TITAN: A Toolbox for Information-Theoretic Analysis of Molecular Networks

Tido Bergmans[1] and Tansu Celikel[1,2]

1) Department of Neurophysiology, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands
2) School of Psychology, Georgia Institute of Technology, Atlanta - GA, USA

Correspondence should be addressed to {t.bergmans, t.celikel}@neurophysiology.nl

Abstract

Background

Reconstruction of molecular networks allows structural and functional characterization of cells, tissues, and organs. A network analytical approach to statistical correlations in molecular data can provide new insight into the underlying system, but currently there is not any analytical approach that is generalizable across different types of data. Here we present TITAN, an open-source toolbox in MATLAB and Octave for the reconstruction and graph analysis of molecular networks. Using an information-theoretical approach TITAN reconstructs networks from transcriptional data, revealing the topological structure of correlations in biological systems.

Findings

The toolbox includes a data processing pipeline to construct gene co-expression networks from transcriptional data, mutual information (MI) and variation of information (VI) to quantify connectivity and graph-based analysis for molecular network analysis. We show that the toolbox can accurately reconstruct network structure from synthetic data with predetermined network topology. We use the toolbox to construct the gene co-expression network of S. cerevisiae and show that the resulting network is indistinguishable from gene co-expression networks previously identified experimentally.

Discussion

TITAN is a general tool for network reconstruction in high dimensional datasets. It could be used to study co-transcriptional networks and reveal connectivity patterns in any large omics datasets. Considering the universal nature of MI as a measure of correlation, TITAN can potentially be used to vertically integrate datasets for a cross-omics analysis that covers a system on every level, from genes to metabolites. This universality can be extended to cross-species analysis and evolutionary genetics. Other applications include molecular fingerprinting of cells, tissues, and organs, and tracing developmental trajectories for gene and protein networks.

Keywords

mutual information, gene co-expression networks, molecular networks, network sparsification, omics data

Background

Thanks to the increased accessibility of the next-gen sequencing and high throughput -omics approaches, it is now possible to study molecular network interactions systematically from genes to metabolites; even though the computational methods to reconstruct molecular networks are still in their infancy, systems theoretical approaches already contribute to our improved understanding of genetic disorders [1], oncogenomics [2], brain plasticity [3–5] and cellular circuits [6–8]. By studying the molecular mechanisms of a cell as a network, the focus of the analysis can be shifted from differential expression of individual genes to functional correlations between groups of genes, revealing genetic pathways and their interactomes.

Molecular network inference is typically performed by calculating some similarity measure for every pair of genes in a sampled dataset [9]. Since similarity measures determine what genes are connected, the choice of similarity measure has a great influence on the resulting network. Pearson correlation, for example, is often used as a similarity measure because it is computationally efficient, but it can only recognize linear dependencies and is sensitive to outliers [10]. Mutual information (MI) is defined as the shared information between two statistical variables and can be expressed in terms of Shannon entropy [11]. MI, as a similarity measure, is flexible, i.e. applicable to datasets with any type of statistical dependency, not just linear relations, and is invariant under reparametrization. Therefore any prior transformations of the data (e.g. log-transformation or discretization methods) do not influence the resulting (computed) networks. This makes MI a useful tool in cross-analysis of molecular data from multiple datasets.

Previous approaches to MI-based molecular network inference include ARACNE [12], *parmigene* [13] and PANA [14]. These approaches construct gene co-expression networks from transcriptional data using MI to infer connectivity between any given pairs of genes. While ARACNE uses the Data Processing Inequality (DPI) to remove indirect edges from the resulting network, *parmigene* seeks minimal computational costs by using parallel computing, and the

PANA algorithm uses functional gene clusters as a basis for network inference. Commonly, these methods allow visualization of binary networks where two nodes are connected when their MI is above a certain threshold. This threshold may be established by shuffling the transcriptional data and determining the p-value from the resulting MI values. Such a binary or unweighted network can be used to discover patterns of connectivity and find clusters of nodes that are co-expressed.

Biological networks, however, are not binary networks. The weight of the connection between any given two nodes carries information about functional and structural connectivity in the network. Using MI as a measure of similarity, and calculating Variation of Information (VI) rather than distance [15], can provide the necessary weight information and enable quantitative analysis of connectivity in weighted and directed networks. Because network statistics such as average path length and betweenness centrality can only be computed from a weighted network where edge weight corresponds to distance, this expansion of the MI would offer new opportunities to integrate graph-based analytical methods to study biological networks. TITAN, a plug-and-play toolbox for the construction of functional networks at the molecular level, provides the aforementioned functionality. The toolbox, natively compatible with MATLAB and Octave, includes a data processing pipeline that uses information theory to infer networks from sampled data. Functions could be called in Python using Oct2Py [16]. The toolbox can be used for analysis and visualization of the resulting network, including pathway analysis, clustering, and export of data to network visualization software.

TITAN uses MI and VI to find correlations in molecular data (e.g. gene expression data) and construct a network. The properties of MI and VI allow for a fast data processing pipeline that is robust and broadly applicable to study biological networks. We confirm the sensitivity and selectivity of the toolbox using the synthetic (artificial) data with known network topology, then demonstrate its utility in determining transcriptional networks in various species of yeast cells [17]. From this data, we construct a co-expression network and compare the resulting network to previously established genetic networks in yeast. We show that TITAN is able to accurately reconstruct genetic networks from gene expression data independent of the choice of parameters.

Methods

TITAN is an Octave/MATLAB toolbox that can be used for the construction and analysis of correlation networks (Figure 1). The toolbox has a modular organization. After data import, normalization and entropy calculations, mutual information between any given node (e.g. transcript) in the network is calculated. “Network analysis” routines allow sparsification of the network by identifying indirect connections and provide tools for graph-theory based network analysis. Users can optionally export the data for further analysis, e.g. in Gephi [18], for custom visualizations.

