AI-based Integrated Approach for Gene Protein Regulatory Network

Md Wasi Ul Kabir[¶], Duaa Alawad[¶], Krishna Shah, Nayan Howladar, Muhammad Farooq, Sofiane Benkara, Amrit Rajbhandari, Dang Pham, Joshua Pierre, Md Tamjidul Hoque[*]

Department of Computer Science, University of New Orleans, New Orleans, LA, USA.

*To whom correspondence should be addressed: email: thoque@uno.edu (MTH)

[¶]These authors contributed equally to this work as first authors.

Abstract

Gene regulatory networks (GRNs) play an important function in a variety of cellular processes and pathways. Recent advancements in high-throughput biological data gathering have provided fresh platforms for studying GRNs, generating significant interest in the mathematical modeling of biological networks. An effective GRN inference algorithm can find proper regulatory connections among genes, greatly facilitating discovering the basics of how a cell functions and functions, allowing for a unique and improved understanding of disease initiation and progression. However, the involvement of protein level interactions can refine and validate some of the interactions that are inferred from gene-gene interactions, and hence we call our focus a gene-protein regulatory network (GPRN). Consequently, this paper proposes a method to select genes that produce proteins that bind to DNA (deoxyribonucleic acid), RNA (ribonucleic acid), or other proteins and carbohydrate. The selected genes are used to create three different gene protein regulatory networks based on different bindings. Finally, we show a method to filter those relations(edges) by using pairwise protein-DNA interaction, protein-RNA interaction and protein-protein interaction tools to reduce the noise of the regulatory network.

Introduction

The nucleus of a cell contains all a live organism's genetic information (genome). That genome is kept in chromosomes, which are lengthy DNA molecules (deoxyribonucleic acid). In human cells, there are 23 pairs of chromosomes [1]. DNA is the material that makes up genes. The basic physical and functional unit of heredity is the gene. Human beings have between 20000 and 25000 genes [2]. Only 3% of the DNA in their bodies was translated into proteins, and the remaining genes are thought to be involved in regulating other genes [2]. Genes store information about how we look and how our cells function inside our bodies. Some genes serve as blueprints for creating protein-like molecules. Many genes, on the other hand, do not produce proteins. Instead, each gene performs a unique function in our body. A gene's DNA contains instructions for the cell's production of proteins. Figure 1 shows an illustration of cells, chromosomes, and genes.

Proteins are the fundamental building components of our body. Proteins are found in bones, teeth, hair, muscles, and blood. Those proteins assist our bodies in growing, functioning effectively, and remaining healthy[3]. Each gene in the human may produce up to ten distinct proteins[4]. That means the human will have over 300,000 proteins [5].

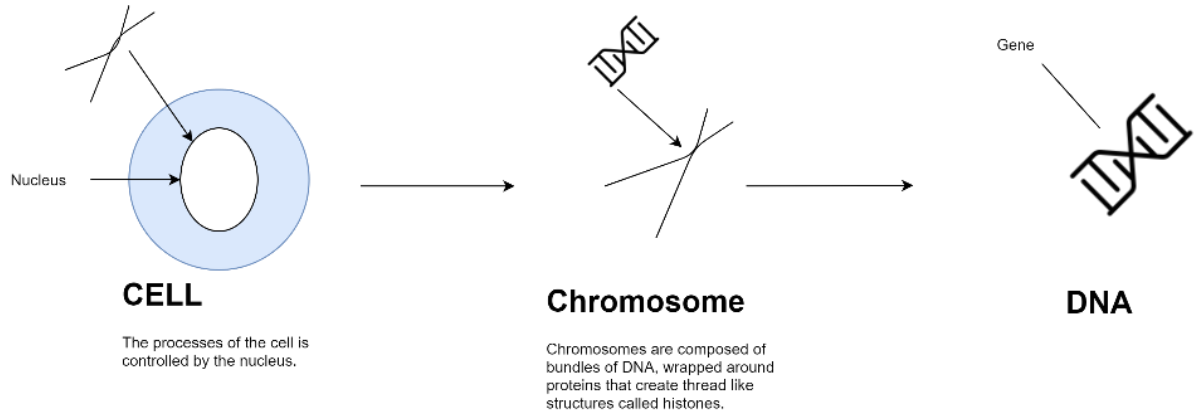

Figure 1. Illustration of cell, chromosomes, and genes.

The transfer of genetic information in cells from DNA to messenger RNA (mRNA) to protein is described by the central dogma of molecular biology. According to the theory, genes determine the sequence of mRNA molecules, which specify the sequence of proteins. Because the information held in DNA is important to cellular function, the cell safeguards it by copying it as RNA [6]. For every nucleotide it reads in the DNA strand, an enzyme adds one to the mRNA strand. Because three mRNA nucleotides correspond to one amino acid in the polypeptide sequence, translating this information to a protein is more complicated. The conversion process of DNA to protein is described in the following paragraphs.

