



# Inferring Causation in Yeast Gene Association Networks With Kernel Logistic Regression

Amira Al-Aamri<sup>1</sup> , Kamal Taha<sup>2</sup>, Maher Maalouf<sup>3</sup>, Andrzej Kudlicki<sup>4</sup>  and Dirar Homouz<sup>5</sup>

<sup>1</sup>Department of Biomedical Engineering, Khalifa University of Science and Technology, Abu Dhabi, UAE. <sup>2</sup>Department of Electrical and Computer Engineering, Khalifa University of Science and Technology, Abu Dhabi, UAE. <sup>3</sup>Research Center of Digital Supply Chain and Operations, Department of Industrial and Systems Engineering, Khalifa University of Science and Technology, Abu Dhabi, UAE. <sup>4</sup>Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX, USA. <sup>5</sup>Department of Physics, Khalifa University of Science and Technology, Abu Dhabi, UAE.

Evolutionary Bioinformatics  
Volume 16: 1–6  
© The Author(s) 2020  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/1176934320920310



**ABSTRACT:** Computational prediction of gene-gene associations is one of the productive directions in the study of bioinformatics. Many tools are developed to infer the relation between genes using different biological data sources. The association of a pair of genes deduced from the analysis of biological data becomes meaningful when it reflects the directionality and the type of reaction between genes. In this work, we follow another method to construct a causal gene co-expression network while identifying transcription factors in each pair of genes using microarray expression data. We adopt a machine learning technique based on a logistic regression model to tackle the sparsity of the network and to improve the quality of the prediction accuracy. The proposed system classifies each pair of genes into either connected or nonconnected class using the data of the correlation between these genes in the whole *Saccharomyces cerevisiae* genome. The accuracy of the classification model in predicting related genes was evaluated using several data sets for the yeast regulatory network. Our system achieves high performance in terms of several statistical measures.

**KEYWORDS:** Bioinformatics, gene co-expression network, transcription factor, predictive model

**RECEIVED:** August 21, 2019. **ACCEPTED:** March 24, 2020.

**TYPE:** Machine Learning Models for Multi-omics Data Integration—Methods and Protocols

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This publication is based on work supported by the Khalifa University of Science and Technology under "Award No. RC2 DSO."

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Dirar Homouz, Department of Physics, Khalifa University of Science and Technology, P.O. Box 127788, Abu Dhabi, UAE. Email: dirar.homouz@ku.ac.ae

## Introduction

Gene expression profiling is a crucial step in identifying regulatory genes, which are genes in control of the functional product of other genes. It is also a step toward the categorization of genes or proteins based on their functionality. Data analysis technologies such as microarray and RNA-Seq are typically used in producing gene expression data. Such techniques infer gene-gene associations deduced from constructed gene networks such as co-expression networks and gene regulatory networks. These networks, in return, extend the use of gene expression data to other utilities such as protein function prediction and Gene Ontology terms cluster analysis.<sup>1–3</sup> Observing co-expressed genes or regulatory genes is fundamental to deepen the understanding of how genes interact to create different proteins in various organisms.<sup>4</sup> Usually, related genes are inferred based on similar expression profiling across multiple samples with different experimental conditions. One notable limitation with expression data analysis technologies is the drop of the co-expression inference accuracy when it is not deduced from enough experimental conditions, leading to increased noise levels in the data. Several approaches have been proposed to use the co-expression data and improve co-expression inference accuracy by increasing the size of data and the number of experimental conditions.<sup>5–7</sup>

Unlike co-expression networks, identifying causal relations based on expression data is a difficult problem, and no reliable,

general methods are known. Most approaches are either using Directed Acyclic Graphs<sup>8</sup> (limited to very small networks only), or using methods derived from Grainger Causation analysis<sup>9</sup> (requiring time-course data sets), or 3-gene interactions.<sup>10</sup> For a review of most important previous attempts to develop a method for inferring causation, see, for example, Glymour et al<sup>11</sup> and Pearl.<sup>12</sup> Causation has also been inferred using joint distribution of variables in some special cases.<sup>13</sup> Higher moments have been successfully used for describing joint probability distribution of variables with dependencies with applications in physics.<sup>14,15</sup>

