nature COMMUNICATIONS

ARTICLE

Received 1 Nov 2012 | Accepted 4 Jul 2013 | Published 6 Aug 2013

DOI: 10.1038/ncomms3241

OPEN

Counting motifs in the human interactome

Ngoc Hieu Tran[1], Kwok Pui Choi[1,2] & Louxin Zhang[2,3]

Small over-represented motifs in biological networks often form essential functional units of biological processes. A natural question is to gauge whether a motif occurs abundantly or rarely in a biological network. Here we develop an accurate method to estimate the occurrences of a motif in the entire network from noisy and incomplete data, and apply it to eukaryotic interactomes and cell-specific transcription factor regulatory networks. The number of triangles in the human interactome is about 194 times that in the *Saccharomyces cerevisiae* interactome. A strong positive linear correlation exists between the numbers of occurrences of triad and quadriad motifs in human cell-specific transcription factor regulatory networks. Our findings show that the proposed method is general and powerful for counting motifs and can be applied to any network regardless of its topological structure.

[1] Department of Statistics and Applied Probability, National University of Singapore (NUS), Singapore 117546, Singapore. [2] Department of Mathematics, National University of Singapore (NUS), Singapore 119076, Singapore. [3] NUS Graduate School for Integrative Sciences and Engineering, Singapore 117456, Singapore. Correspondence and requests for materials should be addressed to L.X.Z. (email: matzlx@nus.edu.sg).

The increasing availability of genomic and proteomic data has propelled network biology to the frontier of biomedical research[1–4]. Network biology uses a graph to depict interactions between cellular components (proteins, genes, metabolites and so on), where the nodes are cellular components and the links represent interactions. Two of the most surprising discoveries from the genome sequencing projects are that the human gene repertoire is much smaller than had been expected, and that there are just over 200 genes unique to human beings[5]. As the number of genes alone does not fully characterize the biological complexity of living organisms, the scale of physiologically relevant protein and gene interactions are now being investigated to understand the basic biological principles of life[6–8]. Although the list of known protein–protein interactions (PPIs) and gene regulatory interactions (GRIs) is expanding at an ever-increasing pace, the human PPI and GRI networks are far from being complete and, hence, their dynamics have yet to be uncovered[9–11].

The feed-forward loop (FFL) and several other graphlets (called motifs) are found to be over-represented in different biological networks[11]. Furthermore, over-represented motifs usually represent functional units of biological processes in cells. Hence, it is natural to ask whether a motif, such as a triangle, appears more often in the interactome of humans than in that of other species, or whether the FFL or the bi-fan appears more frequently in the human gene regulatory network. As the biological networks that have been reported are actually the subnetworks of the true ones and often contain remarkably many incorrect interactions for eukaryotic species, there are two approaches to answering these questions. One approach is to infer spurious and missing links in the entire network[12–14], and then to count motif occurrences. Another approach is to estimate the number of motif occurrences in the interactome from the observed subnetwork data using the same method as that for estimating the size of eukaryotic interactomes[9,10,15]. If we have the number of occurrences of a motif or its estimate in a network, we can determine whether the motif is over-represented or not, based on how often the motif is seen in a random network with similar structural parameters[11,16,17].

In the present work, we take spurious and missing link errors into account to develop an unbiased and consistent estimator for the motif count. The method works for both undirected and directed networks. We derive explicit mathematical expressions for the estimators of five commonly studied triad and quadriad network motifs (Fig. 1). These estimators are further validated extensively for each of the following four models: Erdös–Renyi (ER)[18], preferential attachment[19], duplication[20] and the geometric model[21] (Supplementary Note 1). By applying the method to eukaryotic interactomes, we find that the number of triangles in the human interactome is about 194 times that of the *Saccharomyces cerevisiae* interactome, three times as large as expected. By applying the method to human cell-specific transcription factor (TF) regulatory networks[22], we discover a strong positive linear correlation between the counts of widely studied triads and quadriads. We also notice that the embryonic stem cell's TF regulatory network has the smallest number of occurrences relative to its network size for all the five motifs under study.

Results

Estimating motif occurrences. In this study, we shall consider PPIs and gene regulatory networks. The former are undirected, whereas the latter are directed networks. Consider a directed or undirected network $\mathcal{G}(V, E)$, where V is the set of nodes and E is the set of links. For simplicity, we assume that $\mathcal{G}$ has n nodes and $V = \{1,2,3,\ldots,n\}$. Let $\mathcal{G}^{\text{obs}}(V^{\text{obs}}, E^{\text{obs}})$ be an observed subnetwork of $\mathcal{G}$. Following (ref. 9), we model an observed subnetwork as the outcome of a uniform node sampling process in the following sense. Let X_i be independent and identically distributed Bernoulli random variables with the parameter $p \in (0,1]$ for $i = 1,2,\ldots,n$. We use $X_i = 1$ and $X_i = 0$ to denote the events that node i is sampled and not sampled, respectively. Then V^{obs} is the set of nodes i with $X_i = 1$, and E^{obs} is induced from E by V^{obs}. For clarity of presentation, we first introduce our method for the case when the observed subnetwork is free from experimental errors, and then generalize it to handle noisy observed subnetwork data.

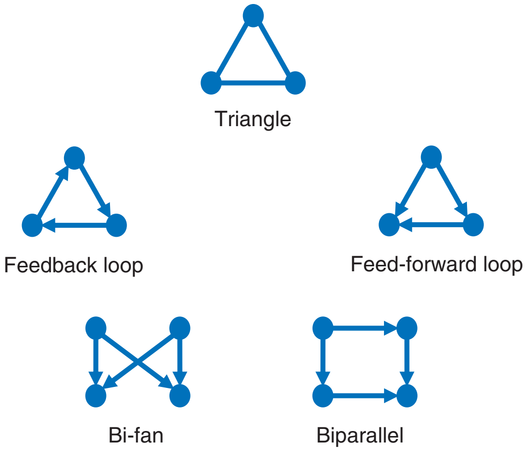


Figure 1 | Network motifs found in biological networks. The feed-forward loop, bi-fan and biparallel are over-represented, whereas feedback loop is under-represented in gene regulatory networks and neuronal connectivity networks[11].

