Journal of Clinical and Translational Science

www.cambridge.org/cts

Identifying large-scale interaction atlases using probabilistic graphs and external knowledge

Sree K. Chanumolu and Hasan H. Otu

Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

Translational Research, Design and Analysis Research Article

Cite this article: Chanumolu SK and Otu HH. Identifying large-scale interaction atlases using probabilistic graphs and external knowledge. *Journal of Clinical and Translational Science* **6**: e27, 1–10. doi: 10.1017/cts.2022.18

Received: 27 September 2021
Revised: 29 December 2021
Accepted: 7 February 2022

Keywords:
Interactome; atlas; gene interaction network; external knowledge; Bayesian networks

Address for correspondence:
H. H. Otu, PhD, Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA.
Email: hotu2@unl.edu

Abstract

Introduction: Reconstruction of gene interaction networks from experimental data provides a deep understanding of the underlying biological mechanisms. The noisy nature of the data and the large size of the network make this a very challenging task. Complex approaches handle the stochastic nature of the data but can only do this for small networks; simpler, linear models generate large networks but with less reliability. *Methods:* We propose a divide-and-conquer approach using probabilistic graph representations and external knowledge. We cluster the experimental data and learn an interaction network for each cluster, which are merged using the interaction network for the representative genes selected for each cluster. *Results:* We generated an interaction atlas for 337 human pathways yielding a network of 11,454 genes with 17,777 edges. Simulated gene expression data from this atlas formed the basis for reconstruction. Based on the area under the curve of the precision-recall curve, the proposed approach outperformed the baseline (random classifier) by ~15-fold and conventional methods by ~5–17-fold. The performance of the proposed workflow is significantly linked to the accuracy of the clustering step that tries to identify the modularity of the underlying biological mechanisms. *Conclusions:* We provide an interaction atlas generation workflow optimizing the algorithm/parameter selection. The proposed approach integrates external knowledge in the reconstruction of the interactome using probabilistic graphs. Network characterization and understanding long-range effects in interaction atlases provide means for comparative analysis with implications in biomarker discovery and therapeutic approaches. The proposed workflow is freely available at http://otulab.unl.edu/atlas.

Introduction

Individual elements of a biological system work in concert at the molecular level, which is best analyzed and explained within the context of networks. Networks involving all direct and indirect interactions between genes and/or gene products (the interactome) can be used to understand biological pathways and disease mechanisms. Such an understanding and tools for in silico manipulation lead to new innovative, noninvasive, cost-effective, and scalable approaches to combat human disease by providing means to manipulate and model molecular mechanisms in an efficient and effective way [1].

Conventional methods for interaction network construction used correlation [2-5] or mutual information [6-9]-based measures. Although dependent- or coexpression may lead to functional similarity [10,11], these approaches produced bulky networks and were based on pairwise associations only [12-15] ignoring higher-level associations that may not be inferred by strong individual, paired associations. More complex methods emerged using Bayesian networks (BN) [16-23], Gaussian graphical models – simultaneous equation models [24-26], state space models [27-29], machine learning [30,31], and statistical methods [32,33]. These methods overcome the limited view of the pairwise approaches that are generally based on linear associations by providing a probabilistic blanket of dependency and coregulation, modeling nonlinear associations. However, they can learn networks for only a limited number of nodes [34] because of their high computational complexity. Indeed, biological systems operate at scales much larger than the network sizes handled by these approaches [35,36], which must resort to dissecting the system into pathways. To perform true system-level analysis, there is a need for tools that infer interaction networks, or interaction atlases, at levels beyond the current pathway views.

While the ever-increasing biological data production in the fields of genomics, transcriptomics, and proteomics has resulted in a plethora of approaches attempting to recover interaction networks in biological systems [37,38], they have not always made efficient use of the vast amount of annotated data available [39]. This valuable resource can be used in a systematic way to guide methods that learn interaction networks [40]. Traditional methods for gene interaction (GI) network construction use linear measures and generate large interactomes that miss nonlinear relationships and ignore external knowledge [4,8,41,42]. Another group of methods either uses a single external knowledge source or requires the exact gold-standard network for the one to be generated [43-45], which are impractical approaches in real settings. A third group

© The Author(s), 2022. Published by Cambridge University Press on behalf of The Association for Clinical and Translational Science. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

of approaches focuses on a limited view of interactions (e.g., finding regulatory elements only, or identifying functional associations as interactions) or uses deterministic mechanisms for external knowledge incorporation [46-48]. These last two groups of methods perform well only for a few hundred to one- or two-thousand genes due to algorithm complexity.

In this paper, we propose a method that uses a diverse set of knowledge bases to infer interaction between two genes based on a stochastic, automated framework. Our approach fuses this information with experimental data in a probabilistic graph representation to generate a large-scale interactome (a few tens of thousands of genes). The proposed approach provides a higher system-level view to understand the biological mechanisms in health and disease. We see limited efforts in this direction using linear models or models that do not incorporate external knowledge in building large-scale networks [49,50]. Although these are helpful, there is a need for advanced computational methods that make use of the existing interaction knowledge that is external to the experimental data.

The algorithm described in this paper uses a novel divide-and-conquer approach to construct interaction atlases. Instead of learning the entire large interaction atlas in one shot, we first divide the nodes into groups based on their expression profiles. The interaction network within each cluster is learned using both experimental data and external knowledge. One representative gene in each cluster is selected to build the network between the cluster representatives. This forms a "network of clusters" and is used to merge clusters that are linked together by building an interaction network using the union of the nodes in the two clusters. The ensemble of the links after the "merge" process yields the final atlas.

The proposed method is distinct from existing approaches that dissect biological networks, such as the module network representation [51], which constrains the nodes in a *module* to having the same parents or tree-based methods [52], which model recovery as a feature selection problem based on ranked lists obtained from regression analysis or the stochastic block model-based approaches where families of distributions are defined for the nodes, resulting in unscalable node classification and parameter estimation problems [53,54]. We establish the proposed method using optimized clustering, structure learning, external knowledge incorporation, representative gene selection, and cluster merging processes. The utility of our approach is demonstrated with simulated data and compared to correlation and information theory-based approaches that are used for large-scale interaction atlas generation.

Materials and Methods

The proposed atlas generation method is shown in Fig. 1. In what follows, we describe each module in Fig. 1 in detail.