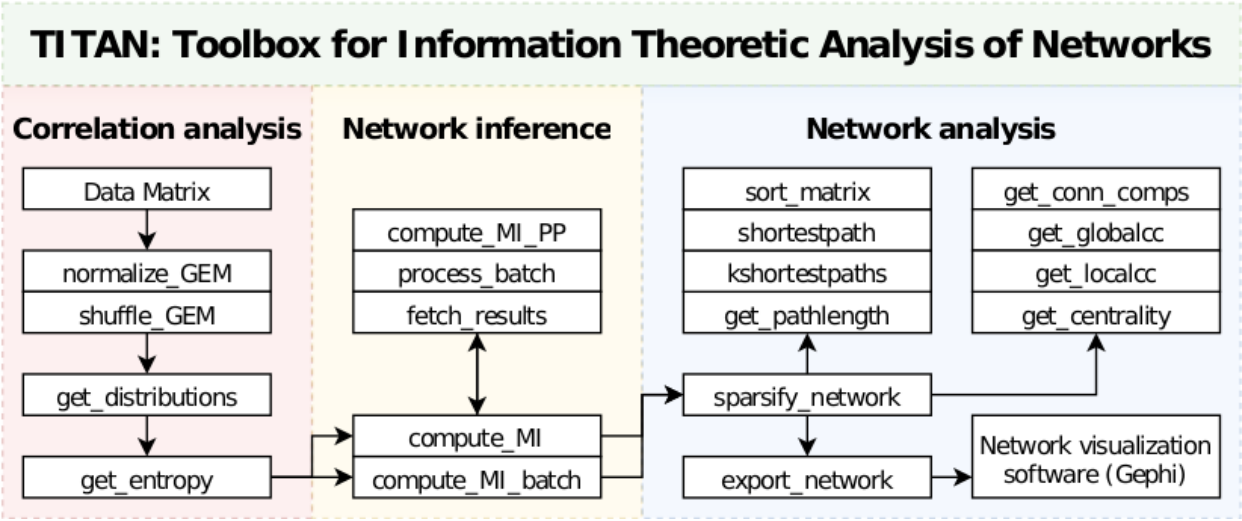

Figure 1: An overview of TITAN functions and workflow. The data processing pipeline can be separated into three modules: data preparation, network inference, and network analysis. Function list in each module and data processing hierarchy are shown.

The toolbox comprises a data processing pipeline that imports a high-dimensional dataset organized as a matrix where columns are samples (e.g. cells) and rows are (molecular) targets (e.g. genes), and constructs a network based on correlations; see Figure 2 for the implementation of the toolbox for network analysis from gene expression data. The toolbox includes functions to evaluate network structure, and identify hubs, clusters, and pathways in the network. The following sections describe the different steps in the pipeline in more detail.

