Published in final edited form as:
Nat Methods. 2013 December ; 10(12): 1169–1176. doi:10.1038/nmeth.2728.

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], and **Albertha J M Walhout**[1,2]

[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

Abstract

Biological networks can be used to functionally annotate genes based on 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 Jaccard and Pearson Correlation Coefficient, and compare their performance in different types of analyses. We introduce a web tool (‘GAIN’ - Guide for Association Index for Networks) to calculate and compare interaction profile similarities, and to define modules of genes with similar profiles.

INTRODUCTION

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’) (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, on the other hand, describe interactions between two different types of nodes (X-type and Y-type) with edges only connecting nodes of a different type (Fig. 1a). These 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 unknown function has a similar interaction profile as a node with a known function, the first node may have a similar function[2,15]. Additionally, network analysis can identify modules: neighborhoods comprised of nodes with similar interaction profiles that can point to functional relationships between larger sets of genes[16,17]. While 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

Correspondence should be addressed to J.I.F.B. (juan.fuxmanbass@umassmed.edu) or A.J.M.W. (marian.walhout@umassmed.edu).

AUTHOR CONTRIBUTIONS

J.I.F.B. and A.J.M.W. conceived the project; J.I.F.B. performed the data analysis with 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.

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 will focus 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'), $|N(A)|$ 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 that can be used, each of which utilizes the variables mentioned above in a different way (see below and Box 2).

Similarity indices reflect the proportion of overlap and only consider the number of shared interactions between two X-type nodes and their individual degrees, but 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 (reviewed in[18], Supplementary Table 1). We will focus on four that are commonly used in genomics and systems biology (Box 2).

The Jaccard index calculates the proportion of shared Y-type nodes between two X-type 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 non-partners 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.

Statistic-based indices employ probability distributions (Chi-square, Fisher's exact test, *etc*.) 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 (Supplementary Table 1)[18]. We will discuss two 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 anti-correlation. 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

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 ($\mathrm{Simpson}_{AB}$ = 1 in example 1) to low similarity ($\mathrm{Hypergeometric}_{AB}$ = 0.146 in example 1). Further, different indices can rank the interaction profile similarity between a pair of nodes in a different order. For instance, while according to most indices the profiles of A and B are most similar in example 3, 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_y) into a single number. For instance, the Jaccard index cannot discriminate between examples 1 and 2 ($\mathrm{Jaccard}_{AB}$ = 0.333), in which the total number of edges is differently distributed between A and B, while the other indices can.

Non-specific interactions can drive interaction profile similarity

The indices mentioned above only consider the similarity in interacting partners between two X-type nodes but do not consider 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 based on the tissues in which they are located. Different types of neurons express common genes. Thus, in a gene-to-tissue network, 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 non-specific interactions by ranking the significance of similarity between two X-type nodes according to the specificity of their shared interaction partners[9]. 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 (Box 2). As originally defined, 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, 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 where 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, and hence this shared interaction is less specific. Applying CSI to these networks alleviates this: when a constant of 0.05 is used, CSI_{AB} = 0.5 whereas CSI_{AC} = 0.17 (Fig. 1c).

GAIN: a web tool for association indices and clustering

We developed a web tool named GAIN (Guide for Association Index for Networks) (http://csbio.cs.umn.edu/similarity_index/login.php) (Fig. 2a). GAIN allows the user to upload an interaction datasets and perform several tasks. First, interactions can be visualized as a heatmap (Fig. 2b) or graph (Fig. 2c). Second, GAIN allows the user to find modules by calculating all pair-wise values with a user-selected association index followed by hierarchical clustering and by displaying a heatmap (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 of interest from all node pairs (Fig. 2f). For instance, in a gene regulatory network this can be used to determine if pairs of highly homologous transcription factors (TFs) have more similar interaction profiles than all possible TF pairs.

Finding network modules

Network modules are groups of nodes with relatively high interaction profile similarity and can point to shared biological function. To compare how different association indices perform to identify network modules, we used two bipartite networks. The first is a subset of a *C. elegans* gene-to-phenotype network that connects 52 essential genes to 94 phenotypic features[9]. Genes that belong to four modules manually determined by the authors of the original paper were selected and served 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 heatmaps (Fig. 3a). Visual inspection shows that the Simpson index is least suitable for the identification of the four modules while CSI performs the best. This observation was confirmed quantitatively by calculating the separation 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. We used the top 10% of the values obtained with each index (Fig. 3b). CSI outperforms the other indices as: (1) it better demarcates the modules, (2) only two genes are not assigned to any module, and (3) only one gene is placed into a different module than 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 pair-wise 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; Supplementary Fig. 1). Importantly, this analysis of a real network further substantiates the notion that different indices can result in different values and ranking of interaction profile similarity, as described theoretically above for a small sample network (Fig. 1b). While the Simpson index and CSI are both not well correlated with any of the other indices, the consequences of this are quite different: for Simpson it results in reduced module demarcation, whereas for CSI it is precisely the opposite. The denominator in the Simpson index only uses 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 TFs and 542 promoters[4]. Association indices were calculated for each pair of promoters according to their shared TFs and values were clustered into heatmaps (Supplementary Fig. 2a). The heatmaps are visually quite similar and numerous modules can be detected. There were no previously benchmarked modules available. However, since genes with similar functions are frequently bound by the same TF(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 for genes involved in oxidation-reduction process was only detected using the Jaccard, Geometric, Hypergeometric and CSI 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

