Mathematical Modelling of Biological Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Guided Research: BioKGC Application of lncRNA Regulation

Hui Cheng

Examiner:
Prof. Dr. Fabian J. Theis

Supervisor:
Prof. Dr. Annalisa Marsico, Dr. Emy Yue Hu

Submitted:
Munich, 2024/04/12

Contents

1 ABSTRACT

Long non-coding RNAs (lncRNAs) are transcripts longer than 200 nucleotides. Although lncRNAs do not directly encode proteins, an increasing body of evidence suggests their role as key regulators of gene expression. They play crucial roles in various biological functions and disease processes, including cancer. Investigating the regulatory mechanisms of lncRNAs can provide valuable insights into the molecular mechanisms underlying tumorigenesis and may aid in identifying potential diagnostic biomarkers and therapeutic targets.

One approach to identifying potential lncRNA regulatory instances is to examine gene regulatory networks using knowledge graphs, which can help identify reasonable missing links. This link prediction task falls under the broader umbrella of knowledge graph completion. While knowledge graph completion has found numerous applications in biological networks, such as protein-protein interaction networks, disease gene association networks, and drug association networks, its application in lncRNA regulation remains underexplored.

To address this gap, we construct a lncRNA-target gene regulation knowledge graph using LncTarD 2.0 [29] in this project. Subsequently, we employ the state-of-the-art path-based graph neural network BioKGC, designed explicitly for biological networks, to predict potential novel lncRNA regulations.

2 INTRODUCTION

2.1 LncRNA Regulation Mechanisms

The gene is widely regarded as the fundamental unit of inheritance. Most genes encode specific proteins or segments that carry out cellular functions. Gene expression involves transcribing DNA into messenger RNA (mRNA), subsequently translated into proteins. During transcription, RNA polymerase identifies specific sequences in the DNA known as promoters. It transcribes the gene sequence into a primary mRNA molecule that contains both coding regions called introns, interspersed in protein-coding regions known as exons. After modifications such as capping, splicing, and polyadenylation, the mature mRNA transcripts leave the nucleus and enter into the cytoplasm, where they bind to ribosomes and then are translated into specific amino acid sequences, ultimately generating functional proteins [2].

Long non-coding RNAs (lncRNAs) are a type of RNA molecule that, unlike mRNA, does not code for proteins but plays crucial regulatory roles at various levels of gene expression, including epigenetic, transcriptional, post-transcriptional, translational, and post-translational regulation [28].

lncRNAs primarily interact with mRNA, DNA, protein, and miRNA in the form of RNA. Long non-coding RNAs (lncRNAs) intricately regulate gene expression through multiple mechanisms in epigenetics. Firstly, lncRNAs modulate histone modifications, such as methylation or acetylation (Figure 2.2.A&B). Secondly, they also play a role in chromatin looping, either promoting or inhibiting interactions between distant genomic regions. Moreover, lncRNAs can directly bind to DNA or recruit DNA methyltransferases to regulate the transcription of target genes (Figure 2.2.C). In the regulation between lncRNAs and mRNAs, lncRNAs can influence mRNA stability by directly interacting with mRNAs (Figure 2.2.L) or by modulating RNA methylation, thereby affecting gene expression at the post-transcriptional level (Figure 2.2.N). LncRNAs also engage in diverse interactions with proteins: they interact with transcription factors to either activate or inhibit the transcription of downstream genes (Figure 2.2.E&F), with splicing factors to regulate mRNA alternative splicing at the post-transcriptional level (Figure 2.2.G), and with various proteins to regulate protein localization, phosphorylation, ubiquitination, and acetylation, thus influencing protein degradation or formation, and ultimately affecting protein expression at the post-translational level (Figure 2.2.R). Furthermore, lncRNAs function as competitive endogenous RNAs (ceRNAs), sequestering microRNAs and

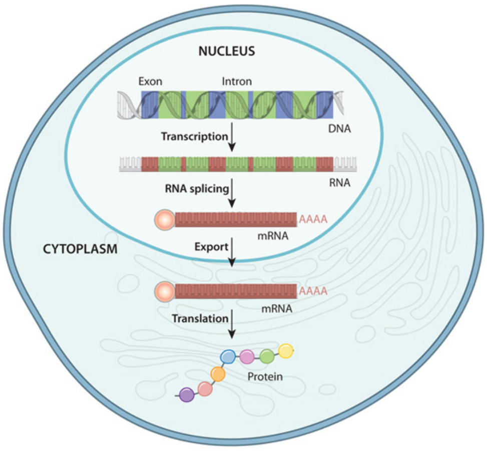


Figure 2.1 An overview of information flow from DNA to proteins. DNA is transcribed into mRNA in the cell nucleus, and the mRNA is processed and transported to the cytoplasm, where it is translated into proteins [2].

