Bull Math Biol (2012) 74:356–374
DOI 10.1007/s11538-011-9680-2

ORIGINAL ARTICLE

Assessing Coverage of Protein Interaction Data Using Capture–Recapture Models

W.P. Kelly · M.P.H. Stumpf

Received: 27 August 2010 / Accepted: 14 July 2011 / Published online: 26 August 2011

Abstract Protein interaction networks comprise thousands of individual binary links between distinct proteins. Whilst these data have attracted considerable attention and been the focus of many different studies, the networks, their structure, function, and how they change over time are still not fully known. More importantly, there is still considerable uncertainty regarding their size, and the quality of the available data continues to be questioned. Here, we employ statistical models of the experimental sampling process, in particular capture–recapture methods, in order to assess the false discovery rate and size of protein interaction networks. We uses these methods to gauge the ability of different experimental systems to find the true binary interactome. Our model allows us to obtain estimates for the size and false-discovery rate from simple considerations regarding the number of repeatedly interactions, and provides suggestions as to how we can exploit this information in order to reduce the effects of noise in such data. In particular our approach does not require a reference dataset. We estimate that approximately more than half of the true physical interactome has now been sampled in yeast.

Keywords Capture–recapture · Networks · Error rates · Protein interactions · Sampling

W.P. Kelly · M.P.H. Stumpf
Centre for Bioinformatics, Imperial College London, London, UK

W.P. Kelly
e-mail: william.kelly04@imperial.ac.uk

M.P.H. Stumpf (✉)
Institute of Mathematical Sciences, Imperial College London, London, UK
e-mail: m.stumpf@imperial.ac.uk

1 Introduction

The biological structure and function of organisms at the cellular level are the result of interactions between proteins, their various isoforms and other molecules. The resulting networks of biological interactions found in an organism have been studied using concepts from graph theory, and the quantitative analysis of biological networks has become important aspect of the description of biological systems (Alm and Arkin 2003; de Silva and Stumpf 2005; Schlitt and Brazma 2005; Heo et al. 2011). Here, an interaction between two proteins A and B is represented by an undirected edge in the interaction graph.

There have been numerous reports over the last decade highlighting the use of protein–protein interaction (PPI) network data, including how these can be used to understand molecular processes, disease phenotypes, and evolutionary properties of biological systems (e.g. Brun et al. 2003; Drees et al. 2005; Thorne et al. 2011). But PPI data have also been found to suffer from either abundant noise or potential experimental bias as a consequence of prior biological knowledge (Bader et al. 2004; de Silva et al. 2006), as well as being incomplete (Stumpf et al. 2005; Yang et al. 2008). Furthermore, PPIs are observed under a variety of experimental conditions and using different experimental methods. These data are then used to construct the protein interaction networks (PIN) of a species. The PIN forms the collection of the possible physical protein–protein interactions that can occur within the system. The experimental techniques employed to test for interactions will undoubtedly present differing amounts of noise and vary in which parts of the interactome they can map. There is thus a continuing need to reassess the reliability of PIN data as more interaction studies are published (von Mering et al. 2002). Previously, studies have estimated the false-discovery rate (FDR), other error rates, and interactome size of *S. cerevisiae* using similar methods. For instance, D'haeseleer and Church (2004) presented an *overlap method* (see Fig. 1 for an illustration of this approach) for estimating error rates in PIN data sets and the size of the *S. cerevisiae* PPI interactome. The FDR is estimated from three data sets: a reference set (taken from a "reliable" reference PPI source) along with two other large experimental sets. Using the overlaps between the sets, and assuming that these overlaps are error free (both as a consequence of the validation and also the assumption that the reference is highly accurate) and unbiased the ratio that should occur if the other sets are error free can be found.

The coverage of each experiment is of vital importance when considering error rates. Each experimental technique may be associated with different types of error or noise. Methods are also only applicable to a restricted subspace of the protein pairs in a given system, or may only be able to test a small proportion of the possible proteins that may interact with one another. Experimental noise and bias in which protein pairs are being assessed have been considered in the literature in order to produce FDR and interactome size estimates (Chiang et al. 2007; Huang et al. 2007; Gentleman and Huber 2007). Different false-positive rates and differences in the ability to test certain protein pairs, may not be identical across experiments and this will influence the error estimates obtained by overlap methods.

Grigoriev (2003) also assessed the size of the *S. cerevisiae* PIN, and estimated that each protein has somewhere between 3 and 5 distinct interactions. Hart et al.

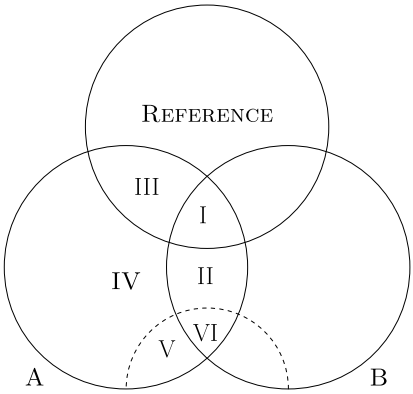


Fig. 1 Overlap method. The overlap found between three different interaction datasets, two of which are being compared against a reference set, can be used to estimate error rates. The FDR is estimated using the ratios of the number of interactions, I–VI, found for the 3 datasets. A and B are two experimental datasets whilst REFERENCE is a set of true interactions and assumed to be completely free of false–positive interactions. The *area separated by the dashed line* represents potential noise in the interaction data: not all reported interactions are true interactions, and this needs to be incorporated into the analysis. Even interactions found in several datasets, cf. interactions in VI, may be false–positive, although this is less likely the more often an interactions has been reported

(2006) found that earlier studies systematically underestimated the interactome size as each dataset probed different sets of protein pairs. Using estimated error rates and intersection of datasets they proceeded to infer which protein pairs had been tested. The estimated size is then scaled to take account of this coverage; based on this the *S. cerevisiae* PIN was predicted to have 38,000–76,000 interactions. Stumpf et al. (2008) assessed the size of *S. cerevisiae* and other interactomes. The authors assessed the overall size of the interactome by modelling the effect of sub-sampling from the complete true network and how this would affect the overlap between pairs of datasets. For the *S. cerevisiae* interactome, they obtained estimates of approximately 24,000–26,000 interactions.