Frequently, biologists identify a single gene of unknown function, for instance in a genetic screen. So far, we have discussed network modules as a starting point for functional annotation. However, when only a single gene is being analyzed, there is no need to first comprehensively identify network modules. Moreover, modules are not always suitable to annotate gene function because a gene may not belong to a clearly defined module and it may have more than one function. An intuitive way to annotate gene function is to use the guilt-by-association principle, which postulates that 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 (1) the top k association index values between that gene and the gene(s) that are known to have that function, and (2) 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, while 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, while the Simpson index performed the worst (Fig. 5b, c). This 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: (1) whether directly interacting pairs of nodes in one network also tend to interact in another (Fig. 6a, note that this involves two monopartite networks), (2) whether interacting nodes in one monopartite network have similar interaction profiles in another, bipartite network, or (3) whether pairs of nodes with similar interaction profiles in one bipartite network are also similar in another (Fig. 6c). An example of the first type of question is whether 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 TFs 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, based on 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 or Simpson are most suitable and the Hypergeometric index can be used to determine significance. One can also determine the overlap between modules in one network versus another. Such 'cross-network module preservation' has been evaluated elsewhere[21].

To integrate a monopartite and bipartite (Fig. 6b), or two bipartite networks (Fig. 6c), the biological question should inform index selection. We illustrate this with two datasets. The first is a multiparameter, integrated *C. elegans* basic helix-loop-helix (bHLH) network comprised of 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, Supplementary Fig. 3a). This is because a few bHLH proteins that bind many partners are broadly expressed, while each of its partners is 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 dataset 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, while the Simpson index was least efficient (Supplementary Fig. 3b). The relatively poor performance of the Simpson index is because it only considers the degree of the least connected promoter. As a consequence, a promoter bound by many TFs may be regarded as similar to a promoter bound by few, some of which are shared. However, differences between TFs 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 to predict gene function and identify modules. However, CSI levels the similarities between modules, which is a disadvantage to compare 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. Conversely, if one wants 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 suited to calculate the statistical significance of interaction profile overlap.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the Walhout 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 National Science Foundation.

References

1. Walhout AJM, et al. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science. 2000; 287:116–122. [PubMed: 10615043]

2. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat. Genet. 2000; 18:1257–1261.

3. Costanzo M, et al. The genetic landscape of a cell. Science. 2010; 327:425–431. [PubMed: 20093466]

4. Harbison CT, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004; 431:99–104. [PubMed: 15343339]

5. Walhout AJM. Unraveling Transcription Regulatory Networks by Protein-DNA and Protein-Protein Interaction Mapping. Genome Res. 2006; 16:1445–1454. [PubMed: 17053092]

6. Reece-Hoyes JS, et al. Extensive rewiring and complex evolutionary dynamics in a *C. elegans* multiparameter transcription factor network. Mol Cell. 2013; 51:116–127. [PubMed: 23791784]

7. Bordbar A, Palsson BO. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. J Intern Med. 2012; 271:131–141. [PubMed: 22142339]

8. Watson E, MacNeil LT, Arda HE, Zhu LJ, Walhout AJM. Integration of metabolic and gene regulatory networks modulates the *C. elegans* dietary response. Cell. 2013; 153:253–266. [PubMed: 23540702]

9. Green RA, et al. A high-resolution C elegans essential gene network based on phenotypic profiling of a complex tissue. Cell. 2011; 145:470–482. [PubMed: 21529718]

10. Su AI, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 2004; 101:6062–6067. [PubMed: 15075390]

11. Fowlkes CC, et al. A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. Cell. 2008; 133:364–374. [PubMed: 18423206]

12. Martinez NJ, Ow MC, Reece-Hoyes J, Ambros V, Walhout AJ. Genome-scale spatiotemporal analysis of *Caenorhabditis elegans* microRNA promoter activity. Genome Res. 2008; 18:2005–2015. [PubMed: 18981266]

13. Grove CA, et al. A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. Cell. 2009; 138:314–327. [PubMed: 19632181]

14. Ritter AD, et al. Complex expression dynamics and robustness in *C. elegans* insulin networks. Genome research. 2013; 23:954–965. [PubMed: 23539137]

15. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome research. 2011; 21:1109–1121. [PubMed: 21536720]

16. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. Science. 2002; 297:1551–1555. [PubMed: 12202830]

17. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100:12123–12128. [PubMed: 14517352]

18. Hayek, L-AC. Analysis of amphibian biodiversity data. Chapter 9. In: Heyer, WR., editor. Measuring and monitoring miological diversity. Standard methods for amphibians. Washington D.C.: Smithsonian Institution; 1994. p. 207-269.