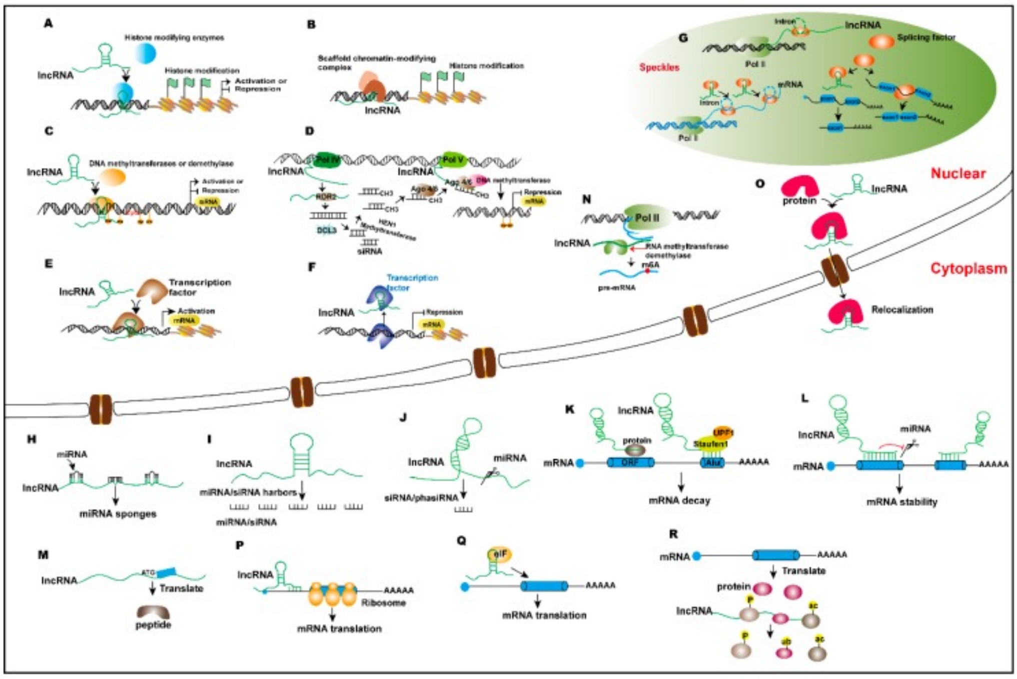


Figure 2.2 Regulatory mechanisms of lncRNAs at the genome level. (A) lncRNAs interact with histone-modifying enzymes to activate or repress gene transcription. (B) lncRNAs recruit histone-modified complexes or act as scaffolds for multiple histone modifiers to regulate histone modification of genes and thereby regulate gene transcription. (C) lncRNAs recruit DNA methyltransferases or demethylases to regulate the target gene transcription. (D) Pol IV/V transcribed lncRNAs are involved in RNA-dependent DNA methylation, thus activating or repressing gene transcription. (E,F) lncRNAs interact with transcription factors to activate or repress gene expression. (G) lncRNAs interact with splicing factors or proteins to regulate the mRNA alternative splicing; splicing factors directly regulate the lncRNA's alternative splicing in speckles. (H) lncRNAs act as miRNA sponges that regulate target gene expression. (I) lncRNAs act as miRNA or small interfering RNAs (siRNA) precursors. (J) miRNAs target lncRNAs to produce siRNA or phased small-interfering RNAs (phasiRNAs). (K) lncRNAs are involved in the Staufen1-mediated mRNA decay, and lncRNAs bind to proteins and mediate mRNA decay. (L) lncRNAs directly bind to mRNA and regulate mRNA stability or competitively bind to mRNA to improve mRNA stability. (M) lncRNAs can be translated to peptides. (N) lncRNAs interact with RNA methyltransferases or demethylases and thus regulate mRNA expression. (O) lncRNAs combine with proteins to regulate protein localization. (P) lncRNAs interact with mRNAs and affect mRNA translation. (Q) lncRNAs bind the translation initiation complex eIF (eukaryotic initiation factor) to regulate mRNA translation. (R) lncRNAs interact with proteins and control protein phosphorylation, acetylation, and ubiquitination at the post-translation level [28]. Among these, (D) is restricted to plants.

modulating their activity (Figure 2.2.I). Recent studies have also demonstrated the ability of lncRNAs to encode micro-peptides, further expanding their biological repertoire [28].

In recent studies, the regulation of lncRNAs has been shown to influence the development of various human diseases, including cancer. The well-known lncRNA MALAT1 is highly expressed in human tumors, promoting the proliferation and metastasis of various tumors and hematological malignancies. For instance, MALAT1 regulates EZH2 expression to facilitate colorectal cancer progression [23]. As the understanding of the precise roles of lncRNAs in disease pathobiology continues to expand, there is a growing exploration of therapeutic agents targeting oncogenic lncRNAs in tumor cells [9].

2.2 Knowledge Graph Completion

In computer science, a graph, denoted as $G = (V, E)$, where V is the set of vertices and E is the set of edges, is defined as a classical data structure used to represent relationships between entities. With the increasing integration of computer science in biology, graphs are also commonly employed to depict biological networks. There are many types of graphs. Considering edges, depending on whether the relations between vertices are symmetric or not, graphs can be classified into undirected and directed graphs. Edges can also be weighted or unweighted; in a weighted graph, the weights of the edges measure the strength or density of the relationships between the vertices, while in an unweighted graph, all edges have the same weight. Considering vertices, graphs can also be classified as homogeneous and heterogeneous based on the number of vertex and edge types; a homogeneous graph consists of a single vertex type, whereas a heterogeneous graph contains different vertex and edge types, allowing more diverse and nuanced information [17]. A finite sequence of consecutive distinct vertices and the edges between them form a path, the fundamental element in understanding the connectedness of graphs. Strongly connected components refer to subsets of vertices within a graph where each vertex is reachable from every other vertex within the subset. Conversely, weakly connected components represent subsets of vertices in which connectivity may not necessarily be direct but can be achieved through a sequence of edges regardless of direction [14].

A knowledge graph $KG = (\mathcal{E}, \mathcal{R})$ can be regarded as a heterogeneous directed graph, with the exception that vertices are derived as entities $\mathcal{E}$ and edges are derived as relationships $\mathcal{R}$ [8]. Knowledge graphs are usually represented as sets of triplets. A triplet (h, r, t) stands for head, relation and tail, where $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$. In this way, the problem of knowledge graph completion can be transformed into estimating the missing components of triplets, delineating three specific tasks based on the missing parts: 1) tail prediction $(h, r, ?)$ - given the head entities and relationships in a triplet, predict the corresponding tail entities; 2) head prediction $(?, r, t)$ - given the relationship and tail entities, predict the corresponding head entities; 3) relation prediction $(h, ?, t)$ - given the head and tail entities, and predict the relationship between them [12].

Knowledge graph completion methods can broadly be categorized into embedding-based and path-based approaches. Embedding-based approaches adopt encoding models, ranging from simple linear models to more complex neural networks, to learn feature representations of entities and relationships in a knowledge graph by minimizing the distance between the head and tail entity embeddings and the relationship embeddings, e.g., TransE [10], or maximizing the similarity between the embeddings of head entities, relations, and tail entities in the knowledge graph, e.g., DistMult [25]. Embedding-based approaches have demonstrated significant performance in several benchmark tests but are still constrained on one-hop relations. In many real-world scenarios, particularly in large and interconnected knowledge graphs, relationships between entities are intricate and may involve multi-hop paths. To address this challenge, there has been a growing interest in path-based approaches that aim to capture structural information of knowledge graphs and thereby improve model performance and interpretability. There are many path-based methods; in this project, we focus on NBFNet and BioKGC, which have improved from NBFNet. There are many path-based methods, but the focus of this project is limited to the Neural Bellman-Ford Networks (NBFNet) [31] and the BioKGC model.

2.2.1 NBFNet

NBFNet is a general graph neural network framework that learns a representation for each path from the query entity to potential tail entities based on the relations along the path, inspired by the Bellman-Ford algorithm, a classic method used to find the shortest paths from a single source vertex u to all other vertices v in a weighted graph. Expanding the Bellman-Ford algorithm's formula of Equation (2.1) by generalizing the addition operator $+$ to any summation operator $\oplus$, and the minimum operation $\min$ to any multiplication operator $\otimes$, while ensuring that $\oplus$ and $\otimes$ satisfy a semiring system, yields a generalized Bellman-Ford algorithm of Equation (2.2). The NBFNet parameterizes the generalized Bellman-Ford algorithm with 3 neural components: INDICATOR, MESSAGE, and AGGREGATE functions, as shown in Equation (2.3).

$$d[v] = \min(d[u] + \omega(u, v)) \tag{2.1}$$

$$\begin{aligned} &h_q^{(0)}(u, v) \leftarrow \mathbb{1}_q(u = v) \\ &h_q^{(t)}(w, v) \leftarrow (\oplus_{(x,r,v)\in\mathcal{E}(v)} h_q^{(t-1)}(u, x) \otimes w_q(x, r, v)) \oplus h_q^{(0)}(u, v) \end{aligned} \tag{2.2}$$

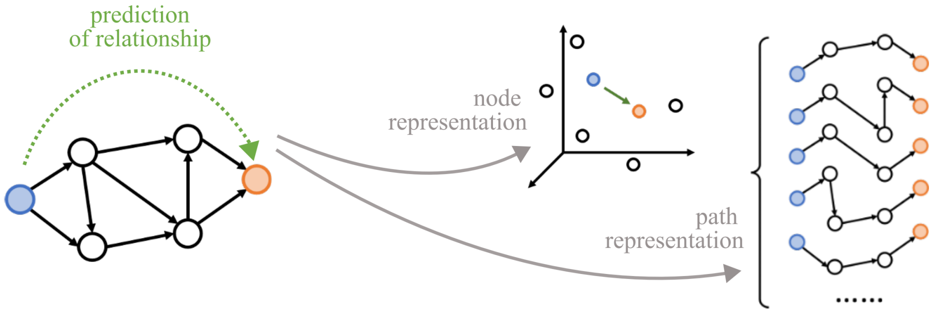


Figure 2.3 Two approaches to knowledge graph completion: one is embedding-based methods, which learn vector representations of nodes in a specific space, and the other is path-based methods, which learn path representations.

