trend are *S. cerevisiae* X, *S. cerevisiae* V and *S. cerevisiae* VII, with strong positive assortativity. This is consistent with Fig. 3, where the *S. cerevisiae* X dataset is indeed distinguished by the presence of consistently high values of $\Pi(k, k')$ along the main diagonal, signalling a strong preference for interactions between nodes with similar degrees. The assortativities of the null models (in green) are expected to be closer to zero than those of the real PPINs. This is indeed true for most cases, although for some networks (*e.g. M. loti*, *P. falciparum*, *E. coli*, and *S. cerevisiae* VIII) the assortativity differences between the original networks and their null models are negligible. In sufficiently large networks, all correctly generated null models would exhibit zero assortativity, so any deviation of the green line from zero in Fig. 4 must reflect finite size effects or effects caused by slow relaxation (see 'Definition and generation of null models' in the Methods) during the randomisation process (or both). In Fig. 4 the deviations are most likely due to finite size effects; this can be concluded upon measuring the Hamming distances between the original networks and their null models (which measure the extent of microscopic dissimilarity between the two, see Fig. S1), which show no evidence for insufficient relaxation in the null model generation.

Degree complexity and wiring complexity

We now turn to information-theoretic quantifiers of PPIN structure, applying the methods developed in Annibale et al. [22]. One of these is the network complexity, which (modulo finite size effects) measures the amount of topological information contained in a network's degree statistics and DDCs. It consists of two contributions, both of which can be expressed explicitly in terms of the functions $p(k)$ and $\Pi(k, k')$. The first is the degree complexity, measuring the information revealed by knowledge of $p(k)$ alone. The second is the wiring complexity, measuring the information revealed by subsequent knowledge of $\Pi(k, k')$. See the Methods section for precise definitions. In Fig. 5A we plot the wiring complexities, as black bars, for our experimental datasets (Original), together with those of their randomisations (null models) in grey (Reshuffled). In panel (a) the network complexity is computed 'per node' as given in equation 2 (Methods) and it takes into account the average degree of the network, while the complexity 'per link' in panel (b) is independent of the average degree. For our dataset it appears more appropriate to use the 'per link' complexity, because the relative differences between Y2H and AP–MS*S. cerevisiae* networks are smaller. The AP–MS networks tend to have higher wiring complexities than the Y2H ones, although less so in the 'per link' plot, except for *C. jeuni* and *T. pallidum*. The latter, however, are special in that around 80% of their predicted encoded proteins have at least one assigned interaction (this percentage can be seen as an approximation of the 'coverage' of the network), which is the highest value among all datasets studied; this may explain differences between these two networks and other Y2H datasets. Similarly to what was observed for assortativities, *S. cerevisiae* X is again seen to stand out with an extremely high wiring complexity, consistent with the strong degree correlations observed earlier in Fig. 3.

Information-theoretic clustering

A second information-theoretic tool derived in Annibale et al. [22] is a transparent formula for an information-theoretic 'distance' between any two networks, once more expressed explicitly in terms of the functions $p(k)$ and $\Pi(k, k')$ of the networks concerned. This network distance is the symmetrised Kullback-Leibler divergence of the maximum entropy graph distributions with degree distributions and degree correlations identical to those of the two networks. We can use this mutual distance measure to cluster our PPIN datasets, and construct dendrograms analogous to phylogenetic trees. In Fig. 6A we show the resulting information-theoretic dendrogram for the full collection of all our PPIN data sets. The pariwise distance matrices of the AP-MS and Y2H data sets are provided in Tables S1 and S2, respectively. In Fig. 6B we limit our analysis to single-technique *S. cerevisiae* data sets only (excluding *S. cerevisiae* V and *S. cerevisiae* VII, which are the result of integrating datasets detected by a variety of different techniques). The results of these analyses are quite revealing. Those data sets which were most strongly criticised in the past for having worryingly small overlaps [29], for example the Y2H data sets *S. cerevisiae* I *versus* II and *H. sapiens* I *versus* II, are now unambiguously found to be topologically similar after all.

We also observe that the full collection of PPINs group primarily by detection method, so at least for the presently available PPIN datasets, any biological similarities (whether or not based on evolutionary relationship) are overshadowed by methodological biases. This is particularly evident in the central subgroup (central pink leaves) in Fig. 6A, which clusters almost exclusively Y2H datasets and comprises a wide range of species. The methodological biases are also obvious in the intra-species comparison of *S. cerevisiae* depicted in Fig. 6B. The largest sub–group distance within this *S. cerevisiae* tree is the one between two AP–MS datasets that have been post–processed differently (the top two within the green box). Also, the single PCA network is separated from the AP–MS and Y2H subgroups. We can now summarise the two, in our view, most important observations:

- PPINs of the same species and measured *via* the same experimental method are statistically similar, and more similar than networks measured *via* the same method but for different species. Apparently, the former exhibit similar *macroscopic* topological features, despite the small microscopic overlap of the individual PPINs. The information-theoretic network distance is therefore a useful macroscopic descriptor of similarity.
- PPINs measured *via* the same experimental method cluster together, revealing a bias introduced by the methods that is seen to overrule species-specific information. Although methodological biases have been acknowledged in the literature, we are now in a position to quantify their impact: the bias is proportional to the excess distances between the *S. cerevisiae*