19. Goldberg DS, Roth FP. Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci U S A. 2003; 100:4372–4376. [PubMed: 12676999]

20. Gunsalus KC, et al. Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. Nature. 2005; 436:861–865. [PubMed: 16094371]

21. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? PLos Computational Biology. 2011; 7:e1001057. [PubMed: 21283776]

22. Huttenhower C, Hibbs M, Myers C, Troyanskaya OG. A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics. 2006; 22:2890–2897. [PubMed: 17005538]

Box 1

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 (*e.g.*, only X-type nodes) are connected by an edge if their similarity exceeds a selected threshold.

Box 2

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.

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

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

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

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

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

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 **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.

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

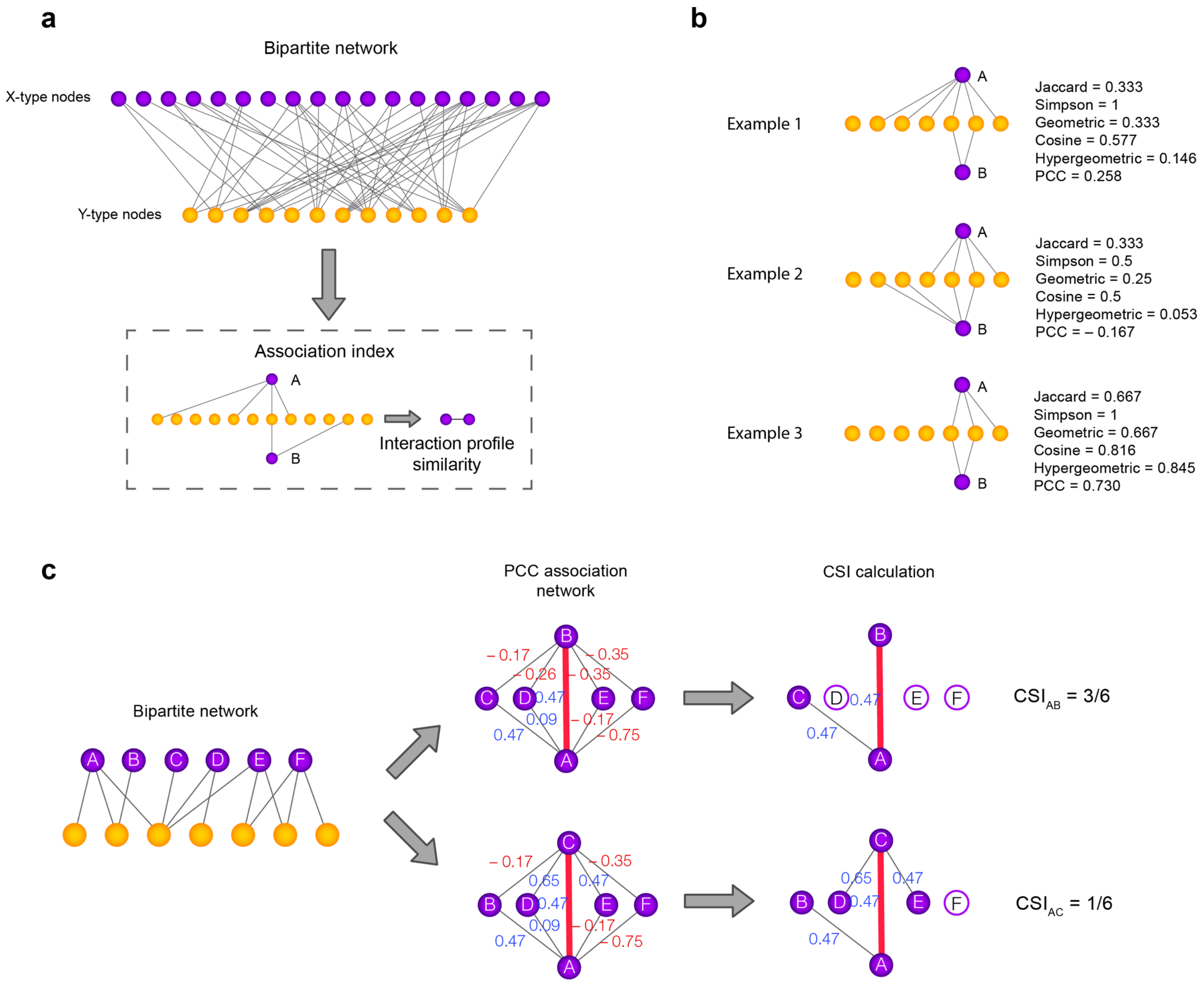
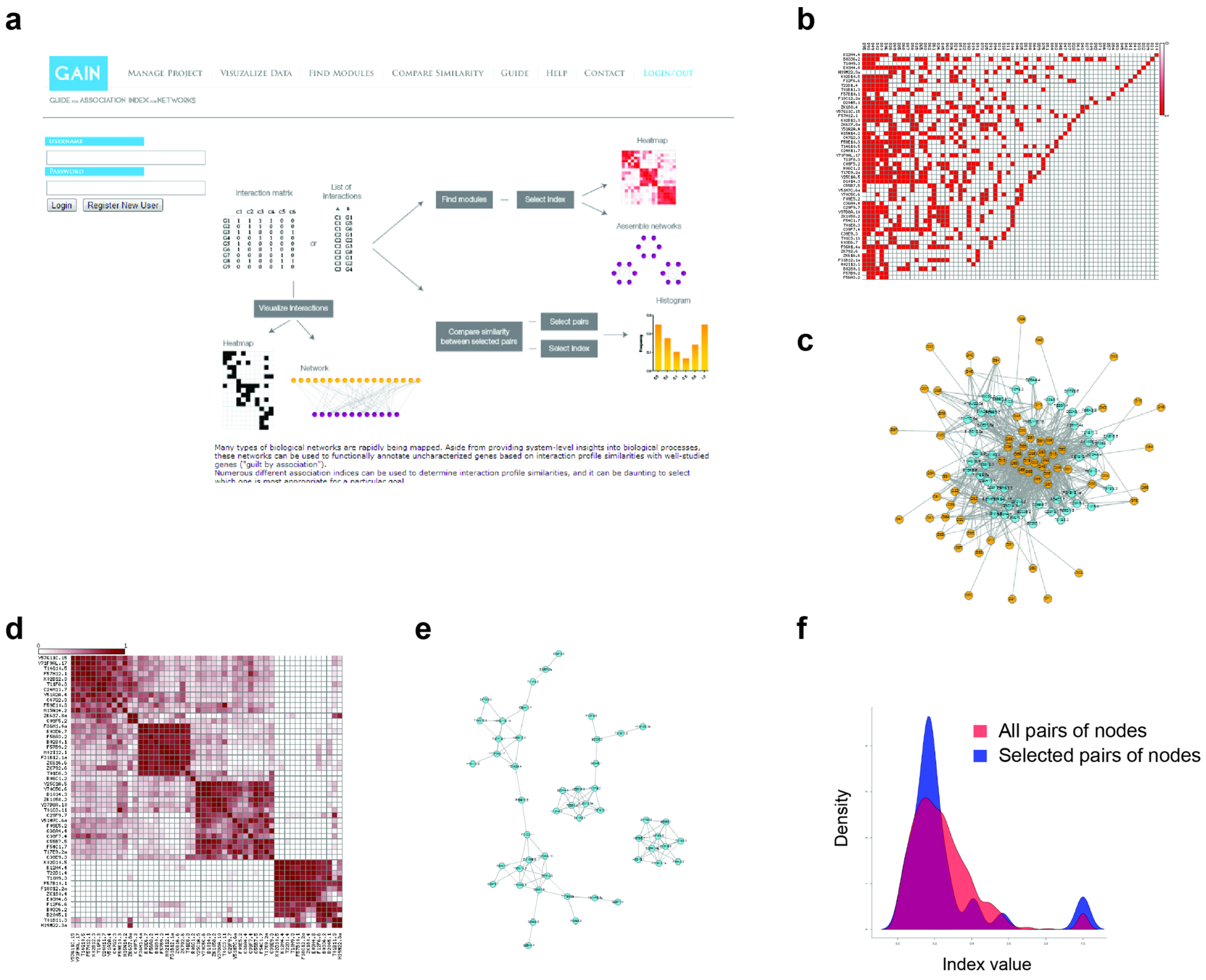


Figure 1.