Consider a motif $\mathcal{M}$. We use $N_{\mathcal{M}}$ and $N_{\mathcal{M}}^{\text{obs}}$ to denote the number of occurrences of $\mathcal{M}$ in $\mathcal{G}$ and $\mathcal{G}^{\text{obs}}$, respectively. We assume that the number of nodes, n, is known, but only links in G^{obs} are known. We are interested in estimating $N_{\mathcal{M}}$ from the observed subnetwork $\mathcal{G}^{\text{obs}}$. As $\mathcal{G}^{\text{obs}}$ is assumed to be free from experimental errors, we can obtain $N_{\mathcal{M}}^{\text{obs}}$ simply by enumeration. Let us define the following:

$$\widehat{N}_{\mathcal{M}} = \frac{\binom{n}{m}}{\binom{n^{\text{obs}}}{m}} N_{\mathcal{M}}^{\text{obs}}, \tag{1}$$

where m and n^{obs} are the number of nodes in $\mathcal{M}$ and $\mathcal{G}^{\text{obs}}$, respectively.

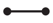
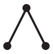
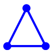
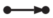
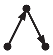
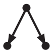
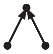
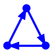
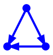
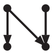
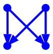
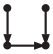
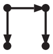
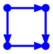
Let $\mathbf{A} = [a_{ij}]_{1 \le i,\, j \le n}$ denote the adjacency matrix of $\mathcal{G}$, where $a_{ij} = 1$ if there is a link from i to j, and $a_{ij} = 0$ otherwise. Furthermore, for a subset $J \subseteq \{1,2,\ldots,n\}$, $\mathbf{A}[J]$ denotes the submatrix consisting of entries in the rows and columns indexed by J. We write $N_{\mathcal{M}}$ as a function of $\mathbf{A}$ and $N_{\mathcal{M}}^{\text{obs}}$ as a function of $\mathbf{A}$ and the random variables X_i. We also assume the following:

$$N_{\mathcal{M}} = \sum_{i_1 < i_2 < \cdots < i_m} f_{\mathcal{M}}(\mathbf{A}[i_1, i_2, \ldots, i_m]), \tag{2}$$

$$N_{\mathcal{M}}^{\text{obs}} = \sum_{i_1 < i_2 < \cdots < i_m} f_{\mathcal{M}}(\mathbf{A}[i_1, i_2, \ldots, i_m]) X_{i_1} X_{i_2} \ldots X_{i_m}, \tag{3}$$

where $f_{\mathcal{M}}()$ is a function chosen to decide whether $\mathcal{M}$ occurs among nodes $i_1, i_2, \ldots, i_m$ or not. For the motifs listed in Table 1, their corresponding functions $f_{\mathcal{M}}()$ are given in Supplementary Table S1.

Table 1 | Bias-corrected estimators for 14 motifs.

	Motif	Bias-corrected estimator
1		$\widetilde{N}_1 = \frac{1}{r}\left[\widehat{N}_1 - \binom{n}{2} r_+\right]$
2		$\widetilde{N}_2 = \frac{1}{r^2}\left[\widehat{N}_2 - 2(n-2)r_+ r\widetilde{N}_1 - 3\binom{n}{3} r_+^2\right]$
3		$\widetilde{N}_3 = \frac{1}{r^3}\left[\widehat{N}_3 - r_+ r^2\widetilde{N}_2 - (n-2)r_+^2 r\widetilde{N}_1 - \binom{n}{3} r_+^3\right]$
4		$\widetilde{N}_4 = \frac{1}{r}\left[\widehat{N}_4 - 2\binom{n}{2} r_+\right]$
5		$\widetilde{N}_5 = \frac{1}{r^2}\left[\widehat{N}_5 - 2(n-2)r_+ r\widetilde{N}_4 - 6\binom{n}{3} r_+^2\right]$
6		$\widetilde{N}_6 = \frac{1}{r^2}\left[\widehat{N}_6 - (n-2)r_+ r\widetilde{N}_4 - 3\binom{n}{3} r_+^2\right]$
7		$\widetilde{N}_7 = \frac{1}{r^2}\left[\widehat{N}_7 - (n-2)r_+ r\widetilde{N}_4 - 3\binom{n}{3} r_+^2\right]$
8		$\widetilde{N}_8 = \frac{1}{r^3}\left[\widehat{N}_8 - r_+ r^2\widetilde{N}_5 - (n-2)r_+^2 r\widetilde{N}_4 - 2\binom{n}{3} r_+^3\right]$
9		$\widetilde{N}_9 = \frac{1}{r^3}\left[\widehat{N}_9 - r_+ r^2(\widetilde{N}_5 + \widetilde{2\mathrm{N}}_6 + 2\widetilde{N}_7) - 3(n-2)r_+^2 r\widetilde{N}_4 - 6\binom{n}{3} r_+^3\right]$
10		$\widetilde{N}_{10} = \frac{1}{r^3}\left\{\widehat{N}_{10} - 2r_+ r^2\left[\binom{\widetilde{N}_4}{2} + (n-3)(\widetilde{N}_6 + \widetilde{N}_7)\right] - 6\binom{n-2}{2} r_+^2 r\widetilde{N}_4 - 24\binom{n}{2} r_+^3\right\}$
11		$\widetilde{N}_{11} = \frac{1}{r^4}\left\{\widehat{N}_{11} - r_+ r^3\widetilde{N}_{10} - r_+^2 r^2\left[\binom{\widetilde{N}_4}{2} + (n-3)(\widetilde{N}_6 + \widetilde{N}_7)\right] - 2\binom{n-2}{2} r_+^3 r\widetilde{N}_4 - 6\binom{n}{4} r_+^4\right\}$
12		$\widetilde{N}_{12} = \frac{1}{r^3}\left\{\widehat{N}_{12} - r_+ r^2\left[2\binom{\widetilde{N}_4}{2} + (n-3)(\widetilde{N}_5 + 2\widetilde{N}_7)\right] - 6\binom{n-2}{2} r_+^2 r\widetilde{N}_4 - 24\binom{n}{4} r_+^3\right\}$
13		$\widetilde{N}_{13} = \frac{1}{r^3}\left\{\widehat{N}_{13} - r_+ r^2\left[2\binom{\widetilde{N}_4}{2} + (n-3)(\widetilde{N}_5 + 2\widetilde{N}_6)\right] - 6\binom{n-2}{2} r_+^2 r\widetilde{N}_4 - 24\binom{n}{4} r_+^3\right\}$
14		$\widetilde{N}_{14} = \frac{1}{r^4}\left\{\widehat{N}_{14} - r_+ r^3(\widetilde{N}_{12} + \widetilde{N}_{13}) - r_+^2 r^2\left[2\binom{\widetilde{N}_4}{2} + (n-3)(\widetilde{N}_5 + \widetilde{N}_6 + \widetilde{N}_7)\right] - 4\binom{n-2}{2} r_+^3 r\widetilde{N}_4 - 12\binom{n}{4} r_+^4\right\}$

n (n^{obs}), the number of nodes in the entire network (respectively, the observed subnetwork).
m_i, the number of nodes in motifs of type-i.
N_i^{obs}, the number of occurrences of motifs of type-i observed in the subnetwork data.
$r = 1 - r_- - r_+$.
$\widehat{N}_i = \binom{n}{m_i} N_i^{\mathrm{obs}} / \binom{n^{\mathrm{obs}}}{m_i}$, $1 \leqslant i \leqslant 14$.

