In-silico Gene Annotation Prediction Using the Co-expression Network Structure

Miguel Romero(✉), Jorge Finke, Mauricio Quimbaya, and Camilo Rocha

Pontificia Universidad Javeriana Cali, Cali, Colombia
{miguel.romero,jfinke,maquimbaya,camilo.rocha}@javerianacali.edu.co

Abstract. Identifying which genes are involved in particular biological processes is relevant to understand the structure and function of a genome. A number of techniques have been proposed that aim to annotate genes, i.e., identify unknown biological associations between biological processes and genes. The ultimate goal of these techniques is to narrow down the search for promising candidates to carry out further studies through in-vivo experiments. This paper presents an approach for the in-silico prediction of functional gene annotations. It uses existing knowledge body of gene annotations of a given genome and the topological properties of its gene co-expression network, to train a supervised machine learning model that is designed to discover unknown annotations. The approach is applied to *Oryza Sativa Japonica* (a variety of rice). Our results show that the topological properties help in obtaining a more precise prediction for annotating genes.

Keywords: Co-expression network · Topological properties · *Oryza Sativa Japonica* · Machine learning · Functional gene annotation

1 Introduction

Available genome data has grown exponentially in the last decade, mainly due to the development of new technologies, including gene expression profiles generated with RNA sequencing [17]. Intuitively, genes are said to co-express whenever they are active simultaneously, indicating that they are associated to the same biological processes. Co-expression networks have been used widely to predict biological information (specific biological functions and processes) based on the interactions of the genes [16,22,24,25]. The working hypothesis of correlated expression implying a relevant biological relationship has resulted in a promising strategy to perform functional genome annotation.

Co-expression networks are generally, represented as undirected, weighted graphs built from empirical data. Vertices denote genes and edges indicate a weighted relationship about their co-expression. Since co-expression networks include all correlated expression patterns between genes, a detailed analysis of the network topology –in addition to node-to-node relationships– may provides

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2019, SCI 882, pp. 802–812, 2020.
https://doi.org/10.1007/978-3-030-36683-4_64

insights into the network structure and organization. An approach based on co-expression networks ultimately provides additional information to build up novel biological hypotheses. It remains an open challenge to develop models that combine ideas from network theory and machine learning, and take advantage of the co-expression network structure for predicting functional gene annotations.

This paper presents an approach for predicting gene annotations based on the topological properties of the gene co-expression network of a given genome. The main idea is to combine the co-expression information available for the genome, the topological properties, and the body of known annotations (experimentally verified). The goal is to predict unknown annotations for genes. By taking advantage of the co-expression network structure, this approach aims to exploit additional information for the prediction that helps to establish strong functional associations between genes and the biological processes in which they are involved.

The proposed approach is showcased to predict gene annotations for the *Oryza Sativa Japonica* species, a variety of rice. The rice co-expression network is built from information available at ATTED-II [4] and a body of annotations gathered from RAP-DB [19]. The supervised machine learning technique XGBoost is used for the prediction of 141 functional gene annotations. For each annotation, a model with and without topological measures are trained. Their performance is compared to identify how topological measures can improve the annotation prediction in terms of precision. The experiments show that there are promising candidates to carry out further studies through in-vivo experiments, i.e., there exists set of genes that are consistently predicted to have a given annotation.

The remainder of the paper is organized as follows. Section 2 presents an overview on gene annotation and techniques used to perform functional genome annotation. Section 3 describes co-expression networks and a network-based approach for predicting gene annotation. Section 4 presents a case study for the *Oryza Sativa Japonica* species. Section 5 draws some conclusions and presents future research directions.

2 Gene Annotation

The goal of gene annotation is to determine the structural organization of a genome and discover sets of gene functions, i.e., the locations of genes and coding regions in a genome that determine what genes do [18,26]. Once a genome is sequenced, it is annotated to understand its structure and how it encodes biological function. Though several organism have been completely sequenced, genome annotation remains a significant challenge, mainly due to its extreme combinatorial nature.

Genome annotation focuses on two complementary processes. First, the genome structure is defined, that is genes are identified and intergenic regions are characterized from specific sequences that are associated with genomic structures (particular promoter motifs or repetitive signatures). Second, putative functions of genes are assigned to establish gene and, as a whole, genome functional characterizations [10]. While genomic structure can be determined by the detection

of specific genomic elements inserted in the sequence itself, genome functional annotation is more laborious. It generally depends on several annotation strategies that combine alignment-based information with experimental evidence associated with gene functional predictions. Often extensive in-vivo experimentation is required to gain certainty on processes associated to genes [28]. The rapid accumulation nowadays of genome-wide data describing both, genome sequences and functional properties of genes, has facilitated the novel development of integrative approaches to target genome annotation.