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, B) is determined based on the number of shared Y-type nodes and the total number of Y nodes connected to A and B. (**b**) Association index comparison. For each pair of X-type nodes the Jaccard, Simpson, Geometric, Cosine, Hypergeometric indices and PCC were calculated based on 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. CSI was also calculated between A and C.



	c1	c2	c3	c4	c5	c6
G1	1	1	1	1	0	0
G2	0	1	1	0	0	0
G3	1	1	0	0	0	1
G4	0	0	1	1	0	0
G5	1	0	0	0	0	0
G6	1	0	0	1	0	0
G7	0	0	0	0	1	0
G8	0	1	0	0	1	1
G9	0	0	0	0	1	0

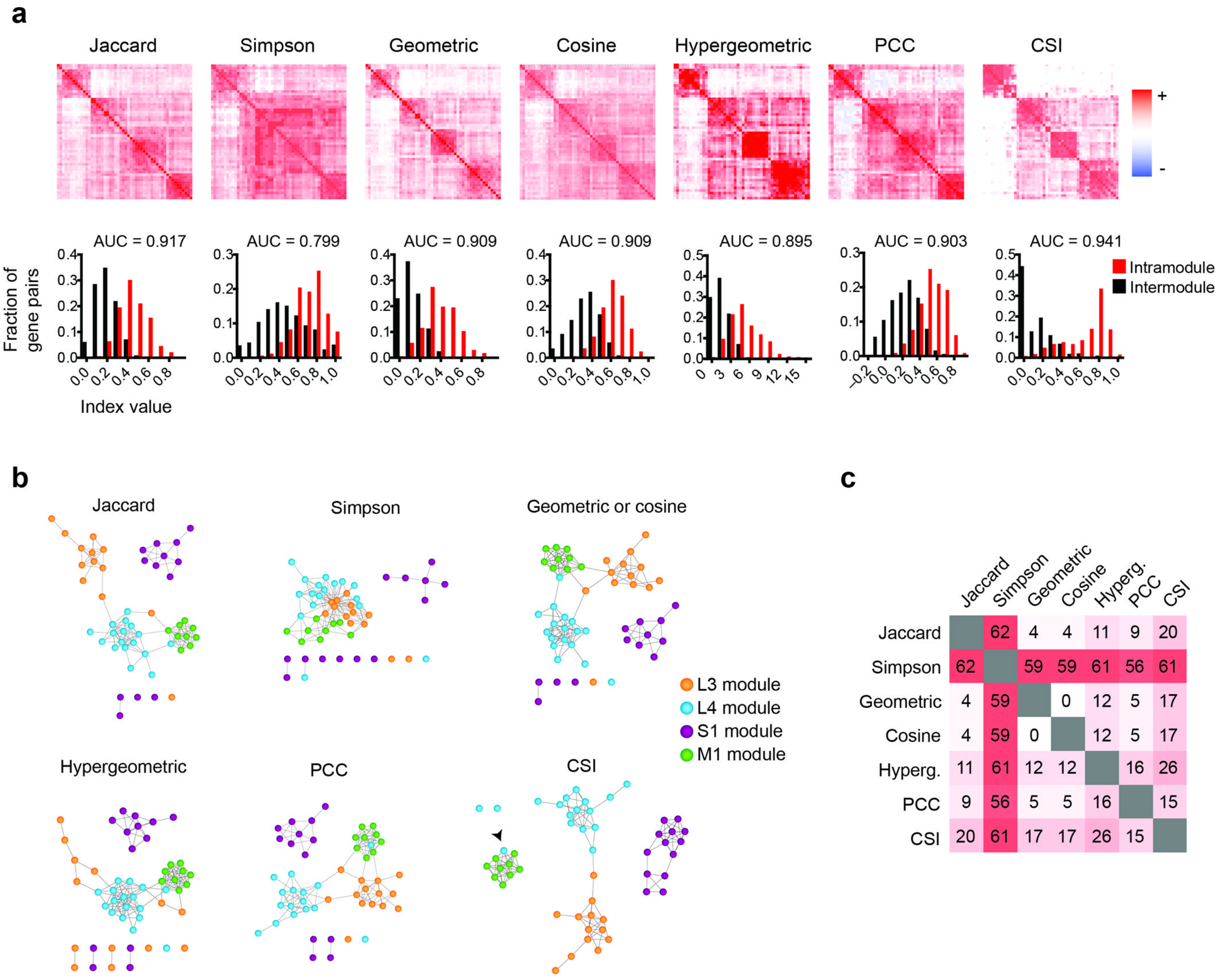
A	B
C1	G1
C1	G5
C1	G6
C2	G1
C2	G2
C2	G3
C2	G8
C3	G1
C3	G2
C3	G4

Many types of biological networks are rapidly being mapped. Aside from providing system-level insights into biological processes, these networks can be used to functionally annotate uncharacterized genes based on interaction profile similarities with well-studied genes ("guilt by association").

Numerous different association indices can be used to determine interaction profile similarities, and it can be daunting to select which one is most appropriate for a particular goal

Figure 2.