A) Transcription:

Transcription is the process by which information is transferred from one strand of DNA to RNA by the enzyme RNA Polymerase. The DNA strand that goes through this process is made up of three parts: a promoter, a structural gene, and a terminator [7]. The strand of DNA that synthesizes the RNA is known as the template strand, while the other strand is known as the coding strand. The promoter is bound by the DNA-dependent RNA polymerase, which catalyzes polymerization in the 3′ to 5′ direction. RNA polymerase reads a DNA sequence during transcription and generates a corresponding, antiparallel RNA strand. Post-transcriptional alterations are performed on the freshly released RNA strand [1].

Unlike DNA replication, transcription produces an RNA complement that replaces the RNA uracil (U) in all places where the DNA thymine (T) would have occurred. Thus, transcription is the initial step in gene expression. A transcript is a segment of DNA that has been transcribed into an RNA molecule. Some transcripts serve as structural or regulatory RNAs, while others encode one or more proteins. If the transcribed gene encodes a protein, the outcome of transcription is messenger RNA (mRNA), which is subsequently utilized to produce that protein during the translation process.

B) Translation

The translation is how mRNA is decoded and translated into a polypeptide sequence, also known as a protein. This type of protein synthesis is controlled by the mRNA and performed with the assistance of a ribosome, a huge complex of ribosomal RNAs (rRNAs) and proteins. During translation, a cell decodes the genetic message of the mRNA and assembles the brand-new polypeptide chain [6]. Transfer RNA, or tRNA, translates the sequence of codons on the mRNA strand. The primary purpose of tRNA is to transport a free amino acid from the cytoplasm to a ribosome connected to the developing polypeptide chain. tRNAs continue to add amino acids to the expanding end of the polypeptide chain until they reach a stop codon on the mRNA. The ribosome then delivers the finished protein into the cell. Figure 2 shows the steps of transcription and translation to produce proteins from genes.

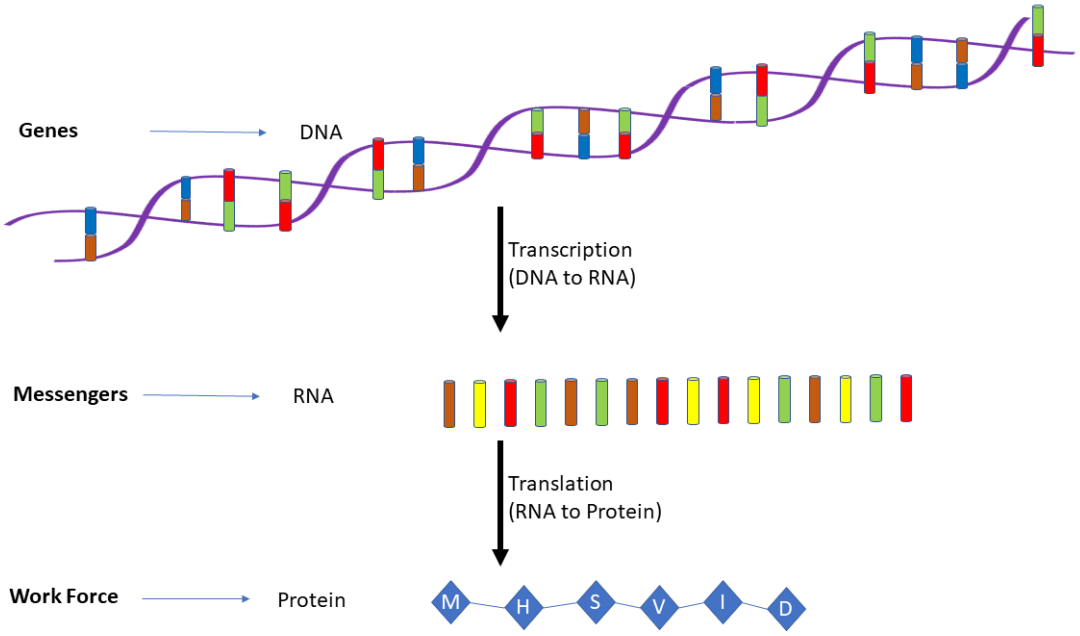

Figure 2. Steps of transcription and translation of genes.

Gene Regulatory Network (GRN)

Gene regulation is the process of regulation in which a set of genes in a cell are expressed (turned on) to make a functional product, i.e., proteins. Though almost all the cells in our body have the same DNA, each cell has a different set of active genes, transforming into different sets of functional RNAs and proteins to do some specialized tasks. For example, the human stomach cell creates acids, whereas the liver cell produces enzymes that break alcohol into a non-toxic molecule.

