Estimating the size of the human interactome

Michael P. H. Stumpf[†‡§], Thomas Thorne[†], Eric de Silva[†], Ronald Stewart[†], Hyeong Jun An[¶], Michael Lappe[¶], and Carsten Wiuf[§‖]

[†]Division of Molecular Biosciences, Imperial College London, Wolfson Building, London SW7 2AZ, United Kingdom; [‡]Institute of Mathematical Sciences, Imperial College London, London SW7 2AZ, United Kingdom; [‖]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; and [¶]Bioinformatics Research Center, University of Aarhus, 8000 Aarhus C, Denmark

Edited by Burton H. Singer, Princeton University, Princeton, NJ, and approved February 19, 2008 (received for review August 27, 2007)

After the completion of the human and other genome projects it emerged that the number of genes in organisms as diverse as fruit flies, nematodes, and humans does not reflect our perception of their relative complexity. Here, we provide reliable evidence that the size of protein interaction networks in different organisms appears to correlate much better with their apparent biological complexity. We develop a stable and powerful, yet simple, statistical procedure to estimate the size of the whole network from subnet data. This approach is then applied to a range of eukaryotic organisms for which extensive protein interaction data have been collected and we estimate the number of interactions in humans to be ≈650,000. We find that the human interaction network is one order of magnitude bigger than the *Drosophila melanogaster* interactome and ≈3 times bigger than in *Caenorhabditis elegans*.

evolutionary systems biology | network inference | network sampling theory | network evolution

One of the perhaps most surprising results of the genome-sequencing projects was that the number of genes is much lower than had been expected and is, in fact, surprisingly similar for very different organisms (1, 2). For example, the nematode *Caenorhabditis elegans* appears to have a similar number of genes as humans, whereas rice and maize appear to have even more genes than humans. It was then quickly suggested that the biological complexity of organisms is not reflected merely by the number of genes but by the number of physiologically relevant interactions (1, 3). In addition to alternative splice variants (4), posttranslational processes (5), and other (e.g., genetic) factors influencing gene expression (6, 7), the structure of interactome is one of the crucial factors underlying the complexity of biological organisms. Here, we focus on the wealth of available protein interaction data and demonstrate that it is possible to arrive at a reliable statistical estimate for the size of these interaction networks. This approach is then used to assess the complexity of protein interaction networks in different organisms from present incomplete and noisy protein interaction datasets.

There are now fairly extensive protein interaction network (PIN) datasets in a number of species, including humans (8, 9). These have been generated by a variety of experimental techniques (as well as some *in silico* inferences). Although these techniques and the resulting data are (*i*) notoriously prone to false positives and negatives (10, 11), and (*ii*) result in highly idealized and averaged network structures (12), such interaction datasets are increasingly turning into useful tools for the analysis of the functional (e.g., ref. 13) and evolutionary properties (14) of biological systems. In particular, in *Saccharomyces cerevisiae* we are beginning to have a fairly complete description of the protein interaction network that is accessible with current experimental technologies; the recent high-quality literature-curated dataset of Reguly *et al.* (15) provides us with a dataset that should be almost completely free from false positives. For most other organisms, however, interaction data are still far from complete and it has recently been shown that subnetworks, in general, have qualitatively different properties from the true network (16–18). Although the importance of network-sampling properties had only been realized relatively recently, this aspect of most systems biology data are increasingly being recognized (11, 19) as important.

There are, however, some properties of the true network that can be inferred even from subnet data, and here we show that the total network size is one property for which this is the case. Present protein-interaction datasets enable us to estimate the size of the interactomes in different species by using graph theoretical invariants. This is particularly interesting for species where more than one experimental dataset is available. Below we first describe a robust and very general estimator of network size from partial network data that overcomes this problem. We then apply it to available PIN data in a range of eukaryotic organisms. In supporting information (SI) *Text* we demonstrate the power of this approach by using extensive simulation studies.

Estimating Interactome Size

Here, we develop an approach for estimating the size of a network from incomplete data. We will show below (and by using extensive simulations in *SI Text*) that for a given species estimates from different independent datasets—generated by different methods such as yeast-two-hybrid and TAP tagging—yield estimates for the interactome size that are in excellent agreement.

We are concerned with a true network, $\mathcal{N}$, which has $N_{\mathcal{N}}$ nodes and $M_{\mathcal{N}}$ edges. The sets of nodes and edges are given by $\mathcal{V}_{\mathcal{N}}$ and $\mathcal{E}_{\mathcal{N}}$, respectively; these define the graph representation of the true network:

$$G_{\mathcal{N}} = (\mathcal{V}_{\mathcal{N}}, \mathcal{E}_{\mathcal{N}}). \quad [1]$$

We pick a subset of nodes $\mathcal{V}_{\mathcal{S}} \subseteq \mathcal{V}_{\mathcal{N}}$ and study properties of the subgraph $G_{\mathcal{S}}$ induced by the nodes in $\mathcal{V}_{\mathcal{S}}$

$$G_{\mathcal{S}} = (\mathcal{V}_{\mathcal{S}}, \mathcal{E}_{\mathcal{S}}), \quad [2]$$

where the set of edges observed in the $\mathcal{S}$ is a subset of the total set of edges, $\mathcal{E}_{\mathcal{S}} \subseteq \mathcal{E}_{\mathcal{N}}$. Our aim is to predict the number of interactions in the true network $G_{\mathcal{N}}$ based on the available data in the subnet, $G_{\mathcal{S}}$.

