BIOINFORMATICS INSTITUTE 2021/22

Spring term research projects

Saint Petersburg

2022

ISBN 978-5-7422-7814-6

BIOINFORMATICS INSTITUTE

2021/22. Projects Abstracts

Saint Petersburg, 2022

Table of contents

SPRING 2022

УДК 577.22

Age patterns in gene regulatory networks

Y. Burankova[1], E. Zhivkoplias[2]

[1] *Bioinformatics Institute, Kantemirovskaya street, 2A, 197342, Saint-Petersburg, Russia*
[2] *Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Solna, Sweden*

Introduction

One of the problems in systems biology remains the lack of understanding of the large-scale biological relationships between genes and the proteins they encode. The wide availability of system-level gene expression datasets makes it possible to reconstruct hidden regulatory relationships between gene-gene and gene-protein, or to reverse-engineered gene regulatory networks (GRNs) [1]. GRN comprises nodes (the genes and their regulators) and edges (the regulatory relationships between the nodes). It is usually represented mathematically as an oriented graph. The nature of the interactions in GRNs distinguishes it from other networks in biological systems. The interactions between molecules in GRNs usually involve the indirect regulatory interaction through the biological molecules, which are hard to detect and quantify. Consequently, GRNs are harder to validate.

The GRNs we know are the result of a long biological evolution. The phylogenomic analysis makes it possible to classify genes based on the oldest species that carry orthologous genes [2, 3]. For protein-protein interaction (PPI) networks in yeast and human, it was shown that proteins of the same age tend to interact more [4, 5].

This project aims to explore if gene interaction preference for genes of similar age holds in gene regulatory networks, particularly in those that describe direct regulatory interaction (transcription factor-target gene). Existing network prediction methods rely primarily on expression data. If gene interaction preference for genes of similar age holds in gene regulatory networks, incorporating biological knowledge into network inference methods could help to improve the reliability of the GRNs inferred from expression data.

Materials and methods

For the analysis, we used three gene regulatory networks. Yeast GRN is a complete transcriptional regulatory network (Tnet) [6]. The other two, Mouse GRN and Human GRN, are manually curated databases (TRRUST v2) [7]. Data contain the list of links between transcription factors (TF) and corresponding target genes (TG). All edges have been experimentally confirmed earlier.

First, we studied the GRNs structure using NetworkX 2.8.1 [8] and pandas 1.4.2 Python 3.10.1 libraries [9].

Yeast GRN has 4 441 genes with 12 873 interactions. Of these, 157 genes are TF, and 4 410 are targets. The average number of interactions for nodes is 2.8987. Mouse GRN has 2 456 genes with 7 057 interactions. Of these, 827 genes are TF, and 2 092 are targets. The average number of interactions for nodes is 2.6425. Human GRN has 2 862 genes with 8 427 interactions. Of these, 795 genes are TF, and 2 492 are targets. The average number of interactions for nodes is 2.9444. We used three methods to obtain age classes: protein age classes [2], GenOrigin database [10] and calculated using a phylostratigraphy approach [3].

Protein age classes [2] were translated into gene age classes using protein-gene name matching from the YeastGenome [11] and UNIPROT [12] databases. Interaction maps of TF and targets and TG/TF heatmaps were built for each GRN. Finally, the "difference" of ages in relationships was calculated. The number is the difference between the ages; the smaller, the closer the ages of the interacting genes.

We used the gene ages from the GenOrigin [10] database to calculate the same parameters as for protein classes for Yeast GRN parameters. We used the GenOrigin phylogenetic tree to convert a numerical age into an age class.

We used a phylostratigraphy approach [2] to determine the age of yeast genes in GRN. The iTOL tree [13] phylogeny was used in the analysis to truncate the swiss DB. We compared 4 184 yeast gene sequences by BLAST (blastx) against truncated the Swiss-prot [14] database (94 268 sequences, (10-3 E-value cutoff).

We tested the possibility of randomly obtaining the derived age class ratios in the gene regulation network. We randomly reassigned age classes to 1000 yeast, mouse, and human GRNs to do this. The percentage of each "age" interaction distance for each network was calculated. For each resulting age distance distribution, the standard deviation was counted.

The workflow is represented in the .ipynb files and available in the GitHub repository https://github.com/Freddsle/age_patterns.

Results and Discussion

After translation and mapping protein age classes to GRNs, age was determined for 3 437 (77.4%) genes in Yeast GRN, for 2 287 - (93.1%) in Mouse GRN, and 2 855 (99.8%) - in Human GRN. For the genes, 8 age classes were identified for each GRN. Cellular_organisms, Euk+Bac, Euk_Archaea, Eukaryota, Opisthokonta classes were found in all three networks. Dikarya, Ascomycota, Saccharomyceta classes present in Yeast GRN, and Eumetazoa, Mammalia, Vertebrata in Mouse and Human GRNs.