A gene regulatory network (GRN) represents gene regulation using a set of genes and their relevant regulatory molecules that interact with each other to determine the function of a cell. GRN represents a directed graph that the regulators of gene expression are connected to target gene nodes, and those interactions can be represented as edges[8]. Regulators of gene expression constitute transcription factors (TF) that can be worked as activators and repressors, RNA binding proteins, regulatory RNAs, etc. [9]. It is important to determine regulatory relationships between transcription factors and their target genes to

understand biological phenomena ranging from cell growth and division to cell differentiation and development [10]. Therefore, reconstruction of GRNs is essential to understand how gene expression dysregulation causes some diseases such as cancer [11]. Figure 3 shows the gene expressions and the gene regulatory network.

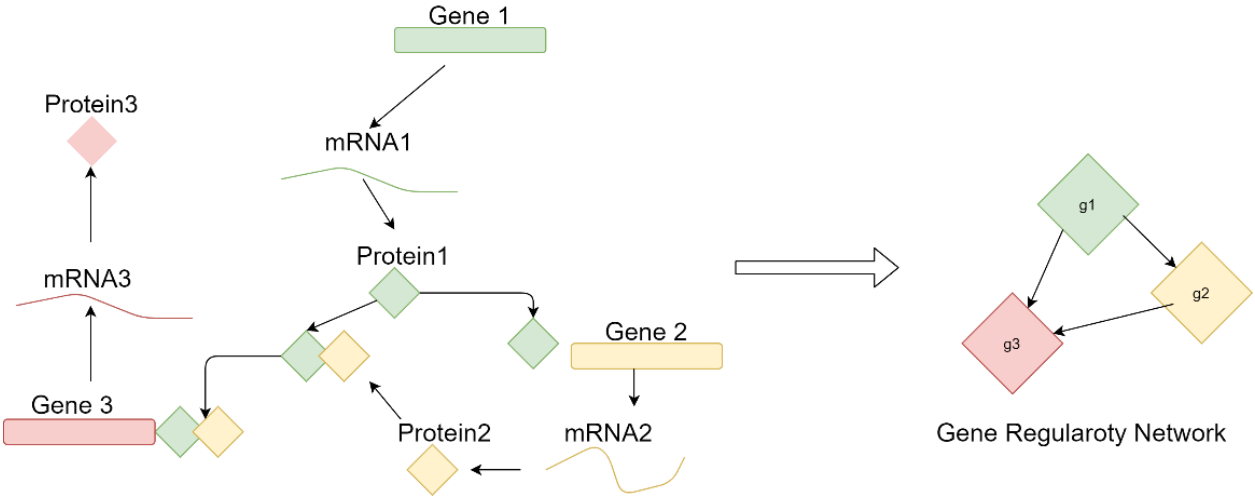

Figure 3. Example of a gene regulatory network.

Gene Protein Regulatory Network (GPRN)

Gene Regulatory Network (GRN) can be a complex network. Each cell has a different network as they have different regulations depending on the cell function. This paper proposed a gene protein regulatory network (GPRN) based on the protein interactions with DNA, RNA, and other proteins. Genes are translated into proteins, and the proteins can be classified into different classes such as DNA binding proteins, RNA binding proteins, and protein-binding proteins. Based on the binding information, we can select a subset of active genes or regulating genes and find the interaction between regulating genes with other genes. The proposed method will generate three different networks for three types of bindings. In the following paragraphs, we describe three different gene regulatory networks.

GPRN based on Protein-DNA interactions

The first network is based on the Protein-DNA interactions. First, we extract the canonical protein sequence from a set of genes. A canonical sequence is DNA, RNA, or amino acids that reflect the most common choice of base or amino acid at each position. In canonical sequence, each gene is represented by one single protein sequence. We extract the canonical sequence from the UniProtKB dataset. After extracting the protein sequence, the DNA binding prediction tools [12] are used to select the genes which are producing DNA binding proteins or transcription factors. The target genes are turned on or off by these DNA binding proteins/transcription factors. Then, a gene regulatory network is inferred using the same idea from the author of the GENIE3 algorithm [13]. The GENIE3 algorithm infers the regulators for each target gene from the expression data. The importance of each of the transcript factors for activating a certain gene is calculated using a tree-based regression method. The inferred network can be noisy as it is created based on the expression data only, and it may contain many edges with false positives. To further purify the network, we can obtain the pairwise Protein-DNA interaction probability using the existing tools, i.e.,

DNAproDB [14], FoldX [15]. Figure 4 shows the steps of the gene protein regulatory network based on Protein-DNA interactions.