The need to assess the coverage of each study was highlighted further by Gentleman and Huber (2007). Direct comparison of interaction data fails to take into account that not every possible protein-pair is being tested and negative results are hardly ever reported. Accordingly, unless this is explicitly considered, the overlap between studies will appear lower, thereby increasing the reported or apparent error rates. Chiang et al. (2007) performed error analysis using: the number of repeated reported interactions (ρ); the number of reported protein pairs for which no interaction exists (ϕ); and the number of non-repeated reported interactions (ζ). The error rates, for a pair of datasets, were found using the following relationships,

$$\begin{aligned}
\mathbb{E}(\rho) &= m(1-p_{\mathrm{FN}})^2 + m^* p_{\mathrm{FP}}^2, \\
\mathbb{E}(\phi) &= m\, p_{\mathrm{FN}}^2 + m^*(1-p_{\mathrm{FP}})^2, \\
\mathbb{E}(\xi) &= 2mp_{\mathrm{FN}}(1-p_{\mathrm{FN}}) + m^* p_{\mathrm{FP}}(1-p_{\mathrm{FP}}),
\end{aligned} \tag{1}$$

where m is the number of true interactions (observable interactome size) and m^* is the number of false interactions. If we know the protein set, V, tested against each other then a further condition is $\binom{|V|}{2} = m + m^*$. These equation reveal the trade-off between p_{FP} and p_{FN} when comparing two datasets. If a reference set is used to determine the FDR in experimental data then careful consideration of coverage and false–negative rates are required for accurate results.

Overall, there have been a number of FDR estimates for individual *S. cerevisiae* datasets, ranging from around 0.15 to 0.90 for interaction sets (Chiang et al. 2007; Gentleman and Huber 2007; Huang et al. 2007). As time has passed, and more data have become available, the size estimates have tended to increase. However, the current consensus appears to suggest that the number of distinct PPIs is between 20,000–40,000 (Hart et al. 2006; Stumpf et al. 2008).

2 Assessing Completeness of Interactome Data

2.1 Experimental Protein Interaction Data

In order to assess the interactome size of *S. cerevisiae* (or any other organism for which a sufficiently large number of studies have been performed), a collection of interaction studies is required. Capture–recapture methods (Shokouhi et al. 2006) rely on using the overlap between datasets in order to estimate the overall properties of the sampled set. However, in our case, we are sampling truly interacting and false–positive protein pairs from the possible set of distinct protein pairs: i.e. any reported interaction *may* be the consequence of systematic or stochastic error from the experimental methods (both of which are considered here) rather than a real physical interaction.

Here, we use the *S. cerevisiae* data collected in BioGRID, and the methods used to estimate the false-discovery rate (FDR) and interactome size are assessed on the complete set of data from this resource (version 2.0.60). The information required for the urn model used below consists of: the number of different protein pairs observed (m_{obs}); the number of interactions reported (s_{obs}); the number of distinct interactions reported (i_{obs}); and a list of experiment sizes ($\{r_1^{\mathrm{obs}}, r_2^{\mathrm{obs}}, \ldots, r_q^{\mathrm{obs}}\}$). Tables 1 and 2 summarize this information, which is used in order to assess the interactome size and FDR, and to compare small scale (SSE) and high-throughput (HTP) data.

2.2 Coverage of Current Protein-Interaction Data

The majority of the PPIs are found in experiments reporting more than 1,000 interactions. Accordingly, the HTP data used here are presumed to have been independently sampled from the sets of true or false interactions. Small-scale experiments (SSEs) make up the majority of the studies although they only produce a small proportion of reported interactions. They have been viewed as more reliable but are difficult to summarise from a sampling point of view; an independent sampling approach can be considered as the mean-field approximation to non-independent and non-random sampling processes (Stumpf et al. 2008, Supporting Information).

Table 1 Interaction datasets. The different subsets of physical PPIs from BioGRID used to find FDR, κ, and interactome size. The protein data exclude proteins that have only been reported as self-interacting; both subsets ($\geq$ and $<$) compared in the study are shown

Dataset		Model parameters		
Size, r_k	Experiments	All, s_{obs}	Distinct, i_{obs}	Proteins, $\check{n}$
All PPI	4,900	95,595	60,068	5,435
$\geq$5	1,878	89,688	58,518	5,367
$<$5	3,022	5,907	3,249	1,802
$\geq$10	871	83,347	56,320	5,312
$<$10	4,029	12,248	6,349	2,409
$\geq$100	51	65,520	48,976	5,222
$<$100	4,849	30,075	15,444	3,463
$\geq$1,000	15	55,136	42,288	4,997
$<$1,000	4,885	40,459	22,843	4,173

Table 2 PPI experimental method datasets. The experimental techniques data totals for each methodology in BioGRID. These are used to find the coverage, and estimated total number of putative interactions

Method	All, s_{obs}	Distinct, i_{obs}	Proteins, $\check{n}$	Experiments
Affinity Capture Luminescence	53	25	15	2
Affinity Capture MS	43,660	27,104	3,789	297
Affinity Capture Western	22,538	16,725	3,926	1,930
Biochemical Activity	5,822	5,518	1,760	313
Co-crystal Structure	298	206	237	197
Co-fractionation	623	543	370	98
Co-localization	466	380	239	121
Co-purification	2,506	2,239	1,039	195
Far Western	68	46	36	13
FRET	163	124	80	20
PCA	2,402	2,402	794	3
Protein-peptide	231	198	117	32
Reconstituted Complex	3,396	2,335	1,419	855
Two-hybrid	13,369	9,956	3,122	824

If there are $\check{n}$ proteins in a study then $m_{\text{obs}} = \binom{\check{n}}{2}$, protein pairs have potentially been assessed. A scaling factor, ρ, is defined so that an estimate of the complete interactome size can be found. This is necessary as each experimental technique may only be able to probe a subset of the complete set of protein-pairs, and there is no evidence regarding possible interactions between other unobserved proteins. Assuming a uniform distribution of true interactions across the proteome, the scaling factor to

find the complete interactome size as used in Stumpf et al. (2008) is

$$\rho = \frac{\binom{n}{2}}{\binom{\check{n}}{2}} = \frac{n(n-1)}{\check{n}(\check{n}-1)}, \tag{2}$$

where n is the number of proteins in yeast, here taken to be 5,800 (Hirschman et al. 2006).

The false–negative rate (FNR) and the false discovery rate (FDR) are defined as

$$\begin{aligned} \mathrm{FDR} &= \frac{\#\mathrm{FP}}{\#\mathrm{FP} + \#\mathrm{TP}}, \\ \mathrm{FNR} &= \frac{\#\mathrm{FN}}{\#\mathrm{FN} + \#\mathrm{TN}}, \end{aligned} \tag{3}$$

where FP is the set of false–positive interactions, TP the true–positives, FN the false–negatives, and TN the set of true–negatives found in the data. In order to estimate the negative set of results, we need to know all protein pairs that have been tested (information that is not always reported): We assume here that the pairwise interactions among all pairs of proteins which are reported to have an interaction have been tested: we then have $\binom{r_i^{\mathrm{obs}}}{2}$ potential pairs, where r_i^{obs} is the number of distinct proteins reported in experiment i. This will be more accurate for larger experiments, where pairs are tested comprehensively. Thus, for the overall data, this will give a good approximation to the amount of testing completed since the HTP tests contribute a majority of the positive, and similarly negative, interaction results.

