APPLICATION OF GRAPH MODELS IN BIOINFORMATICS

THESES OF THE PH.D. DISSERTATION

Balázs Ligeti

Roska Tamás Doctoral School of Sciences and Technology
Pázmány Péter Catholic University,
Faculty of Information Technology and Bionics

Supervisor:
Prof. Sándor Pongor

Budapest, 2016

1. Introduction

The network view on biological data has profoundly influenced the ways we are looking at problems of diagnosis and therapy in life sciences today. In traditional paradigms, we used to look at data as isolated entities stored in organized databases. Today, we increasingly consider data as an interconnected network. There are many kinds of connections – for instance drugs can be connected to diseases, to their protein targets, to genes producing the targets, or to drugs they can replace or antagonize. In a similar manner, proteins can be linked to other proteins they physically contact, to genes they regulate, to diseases they play a role in, etc. This is a very complex picture, because we have many types of entities and relationships which are defined in separate ontologies that in turn can be considered as networks of terms. The storage and manipulation of such a large body of data is clearly too demanding for current computers. In addition, such data networks are both incomplete and noisy. Namely, we have a seemingly large number of proteins, but the knowledge of proteins is rarely validated by experiment, and a large part of the annotations are just taken over from homologous proteins of various organisms. Also, we cannot be sure whether or not two proteins are linked in all tissues and/or in all phases of the cell cycle. The solution of these problems is to break down the hypothetical data-network into specific - disease-specific, tissue-

specific, pathogen-specific etc. - manually curated parts which contain reliable information on a given problem. This is a tedious and labor-intensive solution, which is justified only in highly significant fields. Cancer-specific data networks are an example of this approach. In addition, there are two major information sources that can help data-sparsity problems. On the one hand, various high-throughput experimental methods (two-hybrid systems, DNA sequencing, Chip-seq, etc.) provide novel kinds of molecular interaction data,that in principle can be easily added to the existing databases. However, high-throughput data are most often laden with noise, which has to be handled. In such cases hierarchical data networks (i.e. ontologies) may offer a good framework to balance between the reduction of noise and sensitivity to discover novel data links from the experiments.

On the other hand, literature databases containing abstracts or full text of scientific papers provide a large body of new knowledge that can in principle be linked to molecular data. Again, the process is not trivial: scientific texts use natural language, and concepts are often not analogous to the ones used in other texts or in molecular databases.

Disease-specific databases and tools represent a current approach where the above problems are tackled by large communities of scientists. Cancer databases and tools are a typical example, since cancer is one of the most severe complex diseases,

which is responsible for ~15% of all human deaths, and which has >100 more-or-less well-characterized types and >500 human genes associated with it [1]. Oncologists use a variety of traditional databases, but there are a number of data-collection efforts dedicated to the gathering of data on various cancer types. All this provides a solid knowledge base for designing integrated data-networks in which novel questions related to cancer therapy can be answered.

Here I am concerned with three types of questions that can be addressed via integrated data networks: i) finding drug combinations potentially useful for cancer therapy. I tackle this problem by using a simple network overlap measure applied to data networks. And ii) finding novel gene-disease associations in ovarian cancer for generating a list of potential biomarkers. I approach this problem by using a text mining approach applied to MEDLINE abstracts [2] as well as the STRING database [3]. iii) Finally, I present a practical application by testing a dedicated, data subnetwork in accelerating the taxonomic identification. Here I take advantage of the fact that taxonomic and even functional subnetworks are hierarchical graphs, which allows a substantial speedup with respect to current algorithms.

2. Methods

From a logical point of view, all interaction networks and data networks are graphs in which nodes are entities such as molecules, diseases, i.e. biological, physical, as well as conceptual objects, while the edges or links between nodes are relationships, such as molecular interactions, drug-disease connections, drug compatibilities, etc.

This work is concerned with the concept of network neighborhood that can be defined as a subnetwork or subgraph around a selected node. Defining a subnetwork in a data-network can be carried out either by static or dynamic methods using probabilistic approach.

Here it is assumed that an effect propagates from a central node such as a drug target. This is a dynamic approach since the nodes of the network get weighted in an iterative fashion during propagation, and at the end one can select those nodes that have weights exceeding some threshold value. We are concerned with two kinds of propagation algorithms used in several fields of computer science, PageRank [4-6] and diffusion [7-10].

