Using networks to measure similarity between genes: association index selection

Juan I Fuxman Bass[1,2], Alos Diallo[1,2], Justin Nelson[3], Juan M Soto[1,2], Chad L Myers[3] & Albertha J M Walhout[1,2]

Biological networks can be used to functionally annotate genes on the basis of interaction-profile similarities. Metrics known as association indices can be used to quantify interaction-profile similarity. We provide an overview of commonly used association indices, including the Jaccard index and the Pearson correlation coefficient, and compare their performance in different types of analyses of biological networks. We introduce the Guide for Association Index for Networks (GAIN), a web tool for calculating and comparing interaction-profile similarities and defining modules of genes with similar profiles.

Biological processes are orchestrated through complex interaction networks. Networks are modeled as graphs that depict interactions ('edges') between biological entities such as genes, tissues, proteins and metabolites ('nodes'; see **Box 1**). If only one type of node is involved, as in protein-protein[1,2] or genetic interaction networks[3], the graph is defined as monopartite. Bipartite graphs, by contrast, describe interactions between two different types of nodes (X-type and Y-type), with edges connecting only nodes of different types (**Fig. 1a**). Bipartite graphs include protein-DNA interaction networks[4–6], metabolic networks[7,8], phenotypic networks[9] and expression networks[10–14].

Networks are powerful tools for gene function annotation. For instance, the 'guilt-by-association' principle postulates that if a node with an unknown function has an interaction profile similar to that of a node with a known function, its function may be similar as well[2,15]. Additionally, network analysis can identify modules—neighborhoods comprising nodes with similar interaction profiles that can point to functional relationships between larger sets of genes[16,17]. Although seemingly intuitive, it is not trivial to know how to best capture interaction-profile similarity between nodes, as numerous metrics, or association indices, can be used, and because each index can provide different values and rank similarity between pairs of nodes in a different order. Here, we provide an overview of commonly used association indices. We discuss the differences and similarities between association indices and provide a set of guidelines and a web tool for their selection for different applications.

Types of association indices

We focus here on bipartite networks that connect X-type nodes to Y-type nodes (**Fig. 1a**). In these networks, association indices can be used to measure shared Y-type nodes between two X-type nodes, or vice versa. An association index can measure interaction-profile similarity between X-type nodes A and B by calculating the shared partners ($|N(A) \cap N(B)|$), in relation to their total number of interactions ('node degree'), defined as $|N(A)|$ and $|N(B)|$, and the total number of Y-type nodes in the network (n_y) (**Fig. 1a**). There are three main types of indices, each of which uses the variables mentioned above in a different way (see **Box 2**).

Similarity indices reflect the proportion of overlap and consider only the number of shared interactions between two X-type nodes and the individual degrees of these nodes, but they do not take the total number of Y-type nodes in the network into account. There are many similarity indices, most of which scale interaction-profile similarity between 0 and 1 (ref. 18) (**Supplementary Table 1**). We will focus on four that are commonly used in genomics and systems biology (see **Box 2**). The Jaccard index calculates the proportion of Y-type nodes shared between two X-type

[1]Program in Systems Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [2]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [3]Department of Computer Science and Engineering, University of Minnesota–Twin Cities, Minneapolis, Minnesota, USA. Correspondence should be addressed to J.I.F.B. (juan.fuxmanbass@umassmed.edu) or A.J.M.W. (marian.walhout@umassmed.edu).

RECEIVED 14 JANUARY; ACCEPTED 22 JULY; PUBLISHED ONLINE 26 NOVEMBER 2013; DOI:10.1038/NMETH.2728

BOX 1 GLOSSARY OF TERMS

A **graph** is a pair $G = (N, E)$ comprising a set N of nodes connected by a set E of edges.
The **degree** of a node A ($|N(A)|$) is defined as the number of nodes with which it interacts.
Hubs are nodes with a disproportionately high degree.
A **module** is a set of highly interconnected nodes.
A **monopartite** graph contains only one type of node.
A **bipartite** graph contains two types of nodes (X-type and Y-type nodes), and connections occur only between nodes of a different type.
An **association index** is a measure that quantifies interaction-profile similarity.
An **association network** is a network in which two nodes of the same type (for example, only X-type nodes) are connected by an edge if their similarity exceeds a selected threshold.

nodes relative to the total number of Y-type nodes connected to either X-type node. The Simpson index (equal to the meet/min index[19] and similar to the topological overlap coefficient[16]) considers the number of shared Y-type nodes relative to the smallest degree of either X-type node. The geometric index calculates the square of the number of shared interactions between two X-type nodes, divided by the product of their individual degrees. Finally, the cosine index corresponds to the square root of the geometric index.

Unlike similarity indices, matching indices, such as the simple matching coefficient and the Hamann index (**Supplementary Table 1**), consider the proportion of shared Y-type nodes as well as Y-type nodes that are not connected to either of the two X-type nodes. Because biological networks are sparse, shared nonpartners can contribute more to the similarity between two nodes than shared partners. Therefore, matching indices are not appropriate for the analysis of most biological networks and will not be discussed further.