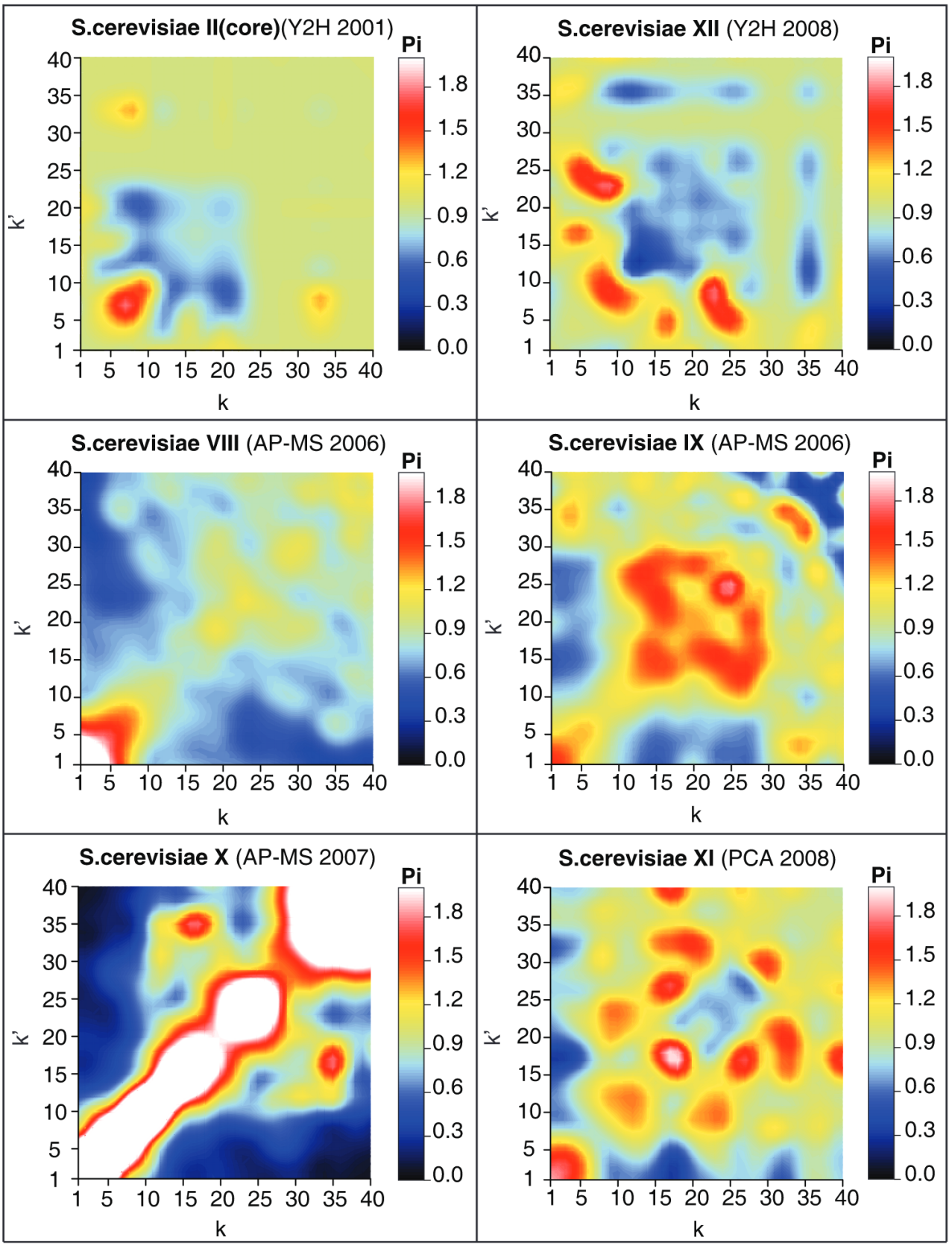

S.cerevisiae II(core)(Y2H 2001)
S.cerevisiae XII (Y2H 2008)
S.cerevisiae VIII (AP-MS 2006)
S.cerevisiae IX (AP-MS 2006)
S.cerevisiae X (AP-MS 2007)
S.cerevisiae XI (PCA 2008)
Pi
1.8
1.5
1.2
0.9
0.6
0.3
0.0
k'
k
1 5 10 15 20 25 30 35 40

Figure 3. Heat map representations of the normalised DDC function $\Pi(k, k')$ of yeast PPINs. Definitions and conventions are identical to those of Fig. 2.
doi:10.1371/journal.pone.0012083.g003

networks measured by AP-MS compared to those measured by Y2H (Fig. 6B).

Therefore, a species tree based on data from two different experimental methods yields an inconsistent picture (see Fig. 6A), in which the wanted contributions of the 'species distance' are modulated with the unwanted contributions of the 'methodological distance' originating from sampling biases.

A clearer picture is obtained when trees based on data from a single experimental method are constructed. Since the Y2H data appear less biased than the AP-MS data, a multi-species tree of Y2H networks is shown in Fig. 7A juxtaposed to a reference tree (Fig. 7B). Comparison of the trees reveals that the network distance measure correctly assigns short distances between *H. sapiens*, *C. elegans* and *S. cerevisiae*, but misassigns short distances between these species and *M. loti* or *Synechocystis*. Therefore one may deduce that the 'biological signal' captured by the network complexity difference is indeed strong enough for some of the networks in our data set to place them correctly on the tree, while others would require more data completeness to reach a comparable signal. For example, the two correctly placed bacteria *C. jejuni* and *T. pallidum* show a relatively high complexity (Fig. 5B).

Finally, in order to separate the contributions of the degree distribution and the DDCs to the distance information in generating the dendrograms shown in Fig. 6, we have also performed the same computations on the basis of a simplified information-theoretic distance measure for PPINs, which would have been the result of characterising all PPINs by their degree distributions alone. The result is shown in Fig. S2.

Network size effects

The mathematical framework operates on statistical grounds and the precision of the results depends on the sample size. We have already pointed out finite size effects in the assortativity of some reshuffled networks. To assess the robustness of the macroscopic network properties *versus* the variation of the sampled data and the sample size, we have performed a sub-sampling experiment. Each network was modified by randomly removing a

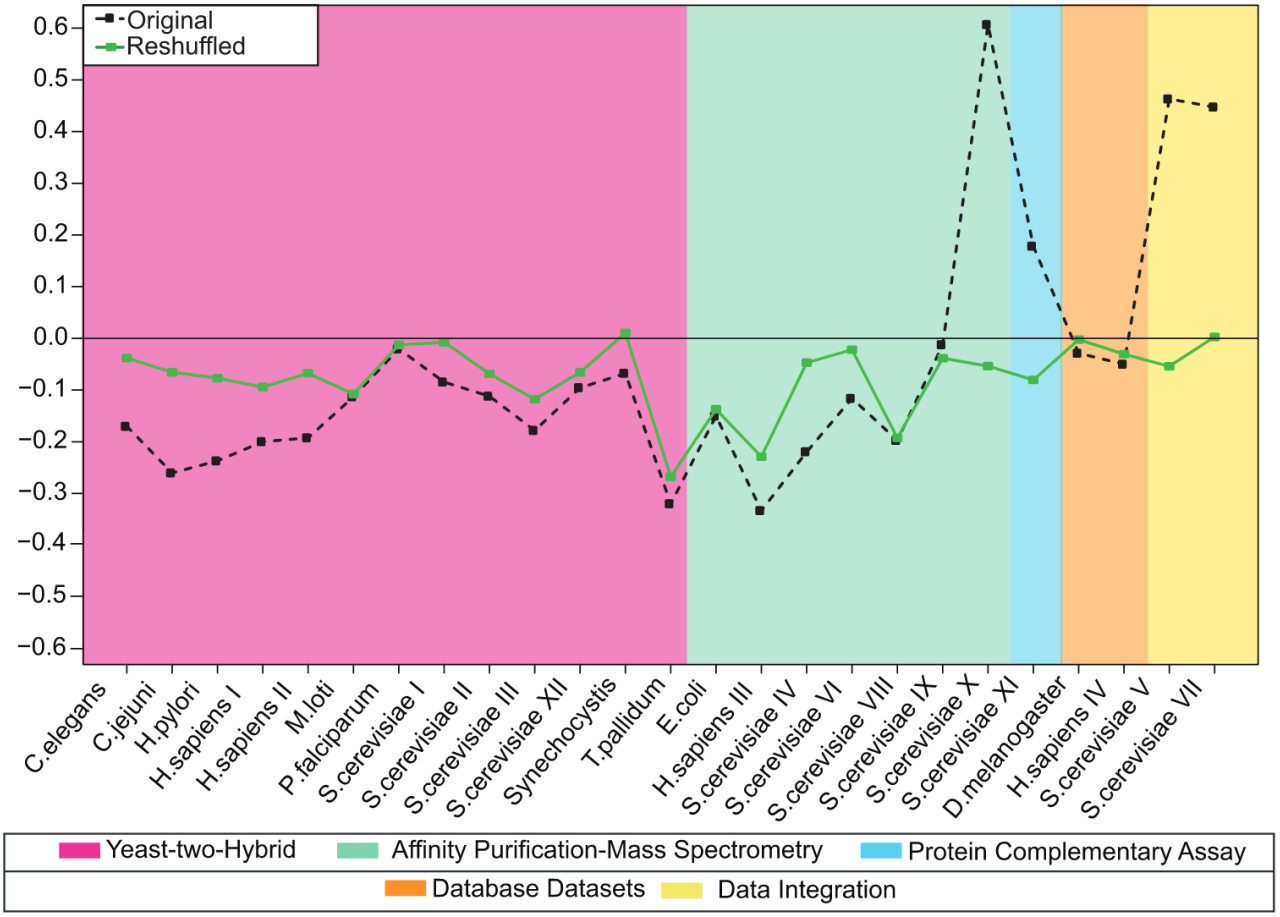
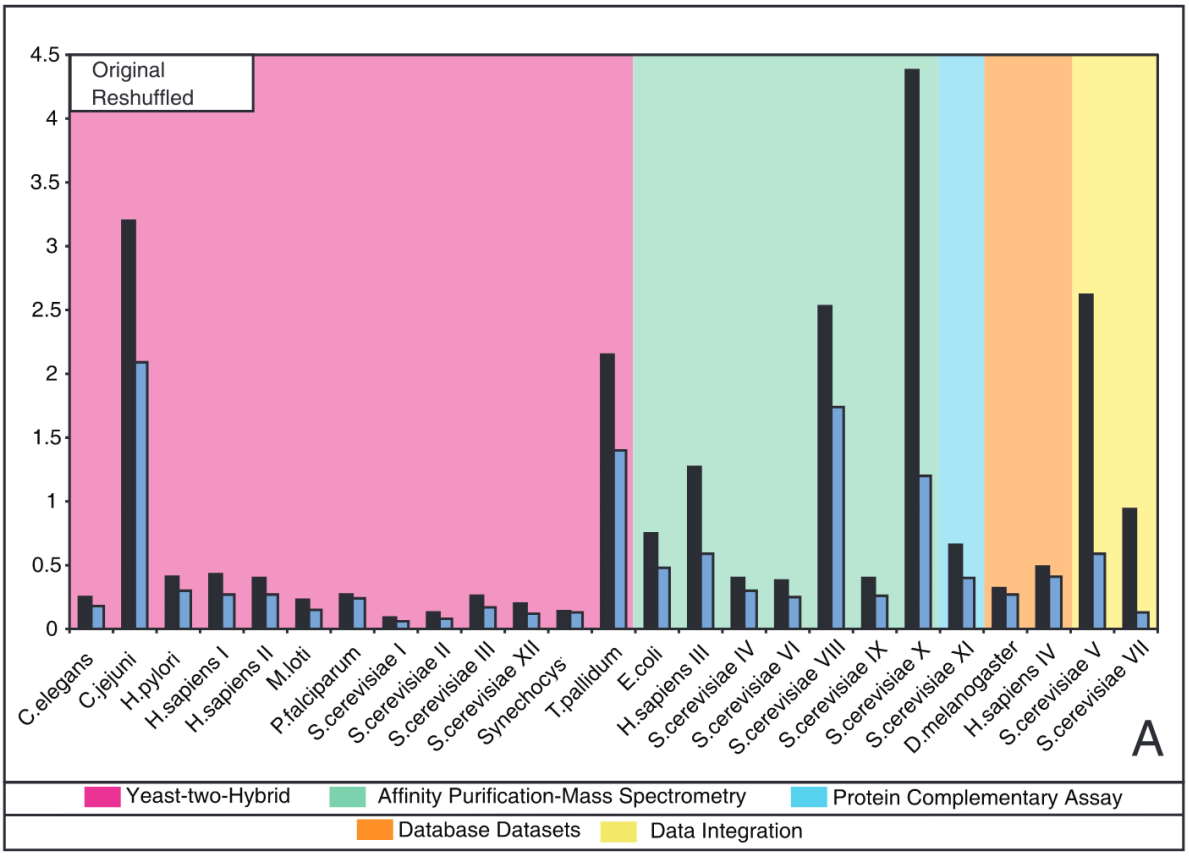