From equations (1) and (3), we have

$$E(\widehat{N}_{\mathcal{M}}) = \binom{n}{m} \sum_{1 \le i_1 < i_2 < \ldots < i_m \le n} f_{\mathcal{M}}(\mathbf{A}[i_1, i_2, \ldots, i_m]) \times E\left(\frac{X_{i_1} X_{i_2} \cdots X_{i_m}}{\binom{n^{\mathrm{obs}}}{m}}\right),$$

where n^{obs} is a random variable such that

$$n^{\mathrm{obs}} = X_1 + X_2 + \cdots + X_n. \tag{4}$$

As the random variables X_i are independent and identically distributed, for any $1 \le i_1 < i_2 < \ldots < i_m \le n$, we also have

$$E\left(\frac{X_{i_1} X_{i_2} \cdots X_{i_m}}{\binom{n^{\mathrm{obs}}}{m}}\right) = E\left(\frac{X_1 X_2 \cdots X_m}{\binom{n^{\mathrm{obs}}}{m}}\right).$$

Hence, by equation (2),

$$E(\widehat{N}_{\mathcal{M}}) = \binom{n}{m} N_{\mathcal{M}} E\left(\frac{X_1 X_2 \cdots X_m}{\binom{n^{\mathrm{obs}}}{m}}\right)$$

$$= n(n-1) \cdots (n-m+1) N_{\mathcal{M}}$$

$$\times E\left(\frac{X_1 X_2 \cdots X_m}{n^{\mathrm{obs}}(n^{\mathrm{obs}} - 1) \cdots (n^{\mathrm{obs}} - m + 1)}\right).$$

By conditioning on the event that $X_1 = X_2 = \cdots = X_m = 1$, we rewrite equation (4) as

$$n^{\mathrm{obs}} = Z + m,$$

where $Z \sim \text{Binomial}(n-m,p)$, and hence

$$E\left(\frac{\widehat{N}_{\mathcal{M}}}{N_{\mathcal{M}}}\right)=n(n-1)\cdots(n-m+1)p^{m}$$

$$\times E\left(\frac{1}{(Z+m)(Z+m-1)\cdots(Z+1)}\right).$$

As

$$E\left(\frac{1}{(Z+m)(Z+m-1)\cdots(Z+1)}\right)$$

$$=E\left(\int_0^1 \frac{(1-u)^{m-1}}{(m-1)!}u^{Z}du\right)$$

$$=\int_0^1 \frac{(1-u)^{m-1}}{(m-1)!}E(u^{Z})du,$$

we have

$$E\left(\frac{\widehat{N}_{M}}{N_{\mathcal{M}}}\right)=1-\sum_{j=0}^{m-1}\binom{n}{j}p^{j}q^{n-j} \qquad (5)$$

by applying integration by parts and simplification. Therefore, we have obtained the following theorem.

Theorem 1: Let $\mathcal{G}$ be a network of n nodes. Assume $\mathcal{G}^{\text{obs}}$ is a subnetwork of $\mathcal{G}$ obtained by a uniform node sampling process that selects a node with probability p. For any motif $\mathcal{M}$ of m nodes, the estimator $\widehat{N}_{\mathcal{M}}$ defined in equation (1) satisfies equation (5). Therefore, $\widehat{N}_{\mathcal{M}}$ is an asymptotically unbiased estimator for $N_{\mathcal{M}}$ in the sense that $E(\widehat{N}_{\mathcal{M}}/N_{\mathcal{M}}) \rightarrow 1$ as n goes to infinity. Moreover, the convergence is exponentially fast in n.

When the estimator (1) is applied to estimate the number of links in an undirected network $\mathcal{G}$, the variance has the following closed-form expression:

$$\text{Var}\left(\frac{\widehat{N}_1}{N_1}\right)=\left(\frac{2qN_2}{pN_1^2}+\frac{1-p^2}{p^2N_1}\right)(1+O(n^{-1}))+O(n^{-1}),$$

where N_1 and N_2 are, respectively, the number of links and three-node paths in $\mathcal{G}$ (Supplementary Methods). This leads to our next theorem.

Theorem 2: When $\mathcal{G}$ is generated from one of the ER, preferential attachment, duplication or geometric models, $\text{Var}(\widehat{N}_1/N_1) \rightarrow 0$ as n goes to infinity.

Theorem 2 says that $\widehat{N}_1$ is consistent. For an arbitrary motif, it is much more complicated to derive the variance of the estimator (1). Nevertheless, our simulation shows that for all the motifs in Fig. 1, the variance of the estimator converges to zero as n goes to infinity and, hence, it is consistent (Fig. 2 and Supplementary Figs S1–S8). We wish to point out that the notions 'asymptotically unbiased' and 'consistent' are not used in the usual statistical sense where the population is fixed and the number of observations increases to infinity.