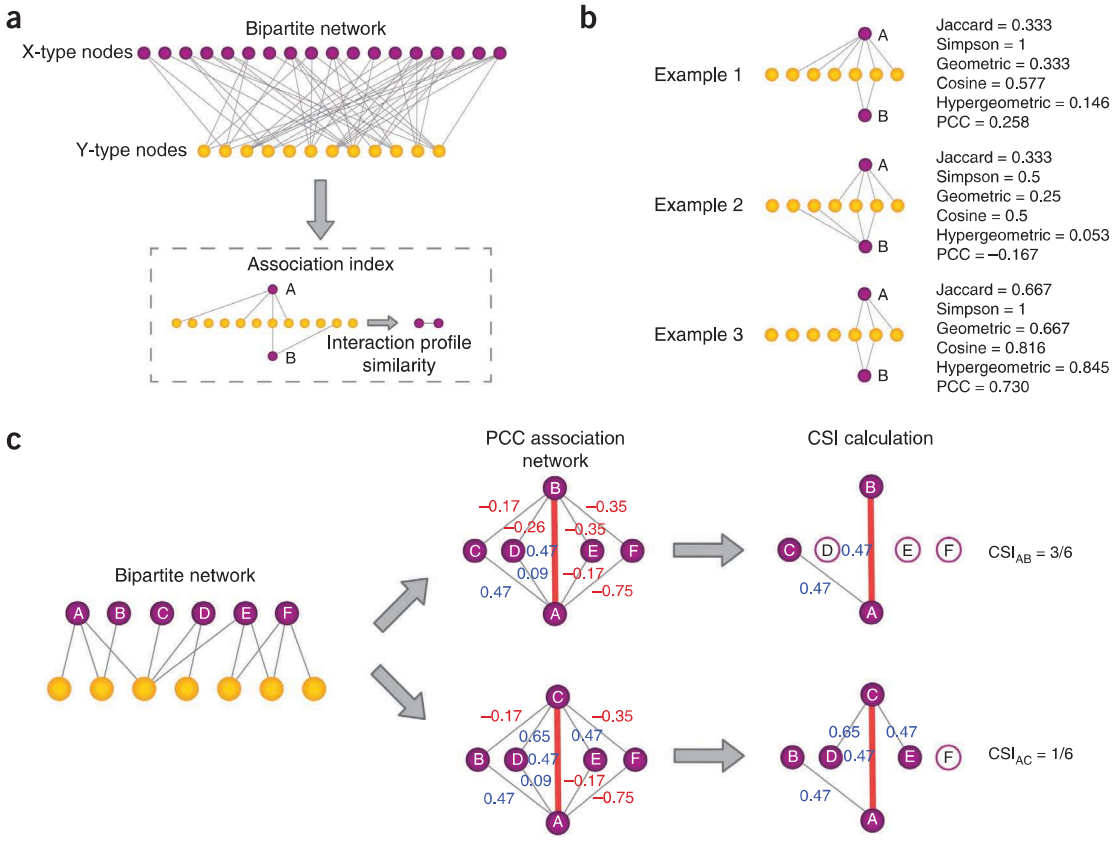

Figure 1 | Measuring interaction-profile similarity between two nodes using association indices. (**a**) Bipartite graphs connect two types of nodes: X-type (purple) and Y-type (yellow). The interaction-profile similarity between a pair of X-type nodes (A and B) is determined on the basis of the number of shared Y-type nodes, the total number of Y-type nodes connected to A and B, and the total number of Y-type nodes in the network. (**b**) Association index comparison. For each pair of X-type nodes, the Jaccard, Simpson, geometric, cosine and hypergeometric indices and PCC were calculated on the basis of their interactions with Y-type nodes. (**c**) CSI calculation between nodes A and B for a bipartite network involving six X-type nodes (purple) and seven Y-type nodes (yellow). For each pair of X-type nodes the PCC was calculated (blue, positive values; red, negative values). In the PCC association network all the edges connected to A or B are highlighted. CSI_{AB} represents the fraction of X-type nodes connected to both A and B, with PCC < PCC_{AB} − 0.05. The CSI was also calculated between A and C.

BOX 2 DEFINITIONS OF ASSOCIATION INDICES

The **Jaccard** index is the proportion of shared nodes between A and B relative to the total number of nodes connected to A or B.

$$J_{AB} = \frac{|N(A) \cap N(B)|}{|N(A) \cup N(B)|}$$

The **Simpson** index is the proportion of shared nodes relative to the degree of the least-connected node.

$$S_{AB} = \frac{|N(A) \cap N(B)|}{\min(|N(A)|, |N(B)|)}$$

The **geometric** index corresponds to the product of the proportion of shared nodes between A and B.

$$G_{AB} = \frac{|N(A) \cap N(B)|^2}{|N(A)| . |N(B)|}$$

The **cosine** index is the geometric mean of the proportions of shared nodes between A and B.

$$C_{AB} = \frac{|N(A) \cap N(B)|}{\sqrt{|N(A)| . |N(B)|}}$$

The **Pearson correlation coefficient** is the correlation between the interaction profiles of A and B.

$$PCC_{AB} = \frac{|N(A) \cap N(B)| . n_y - |N(A)| . |N(B)|}{\sqrt{|N(A)| . |N(B)| . (n_y - |N(A)|).(n_y - |N(B)|)}}$$

The **hypergeometric** index is the log-transformed probability of having an equal or greater interaction overlap than the one observed between A and B.

$$H_{AB} = -log \sum_{i=|N(A) \cap N(B)|}^{\min(|N(A),N(B)|)} \frac{\binom{|N(A)|}{i} \cdot \binom{n_y - |N(A)|}{|N(B)| - i}}{\binom{n_y}{|N(B)|}}$$

The connection specificity index (**CSI**) is defined as the fraction of X-type nodes that have an interaction profile similarity with A and B that is lower than the interaction profile similarity between A and B itself.

$$CSI_{AB} = 1 - \frac{\#\text{nodes connected to A or B with PCC} \geq PCC_{AB} - 0.05}{n_y}$$
$$= \frac{\#\text{nodes connected to A and B with PCC} < PCC_{AB} - 0.05}{n_y}$$

Statistic-based indices employ probability distributions (such as chi-square and Fisher's exact test) to determine the likelihood of observing a certain overlap between the interaction profiles of two X-type nodes given their degree and the total number of Y-type nodes in the network[18] (**Supplementary Table 1**). We will discuss two of the most commonly used statistic-based indices.

