# Complex Network Analysis and Big Omics Data

[1] Blagoj Ristevski [0000-0002-8356-1203] , [2] Snezana Savoska [0000-0002-0539-1771] and

[3] Pece Mitrevski [0000-0002-0300-7115]

[1, 2, 3] "St. Kliment Ohridski" University - Bitola, Faculty of Information and Communication Technologies - Bitola, ul. Partizanska bb 7000, Republic of Macedonia
[1] blagoj.ristevski@uklo.edu.mk, [2] snezana.savoska@uklo.edu.mk
[3] pece.mitrevski@uklo.edu.mk

**Abstract.** In this paper we describe various biological omics data (e.g. genomics, epigenomics, transcriptomics, proteomics, metabolomics and microbiomics) generated using high-throughput sequencing technologies. These omics data are generated in huge amounts and they follow the 6 V's properties of big data. To discover hidden knowledge from this big omics data, complex network analysis is used. Biological complex networks such as gene regulatory and protein-protein interaction networks are appreciated resources for discovering disease genes and pathways, to investigate topological properties of the most important genes associated with particular disease. The inference of biological regulatory networks from different high dimensional omics data is a fundamental and very complex task in bioinformatics. Various big omics data have a great potential to uncover diverse perspectives of biological complex networks. Taking into account the properties of the big omics data, we suggest some direction for further works.

**Keywords:** Complex Networks · Big Data · Omics Data · Biological Networks.

## 1 Introduction

Currently, there are several omics technologies for the analysis of biological samples such as: genomics, epigenomics, transcriptomics, proteomics, metabolomics, which make analysis on the level of genes, epigenome, transcriptome, proteome and metabolome, respectively [9]. Advances in these omics technologies have enabled personalized medicine at an extraordinarily detailed molecular level [10].

Recent studies have shown that miRNAs are one of the key player of regulation (i.e., many biological processes in metabolism, proliferation, differentiation, development, apoptosis, cellular signaling, cancer development and metastasis).

One of the most essential regulatory role of proteins is transcription regulation. Proteins that bind to DNA sequences and regulate the transcription of DNA, and hence gene expression, are called transcription factors. Transcription factors can inhibit or activate the expression of target genes. Techniques to measure the gene expression level in biological samples span from gene-specific techniques such as quan-

titative polymerase chain reaction (qPCR) to high-throughput technologies such as microarray, serial analysis of gene expression (SAGE) or next-generation sequencing (NGS) (i.e., RNA-Seq) [11].

The inference of biological regulatory networks from different high dimensional omics data is an essential and very complex and computationally demanding issue in bioinformatics that demands high performance computing and development of suitable algorithms. These huge amounts of experimental omics data have the 6 V's properties of big data (volume, value, velocity, variety, veracity, variability) and hence they can be named as big omics data.

Volume refers to the amount of generated and collected data, while the value refers to their coherent analysis, which is valuable to the researchers. Velocity refers to data in motion as well as and to the speed and frequency of data creation, processing and analysis. Complexity and heterogeneity of multiple datasets define the variety of big data. Data quality, relevance, uncertainty, reliability and predictive value describe the veracity of big data, while variability is related to the consistency of data over time [8].

The remainder of this paper is structured as follows. Section 2 describes the concept related to big omics data. The subsequent section describes the complex networks, whereas Section 4 describes the analysis of complex networks in biology, which is based on experimental high-throughput data. The last section provides conclusions and suggestions for future works.

## 2 Big Omics Data

Gene-specific techniques and high-throughput experimental technologies used in bioinformatics enable to be generated a huge amount of various biological omics data. These omics data and their mutual interactions can elucidate at systems and molecular level human diseases, such as cancer, and therefore provide the knowledge necessary to control clinical symptoms, to make better diagnosis, prognosis and to develop more effective treatment of diseases. The combination of these different layers of experimental omics data can help to identify candidate target genes involved in cancer and other diseases.

One of the most suitable manner to obtain better insights into this data, to infer non-observable interactions and to visualize connections, links and groups of entities is network analysis. Using network analysis enables obtaining multi-networks, which are defined as a set of $N$ nodes that interact mutually in $M$ different layers, where each layer reflects a distinct interaction that links the same nodes' pair [1]. Multi-networks combine different layers of experimental data such as genomic, proteomic and molecular interaction data and can be gene regulatory networks, microRNA-target networks, transcription factor-target and protein-protein interaction networks [1]. The aim of multi-network analysis is to discover unknown information on the structure and dynamics of the complex biological systems, as well as to discover which entities or genes are involved in the oncogenic processes or in other diseases [1].