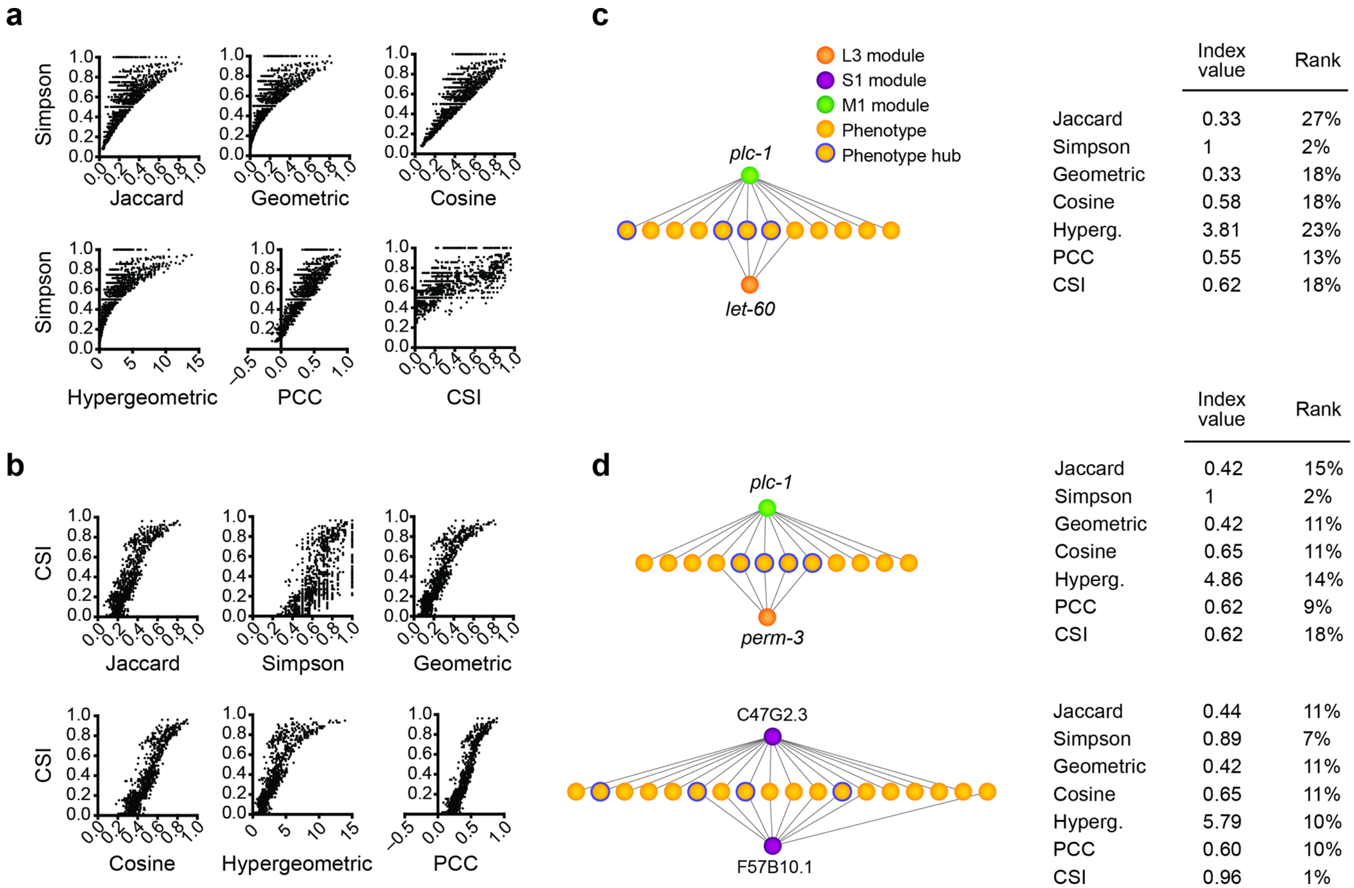
GAIN web tool for the calculation and clustering of association indices. (**a**) Screenshot of GAIN's main window. (**b**) Visualization of a bipartite network as an interaction heatmap or (**c**) as a graph. (**d**) Clustered association index displayed as heatmap and (**e**) association network. (**f**) 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).



	Jaccard	Simpson	Geometric	Cosine	Hyperg.	PCC	CSI
Jaccard		62	4	4	11	9	20
Simpson	62		59	59	61	56	61
Geometric	4	59		0	12	5	17
Cosine	4	59	0		12	5	17
Hyperg.	11	61	12	12		16	26
PCC	9	56	5	5	16		15
CSI	20	61	17	17	26	15	

Figure 3.

Using association indices to identify modules in a gene-to-phenotype network. (**a**) Clustered association index heatmaps for a *C. elegans* gene-to-phenotype network. 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 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**) Association networks were assembled by linking genes that have a top 10% phenotypic profile similarity value. 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**) Pair-wise comparison of the percentage of differences in the edges included in each of the association networks.