2.3 Capture–Recapture Models for Protein Interaction Data

Here, we consider protein-pairs to be observed independently. This allows us to model repeated observations of such molecular interactions using techniques that had previously been used in ecology. The basic *capture–recapture* approach (Burnham and Overton 1978; Bunge and Fitzpatrick 1993; Chao 2001) applies when we have access to repeated samples from a fixed set of objects, here a set of interactions. These approaches have commonly been employed in order to find a population's size, or to elucidate its class structure.

2.3.1 Single Urn Models for Protein Interaction Data

The overlap found between two samples (i.e. the number of items recaptured) is used to estimate the complete population's size. Multiple capture–recapture (Shokouhi et al. 2006) is an extension of this approach to account for any number of samples. This has been used to estimate the size of different populations by observing the overlap between different samples (Xu et al. 2007). A *single urn* model is used to assess the overall population size, by relating the total observed sample to the number of distinct items sampled. The population considered with this model consists of only one class, so we cannot consider samples of both true and false interactions.

Suppose that interactions are *sampled* (where each sample reports one interaction) with replacement from an *urn* containing m different interactions. Having sampled i distinct interactions, the probability that the next sampled interaction has not been observed before is

$$\mathbb{P}(\text{novel interaction sampled} \mid i \text{ distinct interactions}) = \frac{m-i}{m}. \tag{4}$$

The number of samples drawn before encountering a novel interaction, given that $i \geq 0$ have already been collected, is geometrically distributed with parameter $\theta = \frac{m-i}{m}$. The expected number of samples required to find a novel interaction is then

$$\mathbb{E}(\text{samples, to find novel interaction}) = \frac{1}{\theta} = \frac{m}{m-i}.$$

Thus, using the linearity of expectations, the expected number of samples to collect i distinct interactions is

$$\begin{aligned}\mathbb{E}(\text{samples, to find } i \text{ distinct interactions}) &= \sum_{k=0}^{i-1} \mathbb{E}(\text{novel sample} \mid k \text{ distinct}) \\ &= 1 + \frac{m}{m-1} + \cdots + \frac{m}{m-i+1} \\ &= m \sum_{k=0}^{i-1} \frac{1}{m-k}. \end{aligned} \tag{5}$$

When m is known, then (5) can be used to estimate the number of samples necessary to find all of the distinct interactions. Alternatively, given the number of distinct interactions, i, and the number of interactions sampled, $\hat{s}$, is used to estimate m,

$$\hat{s} = m \sum_{k=0}^{i-1} \frac{1}{m-k}. \tag{6}$$

While this simple model cannot be used with noisy data it sheds light on the potential number of putative interactions that each experimental methodology can be expected to report. If we assume that each methodology is subject to systematic biases (rather than stochastic noise) then an estimate of the population size will hold information about the number of different putative interactions that may be expected to be reported. This, in turn, can give us an indication of the potential systematic bias.

2.3.2 Multiple Urn Model for Protein Interaction Data

A *multiple urn* model introduces the ability to sample from several sources of interactions as shown in Fig. 2: (i) true binary physical interactions and (ii) false interactions which are mis-reported in the data. Firstly, suppose that PPIs are reported from a set of protein pairs, E_{obs}, which contains n different proteins. Let m_{obs} be the size of

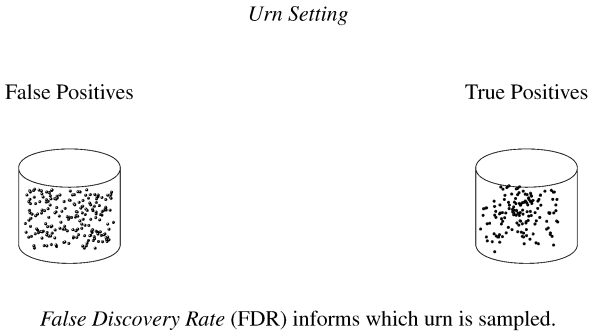


Fig. 2 Mixture urn model. The interactions are sampled from two urns, one containing true interactions and the other urn containing all those protein pairs which can be reported erroneously in the putative PPI data which is being analysed. The rate at which we sample from each urn is determined by the false-discovery rate (FDR)

E_{obs}. The reported PPIs are then either edges of the true interaction graph or the false interaction graph. These may be considered as being drawn from two urns containing either PPIs, $e_a = (v_i, v_j)$, found in $E \cap E_{\text{obs}}$, or false interaction protein pairs, $e_b = (v_k, v_h)$, found in $E' \cap E_{\text{obs}}$. Let m be the size of $E \cap E_{\text{obs}}$ and m' the size of $E' \cap E_{\text{obs}}$. The *proportion* of reported data that are found in E' is also the FDR, κ. Now let s be the number of interactions sampled from E and s' be sampled from E'. Then suppose s with $0 \leq s \leq s_{\text{obs}}$ is fixed,

$$s = \|(1-\kappa)s_{\text{obs}}\|, \tag{7}$$

and also trivially, $s' = \|\kappa s_{\text{obs}}\|$, where $\|x\|$ is the integer closest to x.

The observed number of distinct interactions, i_{obs}, is made up of those sampled from E and those from E'. Let i be sampled from E and i' be from E'; then because $E \cap E' = \emptyset$ we have $i_{\text{obs}} = i + i'$, and

$$\begin{aligned} s_{\text{obs}} &= s' + s, \\ i_{\text{obs}} &= i' + i, \\ m_{\text{obs}} &= m' + m. \end{aligned} \tag{8}$$

Here, s and s' are assumed to be the expected number of samples necessary to find i and i' interactions (as in Sect. 2.3.1). So for a given κ solutions for m and i are sought which satisfy (8) and

$$\begin{aligned} s &= m \sum_{k=0}^{i-1} \frac{1}{m-k}, \\ s' &= m' \sum_{k=0}^{i'-1} \frac{1}{m'-k}. \end{aligned} \tag{9}$$

Thus, in order to estimate the observable interactome size, FDR, and proportion of the true interactome currently sampled, we require that $g(m, i)$, is zero,

$$
\begin{aligned}
g(m,i) &= s_{\text{obs}} - m\sum_{k=0}^{i-1}\frac{1}{m-k} - m'\sum_{k=0}^{i'-1}\frac{1}{m'-k},\\
0 &= s_{\text{obs}} - m\sum_{k=0}^{i-1}\frac{1}{m-k} - (m_{\text{obs}}-m)\sum_{k=0}^{i_{\text{obs}}-i-1}\frac{1}{m_{\text{obs}}-m-k}.
\end{aligned}
\tag{10}
$$