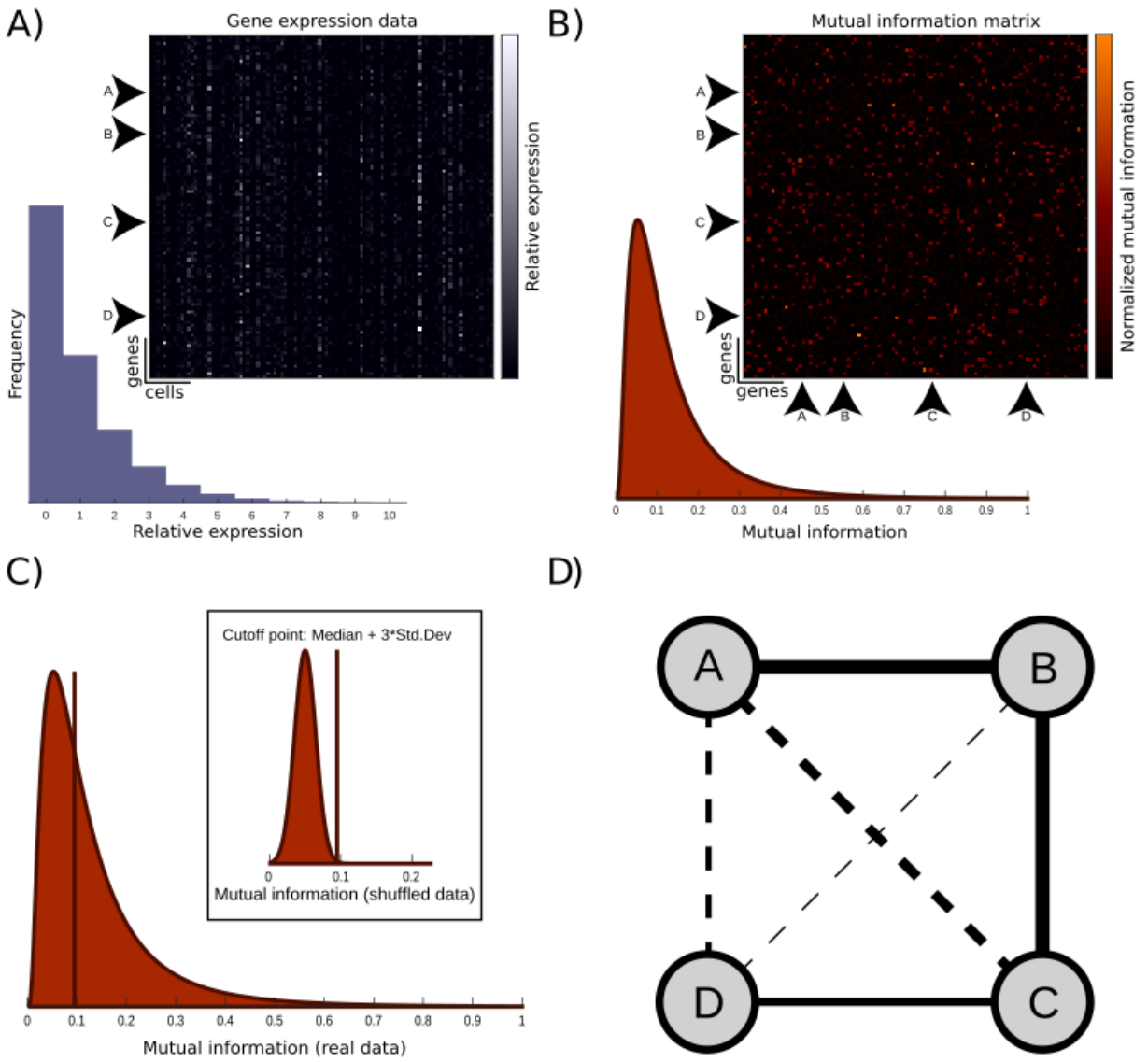

Figure 2: Basic processing steps for quantification of co-expression network from gene expression data. (A) The pipeline starts with a data set organized as a matrix, e.g. RNA-sequencing data, where columns are samples (cells, repetitions) and rows are targets (e.g. transcript names). (B) From the expression data, the mutual information (MI) matrix is computed which stores pairwise correlations between genes. This matrix is interpreted as the adjacency matrix of a weighted, undirected network. (C) In order to establish a significance threshold for MI, a shuffle correction is performed for which MI matrices are computed from randomly shuffled gene expression data (see section on “Shuffle correction”). A threshold is then based on the median and standard deviation of the shuffled MI distribution. (D) To remove spurious correlations from the network, a network sparsification algorithm is applied. This algorithm iterates over the network connections, removing the edges that are the result of an indirect correlation (see section on “Sparsification by data processing inequality”).

Gene expression data normalization

The gene expression data contains transcriptional data from a set of genes over a set of cells. This data is stored in a matrix X where X_{ij} represents the transcript count of gene i in cell (or sample) j. The transcription data is log-transformed and/or normalized with respect to total transcription per cell to account for differences in cell size and metabolic rate. The toolbox provides functions for shuffling or subsampling the data for shuffle correction and bootstrapping.

Network inference

The inference of gene co-expression networks is based on statistical dependencies between transcription levels of individual genes. Mutual information (MI) is used as a measure of correlation to describe these dependencies and quantifies the amount of information obtained about a stochastic variable through the information describing the other variable [11]; [19]; see [20] for a recent expansion. It captures any type of statistical dependency, not just positive or negative linear relations, and is invariant under reparametrization.

Mutual information between two genes is computed using the probability distributions of their respective transcription levels, as well as the joint probability distribution. The mutual information is then normalized by the joint entropy of the two genes. The joint entropy is a theoretical maximum for the mutual information so this constrains the normalized mutual information to the interval [0,1].

The probability distributions of the genes are obtained by discretizing the transcription data using Sturges' method [21]. From these distributions the entropy is calculated according to the equation:

$$H(A) = -\sum_{x} p_A(x) \log(p_A(x))$$

Where H(A) is the entropy of gene A, p_A the transcription distribution of A, and x the discretized transcription levels. From the same principle the mutual information I(A,B) between genes A and B is calculated from the distributions p_A, p_B, and joint distribution p_{AB}, and respective transcription levels x, y:

$$I(A,B) = \sum_x \sum_y p_{A,B}(x,y) \log \left(\frac{p_{A,B}}{p_A(x), p_B(y)} \right)$$

The pairwise mutual information is stored in a matrix (MI matrix) where every entry (A,B) is the mutual information between genes A and B. This matrix is normalized with respect to the joint entropy of A and B, resulting in a symmetric network matrix with values between zero and one representing weighted connections in the gene co-expression network. The joint entropy can be calculated from the single gene entropies and the mutual information:

$$H(A,B) = H(A) + H(B) - I(A,B)$$

Shuffle correction

In order to identify random correlations in the network, a significance threshold is established using shuffle correction. Shuffle correction is performed for each column across rows so that overall transcription per cell is preserved, but any correlations between genes are removed from the data. From this random gene expression matrix, the MI-based network is inferred according to the steps described in the previous section. This step is repeated N times (default: 10). With the distribution of MI values resembling a normal distribution the significance threshold can defined using the median and standard deviation of the distribution, e.g. Median+3*S.D. Connections with a mutual information value below this statistically defined cutoff point are considered random correlations and removed from the network.

Sparsification by Data Processing Inequality

The next step is to remove connections from the network that were produced by indirect correlations. Indirect correlations appear when the transcription of two genes, e.g. A and C, are co-regulated with a third gene B but are otherwise independent. Since there is a correlation in transcriptional behavior between A and B, as well as between B and C, a spurious correlation will also be found between A and C, even though they don't interact and therefore shouldn't be connected in the network. These indirect connections are removed using the principle of Sparsification by Data Processing Inequality [12]. The Data Processing Inequality states that the mutual information between A and C will be lower than the mutual information between A and B or B and C:

$$I(A,C) < I(A,B)$$
$$I(A,C) < I(B,C)$$

The sparsification is carried out by iterating over all first-order loops in the network (i.e. all trios of genes that are fully connected) and removing the weakest connection in the loop.

Network analysis

A number of statistics, all of which are included in TITAN, can be computed that help identify genes of interest and understand the overall structure of the network. At the node level, statistics such as degree, local clustering coefficient, and betweenness centrality can be used to identify genes that have an important role in the network, and potentially in transcriptional regulation. On a structural level, Dijkstra's algorithm [22] can be used to compute the (weighted) shortest path from one node to another, or the degree of separation (DOS), i.e. the minimum number of edges needed to get from one node to another. The degree of clustering in the network can be assessed by the global clustering coefficient, small world coefficient, or PCA-related methods available in the toolbox. Finally, the toolbox contains functionality for exporting the network in GEXF format, a common data format that is used by network visualization softwares, such as Gephi [18].