	Index value	Rank
Jaccard	0.33	27%
Simpson	1	2%
Geometric	0.33	18%
Cosine	0.58	18%
Hyperg.	3.81	23%
PCC	0.55	13%
CSI	0.62	18%

	Index value	Rank
Jaccard	0.42	15%
Simpson	1	2%
Geometric	0.42	11%
Cosine	0.65	11%
Hyperg.	4.86	14%
PCC	0.62	9%
CSI	0.62	18%

Jaccard	0.44	11%
Simpson	0.89	7%
Geometric	0.42	11%
Cosine	0.65	11%
Hyperg.	5.79	10%
PCC	0.60	10%
CSI	0.96	1%

Figure 4.

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

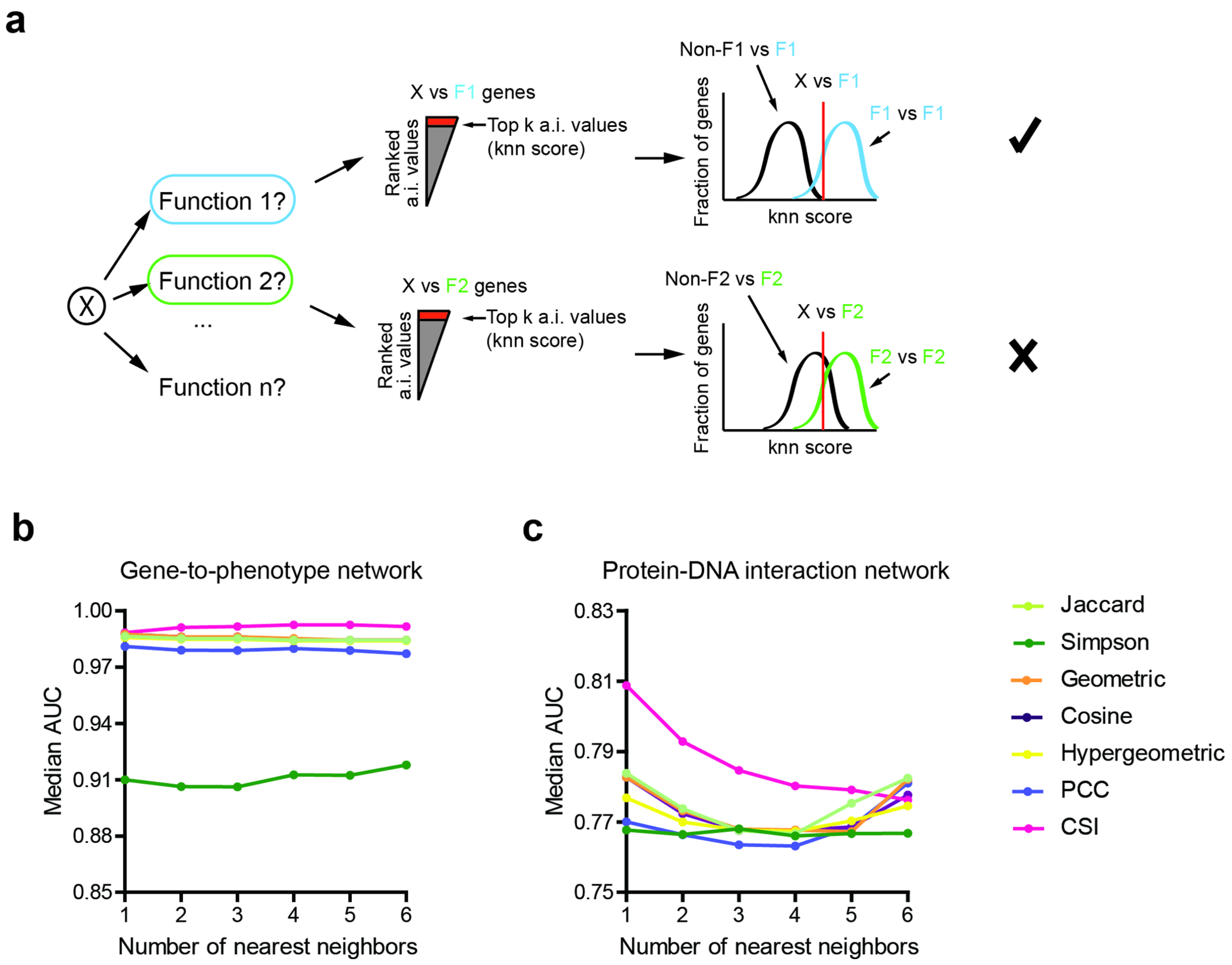


Figure 5.

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 if 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 with that function (blue and green) and for genes that do not have that function (black curves). To assign a function to gene X, the knn scores for genes that have that function and those that do not should be well separated. This separation was determined by calculating the AUC. 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. The panel illustrates a case in which gene X can be assigned function 1 (F1, blue) but not function 2 (F2, green). (**b**) The median AUC calculated for the four functional classes in the *C. elegans* gene-to-phenotype network was determined for each value of k = 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 each value of k = 1 to 6 (number of nearest neighbors).