Most common human disease such as obesity, autism, diabetes and schizophrenia are complex diseases, that is they are result of the combination of multiple genetic and

environmental factors [10]. Moreover, also the microbiome has been related to many human diseases, mainly of metabolic origin, and for this reason it is attracting growing interest of scientists for the information it can provide about the onset and progression of these diseases.

The inference of biological regulatory networks from different high-dimensional omics data is a very complex issue in bioinformatics because of the requirement for high-performance computing and suitable algorithms that can be able to work in parallel.

## 3 Complex Networks

Networks are essential tools for modelling and analysis of complex interactions between entities in biology as well as in medicine, physics, neuroscience and technical and social networks. Once extracted from suitable data, networks enable understanding the fundamental structures that control many complex systems by detection of high-order connectivity patterns [2]. Typical represents of these densely connected patterns or network subgraphs, are called networks motifs that can be categorized as feed-forward loops, open bidirectional edges, triangular motifs or two-hop paths.

For instance, biological complex systems transit from one to another state, underlying a variety of phenomena from cell differentiation to recovery from disease [4]. To model behavior of biological systems, the most suitable manner is by using networks, whose nodes stand for dynamic entities such as genes, proteins, microRNAs, metabolites, miRNA-protein complexes, while the links and interactions among them are represented by network edges [4].

Biological complex networks such as gene regulatory and protein-protein interaction networks are appreciated resources for discovering disease genes and pathways, to investigate topological properties of the most important genes associated with particular disease.

Hybrid Functional Petri Nets (HFPNs) have been deliberately introduced to model biological networks [15]. Besides the coexistence of both continuous places associated with real variables (e.g. concentration levels) and discrete places (marked with tokens) allowed in Hybrid Petri Nets (HPNs), several additional features have been included: continuous transition firing rates can depend on the values of the input places and the weights of arcs can be defined as a function of the markings of the connected places [14]. These, in turn, increase the modelling power of HPNs and their application in biological network analysis.

Gene regulatory networks can indeed provide significant insights into the mechanisms of complex diseases such as cancer. However, there are still challenging tasks to reconstruct these networks because of the complex regulatory mechanisms by which genes are influenced such as transcription factors concentration or availability, non-coding RNAs with different regulatory functions but also by the presence of gene or genomic modifications such as mutations, single nucleotide polymorphisms (SNP), copy number variations (CNV) or epigenetic modifications [6] [7]. Taking into account that gene expression is a product of complex interactions of many biological

entities and processes, multi-omics data are required for more reliable networks reconstruction and for better performances.

Since HPNs provide an appropriate way to represent protein concentration dynamics being coupled with discrete switches, HPN modelling of gene regulatory networks has been considered by Matsuno et al. [16]. DNA modification, transcription, translation, post-transcriptional and translational modifications, are a number of stages whose representation would be required by a detailed modelling of gene regulatory networks. Regulatory networks are close controlled at transcriptional and post-transcriptional level and their changes invoke changes in expression levels of genes that participate in protein-protein interactions or that are targets of miRNAs or of other regulatory non-coding RNAs. Indeed, the mutual interaction between miRNAs and their target genes or transcription factors makes miRNAs the most important players in gene regulation. For this reason miRNAs have a primary role in many human diseases, such as cancer, neurological disorders or syndromes, rheumatic, cardiovascular and metabolic diseases [12].

## 4 Complex Networks Analysis

Two main stages of network analysis are network reconstruction and network interrogation. Reconstruction or reverse engineering of biological networks is a data-driven inference of nodes (e.g. genes, proteins, miRNAs, metabolites, miRNA-protein complexes) and edges among nodes. The aim of systematic network analysis, called network interrogation, is to obtain optimal information insights from reconstructed networks. All data-driven interactions could be done using supervised, unsupervised and semi-supervised machine learning techniques [5]. Using these techniques could help to discover hidden patterns from big omics data and to predict complex phenotypes. Reverse engineering of gene regulatory networks, which employs high dimensional gene expression data, demands developing of robust and computational efficient algorithms for network inference. These algorithms should resolve scalability and usability of gene regulatory networks inferred from experimental omics data [3].

Biological networks are evaluated through the analysis of their properties, such as community detection and link prediction.