These approaches provide a rationale that a more general criterion for causal interactions may be formulated, and that such criterion will be conveniently expressed in terms of the moments of the joint distribution of normalized expression,  $\langle x^k y^l \rangle$ . It is expected that such a method will not suffer from the limitations of graph-based or time-course-based approaches. In our formalism, every directed pair of genes is represented by 1 point in the space of moment expansions. To produce a predictor of causal (regulatory) interaction, we explore the parameter space of the moments of joint distribution, to identify regions within the parameter space that are enriched in causally linked pairs. Because the number of known regulations (size of training set) is a very small fraction of the set of all directed pairs of genes, we decided to employ the Rare Event Weighted Kernel Logistic Regression (RE-WKLR)



classification algorithm, that has been shown to perform well when trained on rare events.<sup>16-19</sup>

In this article, we detect regulatory associations for pairs of genes using microarray expression data obtained from a wide range of experimental conditions. Our method aims to construct causal co-expression network for the whole *Saccharomyces cerevisiae* genome. We create a multidimensional space of values to represent each pair of genes and categorize them into either connected or nonconnected class. We use a rare-event algorithm for the classification approach that represents the sparsity of data that reflects all possible pairs of genes. Connected genes were evaluated using repositories for yeast regulatory association network. Our system performs well in terms of high degree accuracy and recall. The rest of the article is structured as follows. In the “Methods” section, we give a general description of our proposed method. We discuss the experimental results in the section “Experimental Results and Discussion.” Finally, we provide conclusions in the last part of the article.

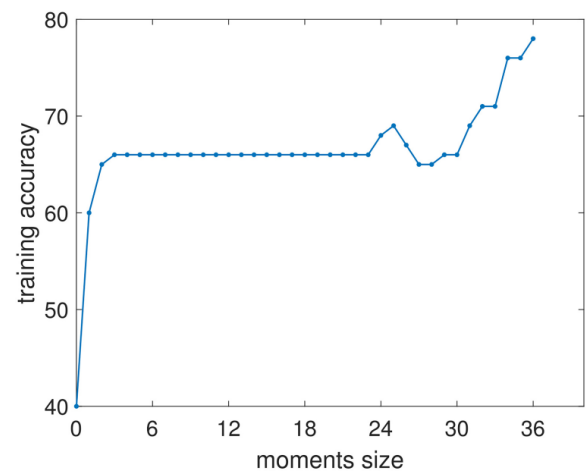
## Methods

We use co-expression microarray data and gene regulatory networks to train a rare-event classification model and categorize each pair of genes into either connected (positive) or nonconnected (negative) class. The pair-wise relations between pairs of genes are measured by calculating the moments of the joint probability distribution of each pair,  $E[x^n y^m]$ , where the moment expansion contains asymmetric terms  $m \neq n$  that are required to infer direction of arrow in regulatory network.

The positive group in this framework denotes pairs of genes with a regulatory association (ie, either gene in the pair is a transcription factor [TF]). The process of the predictive classification model consists of data preparation, training, and testing. Below are the methods and details for each step.

### Data preparation

The Gene Expression Omnibus (GEO) was used to obtain the yeast microarray expression data. The GEO is an online public database that holds various genomics data. It provides tools and downloadable content of selected gene expression profiles and several sets of genomic experiments.<sup>20</sup> We downloaded the microarray data sets for the Affymetrix Yeast Genome S98 Array. The S98 array complies with 9335 probe sets for all known 6400 yeast genes in the complete *Saccharomyces Genome Database* (SGD). Specifically, we use the provided probe set data by the GPL90 platform file that is represented by different experimentation conditions (series) and a total of 1496 samples (the list of GEO accession numbers is provided in the Supplementary Table S1). In many cases, the different probe sets correspond to expression of alternative isoforms, or splicing variants of a gene. Different isoforms may be regulated differently and may have different impact on downstream regulatory pathways; therefore, we find it justified to consider