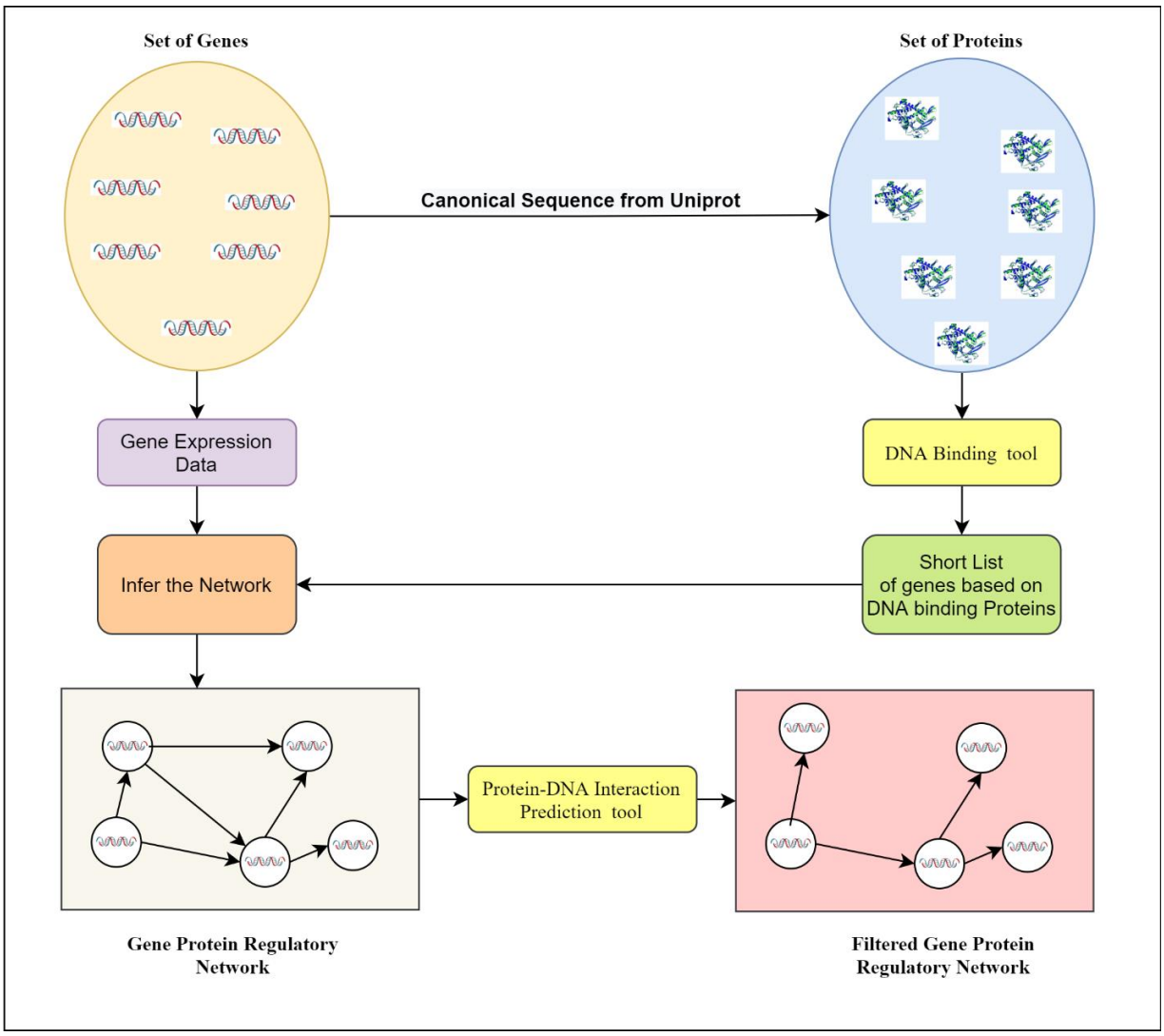

Figure 4. Gene Protein Regulatory Network based on Protein-DNA interactions.

GPRN based on Protein-RNA interactions

The second network is based on the Protein-RNA interactions. Similar to previous architecture, we extract the canonical sequence from the set of genes. Then, a set of active genes based on the interaction between a protein with RNA using the RNA binding prediction tool [16]. Likewise, we can create the network from the expression data and short-listed genes. Further, the network can be filtered using the Protein-RNA interaction probabilities. We can obtain the probabilities using Protein-RNA interaction prediction tools, i.e., RNAct [17]. Figure 5 shows the steps of the gene protein regulatory network based on Protein-RNA interactions.