Community detection in a complex network identifies groups of nodes that are densely connected to each other, but sparsely connected to the nodes that belong to the other groups in the network. To analyze the real network structure and the features of the complex networks as well as the community detection algorithms increasingly become a research challenging topic in the field of complex networks and big data. The identification of resulting communities of densely connected nodes is essential as they may help to reveal a priori unknown functional modules. Most commonly used community detection algorithms are Louvain, Modularity optimization and Infomap. The aim of the Louvain algorithm is to maximize the outcomes of modularity of entire community partition.

In biological networks, the existence of a link between two nodes must be proved by usually very costly laboratory experiments [13]. On the other hand, omics data used in networks reconstruction may contain inaccurate information, which leads to

erroneous link detection. Link prediction algorithms are used to identify these spurious links and to predict actual network links.

## 5 Conclusion

The development of new machine learning techniques and approaches make commonly used machine learning algorithms easy to be adopted by bioinformaticians and to become essential tools for the analysis of big omics data.

As multiple omics technologies can provide a clearer picture of cell functions but also of alterations due to diseases, integration of these technologies with traditional clinical tests will become a routine in future clinical health and disease investigations.

Big omics data of different nature have great potential to analyze complex biological networks under different perspectives. Therefore, the integration of heterogeneous data for the inference of regulatory networks from available biological knowledge actually represents a theme of great interest for computer scientists.

## Acknowledgement

This paper was supported by the Ministry of Education and Science of the Republic of Macedonia and the Ministry of Science and Technology (MOST) of the Government of the People's Republic of China within the bilateral project "Modeling, Simulation and Analysis of Complex Networks in Biology".

## References

1. Cantini L, Medico E, Fortunato S, Caselle M. Detection of gene communities in multinetworks reveals cancer drivers. Scientific reports; 5:17386. (2015)
2. Benson AR, Gleich DF, Leskovec J. Higher-order organization of complex networks. Science;353(6295):163-6. (2016)
3. Fronczuk M, Raftery AE, Yeung KY. CyNetworkBMA: a Cytoscape app for inferring gene regulatory networks. Source code for biology and medicine; 10(1):11. (2015)
4. Dong X, Yambartsev A, Ramsey SA, Thomas LD, Shulzhenko N, Morgun A. Reverse enGENEering of regulatory networks from big data: a roadmap for biologists. Bioinformatics and biology insights;9:BBI-S12467. (2015)
5. Noor E, Cherkaoui S, Sauer U. Biological insights through omics data integration. Current Opinion in Systems Biology. (2019)
6. Zarayeneh N, Ko E, Oh JH, Suh S, Liu C, Gao J, Kim D, Kang M. Integration of multi-omics data for integrative gene regulatory network inference. International journal of data mining and bioinformatics;18(3):223. (2017)
7. Yuan L, Guo LH, Yuan CA, Zhang Y, Han K, Nandi AK, Honig B, Huang DS. Integration of Multi-omics Data for Gene Regulatory Network Inference and Application to Breast Cancer. IEEE/ACM transactions on computational biology and bioinformatics;16(3):782-91. (2018)
8. Ristevski B, Chen M. Big data analytics in medicine and healthcare. Journal of integrative bioinformatics;15(3). (2018)

9. Gallagher IJ, Jacobi C, Tardif N, Rooyackers O, Fearon K. Omics/systems biology and cancer cachexia. In: Seminars in cell & developmental biology; vol. 54, pp. 92-103. Academic Press. (2016)
10. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nature Reviews Genetics;19(5):299. (2018)
11. Ristevski B. Overview of computational approaches for inference of microRNA-mediated and gene regulatory networks. In Advances in Computers; vol. 97, pp. 111-145. Elsevier. (2015)
12. Ristevski B. A survey of models for inference of gene regulatory networks. Nonlinear Anal Model Control; 18(4):444-65. (2013)
13. Lü L, Zhou T. Link prediction in complex networks: A survey. Physica A: statistical mechanics and its applications; 390(6):1150-70. (2011)
14. Alla H, David R. Continuous and hybrid Petri nets. Journal of Circuits, Systems, and Computers; 8(01):159-88. (1998)
15. Matsuno H, Tanaka Y, Aoshima H, Matsui M, Miyano S. Biopathways representation and simulation on hybrid functional Petri net. In silico biology. 3(3):389-404. (2003)
16. Matsuno H, Doi A, Nagasaki M, Miyano S. Hybrid Petri net representation of gene regulatory network. In: Biocomputing; pp. 341-352. (2000)