We assume that the network, $G_{\mathcal{N}}$, is generated according to some (unknown) model characterized by a parameter (vector) θ, and subsequently the observed network, $G_{\mathcal{S}}$ is sampled from it. Then

$$P_{\theta,p}(G_{\mathcal{S}}) = \sum_{G_{\mathcal{N}} \supseteq G_{\mathcal{S}}} P_p(G_{\mathcal{S}}|G_{\mathcal{N}}) P_\theta(G_{\mathcal{N}}), \quad [3]$$

Author contributions: M.P.H.S., M.L., and C.W. designed research; M.P.H.S., T.T., E.d.S., M.L., and C.W. performed research; M.P.H.S., T.T., E.d.S., R.S., H.J.A., and C.W. analyzed data; and M.P.H.S., M.L., and C.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

See Commentary on page 6795.

[§]To whom correspondence may be addressed. E-mail: m.stumpf@imperial.ac.uk or wiuf@birc.au.dk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0708078105/DCSupplemental.

where it is assumed that the sampling is independent of the network-generating model. The parameter p refers to a general sampling process, and not only independent node sampling. Furthermore, we assume the order $N_{\mathcal{N}}$ of the network is known and allow nodes to be annotated with information not related to the wiring of the network (e.g., GO terms or protein family classes). Consequently, the sum is over networks, $G_{\mathcal{N}}$, with $N_{\mathcal{N}}$ (labeled) nodes only. For convenience, we take labeling information to be included in $N_{\mathcal{N}}$ and $N_{\mathcal{S}}$ (the order of $G_{\mathcal{S}}$).

If sampling only depends on the nodes in the network and not on their connections, then $P_p(G_{\mathcal{S}}|G_{\mathcal{N}})$ splits into a product of two terms,

$$P_{\theta,p}(G_{\mathcal{S}}) = Q_p(N_{\mathcal{S}}) \sum_{G_{\mathcal{N}} \supseteq G_{\mathcal{S}}} q(G_{\mathcal{S}}, G_{\mathcal{N}}) P_\theta(G_{\mathcal{N}}), \qquad [4]$$

where $Q_p(N_{\mathcal{S}})$ is a term denoting the probability of sampling the nodes in the observed PIN and $q(G_{\mathcal{S}}, G_{\mathcal{N}})$ denotes how many ways this can be done given the (labeled) nodes in $G_{\mathcal{N}}$—by assumption, labeling of nodes is the same in all possible $G_{\mathcal{N}}$s. For example, if the nodes are unlabeled and have degree zero, then $q(G_{\mathcal{S}}, G_{\mathcal{N}}) = \binom{N_{\mathcal{N}}}{N_{\mathcal{S}}}$. If all nodes have degree one, a similar factor can be derived based on the number of degree one (d_1) and degree zero (d_0) nodes that are observed in the PIN: $q(G_{\mathcal{S}}, G_{\mathcal{N}})$ is the number of ways one can choose d_0 and d_1 out of the $N_{\mathcal{N}}/2$ pairs of connected nodes in $G_{\mathcal{N}}$. If all nodes are labeled then $q(G_{\mathcal{S}}, G_{\mathcal{N}}) = 1$, because one can only select the nodes in the PIN in one way.

It follows that $Q_p(N_{\mathcal{N}})$ is sufficient for inference on p (the remaining part of the likelihood does not depend on p). In the case of independent node sampling, each node with probability p, we have $Q_p(N_{\mathcal{S}}) = p^{N_{\mathcal{N}}}(1-p)^{N_{\mathcal{N}} - N_{\mathcal{S}}}$ and the maximum likelihood estimate of p is

$$\hat{p} = \frac{N_{\mathcal{S}}}{N_{\mathcal{N}}} \qquad [5]$$

which is unbiased and consistent.

From the likelihood Eq. **4** it follows that

$$P_\theta(G^*_{\mathcal{N}}|G_{\mathcal{S}}) = \frac{q(G_{\mathcal{S}}, G^*_{\mathcal{N}}) P_\theta(G^*_{\mathcal{N}})}{\sum_{G_{\mathcal{N}} \supseteq G_{\mathcal{S}}} q(G_{\mathcal{S}}, G_{\mathcal{N}}) P_\theta(G_{\mathcal{N}})}, \qquad [6]$$

where $G^*_{\mathcal{N}}$ is a specific network (to distinguish it from the sum over all networks in the denominator). Note that this conditional probability does not depend on p and that, in principle, we can only gain knowledge about the interactome if something is assumed about the network-generating model. Note also that this is a general restriction that is not related to independent node sampling alone.