The PageRank algorithm is a special case of random walk on data network: a walker starts at a certain data node, then randomly selects the next node from its neighbor, then moves there, and so on. In the case of PageRank the walker not only selects a neighboring node randomly, but it can also move to any other

nodes with a certain probability (“restart probability”). If the walker is only allowed to move to specific set of nodes or to the neighboring nodes, then this is the PageRank with prior algorithm [5, 6, 11]. If there is prior knowledge available about which nodes are more relevant, then one can use this information to bias the original PageRank scores. Other well-known algorithms based on random walks include k-step Markov [11], HITS [12], and HITS with Prior [11].

Diffusion is a physical metaphor used to model transport phenomena on networks. In our case, we assign an imaginary quantity, such as “energy” or “drug action” to one node of the network – for instance the gene targeted by the drug – and then use an iterative process to compute how this quantity diffuses along the network.

In a similar way to PageRank with prior,? it is possible to incorporate prior knowledge about the data network, i.e. relevant drugs to a disease by regularizing the Laplacian matrix [7]. The regularization could be interpreted as alteration of diffusion process by i.) controlling (increasing or decreasing) the energy loss of a node, ii.) altering (increasing or decreasing) the input energy flow on certain edges, iii.) both of the above. All of the mentioned alterations can be described with different regularization parameters.

The evaluation of a large system of ordinary differential equations could be a challenging task, like it is in the case of diffusion; however, by using sparse linear algebra and leveraging the sparseness of a typical data network, the solution could be computed in reasonable time. Instead of computing the matrix exponential, one could focus on the approximation of the matrix-vector product gaining a significant speedup. The expression $e^{-Lt}x(0)$ could be approximated by using iterative methods such as Arnoldi algorithm [13-15].

Both PageRank and diffusion methods require the estimation of a threshold value below the nodes (and their respective edges) are omitted from the network neighborhood. This can be carried out by standard Monte-Carlo simulations in which a large number, say ten thousand of iterations are started from randomly selected nodes of the network, and values significantly, say $p<<0.05$, higher than the background are selected as members of the neighborhood.

3. New Scientific Results

I. Prediction of efficient drug combinations

Related publications of the author: [J1][J3][C1]

Drug combinations are highly efficient in systemic treatment of complex multigene diseases such as cancer, diabetes, arthritis, or hypertension. Most of the currently used combinations were found in empirical ways, which limits the speed of discovery for new and more effective combinations.

THESIS I.1. I have developed a novel drug combination prediction method based on the assumption that a perturbation generated by multiple drugs propagates through an interaction network and the drugs may have unexpected effect on targets not directly targeted by either of them (Figure 1). I have introduced a new index, the so-called Target Overlap Score (TOS), to capture this phenomenon. The score quantifies the potential amplification effect as the overlap between the affected subnetworks. The score is computed as the Jacquard or Tanimoto coefficient between the sets of nodes in the subnetworks, net_1 **and** net_2 **:**

$$TOS(net_1, net_2) = \frac{\left|V_{net_1} \cap V_{net_2}\right|}{\left|V_{net_1} \cup V_{net_2}\right|}$$

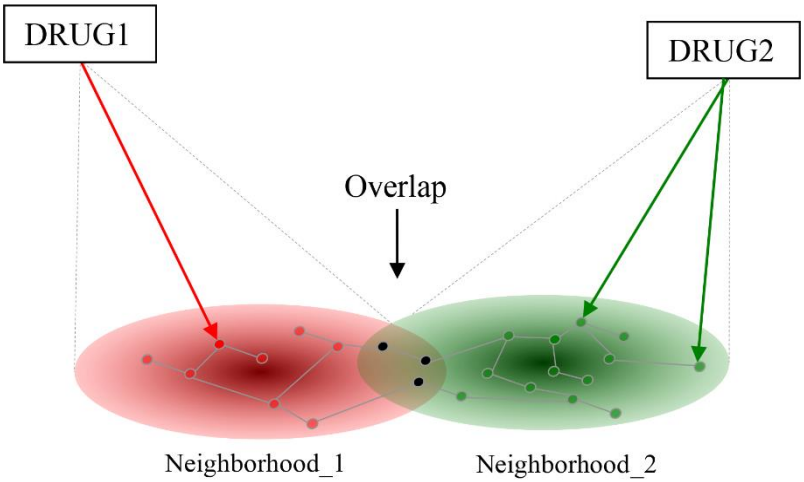

Figure 1. *The effect of two drugs (Drug$_1$, Drug$_2$) reaches their imminent targets first (arrows) and the effect will then propagate to their network neighborhoods (subnetworks) indicated in red and green, respectively. Targets in the overlap are affected by both drugs, and we suppose that drugs affecting a number of common targets will influence the effects of each other. The overlap is quantified as the proportion of jointly affected targets within all affected targets (in set theory terms: intercept divided by union).*

THESIS I.2. I have showed that by using the TOS score it is possible to distinguish both the drug-drug interactions and the drug combinations from random combinations. I also presented that this measure is correlated with the known effects of beneficial and deleterious drug combinations taken from the DCDB, TTD and Drugs.com databases (Figure 2).