Global analysis of similarity in gene expression patterns has been used to infer specific regulatory networks by analysis of gene co-expression analysis. Different techniques and tools, mostly supported by statistical inference, have been proposed to suggest putative biological processes to genes whose functional annotation is partially or completely unknown [14,25].

3 Prediction Based on Co-expression Network Structure

Here gene co-expression networks are represented as undirected graphs where each vertex identifies a gene and an edge the level of co-expression between two genes.

Definition 1. *Let V a set of genes, E a set of edges that connect pairs of genes and w a weight function. A* (weighted) gene co-expression network *is a weighted graph $G = (V, E, w : E \to \mathbb{R}_{\geq 0})$.*

The set of genes V in a co-expression network is particular to the genome under study. The correlation of expression profiles between each pair of genes is measured, commonly, with help of the Pearson correlation coefficient. Every pair of genes is assigned and ranked according to a relationship measure, and a threshold is used as a cut-off measure to determine E. The weight function w denotes how strongly co-expressed are each pair of genes in V. For any pair of genes $u, v \in V$, $w(u, v)$ is usually inversely proportional to the measure of *mutual rank* (MR) between genes u and v. Note that a value of 0 would be assigned to the strongest connections [13].

There are gene co-expression network databases containing several expression profiles obtained from cDNA microarrays and RNA sequencing. Each profile indicates how gene expression is perturbed when the subject organism is exposed to multiple types of stress (for example, to biotic and abiotic stresses). The correlation of expression for a set of genes under multiple conditions may suggest their functional relation, thus offering information on how genes can be related in terms of biological function.

In an annotated gene co-expression network, each gene is associated with the collection of biological functions to which it is related (e.g., through in-vivo experiments).

Definition 2. *Let A be a set of biological functions. An* annotated gene co-expression network *is a gene co-expression network $G = (V, E, w)$ complemented with an annotation function $\phi : V \mapsto 2^A$.*

The network-based approach to gene annotation proposed in this paper can now be explained in detail. Given an annotated co-expression network $G = (V, E, w)$ with annotation function ϕ, the goal is to use the information represented by ϕ together with topological properties of G to obtain a function $\psi : S \mapsto 2^A$. Function ψ predicts associations between annotations and genes based on a supervised machine learning technique.

The overall success of this approach is evaluated in two complementary ways. On the one hand, this approach would be successful if higher precision is achieved for a suggestion $a \in \psi(v)$ to annotate gene $v \in V$, when compared to other approaches (e.g., to a suggestion in which only the information represented by ϕ and the edge structure of G are taken into account). On the other hand, this approach would also be successful if genes $v \in V$ are found for which $\psi(v) \backslash \phi(v) \neq \{\}$, meaning that new annotation suggestions have been found for the candidate gene v. This latter situation is desirable in practice to reduce time and costs associated to laboratory experimentation. In particular, for a biological function $a \in \psi(v)$, with $\psi(v) \backslash \phi(v) \neq \{\}$, laboratory experimentation can then increase the focus on specific biotic and abiotic stresses to see if gene v is actually associated to the biological function a in the genome under study.

4 In-silico Experimentation with Oryza Sativa Japonica

This section describes an in-silico experimentation case study of gene annotation prediction for *Oryza Sativa Japonica* (Osa). It follows the network-based approach proposed in Sect. 3, and explains how the gene co-expression network is built, how it is initially annotated, and how –with the help of topological properties– machine learning techniques are used to improve gene annotation.

4.1 The Co-expression Network and Gene Annotations

The co-expression information used in this paper is taken from the ATTED-II database [4,12,15]. The gene co-expression network $G = (V, E, w)$ comprises 19 665 vertices (genes) and 553 125 edges. The weight function $w : E \mapsto \mathbb{R}_{\geq 0}$ measures the mutual rank (MR) between any pair of genes; it assigns smaller values to stronger links. A MR threshold of 100 is used as the cut-off measure for G, i.e., E contains edges e that satisfy $w(e) \leq 100$.