A reasonable estimate of the edge probability in $G_{\mathcal{N}}$ is

$$\hat{\pi} = \frac{2M_{\mathcal{S}}}{N_{\mathcal{S}}(N_{\mathcal{S}}-1)}, \qquad [7]$$

where $M_{\mathcal{S}}$ is the number of edges in the PIN. It leads to the following estimate of the interactome size:

$$\hat{M}_{\mathcal{N}} = M_{\mathcal{S}} \frac{N_{\mathcal{N}}(N_{\mathcal{N}}-1)}{N_{\mathcal{S}}(N_{\mathcal{S}}-1)}, \qquad [8]$$

The estimate is unbiased and consistent provided the network-generating mechanism ensures some form of uniformity, as is the case for random graphs (Figs. S1 and S2). For example, if $G_{\mathcal{N}}$ has a star topology with one node of degree $N_{\mathcal{N}} - 1$ and the remaining of degree 1, then $M_{\mathcal{S}} = 0$ with probability $1-p$ and $M_{\mathcal{S}} = (N_{\mathcal{N}}-1)/\hat{p}$ with probability p; hence, $\hat{M}_{\mathcal{N}}$ is not consistent. We will demonstrate below that the assumption of independent sampling of nodes is not too restrictive and should apply to many, in particular, high-throughput, experimental studies.

So far we have assumed that the number of ORFs, $N_{\mathcal{N}}$ in an organism is known from genome surveys. Total genome size is, however, still not precisely known in most organisms. Uncertainty in the $N_{\mathcal{N}}$ is, however, easily incorporated. Assume that the value $N_{\mathcal{N}}$ is associated with an error or uncertainty ε (i.e., if the genome contains N_0 protein-coding genes of which $N_{\mathcal{N}}$ are known, then $\varepsilon = (N_0 - N_{\mathcal{N}})/N_0$. Then let $N_{\mathcal{N}} := N_0(1 \pm \varepsilon)$ and for $\varepsilon \lesssim 0.1$ we have

$$\tilde{P} = \frac{N_{\mathcal{S}}}{N_0} \approx (1 \pm \varepsilon)\frac{N_{\mathcal{S}}}{N_{\mathcal{N}}} = (1 \pm \varepsilon)\hat{p}. \qquad [9]$$

Replacing $\hat{p}$ in Eq. **8** with $\tilde{p}$ yields the error-corrected estimate for the true network size

$$\tilde{M}_{\mathcal{N}} = \frac{M_{\mathcal{S}}}{\tilde{p}^2} \approx (1 \pm 2\varepsilon)\frac{M_{\mathcal{S}}}{\hat{p}^2}. \qquad [10]$$

Thus, an uncertainty of ε in the number of nodes in the true network results in an uncertainty of 2ε for the number of edges in the true network.

To assess the variability of the estimator we can construct approximate bootstrap confidence intervals (CI) (20). The number of edges is given by

$$M_{\mathcal{S}} = \frac{1}{2}\sum_{i \in \nu_{\mathcal{S}}} d_i, \qquad [11]$$

in terms of the degree sequence. Now let $\mathbf{d} = \{d_1, d_2, \ldots, d_{N_{\mathcal{S}}}\}$ be the set of degrees of all of the nodes in the graph $G_{\mathcal{S}}$ describing the subnet. Then we generate bootstrap replicates, $\mathbf{d}^*$, by sampling the degrees of the nodes in the sample with replacement $N_{\mathcal{S}}$. For each bootstrap replicate, $\mathbf{d}^*$, we obtain an estimate $M^*_{\mathcal{S}}$ (which may be a noninteger because of the factor 1/2 in Eq. **11**; this does not affect the estimator). Creating a sufficiently large number of bootstrap replicates, $\mathbf{d}^*$, thus allows us to calculate the bootstrap CIs; these have very good coverage properties, as shown in Figs. S3 and S4.

The derivation of Eq. **8** does not depend on any restrictive assumptions (see *SI Text*) but is a generic property of random graphs and their subnets. Crucially Eq. **8** is valid irrespective of the degree sequence or other summary statistics of the networks[††]; confidence intervals (CI) and their coverage properties (20) may, however, depend on the degree sequence or network structure. Because there is no sufficient statistic for general networks (17) [i.e., a summary statistic that would include all information about the likelihood (21) of a network] it is also not possible to improve on these estimators by, for example., including the numbers of observed triangles or the clustering coefficient. The only limitation is the assumption of independent sampling. This is, however, also implicit in all previous attempts at estimating interactome sizes (22–24). Below we show how nonrandom sampling schemes can be described and how false-positive and false-negative rates of PIN data affect our estimate.