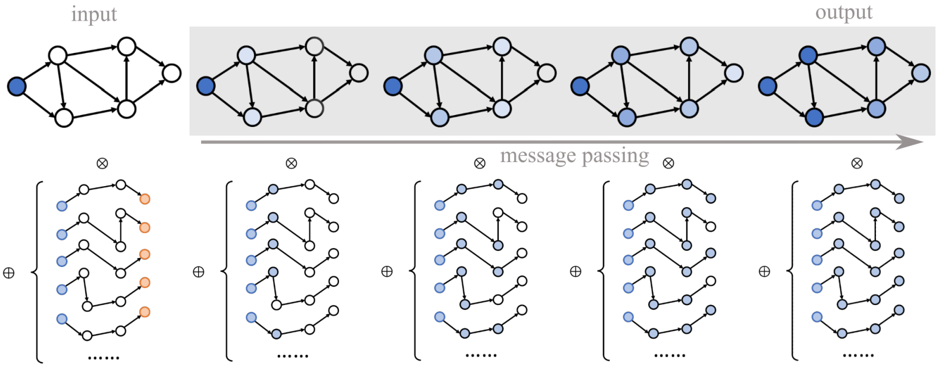


Figure 2.4 In the path-representation learning, information iteratively propagates to adjacent nodes from a single source vertex.

$$\begin{aligned} h_v^{(0)} &\leftarrow \text{INDICATOR}(u, v, q) \\ h_v^{(t)} &\leftarrow \text{AGGREGATE}(\{\text{MESSAGE}(h_x^{(t-1)}, w_q(x, r, v)) | (x, r, v) \in \mathcal{E}(v)\} \cup \{h_v^{(0)}\}) \end{aligned} \tag{2.3}$$

The INDICATOR function initializes a representation on each node, which is taken as the boundary condition of the generalized Bellman-Ford algorithm. The MESSAGE function learns the multiplication operator and can be defined according to the relational operators in knowledge graph embeddings. The AGGREGATE function learns the summation operator, which could be sum, mean, or max, followed by a linear transformation and a non-linear activation.

The learned pair representations, denoted as $h_q(u, v)$, facilitate the computation of conditional likelihoods for tail entity v as $p(v|u, q) = \sigma(f(h_q(u, v)))$, where $\sigma(\cdot)$ represents the sigmoid function and $f(\cdot)$ denotes a feed-forward neural network, typically a 2-layer multi-layer perception (MLP). Similarly, the conditional likelihood of the head entity u is predicted as $p(u|v, q^{-1}) = \sigma(f(h_{q^{-1}}(v, u)))$. NBFNet is trained to minimize the loss of Equiation (2.4), specifically the negative log-likelihood of positive and negative samples generated by cropping one of the entities in a positive triplet.

$$\mathcal{L}_{KG} = -\log p(u, q, v) - \sum_{i=1}^{n} \frac{1}{n} \log(1 - p(u_i', q, v_i')) \tag{2.4}$$

2.2.2 BioKGC

BioKGC is a model derived from NBFNet that is explicitly tailored to biomedical knowledge graphs. BioKGC adopts a more stringent negative sampling strategy, wherein negative samples are exclusively drawn from the same node type as the positive samples. This approach ensures that the negative samples are sufficiently difficult for the model to learn a good decision boundary. Furthermore, BioKGC leverages an

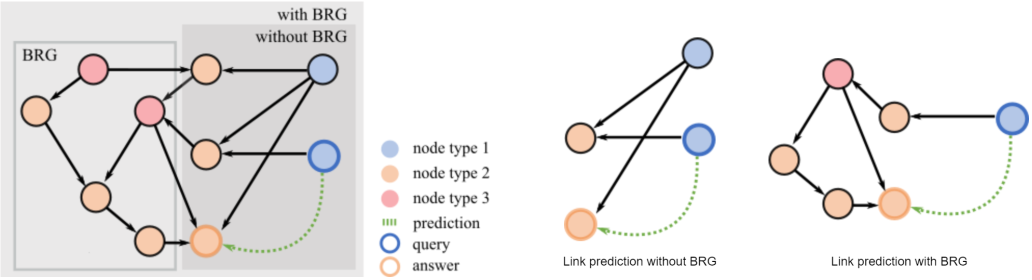


Figure 2.5 BRG can enhance predictions by leveraging additional knowledge bases to provide further insight. Without the BRG, link prediction relies on bipartite graphs where relationships between query and target nodes are inferred from the intersections of nodes similar to the target and query nodes. With the inclusion of the BRG, graph connectivity is boosted, and paths between query and target may exist, allowing direct message propagation.

external background regulatory graph (BRG), augmenting the connectivity of entities within the knowledge graph and thereby enhancing prediction accuracy by leveraging supplementary information from external knowledge bases. During negative sampling, entities from the BRG are not considered; thus, BRG only serves to message passing and does not influence loss.

3 METHODOLOGY

3.1 Data Preprocessing

LncTarD 2.0 is a manually curated database of 8360 experimentally supported functional lncRNA-target regulations in human diseases, categorized into ceRNA or sponge, chromatin looping, epigenetic regulation, expression association, interact with mRNA, interact with protein and transcriptional regulation [29]. Nevertheless, the database exhibits incomplete gene information, as Ensembl IDs are missing for certain genes. Moreover, there are inconsistencies regarding gene names and gene types. For example, Caspase3 and CASP3 - two aliases for the same gene encoding a cysteine-aspartate protease - can both be found in the target gene [1], while SIRT1 is classified as both a lncRNA and a protein-coding gene. Although aliases are quite common, and it has been experimentally demonstrated to be possible for a gene to encode both protein and lncRNA under certain circumstances [22], unifying gene names and types is essential for the construction of a knowledge graph.