The Pearson correlation coefficient (PCC) was originally developed to measure the linear relationship between two continuous variables, such as protein and mRNA levels. This metric can also be applied to bipartite networks where interactions are either present or absent. The PCC provides a value between −1 and 1 that describes how well the interactions overlap. A PCC of 1 indicates a perfect overlap, 0 corresponds to the number of shared interactions expected by chance and −1 depicts perfect anticorrelation. The hypergeometric index calculates the log-transformed probability of observing an equal or greater number of shared nodes by chance and, therefore, measures the significance rather than the magnitude of the overlap.

Comparing association indices

Different association indices can provide different values of interaction-profile similarity. We illustrate this using three small example networks in which two X-type nodes, A and B, share different numbers of Y-type nodes, out of a total of seven (**Fig. 1b**). In each example, different indices can provide different values, ranging from perfect similarity ($Simpson_{AB}$ = 1 in example 1) to low similarity ($hypergeometric_{AB}$ = 0.146 in example 1). Further, different indices can rank the interaction-profile similarity between a pair of nodes in different orders. For instance, according to most indices the profiles of A and B are most similar

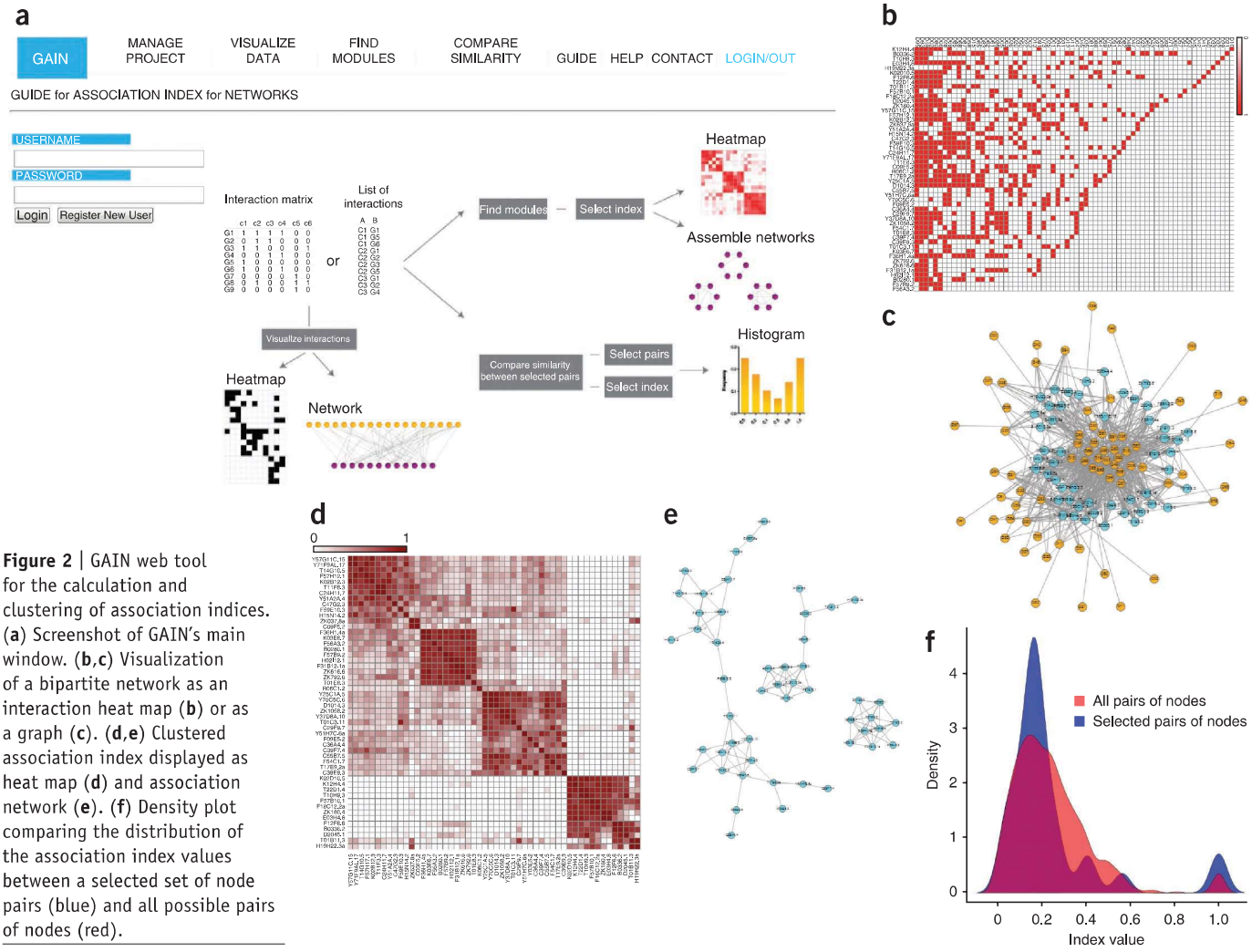

Figure 2 | GAIN web tool for the calculation and clustering of association indices. (**a**) Screenshot of GAIN's main window. (**b**,**c**) Visualization of a bipartite network as an interaction heat map (**b**) or as a graph (**c**). (**d**,**e**) Clustered association index displayed as heat map (**d**) and association network (**e**). (**f**) Density plot comparing the distribution of the association index values between a selected set of node pairs (blue) and all possible pairs of nodes (red).