Other Node-Sampling Schemes

The above approach can be generalized for datasets that are ascertained in certain ways and can thus also deal with experimental bias.

[††]Eq. **8** is a general result for general (random) graphs; it is equally true for all ensembles of random graphs such as Erdös–Rényi and scale-free random graphs. In *SI Text* we further illustrate the simple quadratic relationship by using simulations.

Table 1. Dataset properties and predicted interactome sizes

Species	Dataset	Nodes	Edges	$\hat{M}_{\mathcal{N}}$	95% CI
S. cerevisiae	Uetz *et al.* (29)	1,328	1,389	28,472	26,650–30,460
	Ito *et al.* (30)	3,245	4,367	14,940	13,500–16,650
	Ho *et al.* (31)	871	694	33,234	31,750–34,810
	Gavin *et al.* (32)	726	367	25,391	23,280–27,710
	DIP	4,959	17,226	25,229	24,100–26,440
D. melanogaster	Giot *et al.* (34)	6,991	20,240	75,506	72,700–78,400
	Stanyon *et al.* (35)	362	1,611	2,505,545	2,192,900–2,843,800
	Fromstecher *et al.* (36)	1,200	1,657	211,877	180,419–248,640
	DIP	7,451	22,636	74,336	71,700–77,100
C. elegans	Li *et al.* (33)	2,622	3,955	242,578	221,850–265,700
	DIP	2,638	3,970	240,544	220,030–263,270
H. sapiens	Stelzl *et al.* (8)	1,665	3,083	646,557	588,990–706,640
	Rual *et al.* (9)	1,527	2,529	631,646	564,460–703,830
	DIP	1,085	1,346	672,918	625,170–722,670

The datasets by Stanyon *et al.* (35) and Formstecher *et al.* (36) were detailed and highly focused studies of cell regulation and cancer/signaling-related proteins in *D. melanogaster*, respectively; because only interactions among restricted sets of proteins were studied, these findings have resulted in very detailed but highly localized maps that result in overestimates for the size of the global PIN. Some of the different estimates are not truly independent; for example, DIP contains supersets of various published datasets for the different organisms (this is true, in particular, for *C. elegans*). As such, we would expect it to yield similar results to some of the individual datasets. The exceptions to this are the human datasets where very little overlap was found. Datasets of the individual publications (8, 9, 29–36) were downloaded from the IntAct database resource at European Bioinformatics Institute (EBI) (see *SI Text*; numbers in the IntAct database differ slightly from those in the original publications and, furthermore, we have removed all self-interactions).

Independent but Nonuniform Sampling. We assume independent sampling of nodes. Let node i have a probability p_i for being included in the subnet. We allow $p_i \neq p_j$ and only assume that the p_i values are drawn independently from the same probability distribution,

$$p_i \sim F(\alpha), \quad [12]$$

where α is a parameter (potentially vector valued). The properties of F are not of importance. It follows that $\hat{p}$ is unbiased

$$E(\hat{p}) = \frac{E(M_{\mathcal{S}})}{M_{\mathcal{N}}} = E(p_i) = \langle p \rangle, \quad [13]$$

and also consistent, because

$$\mathrm{Var}(\hat{p}) = \frac{\mathrm{Var}(p_i)}{M_{\mathcal{N}}} = \frac{p(1-p)}{M_{\mathcal{N}}} \to 0, \quad [14]$$

for large networks. Now consider an edge e_{ij} $(i \neq j)$; then the probability of observing this edge in the subnet is

$$\pi_{ij} = p_i p_j, \quad [15]$$

and

$$E[\pi_{ij}] = \langle p \rangle^2. \quad [16]$$

Likewise $\hat{p}^2$ is consistent (25), hence also unbiased for large networks.

Dependent Sampling. Here, we assume as above that p_i is drawn from some probability distribution, $p_i \sim F_i(\alpha)$ that might, however, depend on information related to node i, for example, the degree or functional classification of i; that is, $F_i(\alpha) = F(\alpha; D_i)$, where D_i denotes this information. Although measures for expression abundance may be such a factor, this appears not to be the case for the datasets considered here (Fig. S5). Hence, we might take D_i as an additional parameter in the function F.

In addition, we assume the network is uncorrelated with respect to this information, that is, $P(D_i, D_j) = P(D_i)P(D_j)$; and, given the probabilities p_i, we assume nodes are drawn independently of each other. This assumption is justified for all networks in which the degree–degree correlation of interacting nodes is determined by the degree distribution. This is approximately the case for the networks considered here[‡‡]. It follows that

$$E(\hat{p}) = \frac{E(M_{\mathcal{S}})}{M_{\mathcal{N}}} = E(p_i) = \langle p \rangle, \quad [17]$$

that is, $\hat{p}$ is unbiased. Note that

$$\begin{aligned} E(p_i p_j) &= \int_{D_i, D_j} E(p_i p_j | D_i, D_j) dP(D_i, D_j) \\ &= \int_{D_i} E(p_i | D_i) dP(D_i) \int_{D_j} E(p_j | D_j) dP(D_j) \\ &= E(p_i)E(p_j) = \langle p \rangle^2, \end{aligned} \quad [18]$$

which in turn leads to

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{M_{\mathcal{N}}} \to 0, \quad [19]$$

and consequently consistency. Likewise, it follows that $E(\pi_{ij}) = E(p_i p_j) = \langle p \rangle^2$, and that the edge sampling probability consistently is estimated by $\hat{p}^2$.

