Relation Prediction of Co-morbid Diseases Using Knowledge Graph Completion

Saikat Biswas, Pabitra Mitra, and Krothapalli Sreenivasa Rao

Abstract—Co-morbid disease condition refers to the simultaneous presence of one or more diseases along with the primary disease. A patient suffering from co-morbid diseases possess more mortality risk than with a disease alone. So, it is necessary to predict co-morbid disease pairs. In past years, though several methods have been proposed by researchers for predicting the co-morbid diseases, not much work is done in prediction using knowledge graph embedding using tensor factorization. Moreover, the complex-valued vector-based tensor factorization is not being used in any knowledge graph with biological and biomedical entities. We propose a tensor factorization based approach on biological knowledge graphs. Our method introduces the concept of complex-valued embedding in knowledge graphs with biological entities. Here, we build a knowledge graph with disease-gene associations and their corresponding background information. To predict the association between prevalent diseases, we use ComplEx embedding based tensor decomposition method. Besides, we obtain new prevalent disease pairs using the MCL algorithm in a disease-gene-gene network and check their corresponding inter-relations using edge prediction task.

Index Terms—Disease Co-morbidity, knowledge graph, Markov clustering, embedding, protein-protein interaction.

1 Introduction

Recent years have witnessed a rapid advancement in the field of knowledge graph (KG) construction and completion. A KG is represented by a multi-relational graph composed of entities (nodes) and relationships (different types of edges) among them. Every edge in the KG is represented as a triple, composed of *head entity*, *relation* and *tail entity*. Web-based knowledge graphs(KG) provide a structured representation regarding world knowledge, such as Google Knowledge Vault [1]. Many real-world applications namely semantic parsing, named entity disambiguation and information extraction tasks have been performed from a large number of popular KGs such as Freebase [2], Wordnet [3], DBpedia [4], YAGO [5] and NELL [6]. As these KGs are very logical and robust structures to make data machine understandable and processable, they seemed to be quite successful in biology and biomedicine [7].

A Knowledge graph is expressed as a directed graph with relations between the entities (nodes). Through the task of link prediction, we can predict the missing link between the nodes/entities in a knowledge graph. For an example, let's say we consider two relations as, AgeOfPerson and DOBOfPerson in a knowledge graph. Here, if AgeOfPerson is not recorded for all entries in the knowledge graph, it can be inferred from DOBOfPerson relation. The aim of link prediction is inferring such relations with automatic knowledge discovery. There are three different approaches namely, weight rule learning, graph random walk and tensor factorization used for link prediction task. In weight rule learning model [8], [9] a weight is set as a confidence for the set of first-order-logic rules, that represent the structure of the data. Lao et al. [10] proposed another method namely, graph random walk, which elaborates a strategy as walking through a path (probabilistically) while starting from a node and reach to the desired node. The final is the tensor factorization approach [11], [12], [13], which predicts the relation between head and tail entity by learning the embeddings for each entity and their corresponding relationships. Each of these methods has its own merits and demerits. Lin et al. [14] have shown the combination of all these categories sometimes results better in prediction in comparison to an only single category. In our paper, we use the tensor factorization approach. This tensor factorization model can be grouped into three main approaches: translation based, deep learning based and multiplicative based. The translational approach is proposed by Bordes et al. [12], where an additive function is applied over the generated embedding. A typical deep neural net model [15] is used in deep learning based link prediction method. Multiplicative approach is basically a product based technique over the embeddings [11], [13], [16], [17]. The canonical polyadic (CP) decomposition is one of the first multiplicative approaches [16], which produces two different entity embeddings. In CP decomposition, learning is done independently, and hence it gives a poor performance for KG completion task. As described in Trouillon et al. [18], instead of using real numbers as embeddings, the standard product of complex embeddings gives better result in link prediction task in knowledge graph completion. With this approach, the relations namely, reflexive, irreflexive, symmetric and transitive all can be expressed properly. Besides that, there is another complex vector embedding, referred to as conjugate of complex vector and it is known that the dot product of a complex vector with its conjugate transpose is defined as *Hermitian* (or sesquilinear) dot product. This Hermitian product of vector embeddings represents also the anti-symmetric/asymmetric

- *Saikat Biswas is with Advanced Technology Development Centre, Indian Institute of Technology, Kharagpur, 721302, India.*
E-mail: saikatbiswas17@iitkgp.ac.in, pabitra@cse.iitkgp.ac.in
- *Pabitra Mitra and Krothapalli Sreenivasa Rao are with Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, 721302, India.*

Manuscript received xxxxx xx, xxxx; revised xxxxx xx, xxxx.

relations in knowledge graph, which is the most effective advancement in ComplEx embedding technique in comparison to DistMult embedding [17] technique. So, a robust learning and efficient representation approach of ComplEx embedding makes it an effective method for KG completion task.