Since mutual information is a similarity (or "closeness") metric, it can not be used to describe distances between network nodes. To compute the shortest path from one node to another, the variation of information (VI) was used. Variation of information quantifies the information difference between two distributions [15]; if MI describes how close two nodes are in the network, then VI describes how far they are from one another. The variation of information can be calculated from the joint entropy and mutual information of a gene pair:

$$VI(A,B) = H(A,B) - I(A,B)$$
$$VI_n(A,B) = 1 - I_n(A,B)$$

where $I_n(A,B)$ is the normalized mutual information. The maximum value for VI is 1 which is the case when MI equals zero and two genes are completely uncorrelated. Since two disconnected nodes in the network should have a "distance" of infinity (instead of 1), the distance metric in the network was defined as the ratio of VI to MI:

$$dist(A, B) = \frac{VI(A, B)}{I(A, B)} = \frac{1 - I_n(A, B)}{I_n(A, B)}$$

Findings

In the following two sections we assess the quality of networks that were constructed with the toolbox. First, a network was constructed from artificial (synthetic) data with a known ground truth network organization. Expression profiling was performed on the synthetic data and the underlying network was inferred using TITAN. Second, a network was constructed from RNA-sequencing data from Saccharomyces cerevisiae, a species of yeast whose genome has been studied extensively [23–25]. The resulting network was compared to five previously found co-expression networks in S. cerevisiae, and a network pathway analysis was done to identify potential genes of interest.

Ground truth network

A topologically simple, undirected, unweighted ground truth network was created and used to generate synthetic expression data. The ground truth network consists of 100 disjoint graphs of 10 nodes connected in a linear pattern. From a random seed, expression data was generated that preserves this topology. The data was fed into the data processing pipeline to infer a co-expression network. The resulting network was shuffle-corrected with a cutoff point of Median+3*S.D., sparsified and then compared to the ground truth network to assess its accuracy and to evaluate the limits of the method in terms of minimum data dimensionality. This was repeated while varying the number of samples in the expression data, as well as the levels of noise added to the expression data. The number of samples ranged from 100 to 8000 and noise was sampled from a normal distribution with standard deviation of 10% to 50% of the mean. To assess the quality of the inferred networks as a function of data density and number of samples, the true/false positive rates were used, where a false positive is a connection found between two nodes that were not connected in the original network.

Results (expectedly) showed that network quality increases with the number of samples and decreases with the amount of noise added to the system (Figure 3A). The distributions of MI values in the 8000 sample-networks as a function of degree of separation (DOS) in the original network revealed that directly connected nodes have a median MI >0.2 for all noise levels, with

MI decaying as the distance between nodes increases (Figure 3B). Decay rate increases with the amount of added noise, and disconnected nodes always have MI around zero.