Effects of Uncertain Data on Estimated Interactome Sizes

So far we have assumed that the interaction data are correct. This is not the case for protein interaction data (10–12, 26–28). Here, we show that it is possible to include noisy data and that the estimates given in Table 1 (see also Fig. 2) are not likely to change severely (e.g., by an order of magnitude) for realistic rates

[‡‡]The degree–degree distribution is not significantly different from the product degree distribution (by using the Kolmogorov–Smirnov test); that is, $P(k, l) \approx P(k)P(l)$ for the datasets considered here.

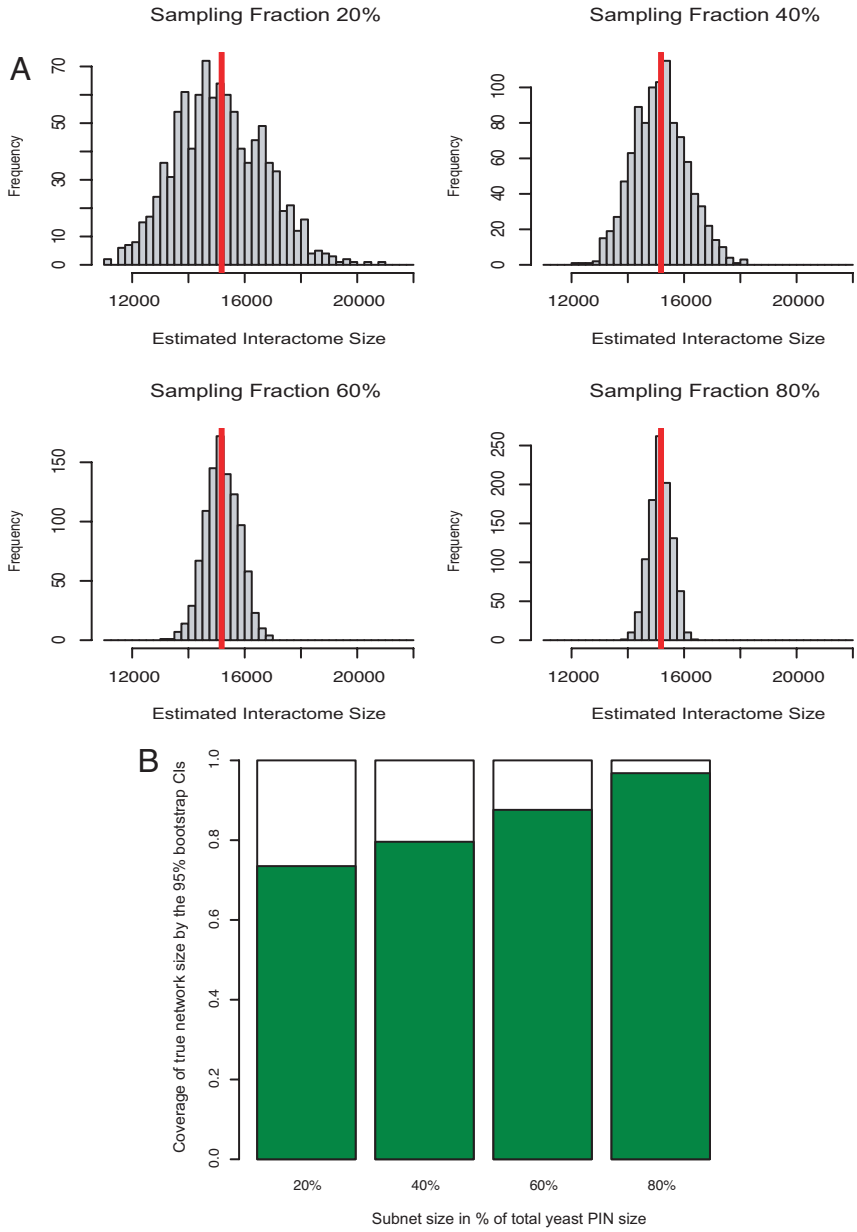

Fig. 1. Performance of the estimator, Eq. **8**, for the yeast network. Here, the DIP dataset was taken as a gold-standard ''true'' interaction network. (*A*) True network size (red bars) and histograms of predicted sizes for subnets that were created by sampling 20%, 40%, 60%, and 80% of nodes with equal probability. (*B*) Fraction of estimates obtained from 1,000 independent subnets (covering 20%, 40%. 60%, and 80% of the nodes in the true network) where the empirical 95% bootstrap confidence interval (based on 1,000 replicates) contains the true value (green).

of false positives and false negatives. We note that the sampling theory developed in the previous sections needs modification to take false positives and false negatives into account; for example, the sum in Eq. **4** should be over all possible networks and not just those containing the observed PIN data.