Experimental Data

To obtain simulated data that follow an interaction network representing true mechanisms, we considered the human pathways found in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database [55]. KEGG pathways involve gene products, compounds, maps (a.k.a. pathways), DNA, RNA, and other molecules. In total, we obtained 337 human pathways, and for each pathway, we deduced all of the direct and indirect gene-GIs using KEGG2Net [56]. We analyzed the KGML files for these pathways to extract the map (pathway) entries that the genes are directly or indirectly linked to. We merged the KEGG2Net GI networks obtained for each pathway following the direct/indirect links

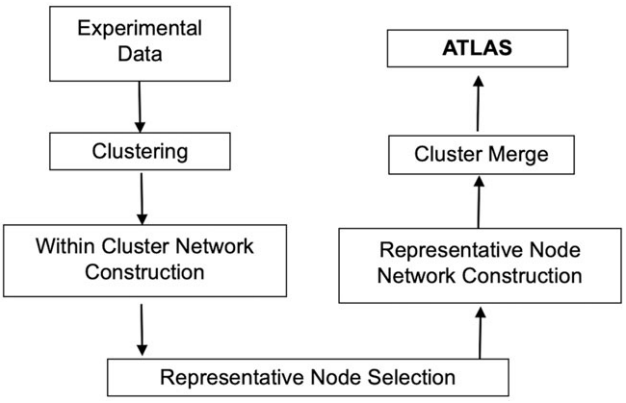


Fig. 1. Workflow for atlas generation.

between the gene and map entries in the KEGG pathways. For example, let the set G_{XY} represent all of the genes in pathway X that has a direct/indirect link to the map (pathway) Y represented as a "node" in pathway X. Similarly, let the set of genes G_{YX} represent all of the genes in pathway Y that has a direct/indirect link to the map (pathway) X represented as a "node" in pathway Y. Then, we establish a link between elements of the gene sets G_{XY} and G_{YX}. By applying this merge method, we obtained an interaction atlas that represents all of the direct and indirect interactions between genes represented by all of the human KEGG pathways, which consisted of 11,454 nodes and 17,777 edges. We used SynTRen v 1.2 [57] to generate simulated transcriptomic data for 20 test and 20 control samples for all of the 11,454 genes represented in our interaction atlas.

Clustering

We clustered the gene expression data matrix using hierarchical clustering, k-means, clusternomics – an integrative context-dependent clustering for biomedical datasets [58], EMMIXgene – a mixture model-based approach to cluster microarray expression data [59], gama – a genetic approach to maximize a clustering criteria [60], DIANA – a divisive, not agglomerative, hierarchical clustering, FANNY – a fuzzy clustering approach, and PAM – partitioning around medoids [61]. The implementations were done in R v 3.6.3 using the packages (functions) stats (hclust, cutree, k-means) v 4.2.0, clusternomics v 0.1.1, EMMIXgene v 0.1.3, gama v 1.0.3, cluster (diana, fanny, pam) v 2.1.2. Validation of the clustering results was performed using the biological homogeneity index (BHI) [62], V-measure [63], and adjusted Rand index (ARI) [64]. These metrics were calculated using R packages clValid v 0.7, saber v 0.3.2, and mclust v 5.4.7, respectively.

Incorporation of External Knowledge

As the number of genes in a cluster is expected to be small (tens to a few hundreds), we learned the networks within a cluster using our previously established tool, Bayesian network prior (BNP) [19]. BNP is a construct that learns an interaction network based on external knowledge and experimental data. As part of this paper, we updated our BNP software by updating the evidence matrix BNP uses to infer interaction of two genes. We gathered information from Gene Expression Omnibus [65], KEGG [55], NCI/Nature Pathway Interaction Database [66], Reactome [67], Biological General Repository for Interaction Datasets [68], FunCoup [69], Hetionet [70], HumanNet [71], RegNetwork

https://doi.org/10.1017/cts.2022.18 Published online by Cambridge University Press

[72], STRING [73], and GeneMANIA [74] data sources that imply an interaction between two genes based on different evidence types, for example, Affinity Capture assays, colocalization, two-hybrid experiments, and coexpression.

We represented the interaction information in the form of an "evidence matrix" where the columns were the evidence types, and the rows were the pairs of genes. If a pair of genes was labeled as interacting by a data source based on an evidence type, we placed a "1" in that location, which was otherwise left as a "0." When all the data sources were combined, we obtained an evidence matrix that contained interaction information for 15,725,553 unique pairs of genes. We added a "GI" column to this evidence matrix, and if a pair of genes had two or more evidence types based on which they were known to be interacting, we labeled the GI entry for that pair as "1" and otherwise labeled it as "0."

BNP is a BN representing the dependency structure between "different experimental evidence types that imply GI" and the "event, GI." Using our evidence matrix, we learned BNP with the bnlearn R package v 4.6.1 [75] based on the hill climbing structure learning approach [76] utilizing the Bayesian information criterion score [77]. The consensus network was obtained based on 1000 bootstrapped datasets where model averaging was used to calculate the strength of links between the nodes of BNP. The final BNP graph was obtained by only retaining the edges that have significant strength values [78]. Therefore, BNP is itself a BN with one node representing "GI" and the remaining nodes representing "different evidence types." BNP reflects the distilled representation of acquired scientific knowledge and can be used to calculate the probability of interaction between two genes using a fusion of external and experimental data. The updated version of the BNP used for this paper can be found at http://otulab.unl.edu/BNP.

Within Cluster Network Construction

Given an expression dataset, the interaction network for the genes in a cluster was calculated using BNP as previously described [19]. BNP uses experimental designs with two groups of samples (e.g., cancer vs. normal) where the set of observations for structure learning is obtained by pairwise comparison of samples in the two groups. As detailed previously [79], this preprocessing step of the expression profiles provides a distribution of expression fold change between the two groups for each node (gene) in the network and has proven to be a reliable and robust way to obtain input data for network learning.

Given two genes, BNP is instantiated with their expression profiles to obtain the value of its GI node that represents their interaction probability based on external knowledge and the supplied experimental data. This probability is calculated for each pair in a set of genes for which an interaction network is to be constructed and incorporated in the structure learning phase to calculate the probability of the candidate graphs. This way the optimum "maximum *a posteriori*" measure is maximized instead of the suboptimum "maximum likelihood" parameter, optimizing the search process as BNP allows for calculation of P(G), the probability of the candidate graph in the search. In the end, the networks learned by BNP represent the GI dynamics for the case under investigation (e.g., cancer) that is used to obtain the experimental data.

Representative Node Selection

We analyzed the GI networks generated for each cluster by BNP using the central informative nodes in network analysis (CINNA) R package v 1.1.54 [80]. Given a network topology, CINNA first identifies the appropriate centrality measures (out of ~50 such measures) for the input network. Next, dimension reduction techniques are used to identify the most informative centrality measure. For each network generated by BNP, we applied CINNA to identify the most informative centrality measure for the network and then used that measure to identify the most central node in the network. We chose these nodes as the representative nodes for their clusters.

Cluster Merge

We sifted the expression values for the representative nodes (genes) of each cluster and used BNP to learn an interaction network for these nodes. Representing each cluster with a gene enabled us to use BNP to build the interaction network of the representative genes. As BNP uses external knowledge to build an interaction network, using a hypothetical gene, for example, the average of all genes in a cluster, or an eigengene in a cluster, would strip BNP off of this feature. The interaction network of representative genes is regarded as a "network of clusters" as each node (gene) represents a cluster. We merged clusters that are linked together by building an interaction network using the union of the genes in the two clusters. The ensemble of the links after the "merge" process yielded the final atlas.