For realistic estimation, one has to take error rates into account, as detecting PPIs or GRIs is error prone to some degree. PPIs or gene regulatory networks have spurious interactions (that is, false positives) and missing interactions (that is, false negatives). We define the false-positive rate r_+ to be the probability that a non-existing link is incorrectly reported, and the false-negative rate r_- to be the probability that a link is not observed. Using the independent random variables $F^{+}_{i_1i_2} \sim \text{Bernoulli}(r_+)$ and $F^{-}_{i_1i_2} \sim \text{Bernoulli}(r_-)$ to model spurious and missing interactions in the observed subnetwork $\mathcal{G}^{\text{obs}}$, we can represent the effect of experimental errors on an ordered pair of nodes (i_1,i_2) as

$$\widetilde{a}_{i_1i_2}=a_{i_1i_2}(1-F^{-}_{i_1i_2})+(1-a_{i_1i_2})F^{+}_{i_1i_2}. \qquad (6)$$

In other words, for any two nodes $i_1,i_2 \in V^{\text{obs}}$, a link (i_1,i_2) is observed in the subnetwork $\mathcal{G}^{\text{obs}}$ (that is, $\widetilde{a}_{i_1i_2}=1$) if (i) there is a link (i_1,i_2) in the real network $\mathcal{G}$ (that is, $a_{i_1,i_2}=1$) and there is no false negative (that is, $F^{-}_{i_1i_2}=0$) or (ii) the link (i_1,i_2) does not exist in the real network $\mathcal{G}$ (that is, $a_{i_1,i_2}=0$) but a false positive occurs (that is, $F^{+}_{i_1i_2}=1$).

To take error rates into account, we simply replace each entry $a_{i1,i2}$ in the adjacency matrix $\mathbf{A}$ with $\widetilde{a}_{i_1i_2}$ to obtain a new matrix, $\widetilde{\mathbf{A}}$, and then replace $\mathbf{A}$ with $\widetilde{\mathbf{A}}$ in equation (3). For any motif $\mathcal{M}$ in Table 1, the expectation of the estimator $\widehat{N}_{\mathcal{M}}$ in equation (1) can be expressed as (Supplementary Methods)

$$E(\widehat{N}_{\mathcal{M}})=\left(1-\sum_{j=0}^{m-1}\binom{n}{j}p^{j}q^{n-j}\right)\times[(1-r_+-r_-)^{s}N_{\mathcal{M}}+W_{\mathcal{M}}],$$

where s is the number of links that $\mathcal{M}$ has and $W_{\mathcal{M}}$ is a function of n, r_-, r_+, and $N_{\mathcal{M}'}$ for all proper submotifs $\mathcal{M}'$ of $\mathcal{M}$ (Supplementary Table S2). Thus, to correct the bias caused by link errors, we derive $\widetilde{W}_{\mathcal{M}}$ from $W_{\mathcal{M}}$ by replacing $N_{\mathcal{M}'}$ with $\widetilde{N}_{\mathcal{M}'}$ for all submotifs of $\mathcal{M}$, and use the following formula to estimate $N_{\mathcal{M}}$:

$$\widetilde{N}_{\mathcal{M}}=\frac{1}{(1-r_+-r_-)^{s}}(\widehat{N}_{\mathcal{M}}-\widetilde{W}_{\mathcal{M}}). \qquad (7)$$

For the motifs listed in Fig. 1, the corresponding bias-corrected estimators are given in Table 1.

We examined the accuracy of the proposed estimators on networks generated by a random network model. As these estimators are asymptotically unbiased, we used the mean square error (MSE) of the ratios $\widehat{N}_{\mathcal{M}}/N_{\mathcal{M}}$ and $\widetilde{N}_{\mathcal{M}}/N_{\mathcal{M}}$, defined later in

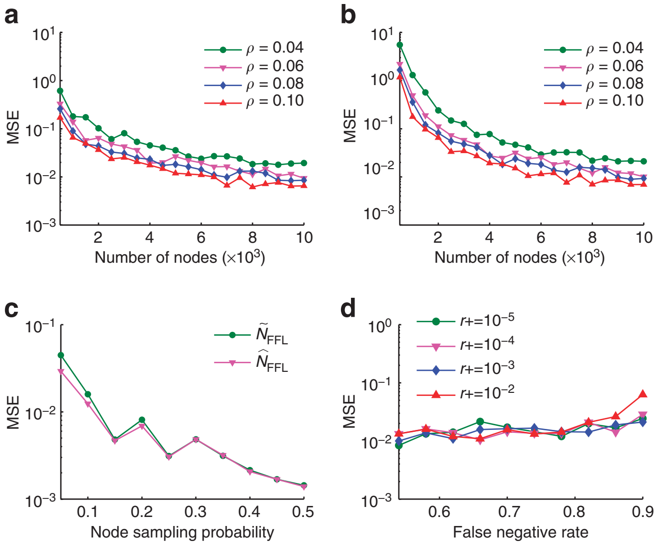


Figure 2 | Plots of MSE($\widehat{N}_{\text{FFL}}$) and MSE($\widetilde{N}_{\text{FFL}}$) for counting the occurrences of FFL. The random networks of n nodes and edge density ρ are generated from the preferential attachment model. Both MSE($\widehat{N}_{\text{FFL}}$) and MSE($\widetilde{N}_{\text{FFL}}$) depend on n, ρ and the node sampling probability p. MSE($\widetilde{N}_{\text{FFL}}$) also depends on the link error rates r_- and r_+. (**a**) MSE($\widehat{N}_{\text{FFL}}$) changes with n and ρ when $p=0.1$. (**b**) MSE($\widetilde{N}_{\text{FFL}}$) changes with n and ρ when $p=0.1$, $r_-=0.85$ and $r_+=0.00001$. (**c**) MSE($\widehat{N}_{\text{FFL}}$) and MSE($\widetilde{N}_{\text{FFL}}$) change with p when $n=5{,}000$, $\rho=0.1$, $r_-=0.85$ and $r_+=0.00001$. (**d**) MSE($\widetilde{N}_{\text{FFL}}$) changes with r_+ and r_- when $n=5{,}000$, $\rho=0.1$ and $p=0.1$.

equation (9), to measure their accuracy (see Methods section). Figure 2 summarizes the simulation results for the FFL motif in random networks generated from the preferential attachment model[19]. (The results for other motif network model combinations are similar and can be found in Supplementary Figs S1–S8.) First, when the edge density ρ is fixed, the MSE of the estimators for FFL decreases and converges to zero as n increases (Fig. 2a,b). Second, the MSE decreases as the edge density increases, suggesting that the estimators are even more accurate when applied to less sparse networks. Third, the MSE of the estimators decreases as p increases (Fig. 2c). Finally, the MSE increases with r_- and r_+ (Fig. 2d). Altogether, our simulation tests confirm that the proposed estimators are accurate for any underlying network.