The first step is identifying the genes in common human gene databases, e.g., Ensembl, NCBI gene, or HGNC. For genes with Ensembl IDs provided in LncTarD 2.0, the name and type of the gene can be directly mapped to the Ensembl human gene database, Homo_sapiens.GRCh38.110 [5], whereas genes without IDs need to be manually looked up name one by one on GenCards, a searchable human genetics database integrating data from multiple sources [3], to get its ID, most commonly used alias and type. For the latter category, four distinct scenarios arise. Firstly, the names of some targets contain more than one gene; for example, a target named "Snail1/2" refers to two genes, SNAI1 and SNAI2, so they need to be separated into two records. Secondly, some genes are not included in the Ensembl database, such as CCAT1, for which information from other databases, such as the NCBI gene, can be used instead. The third problem is commonly associated with miRNAs, many of which have different precursors, for example, miR-194-5p, which has two precursors named MIR194-1 and MIR194-2 in the Ensembl database. Lastly, some genes are only mentioned in some papers but cannot be found in Gencard, such as XLOC_006753 and ENST4, which can only be excluded.

The second step is to resolve the conflict of gene names and types between LncTarD 2.0 and the common human gene databases, with the main principle of preserving most of the names and types from the latter while keeping the type of transcription factor from the former. Transcription factors are proteins specifically involved in the process of transcribing DNA into RNA [4]. They are distinguished from protein-coding genes in LncTarD 2.0 but not in Ensembl, NCBI genes, or HGNC. Considering that more than three hundred genes are classified as transcription factors in LncTarD 2.0, it would be good to keep this type. In addition, some of the genotypes with a relatively small percentage of genes were reclassified. For example, the Ensembl human sapiens database contains several types of pseudogenes, including processed, unprocessed, transcribed processed, transcribed unitary, and transcribed unprocessed pseudogenes; these genes were grouped under the broad category of pseudogene. PRAL, PWAR5, and GPC3-AS1 were originally of the TEC type and are now classified as lncRNAs. TRGJP2, RN7SKP4, and NRON were reclassified as protein-coding genes, pseudogenes and lncRNAs.

Compared to the corrections applied to gene names and types, adjustments to the regulatory mechanisms of lncRNA-target interactions were comparatively minor. The LncTarD 2.0 database introduces seven mechanisms of lncRNA-target regulation: ceRNA or sponge, chromatin looping, epigenetic regulation, expression association, interaction with mRNA, interaction with protein, and transcriptional regulation. Of these, only 12 pairs of regulations belonged to chromatin looping regulation, which was then classified as transcriptional regulation after double-checking their corresponding experimental support, leaving the remaining regulation mechanisms unchanged.

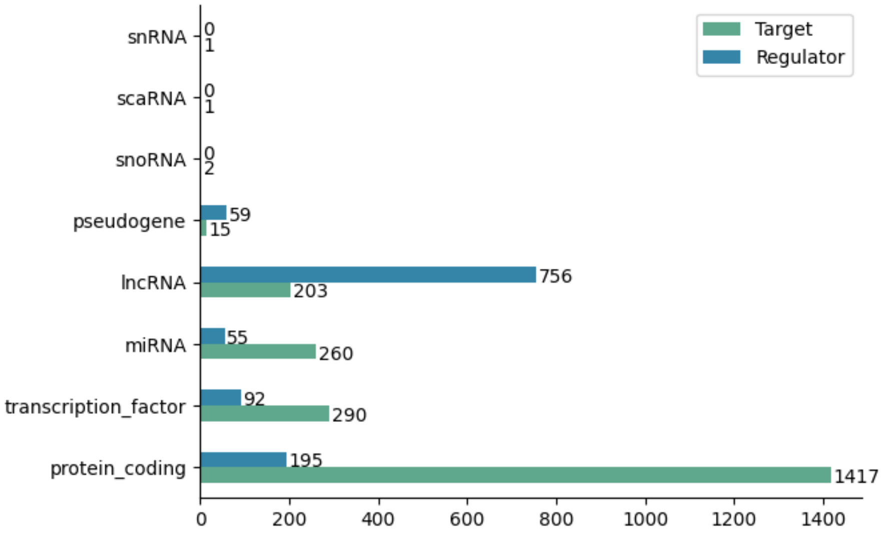


Figure 3.1 Gene type distribution for regulators and targets in knowledge graph constructed from LncTarD 2.0.

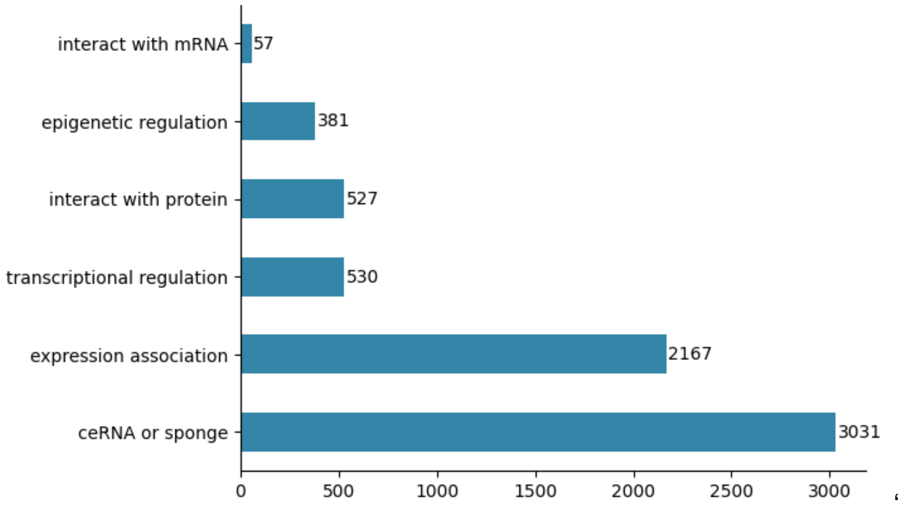


Figure 3.2 Regulatory mechanisms type distribution in the knowledge graph constructed from LncTarD 2.0.

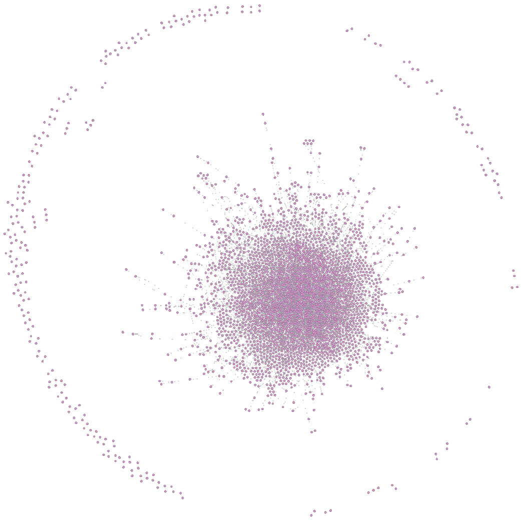


Figure 3.3 Visualization of LncTarD 2.0 with a limit of 3000 nodes and 5000 edges. In the center is a large weakly connected component, surrounded by many small independent connected components containing only two or three nodes.