Results

BNP Construction

We assessed the validity of our BNP construction using fivefold cross-validation on the GI inference. At each iteration, we left out 20% of the gene pairs from our evidence matrix and learned BNP as described with the remaining 80% of the data. For the left-out pairs, we instantiated BNP using their evidence vector and inferred the GI node as a probability value. We were able to predict the GI node's state with an area under the curve (AUC) value of 94%. The final BNP model was developed using all of the gene pairs in the evidence matrix as described. We tested BNP on 167 KEGG human pathways that have less than 40 nodes to obtain networks with reasonable complexity. For each pathway, the corresponding GI network and simulated dataset were obtained using KEGG2Net and SynTRen, respectively. The networks learned with BNP and the corresponding conventional structure learning approach that does not use any external knowledge were compared with the true networks. The results, summarized in the Supplementary Data, show that BNP on average attained a 95.52% AUC whereas the structure learning approach that did not utilize external knowledge attained an average AUC of 67.09%. These results demonstrated confidence in the construction and application of BNP to learning GI networks from experimental data using external knowledge. The updated implementation of BNP is freely available at http://otulab.unl.edu/BNP.

Clustering

For each of the eight clustering algorithms, we had six runs where we used the following expected number of genes in a cluster: 3, 6, 12, 25, 45, 70. We used the default values for all other hyperparameters required by the algorithms. Our goal was to sample the neighborhood of the average number of genes in a cluster, ~34, in both extremities as our simulated data involved 11,454 genes from 337 pathways. For each of the 48 clustering results, we

https://doi.org/10.1017/cts.2022.18 Published online by Cambridge University Press

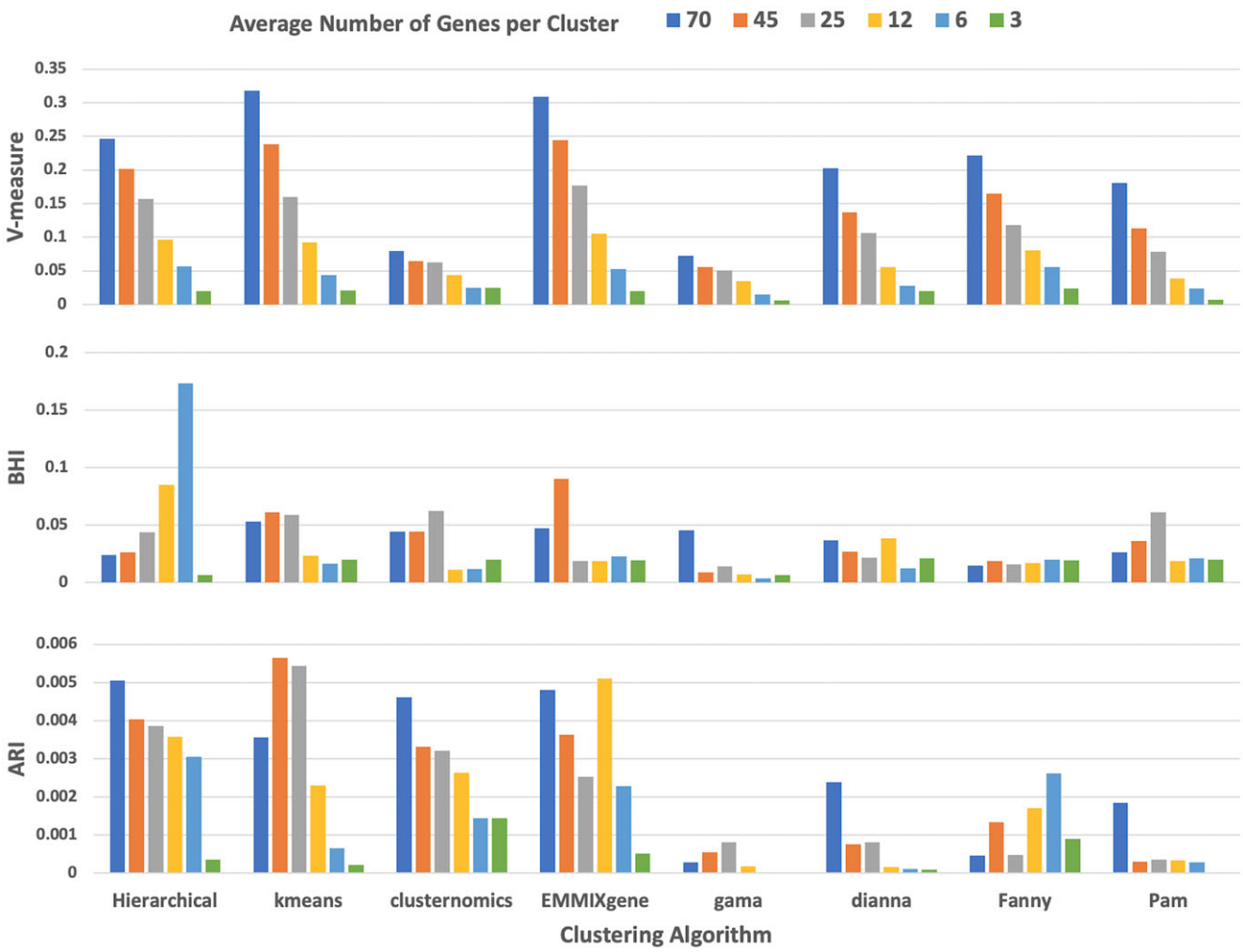


Fig. 2. V-measure, biological homogeneity index (BHI), and adjusted Rand index (ARI) values for the eight clustering algorithms using a range of average number of genes per cluster. The input data are the simulated gene expression data that is generated from the human atlas with 11,454 genes representing 337 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways.

evaluated their performance based on the BHI, V-measure, and ARI. The results are summarized in Fig. 2.

Although V-measure showed a strong correlation with the average number of genes in a cluster, BHI and ARI measures did not render their best performances at 70 average genes per cluster. This was more plausible as the average number of genes in the pathways used to generate the simulated data was ~34. Overall, hierarchical, k-means, and EMMIXgene turned out to be the three top performers when different metrics and average number of genes per cluster were considered, followed by clusternomics and Pam. We continued to experiment with our atlas generating algorithm using the top three clustering methods.

Cluster Merge

We clustered the simulated gene expression data with an expected number of genes per cluster of 25. For each cluster, we learned the GI network using BNP and noted the strength values among the genes within a cluster. The strength values were based on model averaging of 1000 bootstrap datasets. We then picked the representative genes for each cluster and generated the interaction network for the representative genes, again using the BNP approach with 1000 bootstrap datasets. The edges that have significant strength values were retained to determine the final network among the representative genes.