**Figure 1.** The training accuracy increases as the moments represented by each pair of probes increase.

them separately. We use these data to create a multidimensional space of values for all pairs of the 9335 probes by first normalizing the data samples by dividing by the sample mean on a linear scale. Second, we calculate the moments ( $E[x^n y^m]$ ) of the joint probability distribution of all pairs. We formalize the concept of representing each pair by a moment vector (MV) of all moments in Definition 1:

*Definition 1:*

Let  $P_{ij}$  be a pair of 2 probes and represented by MV  $\cdot E[x^n y^m] \in MV; 0 \leq n \leq 6, 0 \leq m \leq 6$ , both  $i, j \in [0 - 9334]$  (1)

We apply the principal component analysis (PCA) concept to reduce and control the vast resulted multidimensional space of moments. The PCA is a linear feature conversion technique for reducing data dimensionality without compromising the nature of the information. It uses an orthogonal model to transform variables into principal components.<sup>21</sup> The principal components of the moments created new elements in the MV for each pair of probes and reduced them from 49 to 36 moments. That is still quite a large number of moments; however, we show in Figure 1 how the number of moments affects the training accuracy. Next, we analyze the data of pairs, each represented by a vector of variables by feeding the data into a rare-event logistic regression (LR) model.

### Training

To prepare the training data, we first classify the pairs to either connected or nonconnected pairs, based on regulatory association data sets. There are several sources for yeast regulatory networks and TFs such as YEASTRACT,<sup>22</sup> YTRP,<sup>23</sup> RegulatorDB,<sup>24</sup> and Harbison data.<sup>25</sup> In this work, we use a curated database for Yeast Transcriptional Regulatory Pathway (YTRP) to train the pairs of probes.<sup>23</sup> Yeast Transcriptional Regulatory Pathway repository identifies target genes (TG) by

**Table 1.** The “0” indicates a nonconnected class, and “1” shows a connected class.

PROBE PAIRS	MOMENTS VECTOR (MV)	CLASS
$P_{00}$	$\langle m_1, m_2, \dots, m_{36} \rangle$	0
$P_{01}$	$\langle m_1, m_2, \dots, m_{36} \rangle$	1
$P_{02}$	$\langle m_1, m_2, \dots, m_{36} \rangle$	1

employing different TFs’ alteration experiments. The training and testing data were selected from pairs of probes that correspond to different genes. The data in the repository are publicly available for downloading as flat files of regulatory pairs (TF-TG). We downloaded to a local SQL database a total of 213 806 pairs from the YTRP TF-gene direct regulatory network file. Table 1 shows an example of the first 3 pairs in the data fed to the classification algorithm.

*Standardization of gene names.* To maintain data consistency throughout the system processes, we used an annotation resource to get standard gene names for both the probes’ genes and YTRP genes. There are many resources, such as UniProt Knowledgebase (UniProtKB),<sup>26</sup> GeneCards,<sup>27</sup> and GeneMania.<sup>28</sup> In this work, we use UniProtKB due to the easy access of downloadable material and details of all yeast genes’ names that are organized by primary name (official gene symbol), synonyms (all other names), and ordered locus names. We have converted the probes’ gene names and all YTRP genes to the primary name of genes, according to UniProtKB. This process made it innovative and more efficient to match the probe pairs to the YTRP data and classify them to either class (connected) or (nonconnected).

*The zero challenge.* As stated, we use YTRP to classify the probe pairs to either class (0 or 1). It is very straightforward to assign a pair to class “1” as long as it is part of the YTRP data. However, it is a nondeterministic approach to assign a pair of probes to class “0” if it is not found in the YTRP data. It is also very challenging to assign pairs to class “0” as the pair could hold a regulatory association, but not yet updated in the association databases. The framework of our system, that is particularly important as the choice of nonconnected pairs that are used for the training, will have a significant effect on testing and determining the prediction parameters later on. One way to tackle this is to follow a heuristic method that decides on whether a pair belongs to class “0.” An example is to look for pairs in multiple databases of regulatory associations, and if it appears to be missing from many databases, then it could be classified to the nonconnected class. Another way is to look for pairs that are certain to be not connected according to the literature. In this work, we assign a probe pair to class “0” if the pair is among the top least correlated pairs. The correlation of each pair is calculated using the moments of joint probability distribution values. In general, there is a research gap regarding