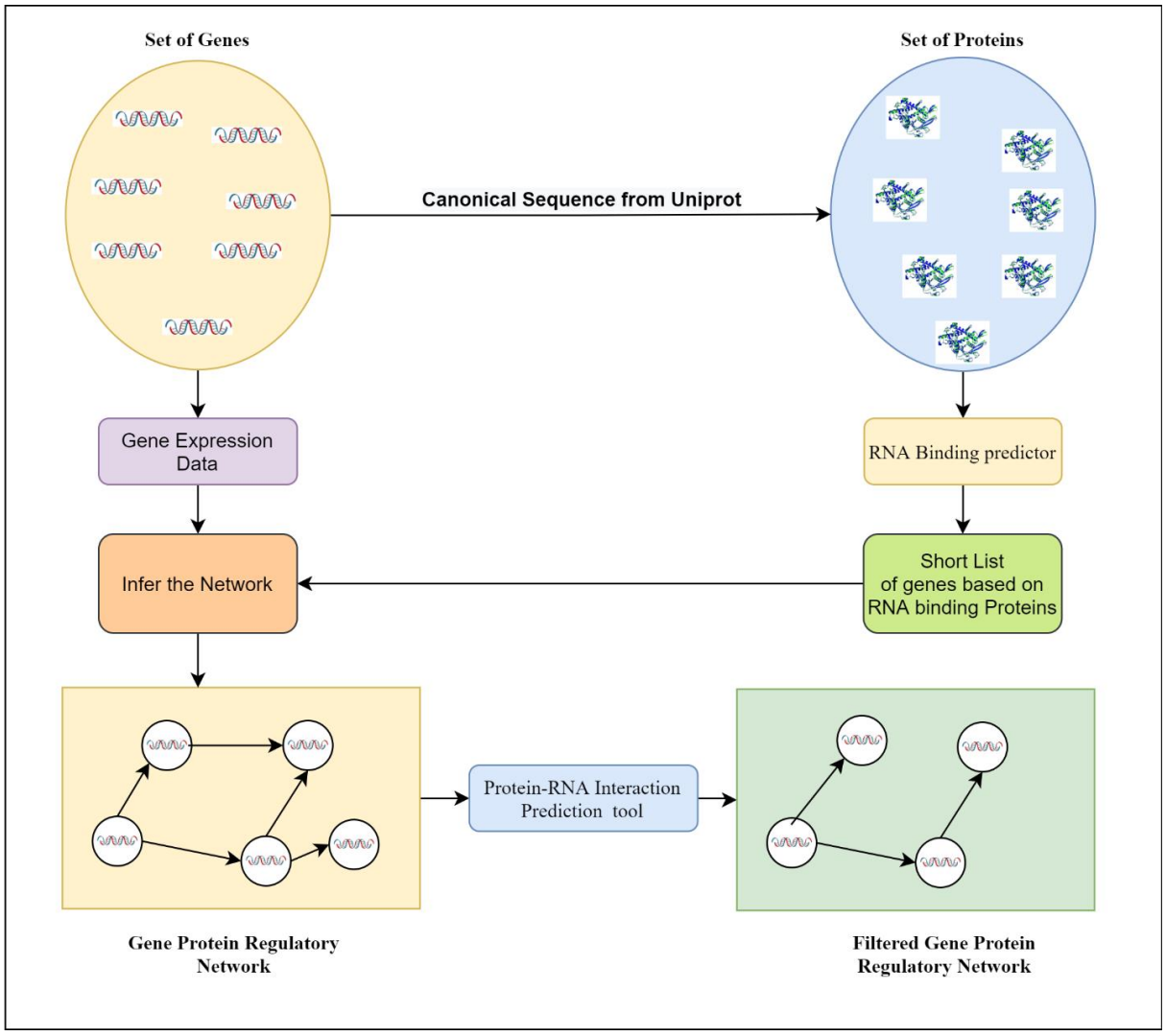

Figure 5. Gene Protein Regulatory Network based on Protein-RNA interactions.

GPRN based on Protein-Protein interactions

The third network is based on the Protein-Protein interactions. Like the previous two architectures, we extract the canonical sequence from the set of genes. First, we select a set of active genes based on the interaction between a protein with other proteins using the protein binding prediction tool [18]. Then, we create the network from the expression data and short-listed genes that produce proteins that bind with other proteins. Further, the network can be filtered using the Protein-Protein interaction probabilities. We can obtain the probabilities using Protein-Protein pairwise interaction prediction tools, i.e., published in [19-21]. Figure 6 shows the steps of the gene protein regulatory network based on Protein-RNA interactions.