The interaction network among the representative genes was used as a map to guide the cluster merge step. If two representative genes were linked, we combined the genes in the two corresponding clusters and learned an interaction network for this union set using the BNP approach with 1000 bootstrap datasets. This resulted in a strength value between the genes that are in two different clusters. However, if a cluster $\boldsymbol{C}$'s representative node was linked to $\boldsymbol{k}$ other representative nodes, then the genes in this cluster $\boldsymbol{C}$ went through $\boldsymbol{k}$ cluster merge processes. This implied that there were $\boldsymbol{k}$ different strength values calculated for the genes that are in cluster $\boldsymbol{C}$, one for each of the cluster merge processes. We noted all of these $\boldsymbol{k}$ strength values for downstream analysis.

We analyzed merged clusters to see if they contained genes that belonged to the same pathway out of the 337 pathways used to construct the atlas. In 78% of the cases, the two merged clusters contained genes from the same pathway. This implied that using the interaction network of the representative genes to merge clusters enabled us to bring together the genes that belonged to the same pathway. These genes were separated in the clustering phase but now would be combined in the cluster merge phase to better capture the original pathway structure.

https://doi.org/10.1017/cts.2022.18 Published online by Cambridge University Press

Table 1. *Area under the curve of precision-recall curve (AUC of PRC) ($\times 10^{-4}$) values for the atlases generated using the proposed approach based on k-mean, hierarchical, or EMMIXgene clustering algorithms*

Clustering method Link strengths	k-means	Hierarchical	EMMIXgene
First cluster	3.8	5.0	3.0
Minimum	29.3	39.0	13.0
Maximum	24.6	34.6	11.9
Mean	29.9	**39.7**	13.1
Median	26.2	36.3	12.7
Tukey	26.0	36.0	12.5

Strength values for the genes in a cluster were calculated either based on the strength value when only the genes in the cluster are used for network generation (First cluster) or the minimum, maximum, mean, median, and Tukey's bi-weight average of the strength values obtained during the cluster merge process. For a pair of genes within a cluster, there were as many strength values as the number of times the cluster has gone through a merge process with another cluster. Best performing combination is highlighted with boldface and shaded background.

Atlas Generation

We ran the complete workflow for our simulated dataset of 11,454 genes where we used hierarchical, k-means, and EMMIXgene clustering approaches for comparison purposes. Furthermore, the strength values among the genes in a cluster were taken to be either the strength value that is calculated when only the genes in the cluster were used to identify the GI network or the mean, median, minimum, maximum, one-step Tukey's bi-weight average [81] of the k strength values obtained during the cluster merge step. Note that k represents the number of times a cluster goes through a merge process.

The true atlas used to generate the simulated data consisted of 17,777 edges. However, there are 65,591,331 possible edges between the nodes that make up the true interaction atlas. Therefore, there are far fewer "true edges" than there are "false edges." Hence, for any accuracy assessment, in the context of identifying an edge as a "true" or "false" edge, the receiver's operating curve (ROC) approach would not be appropriate. It has been suggested that for imbalanced datasets like this one, where the proportion of the true and false class labels are disproportionate, AUC of the precision-recall curve (AUC of PRC) is a better metric than the AUC of ROC [82].

In Table 1, we list the AUC of PRC values for our proposed atlas generation workflow using the three alternative clustering methods and six different strength value calculations for the genes in a cluster. While the baseline (pure random) performance for the AUC of ROC measure is 0.5, the baseline value for the AUC of PRC measure is the ratio of the true class. In our case, the baseline AUC of PRC measure is $\sim 2.7 \times 10^{-4}$. As seen in Table 1, the hierarchical clustering approach outperformed other clustering methods yielding the highest AUC of PRC regardless of the strength calculation method. Furthermore, using the mean of strength values for a pair of genes within a cluster has consistently resulted in the best AUC of PRC value. Therefore, in our final atlas generation model we adopted this best performing approach (hierarchical + mean). The proposed workflow can be accessed at http://otulab.unl.edu/atlas.

In order to compare our proposed workflow with other approaches, we used two main metrics that are used to generate large-scale interaction networks: correlation and average mutual information. For the proposed approach, we also tried the "perfect clustering" case where the 11,454 genes that make up the atlas were clustered into the 337 pathways that the genes came from. The results, summarized in Table 2, show that the proposed method outperforms conventional metrics in identifying the large interaction atlas.

It is of note that the proposed algorithm with perfect clustering results in an AUC of PRC value that is about 130 times better than hierarchical clustering used in the proposed workflow. Although the proposed approach outperforms existing methods even with hierarchical clustering, there is still room for improvement in the clustering phase. The better the clustering results approximate the underlying biological organization, the more the generated atlas becomes similar to the true interactome.

Clustering Effect on Learning

To understand the effect of clustering on learning, we generated 10 subnetworks from our large, true interaction atlas obtained from KEGG, each containing approximately 500 nodes. The subnetwork size was chosen such that it is not too large for the BNP approach (no clustering) to handle; and it is also large enough to justify the atlas approach (clustering) for network learning. We summarize the results in Table 3.

Our results indicate that both BNP and atlas approaches achieve AUC of PRC values well above the pure random performance (baseline). The effect of clustering represented by the atlas approach renders about an 85% reduction in the overall performance. Furthermore, the average number of edges in the networks learned by the atlas approach was about 10% more than those learned by the BNP approach. This is potentially due to the cluster merge steps involved in the atlas approach that is likely to add more edges. Nevertheless, the performance obtained by our atlas approach significantly exceeds that of the baseline's and is within a reasonable distance of BNP's performance, which is promising, considering its ability to handle very large networks that are otherwise impossible to reconstruct without clustering.

Application to Real Expression Data

We applied the proposed workflow to our previously established renal cell cancer (RCC) gene expression data that contained 23 normal and 32 clear-cell RCC samples [83]. We focused on 10 pathways, listed in Table 4, that have common genes and have been found to be associated with RCC using an experimental proteomic-based approach [84]. We had analyzed our expression dataset along with six other RCC datasets to infer active pathways using a Bayesian pathway analysis and these 10 pathways were found to be regulated [79].

We constructed a "mini-atlas" with 213 genes and 892 edges by merging the 10 overlapping pathways listed in Table 4. Based on the expression of these 213 genes from our RCC dataset, we reconstructed the test atlas using the proposed workflow. Our approach generated nine clusters during the first iteration. After a network is learned for each cluster and clusters were merged using the representative GI map, we ended up with a network of 978 edges. The AUC of PRC for the learned network was 67.8×10^{-3}, which was ~17-fold better than the baseline AUC value of 4.0×10^{-3}.