Motif richness in the human interactome. The entire interactomes for eukaryotic model organisms such as *S. cerevisiae*, *Caenorhabditis elegans*, *Homo sapiens* and *Arabidopsis thaliana* are not fully known. We estimated the interactome size (that is, the number of interactions) and the number of triangles in the entire PPI network for *S. cerevisiae*, *C. elegans*, *H. sapiens* and *A. thaliana*, using the data sets CCSB-YI1 (ref. 23), CCSB-WI-2007 (ref. 24), CCSB-HI1 (refs 25,26) and CCSB-AI1-Main[27]. These data sets were produced from yeast two-hybrid experiments and their quality parameters are summarized in Table 2 for convenience.

First, we re-estimated the size of four interactomes using the bias-corrected estimator $\widetilde{N}_1$ (Table 1). To test all possible interactions between selected proteins, the sets of bait and prey proteins should be exchanged in the two rounds of interaction mating in a high-throughput yeast two-hybrid experiment[28]. However, this is only true for the *C. elegans* and *H. sapiens* data sets (CCSB-WI-2007 and CCSB-HI1, respectively). For the *S. cerevisiae* and *A. thaliana* data sets (CCSB-YI1 and CCSB-AI1-Main, respectively), the set of bait proteins are slightly different from the set of prey proteins. For these two cases, we applied our estimator to the subnetwork induced by the intersection of the bait and prey protein sets.

The following estimator was proposed by Stumpf *et al.*[9] for the size of an interactome and was later used to estimate the size of the eukaryotic interactomes[23,24,26,27]:

$$\frac{(\text{No. of observed interactions})\times\text{Precision}}{\text{Completeness}\times\text{Sensitivity}}, \qquad (8)$$

where ‘completeness’ is the fraction of all possible pairwise protein combinations that have been tested. In our notation,
(No. of observed interactions) $= N_1^{\text{obs}}$,
Sensitivity $= 1 - r_-$,
Precision $= 1 - r_d$,
Completeness $= \binom{n^{\text{obs}}}{2} \Big/ \binom{n}{2}$, where r_d is the proportion of spurious links among detected links and is called the false discovery rate. (Note that r_d was called the false-positive rate in ref. 9.) Thus, the estimator (8) becomes

$$\frac{1}{1-r_-}\left(\frac{\binom{n}{2}}{\binom{n^{\text{obs}}}{2}}N_1^{\text{obs}} - \frac{\binom{n}{2}}{\binom{n^{\text{obs}}}{2}}r_d N_1^{\text{obs}}\right).$$

For PPI data sets, r_+ is about 10^{-4} and thus $1-r_- \approx 1-r_- -r_+$. As r_d is also small, our estimator $\widetilde{N}_1$ handles errors differently but is quite close to the estimator (8). In particular, when the precision is 100% or, equivalently, when $r_d = r_+ = 0$, these two estimators are equal (Supplementary Note 2 and Supplementary Fig. S9). Indeed, our estimates for interactome size agree well with those obtained from equation (8) (Table 2). Such an agreement demonstrates again that our estimators for counting motifs are accurate.

We proceed further to estimate the number of triangles in each of the interactomes using the corresponding bias-corrected estimator $\widetilde{N}_3$ in Table 1. For each interactome, we estimated the number of triangles from the observed subnetwork data directly and from sampling the observed subnetwork repeatedly. The two estimates agree well (Table 2).

Our estimation shows that although the size of the *A. thaliana* interactome is about 1.8 times that of the human interactome, it

Table 2 | The interactome size and the number of triangles in the PPI networks of four species in our study.

	S. cerevisiae	***C. elegans***	***H. sapiens***	***A. thaliana***
Total no. of proteins	6,000	20,065	22,500	27,029
No. of proteins screened*	3,676	9,906	7,194	7,108
No. of links detected*	967	1,816	2,754	4,890
*Quality parameters**				
Precision†	0.9400	0.8600	0.7940	0.8030
Sensitivity	0.1700	0.0496	0.0950	0.1570
False-negative rate (r_-)	0.8300	0.9504	0.9050	0.8430
False positive rate (r_+)	0.8×10^{-5}	0.5×10^{-5}	2×10^{-5}	3×10^{-5}
Interactome size				
CCSB estimate*	18,000 ± 4,500	116,000 ± 26,400	130,000 ± 32,600	299,000 ± 79,200
Our estimate‡	14,000	121,000	210,000	377,000
Mean ± s.d.§	15,000 ± 2,700	122,000 ± 16,600	214,000 ± 32,200	376,000 ± 45,600
Link density	8×10^{-4}	6×10^{-4}	8×10^{-4}	10×10^{-4}
No. of triangles				
Our estimate‡	53,000	6,263,000	10,270,000	10,697,000
Mean ± s.d.§	61,000 ± 33,800	5,971,000 ± 3,593,800	11,255,000 ± 4,717,100	10,158,000 ± 4,289,000
Triangle density	1×10^{-6}	5×10^{-6}	5×10^{-6}	3×10^{-6}

CCSB, Center for Cancer Systems Biology; PPI, protein–protein interaction.
*Reported in refs 23–27.
†False discovery rate = 1 − precision.
‡Estimates have been calculated from the observed PPI subnetworks.
§Mean and s.d. of the estimates have been calculated by sampling 100 sub-data sets from the observed subnetwork data using the node sampling probability 0.1.

contains fewer triangles than the human interactome does. The triangle density of the human and *C. elegans* interactomes are similar and are 1.7 times that of the *A. thaliana* and 5 times that of *S. cerevisiae*. The size of the human interactome is only 15 times that of the *S. cerevisiae* interactome, yet the number of triangles in the former is about 194 times that in the latter, 3 times as large as expected.

Correlation between motif counts in TF regulatory networks. Recently, the TF regulatory networks of 41 human cell and tissue types were obtained from genome-wide *in vivo* DNaseI footprints map[22]. In these networks, the nodes are 475 TFs and the regulation of each TF by another is represented by network-directed links. Motif count analysis showed that the distribution of the motif count is unimodal, with the peak corresponding to the mean value for each motif (diagonal panels in Fig. 3). Surprisingly, there is a very strong linear correlation between the counts in the TF regulatory networks of different cell types, even for the triad and quadriad motifs that are topologically very different (Fig. 3).