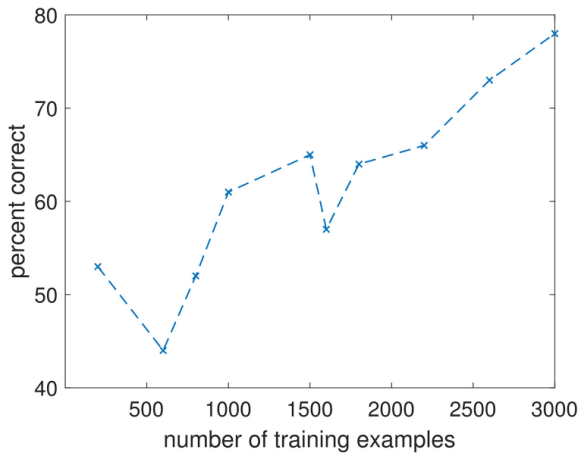
this challenge that limits the ability for nonconnection decision-making.

*Rare-event LR.* The data of MVs and class classification indicator (0 or 1), similar to Table 1, are fed into a RE-WKLR. The RE-WKLR is a rare-event classifier that best characterizes the nature of data in this work.<sup>17</sup> Rare-event classification considers the sparsity of data as the connected pairs of probes are rare compared with the total number of pairs for all 9335 probes. The RE-WKLR is a weighted (W) kernel (K) version of LR, where the weight indicates the proportion of the connected and nonconnected pairs in the data. The kernel reflects the nonlinear alternative of LR and represents the data in a higher dimensional space allowing for a better understanding of the data behavior. The performance of this algorithm was compared with other classification models such as support vector machine in previous work.<sup>17-19</sup> The Gaussian radial basis function kernel is used in this study (see equation (2)). The data set of 12 000 pairs, each represented by 36 moments, is fed to the classification algorithm. The data set is denoted by the kernel matrix  $K=k(X_i, X_j)$ , and  $k_i$  is the  $i$  th row in the matrix that represents 1 pair of probes. The kernel parameter  $\sigma$  indicates the width of the kernel. We also use a regularization variable ( $\lambda$ ) that ensures the prevention of data overfitting, which is also used to calculate the regularized log-likelihood of the classifier (see equation (3)).  $\alpha$  in equation (3) is the dual variable (vector) that also indicates the separation of events and nonevents. It is estimated by maximizing the log-likelihood and then used later on for prediction and testing:

$$k(X_i, X_j) = e^{\left( \frac{-1}{2\sigma^2} \|x_i - x_j\|^2 \right)} \quad (2)$$

$$\ln L_w(\alpha) = \sum_{i=1}^n w_i \ln \left( \frac{e^{y_i k_i \alpha}}{1 + e^{k_i \alpha}} \right) - \frac{\lambda}{2} \alpha^T K \alpha \quad (3)$$

The training data for each pair consist of a feature vector and output class. The elements of the feature vector are calculated from the top 36 PCA components of the moments of joint probability distribution of expression levels. The class assigned to each pair is based on yeast regulatory network database (YTRP). The 6000 pairs with confirmed YTRP connections were selected randomly to represent class “1.” An additional 6000 pairs with no confirmed YTRP connections and with lowest possible expression correlations were selected to represent class “0.” Both  $\sigma$  and  $\lambda$  are user-defined values that are shortlisted based on the classification accuracy. Training the classifier is repeated over several rounds until a maximum probability for classification is obtained. We use the bootstrap technique to resample different variations of the data while tuning the values of  $\sigma$  and  $\lambda$  over several rounds (up to 100 bootstrap rounds). At each round, the pairs are classified into 2 classes: connected (pairs with the regulatory association) and nonconnected (pairs with no regulatory association). The



**Figure 2.** As the number of training examples increases, the percent of correct predicted instances increases.

accuracy of classification is measured at each round and for every tuned value of  $\sigma$  and  $\lambda$ . The ideal parameters  $\sigma$  and  $\lambda$  are the values that yield the maximum training accuracy and the best fit vector  $\alpha$  that is used later on for prediction. Training the data of 12 000 pairs resulted in an accuracy of 78% at  $\sigma = 1.4$  and  $\lambda = 0.1$ .