Figure 4. Assortativities as calculated for both the biological PPINs (black line) and their randomised versions (or 'null models', green line). The randomised networks have degree distributions identical to their biological counterparts, but are otherwise fully random (the randomisation seeks to remove any DDCs initially present). A positive assortativity implies that nodes prefer to interact with other nodes of similar degree, whereas a negative assortativity implies that high degree nodes prefer to interact with low degree nodes.
doi:10.1371/journal.pone.0012083.g004


Original
Reshuffled
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
C.elegans
C.jejuni
H.pylori
H.sapiens I
H.sapiens II
M.loti
P.falciparum
S.cerevisiae I
S.cerevisiae II
S.cerevisiae III
S.cerevisiae XII
Synechocys
T.pallidum
E.coli
H.sapiens III
S.cerevisiae IV
S.cerevisiae VI
S.cerevisiae VIII
S.cerevisiae IX
S.cerevisiae X
S.cerevisiae XI
D.melanogaster
H.sapiens IV
S.cerevisiae V
S.cerevisiae VII
A
Yeast-two-Hybrid
Affinity Purification-Mass Spectrometry
Protein Complementary Assay
Database Datasets
Data Integration

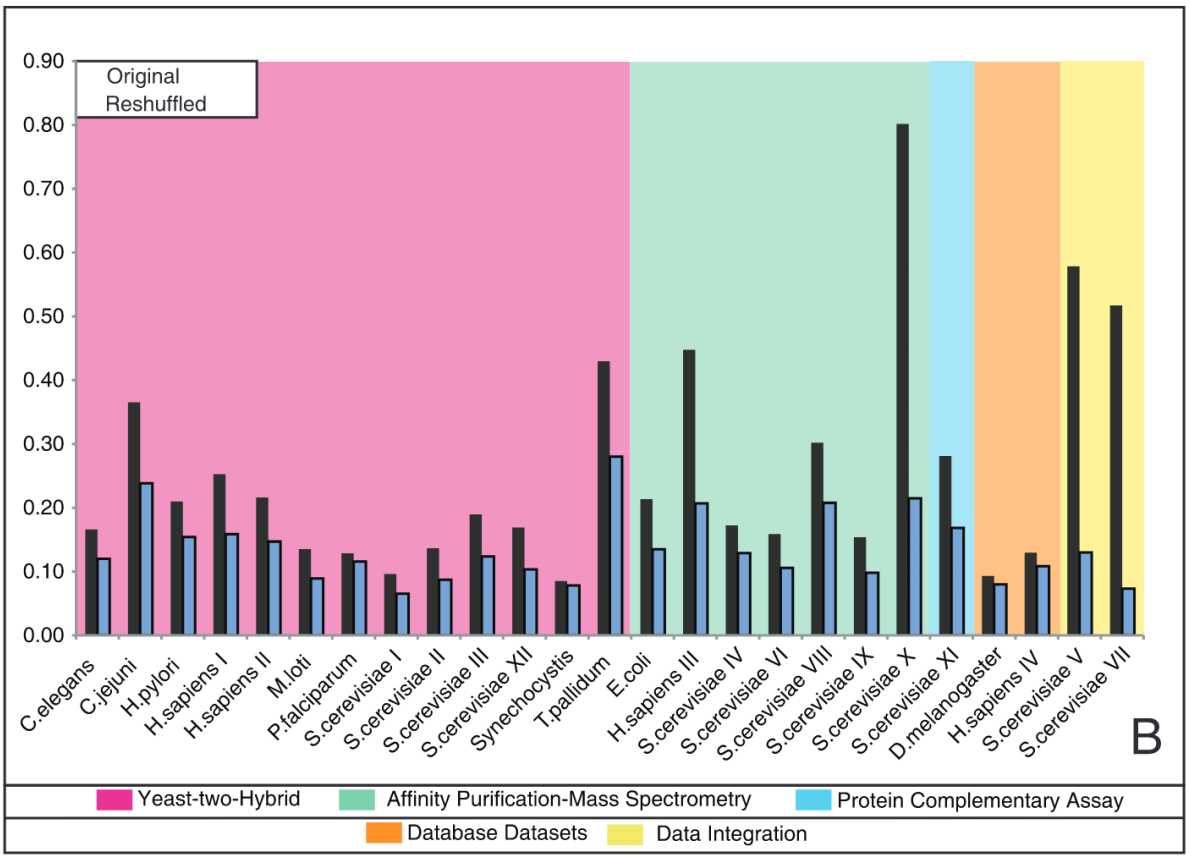

Original
Reshuffled
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
C.elegans
C.jejuni
H.pylori
H.sapiens I
H.sapiens II
M.loti
P.falciparum
S.cerevisiae I
S.cerevisiae II
S.cerevisiae III
S.cerevisiae XII
Synechocystis
T.pallidum
E.coli
H.sapiens III
S.cerevisiae IV
S.cerevisiae VI
S.cerevisiae VIII
S.cerevisiae IX
S.cerevisiae X
S.cerevisiae XI
D.melanogaster
H.sapiens IV
S.cerevisiae V
S.cerevisiae VII
B
Yeast-two-Hybrid
Affinity Purification-Mass Spectrometry
Protein Complementary Assay
Database Datasets
Data Integration