The annotation information for G is taken from the RAP-DB [19] database, a comprehensive set of gene annotations for the genome of rice. Among these annotations, there are 899 for molecular function (i.e., molecular activities of individual gene products), 187 for cellular components (i.e., location of the active gene products), and 633 for biological processes (i.e., pathways to which a gene contributes). It is important to note that genes may be associated to several annotations in each category. Since this work is mainly focused on pathways and large processes, only biological process annotations are consider. The annotation function $\phi : V \to 2^A$ for G associates $|A| = 615$ annotations to $|V'| = 7478$ genes, where $V' \subseteq V$ is the set of genes associated to at least one biological process.

4.2 Topological Properties

Given the co-expression network $G = (V, E, w)$, properties of its network structure are computed for gene annotation prediction. The topological measures considered for each gene u are the following:

- degree: number of edges incident to u;
- eccentricity: maximum shortest distance from u to any vertex in its connected component;
- clustering coefficient: ratio between the number of triangles (3-loops) that pass through u and the maximum number of 3-loops that could pass through it;
- closeness centrality: the reciprocal of the average shortest path length from u;
- betweenness centrality: the amount of control that u has over the interactions of other nodes in the network;
- neighborhood connectivity: the average connectivity of all neighbors of u;
- topological coefficient: the extent to which u shares neighbors with other nodes.

These measures were computed with the help of Cytoscape [21], an open source platform for visualizing and analyzing molecular interaction networks and biological pathways.

4.3 Supervised Training

Two models are trained for predicting gene annotations, one per biological function. Namely, one in which the topological measures of G are used and another one in which they are not. The next paragraphs describe how these models are built, trained, and evaluated.

The dataset summarizes data for the 19 665 genes, 615 annotations, and 7 topological measures. It comprises 19 665 rows and 222 columns. For these experiments, the dataset is heavily imbalanced since 77% of the annotations are related to less than 10 genes each one. In order to counter such an imbalance, two decisions are made. First, only annotations associated with at least 10 genes are considered for prediction, reducing the number of annotations from 615 to 141. Second, the Synthetic Minority Over-sampling TEchnique (SMOTE) is used to over-sample the minority class to potentially improve the performance of a classifier without loss of data [7]. This technique derives the new samples of the minority class from interpolation rather than extrapolation, in order to avoid over-fitting problems.

A supervised machine learning technique for the annotation prediction is used. In particular, the XGBoost implementation of gradient boosted decision trees is used [8]. This technique has recently been dominating applied machine learning competitions for structured or tabular data, and it has implementations in many programming languages, including C++, Java, and Python. In the experiments presented in this section, a Python implementation was used.

Finally, k-fold cross validation is used in the training of the two models with $k = 50$. The number of k is determined for statistical significance in the false positive analysis prediction. The performance of the models is compared using the accuracy, F1-score, and AUC ROC measures.

4.4 Annotation Prediction

Figure 1 presents a summary of the accuracy achieved by the two trained models for predicting gene annotations. In particular, the results are depicted for 32 different annotations. The annotations are sorted in descending order by the