Let the number of true interactions in a network with N nodes be denoted by M; if the data collection process is not perfect, then (assuming independence) the number of reported interactions, $\tilde{M}$ will generally be different from M. Now let M_{TP}, M_{FN}, M_{FP}, and M_{TN} denote the true-positive, false-negative, false-positive, and true-negative results, respectively. We trivially have

$$M = M_{\mathrm{TP}} + M_{\mathrm{FN}} \qquad [20]$$

and

$$\tilde{M} = M_{\mathrm{TP}} + M_{\mathrm{FP}}. \qquad [21]$$

The rates for true positives and false negatives are defined by

$$\mu := \frac{M_{\mathrm{TP}}}{\tilde{M}} \qquad [22]$$

$$\rho := \frac{M_{\mathrm{FN}}}{M}. \qquad [23]$$

Thus, for a given number of reported edges/interactions and estimates of the true-positive and false-negative rates, $\hat{\mu}$ and $\hat{\rho}$, we obtain an estimate for the true number of interactions

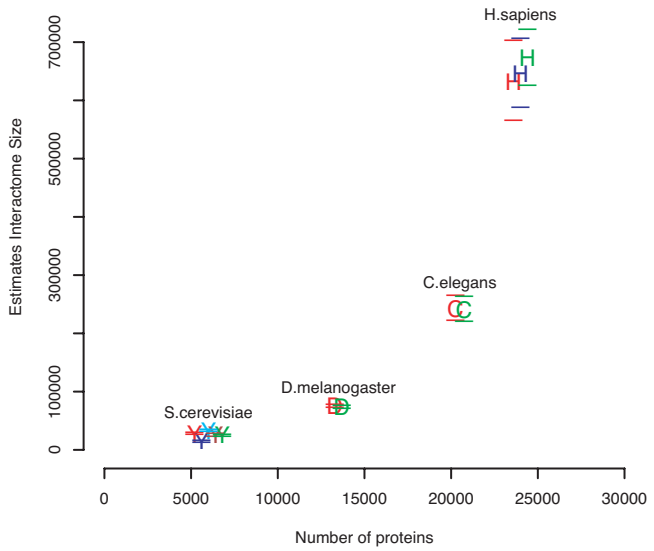

Fig. 2. Estimated interactome sizes for humans and three other eukaryotic species for which high-throughput interaction data are available. The letters denote the approximate position of the point estimate, $\hat{M}_{\mathcal{N}}$, and the horizontal bars indicate the range of the approximate 95% CIs (obtained from 10,000 bootstrap replicates; see *SI Text* for details). (The yeast and human datasets are largely independent but there is large overlap between the datasets for *D. melanogaster* and especially *C. elegans*.)

$$\hat{M} = \frac{\mu}{1-\rho}\tilde{M}. \qquad [24]$$

Thus, for a fixed network (or subnet) the false-positive and false-negative rates affect the estimates of the true number of interactions in a simple linear manner (see Fig. S6).

Results

We use Eq. **8** to estimate interactome sizes in humans and three other eukaryotic organisms: *S. cerevisiae* (29–32), *C. elegans* (33), and *D. melanogaster* (34–36). But we begin with an illustration of the power of this simple estimator by applying it to *S. cerevisiae* PIN data; here, we have treated the presently available PIN data as a proxy for a complete "interaction network" whose size we are trying to predict. In Fig. 1*A* we show the distributions of estimates obtained from 1,000 randomly chosen subnets covering 20%, 40%, 60%, and 80% of the available PIN data [taken from the Database of Interacting Proteins (DIP) (37)]. In Fig. 1*B* we show the coverage properties of the bootstrap 95% CIs for sampling the same sampling fractions. Together with the simulation studies discussed in *SI Text*, the results in Fig. 1 suggest that the estimator $\hat{M}_{\mathcal{N}}$ provides an accurate and reliable way of estimating interactome sizes from present data. Interactome size estimates and their CIs for experimental PIN datasets are shown in Table 1 and Fig. 2 for the organisms considered here. The DIP datasets (always shown in green) are mainly based on high-throughput studies, supplemented by interactions collected from the literature; as such, they generally cannot be treated as independent from the other datasets. For humans, however, there is negligible overlap between the DIP databases and the two recent high-throughput surveys and we can treat the three estimators as approximately independent.