The knowledge graph data representation in connected information systems is defined by the standard world wide web consortium (W3C) recommended structure namely, Resource Description Framework (RDF). All relational databases can certainly be translated into their RDF structure. Recently, RDF based linked data format is used by many major bioinformatics data hubs. In that linked data structure the biological entities are connected through a unique identifier (IRI/ an international resource identifier) and inter-related connections between them are represented by standardized relations [19], [20]. Some of the popular biological data is available as an initiative of Uniprot RDF [21] and Bio2RDF project [22]. Among these two projects, Uniprot primarily focuses on storing data within a single database, whereas, Bio2RDF aims to integrate the different databases and produced in a single linked database structure. Additionally, major available linked databases are provided by European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI) [23], [24]. Moreover, a major development in the field of knowledge-based biological ontologies conceptualized the structure of domain [25], and they explicitly define some of the classes of entities within a specific domain and their inter-related connections [26]. These ontologies are represented by theoretic semantics with formal languages [27], [28] and that helps to infer logic with automated reasoning. But, the huge volume of ontologies restricts to execute the task of reasoning due to the complexity of the languages.

2 Related Works

We review a few studies, that showed the analysis of co-morbid diseases by protein-protein interactions and their relations with phenotypically similar diseases. Recent studies on co-morbid disease prediction are either based on a single biological entity or a few well known primary diseases. Paik et al. [49] proposed that some co-morbid disease conditions share common protein-protein interaction networks. In another study, Park et al. [50] showed that protein localization can be identified as a principal feature for co-morbid diseases with genetic disorders. Moreover, few researchers proposed that highly correlated diseases often share similar evolutionary constraints [51]. A major difference between these approaches and the proposed method is that, instead of using any single biological entity we apply a novel tensor factorization approach by integrating all disease-gene associated relations namely, pathways, genetic functions, disease phenotypes, and drug indications for predicting coexisting disease pairs.

A computational approach for predicting co-morbid disease pairs was proposed by Moni et al. [48]. Their method computationally measures the coexistence of different acute and chronic diseases using OMIM, KEGG and DO databases. He et al. [47] proposed a method to predict prevalent disease pairs by integrating multi-scale data. Although, their study is quite limited to very well studied primary diseases.

In recent times, from raw data, the features are learned using several machine learning methods [29], which became another popular task. In these methods, recently learning the vector embeddings of knowledge graphs became a popular task and by integrating these learned embeddings the relations between the entities can be predicted. Thus the knowledge and data integration are dominantly used for automatic reasoning to infer new relations/logic between the inter-related biological entities.

Maddouri [30] proposed a method for predicting disease co-morbidities by building an RDF based knowledge graph. Whereas in this method, the author used automated KG completion with ELK reasoner and deep neural nets for final learning of the vector embeddings (obtained using word2vec model). On the contrary, our presented study introduces a ComplEx embedding based computational approach to predict prevalent disease pairs. In this method, we build our own knowledge graph and instead of using any other automated reasoning tool and deep neural based learning, we use ComplEx embedding [13], [18] method for knowledge graph completion and edge prediction/link prediction between entities. Our result reflects a robust prediction of co-morbid diseases and their corresponding inter-relations with other biological and biomedical entities.

3 Materials and Methods

3.1 Data description

A knowledge graph (KG) is a graph, composed of multiple relations, where each node represents the entity and the corresponding relation between these entities is represented by an edge in the graph.

In this work, we develop a knowledge graph using ontologies namely, Gene Ontology (GO) [31], Human Phenotype Ontology (HPO) [32] and Disease Ontology (DO) [33,]. Genes and their corresponding functions are obtained from the GO database. Similarly, Genes and their corresponding phenotypes are obtained from HPO database. The other different biological databases used for the knowledge graph construction are defined below:

- Human GO annotations are obtained from SwissProt [21] and phenotype annotations are taken from HPO database [32]. we included 267510 GO annotations and 146838 phenotype annotations.
- Human protein-protein interactions (PPIs) are obtained from STRING database [34]. We took the interactions with the only confidence score of more than 700. We get a total of 171932 interactions (PPIs).
- Drug side-effects and indications are obtained from SIDER [35]. We get a total of 47904 drugs to human phenotype side effects and total of 6731 drugs to disease indications.
- Diseases and their corresponding phenotypes are obtained from HPO database [32].
- Disease databases OMIM [36] (Online Mendelian Inheritance in Man,December,2017 version) and GAD [37] (Genetic Association Database,2013 version) are downloaded for getting the whole disease-gene associations.