Figure 5. Wiring complexities as calculated for both the biological PPINs (black bars) and their randomised versions (or 'null models', grey bars). The randomised networks have degree distributions identical to their biological counterparts, but are otherwise fully random. The wiring complexity measures the topological information contained in a network's normalised DDC function $\Pi(k, k')$, beyond that in its degree statistics $p(k)$. (A) The network complexity is computed per network node. The average degree of each network contributes to its complexity value. (B) The network complexity is computed per link, which removes the dependency on the average degree and reduces complexity value (note the different ordinate scales (A) and (B)).
doi:10.1371/journal.pone.0012083.g005

certain fraction of its nodes (from 10 to 90% in 10% increments) and the distance between the modified (sub-sampled) and the original network was plotted as a function of the degree of node removal (Fig. 8). The plot shows an exemplary collection of networks. The sensitivity of the networks towards the sample size is correlated with their complexity. Sub-sampled networks with high complexity yield larger distances to their original versions than those with low complexity. This is not surprising, as the distance is a measure of the complexity difference. However, the particular curve shapes in Fig. 8 are related to the distribution of node links. For example, the extreme behaviour of *S. cerevisiae* X is owing to its large number of hub-hub links (see Fig. 3), while the *S. cerevisiae* XI network with a flatter DDC distribution is relatively robust to node removal.

The effects of sub-sampling, random and non-random, on the network statistics have been explored by others [8,30], but without consideration of DDCs. The curves in Fig. 8 provide an error estimate for the sub–sampling effects on the information-theoretic properties discussed in this paper. This regards the DDC plots in Figs. 2,3, the network complexities in Fig. 5 and the trees in Figs. 6,7.

Discussion

In this paper we investigated the potential of recently introduced mathematical framework for quantifying and comparing the topologies of PPINs by systematic application to publicly available PPIN datasets. This framework provides exact and explicit measures of network complexity and information-theoretic distances between any two networks. In addition, in benchmarking empirical measurements on PPINs we used null models generated *via* recently developed rigorously unbiased algorithms.

Our methods involve a macroscopic characterisation of PPINs by their degree statistics and DDCs. Degree correlation properties

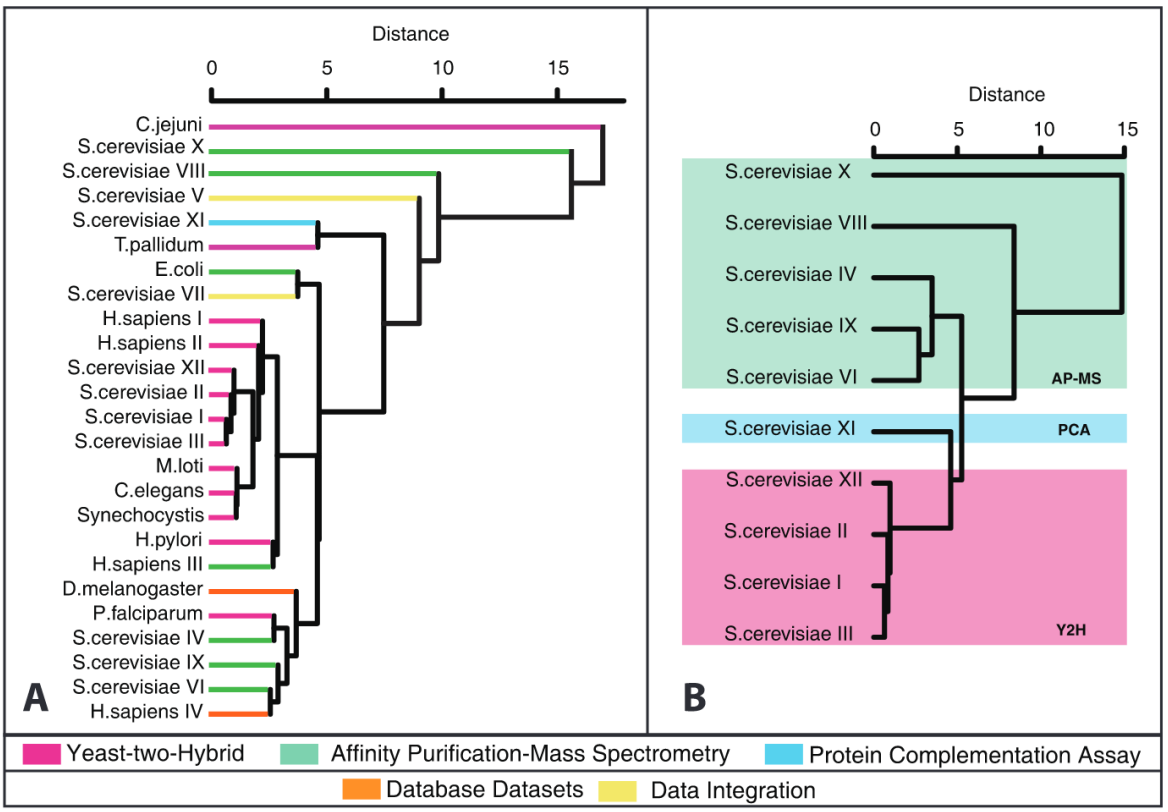


Figure 6. Network comparison by clustering using the full information-theoretic distance measure of equation 9 (see Methods) [22]. This distance is expressed explicitly in terms of the degree statistics and normalised DDC functions of the networks concerned. Panel A: Dendrogram calculated for an extensive collection of PPINs, covering a wide range of species. Panel B: Dendrogram for PPINs of the same species, *viz. S. cerevisiae*. Both trees were constructed using our proposed distance metric in a hierarchical clustering routine (see Methods). In both panels we decorated the clusters using the same colour scheme used throughout this paper to indicate the different experimental detection techniques (see bottom legend of the figure). The data integration datasets (*S. cerevisiae* V and *S. cerevisiae* VII) were excluded from panel b, since they are composed of interactions detected by a variety of different techniques.
doi:10.1371/journal.pone.0012083.g006