Given that human has about 2,886 TF proteins[29], we further estimated the number of occurrences of the 5 motifs for each of the 7 functionally related classes of cells (Fig. 3 and Table 3). This was achieved by simply setting the false-positive and -negative rates to 0, as they are currently unknown. The TF regulatory networks of blood cells have diverse motif counts. Specifically, for all triad and quadriad motifs, the promyelocytic leukemia cell TF regulatory network has the largest number of occurrences, whereas the erythoid cell TF regulatory network has the smallest

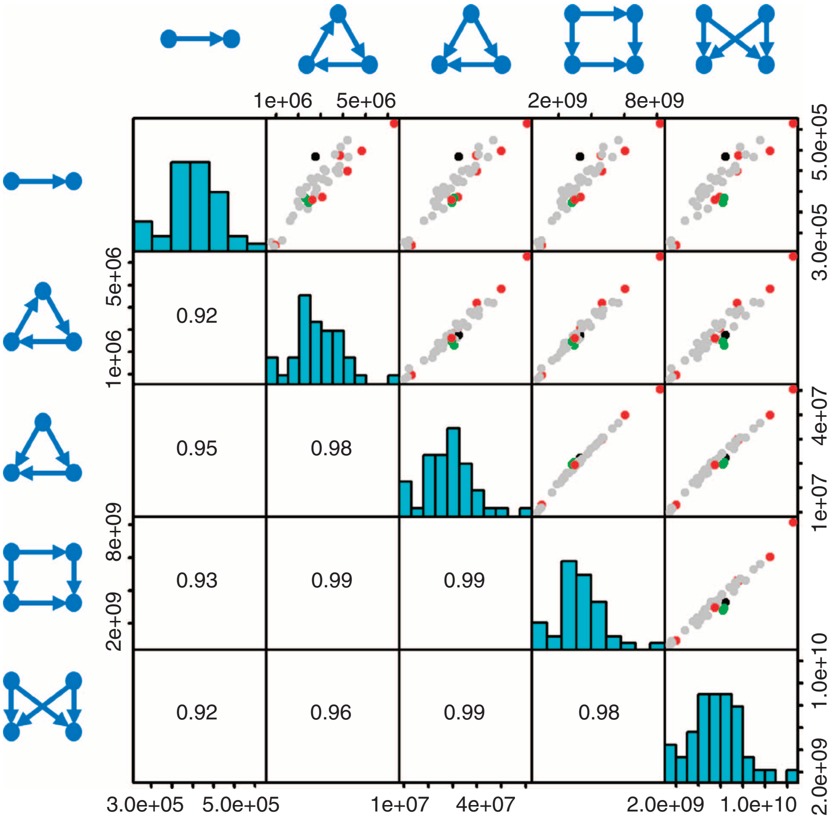


Figure 3 | Correlation of motif counts in 41 human cell-specific TF regulatory networks. The upper triangular panels are scatter plots of the counts of the 5 motifs in the TF regulatory networks of one embryonic stem cell (black), 7 blood cell types (red), 2 cancer cell types (green) and 31 other cell and tissues types (grey)[22]. Here the *x* and *y* axes represent the estimated counts of the two corresponding motifs. Each diagonal panel shows the distribution of these motifs' occurrences, in which the *x* and *y* axes represent the estimated motif count and the number of TF regulatory networks, respectively. The correlation coefficients of the motifs' occurrences are given in the lower triangular panels.

Table 3 | The estimated network size and count of triad and quadriad motifs in seven cell classes.

	No. of links	No. of feedback loop	No. of FFL	No. of biparallel	No. of bi-fan
Epithelia	344 ± 59*	1,896 ± 844	19,901 ± 8,419	1,858,957 ± 1,013,756	3,238,587 ± 1,618,601
Stroma	412 ± 38	2,727 ± 705	29,155 ± 6,290	3,052,803 ± 883,160	5,094,576 ± 1,337,401
Blood	434 ± 97	3,687 ± 1,699	37,884 ± 15,241	4,379,527 ± 2,320,472	7,359,970 ± 3,421,025
Endothelia	447 ± 40	3,160 ± 695	35,314 ± 6,567	3,844,161 ± 948,207	6,877,606 ± 1,540,212
Cancer	380 ± 7	2,378 ± 111	30,122 ± 710	2,862,215 ± 91,628	6,267,987 ± 99,444
Fetal cells	426 ± 70	3,088 ± 998	33,782 ± 9,955	3,660,840 ± 1,500,838	6,498,027 ± 2,284,014
ES cell†	485	2,766	32,400	3,282,473	6,436,708

ES, embryonic stem; FFL, feed-forward loop; TF, transcription factor.
*The motif count for each group is presented in the form mean ± s.d., and the numbers are presented in thousands.
†There is only one ES cell TF regulatory network.

number of occurrences. The embryonic stem cell TF regulatory network has the smallest number of occurrences relative to its network size for all the motifs.

In a random network, the ratio of the FFL count to the feedback loop count is ∼3:1. However, in the human cell-specific TF regulatory networks, the ratio is about 10:1, suggesting FFL is significantly enriched in these networks. Table 3 also suggests that the bi-fan motif is relatively abundant in these networks, as the ratio of the bi-fan count to the biparallel count is roughly 1:2 in a random network.

Discussion

By taking spurious and missing link rates into account, we have developed a powerful method for estimating the number of motif occurrences in the entire network from noisy and incomplete data for the first time. It extends previous studies on interactome size estimation[9,10,23–27] to motif count estimation in a directed or undirected network. Such a method is important because exact motif enumeration is possible only if the network is completely known, which is often not the case in biology. Our proposed method has been proven mathematically as being unbiased and accurate without any assumption at all regarding the topological structure of the underlying networks. Therefore, our proposed estimators can be applied to all the widely studied networks in social, biological and physical sciences.

Interactome size has been estimated from noisy subnetwork data by using equation (8), where the precision (which is $1-r_d$) and sensitivity of the data are taken into account[23,24,26,27]. This approach might yield an inaccurate estimate, as the false discovery rate is often calculated from gold-standard data sets[30–33] and can be quite unreliable, as indicated in ref. 26, in which the false discovery rate for the data set CCSB-HI1 was adjusted from 87% to 93%, to 20.6%, after multiple cross-assay validation. By contrast, our proposed method uses false-positive and false-negative rates for motif count estimation. As the false-negative rate is equal to 1 − sensitivity and the false-positive rate is only about 10^{-4}, our method is more robust than estimations based on the false discovery rate.

Theorems 1 and 2 in the present paper show that motif counting via sampling and then scaling up in a huge network is not merely fast but can also give accurate estimate. Take the triangle motif, for instance. In our validation test, the equation (1)-based sampling achieved less than 1% deviation from the actual count by using no more than 50% of the computing time compared with the naive triangle counting method (Fig. 4 and Supplementary Note 3). As the obtained sampling approach takes positive and negative link-error rates into account, it is a good addition to the methodology for estimating motif count in networks[34,35].