As part of the training phase, we also plot the learning curve of RE-WKLR to show how the training size affects the testing prediction accuracy. As can be seen from Figure 2, the predictive model RE-WKLR improves the prediction of connected instances as the training size increases.

### Validation testing

The optimal parameters obtained from data training are used to predict a data set of microarray probe pairs by providing the best fit vector  $\alpha$ . As seen in equation (4),  $\alpha$  is used to classify a pair to either class: connected (1) or nonconnected (0) based on the value of multiplying the vector  $\alpha$  by the pair's moments vector/kernel row ( $k_i$ ). Testing the data at this point is important to decide on the validity of the prediction parameters along with the threshold limit we consider in this work (threshold = 0.5):

$$y_i = \begin{cases} 0 & P(y_i | k_i \alpha) \leq 0.5 \\ 1 & P(y_i | k_i \alpha) > 0.5 \end{cases} \quad (4)$$

We test random data of pairs using YTRP as the yeast regulatory network source for evaluation. We conducted a test consisting of 10 000 new random pairs. The accuracy (ACC) is measured by observing the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). A description of each measure is explained below:

- TP is the total of correctly predicted connected pairs (found in YTRP);
- FP is the total of negative instances predicted as connected pairs that are not found in YTRP);

**Table 2.** A confusion matrix for the accuracy measures used for the testing data.

N = 10,000 Actual class	PREDICTED CLASS	
	P	N
P	TP = 4998	FN = 2
N	FP = 750	TN = 4250

Abbreviations: FN, false negatives; FP, false positives; TN, true negatives; TP, true positives.

- TN is the number of correctly predicted nonconnected pairs (not found in YTRP and belongs to the 6000 zeros used for training);
- FN corresponds to the number of incorrectly predicted negative pairs (not found in YTRP and not part of the 6000 zeros used for training).

The value of each measure is indicated in Table 2. We also measure the positive predictive value (PPV) and negative predictive value (NPV) to diagnose the performance of testing:

$$PPV = \frac{TP}{TP + FP} = 0.869(86.9\%)$$

$$NPV = \frac{TN}{TN + FN} = 0.999(99.9\%)$$

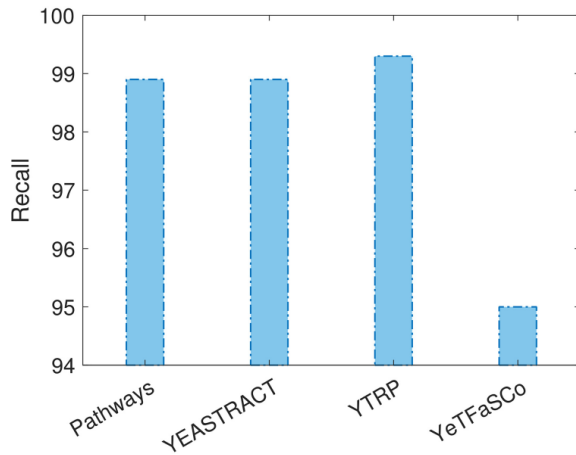
$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} = 0.9248(92.5\%)$$

Justification = centering, margin = 0.6 cm

### Results

We implemented a system that analyzes a multidimensional data of microarray yeast probes and their joint probability distribution moments to identify connected genes and to construct a yeast co-expression network while detecting TFs. This system was executed in C# and Java running on Intel(R) Core i7 processor, with a CPU of 3.4 GHz and 16 GB RAM. The classification model was performed using MATLAB version 2018b. We identified the total number of unique probe pairs using  $(x(x-1))/2$  where  $x = 9335$ . We observed the data for the 43 566 445 probes interactions (pairs) and classified them into connected and nonconnected classes. We trained and tested these interactions using YTRP as the source for yeast regulatory network data. The data used for training and testing are available through <http://ecesrvr.kustar.ac.ae:80/yeast/>.