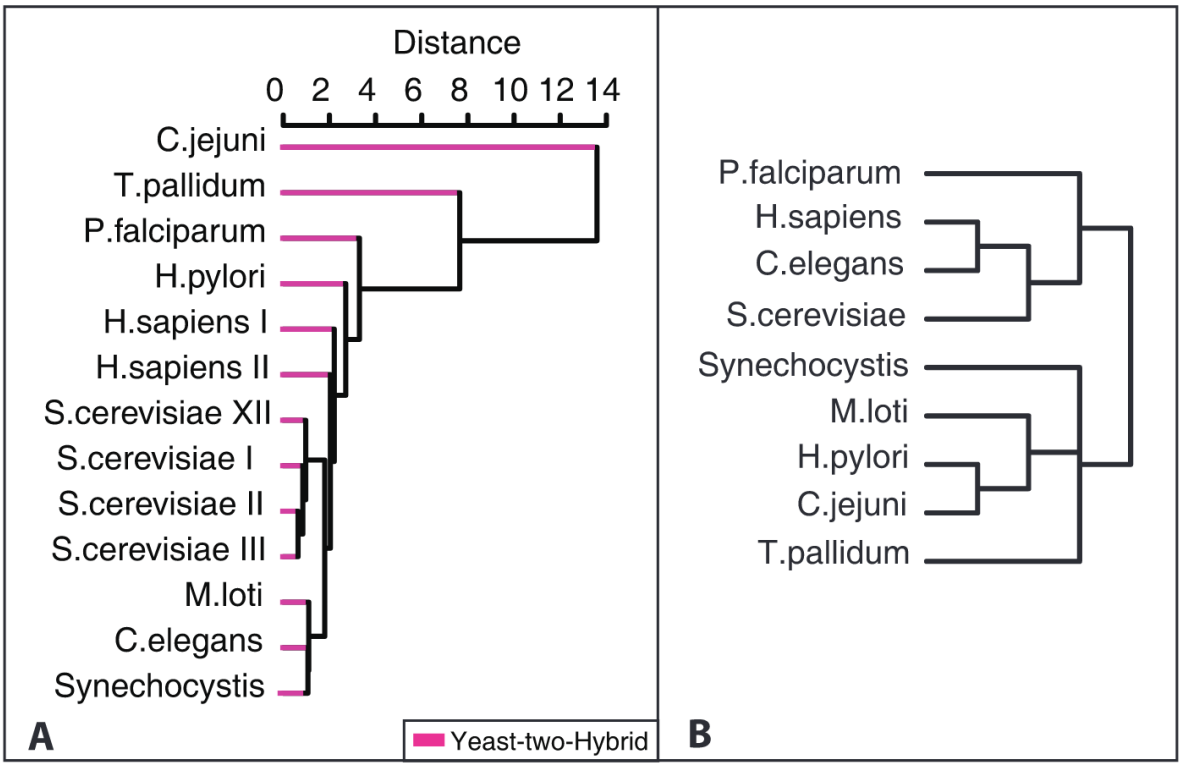


Figure 7. Network comparison by clustering. A: Y2H network cluster using the information-theoretic distance measure. B: Reference species tree provided by the NCBI taxonomy common tree service.
doi:10.1371/journal.pone.0012083.g007

have been used in the past to highlight topological features of networks and to suggest general principles governing functional mechanisms in the interactome [31–35]. These particular studies focused on regularities in interactions between high *versus* low degree nodes; unfortunately they did not agree on the nature and interpretation of such regularities, in particular on the role of high–degree (hub) proteins. Maslov and Sneppen (2002) [31] argued that suppression of hub–hub interactions is a 'universal feature' of robust molecular networks, that reflects compartmentalisation and modularity, characteristics of cellular processes. In contrast, Batada et al. (2006) [32] and Ivanic et al. (2008) [35] did not observe hub–hub interaction suppression, but suggested instead that hub–hub interactions play an important role in the underlying biological processes. An intermediate position was taken by Friedel and Zimmer (2007) [33], who generated artificial versions of biological networks and argued that neither type of degree–weighted behaviour is favoured.

We find that degree-degree correlations provide important and consistent information on PPIN topologies, but it is crucial that they are normalised correctly and that one uses robust and systematic methods for extracting this information. Normalisation of DDCs is usually based on comparison against appropriate randomised networks (null models). The unbiased generation of such null models, however, is nontrivial. Popular randomisation protocols such as 'edge-swapping' are now known to carry the risk of biased sampling, see [36]. The reason why we avoided the inconsistencies of previous studies [31–35] appears to be that, rather than normalising DDCs *via* numerical randomisation, we use an exact mathematical formula for the DDCs of large unbiased random graphs. Our normalised DDCs are by definition unbiased, and not subject to numerical normalisation noise. Where we employ null models for reasons other than normalisation, we use exact algorithms for generating unbiased null models that have only recently become available. Under these improved conditions one does detect reproducible DDC patterns, with an overall preference for high–low degree interactions. However, the variation of DDC patterns, even within the same species and detection method, precludes general conclusions about their origin in the underlying biological mechanisms. This type of inference would require improved (in terms of completeness and error rate) interaction data for several related networks.

The first information-theoretic tool we applied to the PPIN datasets was the formula for a network's complexity recently derived in Annibale et al. [22]. It has two contributions: a term representing the complexity embedded in the degree statistics (the degree complexity), and a second term representing the complexity embedded in the DDCs (the wiring complexity). The wiring complexity quantifies the extent to which DDCs are prominent in a network, similar to the assortativity measuring the nature of the lowest order correlations (if present). The two quantities provide complementary information. One can easily imagine higher order DDCs in PPINs (*e.g.* nonlinear relationships between the degrees of preferred protein partners) that could not be picked up by the assortativity but would still be detected by the wiring complexity. In fact this is already visible in the presently analysed PPIN data. Comparison of Fig. 4 to Fig. 5A shows that, while those datasets with nontrivial assortativities also have high wiring complexities, there are several further PPINs with a high wiring complexity but only a relatively modest assortativity.

The second information-theoretic tool we applied was a formula for an information-theoretic distance between networks. Like the complexity, the formula is expressed explicitly in terms of the