By applying our estimation method to PPI subnetwork data for four eukaryotic organisms, we found that the numbers of triangles in a eukaryotic interactome differ considerably. For example, the triangle motif is exceptionally enriched in the human interactome. As noted in ref. 9, we have to keep in mind that the estimates in Table 2 are based on PPIs that are detectable, given current experimental methods. However, our estimators will remain correct for any interaction data available in the future.

We also discovered that there is a very strong positive linear correlation between triad and quadriad motif occurrences in human cell-specific TF regulatory networks, and that the TF regulatory network of embryonic stem cells has the smallest number of occurrences relative to its network size for each of the common triad and quadriad motifs. Hence, our study reveals a surprising feature of the TF regulatory network of embryonic stem cells.

Finally, we remark that the proposed estimators for motif counting are derived using the assumption that the subnetwork data is the outcome of a uniform node sampling process. In practice, however, biologists may select proteins for study according to their biological importance. The accuracy of our proposed method was examined for a degree-bias and two other non-uniform node sampling schemes (Supplementary Note 4 and Supplementary Figs S10–S12). In the degree-bias sampling process, a network node is sampled independently with a probability that is proportional to its degree in the underlying network. By the nature of this sampling process, it leads to

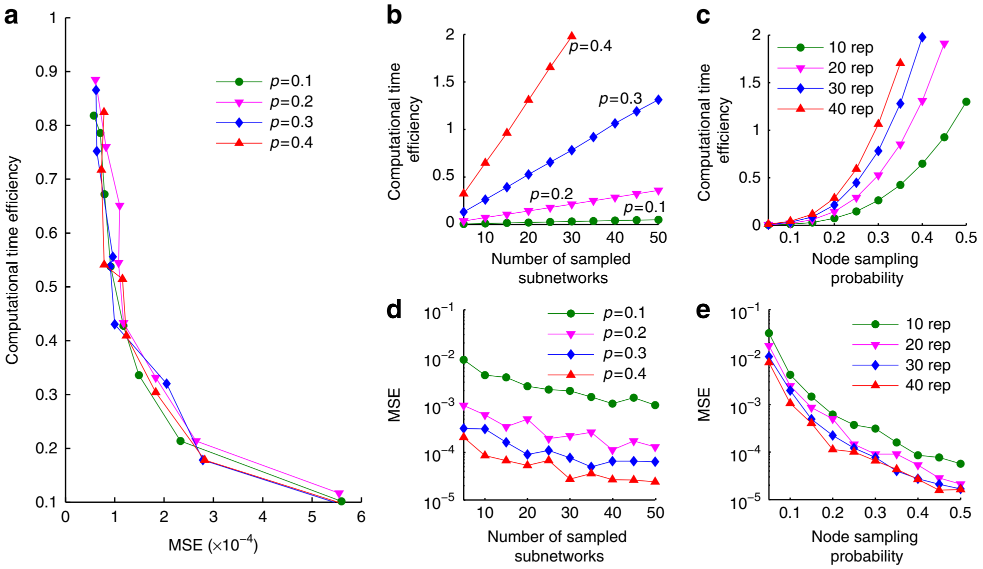


Figure 4 | Computational time efficiency of the proposed sampling approach. The simulation test was conducted on a network of 5,000 nodes with the edge density 0.1. The computational time efficiency of the sampling approach is defined as the ratio of the time taken by our approach to the time used by the direct counting approach, and MSE is defined in equation (9). (**a**) Computational time efficiency versus the MSE for four values of the node sampling probability p. When $p=0.1$, 0.2, 0.3 and 0.4, the number of repetitions was set to $125k$, $25k$, $5k$ and $2k$ ($1\leq k\leq 8$), respectively. (**b**) When the node sampling probability p is fixed, the computational time efficiency increases as a linear function of the number of repetitions. (**c**) When the number of repetitions (rep) is fixed, the computational time efficiency increases as a cubic function of p. (**d**) MSE decreases as the number of repetitions increases. (**e**) MSE decreases as p increases.

overestimation of motif count when our proposed estimator is used. Our simulation tests indicate that its effect on the estimation of motif count depends on the scale-free structure of the underlying network and the proportion of the sampled nodes. In particular, when more than 60% of nodes in a network are sampled, the estimate is no more than five times the actual count. Hence, the triangle counts in the four eukaryotic interactomes are likely less than the estimates listed in Table 2 by a small constant factor. How to correct the overestimation caused by a degree-bias node sampling is challenging and worthy to study in future.

Methods

Interaction data. Human, yeast, worm and *A. thaliana* PPI data sets were downloaded from the Center for Cancer Systems Biology (CCSB) (http://ccsb.dfci.harvard.edu): CCSB-YI1 (ref. 23), CCSB-WI-2007 (ref. 24), CCSB-HI1 (refs 25,26 and CCSB-AI1-Main[27]. TF regulatory interaction data sets were downloaded from the Supplementary Information of ref. 22 in the *Cell* journal website.

Simulation validation for motif estimators. We considered four widely used random graph models: ER[18], preferential attachment[19], duplication[20] and geometric models[21] (Supplementary Note 1). Using each model, we generated 200 random networks by using different combinations of node number $n \in \{500,1{,}000,1{,}500,\ldots,10{,}000\}$ and edge density $\rho \in \{0.01,0.02,\ldots,0.1\}$. Each generated network was taken as the whole network $\mathcal{G}$, from which 100 subnetworks were sampled using the node sampling probability $p \in \{0.05,0.1,0.15,\ldots,0.5\}$. For each motif $\mathcal{M}$ appearing in Fig. 1, we first computed $\widehat{N}_{\mathcal{M}}$ (given in equation (1)) from the motif count in each sampled subnetwork. This was used as an estimate of the number of occurrences of the motif in the error-free case, $N_{\mathcal{M}}$. Spurious and missing interactions were then planted in the sampled subnetworks with the chosen error rates r_+ and r_-. The bias-corrected estimator $\widetilde{N}_{\mathcal{M}}$ (given in Table 1) for $N_{\mathcal{M}}$ was then recalculated. We used the MSE of the ratios $\widehat{N}_{\mathcal{M}}/N_{\mathcal{M}}$ and $\widetilde{N}_{\mathcal{M}}/N_{\mathcal{M}}$ to measure the consistency (and hence accuracy) of $\widehat{N}_{\mathcal{M}}$ and $\widetilde{N}_{\mathcal{M}}$, respectively.