The complete interactome size, m_Ω, is then found for a given solution (which is unique if it exists) using ρ, the scaling factor introduced in (2) to account for the coverage of the sampled data, and m:

$$
\begin{aligned}
m_\Omega &= \|\rho m\|\\
&= \left\|\frac{n(n-1)}{\check{n}(\check{n}-1)}m\right\|.
\end{aligned}
\tag{11}
$$

2.3.3 False Discovery Rates for Protein Interaction Data

Whereas the false discovery rate is explicitly modelled in the multiple urn model, the false negative rate is not captured directly. In order to find the number of false–negative results in the interaction data, we need to know or estimate the number of protein pairs which were assayed, and furthermore the number of interactions, which have been wrongly reported as non-existing. Each experiment produces an interaction graph on a number of different proteins, $\{n_1^{\text{obs}}, n_2^{\text{obs}}, \ldots, n_q^{\text{obs}}\}$. Analogously to how the coverage of the complete dataset is estimated, we assume that for each experiment the complete set of combinations of the observed proteins have been tested, i.e. $\binom{n_i^{\text{obs}}}{2}$ protein pairs. Thus, the number of tested protein pairs t_{obs} is

$$
t_{\text{obs}} = \sum_{i=1}^{q}\binom{n_i^{\text{obs}}}{2},
$$

where the total number of negative results is $t_{\text{obs}} - s_{\text{obs}}$. For uniform sampling (assumed to hold at least approximately within the limits of each experimental technique), we can use the multiple urn model to find the estimate

$$
\#\text{FN} = \frac{m}{m+m'}(t_{\text{obs}} - s_{\text{obs}}).
\tag{12}
$$

The false negative and false discovery rate are presented alongside the interactome estimates for the *S. cerevisiae* data in the next section.

3 Results

We have used MCR methods to assess the potential interactome spaces probed by experimental methods and also to assess the FDR and interactome size for *S. cerevisiae*. The *urn model* is also used to assess the differences between the error rates that are found in SSE and HTP experimental data.

3.1 The Effects of Different Experimental Methods

The BioGRID data are split according to experimental methodology, as shown in Table 2. These are used to assess sampling fractions, from which we then estimate the possible number of putative interactions that each methodology may be sampling from; this should give an indication of the potential levels of false–positive noise inherent to each method. Clearly, each methodology is attempting to report true interactions, but is actually sampling from a set of interactions that at best include the complete interactome and then also a number of false–positive interactions from the remaining possible set of protein pairs.

In order to enable a comparison across methodologies of the potential reporting amounts, the results need to be normalised using the scaling factor introduced in the methods section. Whilst it is in principle possible to produce an estimate for any of the methodologies (assuming some level of overlap in the experimental results), the techniques with a small number of experiments or very small sample sizes have been excluded. The primary focus is to attempt to compare HTP technologies from a global perspective and without using gold standard reference sets. To restrict the datasets, all techniques with fewer than 100 repeats are not considered further. Figure 3 shows the estimated number of interactions probed by the methods satisfying these criteria.

Figure 3 shows the potential observable number of interactions for each experimental PPI assay. If we assume that they are probing the same set of true PPIs, then the estimates provide evidence for the false-discovery rates for all methods; yeast two hybrid and affinity capture western appear to be the most reliable techniques. However, the huge difference between biochemical activity estimates and the other methods is probably an indication that the techniques are probing significantly different

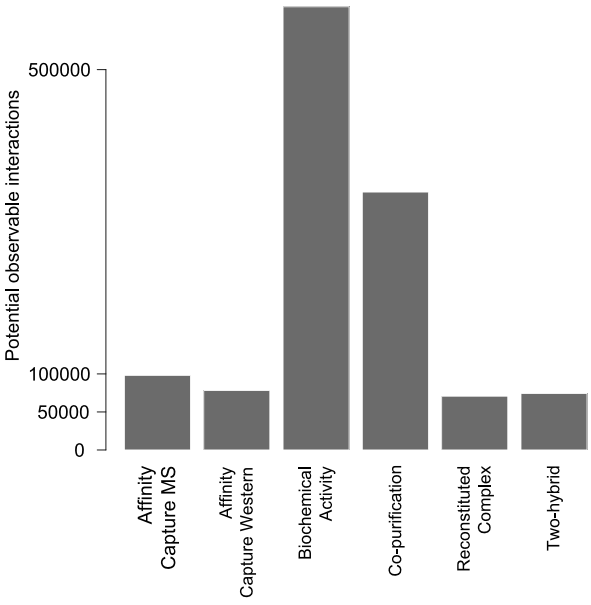


Fig. 3 *S. cerevisiae* technique interaction sets. These figures show the estimated number of interactions from which the methods are sampling. The sets will be a combination of true and false interactions; therefore, the relative sizes of the estimates can provide evidence for how much error the methodologies may be subject to, or, alternatively, whether they are probing comparable sets of interactions. The biochemical activity results appear distinct to the other physical techniques

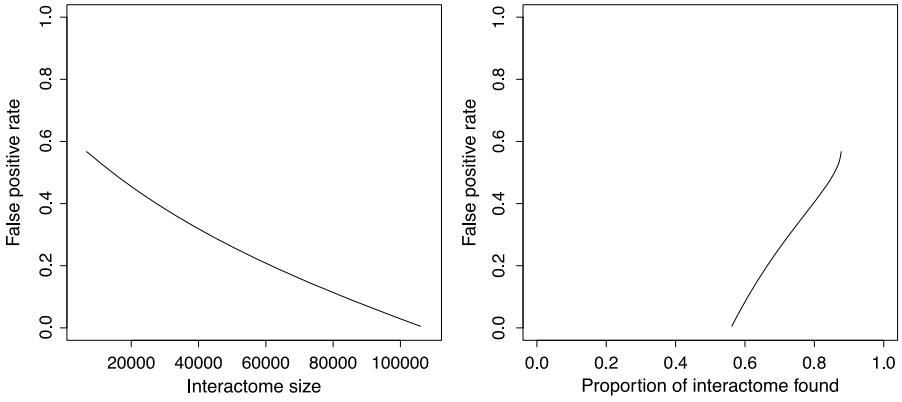
sets of interactions and so should be treated with care when combined for analysis. In particular, kinase interactions are not detected by any other experimental technique.