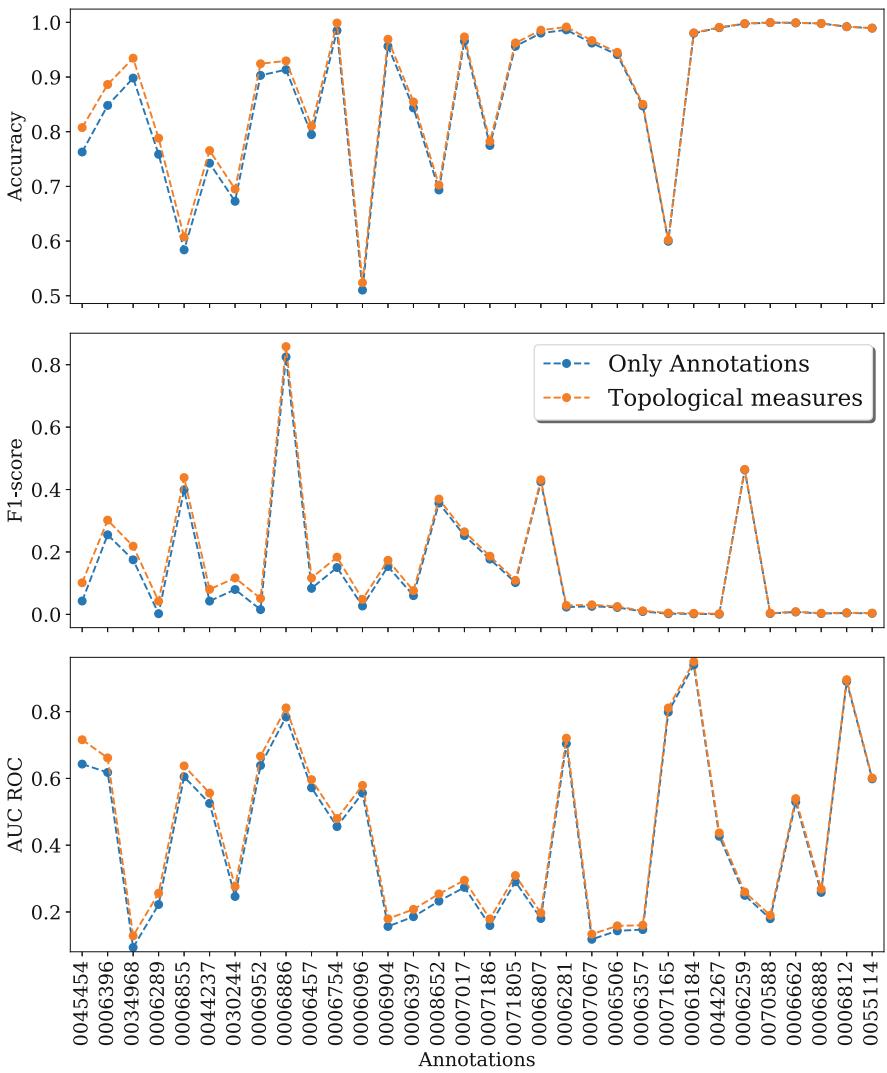

Fig. 1. Performance accuracy, F1-score, and AUC ROC measures for the prediction of 32 annotations with the two trained models (with and without topological measures).

Table 1. Number of genes most frequently annotated as false positives for the 32 annotations by the model trained with topological measures. The 'Max FP' column summarizes the number of times (out of a total of 50) such an annotation is suggested for a gene, while the 'FP' column identifies the number of genes that are consistently given such an annotation.

ID	Biological process	# Genes	Max FP	# FP
0006807	Nitrogen compound metabolic process	15	41	1
0006289	Nucleotide-excision repair	20	46	1
0006397	mRNA processing	17	48	1
0007017	Microtubule-based process	18	49	1
0070588	Calcium ion transmembrane transport	10	36	1
0006184	GTP catabolic process	49	47	1
0044267	Cellular protein metabolic process	25	49	1
0007186	G-protein coupled receptor protein signaling	11	50	1
0006281	DNA repair	62	50	2
0006754	ATP biosynthetic process	24	49	3
0006904	Vesicle docking involved in exocytosis	11	50	4
0055114	Oxidation-reduction process	870	47	5
0006886	Intracellular protein transport	135	50	19
0006855	Drug transmembrane transport	32	50	21
0006662	Glycerol ether metabolic process	28	50	27
0006888	ER to Golgi vesicle-mediated transport	16	50	29
0006259	DNA metabolic process	15	50	32
0007067	Mitosis	11	48	33
0008652	Cellular amino acid biosynthetic process	18	50	52
0030244	Cellulose biosynthetic process	23	50	64
0034968	Histone lysine methylation	11	50	93
0006812	Cation transport	62	50	96
0045454	Cell redox homeostasis	83	49	103
0006506	GPI anchor biosynthetic process	12	50	284
0007165	Signal transduction	104	50	370
0071805	Potassium ion transmembrane transport	24	50	570
0006357	Regulation of transcription from RNA polymera	12	50	1199
0006396	RNA processing	58	50	1212
0044237	Cellular metabolic process	75	50	1318
0006457	Protein folding	162	50	2358
0006952	Defense response	133	50	2679
0006096	Glycolysis	50	50	2875

performance difference between the models. This plot shows that the model trained with the additional information of the topological measures can be more reliable in these cases. The results for the remaining annotations are omitted,

but in these other cases the additional information provided by the network structure does not result in a better prediction performance.