3.2 Knowledge Graph Construction

Given LncTarD 2.0, a knowledge graph can be naturally constructed where entities are the involved genes and relations are the regulatory mechanisms. For a triplet (h, r, t) that constitutes the knowledge graph, the head h corresponds to the regulator, the relation r to the regulation mechanisms, and the tail t to the target gene. The knowledge graph constructed by LncTarD 2.0, as shown in Figure 3.3, consists mainly of a densely connected subgraph containing 2646 genes and 6084 involved regulations, surrounded by many isolated subgraphs containing only two or three genes. Since path-based methods learn a representation of the paths from the query entity to the potential trail entities, where connectivity is very important, therefore only the largest weakly connected component of the constructed knowledge graph was kept and divided into training, validation, and testing sets using the ‘split‘ function from the PyKeen Python package [7]. The default split method, known as coverage splitting, initially allocates triplets to the training set in a greedy manner. Subsequently, the remaining triplets are partitioned to ensure that each node is represented by at least one triplet in the training set. In this project, the triplets were partitioned at a ratio of $(0.8, 0.1, 0.1)$, and the random state was set to 1234. The specific numbers of nodes and edges for each subgraph are detailed in Table 3.1.

Table 3.1 Number of nodes and edges in the training, validation, testing sets, and BRG

	Train	Valid	Test	BRG
# nodes	2646	533	535	15806
# edges	4867	608	609	1035133

The above three datasets derived from LncTarD 2.0 are sufficient for NBFNet, whereas BioKGC requires additional data, the BRG, for message passing. A strongly connected component containing 15806 nodes and 1035133 edges, extracted from a protein-protein interaction (PPI) dataset, is the BRG in our utilization. On the one hand, the substantial overlap of protein-coding genes between the PPI dataset and LncTarD 2.0 allows for the seamless integration of these two knowledge graphs. On the other hand, in a strongly connected component, every node can be reached from any other node via directed paths; consequently, even if there were no direct regulation between two genes in the original knowledge graph if both of them are individually connected to the BRG, their connection can be then guaranteed. This approach ensures that the BRG effectively enriches the original knowledge graph with additional connectivity, thereby improving the overall performance of BioKGC. To avoid negative sampling of protein-coding genes presented only in the BRG, they were given a new type, "protein_coding_ppi", to distinguish them from protein-coding genes in LncTarD 2.0.

4 RESULT

4.1 Main Results

In this project, the models mentioned in Chapter 2, TransE, DistMult, NBFNet, and BioKGC, were employed for the task KGC on the knowledge graph constructed from LncTarD 2.0. For each positive sample, 32 negative samples were sampled. All training was optimized using an Adam with a learning rate of 0.005.

As embedding-based methods were not the focus of this project, only the most basic implementations from PyKeen were employed, which take pure triplets as input and do not consider the gene type information of the entities. Their training and evaluation utilized the PyKeen framework, fine-tuning embedding dimensions among 256, 512, 1024. The early stopping mechanism was incorporated to mitigate overfitting, whereby training was stopped if losses did not decrease significantly over ten consecutive epochs. To ensure reproducibility, the random seed was set to 1.

The path-based model implementations were built upon the official NBFNet codebase[1] and official BioKGC codebase[2], both of them comprising 6 layers with 32 hidden units per layer and a feed-forward network configured as a 2-layer MLP with 64 hidden units. Models obtained at each epoch during training were saved, and the final model was selected based on performance on the validation set. All computations were conducted on a server container equipped with 14 vCPUs Intel Xeon Gold 6330 at 2.00GHz, 46GB RAM, and one NVIDIA RTX 3090 GPU with 24GB VRAM, operating on Ubuntu 20.04 with Python 3.8.10, CUDA 11.8, and PyTorch 2.0.0. The summarized results are presented in Table 4.1.

Table 4.1 Performance comparison of embedding-based and path-based knowledge graph completion methods in terms of MR, MRR, and Hits@k for k = 1, 3, and 10.

	Model	MR	MRR	Hits@1	Hits@3	Hits@10
Embedding-based Methods	TransE	1329.90	0.0045	-	0.0049	0.0074
	DistMult	1364.95	0.0041	0.0016	0.0016	0.0033
Path-based Methods	NBFNet	168.30	0.1378	0.0599	0.1445	0.3062
	BioKGC	**86.84**	**0.1855**	**0.0870**	**0.2036**	**0.3966**

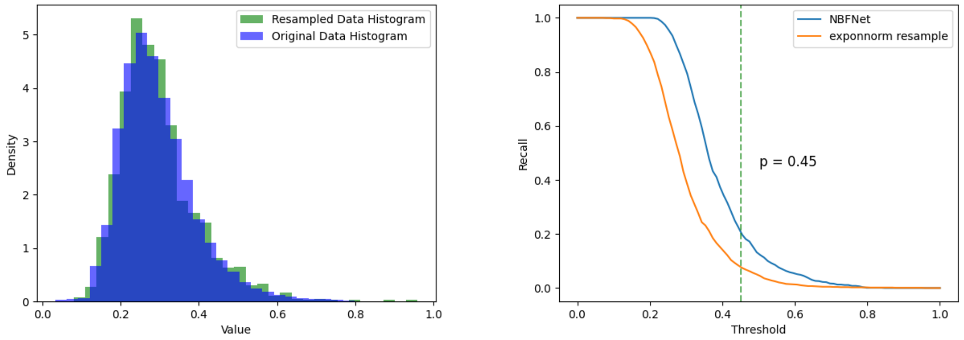
In comparing the two embedding-based methods, TransE and DistMult, their results exhibit marginal differences. Overall, TransE demonstrates slightly superior performance compared to DistMult. However, both methods fall significantly short compared to path-based methods, demonstrating a substantial performance gap. When comparing two path-based methods, BioKGC significantly outperforms NBFNet across all metrics, demonstrating the power of its adapted negative sampling strategy and the integration of the external BRG.

4.2 Recall Curve

To further assess BioKGC, 1248 additional trustworthy lncRNA-target pairs (h, r) sourced from [18] and [16] were introduced to compute recall. Assuming a knowledge graph $KG = (\mathcal{E}, \mathcal{R})$ where $h \in H \subset \mathcal{E}$ represents CRISPR and Enhancer lncRNAs, $t \in T \subset \mathcal{E}$ represents the target genes regulated by them, and there exists no $r \in \mathcal{R}$ so that $(h, r, t) \in KG$. Since the conditional probabilities, $p(e|h, r), h \in H, r \in$

[1] `https://github.com/DeepGraphLearning/NBFNet`

[2] `https://github.com/emyyue/NBFNet`



(a) Distribution of model-predicted conditional probabilities and resampled values.

(b) Recall curve of model predictions and random recall curve.

Figure 4.1 Data distribution and recall curves.

$\mathcal{R}, t \in \mathcal{E}$ output by the model cannot directly answer the binary classification question of whether there exists a relationship between lncRNA h and gene t, it is necessary to establish a threshold value p. When there exists $r \in \mathcal{R}$ such that $p(t|h,r) > p$, it can be inferred that there exists a relationship between h and t. Specifically, true positives (TP) are triplets for which $\max(p(t|h,r)) \leq p$, while false negatives (FN) are those for which $\max(p(h|t,r)) < p$. By varying the threshold value, a continuous recall curve can be obtained. A reference threshold value, around 0.45, is the average of the conditional probabilities of all triplets appearing in the training set predicted by the models, denoted as avg$(p(e|h,r))$, where $(h,r,e) \in KG_{\text{train}}$.