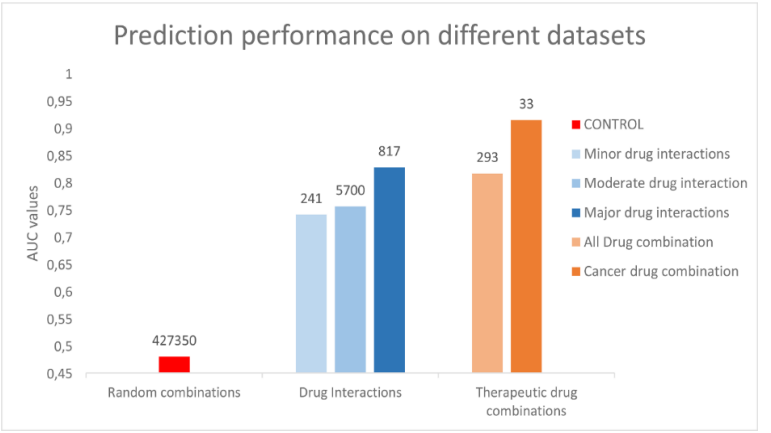

Figure 2. *The prediction performance was measured on several different training sets, cancer related drug-drug interactions and drug combinations. The prediction method is based on a simple measure, the Target Overlap Score (TOS). The prediction procedure was repeated 100 times with different negative sets, then the average value was reported. The standard deviation of AUC values (not shown) are between 0.0001 and 0.006 for the different datasets.*

Here the prediction is a ranking, in which the efficient combinations are expected to be in the top of the list. The performance, namely how good a ranking is, was characterized with the AUC value. This score is 1 if the ranking is perfect (i.e. all efficient combinations ranked at the top), 0.5 if it is random. Drug - drug interactions are often considered as harmful "negative combinations", since they increase the risk of side effects and may cause "overdose". On the other hand, drug combinations are considered to be desirable (positive) since they can be efficiently used in the treatment of complex diseases. We could show that a simple network overlap measure is well correlated with the

intensity of positive and negative drug interactions as well as with clinical data.

THESIS I.3. I have also investigated that combining two frequently used drug-drug similarity measures with TOS - namely the functional similarity of drugs computed based on their imminent targets, and their therapeutic similarity quantified by using the anatomical therapeutic chemical (ATC) classification system - does not improve the classification performance.

The prediction of drug-drug interactions is expected to improve by incorporating more and more information about drugs. One of the most often successfully used descriptors for drugs are functional annotations (i.e. assigning ontology terms to drugs via their targets), and the therapeutic similarity quantified based on ATC code annotation. The trained classifier (logistic regression model) with various measures (TOS, GO, ATC) does not show any improvement in terms of ranking performance compared to the ranking defined by TOS alone.

THESIS I.4. I have demonstrated the utility of TOS by correlating the score to the outcome of recent clinical trials evaluating trastuzumab, an effective anticancer agent used in combination with anthracycline- and taxane-based systemic chemotherapy in HER2-receptor (erb-b2 receptor tyrosine kinase 2) positive breast cancer.

I have compared the combinations proposed in the treatment of breast cancer being under clinical research (Phase II, Phase III, Phase IV studies) with the predicted TOS score. I narrowed the search to the investigation that implemented the RECIST (Response Evaluation Criteria In Solid Tumors). The TOS showed good correlation with several investigated response variables such as overall response ($r=0.64$; $p=0.0028$), overall survival rate ($r=0.87$; $p=0.017$), and confirmed clinical benefit ($r=0.84$; $p=0.0021$).

II. Prediction of cancer biomarkers by integrating text and data networks

Related publications of the author: [J1][J6]

Text mining methods can facilitate the generation of biomedical hypotheses by suggesting novel associations between diseases and genes. Previously, we had developed a rare-term

model called RaJoLink ([16]) in which hypotheses are formulated on the basis of terms rarely associated with a target domain.