Table 1 presents details about the prediction made by the model trained with the topological data of the co-expression network. The annotations listed in this table correspond to the same 32 annotations included in Fig. 1. For each annotation, the table includes its gene ontology term (ID), its associated biological process, and the total number of genes known to be associated with it in the co-expression network. A false positive analysis is applied to the annotation predictions: the idea is to identify the genes that tend to be classified as a false positive because they are the candidate genes on which lab experimentation can focus on. The 'Max FP' column summarizes the number of times (out of a total of 50) such an annotation is suggested for a gene, while the 'FP' column identifies the number of genes that are consistently given such an annotation.

Note that the annotations in Table 1 are sorted in ascending order by the number of genes most frequently classified as false positive. This set of genes is considerably small for the first 12 annotations of the table and therefore can be seen as good candidates for experimental verification. For example, the only gene associated to the nitrogen compound metabolic process (0006807) is proline dehydrogenase, identified as Os10g0550900, which is related to the functional annotation proline catabolic process to glutamate (0010133).

5 Related Work and Conclusion

Complex network structure has been widely used for the enrichment of analysis techniques from different perspectives and in different domains. A modest summary of the enormous body of work for, mainly, biological predictions is presented next.

The application of networks in biology has grown exceptionally in the last decade due to the large amount of molecular interaction data available [5]. There are two main types of biological networks that are a current focus of research. The first group is of molecular networks. It includes protein interaction networks where proteins are represented as vertices that are connected by physical interactions, metabolic networks where metabolites are vertices connected by co-appearance in biochemical reactions, regulatory networks where connections are regulatory relationships between transcription factors and genes, and RNA networks. The second group is of genetic networks. It includes co-expression networks, in which genes are vertices that are connected by similar expression patterns, such as the ones studied in the present work. This latter group has been used to identify the function of a large set of genes and their role in specific biological processes in different species [5,25]. In particular, the co-expression networks have helped to address the problem of identifying the role of the genes in biological systems. This work takes a step forward in exploiting the rich structural property of a network with the goal of increasing annotation prediction precision.

Studying the link prediction problem is one of the most common applications of the topological properties of networks. Tan et al. [23] examine the role of

network topology in predicting missing links from the perspective of information theory. They present a practical approach based on the mutual information of network structures. Naaman et al. [11] show that edges with similar network topology, as defined by a combination of network measures having similar signs, can be used to predict edge sign based on correlation measures on the network topology.

In biological networks, the topological properties have been used for the prediction of interactions between genes or proteins. Lobato et al. [2,3] use the topology of biological interactomes for the prediction of interactions in biological networks. Santolini et al. [20] use biochemical networks, with experimentally measured kinetic parameters, to predict the impact of perturbation patterns in biological interactome networks. They approximate perturbation patterns using increasingly accurate topological models. Benstead-Hume et al. [6] explore computational approaches to identify genes that have become essential in individual cancer cell lines. They use machine learning techniques, the protein-protein interaction network, and the network topology to classify genes that can be essential to human cancer processes.

Within the broader picture of network-based analysis techniques, there is some recent work in other domains. For instance, Abeysinghe et al. [1] study the topological properties of real-world electricity distribution networks at the medium voltage level by employing the techniques from complex networks analysis and graph theory. Jiang et al. [9] use large urban street networks for topological analysis and show that these networks have the small-world property, but do not exhibit scale-freeness. Zhang et al. [27] use topological properties to better understand bus networks in large cities to optimize the bus lines and transfers.

This paper presented a network-based approach for annotation prediction of genes. It uses the information of the co-expression network of the genome under study to build a predictive model using machine learning techniques. When trained with the topological measures as part of the data set, the model is shown by a series of in-silico experiments to improve the accuracy, F1-score, and AUC ROC in comparison to a model trained without the topological measures of the co-expression network. Each pair of models is trained for predicting a particular gene annotation. By measuring the number of genes most frequently classified as false positive by the prediction model, a small number of genes is identified for 12 biological processes in Oryza Sativa Japonica: these genes are good candidates for experimental verification.

As usual, significant work remains to be done. A next step is to perform experimental evaluation in the laboratory to validate the in-silico predictions. Also, more in-silico experimentation can be used to predict gene annotations in other species, such as sugar cane.

Acknowledgments. The authors would like to thank the anonymous referees for their helpful comments. This work was funded by the OMICAS program: Optimización Multiescala In-silico de Cultivos Agrícolas Sostenibles (Infraestructura y Validación en Arroz y Caña de Azúcar), sponsored within the Colombian Scientific Ecosystem by

The World Bank, Colciencias, Icetex, the Colombian Ministry of Education and the Colombian Ministry of Industry and Turism under Grant FP44842-217-2018.