in example 3, but the Simpson index ranks examples 1 and 3 as equally similar. Finally, even for pairs of nodes that have different overlap and/or node degree, an index may output identical values as it condenses four variables (overlap, degree of A, degree of B and n_γ) into a single number. For instance, the Jaccard index cannot discriminate between examples 1 and 2 ($Jaccard_{AB}$ = 0.333), in which the total number of edges is differently distributed between A and B, whereas the other indices can.

Nonspecific interactions can drive similarity

The indices mentioned above consider the similarity in interacting partners between two X-type nodes but not the interaction specificity. Two issues need to be considered. First, Y-type hubs may confer artificially high levels of interaction-profile similarity: if half of all X-type nodes bind a Y-type hub, this overlap is not very informative. Second, not all Y-type nodes are independent, which may also confer exaggerated levels of interaction-profile similarity. For instance, neurons can be classified into different categories on the basis of the tissues in which they are located. Different types of neurons express common genes. Thus, in a gene-to-tissue network where genes are connected to the tissues in which they are expressed, neuronal genes may be connected to many classes of neurons, artificially increasing their similarity.

The connection specificity index (CSI) provides a context-dependent measure that mitigates the effect of nonspecific interactions by ranking the significance of similarity between two X-type nodes according to the specificity of their shared interaction partners[9]. The CSI between two nodes A and B is defined as the fraction of X-type nodes that have an interaction-profile similarity with A and B that is lower than the interaction-profile similarity between A and B themselves (see **Box 2**). As originally defined, the CSI employs the PCC as a first-level association index to rank the similarity between nodes, and then uses a constant of 0.05 to define the lower boundary of interaction-profile similarity[9]. When the constant is increased, the CSI provides a more stringent measure. Other association indices may also be used for a first-level ranking of interaction-profile similarity. **Figure 1c** illustrates an example in which the CSI reduces the influence of hubs. In this network, A and B interact with three and one Y-type nodes, respectively, and share one Y-type node, resulting in a PCC_{AB} = 0.47 (**Fig. 1c**). A and C also share one interaction partner and therefore PCC_{AC} = 0.47 as well. However, many other X-type nodes interact with the Y-type node connected to both A and C, hence this shared interaction is less specific. Applying CSI to these networks alleviates this problem: when a constant of 0.05 is used, CSI_{AB} = 0.5, whereas CSI_{AC} = 0.17 (**Fig. 1c**).

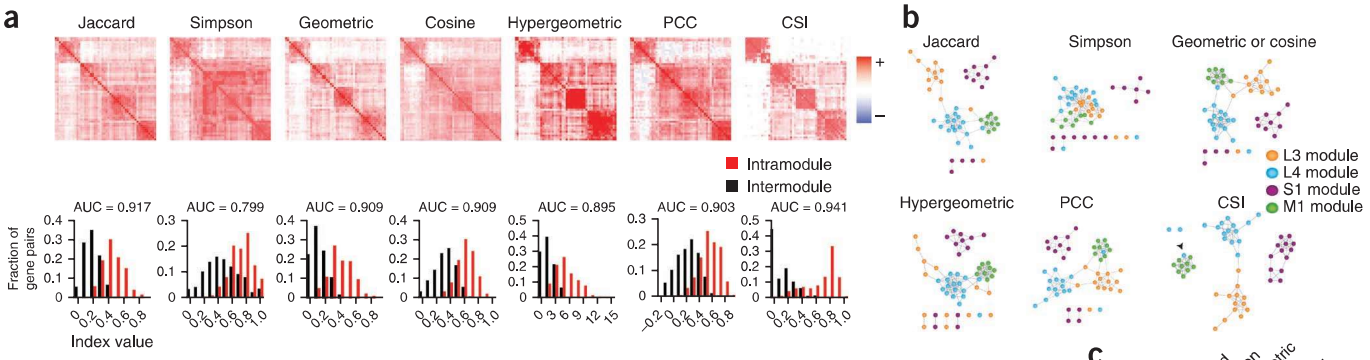

Figure 3 | Using association indices to identify modules in a gene-to-phenotype network. (**a**) Clustered association index heat maps for a *C. elegans* gene-to-phenotype network (top). The association index was calculated for each pair of genes according to shared phenotypic features and then clustered using hierarchical clustering. The distribution of index values for gene pairs that belong to the same module (intramodule, red) is plotted (bottom) against the values of gene pairs that belong to different modules (intermodule, black). The area under the receiver operating characteristic curve (AUC) measures the separation between the two distributions. (**b**) The association networks shown were assembled by linking genes that have a top 10% phenotypic profile similarity value using the indicated indices. A force-directed layout was generated by Cytoscape 2.8.1. The colors of the nodes represent a manual partitioning of the genes in modules[9]. (**c**) Pairwise comparison of the percentage of differences in the edges included in each of the association networks. Hyperg., hypergeometric.