Based on the results in Table 1 and Fig. 2, we would therefore expect—given present experimental methods and ignoring multiple splice variants—the human interactome to contain ≈650,000 protein interactions. Thus, it is approximately an order of magnitude larger than the estimated *D. melanogaster* interactome, and a factor of 3 more complex than the estimated *C. elegans* interactome; this contrasts with relative genome sizes of ≈1.8 and ≈1.2, respectively. The results for the *S. cerevisiae* PIN suggest that it will ultimately contain ≈25,000–35,000 interactions (see also Table 1); this agrees well with previous estimates (22, 23). It also agrees well with estimates obtained from the recent data generated by Reguly *et al.* (15): for the pure literature-curated set we obtained 37,000 interactions; for the complete network data we obtained an estimate of ≈35,000 interactions in the yeast PIN. These two datasets were, however, collected from the literature and the sampling process is thus much harder, perhaps even impossible, to model accurately.

By using Eq. **24** the impacts of false-positive and false-negative rates are easily assessed (see also ref. 38). We find that the linear effect of the error rates on the estimated number of true interactions results in a comparatively modest effect. The estimates of the true-positive rates in PIN datasets range from 35% (33) to 84% (34); there are fewer estimates for the false-negative rate that are on the order of 20–40% (10) obtained for different *S. cerevisiae* datasets. It appears that, for realistic rates of true positive and false positive, the estimate of the human interactome size remains very similar compared to the simple estimate obtained in this article of ≈650,000 protein–protein interactions. Similar curves can be drawn for the other species, too, and in each case we obtain comparable values for most combinations of realistic error rates. Thus, we believe that error rates exert a comparatively moderate effect on the estimator (Eq. **8**).

Overall, it therefore appears that estimates obtained from Eq. **8** should be accurate to within less than an order of magnitude even under the very worst circumstances. A much more realistic estimate, however, can be obtained from comparing the different and essentially independent estimates for *S. cerevisiae*. These findings suggest that an accuracy of approximately a factor of 2 is more realistic. Reassuringly, these results are confirmed when applying a recent multimodel inference procedure (39) that deals with incomplete network data.

Discussion

We have shown that it is possible to estimate the size of interactomes reliably from present partial interaction data. Our estimator is powerful and robust, relying on assumptions that appear to be met by typical systematic high-throughput studies. Unlike the previous approach of Hart *et al.* (24), who implicitly assume that interactions do not occur between surveyed proteins and those not yet surveyed, our estimate deals with missing data in a coherent and statistically meaningful manner; the route taken by Grigoriev (23) can be understood as a special case of the present approach when two or more datasets are available. Moreover, noise and different sampling/ascertainment strategies are straightforwardly included in the analysis (38, 40). We have illustrated the power of this approach by using simulated sampling processes in *S. cerevisiae* and have found that the estimator, Eq. **8**, and the bootstrap confidence intervals have very good coverage properties. We have then applied this inferential framework to published datasets in four eukaryotic organisms. We found that the predicted interactome sizes differ quite considerably between these species. For example, the human interactome appears to be an order of magnitude larger than the *D. melanogaster* interactome. Unfortunately, for maize and rice, which have comparable or even larger number of genes to humans, only tiny PIN datasets are available and we cannot obtain useful estimates for their respective interactome sizes. If conventional assumptions about the different complexity of organisms are indeed correct, and if interactome size does reflect organismic complexity (1–3, 41), then we would expect these organisms to have smaller interactomes than humans. The increase of interactome size with number of proteins/ORFs should thus not be uniform or even monotonic. We note that the

estimate of ≈ 650,000 interactions means that the human PIN will still be relatively sparse: this corresponds to only ≈0.2% of all possible pairwise interactions being present; for most other species, however, the network is even sparser.

There are a number of other factors that may contribute to an explanation of the increase in phenotypic *bauplan* complexity between species: the diversity of the transcriptome (42) and protein-domain architecture (43) have all been implicated in the literature. Here, we have demonstrated that interactome sizes are consistent with biological intuition about the complexity of eukaryotic organisms. We note that our estimator is very flexible and reflects the quality of present data: we predict the number of interactions that are detectable given present experimental technology. For example, we have not considered (physiologically probably very important) transient or condition-specific interactions. Should more sensitive and reliable experimental methodologies or better estimates of experimental error rates become available in the future, then Eq. **8** can, of course, be used to predict an updated number of protein–protein interactions for an organism. Our formalism is also readily extended to directed network data (such as gene-regulation networks).

As a final note, we want to stress that the estimates necessarily reflect experimental technology. Thus, the estimates in Table 1 refer only to the types of interactions that are detectable given present experimental methods and protocols. The estimator for the size of the true network, however, will remain universally correct for suitable datasets and for all types of networks. We will thus be able to use it in the future and apply it to other network datasets as well.

ACKNOWLEDGMENTS. This work was supported by the Wellcome Trust (M.P.H.S., E.d.S., and T.T.), the Royal Society and the Carlsberg Foundation (M.P.H.S. and C.W.), and an EMBO Young Investigator fellowship (to M.P.H.S.). C.W. is supported by the Danish Research Council.