3.2 Estimating the Yeast Interactome Size

The BioGRID physical interaction data, Table 1, are used to find the results for FDR and interactome sizes shown in Fig. 4. The scaling factor, ρ, is approximately 1.36. Figure 4(a) shows the relationship between FDR, κ, and interactome size, ρm. This indicates that the FDR for the complete data should be between 0 and 0.6, whilst the interactome has fewer than 100,000 interactions. Using interactome size estimates guided by the literature of 20,000–40,000 interactions produces an estimated FDR across the complete data of 0.32–0.47. Similarly, using FDR estimates from the literature (which have predicted an FDR of larger than 0.2 in general) suggests that the interactome size has fewer than 60,000 interactions.

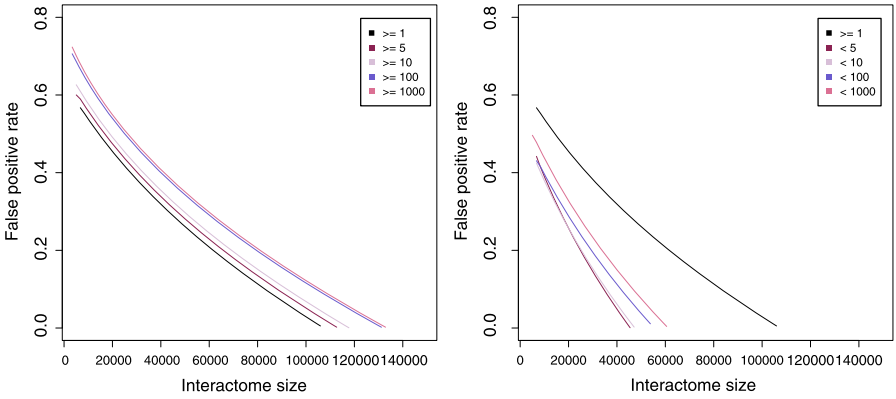
Figure 4(b) shows the proportion of the interactome, $\frac{i}{\rho m}$, that has been reported for the range of FDR estimates. Somewhere between 40% and 80% of the true yeast interactome has already been mapped, depending on the FDR. A higher FDR, due to its associated lower interactome size (in Fig. 4(a)), means that a higher proportion of the interactome has already been determined; and there will be fewer unseen true interactions if there is more noise (interactions sampled from false interaction urn, E'), a result consistent with how validation information is used to find the FDR and interactome size in the urn models.

The experimental data also provide an estimate for the completeness of the *S. cerevisiae* physical interactome sampled. The false negative rate can be estimated and used to give an estimate of the number of falsely unreported interactions. Our estimate is 0.90, which means that the experimental methods have failed to detect a number of interactions that is comparable to the true size of the yeast interactome, or



(a) Size, m, and FDR, κ. (b) Proportion known, $\frac{i}{\rho m}$, and FDR, κ.

Fig. 4 *S. cerevisiae* physical interactome size. The results found for *S. cerevisiae* using the single urn model are shown in the two plots. Figure 4(a) displays the estimated FDR and size. Figure 4(b) shows how the FDR relates to the proportion of the complete interactome that has been reported



(a) $HTP_\epsilon = \{P_k : r_{k,obs} \geq \epsilon\}$ (b) $SSE_\epsilon = \{P_k : r_{k,obs} < \epsilon\}$

Fig. 5 Experiment and interactome size. Plots show the interactome size, ρm, and FDR, κ, estimates found for SSE_ϵ and HTP_ϵ data using the single urn model. The HTP_ϵ data are less reliable than the complete data whilst the SSE_ϵ data are more reliable, although there is a general lack of validations in the smallest datasets (SSE_5) which may suggest either a higher FDR or poor modelling performance

somewhere between 15,000–40,000 true PPIs, depending on experimental technology.

3.3 The Role of Experiment Size

Each type of dataset may have a different FDR. To compare the noise found in HTP and SSE data, the complete interactome data are split by experiment, P_k, according to experiment size, $r_{k,\text{obs}}$. The estimated FDR, κ, for a given interactome size is then compared between $\text{SSE}_\epsilon = \{P_k : r_{k,\text{obs}} < \epsilon\}$ and $\text{HTP}_\epsilon = \{P_k : r_{k,\text{obs}} \geq \epsilon\}$ data for $\epsilon \in \{5, 10, 100, 1000\}$.

Figure 5 shows the relationship between FDR, κ, and interactome size, ρm, for SSE and HTP datasets. In general, for a fixed interactome size, SSE experiments have a lower FDR than the HTP data (when defined using the same ϵ) and HTP data produce a wider range of possible interactome sizes. The highest estimates obtained from HTP_{1000} (containing the 15 studies reporting more than 1,000 interactions) suggest a maximal yeast interactome size of around 140,000 pairwise interactions, provided that the FDR is negligible.

4 Discussion

Capture–recapture models have been used to find the relationship between FDR and interactome size. Firstly, we assessed how many interactions are sampled by the different techniques. This suggests that, under the assumption that different methodologies sample the same underlying true interactions, two-hybrid and affinity capture western techniques are the least error prone of the physical interaction assays. Secondly, a more detailed urn model was used to estimate the combination of false and

true interactions found in the complete (BioGRID) *S. cerevisiae* interaction data for physical interactions. These provide evidence that around half of the complete physical interactome is already in the BioGRID dataset. As more replicates are generated, these true interactions should start to stand out more from the background noise.

The models require substantial numbers of validated interactions in order to provide reliable estimates; this is already the case for the available *S. cerevisiae* interaction data. Our results suggest that the FDR rate for the physical data is at most ≈ 0.6, and, given an interactome size of 20,000–40,000, we estimate the FDR to be in the range of 0.32–0.47. The multiple urn model requires FDR estimates in order to allow us to predict the size of the whole interactome. However, the model can be used to assess published estimates for either FDR or interactome size, and check for inconsistencies. For instance, reported FDRs of 90% have been published for HTP data. Our analysis suggests strongly that this is not possible and that the FDR is certainly less than 0.8 even under the most extreme conditions.

Repeatedly reported interactions can be used to construct a reference set of PPIs, and our urn models allow us to assess this in more detail and recent HTP techniques present an opportunity for all the observable protein pairs to be tested in this manner. The assumption of uniform sampling of interactions is at least correct in a mean-field sense, and is increasingly supported by the technical set up of the larger experiments that contributed the majority of the PPI data. However, the role of systematic error in any of the experimental methods has been ignored. If the sampling is significantly skewed toward a particular subset of proteins, or particular interactions, then the overall interactome size estimates will be lower than in found here, as will the FDR estimates.

Differences in how the protein pairs have been sampled may make also affect comparisons of error rates between SSE and HTP studies. If the SSE size estimates are too low, then the relative FDR will increase for a given interactome size, further reducing the difference between the FDR seen for HTP and SSE data. Overall, the FDR is found to be up to 50% larger in the biggest HTP experiments compared to the smallest SSE.