GAIN: a web tool for association indices and clustering

We developed GAIN (http://csbio.cs.umn.edu/similarity_index/login.php) (**Fig. 2a** and **Supplementary Methods**), which allows a user to upload an interaction data set and perform several tasks. First, interactions can be visualized as a heat map (**Fig. 2b**) or graph (**Fig. 2c**). Second, GAIN allows the user to find modules by calculating all pairwise values with a user-selected association index followed by hierarchical clustering and by displaying a heat map (**Fig. 2d**) or association network (**Fig. 2e**). Association networks contain one node type connected by an edge only when their interaction-profile similarity exceeds a user-selected threshold. Finally, GAIN can display a density plot to determine whether an association index can discriminate the interaction-profile similarity of a particular set of node pairs selected by the user from all node pairs (**Fig. 2f**). For instance, in a gene regulatory network this can be used to determine whether pairs of highly homologous transcription factors have more similar interaction profiles than all possible pairs of transcription factors.

Finding network modules

Network modules are groups of nodes with relatively high interaction-profile similarity and can point to shared biological function between nodes. To compare how different association indices perform in the identification of network modules, we used two bipartite networks. The first is a subset of a *Caenorhabditis elegans* gene-to-phenotype network that connects 52 essential genes to 94 phenotypic features[9]. We used genes that belong to four modules manually determined by the authors of the original paper to benchmark the performance of the different indices. Association indices were calculated for each pair of genes according to their shared phenotypic features and then clustered into heat maps (**Fig. 3a**). Visual inspection shows that the Simpson index is least suitable for the identification of the four modules, and CSI performs the best. Consistent

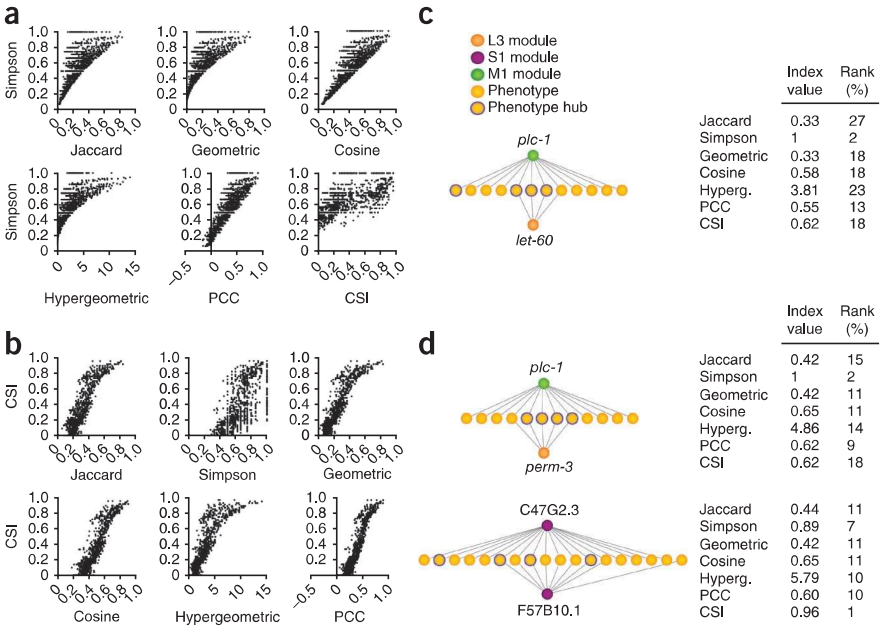

Figure 4 | Comparing association indices in the *C. elegans* gene-to-phenotype network. (**a**,**b**) Pairwise association index values for all genes according to shared phenotypic features for the Simpson index (**a**) and CSI (**b**), plotted against the values determined with the other indices. (**c**,**d**) The interaction-profile similarity between *plc-1* and *let-60* (belonging to different modules) (**c**), *plc-1* and *perm-3* (belonging to different modules) and between C47G2.3 and F57B10.1 (belonging to the same module) (**d**) were determined for all the association indices. The ranking of interaction-profile similarity across the entire network (in the top *x*% values from most to least similar) is indicated (right). Yellow nodes indicate phenotypes. Phenotype hubs (connected to more than 40% of the genes) are indicated with a blue outline. Hyperg., hypergeometric.