For the estimator Y of a parameter θ, the MSE of Y in estimating θ is defined as

$$\mathrm{MSE}(Y) = E((Y-\theta)^2).$$

This expression can be used to measure the MSE made in the estimation. In our validation test, we sampled 100 subnetworks from a network $\mathcal{G}$ to evaluate the consistency of the estimator $\widehat{N}_{\mathcal{M}}$ of a motif $\mathcal{M}$. As $E(\widehat{N}_{\mathcal{M}}/N_{\mathcal{M}})$ approaches to 1 when n is large (Theorem 1), the $\mathrm{MSE}(\widehat{N}_{\mathcal{M}}/N_{\mathcal{M}})$ was approximately computed as

$$\mathrm{MSE}\left(\frac{\widehat{N}_{\mathcal{M}}}{N_{\mathcal{M}}}\right) = \frac{1}{100}\sum_{1\le i\le 100}\left(\frac{\widehat{N}_{\mathcal{M},i}}{N_{\mathcal{M}}}-1\right)^2, \tag{9}$$

where $\widehat{N}_{\mathcal{M},i}$ is the estimate calculated from the i^{th} subnetwork using $\widehat{N}_{\mathcal{M}}$, $1 \le i \le 100$. Computing $\mathrm{MSE}(\widetilde{N}_{\mathcal{M}}/N_{\mathcal{M}})$ is similar.

References

1. Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5,** 101–113 (2004).
2. Ideker, T., Dutkowski, J. & Hood, L. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* **144,** 860–863 (2011).
3. Vidal, M., Cusick, M. E. & Barabasi, A.-L. Interactome networks and human disease. *Nat. Rev. Genet.* **12,** 56–68 (2011).
4. Barabasi, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Cell* **144,** 986–998 (2011).
5. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
6. Jeong, H., Mason, S. P., Barabasi, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411,** 41–42 (2001).
7. Hahn, M. W. & Kern, A. D. Comparative genomics of centrality and essentiality in three eukaryotic protein- interaction networks. *Mol. Biol. Evol.* **22,** 803–806 (2004).
8. He, X. & Zhang, J. Why do hubs tend to be essential in protein networks? *PLoS Genet.* **2,** e88 (2006).
9. Stumpf, M. P. H. *et al.* Estimating the size of the human interactome. *Proc. Natl Acad. Sci. USA* **105,** 6959–6964 (2008).
10. Rottger, R., Ruckert, U., Taubert, J. & Baumbach, J. How little do we actually know? On the size of gene regulatory networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **9,** 1293–1300 (2012).
11. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298,** 824–827 (2002).
12. Deng, M., Mehta, S., Sun, F. & Chen, T. Inferring domain-domain interactions from protein-protein interactions. *Genome Res.* **12,** 1540–1548 (2002).
13. Liu, Y., Liu, N. & Zhao, H. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* **21,** 3279–3285 (2005).
14. Guimera, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl Acad. Sci. USA* **106,** 22073–22078 (2009).
15. Sambourg, L. & Thierry-Mieg, N. New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. *BMC Bioinformatics* **11,** 605 (2010).
16. Kashtan, N., Itzkovitz, S., Milo, R. & Alon, U. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* **20,** 1746–1758 (2004).
17. Picard, F., Daudin, J.-J., Koskas, M., Schbath, S. & Robin, S. Assessing the exceptionality of network motifs. *J. Comput. Biol.* **15,** 1–20 (2008).
18. Erdos, P. & Renyi, A. On the strength of connectedness of a random graph. *Acta Math Hung.* **12,** 261–267 (1960).
19. Barabasi, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).
20. Chung, F., Lu, L., Dewey, T. G. & Galas, D. J. Duplication models for biological networks. *J. Comput. Biol.* **10,** 677–687 (2003).
21. Przulj, N., Corneil, D. G. & Jurisica, I. Modeling interactome: scale-free or geometric? *Bioinformatics* **20,** 3508–3515 (2004).
22. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150,** 1274–1286 (2012).
23. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322,** 104–110 (2008).
24. Simonis, N. *et al.* Empirically controlled mapping of the *Caenorhabditis elegans* protein–“protein interactome network. *Nat. Methods* **6,** 47–54 (2009).
25. Rual, J. F. *et al.* Towards a proteome-scale map of the human protein–“protein interaction network. *Nature* **437,** 1173–1178 (2005).
26. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* **6,** 83–90 (2009).
27. *Arabidopsis* Interactome Mapping Consortium. Evidence for network evolution in an *Arabidopsis* interactome map. *Science* **333,** 601–607 (2011).
28. Stelzl, U. *et al.* A human protein protein interaction network: a resource for annotating the proteome. *Cell* **122,** 957–968 (2005).
29. Wilson, D. *et al.* DBD-taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.* **36,** Database issue D88–D92 (2008).
30. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein protein interactions. *Nature* **417,** 399–403 (2002).
31. D'haeseleer, P. & Church, G. M. Estimating and improving protein interaction error rates. *Proc. IEEE Comput. Syst. Bioinform. Conf.* 216–223 (2004).
32. Hart, G. T., Ramani, A. K. & Marcotte, E. M. How complete are current yeast and human protein–interaction networks? *Genome Biol.* **7,** 120 (2006).
33. Reguly, T. *et al.* Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* **5,** 11 (2006).
34. Alon, N., Dao, P., Hajirasouliha, I., Hormozdiari, F. & Sahinalp, S. C. Biomolecular network motif counting and discovery by color coding. *Bioinformatics* **24,** i241–i249 (2008).
35. Gonen, M. & Shavitt, Y. Approximating the number of network motifs. *Internet Math.* **6,** 349–372 (2010).

Acknowledgements

We thank the two reviewers of this manuscript for valuable comments. We also thank Michael Calderwood, Jean-Francois Rual and Nicolas Simonis for their help in the analyses of interactome data. This work was supported by fund provided by Ministry of Education (Tier-2 grant R-146-000-134-112).

Author contributions

Theoretical study and data analyses: N.H.T. and K.P.C. Writing: N.H.T., K.P.C. and L.X.Z. Project design: K.P.C. and L.X.Z.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Tran, N. H. *et al.* Counting motifs in the human interactome. *Nat. Commun.* 4:2241 doi: 10.1038/ncomms3241 (2013).