To demonstrate the utility of our approach, we focused on a subnetwork of the reconstructed atlas that contained genes from two networks: hsa00010 (Glycolysis / Gluconeogenesis) and hsa00330 (Arginine and proline metabolism). These two KEGG pathways have one gene, aldehyde dehydrogenase, in common.

https://doi.org/10.1017/cts.2022.18 Published online by Cambridge University Press

Table 2. *Area under the curve of precision-recall curve (AUC of PRC) values with 95% confidence interval for the atlases generated using the correlation and average mutual information (AMI) metrics compared with the proposed approach based on hierarchical clustering and perfect clustering of expression data*

	Correlation	Hierarchical (proposed)	AMI	Perfect clustering (proposed)
AUC of PRC ($\times 10^{-4}$)	7.9 [7.6–8.2]	39.7 [37.8–41.6]	2.4 [2.2–2.6]	5116 [4895–5337]

Table 3. *Subnetwork statistics and the area under the curve of precision-recall curve (AUC of PRC) values for the learned networks using the Bayesian network prior (BNP) and atlas approaches*

				AUC of PRC ($\times 10^{-3}$)		
Subnetwork	No. of nodes	No. of edges	No. of pathways involved	BNP	Atlas	Baseline
1	528	879	11	129.6	112.8	6.3
2	542	840	13	109.8	95.5	5.7
3	496	913	13	113.9	102.5	7.4
4	514	911	8	112.2	96.5	6.9
5	537	850	11	74	59.9	5.9
6	526	897	13	124.8	109.8	6.5
7	494	837	11	124.2	105.6	6.9
8	459	845	14	83.2	69.9	8.0
9	508	912	8	108.8	87.0	7.1
10	462	816	8	79.3	63.4	7.7
Average	507	870	11.0	106.0	90.3	6.8
St. dev.	29.05	36.57	2.31	20.06	19.45	0.90

Each subnetwork was chosen to have ~500 nodes.

Table 4. *List of pathways used to generate the "mini-atlas" for testing the proposed workflow*

KEGG pathway ID	Pathway name	No. of nodes	No. of edges
hsa00010	Glycolysis/Gluconeogenesis	28	78
hsa00020	Citrate cycle (TCA cycle)	16	40
hsa00030	Pentose phosphate pathway	20	61
hsa00061	Fatty acid biosynthesis	12	35
hsa00230	Purine metabolism	48	277
hsa00280	Valine, leucine, and isoleucine degradation	33	116
hsa00330	Arginine and proline metabolism	31	64
hsa00562	Inositol phosphate metabolism	27	75
hsa00620	Pyruvate metabolism	21	51
hsa00640	Propanoate metabolism	20	53

KEGG, Kyoto Encyclopedia of Genes and Genomes.

The subnetwork shown in Fig. 3 consists of 15 nodes and 18 edges. Our method correctly identified 14 edges that exist in these pathways (true positives), missed two edges (false negatives), and suggested two new edges (false positives) that do not exist in the true KEGG pathways.

The subnetwork shown in Fig. 3 demonstrates two important utilities of the proposed method. Firstly, the edge between the genes enolase 1 (ENO1) and dihydrolipoamide S-acetyltransferase (DLAT) do not exist in the KEGG pathway hsa00010 but is suggested by our method as a putative interaction. Indeed, there are studies that show significant association between these genes [85,86]. Therefore, the proposed method can suggest potential interactions that may not exist in current pathways but warrant further study for the given experimental data. Secondly, the edge

https://doi.org/10.1017/cts.2022.18 Published online by Cambridge University Press

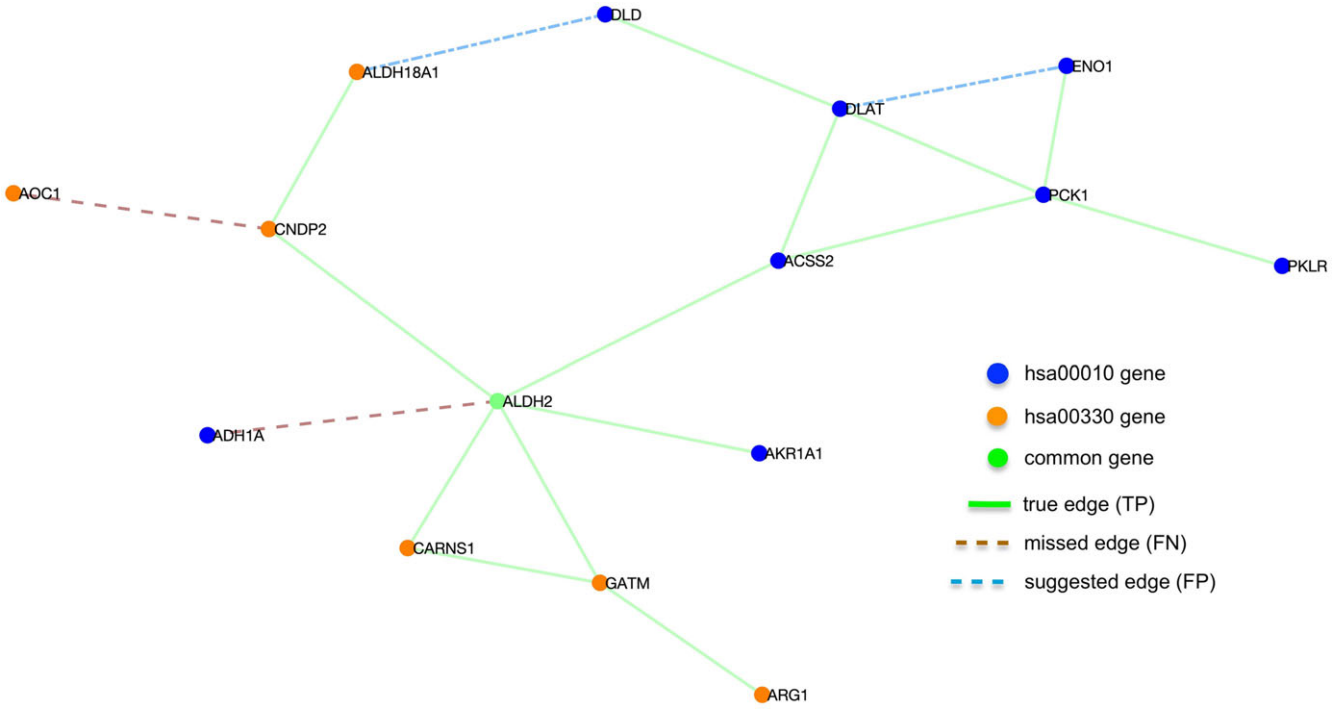


Fig. 3. A subnetwork of the reconstructed test atlas that involves genes from the hsa00010 and hsa00330 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. FN, false negative; FP, false positive; TP, true positive.

between the genes dihydrolipoamide dehydrogenase (DLD) and aldehyde dehydrogenase 18 family member A1 (ALDH18A1) not only suggest a potential interaction that do not exist in the reference databases but also infers an edge that connects two pathways that are otherwise not connected. Like the ENO1–DLAT interaction, there exist studies that imply association between DLD and ALDH18A1 [87,88]. Hence, the proposed method has the potential to merge disconnected pathways and suggest interactions that do not exist in existing pathway databases but may be in play for the given experimental data.