For comparative purposes, a random recall curve was also generated, wherein instead of the actual model-predicted conditional probabilities, 1248 values were randomly sampled from an approximate exponential normal distribution of all predicted conditional probabilities. When the sampled value is higher than p, it is considered a true positive (TP); otherwise, it is considered a false negative (FN).

$$Recall = \frac{TP}{TP + FN} \tag{4.1}$$

4.3 Interpretability

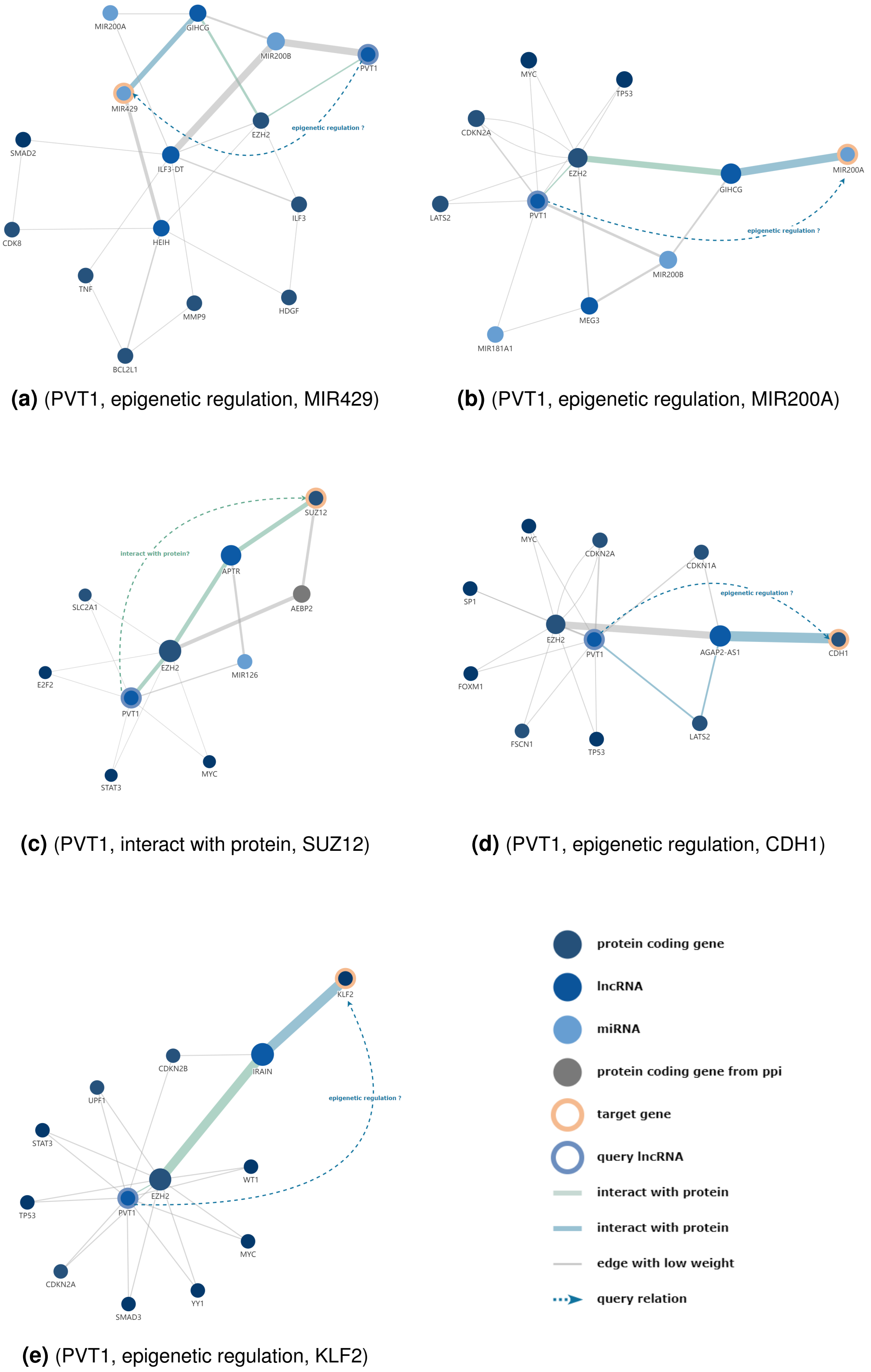
To further demonstrate the ability of BioKGC to uncover potential novel lncRNA regulations, we selected PVT1, "a rising star among oncogenic lncRNAs [13]," as the head entity and computed all conditional probabilities $p(t|\text{PVT1}, r)$ for all entities across all relationship types. The top 5 novel predictions with the highest probabilities are presented in Table 4.2, where "novel" refers to the absence of a direct connection between these two genes in the knowledge graph.

Table 4.2 Top 5 novel predicted regulations of PVT1 with highest conditioned probabilities.

Head h	Relation r	Tail t	Tail gene type	$p(t\|h,r)$
PVT1	epigenetic regulation	MIR429	miRNA	0.881
PVT1	epigenetic regulation	MIR200A	miRNA	0.871
PVT1	interact with protein	SUZ12	protein_coding	0.839
PVT1	epigenetic regulation	CDH1	protein_coding	0.836
PVT1	epigenetic regulation	KLF2	transcription_factor	0.816

BioKGC assigns each edge in the KG an importance score derived from the prediction gradient; the sum of the edge importance approximates the importance of the paths. Figure 4.2 illustrates the top 10 most crucial paths for the top 5 novel predictions, ranked by gradient. The edge width represents the frequency of appearance in paths, and the edge color indicates the importance. Specifically, the three most impotent edges are highlighted: green denotes the regulation mechanism "interact with protein," blue represents "epigenetic regulation," while the remaining edges are depicted in gray. Nodes are color-coded based on their node type: dark blue for protein-coding genes, navy blue for lncRNAs, light blue for miRNAs, and gray for protein-coding genes from BRG.