- Disease to pathway associations is downloaded from Comparative Toxicogenomics Database [38].
- Gene Entrez identifiers and their corresponding pathway ids are downloaded from Reactome database [39], [40].

First, we map all the proteins in the protein-protein interaction (PPIs) to their corresponding Entrez gene identifiers and by that we represent both genes and proteins. We use four different databases, namely OMIM, GAD, HPO and DO to obtain the total disease-gene associations. OMIM database is largely used as a repository of disease-gene associations, which is based on genetic information primarily with Mendelian disorders. Similarly, the genetic associations based on polymorphism data is collected from the GAD database. The DO database is a collective knowledge base of various developmental human diseases and also provides huge cross-platform mappings with MeSH (Medical Subject Headings), ICD (International Classification of Diseases) and OMIM identifiers. As diseases and genes are annotated with different identifiers in different databases, the inconsistency regarding disease identifiers is the normal issue for varied databases. So, to get rid of this issue all the disease data is integrated on the basis of disease name normalization. The whole data integration is done through OMIM identifier. We map each of the databases with their OMIM identifier (MIM identifier). For example, the HPO database provides the mapping between HPO_ ID and OMIM_ ID, GAD database provides the information of mapping between GAD_ ID and OMIM_ ID. Finally, DO database similarly gives mapping information between DO_ ID and OMIM_ ID. Then the ICD-9-CM codes and their respective OMIM_ IDs are collected from two different sources. One source is the manual curation by Goh et al. [41] and the second one is the OMIM_ ID to ICD-9-CM through DO_ ID. Finally, we get the ICD-9-CM to Entrez gene identifier associations with the aforementioned four databases. Then we perform the parsing of DO_ ID to UMLS_ ID mappings and through UMLS_ ID we get another ICD-9-CM code to Entrez ID associations. To get all the disease to gene associations we integrate these two obtained results of ICD-9-CM to Entrez Identifiers, without any duplications.

To get the indications of drug to diseases, we use the SIDER [35] database. From SIDER we get the mapping information between Drug_ ID to UMLS_ ID and from DO database we get the information mapping between ICD-9-CM codes and UMLS_ ID. From this information, through UMLS_ ID we finally get the drug indications to respective ICD-9-CM codes.

The mappings from OMIM_ IDs to HPO_ IDs are obtained from the HPO database. The diseases and their corresponding human phenotypes are mapped in two ways, namely via OMIM_ IDs and DO_ IDs.

Finally, we get a total of 72,383 HPO_ ID and disease (ICD-9-CM) associations.

We get the ICD-9-CM to MeSH_ ID via DO_ ID. From Comparative Toxicogenomics Database, we get the mapping information between MeSH_ ID and Pathway_ ID. Similarly, we also get the mapping information of OMIM_ ID and Pathway_ ID. Finally, through MeSH_ IDs we map the disease (ICD-9-CM) to Pathway_ ID and from previously obtained OMIM_ ID to ICD-9-CM codes, we get ICD-9-CM to Pathway_ ID through OMIM_ ID. Finally, we integrate both of these disease to pathway identifiers and collectively obtain ICD-9-CM to Pathway_ ID mappings.

We collected the Entrez id to pathway identifiers from Reactome Database.

The drug side-effects to human phenotypes are extracted from SIDER database. We get total of 47,904 drug side-effects to human phenotypes.

In our knowledge base, we use the dataset used by Park and his group, which contains the information about how often two diseases appear in health insurance claim [42]. This dataset was basically generated from Medicare claims for hospitalizations for 1990-1993. The raw data was composed of total 32,341,347 claims of approximately 13,039,018 individuals of over the age of 65 years.

3.2 Knowledge Graph building

Knowledge Graph (KG) is a graph based structure composed of entities and the corresponding relations between entities. Knowledge graph is very helpful for search engines and several recommendation systems. Similarly, KG can also be beneficial for biological and biomedical information extraction [43]. Our proposed knowledge graph is based on different biological entities and their ontologies, which define elaborately the background knowledge of those biological entities. The real purpose of our biological knowledge graph is to create an inter-relational structure among biological entities by integrating the existing information. Figure 1 shows our proposed knowledge graph.