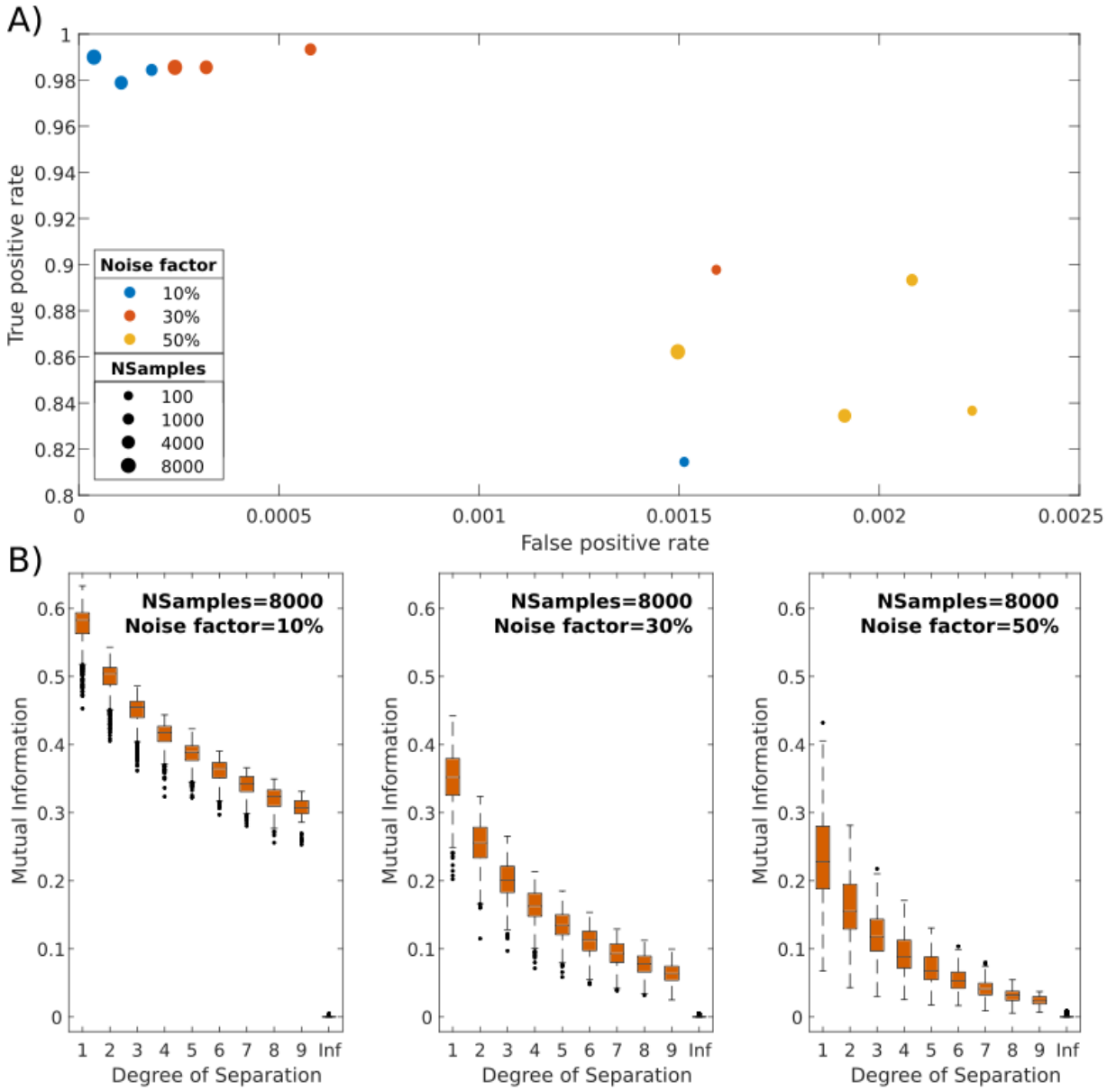

Figure 3: Networks constructed from synthetic gene expression data. Networks were constructed from synthetic expression data with varying numbers of samples and variable amounts of noise added. (A) Accuracy of constructed networks when compared to their ground truth respectives. Marker color denotes noise levels (red=10%, blue=30%, yellow=50%) and marker size denotes the number of samples used (100, 1000, 4000, 8000). (B) MI as a function of separation (DOS) in the networks. DOS is the number of edges between a pair of genes; a value of 1 means they are directly connected; a value of 'inf' means there is no possible path between them.

S. cerevisiae networks

To test TITAN's performance with biological systems, networks were constructed from RNA-sequencing data from S. cerevisiae [17]. A co-regulatory network was constructed and filtered after shuffle correction from 7126 genes sequenced in 672 samples. Subsets of this network were compared to five corresponding co-expression networks found in previous studies that are listed in Table 1. The network data were made available by YeastNet v3 [26].

Table 1. Sources of the co-expression networks in Saccharomyces cerevisiae

No.	Reference	Title
1	Mitchell et al. 2009 [30]	Adaptive prediction of environmental changes by microorganisms.
2	Wade et al. 2009 [31]	The Snf1 kinase and proteasome-associated Rad23 regulate UV-responsive gene expression.
3	Shalem et al. 2011 [32]	Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA Pol II.
4	Yona et al. 2012 [33]	Chromosomal duplication is a transient evolutionary solution to stress.
5	Orlando et al. 2008 [34]	Global control of cell-cycle transcription by coupled CDK and network oscillators.

As expected, compared to the synthetic data, the MI values were significantly smaller in the yeast networks and the median value for normalized MI in each of the 5 networks was less than <0.05 (Figure 4A). To numerically compare TITAN's network reconstruction against previously published network structure in the yeast, we calculated the receiver operating characteristics (ROC). Because ROC is sensitive to the noise level, we calculated three thresholds to binarize the MI matrix (Figure 4A; three candidate thresholds established by shuffle correction are shown in color) and quantified the successful identification (i.e. true positive) against false positive ratio (Figure 4B). The results expectedly showed that with the threshold inversely proportional to the true and false positive discovery rate; with the increased threshold both discovery rates are reduced. Conservatively ~60% of all previously identified networks could be correctly discovered using TITAN with a false discovery rate of ~20% (Figure 4) using conservative normalized MI

threshold (i.e. 11 S.D.). The discovery rate improves to >80% with lower thresholds (e.g. 7 S.D.) at the expense of increased false discovery rate.

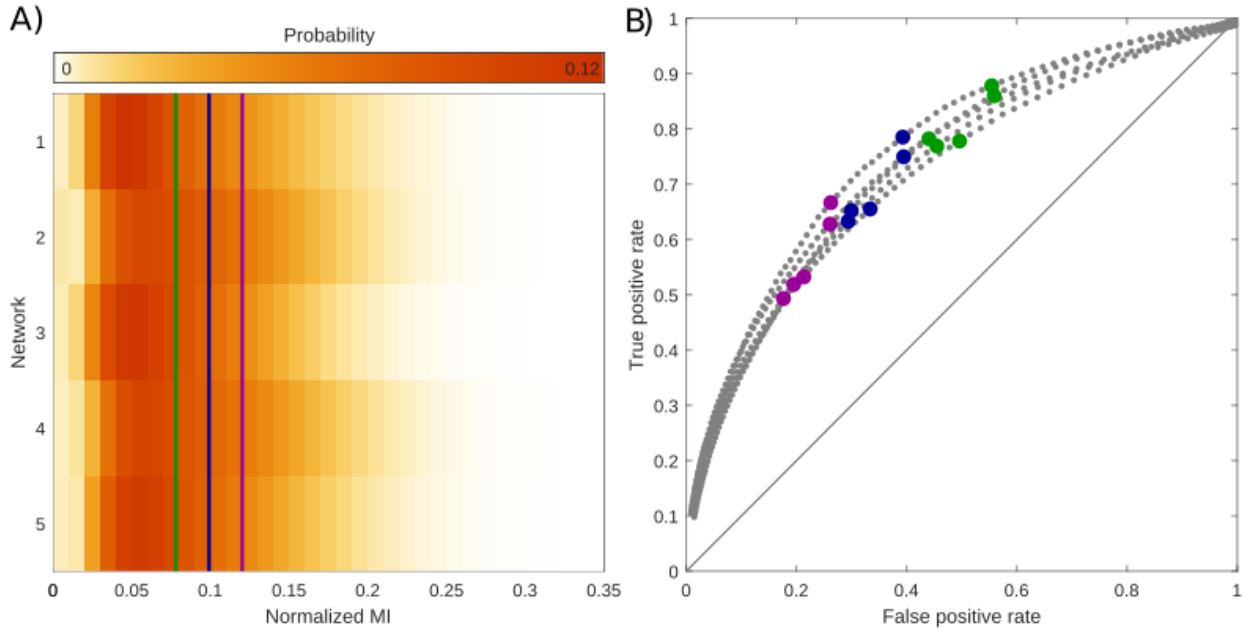

Figure 4: Five gene co-expression networks in S. cerevisiae constructed from RNA-seq data. (A) Distributions of MI values of the constructed networks. The three candidates Median plus 7 (green), 9 (blue) and 11 (purple) times S.D. of the shuffled MI distribution are shown. (B) Receiver operating characteristic (ROC) curve. True and false positive rates show the similarity between the constructed networks and previously established networks (see Table 1) for different significance thresholds. The colored data points correspond with the thresholds shown in (A).