In all five novel predictions involving PVT1, despite the existence of multi-hop paths from PVT1 to the target gene in the knowledge graph, the most important edges identified by the model do not highlight any of them; instead, they constitute a small bipartite subgraph similar to the prediction made without BRG as shown in Figure 2.5 with one set of the bipartite graph comprises PVT1 and another lncRNA, while the other set consists of the target gene and another gene regulated by both lncRNAs. This observation aligns with the assumption in biological research that genes belonging to the same gene cluster tend to be regulated by the same lncRNA. For instance, as depicted in Figures 4.2a and 4.2b, EZH2 is regulated by both PVT1 and GIHCG via the "interact with protein" mechanism, while GIHCG simultaneously epigenetically regulates MIR200A, MIR200B, and MIR429. Based on this, the model infers that MIR200A and MIR429 also receive epigenetic regulation from PVT1. Biologically, MIR200A, MIR200B, MIR429, as well as MIR141 and MIR200C, belong to the miR-200 family, which plays crucial roles in cancer initiation and metastasis [21]. Evidence suggests that PVT1 promotes cervical cancer progression by epigenetically silencing miR-200b through interaction with EZH2, leading to increased histone H3K27 trimethylation on the miR-200b promoter and inhibition of miR-200b expression [27]. PVT1 may also contribute to the tumorigenesis and metastasis of melanoma by binding to EZH2 and regulating the expression of miR-200c [11]. Another intriguing observation is the presence of PVT1-EZH2 regulation in all novel predictions, with four of them highlighted in the plots, showing the importance of EZH2 in PVT1 regulation. This finding is consistent with experimental evidence demonstrating PVT1-EZH2 regulation in various cancers, including gastric cancer [6], thyroid cancer [30], glioma [24], and hepatocellular carcinoma [15].



(a) (PVT1, epigenetic regulation, MIR429)

(b) (PVT1, epigenetic regulation, MIR200A)

(c) (PVT1, interact with protein, SUZ12)

(d) (PVT1, epigenetic regulation, CDH1)

(e) (PVT1, epigenetic regulation, KLF2)

Figure 4.2 Visualizations illustrating the top 5 novel predicted regulatory interactions involving PVT1, as detailed in Table 4.2. Subfigures (a), (b), (d), and (e) depict the genes and regulatory relationships implicated in the prediction of epigenetic regulatory associations between PVT1 and MIR429, MIR200A, CDH1, and KLF2, respectively. Subfigure (c) presents the significant nodes and edges forecasting the regulatory interaction of PVT1 with the protein SUZ12. Each subgraph shows the top 10 paths for the top 5 new predictions, sorted by gradient. Edge width shows frequency, and color shows significance. Green edges indicate "interact with protein," blue is for "epigenetic regulation," and the rest are gray. Nodes are colored as follows: dark blue for protein-coding genes, navy blue for lncRNAs, light blue for miRNAs, and gray for BRG protein-coding genes.

5 DISCUSSION

Despite increasing evidence highlighting the significance of lncRNAs for humans, especially in diseases including cancers, the overall understanding of lncRNA regulatory mechanisms remains largely incomplete compared to other gene products, such as proteins. To deepen the understanding of lncRNA-mediated regulation, this project constructed a knowledge graph of lncRNA regulation using the LncTarD 2.0 dataset and employed the knowledge graph completion methods to identify and infer potential novel regulatory instances, which may help to further reveal new insights and potential treatments for diseases through.

The project compared four knowledge graph completion methods: TransE, DistMult, NBFNet, and the BioKGC model tailored for biological networks. Overall, the experimental results align with the expectation that BioKGC would excel in completing the lncRNA-target regulation knowledge graph and making reasonable predictions.

Compared to embedding-based methods, path-based methods consider the types of genes and capture intricate relational patterns and dependencies more effectively due to their ability to reason directly over paths in the knowledge graph.

Compared to NBFNet, BioKGC employs stricter negative sampling criteria, considering only genes of the same type during negative sampling, which helps enhance the model's performance and accuracy. Additionally, BioKGC introduces the BRG, designed to boost the connectivity between genes in the knowledge graph, thereby facilitating better information propagation between related genes. This path-based representation not only provides improved predictive capabilities but also offers stronger interpretability, allowing researchers to scrutinize the biological plausibility of the model. Researchers can better understand the functions and interactions between genes by analyzing paths in the knowledge graph, thereby advancing biomedical research.

Nevertheless, the model's performance remains significantly constrained by the limitations of the dataset. Firstly, the LncTarD 2.0 dataset is quite small, with the constructed knowledge graph containing only 2646 entities and 6084 triplets. This is not on the same order of magnitude when compared to commonly used benchmark datasets for knowledge graph completion, such as FB15k-237 [20], which contains 310,116 triples with 14,541 entities and 237 relation types, and WN18RR [10], which contains 93,003 triples with 40,943 entities and 11 relation types. The model may not have enough examples to learn from, which can lead to poor performance. In addition, many existing knowledge graph completion models, including BioKGC, are based on the assumption of a closed environment. That is, the completion of the graph can only be achieved by mining the potential relationships between existing entities, and it is impossible to predict weak connections and new entities. Knowledge graph completion in a closed-world environment heavily relies on well-structured and comprehensive knowledge graphs with broad entity coverage [12]. Currently, it is estimated that there are over 100,000 human lncRNAs, with approximately 16,000 recorded by Human GENCODE [19]; LncTarD 2.0 covers only a small fraction of lncRNAs. As a result, models trained on such data can only make limited predictions.

Secondly, BioKGC only supports a one-to-one mapping between genes and gene types. However, in reality, one gene can correspond to multiple gene types due to the diversity of gene expression. The same gene segment can produce different types of products under different genetic and environmental conditions. For example, the UV-induced ASCC3 short isoform functions as a long non-coding RNA [22]. Currently, when multiple gene types are associated with a gene in LncTarD 2.0, the types recognized by Ensembl are selected. This approach inevitably leads to incorrect representation of a small portion of regulations in the knowledge graph. Consequently, the model generates negative samples from genes of the wrong type, thereby affecting the model's overall performance.

Thirdly, the current datasets do not adequately demonstrate the transitivity of the relationship between lncRNAs and their target genes. Many regulatory mechanisms exhibit transitivity, such as "ceRNA or sponge" interactions. For example, lncRNA NEAT1 promotes the proliferation of ovarian cancer cells and angiogenesis of co-incubated human umbilical vein endothelial cells by regulating FGF9 through sponging MIR365 [26]. However, LncTarD 2.0 only includes the relation (NEAT1, ceRNA or sponge, FGF9), omitting the relation from NEAT1 to MIR365 and MIR365 to FGF9. Additionally, MIR365 is not even included in the entity set. Another example is transcriptional regulation, where the lncRNA MALAT1 directly promotes transcription of the LTBP3 gene by recruiting the SP1 transcription factor to the LTBP3 promoter [28]. In the dataset, only the relations (MALAT1, interact with protein, SP1) and (MALAT1, transcriptional regulation, LTBP3) are included, missing the link from SP1 to LTBP3. The lack of transitivity in the pathways within the knowledge graph might force the model to rely more on bipartite graph reasoning rather than path reasoning.

Lastly, calculating metrics other than recall in completing gene regulatory knowledge graphs is quite challenging. Most biological experiments can only prove that two genes are linked in a particular biological process. At the same time, it is hard to prove that two genes are completely unrelated under any circumstances. The absence of a link between two genes in the knowledge graph probably only means that their relationship has simply not been discovered yet, making it impossible to find proper true negative samples for evaluating the model's performance.

In summary, to further improve the performance of the BioKGC model, it is necessary to enhance the lncRNA regulation knowledge graph. LncTarD 2.0, constructed using purely manual methods, requires significant effort from domain experts, and human errors are inevitable. Here, we propose an initial concept for a sustainable semi-automated system for constructing and improving gene regulatory knowledge graphs, comprising the following six subsystems:

1. Knowledge Collection Subsystem: Use web crawler programs to automatically gather articles from PubMed containing keywords related to gene regulation.