Certainly, we use the protein-protein interaction network with gene level information and metabolic pathways of the respective genes. In our KG we mapped the proteins to their Entrez ids to get the genetic level information from the PPI itself. Then we integrate the PPI, with diseases and corresponding drugs to treat those diseases. After mapping each protein of PPI, to their gene ids, we get an entity as a gene. By extracting the corresponding genetic functions of each gene, we create a relation *'has_ function'* from gene entity to its genetic function entity. Similarly, for getting the associations between gene entity and its genetic phenotypes, we create another relation *'has_ gene_ phenotype'* from genes to its phenotypes. The graph is integrated with other entities as disease (ICD-9-CM codes), drug and pathway. We connected the gene to disease entity by the relation *'causes'*. The drugs and diseases are similarly get connected by *'treats'* relation pointing from drugs towards diseases. Some of the drugs also create different phenotypic side-effects. So drugs and their phenotypes are connected by relation *'has_ side_ effect'* from drugs to phenotypes. Similarly, diseases and phenotypes are related through *'has_ disease_ phenotype'* relation, as some of the diseases are genetically triggered and create a certain phenotypic condition in the human body. In the whole graph, a pathway is the common sharing entity between genes and diseases. A disease is related to a pathway through *'disrupts'* relation and pathway relate to gene through *'involves'* relation. In the directed graph some relations are uni-directional, some are bi-directional, and some are with self-loops. Bi-directions are made to avail all the entities in the graph to form a cycle to itself.

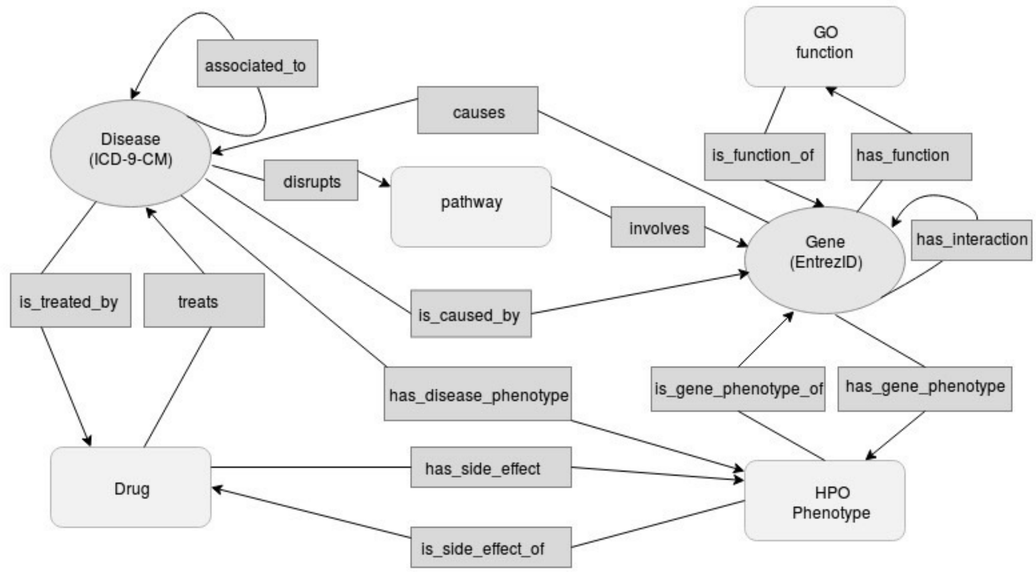

Fig. 1: Proposed Knowledge Graph Schema

For example, if we start traversing from entity *Gene*, we will reach the same node after covering the desired path. The whole KG is based on the disease-gene associations and their corresponding background knowledge, so these two entities are of special types and are connected with the bi-directional relation as well. On the other hand, the relations between disease to a pathway, a pathway to gene and disease to phenotype are uni-directional. There are two entities namely disease and gene, which are also related to itself only by self-loops with the relations *'has_ interaction'* and *'associated_ to'*, respectively. The supplementary Figure S2 shows an example of our proposed knowledge graph with real data.

3.3 Knowledge Graph Completion

3.3.1 Basic Background

Let, for a given knowledge graph, $\mathcal{E}$ represents a set of entities, where $|\mathcal{E}|$ = n and $\mathcal{R}$ represents a set of relations, where $|\mathcal{R}|$ = m. A knowledge graph is composed of a set of triples as, $\mathcal{D}^+ = \{(h, r, t)\}$, where $h \in \mathcal{E}$ is the head, $r \in \mathcal{R}$ is the relation and $t \in \mathcal{E}$ is the tail of the triple instance.