The presented model also assumes that errors are stochastic in nature, rather than systematic. Systematic errors and bias give rise to spurious interactions, as they will almost certainly appear more often than stochastic errors. Ignoring this potential set of errors will increase the interactome size estimate, whilst reducing the FDR presented here. In order to take account of systematic errors from different techniques, the same multiple urn model should be reapplied to all the data from each technique. Given enough data, the amount of systematic error from each technique can then be assessed and the size of the interactome estimated more reliably.

From our analysis, it appears that over half a million sampled interactions may be required to classify all the reported interactions correctly. However, through simple replication, this should be possible without the requirement of having an underlying reference set as a simple statistical approach could separate out the noise and identify any systematic biases. Obviously, the number of repeated samples would be lower if the FDR could be reduced in experimental replicates. Then, if the scaling factor is appropriate, validations enable complete elucidation of the—approximately 73%, given the scaling factor—PPIs that are currently observable. Then further inference

methods using biological characteristics, or new experimental methods, can use this PPI reference set to fully elucidate the *S. cerevisiae* interactome.

Here, we have applied relatively straightforward ideas from sampling and probability theory in order to assess how complete interaction networks are. This, however, opens up are several further avenues to explore: First of all, we have assumed that there is some homogeneity in the way interactions are assayed. With high probability, the variability in the tendency of different proteins to generate false–positives will be considerable. Here, simple urn schemes, even more complicated ones than presented here (data not shown), prove too restrictive to cope with such variability. It would be possible, at least in principle, to use bio-informatics approaches to collate the extensive functional and structural and use mixture models (Marras et al. 2010) to perform more sophisticated analyses than the one presented here. This will come at considerable computational expense (Kelly and Stumpf 2010). Gaining robust insights into the architecture of such networks, their functional organization and evolution is, however, of increasing importance in biomedical and biotechnological problems. The biggest challenge, experimentally and theoretically, may arguably come from the intrinsically dynamic nature of molecular interaction networks (Kelly and Stumpf 2008; Lèbre et al. 2010): descriptions in terms of static (time-independent) graphs provide a poor description and more flexible and powerful descriptors are called for. In light of this, our estimates only refer to the total number of potential interactions, and not those that are realized at any given time point.

Appendix: The Effects of Experiment Size on PIN Estimates

This Appendix contains further information about the multiple capture–recapture methodology, and compare the simple urn model with one where experimental size is taken into account in order to predict the FDR and interactome size from a variety of datasets.

A.1 Uniqueness of Solution

In order to examine the possible uniqueness of i such that $g(m, i) = 0$ take m and i both as positive real number for convenience. The expectation found in (5) is (only in this section) approximated by

$$\begin{aligned}\mathbb{E}(\mathbb{S},\ \text{to find } i \text{ distinct interactions}) &= m\sum_{k=0}^{i-1}\frac{1}{m-k}\\ &= m\sum_{k=1}^{i-1}\frac{1}{m-k}+1\end{aligned}$$

$$\approx m \int_0^{i-1} \frac{1}{m-x}\, dx + 1$$
$$= m \log\left(\frac{m}{m-i+1}\right) + 1. \tag{13}$$

Now, in order to examine the uniqueness of a solution for the urn model, (9) are approximated using (13) as

$$S \approx m \log\left(\frac{m}{m-i+1}\right),$$
$$S' \approx m' \log\left(\frac{m'}{m'-i'+1}\right). \tag{14}$$

$g(m,i)$ now is

$$g(m,i) \approx s_{\text{obs}} - m \log\left(\frac{m}{m-i+1}\right) - m' \log\left(\frac{m'}{m'-i'+1}\right)$$
$$= s_{\text{obs}} - m \log\left(\frac{m}{m-i+1}\right)$$
$$- (m_{\text{obs}} - m) \log\left(\frac{m_{\text{obs}} - m}{(m_{\text{obs}} - m) - (i_{\text{obs}} - i) + 1}\right), \tag{15}$$

and the derivative of $g(m,i)$ with respect to i is

$$\frac{\partial g(m,i)}{\partial i} \approx -\frac{m}{m-i+1} + \frac{m_{\text{obs}} - m}{(m_{\text{obs}} - m) - (i_{\text{obs}} - i) + 1}. \tag{16}$$

This derivative is negative if

$$m\big((m_{\text{obs}} - m) - (i_{\text{obs}} - i) + 1\big) > (m_{\text{obs}} - m)(m - i + 1),$$

which reduces to

$$\frac{m_{\text{obs}}}{m} > \frac{i_{\text{obs}} - 2}{i - 1}. \tag{17}$$

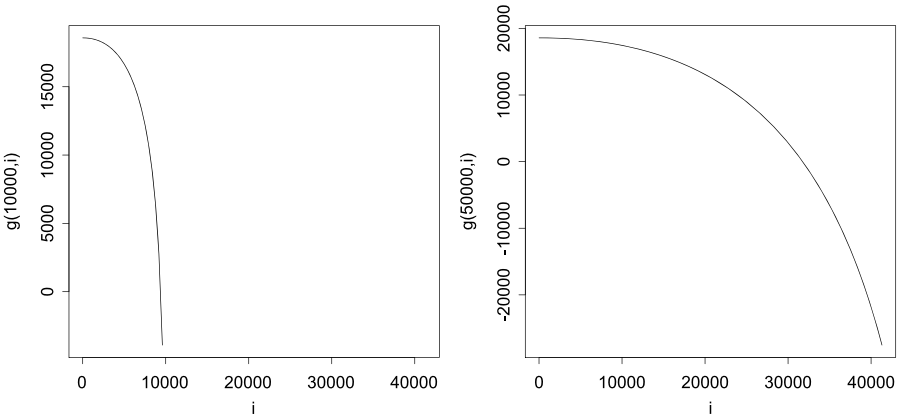
As protein interaction graphs are assumed to be sparse (i.e. $m \ll m_{\text{obs}}$), it follows that $\frac{\partial g(m,i)}{\partial i}$ can be positive only for small i. Figure 6 shows the behaviour of $g(m,i)$ for the physical data parameters taken from Table 1.

Using (10) and setting $i = 1$ for simplicity,

$$g(m,1) = s_{\text{obs}} - (m_{\text{obs}} - m) \log\left(\frac{m_{\text{obs}} - m}{(m_{\text{obs}} - m) - i_{\text{obs}}}\right), \tag{18}$$

which is positive for all parameter sets defined in Table 1 and $m \ll m_{\text{obs}}$.

Further, (16) is decreasing in i, so the second derivative of g with respect to i is negative. Therefore, as $g(m,1)$ for considered m is positive, if an i exists such that $g(m,i) = 0$ then the solution is unique.