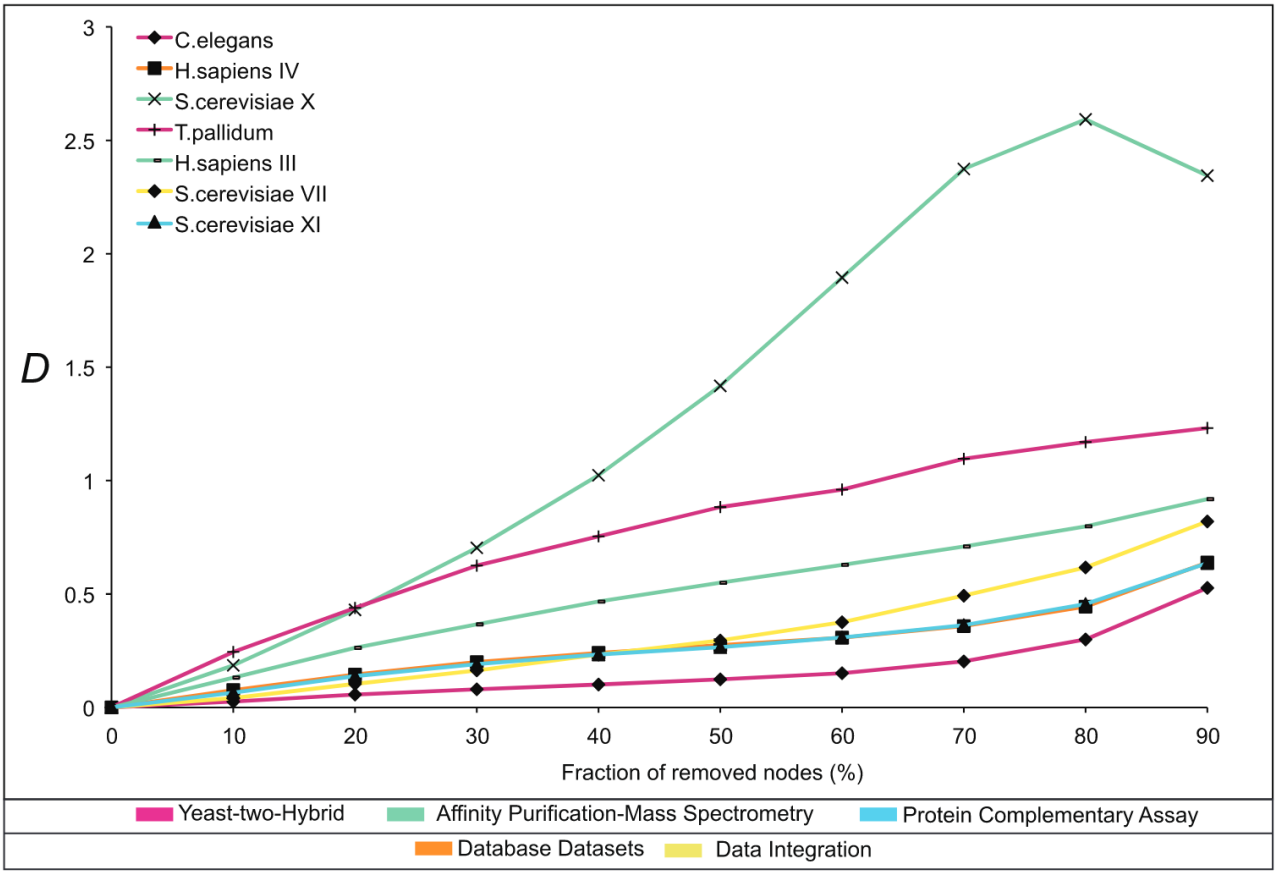


Figure 8. The effect of random sub-sampling on the network topology. Networks were modified by random removal of nodes to various degrees, given as fraction of nodes in the original network in the range from 10 to 90% in 10% increments. The information-theoretic distance between the original network and the sub-sampled network is plotted over the degree of node removal. A selection of typical curves is shown. Networks with high complexity show a large distance to the original network.
doi:10.1371/journal.pone.0012083.g008

networks' degree statistics and DDCs, and, based on macroscopic statistical features, it avoids the problems with the more primitive overlap-based network dissimilarity measures. Application of this second tool to our datasets resulted in a pairwise distance table, which we used to cluster the PPINs. The results, summarised in a dendrogram, are very revealing. Those data sets which were most strongly criticised in the past for having small overlaps, for example the Y2H data sets of *S. cerevisiae*, are now unambiguously found to be topologically similar. Furthermore, our method shows clearly that the PPINs group primarily by detection method, so biological similarities based on evolutionary relationship are presently overshadowed by methodological biases. These biases have been the centre of an active debate in recent years; the problems which they generate and methods to overcome these have been described recently, *e.g.* [9,17,37,38].

In particular, an often overlooked aspect of data derived from AP–MS experiments is the influence of the post-processing protocol on the final binary interaction map. This crucial aspect is now starting to be addressed by different groups [39–41], and we expect more accurate data to emerge in the near future. Our information-theoretic tools are thus very timely: they provide the required resolution and precision in the assessment and comparison of new PPIN data, and in evaluating the progress of the experimental methods. Being able to quantify biases accurately is a prerequisite for their systematic removal.

One should keep in mind that biological systems are not necessarily perfect, and that the presence in PPINs of non-selected, non-functional PPIs is to be expected [42]; the interpretation of interactome data will therefore always have to take account of noise. This again suggests that information-theoretic methods, with their rigorous probabilistic basis, should be seen as the appropriate tools in PPIN analysis. One could also envisage these methods being used to guide experimental efforts aimed at remedying the present under-sampling of PPINs, by predicting on statistical grounds the properties of missing network nodes and interactions.

In conclusion, we believe to have succeeded in

- supplying the biological and bio-informatics communities with a new generation of precise and user-friendly computational tools with which to quantify PPIN topologies and test new protocols for the removal of experimental biases from PPIN datasets, and
- demonstrating by a systematic application of these tools to publicly available datasets that the present protein network data are strongly biased by their experimental methods, while still exhibiting species–specific similarity and reproducibility.

We hope and anticipate that in the near future the accuracy and sensitivity of experiments will improve substantially, alongside a

further sharpening of the mathematical and computational tools for their analysis, allowing for meaningful comparisons of interactomes.

Materials and Methods

The following section gives a complete reference of the formulae used in this study. The central equations 2, 7 and 9 have been published in a recent work by the authors [22] in the context of parametrised random graph ensembles. They are repeated here in commented form to aid the reader.

Mathematical definitions

Degree distribution. Given a protein-protein interaction network with N nodes, we label its proteins by Roman indices $i,j=1\ldots N$, and represent the microscopic interaction information as a symmetric matrix $\boldsymbol{c}$ with entries c_{ij}, where $c_{ij}=1$ if i interacts with j, and $c_{ij}=0$ otherwise (with $c_{ii}=0$ for all i). The degree of a node i is then defined as $k_i(\boldsymbol{c})=\sum_j c_{ij}$, and the degree distribution of the PPIN is defined as

$$p(k)=\frac{1}{N}\sum_i \delta_{k,k_i(\boldsymbol{c})} \tag{1}$$

(here $\delta_{nm}=1$ if $n=m$ and $\delta_{nm}=0$ if $n\neq m$).

Degree–degree correlation (DDC). The average degree in the PPIN is given by $\bar{k}=N^{-1}\sum_i k_i(\boldsymbol{c})=\sum_k p(k)k$. The normalised DDC function $\Pi(k,k')$ of the network is defined as the ratio between the probability that two randomly picked nodes in c with degrees (k,k') are found to be connected, divided by what this probability would have been in large *random* networks with the same degree distribution as $\boldsymbol{c}$. The probabilities for large random networks can be calculated analytically, see *e.g.* [43]. This results in the following definition:

$$\Pi(k,k')=\frac{\sum_{ij} c_{ij}\delta_{k,k_i(\boldsymbol{c})}\delta_{k',k_j(\boldsymbol{c})}}{[p(k)-N^{-1}\delta_{kk'}]p(k')}\frac{\bar{k}}{Nkk'}(1-\frac{1}{N}) \tag{2}$$

This quantity was plotted for our PPIN datasets in heat-map form in Figs. 2 and 3. Any statistically significant deviation from $\Pi(k,k')=1$ signals the presence of non-trivial DDCs. Both $p(k)$ and $\Pi(k,k')$ are macroscopic quantities that can be measured directly and at low computation cost.

Assortativity. The assortativity a (as plotted in Fig. 4 for our PPIN datasets) is defined [26] as the magnitude of the normalised correlations for the joint probability $W(k,k')$ of finding a randomly drawn interaction in the graph $\boldsymbol{c}$ connecting nodes with degrees k and k' respectively, *viz.*

$$W(k,k')=\frac{1}{N\bar{k}}\sum_{ij} c_{ij}\delta_{k,k_i(\boldsymbol{c})}\delta_{k',k_j(\boldsymbol{c})} \tag{3}$$

Upon defining averages over this measure as $\langle f(k,k')\rangle=\sum_{kk'}W(k,k')f(k,k')$, and using the symmetry of $W(k,k')$ as well as the relation $\sum_{k'}W(k,k')=p(k)k/\bar{k}$, one has

$$a=\frac{\langle kk'\rangle-\langle k\rangle^2}{\langle k^2\rangle-\langle k\rangle^2}=\frac{\frac{1}{N\bar{k}}\sum_{ij}c_{ij}k_i(\boldsymbol{c})k_j(\boldsymbol{c})-(\frac{1}{N\bar{k}}\sum_i k_i^2(\boldsymbol{c}))^2}{\frac{1}{N\bar{k}}\sum_i k_i^3(\boldsymbol{c})-(\frac{1}{N\bar{k}}\sum_i k_i^2(\boldsymbol{c}))^2} \tag{4}$$

The relation between $W(k,k')$ and $\Pi(k,k')$ is

$$W(k,k')=\Pi(k,k')p(k)p(k')kk'/\bar{k}^2 \tag{5}$$

which is why the assortativity can be written as a function of $p(k)$ and $\Pi(k,k')$:

$$a=\frac{\sum_{kk'}p(k)p(k')[\Pi(k,k')-1](kk')^2}{\bar{k}\sum_k p(k)k^3-(\sum_k p(k)k^2)^2} \tag{6}$$

Hamming distance. The Hamming distance is defined as $\Delta=(N\bar{k})^{-1}\sum_{i<j}|c_{ij}-c'_{ij}|$, where the binary interaction variables c_{ij} define the original PPIN and the variables c'_{ij} represent its randomisation.