Similarly to Figure 3B for synthetic data, Figure 5 shows the distributions of MI values as a function of DOS in the five S. cerevisiae networks. The bottom-right plot shows the medians of the five distributions plotted together. A DOS of 'Inf' (infinite) occurs when there is no possible path between two nodes. In this case, the two genes represented by these nodes are uncorrelated. The figures show that MI decreases as DOS increases, indicating that MI is indeed a measure of "closeness" of two nodes in the network, which decreases as noise is introduced with every new edge along the path.

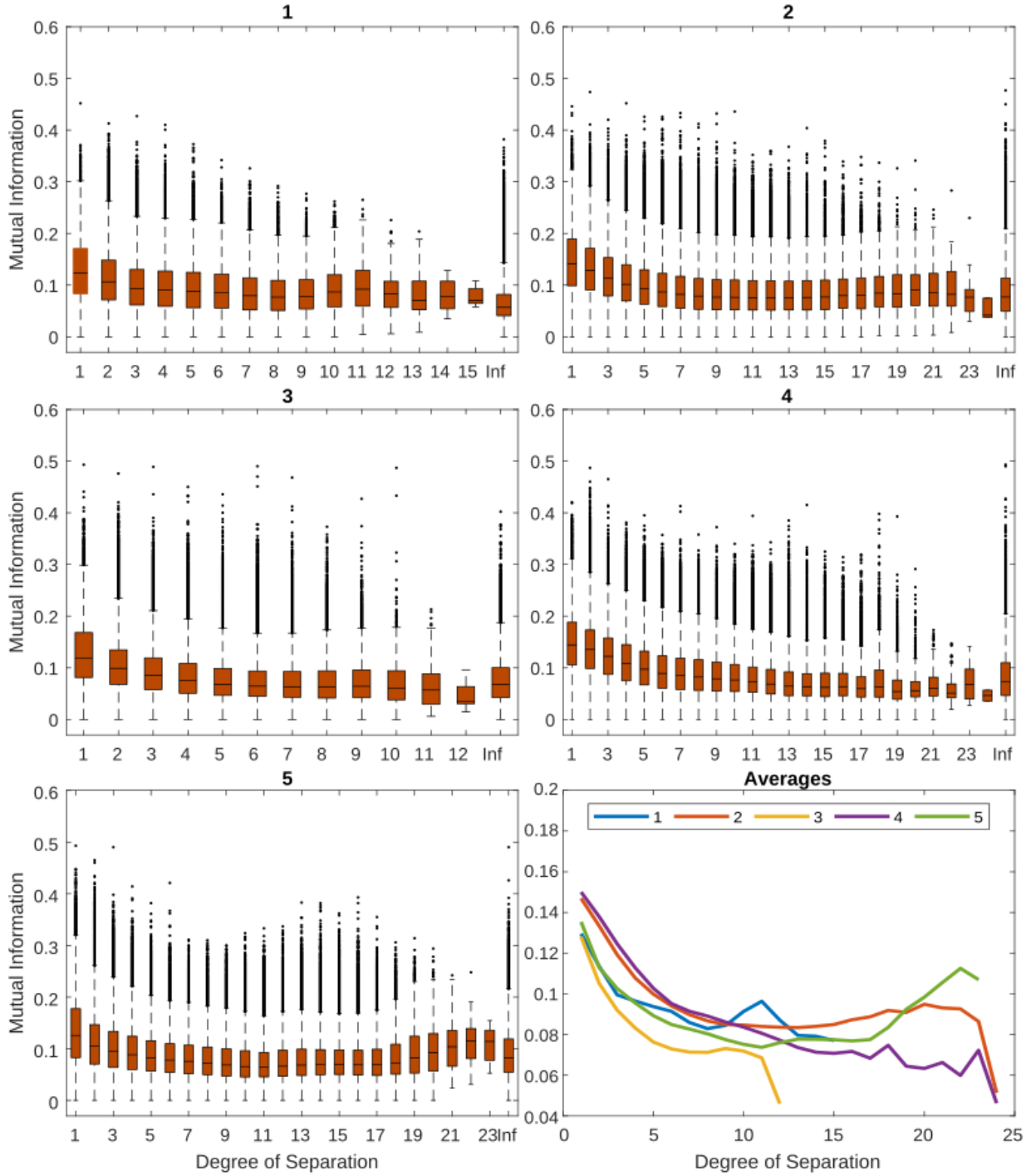

Figure 5: Mutual information (MI) distributions for varying degrees of separation (DOS) in the networks in Saccharomyces cerevisiae. DOS is the network distance between a pair of genes: A value of 1 means they are directly connected; a value of 'inf' means there is no path between them. The distributions are shown for all networks separately, and the distribution medians are plotted in the bottom right subfigure.

TITAN differs from previous approaches to MI-based molecular network inference (ARACNE [12], *parmigene* [13] and PANA [14]) as such it allows weighted network analysis for pathway discovery (see variation of information in Methods). The direct utility of identification of weights in the network is that it allows computation of the shortest path among any given two targets within the network, hence expanding the utility of MI analysis to pathway discovery. This approach can be further expanded to the analysis of each target as a hub by calculation of the betweenness centrality which is defined as the fraction of all shortest paths that go through a particular node [27]. Figure 6 shows the distributions of betweenness centrality for the nodes in the S. cerevisiae networks, as well as the top 5 highest scoring nodes for this metric. Genes that preserve their high centrality across multiple networks are in bold print.