The proportion of the 'Eumetazoa->Eumetazoa' and 'Eumetazoa->Vertebrata' interactions are the largest among all interactions for mouse and human GRNs (each is more than 10%). On average one TF controls more targets (maximum up to 25) in the yeast network than in mouse (up to 6) and human GRN (up to 8). For yeast GRN, younger nodes have more edges to different age nodes in the network than older nodes. For mouse and human GRNs, the differences are less noticeable. There is no such drop in the number of connections with increasing age.

Human and mouse GRNs have demonstrated a tendency for genes from similar age groups to interact more with each other than with more "distant" age groups. For the yeast GRN, this does not seem to be the case.

The gene ages calculated from the protein ages gave different results for human and mouse, and yeast GRNs. Therefore, we decided to use the gene ages from the GenOrigin. After mapping, we determined the age class for 4 184 genes (94.2%) in Yeast GRN.

TF of the 'Dikarya' age class control fewer targets than other TF classes; there are less than six targets per 'Dikarya' TF. Also, targets of 'Dikarya' and 'Opisthokonta' classes are controlled by m0000ore TF than other target classes. There are less than 5 'Opisthokonta' targets per TF. For 'Dikarya' TF and 'Opisthokonta' targets, the proportion of links among all links in the network is minimal for any edges (less than 0.3% for any combination).

Using gene ages from the GenOrigin, there is no significant predominance of interactions between similar age classes in the yeast network. Edges with age distances 0 ("same age") and 1 ("close age") account for less than 35% of all edges.

When using phylostratigraphy, the fraction of "same age" interactions (distance between ages is 0) has increased. However, this observation may be caused by the truncated tree, in which all age classes older than eukaryotes also received the label eukaryotes. Also, even though the 'Opisthokonta' class was sufficiently represented in the truncated Swiss database, the number of targets of this age class turned out to be less than expected. Therefore, we plan to blast GRNs genes to a fine-grained tree with a more uniform representation of nodes across gene classes.

Was it possible to obtain preferences in the interaction in a random network? We determined interaction preference only for certain age distances (distances are 2, 7 or 8) using the method with randomly assigned classes in Yeast GRN.

Conclusion

Unfortunately, we cannot confidently say that our hypothesis about gene interaction preference in GRN has been confirmed. None of the three methods used to obtain gene ages showed that interactions of "same" and "close" age are dominated in yeast GRN. There are no significant differences compared to the model where the

age categories are randomly assigned. We need a more correctly formulated null hypothesis (a method for obtaining a random network) or a more correct phylogenetic resolution (a fine-grained tree with a more uniform representation of nodes across gene classes).

References

1. Davidson, Eric H. "Emerging properties of animal gene regulatory networks." Nature 468.7326 (2010): 911-920.

2. Liebeskind, B.J., McWhite, C.D., and Marcotte, E.M. "Towards consensus gene ages." Genome biology and evolution 8.6 (2016): 1812-1823.

3. Domazet-Lošo, T., Brajković J., and Tautz, D. "A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages." Trends in Genetics 23.11 (2007): 533-539.

4. Chen, C-Y., et al. "Dissecting the human protein-protein interaction network via phylogenetic decomposition." Scientific reports 4.1 (2014): 1-10.

5. Capra, J.A., Pollard, K.S., Singh, M.. "Novel genes exhibit distinct patterns of function acquisition and network integration." Genome biology 11.12 (2010): 1-16.

6. Balaji, S., et al. "Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast." Journal of molecular biology 360.1 (2006): 213-227.

7. Han, H., et al. "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions." Nucleic acids research 46.D1 (2018): D380-D386.

8. Hagberg, Aric, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. No. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

9. Pandas Python Library. Link: https://pandas.pydata.org/.

10. Tong, Y.-B., et al. "GenOrigin: A comprehensive protein-coding gene origination database on the evolutionary timescale of life." Journal of Genetics and Genomics (2021).

11. The Saccharomyces Genome Database. Link: https://www.yeastgenome.org/.

12. The UniProt Consortium. “UniProt: the universal protein knowledgebase in 2021”. Nucleic Acids Res. 49 (2021): D1.

13. Letunic, I., Bork, P.. "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation." Nucleic acids research 49.W1 (2021): W293-W296.

14. Bairoch, Amos, and Rolf Apweiler. "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000." Nucleic acids research 28.1 (2000): 45-48.