Discussion

Biological systems operate through a networked cascade of events and networks involving all direct and indirect interactions between genes and/or gene products describe the functional workflow of an organism's biological machinery. When building the interactome of an organism, the biological databases that provide a vast amount of annotated data can be used in a systematic way. BN have become increasingly popular as they capture both linear and nonlinear interactions, handle stochastic events in a probabilistic framework accounting for noise, and focus on local interactions, which can be related to causal inference. However, interaction network learning algorithms are computationally very intense and feasible for only a limited number of nodes. Current methods have reached a bottleneck in terms of the size of the reconstructed network. In this paper, we proposed an algorithm to fix this bottleneck by developing a modular approach for reconstructing the entire interaction atlas for a complex organism. We demonstrated the effectiveness of our approach by constructing the interaction atlas for humans.

Our goal was to provide a divide-and-conquer approach that first identified groups of molecules that showed dependency based on experimental data. Within each group, or cluster, the corresponding interaction network is learned using external knowledge via the BNP framework. Each network is summarized using one representative node, all of which are used to build a meta-network representing the interactions between the clusters. The union of the nodes in interacting clusters underwent a second learning phase, and the ensemble of all the learned edges represented the final interaction atlas.

Using the simulated data that represented all the human pathways in the KEGG database, we optimized the proposed method for the clustering approach, cluster size, representative node selection, and cluster merge process. Our optimized parameter selection outperformed existing large-scale interactome generating metrics using the AUC of PRC measure. Our results suggested that the accuracy in the clustering phase of the proposed method dramatically impacted the reliability of the reconstructed interaction atlas. In the process of developing the atlas generation workflow, we also updated the BNP method and both the BNP approach, which can be used to reconstruct small interaction networks (with no clustering), and the atlas generation method described in this paper can be found at http://otulab.unl.edu/BNP and http://otulab.unl.edu/atlas, respectively.

Our current workflow has potential limitations. We can only operate on human transcriptomics and proteomics data for now. To apply the current approach to other organisms, we need to build the corresponding knowledge base that collects interaction information for those organisms based on external databases. We also cannot extend the current approach to other omics, such as metabolomics, or to a multiomics approach where an interaction network that involves different omic types is constructed. However, this can be possible when such a knowledge base, which lists interaction information between different omics, is established. Our current implementation does not identify network characteristics and motifs that result from the identified atlas. We hope to address these issues in our future work. Furthermore, despite bringing the ability to

https://doi.org/10.1017/cts.2022.18 Published online by Cambridge University Press

reconstruct very large interaction networks, clustering diminishes the performance of network learning by about 10–15% (Table 3). Despite our attempts to optimize each algorithmic step shown in Fig. 1, our approach still suffers methodological weaknesses. Primarily, there is room for improvement in hyperparameter selection in the employed clustering approaches as opposed to using the default values. Additionally, the network learning approaches, even though proven to be robust for gene expression data, are not able to produce faithful representations of the entire dependency structure but rather a subset of it.

The ability to generate large interaction atlases provides the means to understand the global characteristics and distant influences in the interactome. Most of the existing methods for interaction network generation focus on local modules at the pathway level, which does not provide the overall cause-and-effect mechanisms. Identifying interaction atlases can also be used to understand the characteristics of the interactome from a network science perspective. Coupled with biological interpretation of the interactions, this provides a tool to perform comparative analyses of disease mechanisms with a potential to lead to biomarker discoveries and/or putative therapeutic approaches.

Supplementary Material. To view supplementary material for this article, please visit https://doi.org/10.1017/cts.2022.7.

Acknowledgements. Research reported in this publication was supported by the National Library of Medicine (NLM) of the National Institutes of Health (NIH) under award number R21LM012759 (SKC, HHO). The funding body was not involved in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Disclosures. The authors declare that they have no conflicts of interest.

Data Availability Statement. The software and data used in this publication are available at http://otulab.unl.edu/atlas.

References

1. **Emmert-Streib F, Glazko GV.** Network biology: a direct approach to study biological function. *Wiley Interdisciplinary Reviews Systems Biology and Medicine* 2011; **3**(4): 379–391. DOI 10.1002/wsbm.134.
2. **Carter SL, Brechbuhler CM, Griffin M, Bond AT.** Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 2004; **20**(14): 2242–2250. DOI 10.1093/bioinformatics/bth234.
3. **D'Haeseleer P, Liang S, Somogyi R.** Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 2000; **16**(8): 707–726.
4. **Horvath S, Dong J.** Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology* 2008; **4**(8): e1000117. DOI 10.1371/journal.pcbi.1000117.
5. **Kumari S, Nie J, Chen HS, *et al.*** Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS One* 2012; **7**(11): e50411. DOI 10.1371/journal.pone.0050411.
6. **Butte AJ, Kohane IS.** Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing* 2000, 418–429.
7. **Daub CO, Steuer R, Selbig J, Kloska S.** Estimating mutual information using B-spline functions: an improved similarity measure for analysing gene expression data. *BMC Bioinformatics* 2004; **5**(1): 118. DOI 10.1186/1471-2105-5-118.
8. **Margolin AA, Nemenman I, Basso K, *et al.*** ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 2006; **7**(S1): 14863. DOI 10.1186/1471-2105-7-S1-S7.
9. **Steuer R, Kurths J, Daub CO, Weise J, Selbig J.** The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* 2002; **18**(Suppl 2): S231–S240.
10. **Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS.** Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences of the United States of America* 2000; **97**(22): 12182–12186. DOI 10.1073/pnas.220392197.
11. **Wolfe CJ, Kohane IS, Butte AJ.** Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics* 2005; **6**(1): 227. DOI 10.1186/1471-2105-6-227.
12. **Filkov V, Skiena S, Zhi J.** Analysis techniques for microarray time-series data. *Journal of Computational Biology* 2002; **9**(2): 317–330. DOI 10.1089/10665270252935485.
13. **Gillis J, Pavlidis P.** Guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology* 2012; **8**(3): e1002444. DOI 10.1371/journal.pcbi.1002444.
14. **Kinney JB, Atwal GS.** Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences of the United States of America* 2014; **111**(9): 3354–3359. DOI 10.1073/pnas.1309933111.
15. **Reshef DN, Reshef YA, Finucane HK, *et al.*** Detecting novel associations in large data sets. *Science* 2011; **334**(6062): 1518–1524. DOI 10.1126/science.1205438.
16. **Djebbari A, Quackenbush J.** Seeded Bayesian networks: constructing genetic networks from microarray data. *BMC Systems Biology* 2008; **2**(1): 57. DOI 10.1186/1752-0509-2-57.
17. **Friedman N, Linial M, Nachman I, Pe'er D.** Using Bayesian networks to analyze expression data. *Journal of Computational Biology* 2000; **7**(3-4): 601–620. DOI 10.1089/106652700750050961.
18. **Hartemink AJ, Gifford DK, Jaakkola TS, Young RA.** Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing* 2001, 422–433.
19. **Isci S, Dogan H, Ozturk C, Otu H.** Bayesian network prior: network analysis of biological data using external knowledge. *Bioinformatics* 2014; **30**(6): 860–867.
20. **Li R, Dudek SM, Kim D, *et al.*** Identification of genetic interaction networks via an evolutionary algorithm evolved Bayesian network. *BioData Mining* 2016; **9**(1): 18. DOI 10.1186/s13040-016-0094-4.
21. **Lo LY, Wong ML, Lee KH, Leung KS.** High-order dynamic Bayesian network learning with hidden common causes for causal gene regulatory network. *BMC Bioinformatics* 2015; **16**(1): 395. DOI 10.1186/s12859-015-0823-6.
22. **Hartemink AJ.** Reverse engineering gene regulatory networks. *Nature Biotechnology* 2005; **23**(5): 554–555. DOI 10.1038/nbt0505-554.
23. **Su C, Andrew A, Karagas MR, Borsuk ME.** Using Bayesian networks to discover relations between genes, environment, and disease. *BioData Mining* 2013; **6**(1): 6.
24. **Kramer N, Schafer J, Boulesteix AL.** Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics* 2009; **10**(1): 384. DOI 10.1186/1471-2105-10-384.
25. **Li H, Gui J.** Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* 2006; **7**(2): 302–317. DOI 10.1093/biostatistics/kxj008.
26. **Wille A, Zimmermann P, Vranova E, *et al.*** Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology* 2004; **5**(11): R92. DOI 10.1186/gb-2004-5-11-r92.
27. **Hasegawa T, Yamaguchi R, Nagasaki M, Miyano S, Imoto S.** Inference of gene regulatory networks incorporating multi-source biological knowledge via a state space model with L1 regularization. *PLoS One* 2014; **9**(8): e105942. DOI 10.1371/journal.pone.0105942.
28. **Hirose O, Yoshida R, Imoto S, *et al.*** Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics* 2008; **24**(7): 932–942. DOI 10.1093/bioinformatics/btm639.
29. **Rangel C, Angus J, Ghahramani Z, *et al.*** Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics* 2004; **20**(9): 1361–1372. DOI 10.1093/bioinformatics/bth093.