2. Knowledge Extraction Subsystem: Utilize fine-tuned or prompt-based language models to extract and output triplets of involved gene regulation from the input articles.

3. Knowledge Enhancement Subsystem: Retrieve gene type information from databases such as Ensembl based on gene names.

4. Human Review Subsystem: Before merging new triplets into the existing knowledge graph, perform the manual review to ensure accuracy. The reviewed data can also serve as new training data for upgrading the LMs used for knowledge extraction.

5. Knowledge Graph Query Subsystem: Query and visualize existing relationships in the knowledge graph.

6. Knowledge Graph Completion Subsystem: Apply BioKGC to explore potential new regulatory relationships between genes.

Implementing such a semi-automated system can streamline the knowledge graph construction and refinement process while leveraging automated extraction techniques and human expertise to ensure accuracy and completeness. This approach could lead to a more comprehensive and reliable gene regulatory knowledge graph, thereby enhancing the performance of models like BioKGC in understanding and predicting gene regulatory networks.

Bibliography

[1] CASP3 caspase 3 [Homo sapiens (human)] - Gene - NCBI.

[2] Gene Expression | Learn Science at Scitable.

[3] GeneCards - Human Genes | Gene Database | Gene Search.

[4] General transcription factor / transcription factor | Learn Science at Scitable.

[5] Homo_sapiens - Ensembl genome browser 111.

[6] Long noncoding RNA PVT1 indicates a poor prognosis of gastric cancer and promotes cell proliferation through epigenetically regulating p15 and p16 - PubMed.

[7] Pykeen/pykeen.

[8] Stanford CS224W: Machine Learning with Graphs | 2021 | Lecture 10.2 - Knowledge Graph Completion.

[9] Nicola Amodio, Lavinia Raimondi, Giada Juli, Maria Angelica Stamato, Daniele Caracciolo, Pierosandro Tagliaferri, and Pierfrancesco Tassone. MALAT1: A druggable long non-coding RNA for targeted anti-cancer approaches. 11(1):63.

[10] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

[11] Lamei Chen, Dongmei Ma, Yuanyuan Li, Xiaoying Li, Lei Zhao, Jing Zhang, and Yali Song. Effect of long non-coding RNA PVT1 on cell proliferation and migration in melanoma. 41(3):1275–1282.

[12] Zhe Chen, Yuehan Wang, Bin Zhao, Jing Cheng, Xin Zhao, and Zongtao Duan. Knowledge Graph Completion: A Review. 8:192435–192456.

[13] Teresa Colombo, Lorenzo Farina, Giuseppe Macino, and Paola Paci. PVT1: A rising star among oncogenic long noncoding RNAs. 2015:304208.

[14] Darij Grinberg. An introduction to graph theory.

[15] Jianping Guo, Chong Hao, Congcong Wang, and Luo Li. Long noncoding RNA PVT1 modulates hepatocellular carcinoma cell proliferation and apoptosis by recruiting EZH2. 18:98.

[16] S. John Liu, Max A. Horlbeck, Seung Woo Cho, Harjus S. Birk, Martina Malatesta, Daniel He, Frank J. Attenello, Jacqueline E. Villalta, Min Y. Cho, Yuwen Chen, Mohammad A. Mandegar, Michael P. Olvera, Luke A. Gilbert, Bruce R. Conklin, Howard Y. Chang, Jonathan S. Weissman, and Daniel A. Lim. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. 355(6320):aah7111.

[17] Ahmad F. Al Musawi, Satyaki Roy, and Preetam Ghosh. A Review of Link Prediction Applications in Network Biology.

[18] Evgenia Ntini, Stefan Budach, Ulf A. Vang Ørom, and Annalisa Marsico. Genome-wide measurement of RNA dissociation from chromatin classifies transcripts by their dynamics and reveals rapid dissociation of enhancer lncRNAs. 14(10):906–922.e6.

[19] Luisa Statello, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte. Gene regulation by long non-coding RNAs and its biological functions. 22(2):96–118.

[20] Kristina Toutanova and Danqi Chen. Observed Versus Latent Features for Knowledge Base and Text Inference.

[21] Katrina L. Watson, Robert A. Jones, Anthony Bruce, and Roger A. Moorehead. The miR-200b/200a/429 cluster prevents metastasis and induces dormancy in a murine claudin-low mammary tumor cell line. 369(1):17–26.

[22] Laura Williamson, Marco Saponaro, Stefan Boeing, Philip East, Richard Mitter, Theodoros Kantidakis, Gavin P. Kelly, Anna Lobley, Jane Walker, Bradley Spencer-Dene, Michael Howell, Aengus Stewart, and Jesper Q. Svejstrup. UV Irradiation Induces a Non-coding RNA that Functionally Opposes the Protein Encoded by the Same Gene. 168(5):843–855.e13.

[23] J. J. Xie, W. H. Li, X. Li, W. Ye, and C. F. Shao. LncRNA MALAT1 promotes colorectal cancer development by sponging miR-363-3p to regulate EZH2 expression. 33(2):331–343.

[24] Anqiang Yang, Handong Wang, and Xiaobing Yang. Long non-coding RNA PVT1 indicates a poor prognosis of glioma and promotes cell proliferation and invasion via target EZH2. 37(6):BSR20170871.

[25] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding Entities and Relations for Learning and Inference in Knowledge Bases.

[26] Jialing Yuan, Ke Yi, and Lingyun Yang. LncRNA NEAT1 promotes proliferation of ovarian cancer cells and angiogenesis of co-incubated human umbilical vein endothelial cells by regulating FGF9 through sponging miR-365: An experimental study. 100(3):e23423.

[27] Shaorong Zhang, Guanli Zhang, and Jingying Liu. Long noncoding RNA PVT1 promotes cervical cancer progression through epigenetically silencing miR-200b. 124(8):649–658.

[28] Xiaopei Zhang, Wei Wang, Weidong Zhu, Jie Dong, Yingying Cheng, Zujun Yin, and Fafu Shen. Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. 20(22):5573.

[29] Hongying Zhao, Xiangzhe Yin, Haotian Xu, Kailai Liu, Wangyang Liu, Lixia Wang, Caiyu Zhang, Lin Bo, Xicheng Lan, Shihua Lin, Ke Feng, Shangwei Ning, Yunpeng Zhang, and Li Wang. LncTarD 2.0: An updated comprehensive database for experimentally-supported functional lncRNA-target regulations in human diseases. 51(D1):D199–D207.

[30] Qinyi Zhou, Jun Chen, Jialin Feng, and Jiadong Wang. Long noncoding RNA PVT1 modulates thyroid cancer cell proliferation by recruiting EZH2 and regulating thyroid-stimulating hormone receptor (TSHR). 37(3):3105–3113.

[31] Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. Neural Bellman-Ford Networks: A General Graph Neural Network Framework for Link Prediction.