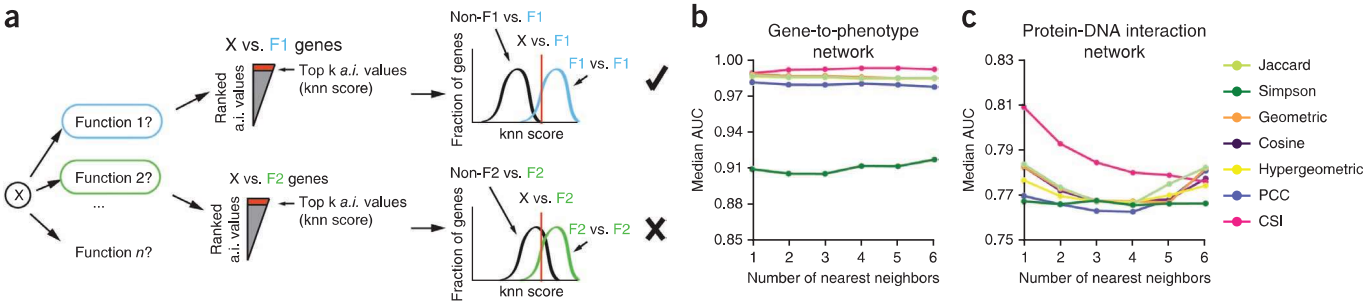

Figure 5 | Predicting gene function. (**a**) A *k*-nearest-neighbor (knn) algorithm was used to evaluate how well each index is able to assign genes to functional classes (F). To determine whether an uncharacterized gene X can be assigned a particular function, a knn score was determined as the average of the top *k* association index (a.i.) values between X and genes with that function. The knn values were then calculated for genes that have that function (blue and green) and for genes that do not (black curves). To assign a function to gene X, these values should be well separated (determined by calculating the AUC). **a** illustrates a case in which gene X can be assigned function 1 (F1, blue) but not function 2 (F2, green). The median AUC determined for all the functional classes was used as a measure of performance of the different association indices to predict gene function. (**b**) The median AUC calculated for the four functional classes in the *C. elegans* gene-to-phenotype network was determined for *k* values of 1 to 6 (number of nearest neighbors). (**c**) The median AUC calculated for the biological process GO slim terms in the yeast protein-DNA interaction network was determined for *k* values from 1 to 6.

with this observation, we found that CSI is best able to discriminate between the interaction-profile similarity between nodes that belong to the same module and that of nodes belonging to different modules (**Fig. 3a**).

Next, we asked which index performs best to delineate association networks for this gene-to-phenotype network. These networks connect nodes that have an interaction-profile similarity above a certain threshold. Therefore, they serve not only to delineate modules but also to identify nodes related to more than one module and nodes that are not related to any module. We used the top 10% of the values obtained with each index (**Fig. 3b**). CSI outperforms the other indices, as it (i) better demarcates the modules, (ii) leaves only two genes not assigned to any module and (iii) places only one gene into a module different from that assigned by manual classification (**Fig. 3b**). Generally, association networks obtained with different indices exhibit a large degree of overlap in the edges included, except for those obtained with the Simpson index and CSI (**Fig. 3c**). Indeed, by determining all pairwise association index values for each pair of indices, i.e., by not limiting to the top 10%, comparisons involving Simpson or CSI were least correlated (**Fig. 4a**,**b** and **Supplementary Fig. 1**). This analysis of a real network further substantiates the notion that different indices can result in different values and ranking of interaction-profile similarity. Neither the Simpson index nor CSI is well correlated with any of the other indices, but the consequences in each case are quite different: for Simpson the poor correlation results in reduced module demarcation, whereas for CSI it results in precisely the opposite. The denominator in