The accuracy of identifying related probes was evaluated using different repositories for gene regulatory networks. We used databases that target the *S cerevisiae* genome. The evaluation criteria are essentially based on the observation of true positives and false negatives to estimate the sensitivity (recall) of data using equation (5). The computation of the precision of data in this work is omitted because of the fact that computing precision depends on false positives. As stated previously in the zero challenge section and



**Figure 3.** The recall accuracy calculated for different regulatory network databases and using a transcription factors repository (YeTFaSCo). YeTFaSCo indicates Yeast Transcription Factor Specificity Compendium.

the testing section, a true nonconnected pair is hard to confirm, and hence it is not practical to identify false positives:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (5)$$

We conduct mainly 2 main experimental tests: (1) using regulatory network databases and (2) using the TFs database.

#### 1. Using regulatory network databases:

- a. YEASTRACT<sup>22</sup>: We use a curated repository called the Yeast Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT) to identify connected pairs of probes. The YEASTRACT holds more than 160 000 yeast regulatory associations between transcription regulators (factors) and TG. The information of genes in YEASTRACT is systematically updated from sources such as the SGD,<sup>29</sup> Gene Ontology (GO) consortium,<sup>30</sup> and Regulatory Sequence Analysis (RSA) Tools.<sup>31</sup>
- b. Biochemical pathways<sup>29</sup>: We used the pathway data from the SGD to test for the prediction of our data. The SGD provides different download categories of topics, data, and format. We used the yeast biochemical pathway files from the customized search tool to differentiate the genes responsible for catalyzing the biochemical reaction. We also used the interaction data available to construct the genes pairs. The database holds about 339 405 interactions in pathways.
- c. YTRP<sup>23</sup>: We also evaluate the prediction of YTRP pairs that were not used in either training or testing. Similar to the training and testing phase, we first converted all the genes names in the databases to one standard name using UniProtKB.<sup>26</sup> For each of the benchmarks above, we conducted several tests, each with a sample of at least 2000 pairs, and calculated the recall. The results shown in Figure 3 are the overall average accuracy for all tests with each benchmark.

#### 2. Using TFs database:

- a. YeTFaSCo<sup>32</sup>: The Yeast Transcription Factor Specificity Compendium is a database of yeast TFs. It holds around 1887 yeast TFs and TF specificities in 2 main formats (ie, position frequency matrix and position weight matrix). The website includes several ways to browse, analyze, and download the data. The available data are all motifs, expert-curated sets, expert-selected motifs, GB tracks, and microarray data. We used YeTFaSCo to evaluate the recall of TFs identified by our system. We compute the number of TFs according to YeTFaSCo that appear in random test samples of connected pairs. We report this recall of TFs also in Figure 3.

## Conclusions

In this article, we have presented an approach to detect yeast genes regulatory associations using microarray expression data acquired using samples with many and diverse experimental conditions. The classification algorithm followed in this work is based on a Weighted and Kernel version of Logistic Regression (RE-WKLR). It has been shown that this classifier presents the rarity nature of connected genes and the abundance of nonconnected genes. Each pair of genes was defined as a vector of 36 features, which are the moments of the joint probability distribution of expression levels of the 2 genes. The prediction accuracy for the connected genes was assessed using different benchmarks holding information of yeast regulatory association network. Our system performed well in terms of high degree accuracy and recall. Also, the system identifies well-known TFs using YeTFaSCo as a source of TFs. The prediction of the connected class was evaluated and found to score more than 95% in most of the yeast regulatory association repositories. We plan to extend our work by including the directionality of the regulatory network using multiclass classification. Also, we will use our proposed method in the future on RNA-Seq data.



## Acknowledgements

The authors would like to acknowledge the support provided by the Office of Research Support at Khalifa University.

## Author Contributions

DH and AK conceived of the presented idea. AA, KT, and DH developed the theory and performed the computations. MM provided the classification algorithms. All authors discussed the results and contributed to the final manuscript.

## ORCID iDs

Amira Al-Aamri  <https://orcid.org/0000-0002-5419-1688>  
Andrzej Kudlicki  <https://orcid.org/0000-0001-8158-9600>

## REFERENCES

1. Chen C, Zhang D, Hazbun TR, Zhang M. Inferring gene regulatory networks from a population of yeast segregants. *Sci Rep.* 2019;9:1197.
2. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform.* 2017;19:575-592.

3. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol.* 2007;3:88.
4. Kort EJ, Norton P, Haak P, Berghuis B, Ramirez S, Resau J. Gene expression profiling in veterinary and human medicine: overview of applications and proposed quality control practices. *Vet Pathol.* 2009;46:598-603.
5. Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief Bioinform.* 2011;13:281-291.
6. Saha A, Kim Y, Gewirtz AD, et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* 2017;27:1843-1858.
7. Haque S, Ahmad JS, Clark NM, Williams CM, Sozzani R. Computational prediction of gene regulatory networks in plant growth and development. *Curr Opin Plant Biol.* 2019;47:96-105.
8. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. *Science.* 2005;308:523-529.
9. Granger CW. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica.* 1969;37:424-438.
10. Li X, Zhu M, Brasier AR, Kudlicki AS. Inferring genome-wide functional modulatory network: a case study on NF-KB/RelA transcription factor. *J Comput Biol.* 2015;22:300-312.
11. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. *Front Genet.* 2019;10:524.
12. Pearl J. Causal inference in statistics: an overview. *Stat Surv.* 2009;3:96-146.
13. Dzhafarov E, Kujala JV. The joint distribution criterion and the distance tests for selective probabilistic causality. *Front Psychol.* 2010;1:151.
14. Cieliegl P, Chodorowski MJ, Kiraga M, Strauss MA, Kudlicki A, Bouchet FR. Gaussianity of cosmic velocity fields and linearity of the velocity-gravity relation. *Mon Not R Astron Soc.* 2003;339:641-651.
15. Kudlicki A, Plewa T, Rozyczka M. CPPA—a new hydrodynamical code for cosmological large-scale structure simulations. <https://arxiv.org/pdf/astro-ph/9609037.pdf>. Updated 1996.
16. Maalouf M, Siddiqi M. Weighted logistic regression for large-scale imbalanced and rare events data. *Knowl-Based Syst.* 2014;59:142-148.
17. Maalouf M, Humouz D, Kudlicki A. Robust weighted kernel logistic regression to predict gene-gene regulatory association. Paper presented at: IIE Annual Conference; May 31-June 3, 2014:1356; Montréal, QC, Canada. Peachtree Corners, GA: Institute of Industrial and Systems Engineers (IISE).
18. Maalouf M, Homouz D. Kernel ridge regression using truncated Newton method. *Knowl-Based Syst.* 2014;71:339-344.
19. Maalouf M, Trafalis TB. Robust weighted kernel logistic regression in imbalanced and rare events data. *Comput Stat Data An.* 2011;55:168-183.
20. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30:207-210.
21. Jackson JE. *A User's Guide to Principal Components*, vol. 587. New York, NY: John Wiley & Sons; 2005.
22. Teixeira MC, Monteiro P, Jain P, et al. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 2006;34:D446-D451.
23. Yang TH, Wang CC, Wang YC, Wu WS. YTRP: a repository for yeast transcriptional regulatory pathways. *Database.* 2014;2014:bau014.
24. Choi JA, Wyrick JJ. RegulatorDB: a resource for the analysis of yeast transcriptional regulation. *Database.* 2017;2017:bax058.
25. Harbison CT, Gordon DB, Lee TI, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature.* 2004;431:99.
26. Boutet E, Lieberherr D, Tognolli M, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. In Edwards D, ed. *Plant Bioinformatics*. New York, NY: Humana Press; 2016:23-54.
27. Safran M, Dalah I, Alexander J, et al. GeneCards version 3: the human gene integrator. *Database.* 2010;2010:baq020.
28. 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 Res.* 2010;38:W214-W220.
29. Cherry JM, Hong EL, Amundsen C, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 2011;40:D700-D705.
30. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 2004;32:D258-D261.
31. Thomas-Chollier M, Sand O, Turatsinze JV, et al. RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.* 2008;36:W119-W127.
32. De Boer CG, Hughes TR. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res.* 2011;40:D169-D179.