THESIS II.1. I have improved the sensitivity of the RaJoLink rare term based algorithm by using network analysis algorithm such as personalized diffusion ranking and PageRank with Prior on the STRING protein-protein association network.

Since many current medical hypotheses are formulated in terms of molecular entities and molecular mechanisms, here we extend the methodology to proteins and genes using a standardized vocabulary as well as a gene/protein network model. The proposed enhanced RaJoLink rare-term model combines text mining and gene prioritization approaches. Its utility is illustrated by finding known, as well as potential gene-disease associations in ovarian cancer using MEDLINE abstracts and the STRING database.

THESIS II.2. Based on the enhanced prediction I proposed 10 novel genes - RUNX2, SOCS3, BCL6, PAX6, DAPK1, SMARCB1, RAF1, E2F6, P18INK4C (CDKN2C), and PAX5 - that are likely to be related to the disease and at the time had not been described as such. Since 2012, two of them (RUNX2, BCL6) have been confirmed.

The RUNX2 transcription factor is a putative tumor suppressor gene. It has also been associated with many cancer types inculding prostate, lung, breast cancer, osteosarcoma, thyroid tumors. The potential of the gene in this association is supported by the prognostic power of hormone receptors in ovarian cancer [17]. In 2012 it was also confirmed that RUNX2 is associated with advanced tumor progression in epithelial OC [18]. In addition, the inhibition of RUNX2 lead to the significant decrease of cell proliferation.

BCL6 (B-cell CLL/lymphoma 6) is another transcription factor found to be frequently mutated in diffuse large-cell lymphoma. The gene was related not only to lymphomas and leukemias, but also to progression to breast, gastric and lung cancer. Interestingly, both BCL6 and RUNX2 are influenced by prolactin secretion. Wang et al. showed the BCL6 is a negative prognostic factor in ovarian cancer [19] and the inhibition of

BCL6 along with NACC1 [20] reduced the invasion capabilities of cancer cells.

III. Fast and sensitive characterization of microbial studies

Related publication of the author: [J5]

Next generation sequencing (NGS) of metagenomic samples is becoming a standard approach to detect individual species or pathogenic strains of microorganisms. Computer programs used in the NGS community have to balance between speed and sensitivity and as a result, species or strain level identification is often inaccurate and low abundance pathogens can sometimes be missed.

THESIS III.1. I have demonstrated that using hierarchical networks such as taxonomy along with fast aligners, i. e. bowtie2, the evaluation of high-throughput sequencing data is feasible in a reasonable time with good classification accuracy. The algorithm assigns the individual reads to the common ancestor of the taxa having its genome hit by the short read (Figure 3).

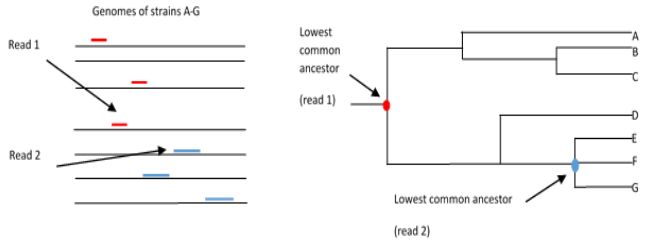

Figure 3. The Taxoner algorithm
In the first step the short reads were mapped to the microbial genomes. In the next step the alignments were preprocessed and only those hits were mapped to the taxonomy tree, which were above a certain threshold. In the classification or the binning step the read was assigned to the lowest common ancestor of the taxa it hit.

THESIS III.2. I have illustrated the applicability of the Taxoner principles on whole genome shotgun sequencing of known or unknown pathogens (*Staphylococcus aureus*, *Bacillus anthracis*). The results suggested that the performance of Taxoner is as good as the state-of-the-art BLAST-based methods, while it is faster by two orders of magnitude. Finally, it is also compatible with various sequencing platforms.

In the recent years various sequencing platforms have been developed. We compared the classification performance of Taxoner, Metaphlan [21] and BLAST [22, 23] combined with Megan [24, 25] on whole genome sequencing datasets of *Staphylococcus aureus* produced by Roche 454, Ion Torrent and

Illumina. The low false negative rates implicate that Taxoner is almost as reliable as BLAST+Megan, however, it requires comparatively less computational power and time.

THESIS III.3. I have proven that using the Taxoner principles it is possible to characterize the microbial communities at the lowest taxonomic level, even in species or strain level.

I analyzed the MOCK dataset representing 22 microbial strains and species in equal amounts provided by the Human Microbiome Project for validation purposes. The dataset consists of 6.5 Illumina short-reads. Taxoner was capable of confidently detecting most of the taxa (14/22) even in strain level.