Information-theoretic tools

Degree complexity and wiring complexity. Using methods from random graph theory and statistical mechanics the following explicit formula was derived for the information-theoretic complexity $C[p,\Pi]$ per node of non-directed networks (such as PPINs) with degree distribution $p(k)$ and normalised degree-degree correlations $\Pi(k,k')$:

$$\begin{aligned}C[p,\Pi]=&\sum_k p(k)\log[p(k)/\pi(k)]\\&+\frac{1}{2\bar{k}}\sum_{kk'}p(k)p(k')kk'\Pi(k,k')\log\Pi(k,k')\end{aligned} \tag{7}$$

where $\pi(k)$ is the Poissonian distribution with average degree $\bar{k}$:

$$\pi(k)=e^{-\bar{k}}\bar{k}^k/k! \tag{8}$$

The first term in (7) is called the degree complexity; the second term, which would be zero in our null models, is called the wiring complexity. This latter quantity was plotted in Fig. 5A.

Network distance. Similar calculations led also to an explicit formula for an information-theoretic distance D_{AB} between any two non–directed networks A and B (such as PPINs), characterised by the structure functions $\{p_A,\Pi_A\}$ and $\{p_B,\Pi_B\}$, respectively:

$$\begin{aligned}D_{AB}=&\frac{1}{2}\sum_k p_A(k)\log\left[\frac{p_A(k)}{p_B(k)}\right]+\frac{1}{2}\sum_k p_B(k)\log\left[\frac{p_B(k)}{p_A(k)}\right]\\&+\sum_{kk'}\frac{p_A(k)p_A(k')kk'}{4\bar{k}_A}\Pi_A(k,k')\log\left[\frac{\Pi_A(k,k')}{\Pi_B(k,k')}\right]\\&+\sum_{kk'}\frac{p_B(k)p_B(k')kk'}{4\bar{k}_B}\Pi_B(k,k')\log\left[\frac{\Pi_B(k,k')}{\Pi_A(k,k')}\right]\end{aligned} \tag{9}$$

with $\bar{k}_A=\sum_k p_A(k)k$ and $\bar{k}_B=\sum_k p_B(k)k$. The distance (9) was used to calculate the dendrogram of Fig. 6. A simplified distance that no longer takes DDC information into account is obtained upon removing the last two lines from (9), leaving an expression that involves only the two degree distributions $p_A(k)$ and $p_B(k)$; this simplified distance definition was used to calculate the dendrogram of Fig. 7.

Definition and generation of null models

Given an observed N-node PPIN with degree distribution $p(k)$, we define its associated null model as a graph drawn randomly and

with uniform probabilities from the set of *all* N-node graphs with degree distribution $p(k)$, so the probability of any graph $\boldsymbol{c}$ being generated as null model for the PPIN under study must be

$$p(c)=\sum_{k_1\ldots k_N}\left[\prod_i p(k_i)\right]\frac{\prod_i \delta_{k_i,k_i(\boldsymbol{c})}}{\sum_{\boldsymbol{c}'}\prod_i \delta_{k_i,k_i(\boldsymbol{c}')}} \tag{10}$$

The issue of sampling uniformly the desired space of graphs is non–trivial. Naive application to the original PPIN of the popular method of 'edge-swapping' (or 'graph shuffling') would indeed upon equilibration produce randomised graphs, but these might not be sampled uniformly; biased sampling would invalidate any inference based on comparing observations in real PPINs to those in randomised graphs. In this paper we used the general and exact Markov chain Monte Carlo (MCMC) method for generating random graphs proposed in Coolen et al. [36], which is based on edge–swaps [44] but involves nontrivial move acceptance probabilities. Most graph randomisation protocols, including the one used in this paper, are defined *via* a degree–preserving MCMC dynamics in the space of graphs, which is defined such that it produces a relaxation towards an equilibrium state where all acceptable graphs are generated with prescribed probabilities. This dynamics must be run for a sufficient duration of time to guarantee that all transients in the MCMC have died down and the desired equilibrium state has indeed been reached. In this paper we have used equilibration times such that the number of accepted transitions in the MCMC exceeded 100 per link, which (upon systematic monitoring of a number of key observables in the graphs) was found to be adequate to ensure equilibration of the Markov chain.

Numerical practicalities

After measuring a graph's degree distribution $p(k)$ and normalised DDC functions $\Pi(k, k')$ we applied to both functions a weak Gaussian smoothening, resulting in the new functions

$$\tilde{p}(k)=\frac{\sum_{m=0}^{k_{\max}} e^{-\frac{1}{2\sigma^2}(k-m)^2}p(m)}{\sum_{m=0}^{k_{\max}} e^{-\frac{1}{2\sigma^2}(k-m)^2}} \tag{11}$$

$$\tilde{\Pi}(k,k')=\frac{\sum_{m,m'=0}^{k_{\max}} e^{-\frac{1}{2\sigma^2}[(k-m)^2+(k'-m')^2]}\Pi(m,m')}{\sum_{m,m'=0}^{k_{\max}} e^{-\frac{1}{2\sigma^2}[(k-m)^2+(k'-m')^2]}} \tag{12}$$

with diffusion width $\sigma=1.5$. The reason for doing this is that it prevents probabilities from being strictly zero, which (while reflecting only finite size effects) would cause problems in the distance measure (9). A further benefit of this smoothening is that it removes some of the finite size noise from the images in Figs. 2 and 3.

Dendrograms, as shown in Fig. 5A and Fig. 6, were computed using information-theoretic network distances (9) in the hierarchical clustering routine 'hclust' of the R environment [45] with the 'average' agglomeration method.

Finally, in plotting the normalised degree-degree correlations of PPINs in Figs. 2 and 3, we chose to limit ourselves to $k\leq 40$. The reason is that while proteins with larger degrees certainly exist in the networks studied, the limited number of these no longer justify the interpretation of quantities such as $\Pi(k, k')$ as clean estimators of (normalised) probabilities; this would require more data points in the large (k, k') regions.

Supporting Information

Figure S1 This figure shows for each PPIN the normalised Hamming distance between the original network and its null model. The null models were obtained for each PPIN by application of the exact Markov Chain Monte Carlo randomisation protocol of Coolen et al. 2009. The Hamming distance Δ is defined in such a way that it equals zero if the two networks are strictly identical, and equals one if the two networks are statistically independent (apart from the values of their degrees, which are preserved by the randomisation). The effect of insufficient equilibration of the randomisation protocol would be marked by Hamming distances significantly less than one. This figure supports our confidence that the equilibration time which we used in the randomisation algorithm, being 100 accepted moves per protein interaction, were adequate.
Found at: doi:10.1371/journal.pone.0012083.s001 (0.45 MB EPS)

Figure S2 Network comparison by clustering using a simplified information-theoretic distance without the contribution of DDCs. Definitions and conventions are identical to those in Figure 6.
Found at: doi:10.1371/journal.pone.0012083.s002 (0.46 MB EPS)

Table S1 Triangular matrix of information theoretic network distance between the AP-MS datasets.
Found at: doi:10.1371/journal.pone.0012083.s003 (0.00 MB TXT)

Table S2 Triangular matrix of information theoretic network distance between the Y2H datasets.
Found at: doi:10.1371/journal.pone.0012083.s004 (0.00 MB TXT)

Author Contributions

Conceived and designed the experiments: AA ACCC FF. Performed the experiments: LPF AA JK ACCC. Analyzed the data: LPF AA JK ACCC FF. Contributed reagents/materials/analysis tools: FF. Wrote the paper: LPF AA JK ACCC FF.