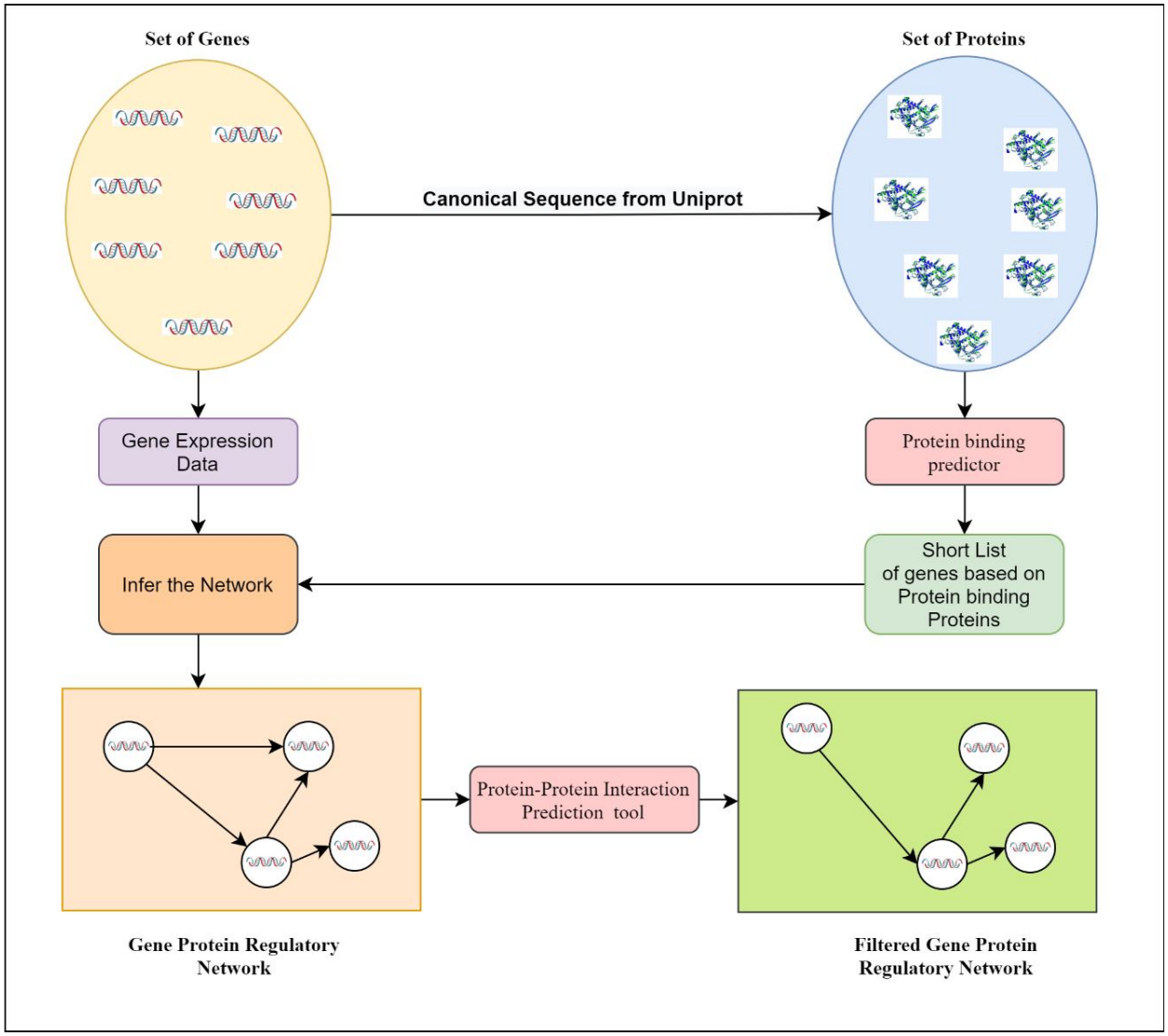

Figure 6. Gene Protein Regulatory Network based on Protein-Protein interactions.

Prediction tools:

We proposed a gene protein regulatory network based on only protein interacting with DNA, RNA, and other proteins. It is possible to create a network based on other types of binding, i.e., Carbohydrate-Binding proteins, or to create a network with the transposable elements. Here is a brief description of some of the tools that can be used to predict the properties of a given protein or genes.

DNA binding proteins

In our lab, we developed a stacking-based machine learning method, called StackDPPred, to effectively predict DNA-binding proteins [12]. DNA-binding proteins play an important role in various biological processes such as DNA replication, repair, transcription, and splicing. The corresponding code and data can be found here: code and data

RNA binding proteins
We developed a method called AIRBP, which is designed using an advanced machine learning technique called stacking to effectively predict RBPs by utilizing features extracted from evolutionary information, physicochemical properties, and disordered properties[16]. RBPs play crucial roles in post-transcriptional control of RNAs and RNA metabolism and have diverse roles in various biological processes such as splicing, mRNA stabilization, mRNA localization and translation, RNA synthesis, folding-unfolding, modification, processing, and degradation. The corresponding code and data can be found here: code and data.

Peptide-Binding proteins
Peptide-recognition domains (PRDs) are critical as they promote coupled binding with short peptide motifs of functional importance through transient interactions. Therefore, we developed a machine-learning-based tool, named PBRpredict, to predict residues in peptide-binding domains from protein sequence alone [18]. The proposed predictor is found competitive based on statistical evaluation. The corresponding code and data can be found here: code and data.

Carbohydrate Binding proteins
The study of protein-carbohydrate interactions at the residue level is useful in treating many critical diseases. We developed a balanced predictor called StackCBPred to predict protein-carbohydrate binding sites[22]. The corresponding code and data can be found here: code and data.