THESIS III.4. Taxoner is sensitive and capable of identifying taxa being present only in small abundance, furthermore, it needs two orders of magnitude less reads to complete the identification than MetaPhLan. In addition, the method is applicable to cases where the genome sequence of the studied microbe is unknown.

In the application to pathogen detection the sensitivity of the analysis is a crucial question. The sensitivity was measured as the

number of reads necessary for detecting a certain species. After randomly sampling an experimental anthrax dataset, the analysis revealed that Taxoner could confidently identify the anthrax from 10 reads, while MetaPhLan needed 200-350 reads.

Sensitive detection of microorganisms with unknown sequence is a crucial question as well, since the majority of them are still unknown. In order to assess the classification performance on unknown species, whose genome sequence is missing from the database, I have analyzed an experimental anthrax dataset (*B. anthracis* strain BA104; NCBI taxon id: Not Available). Taxoner classified the majority of reads (96.50%) as *Bacillus anthracis*, a small portion of 1.2% was classified as other species from the *Bacillus* genus.

4. Application of the results

In my research? I have investigated how different graph models can help in various bioinformatics problems. My research covered the following areas: i) finding novel drug combinations, ii) finding novel, unexpected biomarkers from literature, and iii) improving the classification performance of metagenomics reads.

Network analysis strategies are not only helpful in discovering novel associations between diseases and genes, but could predict new beneficial interactions of drugs as well, thus making it possible to design better treatment for cancer patients.

I have demonstrated that the extension of text mining with network analysis can help in identifying novel biomarkers for ovarian cancer. Indeed, since 2012, two out of ten completely novel associations highlighted by the algorithm have been confirmed by other studies [18-20, 26].

The application of graph models is not limited to exploring the disease-gene-drug relationships. It also includes the analysis strategies of high-throughput data, such as the evaluation of metagenomics datasets.

I have managed to demonstrate that applying network principles may help us with exploring unexpected and non-trivial relationships between drugs, diseases and microbes.

5. Publications

[J1] **Ligeti, B.,** Menyhárt, O., Petrič I., Győrffy, B.; Pongor, S. (2016). Propagation on Molecular Interaction Networks: Prediction of Effective Drug Combinations and Biomarkers in Cancer Treatment. *Current Pharmaceutical Design*, accepted

[J2] **Ligeti, B**.; Vera, R.; Juhász, J.; Pongor, S. (2016). CX, DPX and PCW: Web servers for the visualization of interior and protruding regions of protein structures in 3D and 1D. *Springer Protocols: Methods in Molecular Biology*, in press.

[J3] **Ligeti, B.**; Pénzváltó, Z.; Vera, R.; Győrffy, B.; Pongor, S. (2015). A Network-Based Target Overlap Score for Characterizing Drug Combinations: High Correlation with Cancer Clinical Trial Results. *PLoS One*. **10** (9), e0129267.

[J4] Hudaiberdiev, S.; Choudhary, K.; Vera, R.; Gelencsér, Zs.; **Ligeti**, B.; Lamba, D.; Pongor, S. (2015); Census of solo LuxR genes in prokaryotic genomes. *Front. Cell. Infect. Microbiol*. 5:20. doi:10.3389/fcimb.2015.00020

[J5] Pongor, L. S.; Vera, R.; **Ligeti** B. (2014). Fast and Sensitive Alignment of Microbial Whole Genome Sequencing Reads to Large Sequence Datasets on a Desktop PC: Application to Metagenomic Datasets and

Pathogen Identification. *PLoS One*, published 31 Jul 2014, 10.1371/journal.pone.0103441

[J6] Petrič, I.; **Ligeti**, B.; Győrffy, B.; Pongor, S. (2014). Biomedical Hypothesis Generation by Text Mining and Gene Prioritization. *Protein Pept Lett*. 21 (8), 847-857.

[C1] **Ligeti**, B.; Vera, R.; Lukács, G.; Győrffy, B.; Pongor, S. (2013). Predicting effective drug combinations via network propagation. *Biomedical Circuits and Systems Conference*, 378-381.

[J8] Vera, R.; Perez-Riverol, Y.; Perez, S.; **Ligeti**, B.; Kertész-Farkas, A.; Pongor, S. (2013). JBioWH: an open-source Java framework for bioinformatics data integration. *Database*. 2013, bat051.