References

1. Cesareni G, Ceol A, Gavrila C, Palazzi LM, Persico M, et al. (2005) Comparative interactomics. FEBS Lett 579: 1828–1833.
2. Vidal M (2009) A unifying view of 21st century systems biology. FEBS Lett 583: 3891–3894.
3. Kar G, Gursoy A, Keskin O (2009) Human cancer protein-protein interaction network: a structural perspective. PLoS Comput Biol 5: e1000601.
4. Schadt EE, Zhang B, Zhu J (2009) Advances in systems biology are enhancing our understanding of disease and moving us closer to novel disease treatments. Genetica 136: 259–269.
5. Schadt EE (2009) Molecular networks as sensors and drivers of common human diseases. Nature 461: 218–223.
6. Kiemer L, Cesareni G (2007) Comparative interactomics: comparing apples and pears? Trends Biotechnol 25: 448–454.
7. Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interaction networks? Genome Biol 7: 120.1–120.9.
8. Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, et al. (2008) Estimating the size of the human interactome. Proc Natl Acad Sci U S A 105: 6959–6964.
9. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, et al. (2009) An empirical framework for binary interactome mapping. Nat Methods 6: 83–90.
10. Pržulj N (2007) Biological network comparison using graphlet degree distribution. Bioinformatics 23: e177–183.
11. Milenković T, Pržulj N (2008) Uncovering biological network function *via* graphlet degree signatures. Cancer Inform 6: 257–273.
12. Narayanan M, Karp RM (2007) Comparing protein interaction networks *via* a graph match-and-split algorithm. J Comput Biol 14: 892–907.

13. Kuchaiev O, Pržulj N (2009) Learning the structure of protein-protein interaction networks. Pac Symp Biocomput. pp 39–50.
14. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, et al. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A 100: 11394–11399.
15. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, et al. (2004) Pathblast: a tool for alignment of protein interaction networks. Nucleic Acids Res 32: W83–88.
16. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, et al. (2005) Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A 102: 1974–1979.
17. Wodak SJ, Pu S, Vlasblom J, Seraphin B (2009) Challenges and rewards of interaction proteomics. Mol Cell Proteomics 8: 3–18.
18. Stumpf MPH, Wiuf C, May RM (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. Proc Natl Acad Sci U S A 102: 4221–4224.
19. Yu X, Ivanic J, Wallqvist A, Reifman J (2009) A novel scoring approach for protein co-purification data reveals high interaction specificity. PLoS Comput Biol 5: e1000515.
20. Barabási AL (2009) Scale-free networks: a decade and beyond. Science 325: 412–413.
21. Perez-Vicente CJ, Coolen ACC (2008) Spin models on random graphs with controlled topologies beyond degree constraints. J Phys A: Math Theor 41: 255003.
22. Annibale A, Coolen ACC, Fernandes LP, Fraternali F, Kleinjung J (2009) Tailored graph ensembles as proxies or null models for real networks i: tools for quantifying structure. J Phys A: Math Theor 42: 485001.1–485001.25.
23. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509–512.
24. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74: 509–512.
25. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86: 3200–3203.
26. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89: 208701.
27. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 409–410.
28. Newman MEJ, Leicht EA (2007) Mixture models and exploratory analysis in networks. Proc Natl Acad Sci U S A 104: 9564–9569.
29. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, et al. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322: 104–110.
30. de Silva E, Thorne T, Ingram P, Agrafioti I, Swire J, et al. (2006) The effects of incomplete protein interaction data on structural and evolutionary inferences. BMC Biology 4: 39.
31. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296: 910–913.
32. Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, et al. (2006) Stratus not altocumulus: a new view of the yeast protein interaction network. PLoS Biol 4: e317.
33. Friedel CC, Zimmer R (2007) Influence of degree correlations on network structure and stability in protein-protein interaction networks. BMC Bioinformatics 8: 297.
34. Gallos LK, Song C, Makse HA (2008) Scaling of degree correlations and its influence on diffusion in scale-free networks. Phys Rev Lett 100: 248701.
35. Ivanic J, Wallqvist A, Reifman J (2008) Probing the extent of randomness in protein interaction networks. PLoS Comput Biol 4: e1000114.
36. Coolen ACC, De Martino A, Annibale A (2009) Constrained markovian dynamics of random graphs. J Stat Phys 136: 1035–1067.
37. Ivanic J, Yu X, Wallqvist A, Reifman J (2009) Influence of protein abundance on high-throughput protein-protein interaction detection. PLoS ONE 4: e5815.
38. Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, et al. (2009) An experimentally derived confidence score for binary protein-protein interactions. Nat Methods 6: 91–97.
39. Ramakrishnan SR, Vogel C, Kwon T, Penalva LO, Marcotte EM, et al. (2009) Mining gene functional networks to improve mass-spectrometry-based protein identification. Bioinformatics 25: 2955–2961.
40. Friedel CC, Krumsiek J, Zimmer R (2009) Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. J Comput Biol 16: 971–987.
41. Friedel CC, Zimmer R (2009) Identifying the topology of protein complexes from affinity purification assays. Bioinformatics 25: 2140–2146.
42. Levy ED, Landry CR, Michnick SW (2009) How perfect can protein interactomes be? Sci Signal 25: 193–197.
43. Dorogovstev SN, Mendes JF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford: Oxford University Press.
44. Seidel J (1976) A survey of two-graphs. Colloquio Internazionale sulle Teorie Combinatorie I: 481–511.
45. R Development Core Team (2007) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.
46. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403: 623–627.
47. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98: 4569–4574.
48. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, et al. (2001) The protein-protein interaction map of *Helicobacter pylori*. Nature 409: 211–215.
49. Lacount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, et al. (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum*. Nature 438: 103–107.
50. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957–968.
51. Parrish JR, Yu J, Liu G, Hines JA, Chan JE, et al. (2007) A proteome-wide protein interaction map for *Campylobacter jejuni*. Genome Biol 8: R130.
52. Sato S, Shimoda Y, Muraki A, Kohara M, Nakamura Y, et al. (2007) A large-scale protein protein interaction analysis in *Synechocystis sp. PCC6803*. DNA Res 14: 207–216.
53. Simonis N, Rual JF, Carvunis AR, Tasan M, Lemmens I, et al. (2008) Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. Nat Methods 6: 47–54.
54. Shimoda Y, Shinpo S, Kohara M, Nakamura Y, Tabata S, et al. (2008) A large scale analysis of protein-protein interactions in the nitrogen-fixing bacterium *Mesorhizobium loti*. DNA Res 15: 13–23.
55. Titz B, Rajagopala SV, Goll J, Häuser R, Mckevitt MT, et al. (2008) The binary protein interactome of *Treponema pallidum* – the syphilis spirochete. PLoS ONE 3: e2292.
56. Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141–147.
57. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415: 180–183.
58. Collins SRR, Kemmeren P, Chu X, Greenblatt JFF, Spencer F, et al. (2007) Towards a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. Mol Cell Proteomics 6: 439–450.
59. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440: 637–643.
60. Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, et al. (2006) Large-scale identification of protein-protein interaction of *Escherichia* coli K-12. Genome Res 16: 686–691.
61. Claude A, Aloy P, Grandi P, Krause R, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636.
62. Ewing RM, Chu P, Elisma F, Li H, Taylor P, et al. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol 3: 89.
63. Tarassov K, Messier V, Landry CR, Radinovic S, Molina MM, et al. (2008) An *in vivo* map of the yeast protein interactome. Science 320: 1465–1470.
64. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34: D535–539.
65. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human protein reference database–2009 update. Nucleic Acids Res 37: D767–772.
66. Dong J, Bertin N, Hao T, Goldberg DS, Berriz GF, et al. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430: 88–93.
67. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417: 399–403.