30. **Lee WP, Tzou WS.** Computational methods for discovering gene networks from expression data. *Briefings in Bioinformatics* 2009; **10**(4): 408–423. DOI 10.1093/bib/bbp028.
31. **Libbrecht MW, Noble WS.** Machine learning applications in genetics and genomics. *Nature Reviews Genetics* 2015; **16**(6): 321–332. DOI 10.1038/nrg3920.
32. **Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R.** Gene regulatory network inference: data integration in dynamic models: a review. *Biosystems* 2009; **96**(1): 86–103. DOI 10.1016/j.biosystems.2008.12.004.
33. **Lezon TR, Banavar JR, Cieplak M, Maritan A, Fedoroff NV.** Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 2006; **103**(50): 19033–19038. DOI 10.1073/pnas.0609152103.
34. **Gogoshin G, Boerwinkle E, Rodin AS.** New algorithm and software (BNOmics) for inferring and visualizing Bayesian networks from heterogeneous big biological and genetic data. *Journal of Computational Biology* 2017; **24**(4): 340–356. DOI 10.1089/cmb.2016.0100.
35. **Ballerstein K, Haus UU, Lindquist JA, Beyer T, Schraven B, Weismantel R.** Discrete, qualitative models of interaction networks. *Frontiers in Bioscience* 2013; **5**(1): 149–166.
36. **Barabasi AL, Gulbahce N, Loscalzo J.** Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* 2011; **12**(1): 56–68. DOI 10.1038/nrg2918.
37. **Le Novere N.** Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics* 2015; **16**(3): 146–158. DOI 10.1038/nrg3885.
38. **Wang YX, Huang H.** Review on statistical methods for gene network reconstruction using expression data. *Journal of Theoretical Biology* 2014; **362**(6): 53–61. DOI 10.1016/j.jtbi.2014.03.040.
39. **Rigden DJ, Fernandez-Suarez XM, Galperin MY.** The 2016 database issue of nucleic acids research and an updated molecular biology database collection. *Nucleic Acids Research* 2016; **44**(D1): D1–D6. DOI 10.1093/nar/gkv1356.
40. **Ideker T, Dutkowski J, Hood L.** Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* 2011; **144**(6): 860–863. DOI 10.1016/j.cell.2011.03.007.
41. **Peng J, Wang P, Zhou N, Zhu J.** Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* 2009; **104**(486): 735–746. DOI 10.1198/jasa.2009.0126.
42. **Schafer J, Strimmer K.** An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 2005; **21**(6): 754–764. DOI 10.1093/bioinformatics/bti062.
43. **Ghanbari M, Lasserre J, Vingron M.** Reconstruction of gene networks using prior knowledge. *BMC Systems Biology* 2015; **9**(1): 84. DOI 10.1186/s12918-015-0233-4.
44. **Kpogbezan GB, van der Vaart AW, van Wieringen WN, Leday GGR, van de Wiel MA.** An empirical Bayes approach to network recovery using external knowledge. *Biometrical Journal* 2017; **59**(5): 932–947. DOI 10.1002/bimj.201600090.
45. **Zuo Y, Cui Y, Yu G, Li R, Ressom HW.** Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO. *BMC Bioinformatics* 2017; **18**(1): 99. DOI 10.1186/s12859-017-1515-1.
46. **Greene CS, Krishnan A, Wong AK, *et al.*** Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics* 2015; **47**(6): 569–576. DOI 10.1038/ng.3259.
47. **Lo K, Raftery AE, Dombek KM, *et al.*** Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. *BMC Systems Biology* 2012; **6**(1): 101. DOI 10.1186/1752-0509-6-101.
48. **Shahdoust M, Pezeshk H, Mahjub H, Sadeghi M.** F-MAP: a Bayesian approach to infer the gene regulatory network using external hints. *PLoS One* 2017; **12**(9): e0184795. DOI 10.1371/journal.pone.0184795.
49. **Foroushani A, Agrahari R, Docking R, *et al.*** Large-scale gene network analysis reveals the significance of extracellular matrix pathway and homeobox genes in acute myeloid leukemia: an introduction to the Pigengene package and its applications. *BMC Medical Genomics* 2017; **10**(1): 16. DOI 10.1186/s12920-017-0253-6.
50. **Leal LG, Lopez C, Lopez-Kleine L.** Construction and comparison of gene co-expression networks shows complex plant immune responses. *Peer Journal* 2014; **2**(Suppl 2): e610. DOI 10.7717/peerj.610.
51. **Segal E, Pe'er D, Regev A, Koller D, Friedman N.** Learning module networks. *Journal of Machine Learning Research* 2005; **6**: 557–588.
52. **Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P.** Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 2010; **5**(9): e12776. DOI 10.1371/journal.pone.0012776.
53. **Airoldi EM, Blei DM, Fienberg SE, Xing EP.** Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* 2008; **9**: 1981–2014.
54. **Karrer B, Newman ME.** Stochastic blockmodels and community structure in networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics* 2011; **83**(1): 1981. DOI 10.1103/PhysRevE.83.016107.
55. **Kanehisa M, Goto S.** KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 2000; **28**(1): 27–30. DOI 10.1093/nar/28.1.27.
56. **Chanumolu SK, Albahrani M, Can H, Otu HH.** KEGG2Net: deducing gene interaction networks and acyclic graphs from KEGG pathways. biological pathways; gene interaction networks; bayesian networks; cycle removal. *EMBnet Journal* 2021; **26**: e949. DOI 10.14806/ej.26.0.949.
57. **Van den Bulcke T, Van Leemput K, Naudts B, *et al.*** SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics* 2006; **7**(1): 43. DOI 10.1186/1471-2105-7-43.
58. **Gabasova E, Reid J, Wernisch L.** Clusternomics: integrative context-dependent clustering for heterogeneous datasets. *PLoS Computational Biology* 2017; **13**(10): e1005781. DOI 10.1371/journal.pcbi.1005781.
59. **McLachlan GJ, Bean RW, Peel D.** A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 2002; **18**(3): 413–422. DOI 10.1093/bioinformatics/18.3.413.
60. **Rodrigues J, Vasconcelos G, Tinós R.** Using gama - a genetic approach to maximize clustering criterion. 2019. (https://cran.r-project.org/web/packages/gama/vignettes/gama.html).
61. **Kaufman L, Rousseeuw PJ.** *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc, 1990.
62. **Datta S, Datta S.** Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics* 2006; **7**(1): 397. DOI 10.1186/1471-2105-7-397.
63. **Rosenberg A, Hirschberg J.** V-measure: a conditional entropy-based external cluster evaluation measure. 2007, 410–420.
64. **Hubert L, Arabie P.** Comparing partitions. *Journal of Classification* 1985; **2**(1): 193–218.
65. **Barrett T, Wilhite SE, Ledoux P, *et al.*** NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research* 2013; **41**: D991–D995. DOI 10.1093/nar/gks1193.
66. **Schaefer CF, Anthony K, Krupa S, *et al.*** PID: the pathway interaction database. *Nucleic Acids Research* 2009; **37**(suppl_1): D674–D679. DOI 10.1093/nar/gkn653.
67. **Vastrik I, D'Eustachio P, Schmidt E, *et al.*** Reactome: a knowledge base of biologic pathways and processes. *Genome Biology* 2007; **8**(3): R39. DOI 10.1186/gb-2007-8-3-r39.
68. **Stark C, Breitkreutz BJ, Chatr-Aryamontri A, *et al.*** The BioGRID interaction database: 2011 update. *Nucleic Acids Research* 2011; **39**: D698–D704. DOI 10.1093/nar/gkq1116.
69. **Ogris C, Guala D, Sonnhammer ELL.** FunCoup 4: new species, data, and visualization. *Nucleic Acids Research* 2018; **46**(D1): D601–D607. DOI 10.1093/nar/gkx1138.
70. **Himmelstein DS, Lizee A, Hessler C, *et al.*** Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* 2017; **6**: 27. DOI 10.7554/eLife.26726.
71. **Hwang S, Kim CY, Yang S, *et al.*** HumanNet v2: human gene networks for disease research. *Nucleic Acids Research* 2019; **47**(D1): D573–D580. DOI 10.1093/nar/gky1126.
72. **Liu ZP, Wu C, Miao H, Wu H.** RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database (Oxford)* 2015; **2015**: bav095. DOI 10.1093/database/bav095.