(a) $g\,(10000, i)$ for $i \in [0, i_{obs}]$ (b) $g\,(50000, i)$ for $i \in [0, i_{obs}]$

Fig. 6 Single urn function. $g(m, i)$ for the physical interactome data parameters, $s_{\text{obs}} = 59956$, $m_{\text{obs}} = \binom{4967}{2}$ and $i_{\text{obs}} = 41313$. For $m \in \{10000, 50000\}$ the function can be seen to have a single solution satisfying $g(m, i) = 0$

A.2 Multiple urns

Rather than in a series of independent studies reporting individual interactions, the *S. cerevisiae* data have been published in studies producing multiple interactions. Each study, P_k, contains a set of reported interactions E_{P_k}. A *multiple urn* model has also been used, which assumes that interactions are drawn without replacement from the observable protein pairs, E_{obs}. This differs from the assumption in Sect. 2.3.2 where each interaction is drawn from E_{obs} with replacement.

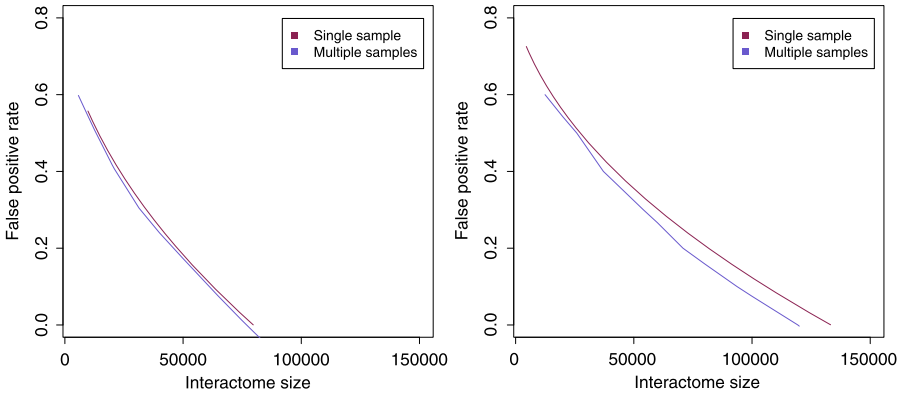
Recall that the number of true interactions, the interactome size, is m. Now suppose that q experiments, $P_1, \ldots, P_q$, are conducted and that the number of *true* interactions reported in experiment P_k is r_k. For each experiment, P_k, let $p_{h,j,k}$ be the probability of drawing $(j - h)$ novel true interactions, given that h distinct true interactions are observed in experiments $\{P_1, \ldots, P_{k-1}\}$. The probability $p_{h,j,k}$ can be described as a transition matrix (each state referring to the number of distinct interactions sampled) where for the kth experiment,

$$p_{h,j,k} = \begin{cases} 0 & \text{if } j < h, \\ \dfrac{\binom{m-h}{j-h}\binom{h}{r_k-j+h}}{\binom{m}{r_k}} & \text{if } j \geq h, \end{cases} \tag{19}$$

which is equivalent to

$$p_{h,j,k} = \left(\frac{(m-h)!h!r_k!(m-r_k)!}{(j-h)!(m-j)!(r_k-j+h)!(j-r_k)!m!} \right) \quad \text{if } j \geq h. \tag{20}$$

To find possible values of κ and m that are consistent with the data found in Table 1, different values of m, s, and i are simulated. Unlike the single urn model, however, the experiments provide s_{obs} samples and in each experiment the reported



(a) $SSE_{100} = \{P_k : r_{k,obs} < 100\}$ (b) $HTP_{100} = \{P_k : r_{k,obs} \geq 100\}$

Fig. 7 Single or multiple urns. Plots show the differences between the interactome size, ρm, and FDR, κ, estimates found using the single and multiple urn models for: (**a**) SSE_{100} and (**b**) HTP_{100} data. The single urn results are shown in *red* and the multiple urn results are shown in *blue*. The effect on SSE data is small between the models, whilst there is a larger difference to the predictions made when considering the HTP experiments

interactions have to be split into true ($e \in E$) and false ($e \in E'$) reported interactions. The complete experiment sizes $\{r_{1,\text{obs}}, r_{2,\text{obs}}, \ldots, r_{q,\text{obs}}\}$ are such that

$$s_{\text{obs}} = \sum_{k=1}^{q} r_{k,\text{obs}}. \tag{21}$$

In order to simulate this model, $\kappa \in \{\frac{1}{s_{\text{obs}}}, \ldots, \frac{s_{\text{obs}}-1}{s_{\text{obs}}}\}$ is chosen, and then the number of interactions drawn from the urns of true interactions and false interactions are uniformly, and at random, selected such that $\{r_1, r_2, \ldots, r_q\}$ are sampled from the interaction urn (E) and $\{r'_1, r'_2, \ldots, r'_q\}$ are sampled from the false interaction urn (E'), such that $r_k + r'_k = r_{k,\text{obs}}\ \forall k \in [1, q]$ and $\sum_{k=1}^{q} r_k = (1-\kappa)s_{\text{obs}}$.

For each possible κ (along with a collection of 1,000 sampled experiment sizes) and each $m \in [0, m_{\text{obs}}]$ the average number of distinct interactions, $\breve{i}$, is found through simulation and forms a possible solution for m and κ only if $\breve{i} = i_{\text{obs}}$. The multiple urn model is simulated in order to assess the effect of sampling from experiments of different sizes, in contrast to the simple with the replacement model in Sect. 2.3.2.

This model is used as a comparison to the simple model on predictions found from high-throughput (HTP) and small-scale experimental (SSE) data. The single urn model, found in the main text, ignores the effect of experiment size. This will have a more profound effect on the HTP results, as the single urn model provides a better description of data produced by smaller experiments. The multiple urn model described above takes explicit account of experiment sizes, $\{r_{1,\text{obs}}, r_{2,\text{obs}}, \ldots, r_{q,\text{obs}}\}$, (at the expense of simplicity) which is now used to estimate the size and FDR for the same datasets.