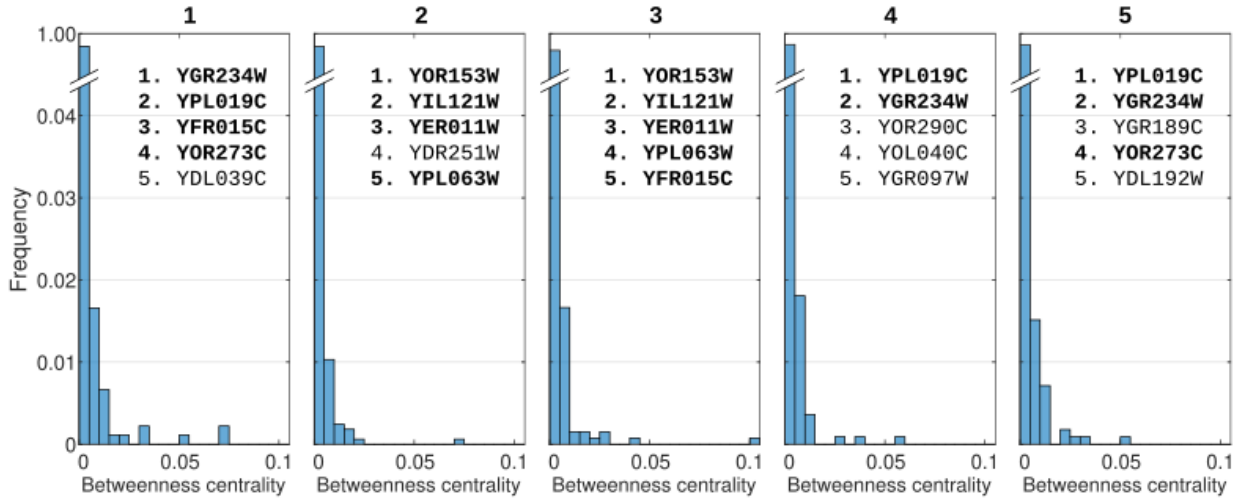

Figure 6: Betweenness centrality distributions for the nodes in the five S. cerevisiae networks. In every plot, the top five genes with the highest centrality are listed. Genes that are preserved across different networks are bold printed.

Discussion

TITAN is an Octave/MATLAB toolbox that can be used for reconstruction of molecular networks. The accuracy of the approach was shown using synthetic (artificial) data and quantified against empirically identified networks in yeast. Expanding co-expression network identification to pathway analysis allows utility of TITAN for identification of nodes in networks (Figure 6).

TITAN works in a plug-and-play manner without the need for additional toolbox installation or complicated setup; the code is available on Github and the expression data can be imported

into Octave or MATLAB from e.g. a comma-separated value (CSV) file. The toolbox offers the flexibility to finetune any parameters in the data processing pipeline, but also works out of the box using default parameters - the latter of which was the case for the analyses described in this article. Network data can be exported to network software tools for visualization and subsequent analysis.

Drawbacks of the MI-based approach include the need for high-dimensional data; Figure 3 shows that true positive rate increases with number of samples and quality of the synthetic data. Especially for higher levels of noise, 1000+ samples are needed for acceptable performance. While MI is computationally heavy, parallelization can speed up the process. For parallel computation of the MI matrix one can use the MATLAB Parallel Processing Toolbox in combination with the parallel processing functions included in TITAN.

Due to the universal nature of MI, TITAN allows for cross-comparison of transcription data between cell types, individuals, populations or species. Rather than comparing individual genes, one can now compare entire transcriptomes by their internal structure, looking for patterns of connectivity that are preserved [28]. These patterns can, for example, shine a light on basic molecular mechanisms at the root of the evolutionary tree, which would be imperceptible through single gene analysis.

Going beyond cross-species transcriptomics, TITAN can be used to integrate multi-omics data covering multiple layers of the cellular process - e.g. genomics, epigenomics, proteomics and metabolomics - to reconstruct network representations of the molecular landscape. Understanding the topology of these networks makes it possible to bridge the gap between genotype and phenotype, and predict downstream effects of genetic defects [29]. It will allow us to link low-level physiological mechanisms to higher-level phenomena such as disease symptoms and behavioral traits, and provide insight into the mechanisms behind cell fate and how to control it, allowing for cellular reprogramming.

Code Availability

- Project name: TITAN
- Project homepage: https://github.com/DepartmentofNeurophysiology/TITAN
- Operating system: Windows/Linux

- Programming language: Octave/MATLAB
- Software requirements: Octave 5.1.0.0 or higher; MATLAB 2015b or higher; MATLAB Parallel Computing Toolbox™ to enable parallel processing functions; Oct2py Python package for implementation in Python 2/3
- License: MIT

Availability of Supporting Code and Data

- See project homepage https://github.com/DepartmentofNeurophysiology/TITAN

Abbreviations

CSV: comma-separated value; DOS: degree of separation; DPI: data processing inequality; MI: mutual information; ROC: receiver operating characteristic; VI: variation of information

Competing Interests

The authors declare that they have no competing interests.

Author's Contributions

T.B. designed and implemented TITAN and wrote the manuscript; T.C. conceived the project, contributed to design and revised the manuscript. All authors approved the final manuscript.

References

1. Lim C-S, Kang X, Mirabella V, Zhang H, Bu Q, Araki Y, et al. BRaf signaling principles unveiled by large-scale human mutation analysis with a rapid lentivirus-based gene replacement method. *Genes Dev*. 2017; doi: 10.1101/gad.294413.116.
2. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, et al. Pathway and network analysis of cancer genomes. *Nat Methods*. 2015; doi: 10.1038/nmeth.3440.
3. Kole K, Komuro Y, Provaznik J, Pistolic J, Benes V, Paul T, et al. Erratum to: Transcriptional mapping of the primary somatosensory cortex upon sensory deprivation. *GigaScience*. 2018; doi: 10.1093/gigascience/giy040.
4. Kole K, Scheenen W, Tiesinga P, Celikel T. Cellular diversity of the somatosensory cortical map plasticity. *Neurosci Biobehav Rev*. 2018; doi: 10.1016/j.neubiorev.2017.11.015.
5. Kole K, Lindeboom RGH, Baltissen MPA, Jansen PWTC, Vermeulen M, Tiesinga P, et al. Erratum to: Proteomic landscape of the primary somatosensory cortex upon sensory deprivation. *GigaScience*. 2018; doi: 10.1093/gigascience/giy041.
6. Huang C, Zeldenrust F, Celikel T. Cortical representation of touch in silico. *bioRxiv*. 2021; doi: 10.1101/2020.11.06.371252.
7. Zeldenrust F, Calcini N, Yan X, Bijlsma A, Celikel T. Cell type specific information transfer for sparse coding. *bioRxiv*. 2020; doi: 10.1101/2020.11.06.371658.
8. Huang C, Englitz B, Reznik A, Zeldenrust F, Celikel T. Information transfer and recovery for