References

1. Abeysinghe, S., Wu, J., Sooriyabandara, M., Abeysekera, M., Xu, T., Wang, C.: Topological properties of medium voltage electricity distribution networks. Appl. Energy **210**, 1101–1112 (2018)
2. Alanis Lobato, G.: Exploitation of complex network topology for link prediction in biological interactomes (2014)
3. Alanis-Lobato, G., Cannistraci, C.V., Ravasi, T.: Exploitation of genetic interaction network topology for the prediction of epistatic behavior. Genomics **102**(4), 202–208 (2013)
4. Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K., Obayashi, T.: ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. Plant Cell Physiol. **57**(1) (2016)
5. Barabási, A.-L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nat. Rev. Genet. **12**(1), 56–68 (2011)
6. Benstead-Hume, G., Wooller, S.K., Dias, S., Woodbine, L., Carr, A.M., Pearl, F.M.G.: Biological network topology features predict gene dependencies in cancer cell lines. Systems Biology (2019, preprint)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
8. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
9. Jiang, B., Claramunt, C.: Topological analysis of urban street networks. Environ. Plan. **31**(1), 151–162 (2004)
10. Mudge, J.M., Harrow, J.: The state of play in higher eukaryote gene annotation. Nat. Rev. Genet. **17**(12), 758–772 (2016)
11. Naaman, R., Cohen, K., Louzoun, Y.: Edge sign prediction based on a combination of network structural topology and sign propagation. J. Complex Netw. **7**(1), 54–66 (2019)
12. Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., Kinoshita, K.: ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. Plant Cell Physiol. **59**(1) (2018)
13. Obayashi, T., Kinoshita, K.: Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Res. **16**(5), 249–260 (2009)
14. Obayashi, T., Kinoshita, K.: COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Res. **39**(Database), D1016–D1022 (2011)
15. Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Aoki, Y., Shirota, M., Kinoshita, K.: ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. Plant Cell Physiol. **55**(1) (2014)
16. Oti, M., van Reeuwijk, J., Huynen, M.A., Brunner, H.G.: Conserved co-expression for candidate disease gene prioritization. BMC Bioinform. **9**(1), 208 (2008)
17. Ranganathan, S., Gribskov, M.R., Nakai, K., Schönbach, C.: Encyclopedia of Bioinformatics and Computational Biology (2019). OCLC: 1052465484

18. Rust, A.G., Mongin, E., Birney, E.: Genome annotation techniques: new approaches and challenges. Drug Discov. Today **7**(11), S70–S76 (2002)
19. Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C., Iwamoto, M., Abe, T., Yamada, Y., Muto, A., Inokuchi, H., Ikemura, T., Matsumoto, T., Sasaki, T., Itoh, T.: Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol. **54**(2) (2013)
20. Santolini, M., Barabási, A.-L.: Predicting perturbation patterns from the topology of biologicalnetworks. Proc. Natl. Acad. Sci. **115**(27), E6375–E6383 (2018)
21. Shannon, P.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. **13**(11), 2498–2504 (2003)
22. Stuart, J.M.: A gene-coexpression network for global discovery of conserved genetic modules. Science **302**(5643), 249–255 (2003)
23. Tan, F., Xia, Y., Zhu, B.: Link prediction in complex networks: a mutual information perspective. PLoS ONE **9**(9), e107056 (2014)
24. van Dam, S., Võsa, U., van der Graaf, A., Franke, L., de Magalhães, J.P.: Gene co-expression analysis for functional classification and gene–disease predictions. Briefings Bioinform. (2017)
25. Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., Van de Peer, Y.: Unraveling transcriptional control in arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol. **150**(2), 535–546 (2009)
26. Yandell, M., Ence, D.: A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. **13**(5), 329–342 (2012)
27. Zhang, H., Zhao, P., Gao, J., Yao, X.-M.: The analysis of the properties of bus network topology in Beijing basing on complex networks. Math. Problems Eng. **1–6**, 2013 (2013)
28. Zhou, Y., Young, J.A., Santrosyan, A., Chen, K., Yan, S.F., Winzeler, E.A.: In silico gene function prediction using ontology-based pattern identification. Bioinformatics **21**(7), 1237–1245 (2005)