Figure 7 shows the difference between the predicted FDR and size values for the multiple and single urn models. The figures show the minimal changes on the solu-

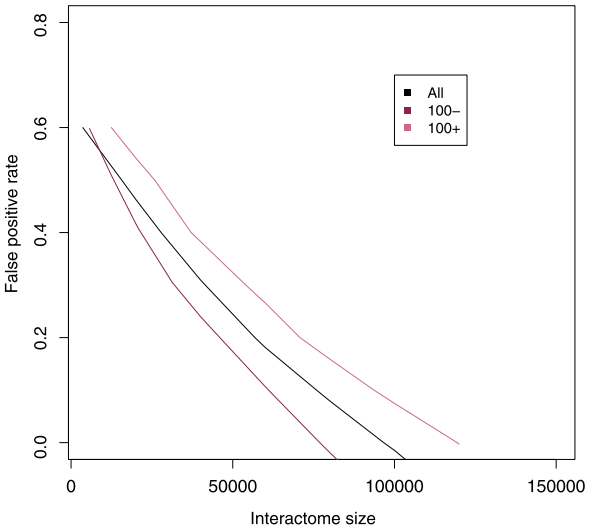


Fig. 8 Multiple urn interactome size results. Plot shows the FDR and size results for multiple urn model. The results are shown for three datasets: complete physical data; SSE_{100}; and HTP_{100}. This shows the differences between the data, with the maximal size, ρm, being over 50% more for the HTP_{100} data in contrast with the SSE_{100} data

tions when only smaller experiments are considered (in this example SSE_{100}), whilst the effect of using the HTP_{100} data is more pronounced.

Figure 8 shows HTP_{100} and SSE_{100} results from the multiple urn model, along with the results for the full physical data shown in black. The complete data results only show minor differences to the estimated FDR and sizes found for the single urn model (shown in Fig. 4). The maximal interactome size is about 50% larger for the HTP_{100} set than found for the SSE_{100}. For interactome size estimates from recent publications of 20,000–40,000 the FDR estimates for each dataset are: 0.31–0.46 (all); 0.38-0.54 (HTP_{100}); and 0.24-0.42 (SSE_{100}). The lower FDR estimates relate to a higher estimated interactome size.

References

Alm, E., & Arkin, A. (2003). Biological networks. *Curr. Opin. Struct. Biol.*, *13*(2), 193–202.

Bader, J. S., Chaudhuri, A., Rothberg, J., & Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.*, *22*(1), 78–85.

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guénoche, A., & Jacq, B. (2003). Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network. *Genome Biol.*, *5*(1), R6.

Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: A review. *J. Am. Stat. Assoc.*, *88*(421), 364–373.

Burnham, K. P., & Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, *65*(3), 625–633.

Chao, A. (2001). An overview of closed capture–recapture models. *J. Agric. Biol. Environ. Stat.*, *6*(2), 158–175.

Chiang, T., Scholtens, D., Sarkar, D., & Gentleman, R. (2007). Coverage and error models of protein–protein interaction data by directed graph analysis. *Genome Biol.*, *8*, R186.

de Silva, E., & Stumpf, M. P. H. (2005). Complex networks and simple models in biology. *J. R. Soc. Interface*, *2*(5), 419–430.

de Silva, E., Thorne, T., Ingram, P. J., Agrafioti, I., Swire, J., Wiuf, C., & Stumpf, M. P. H. (2006). The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol.*, *4*(39), 39.

D'haeseleer, P., & Church, G. (2004). Estimating and improving protein interaction error rates. In *Proceedings of the IEEE computational systems bioinformatics conference*.

Drees, B. L., Thorsson, V., Carter, G. W., Rives, A. W., Raymond, M. Z., Avila-Campillo, I., Shannon, P., & Galitski, T. (2005). Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol.*, *6*(4), R38.

Gentleman, R., & Huber, W. (2007). Making the most of high-throughput protein-interaction data. *Genome Biol.*, *8*(10), 112.

Grigoriev, A. (2003). On the number of protein–protein interactions in the yeast proteome. *Nucleic Acids Res.*, *31*(14), 4157–4161.

Hart, G. T., Ramani, A. K., & Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol.*, *7*(11), 120.

Heo, M., Maslov, S., & Shakhnovich, E. (2011). Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc. Natl. Acad. Sci.*, *108*(10), 4258–4263.

Hirschman, J. E., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hong, E. L., Livstone, M. S., Nash, R., Park, J., Oughtred, R., Skrzypek, M., Starr, B., Theesfeld, C. L., Williams, J., Andrada, R., Binkley, G., Dong, Q., Lane, C., Miyasato, S., Sethuraman, A., Schroeder, M., Thanawala, M. K., Weng, S., Dolinski, K., Botstein, D., & Cherry, J. M. (2006). Genome snapshot: a new resource at the saccharomyces genome database (sgd) presenting an overview of the saccharomyces cerevisiae genome. *Nucleic Acids Res.*, *34*(Database issue), D442–D445.

Huang, H., Jedynak, B. M., & Bader, J. S. (2007). Where have all the interactions gone? estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput. Biol.*, *3*(11), e214.

Kelly, W. P., & Stumpf, M. P. H. (2008). Protein–protein interactions: from global to local analyses. *Curr. Opin. Biotechnol.*, *19*, 396–403.

Kelly, W. P., & Stumpf, M. P. H. (2010). Trees on networks: resolving statistical patterns of phylogenetic similarities among interacting proteins. *BMC Bioinform.*, *11*, 470.

Lèbre, S., Becq, J., Devaux, F., Stumpf, M. P. H., & Lelandais, G. (2010). Statistical inference of the time-varying structure of gene-regulation networks. *BMC Syst. Biol.*, *4*, 130.

Marras, E., Travaglione, A., & Capobianco, E. (2010). Sub-modular resolution analysis by network mixture models. *Stat. Appl. Genet. Mol. Biol.*, *9*(1), 19.

Schlitt, T., & Brazma, A. (2005). Modelling gene networks at different organisational levels. *FEBS Lett.*, *579*, 1859–1866.

Shokouhi, M., Zobel, J., & Scholer, F. (2006). Capturing collection size for distributed non-cooperative retrieval. In *SIGIR proceedings* (pp. 316–323).

Stumpf, M. P. H., Wiuf, C., & May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl. Acad. Sci.*, *102*(12), 4221–4224.

Stumpf, M. P. H., Thorne, T., de Silva, E., Stewart, R., An, H., Lappe, M., & Wiuf, C. (2008). Estimating the size of the human interactome. *Proc. Natl. Acad. Sci.*, *105*(19), 6959–6964.

Thorne, T. W., Ho, H.-L., Huvet, M., Haynes, K., & Stumpf, M. P. H. (2011). Prediction of putative protein interactions through evolutionary analysis of osmotic stress response in the model yeast *Saccharomyces cerevisae*. *Fungal Genet. Biol.*, *48*, 504–511.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., & Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, *417*(6887), 399–403.

Xu, J., Wu, S., & Li, X. (2007). Estimating collection size with logistic regression. In *SIGIR proceedings* (pp. 789–790).

Yang, L., Vondriska, T. M., Han, Z., MacLellan, W. R., Weiss, J. N., & Qu, Z. (2008). Deducing topology of protein–protein interaction networks from experimentally measured sub-networks. *BMC Bioinform.*, *9*, 301.