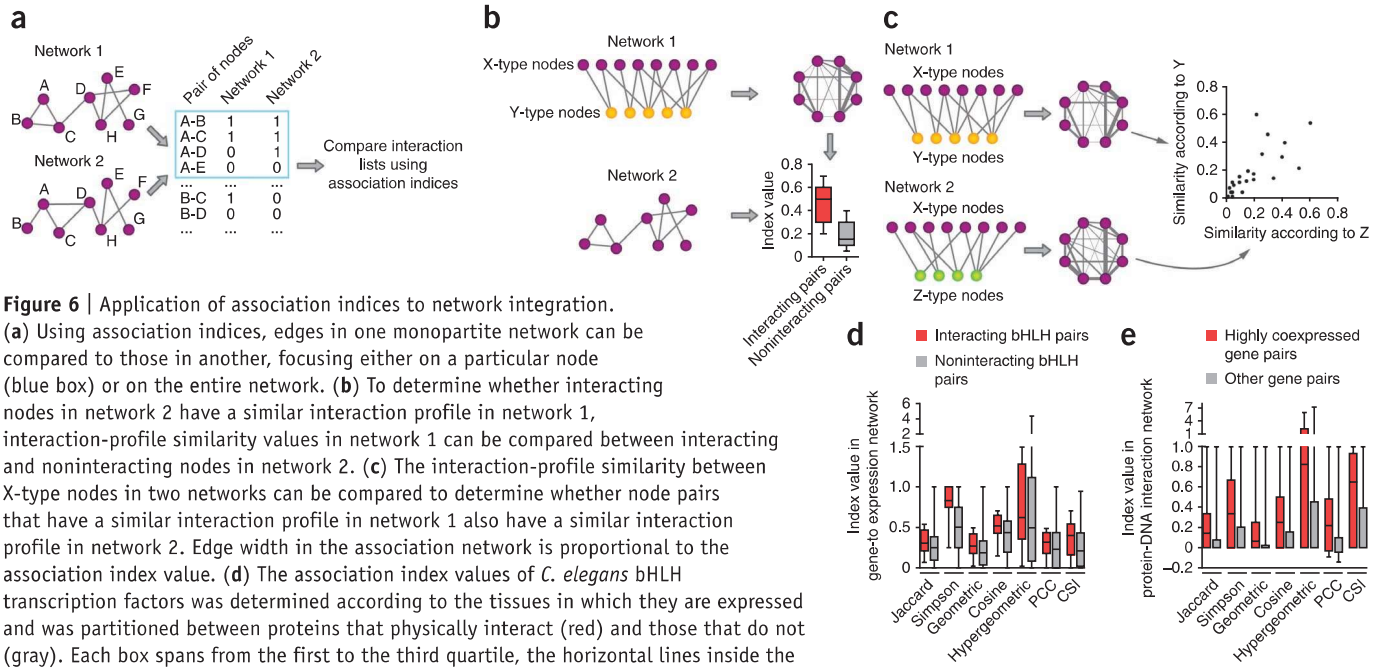

Figure 6 | Application of association indices to network integration. (**a**) Using association indices, edges in one monopartite network can be compared to those in another, focusing either on a particular node (blue box) or on the entire network. (**b**) To determine whether interacting nodes in network 2 have a similar interaction profile in network 1, interaction-profile similarity values in network 1 can be compared between interacting and noninteracting nodes in network 2. (**c**) The interaction-profile similarity between X-type nodes in two networks can be compared to determine whether node pairs that have a similar interaction profile in network 1 also have a similar interaction profile in network 2. Edge width in the association network is proportional to the association index value. (**d**) The association index values of *C. elegans* bHLH transcription factors was determined according to the tissues in which they are expressed and was partitioned between proteins that physically interact (red) and those that do not (gray). Each box spans from the first to the third quartile, the horizontal lines inside the boxes indicate median value and the whiskers indicate minimum and maximum values (**b**,**d**). (**e**) Association index values were determined for pairs of promoters in the yeast protein-DNA interaction network. The values for pairs of highly coexpressed genes (top 1%; red) and other gene pairs (bottom 99%; gray) are plotted.

Table 1 | Association index performance for different applications

	Jaccard	Simpson	Geometric	Cosine	Hypergeometric	PCC	CSI
Identifying network modules	**	*	**	**	**	**	***
Predicting gene function	**	*	**	**	**	**	***
Comparing two sets of node pairs[a]	**	***	**	**	*	***	*
Determining significance of overlap	No	No	No	No	Yes	No	No

Asterisks indicate qualitative strengths, with a greater number indicating greater utility.
[a]Assessment depends on biological question or objective.

the Simpson index uses only the lower of the two overall node degrees, which can lead to artificially high levels of interaction-profile similarity, even for genes belonging to different modules (**Fig. 4c**). In the case of CSI, ranking similarity according to interaction specificity results in a higher value for gene pairs with shared specific phenotypes (C47G2.3 and F57B10.1), and a lower value for gene pairs with shared common phenotypes (*plc-1* and *perm-3*) (**Fig. 4d**).

The second example network contains protein-DNA interactions between 102 yeast transcription factors and 542 promoters[4]. Association indices were calculated for each pair of promoters according to their shared transcription factors, and values were clustered into heat maps (**Supplementary Fig. 2a**). The heat maps are visually quite similar, and numerous modules can be detected. There were no previously benchmarked modules available. However, because genes with similar functions are frequently bound by the same transcription factor(s), we assessed the performance of the different indices by analyzing the biological process Gene Ontology (GO) enrichment in three different modules (**Supplementary Fig. 2b**). Two modules were detected equally well by all association indices (**Supplementary Fig. 2c**); however, for the third module significant enrichment ($P < 0.001$) for genes involved in the oxidation-reduction process was detected only by CSI and the Jaccard, geometric and hypergeometric indices (**Supplementary Fig. 2c**). Thus, association indices can perform differently in different types of networks and even within a network.

Predicting function of individual genes

Biologists frequently identify single genes of unknown functions, for instance in a genetic screen. So far, we have discussed network modules as a starting point for functional annotation. However, for analysis of only a single gene, there is no need to first comprehensively identify network modules. Moreover, modules are not always suitable for annotation of the function of every gene, as a gene may not belong to a clearly defined module and may have more than one function. An intuitive way to annotate gene function is to use the guilt-by-association principle, which postulates that two genes with similar functions have similar interaction profiles. One can assign functions to genes using a variety of different algorithms. Here, we use a *k*-nearest-neighbor algorithm that tests associations between genes and functions (**Fig. 5a**). An unknown gene can be assigned to each function F depending on (i) the top *k* association index values between that gene and the gene(s) that are known to have that function, and (ii) the specificity of the distribution of those values. In the example shown in **Figure 5a**, the highest score for the unknown gene (X) with genes with either known function 1 (F1, blue) or function 2 (F2, green) is similar (red lines). However, for function 1, the two distributions are largely separate, whereas for function 2 the two distributions overlap greatly. Thus, function 1 can be assigned to gene X with greater confidence than function 2. We assessed which association index best predicts function using the two networks described above. Again, CSI was best able to assign genes to functional classes, and the Simpson index performed the worst (**Fig. 5b**,**c**). This result is consistent with the ability of CSI to consider interaction specificity.

Integrated networks

The integration of different types of networks enables the comparison of pairs of nodes across networks[13,20]. Questions that can be answered include (i) whether directly interacting pairs of nodes in one network also tend to interact in another (**Fig. 6a**; note that this involves two monopartite networks), (ii) whether interacting nodes in one monopartite network have similar interaction profiles in a bipartite network (**Fig. 6b**) and (iii) whether pairs of nodes with similar interaction profiles in one bipartite network are also similar in another bipartite network (**Fig. 6c**). An example of the first type of question is whether the genes that encode physically interacting proteins also interact genetically. An example of the second type of question is whether proteins that physically interact tend to share phenotypes. Finally, an example of the third question is whether transcription factors that regulate a shared set of target genes are expressed in the same tissues and/or under the same conditions.

To determine whether interacting nodes in one monopartite network also interact in another network, the overlap between both sets of interactions can be determined, using association indices, on the basis of the number of shared edges between both networks, the number of edges in each network and the total number of node pairs (**Fig. 6a**). To determine the magnitude of similarity, Jaccard and Simpson are most suitable, and the hypergeometric index can be used to determine significance. The same approach can be used to compare interacting node pairs between modules in one network to those in another. Such 'cross-network module preservation' has been evaluated elsewhere[21].