Transposable Elements
Transposable Elements (TEs) or jumping genes are the DNA sequences that have an intrinsic capability to move within a host genome from one genomic location to another. Studies show that the presence of a TE within or adjacent to a functional gene may alter its expression. TEs can also cause an increase in the rate of mutation and can even mediate duplications and large insertions and deletions in the genome, promoting gross genetic rearrangements. We developed a robust approach for the hierarchical classification of TEs, with higher accuracy, using Support Vector Machines (SVM)[22]. The corresponding code and data can be found here: code and data.

The above tools are publicly available in the webserver - https://bmll.cs.uno.edu/. Figure 7 shows the interface which can be used to submit protein sequences for prediction.

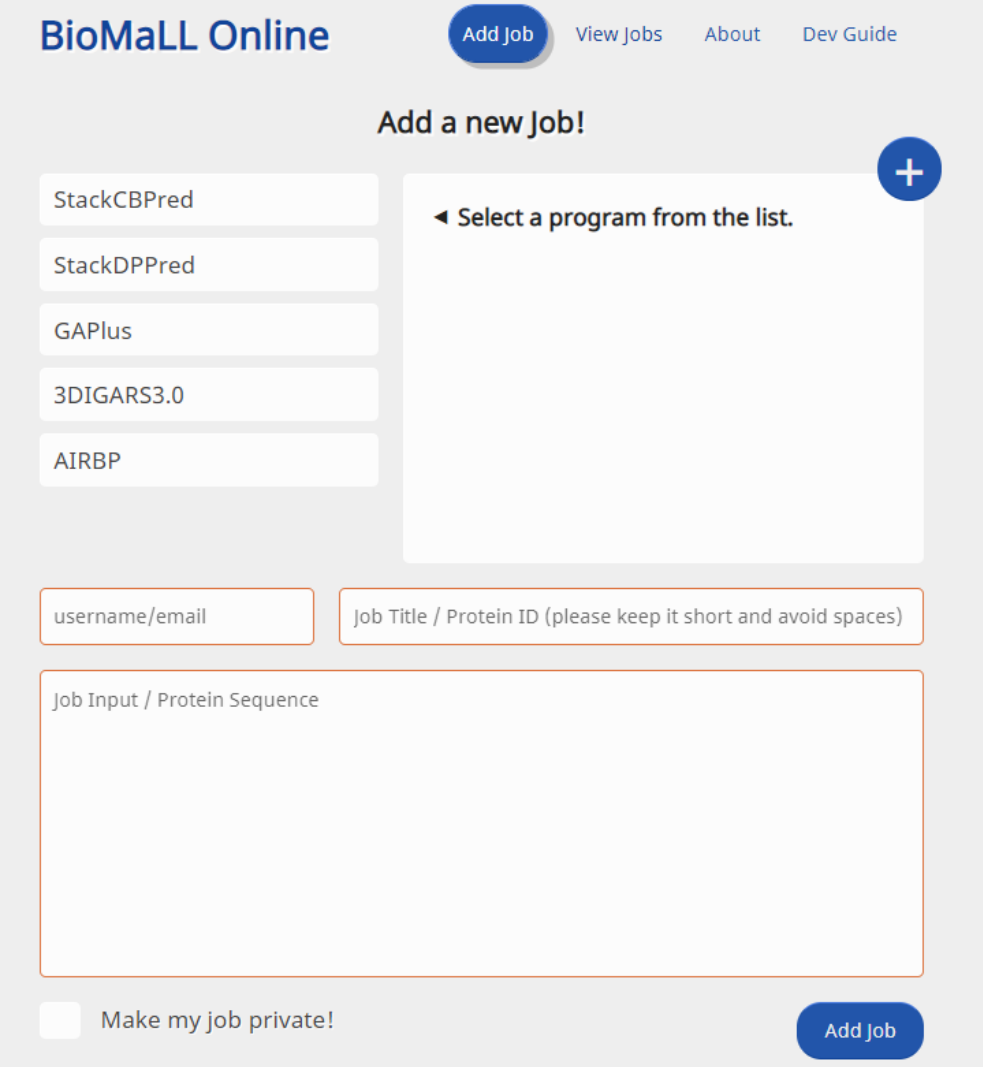

Figure 7. Developed Software in BMLL Lab

Conclusions

Gene protein regulatory network is important to understand the function of a cell. In addition, the network is important to understand the cellular level's dynamic behavior and understand the gene that causes disease. Therefore, an accurate prediction of the regulatory network by protein level interactions will help us understand the complex phenomenon. Further, learning about gene regulation from lab experiments is time-consuming and costly. Therefore, a fast and effective computational method is needed to infer the gene protein regulatory network. We proposed a new approach to discovering the gene regulatory network based on the protein level interactions with other genes in this work. Moreover, we offered a way to reduce the noise of the network by removing edges using different machine learning predictors validating protein level interactions. Though this manuscript would allow us to systematically apply the developed various research tools, in the future, we will merge all these in a pipeline to automate the prediction of the complex interaction between genes-protein in a much convenient form.