1. Lander E, *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
2. Venter J, *et al.* (2001) The sequence of the human genome. *Science* 291:1304–1351.
3. Copley R (2008) The animal in the genome: comparative genomics and evolution. *Philos Trans R Soc London Ser B*, 363:1453–1461.
4. Tian B, Pan Z, Lee JY (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res* 17:156–165.
5. Henikoff S (2005) Histone modifications: Combinatorial complexity or cumulative simplicity? *Proc Natl Acad Sci USA* 102:5308–5309.
6. Hegde RS, Bernstein HD (2006) The surprising complexity of signal sequences. *Trends Biochem Sci*, 31:563–571.
7. Stranger BE, *et al.* (2007) Population genomics of human gene expression. *Nat Genet* 39:1217–1224.
8. Stelzl U, *et al.* (2005) A human protein–protein interaction network: A resource for annotating the proteome. *Cell*, 122:957–968.
9. Rual J, *et al.* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437:1173–1178.
10. Bader JS, Chaudhuri A, Rothberg JM, Chant J (2004) Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* 22:78–85.
11. Deeds EJ, Ashenberg O, Shaknovich EI (2006) A simple physical model for scaling in protein–protein interaction networks. *Proc Natl Acad Sci USA* 103:311–316.
12. de Silva E, Stumpf M (2005) Complex networks and simple models in biology. *J R Soc Interface* 2:419–430.
13. Rives AW, Galitski T (2003) Modular organization of cellular networks. *Proc Natl Acad Sci USA* 100:1128–1133.
14. Stumpf M, Kelly W, Thorne T, Wiuf C (2007) Evolution at the system level: The natural history of protein interaction networks. *Trends Ecol Evol* 22:366–373.
15. Reguly T, *et al.* (2006) Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J Biol* 5:11.
16. Stumpf M, Wiuf C, May R (2005) Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proc Natl Acad Sci USA* 102:4221–4224.
17. Stumpf M, Wiuf C (2005) Sampling properties of random graphs: The degree distribution. *Phys Rev E* 72:036118.
18. Wiuf C, Stumpf M (2006) Binomial subsampling. *Proc R Soc A* 462:1181–1195.
19. Han J, Dupuy D, Bertin N, Cusick M, Vidal M (2005) Effect of sampling on topology predictions of protein–protein interaction networks. *Nat Biotechnol* 23:839–844.
20. Efron B, Tibshirani R (1998) *An Introduction to the Bootstrap* (Chapman & Hall/CRC, New York).
21. Cox D, Hinkley D (1974) *Theoretical Statistics* (Chapman&Hall/CRC, New York).
22. Hazbun T, Fields S (2001) Networking proteins in yeast. *Proc Natl Acad Sci USA* 98:4277–4278.
23. Grigoriev A (2003) On the number of protein–protein interactions in the yeast proteome. *Nucleic Acids Res* 31:4157–4161.
24. Hart G, Ramani A, Marcotte E (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol* 7:120.
25. Silvey S (1975) *Statistical Inference* (Chapman & Hall, New York).
26. von Mering C, *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417:399–403.
27. Lappe M, Holm L (2004) Unraveling protein interaction networks with near-optimal efficiency. *Nat Biotechnol* 22:98–103.
28. Uetz P, Finley R (2005) From protein networks to biological systems. *FEBS Lett* 579:1821–1827.
29. Uetz P, *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403:623–627.
30. Ito T, *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 98:4569–4574.
31. Ho Y, *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415:180–183.
32. Gavin AC, *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141–147.
33. Li S, *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* 303:540–543.
34. Giot L, *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science* 302:1727–1736.
35. Stanyon C, *et al.* (2004) A *Drosophila* protein-interaction map centered on cell-cycle regulators. *Genome Biol* 5:R96.
36. Formstecher E, *et al.* (2005) Protein interaction mapping: A *Drosophila* case study. *Genome Res* 15:376–384.
37. Duan X, Xenarios I, Eisenberg D (2002) Describing biological protein interactions in terms of protein states and state transitions: The LiveDIP database. *Mol Cell Proteomics* 1:104–116.
38. de Silva E, *et al.* (2006) The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol* 4:39.
39. Stumpf M, Thorne T (2006) Multimodel inference of network properties from incomplete data. *J Integr Bioinf* 3:32.
40. Lin N, Zhao H (2005) Are scale-free networks robust to measurement errors? *BMC Bioinformatics* 6:119.
41. Tucker C, Gera J, Uetz P (2001) Towards an understanding of complex protein networks. *Trends Cell Biol* 11:102–106.
42. Carninci P, *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309:1559–1563.
43. Chothia C, Gough J, Vogel C, Teichmann S (2003) Evolution of the protein repertoire. *Science* 300:1701–1703.