the sense of touch. *bioRxiv*. 2020; doi: 10.1101/2020.12.08.415729.
9. Mc Mahon SS, Sim A, Filippi S, Johnson R, Liepe J, Smith D, et al. Information theory and signal transduction systems: From molecular information processing to network inference. *Semin Cell Dev Biol*. 2014; doi: 10.1016/j.semcdb.2014.06.011.
10. Gibson SM, Ficklin SP, Isaacson S, Luo F, Feltus FA, Smith MC. Massive-Scale Gene Co-Expression Network Construction and Robustness Testing Using Random Matrix Theory. *PLOS ONE*. 2013; doi: 10.1371/journal.pone.0055871.
11. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948; doi: 10.1002/j.1538-7305.1948.tb01338.x.
12. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*. 2006; doi: 10.1186/1471-2105-7-S1-S7.
13. Sales G, Romualdi C. parmigene—a parallel R package for mutual information estimation and gene network reconstruction. *Bioinformatics*. 2011; doi: 10.1093/bioinformatics/btr274.
14. Ponzoni I, Nueda MJ, Tarazona S, Götz S, Montaner D, Dussaut JS, et al. Pathway network inference from gene expression data. *BMC Syst Biol*. 2014; doi: 10.1186/1752-0509-8-S2-S7.
15. Meilă M. Comparing clusterings—an information based distance. *J Multivar Anal*. 2007; doi: 10.1016/j.jmva.2006.11.013.
16. Silvester S. oct2py: Python to GNU Octave bridge --> run m-files from python. https://pypi.org/project/oct2py/. Accessed 9 April 2021.
17. Ziemann M, Kaspi A, El-Osta A. Digital expression explorer 2: a repository of uniformly processed RNA sequencing data. *GigaScience*. 2019; doi: 10.1093/gigascience/giz022.
18. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proc Int AAAI Conf Web Soc Media*.
19. Zeldenrust F, de Knecht S, Wadman WJ, Denève S, Gutkin B. Estimating the Information Extracted by a Single Spiking Neuron from a Continuous Input Time Series. *Front Comput Neurosci*. 2017; doi: 10.3389/fncom.2017.00049.
20. Azarfar A, Calcini N, Huang C, Zeldenrust F, Celikel T. Neural coding: A single neuron's perspective. *Neurosci Biobehav Rev*. 2018; doi: 10.1016/j.neubiorev.2018.09.007.
21. Herbert A. Sturges. The Choice of a Class Interval. *J Am Stat Assoc*. 1926;21:65–61926;
22. Dijkstra EW. A note on two problems in connexion with graphs. *Numer Math*. 1959; doi: 10.1007/BF01386390.
23. Münzner U, Klipp E, Krantz M. A comprehensive, mechanistically detailed, and executable model of the cell division cycle in Saccharomyces cerevisiae. *Nat Commun*. 2019; doi: 10.1038/s41467-019-08903-w.
24. Dai Z, Huang M, Chen Y, Siewers V, Nielsen J. Global rewiring of cellular metabolism renders Saccharomyces cerevisiae Crabtree negative. *Nat Commun*. 2018; doi: 10.1038/s41467-018-05409-9.
25. Janjić V, Sharan R, Pržulj N. Modelling the Yeast Interactome. *Sci Rep*. 2014; doi: 10.1038/srep04273.
26. Kim H, Shin J, Kim E, Kim H, Hwang S, Shim JE, et al. YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. *Nucleic Acids Res*. 2014; doi: 10.1093/nar/gkt981.
27. Freeman LC. A Set of Measures of Centrality Based on Betweenness. *Sociometry*. 1977; doi: 10.2307/3033543.
28. Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. *Nat Rev Genet*. 2017; doi: 10.1038/nrg.2017.19.
29. Ma A, McDermaid A, Xu J, Chang Y, Ma Q. Integrative Methods and Practical Challenges

for Single-Cell Multi-omics. *Trends Biotechnol*. 2020; doi: 10.1016/j.tibtech.2020.02.013.
30. Mitchell A, Romano GH, Groisman B, Yona A, Dekel E, Kupiec M, et al. Adaptive prediction of environmental changes by microorganisms. Nature. 2009; doi: 10.1038/nature08112.
31. Wade SL, Poorey K, Bekiranov S, Auble DT. The Snf1 kinase and proteasome-associated Rad23 regulate UV-responsive gene expression. EMBO J. 2009; doi: 10.1038/emboj.2009.229.
32. Shalem O, Groisman B, Choder M, Dahan O, Pilpel Y. Transcriptome Kinetics Is Governed by a Genome-Wide Coupling of mRNA Production and Degradation: A Role for RNA Pol II. PLOS Genet. 2011; doi: 10.1371/journal.pgen.1002273.
33. Yona AH, Manor YS, Herbst RH, Romano GH, Mitchell A, Kupiec M, et al. Chromosomal duplication is a transient evolutionary solution to stress. Proc Natl Acad Sci U S A. 2012; doi: 10.1073/pnas.1211150109.
34. Orlando DA, Lin CY, Bernard A, Wang JY, Socolar JES, Iversen ES, et al. Global control of cell-cycle transcription by coupled CDK and network oscillators. Nature. 2008; doi: 10.1038/nature06955.