To integrate a monopartite and a bipartite (**Fig. 6b**), or two bipartite, networks (**Fig. 6c**), the biological question should inform index selection. We illustrate this with two data sets. The first is a multiparameter, integrated *C. elegans* basic helix-loop-helix (bHLH) network comprising protein-protein interactions and gene-to-tissue expression patterns[13]. Each index revealed that interacting bHLH proteins are more often coexpressed than noninteracting ones (**Fig. 6d**). However, Simpson outperformed the other indices (**Fig. 6d** and **Supplementary Fig. 3a**). This is because a few bHLH proteins that bind many partners are broadly expressed, but each protein's partners are expressed in only a subset of tissues. This is best captured by the Simpson index, as it uses the minimum node degree in the denominator. The second data set is the yeast protein-DNA interaction network,

integrated with a microarray coexpression network[22]. All indices revealed that highly coexpressed genes have higher protein-DNA interaction-profile similarity than other gene pairs (**Fig. 6e**). The PCC best separated the two categories, and the Simpson index was least efficient (**Supplementary Fig. 3b**). The relatively poor performance of the Simpson index is because it considers the degree of only the least connected promoter. As a consequence, a promoter bound by many transcription factors may be regarded as similar to a promoter bound by few, some of which are shared. However, differences between transcription factors bound to promoters are also highly meaningful, as these may contribute to distinct gene-expression profiles.

Conclusions

Different association indices can be used to compare interaction-profile similarity within and across networks, and different indices have strengths and weaknesses for different applications (**Table 1**). CSI is most suitable for predicting gene function and identifying modules. However, CSI levels the similarities between modules, which is a disadvantage in comparing modules. When the main goal is to compare the similarity between node pairs, the biological question should drive index selection. For instance, the Simpson index may be used to avoid penalizing large differences in node degree. If one wants, conversely, to capture this difference, other indices are more appropriate. The hypergeometric index should be used with caution to determine the magnitude of similarity between interaction profiles, as it does not scale linearly with the proportion of overlap. However, only this index is able to calculate the statistical significance of interaction-profile overlap.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank members of A.J.M.W.'s lab, R. McCord and B. Lajoie for discussions and critical reading of the manuscript. We thank J.C. Bare (Institute of Systems Biology) for helpful advice in the development of GAIN. This work was supported by the US National Institutes of Health grants DK068429 and GM082971 to A.J.M.W. J.I.F.B. is partially supported by a postdoctoral fellowship from the Pew Latin American Fellows Program. J.N. and C.L.M. are partially supported by grant DBI-0953881 from the US National Science Foundation.

AUTHOR CONTRIBUTIONS

J.I.F.B. and A.J.M.W. conceived the project; J.I.F.B. performed the data analysis with the assistance of A.D., J.N. and C.L.M.; A.D. and J.I.F.B. developed the GAIN web tool in collaboration with J.N., C.L.M. and J.M.S.; J.I.F.B. and A.J.M.W. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Walhout, A.J.M. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
2. Schwikowski, B., Uetz, P. & Fields, S. A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257–1261 (2000).
3. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
4. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
5. Walhout, A.J.M. Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res.* **16**, 1445–1454 (2006).
6. Reece-Hoyes, J.S. *et al.* Extensive rewiring and complex evolutionary dynamics in a *C. elegans* multiparameter transcription factor network. *Mol. Cell* **51**, 116–127 (2013).
7. Bordbar, A. & Palsson, B.O. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *J. Intern. Med.* **271**, 131–141 (2012).
8. Watson, E., MacNeil, L.T., Arda, H.E., Zhu, L.J. & Walhout, A.J.M. Integration of metabolic and gene regulatory networks modulates the *C. elegans* dietary response. *Cell* **153**, 253–266 (2013).
9. Green, R.A. *et al.* A high-resolution *C. elegans* essential gene network based on phenotypic profiling of a complex tissue. *Cell* **145**, 470–482 (2011).
10. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
11. Fowlkes, C.C. *et al.* A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm. *Cell* **133**, 364–374 (2008).
12. Martinez, N.J., Ow, M.C., Reece-Hoyes, J., Ambros, V. & Walhout, A.J. Genome-scale spatiotemporal analysis of *Caenorhabditis elegans* microRNA promoter activity. *Genome Res.* **18**, 2005–2015 (2008).
13. Grove, C.A. *et al.* A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* **138**, 314–327 (2009).
14. Ritter, A.D. *et al.* Complex expression dynamics and robustness in *C. elegans* insulin networks. *Genome Res.* **23**, 954–965 (2013).
15. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E. & Marcotte, E.M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011).
16. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabasi, A.L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
17. Spirin, V. & Mirny, L.A. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA* **100**, 12123–12128 (2003).
18. Hayek, L.-A.C. in *Measuring and Monitoring Biological Diversity: Standard Methods for Amphibians.* (ed. Heyer, W.R.) Ch. 9, 207–269 (Smithsonian Institution, Washington, DC, 1994).
19. Goldberg, D.S. & Roth, F.P. Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. USA* **100**, 4372–4376 (2003).
20. Gunsalus, K.C. *et al.* Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**, 861–865 (2005).
21. Langfelder, P., Luo, R., Oldham, M.C. & Horvath, S. Is my network module preserved and reproducible? *PLoS Comput. Biol.* **7**, e1001057 (2011).
22. Huttenhower, C., Hibbs, M., Myers, C. & Troyanskaya, O.G. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics* **22**, 2890–2897 (2006).