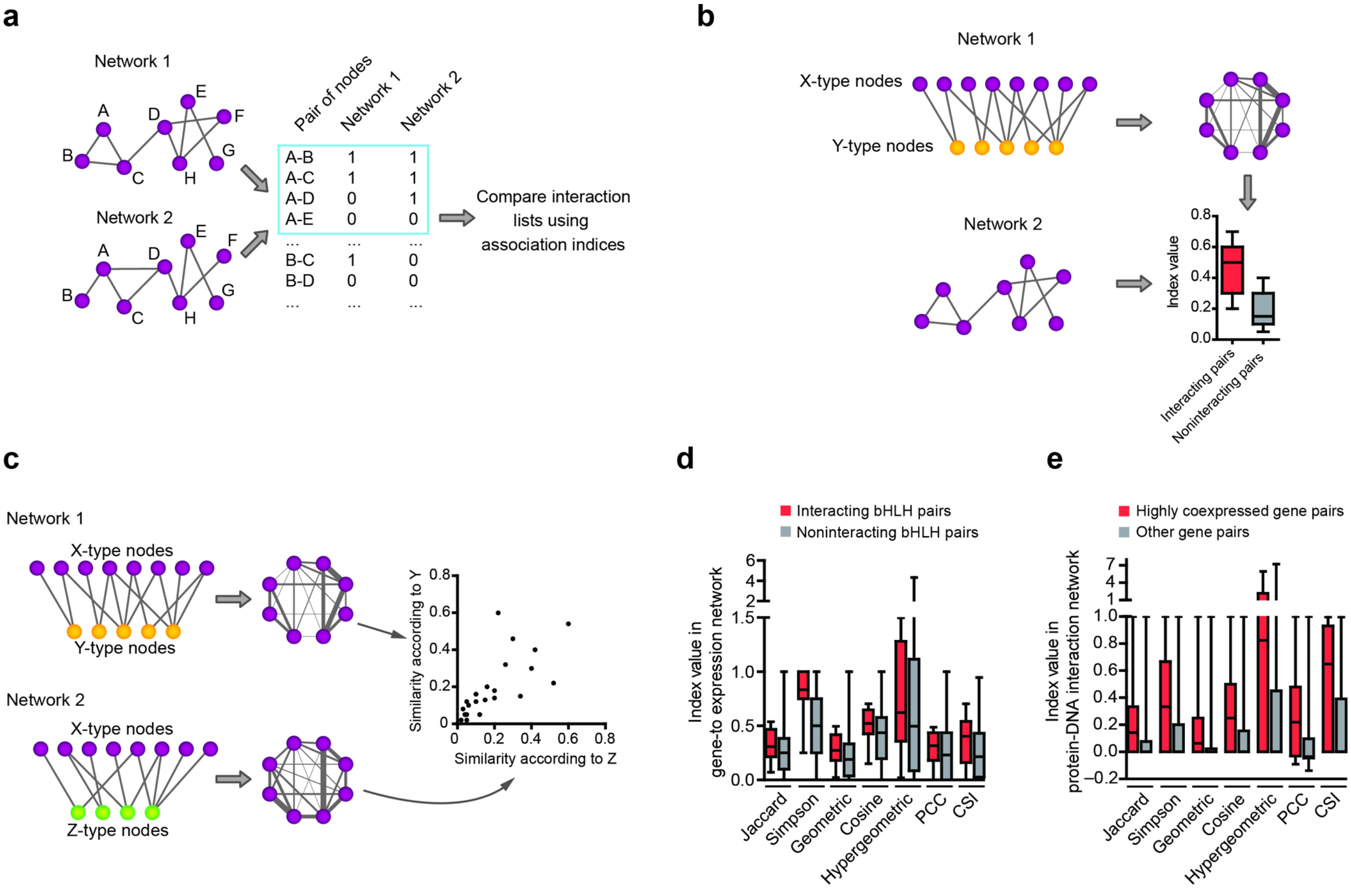


Figure 6.

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**) Comparing interaction profile similarity values in Network 1 between interacting and noninteracting nodes in Network 2. (**c**) The interaction profile similarity of the X-type nodes in Network 1 can be compared to the interaction profile similarity of the same nodes in Network 2. Edge width in the association network indicates is proportional to the association index value. (**d**) The association index values of *C. elegans* bHLH TFs 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 line inside the box indicates the median value and the end of the whiskers indicate the minimum and maximum values. (**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 (red) and other gene pairs (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 nodepairs[a]	**	***	**	**	*	***	*
Determining significance of overlap	no	No	no	no	yes	no	no

Stars indicate qualitative strengths, with a greater number indicating an increase in utility for that application.

[a] Assessment depends on biological question/objective.

See main text for details.