https://doi.org/10.1017/cts.2022.18 Published online by Cambridge University Press

73. **Szklarczyk D, Gable AL, Lyon D**, ***et al.*** STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* 2019; **47**(D1): D607–D613. DOI 10.1093/nar/gky1131.
74. **Warde-Farley D, Donaldson SL, Comes O**, ***et al.*** The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research* 2010; **38**(suppl_2): W214–W220. DOI 10.1093/nar/gkq537.
75. **Scutari M.** Learning Bayesian networks with the bnlearn R Package. *Journal of Statistical Software* 2010; **35**(3): 1–22. DOI 10.18637/jss.v035.i03.
76. **Neapolitan RE.** *Learning Bayesian Networks.* Prentice Hall, 2004.
77. **Schwarz G.** Estimating the dimension of a model. *Annals of Statistics* 1978; **6**(2): 461–464.
78. **Scutari M, Nagarajan R.** Identifying significant edges in graphical models of molecular networks. *Artificial Intelligence in Medicine* 2013; **57**(3): 207–217. DOI 10.1016/j.artmed.2012.12.006.
79. **Isci S, Ozturk C, Jones J, Otu HH.** Pathway analysis of high-throughput biological data within a Bayesian network framework. *Bioinformatics* 2011; **27**(12): 1667–1674. DOI 10.1093/bioinformatics/btr269.
80. **Ashtiani M, Mirzaie M, Jafari M.** CINNA: an R/CRAN package to decipher central informative nodes in network analysis. *Bioinformatics* 2019; **35**(8): 1436–1437. DOI 10.1093/bioinformatics/bty819.
81. **Hoaglin DC, Mosteller F, Tukey JW.** *Understanding Robust and Exploratory Data Analysis.* John Wiley & Sons, 2000.
82. **Saito T, Rehmsmeier M.** The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; **10**(3): e0118432. DOI 10.1371/journal.pone.0118432.
83. **Jones J, Otu H, Spentzos D**, ***et al.*** Gene signatures of progression and metastasis in renal cell cancer. *Clinical Cancer Research* 2005; **11**(16): 5730–5739. DOI 10.1158/1078-0432.CCR-04-2225.
84. **Perroud B, Lee J, Valkova N**, ***et al.*** Pathway analysis of kidney cancer using proteomics and metabolic profiling. *Molecular Cancer* 2006; **5**(1): 64. DOI 10.1186/1476-4598-5-64.
85. **Villeneuve LM, Purnell PR, Stauch KL, Callen SE, Buch SJ, Fox HS.** HIV-1 transgenic rats display mitochondrial abnormalities consistent with abnormal energy generation and distribution. *Journal of NeuroVirology* 2016; **22**(5): 564–574. DOI 10.1007/s13365-016-0424-9.
86. **Yang J, MacDougall ML, McDowell MT**, ***et al.*** Polyomic profiling reveals significant hepatic metabolic alterations in glucagon-receptor (GCGR) knockout mice: implications on anti-glucagon therapies for diabetes. *BMC Genomics* 2011; **12**(1): 281. DOI 10.1186/1471-2164-12-281.
87. **Mahajan UV, Varma VR, Griswold ME**, ***et al.*** Dysregulation of multiple metabolic networks related to brain transmethylation and polyamine pathways in Alzheimer disease: a targeted metabolomic and transcriptomic study. *PLoS Medicine* 2020; **17**(1): e1003012. DOI 10.1371/journal.pmed.1003012.
88. **Marchese L, Nascimento JF, Damasceno FS, Bringaud F, Michels PAM, Silber AM.** The uptake and metabolism of amino acids, and their unique role in the biology of pathogenic trypanosomatids. *Pathogens* 2018; **7**(2): 36. DOI 10.3390/pathogens7020036.

https://doi.org/10.1017/cts.2022.18 Published online by Cambridge University Press