References

[1] MedlinePlus. (1 June 2021) How many chromosomes do people have? *U.S. National Library of Medicine*.

[2] MedlinePlus. (2020) Help ME Understand Genetic Cell and DNA. *U.S. National Library of Medicine*.

[3] Y. Sun and J. R. Dinneny, "Q&A: How do gene regulatory networks control environmental responses in plants?," *BMC Biology,* vol. 16, p. 38, 2018/04/11 2018.

[4] S. Franklin and T. M. Vondriska, "Genomes, proteomes, and the central dogma."

[5] C. Willyard, "New human gene tally reignites debate."

[6] F. CRICK, "Central Dogma of Molecular Biology," *Nature methods,* vol. 227, 1970.

[7] M. D. B. Jingyi Jessica Li, "Statistics requantitates the central dogma," *SCIENCE,* vol. VOL 347, 2015.

[8] D. M. C. Christopher A Jackson, Giuseppe-Antonio Saldi, Richard Bonneau , David Gresham "Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments," *Computational and Systems Biology,* Jan 27, 2020.

[9] L. D. S., "Transcription factors: an overview," *International journal of experimental pathology,*, vol. 74(5), pp. 417–422, 1993.

[10] C. B. G.-B. Sara Aibar, Thomas Moerman, Vân Anh Huynh-Thu,6 Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord,Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts, "SCENIC: Single-cell regulatory network inference and clustering," *Nature methods,* vol. 14, pp. 1083-1086, 2017

[11] C. F. Bram Van de Sande, Kristofer Davie, Maxime De Waegeneer, Gert Hulselmans, Sara Aibar, Ruth Seurinck, Wouter Saelens, Robrecht Cannoodt, Quentin Rouchon, Toni Verbeiren, Dries De Maeyer, Joke Reumers, Yvan Saeys & Stein Aerts, "A scalable SCENIC workflow for single-cell gene regulatory network analysis," *Nature Protocols,* vol. 15, pp. 2247–2276 2020.

[12] A. Mishra, P. Pokhrel, and M. T. Hoque, "StackDPPred: a stacking based prediction of DNA-binding protein from sequence," *Bioinformatics,* vol. 35, pp. 433-441, Feb 1 2019.

[13] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods," *PLOS ONE,* vol. 5, p. e12776, 2010.

[14] J. M. Sagendorf, H. M. Berman, and R. Rohs, "DNAproDB: an interactive tool for structural analysis of DNA–protein complexes," *Nucleic Acids Research,* vol. 45, pp. W89-W97, 2017.

[15] J. D. Blanco, L. Radusky, H. Climente-González, and L. Serrano, "FoldX accurate structural protein-DNA binding prediction using PADA1 (Protein Assisted DNA Assembly 1)," *Nucleic Acids Res,* vol. 46, pp. 3852-3863, May 4 2018.

[16] A. Mishra, R. Khanal, W. U. Kabir, and T. Hoque, "AIRBP: Accurate identification of RNA-binding proteins using machine learning techniques," *Artif Intell Med,* vol. 113, p. 102034, Mar 2021.

[17] B. Lang, A. Armaos, and G. G. Tartaglia, "RNAct: Protein-RNA interaction predictions for model organisms with supporting experimental data," *Nucleic Acids Res,* vol. 47, pp. D601-d606, Jan 8 2019.

[18] S. Iqbal and M. T. Hoque, "PBRpredict-Suite: a suite of models to predict peptide-recognition domain residues from protein sequence," *Bioinformatics,* vol. 34, pp. 3289-3299, 2018.

[19] T. Sun, B. Zhou, L. Lai, and J. Pei, "Sequence-based prediction of protein protein interaction using a deep-learning algorithm," *BMC Bioinformatics,* vol. 18, p. 277, 2017/05/25 2017.

[20] L. Zhang, G. Yu, D. Xia, and J. Wang, "Protein–protein interactions prediction based on ensemble deep neural networks," *Neurocomputing,* vol. 324, pp. 10-19, 2019/01/09/ 2019.

[21] L. Liu, X. Zhu, Y. Ma, H. Piao, Y. Yang, X. Hao, *et al.*, "Combining sequence and network information to enhance protein-protein interaction prediction," *BMC bioinformatics,* vol. 21, pp. 537-537, 2020.

[22] M. Panta, A. Mishra, M. T. Hoque, and J. Atallah, "ClassifyTE: A stacking based prediction of hierarchical classification of transposable elements," *Bioinformatics,* Mar 2 2021.