6. References

[1] A. Pavlopoulou, D. A. Spandidos, and I. Michalopoulos, "Human cancer databases (Review)," *Oncology reports*, vol. 33, pp. 3-18, 2015.

[2] I. Petric, B. Ligeti, B. Gyorffy, and S. Pongor, "Biomedical hypothesis generation by text mining and gene prioritization," *Protein and peptide letters*, vol. 21, pp. 847-857, 2014.

[3] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, *et al.*, "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Res*, vol. 41, pp. D808-15, Jan 2013.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the web," 1999.

[5] T. H. Haveliwala, "Topic-sensitive pagerank," in *Proceedings of the 11th international conference on World Wide Web*, 2002, pp. 517-526.

[6] G. Jeh and J. Widom, "Scaling personalized web search," in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 271-279.

[7] T. Ito, M. Shimbo, T. Kudo, and Y. Matsumoto, "Application of kernels to link analysis," *Proceedings of the eleventh ...,* pp. 586-592, 2005.

[8] J. Kandola, J. Shawe-Taylor, and N. Cristianini, "On the application of diffusion kernel to text data," Technical report, Neurocolt, 2002. NeuroCOLT Technical Report NC-TR-02-1222002.

[9] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*: Cambridge university press, 2004.

[10] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2002, pp. 315-322.

[11] S. White and P. Smyth, "Algorithms for estimating relative importance in networks," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 266-275.

[12] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, pp. 604-632, 1999.

[13] M. Eiermann and O. G. Ernst, "A restarted Krylov subspace method for the evaluation of matrix functions," *SIAM Journal on Numerical Analysis*, vol. 44, pp. 2481-2504, 2006.

[14] C. Moler and C. Van Loan, "Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later," *SIAM review*, vol. 45, pp. 3-49, 2003.

[15] Y. Saad, "Analysis of some Krylov subspace approximations to the matrix exponential operator," *SIAM Journal on Numerical Analysis*, vol. 29, pp. 209-228, 1992.

[16] I. Petrič, T. Urbančič, B. Cestnik, and M. Macedoni-Lukšič, "Literature mining method RaJoLink for uncovering relations between biomedical concepts," *J Biomed Inform*, vol. 42, pp. 219-227, 2009.

[17] T. Fekete, E. Rásó, I. Pete, B. Tegze, I. Liko, G. Munkácsy, *et al.*, "Meta-analysis of gene expression profiles associated with histological classification and survival in 829 ovarian cancer samples," *International Journal of Cancer*, vol. 131, pp. 95-105, 2012.

[18] W. Li, Z. Liu, L. Chen, L. Zhou, and Y. Yao, "MicroRNA-23b is an independent prognostic marker and suppresses ovarian cancer progression by targeting runt-related transcription factor-2," *FEBS letters*, vol. 588, pp. 1608-1615, 2014.

[19] Y.-Q. Wang, M.-D. Xu, W.-W. Weng, P. Wei, Y.-S. Yang, and X. Du, "BCL6 is a negative prognostic factor and exhibits pro-oncogenic activity in ovarian cancer," *American journal of cancer research*, vol. 5, p. 255, 2015.

[20] W. Shan, J. Li, Y. Bai, and X. Lu, "miR-339-5p inhibits migration and invasion in ovarian cancer cell lines by targeting NACC1 and BCL6," *Tumor Biology*, vol. 37, pp. 5203-5211, 2016.

[21] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, "Metagenomic microbial community profiling using unique clade-specific marker genes," *Nat Methods*, vol. 9, pp. 811-4, Aug 2012.

[22] S. A. Shiryev, J. S. Papadopoulos, A. A. Schaffer, and R. Agarwala, "Improved BLAST searches using longer words for protein seeding," *Bioinformatics*, vol. 23, pp. 2949-51, Nov 1 2007.

[23] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences," *J Comput Biol*, vol. 7, pp. 203-14, Feb-Apr 2000.

[24] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Res*, vol. 17, pp. 377-86, Mar 2007.

[25] D. H. Huson, D. C. Richter, S. Mitra, A. F. Auch, and S. C. Schuster, "Methods for comparative metagenomics," *BMC Bioinformatics*, vol. 10 Suppl 1, p. S12, 2009.

[26] Z.-Q. Wang, M. Keita, M. Bachvarova, S. Gobeil, C. Morin, M. Plante, *et al.*, "Inhibition of RUNX2 transcriptional activity blocks the proliferation, migration and invasion of epithelial ovarian carcinoma cells," *PloS one*, vol. 8, p. e74384, 2013.