A relation r is reflexive on set $\mathcal{D}^+$, if (e, r, e) $\in \mathcal{D}^+$ for any entity $e \in \mathcal{E}$. A relation r is called symmetric if (e_1, r, e_2) $\in \mathcal{D}^+ \Leftrightarrow$ (e_2, r, e_1) $\in \mathcal{D}^+$ for any two entities e_1, $e_2 \in \mathcal{E}$ and anti-symmetric if (e_1, r, e_2) $\in \mathcal{D}^+$ then (e_2, r, e_1) $\notin \mathcal{D}^+$. A relation is transitive if (e_1, r, e_2) $\in \mathcal{D}^+ \wedge$ (e_2, r, e_3) $\in \mathcal{D}^+$ $\Rightarrow$ (e_1, r, e_3) $\in \mathcal{D}^+$, for any entities e_1, e_2, $e_3 \in \mathcal{E}$.

An *embedding* defines a function, that maps an entity or a relation to one or more vectors or a matrices of numbers. We represent the embedding corresponding to an entity e by $\mathbf{E}(e)$ and embedding corresponding to a relation r by $\mathbf{E}(r)$.

A *tensor factorization* model primarily defines the followings:

1) The embedding functions for entities and relations.
2) The embeddings $\mathbf{E}(h)$, $\mathbf{E}(r)$ and $\mathbf{E}(t)$ are taken as inputs in a function f and generates a prediction, whether (h, r, t) is in $\mathcal{D}^+$ or not. The final values of these embeddings are learned using the triples in the knowledge graph.

Let, x, y and z be the vectors of length k. We define $\langle x, y, z\rangle$ as the sum of element wise product of these three vectors. That is expressed as,

$\langle x, y, z\rangle = \sum_{j=1}^{k} x_j * y_j * z_j$, where x_j, y_j and z_j represent the j-th element of x, y and z, respectively.

3.3.2 Link Prediction in Knowledge Graph Using ComplEx Embedding

Predicting the incompleteness of knowledge graph is often addressed as link prediction approach in knowledge graph. There are various methods used for this task, but among them tensor factorization methods have proven to be promising enough. We use the tensor factorization method namely, ComplEx embedding for our biological knowledge graph. This ComplEx embedding method is proposed by Troullin and his group in their papers, for link prediction [13] and for knowledge graph completion [18]. Basically, ComplEx embedding is an extended and modified method of DistMult [17] method. The main difference between DistMult and ComplEx embedding is that, DistMult shows good results for symmetric relations but ComplEx embedding shows good results for symmetric and asymmetric relations as well. So, in our knowledge graph we use this ComplEx embedding method. In ComplEx embedding [13] method, embedding of each entity and every relation is considered as a vector of size k of complex numbers (instead of real numbers) . For each entity e, $\mathbf{E}(e)$ is represented as *re(e)* + *im(e)i*, where *re* and *im* stand for *'real'* and *'imaginary'*, respectively and both are of size k and i is the square root of -1. Similarly, for each relation r, $\mathbf{E}(r)$ also can be represented as *re(r)* + *im(r)i*. So, according to ComplEx the scoring function is, $f(\mathbf{E}(h), \mathbf{E}(r), \mathbf{E}(t)) = Real(\sum_{j=1}^{k}(re_j(h) + im_j(h)i) * (re_j(r)+im_j(r)i)*(re_j(t)-im_j(t)i))$, where $Real(a + ib) = a$ (here, h, r, t stand for head, relation and tail, respectively). If we change the sign of *im* vector for tail entity (corresponding to the conjugate of a complex number), ComplEx embedding differentiates between head and tail entities and allows for modelling asymmetric relations. The scoring function of ComplEx embedding can be easily verified and expanded as, $f(\mathbf{E}(h), \mathbf{E}(r), \mathbf{E}(t)) = \langle re(h), re(r), re(t)\rangle + \langle re(h), im(h), im(t)\rangle + \langle im(h), re(r), im(t)\rangle - \langle im(h), im(r), re(t)\rangle$.

3.4 Markov Clustering (MCL) in Disease-Protein-Protein Network

It is shown by Ko et al. [44] that in the context of disease co-morbidity there are a few chances of the prevalence of any disease pairs, which are sharing common genes. So, they showed that a random walk could be considered as another parameter for measuring the correlation between two diseases. Here, a random walk was performed from a set of genes to a gene interaction network. The method simply captures the level of reachability from a set of genes to another. Following that, a random walk based technique can be used to get the co-morbid pairs in a network. As MCL [45] follows the same random walk logic for graph clustering, in the proposed work it is used to get the cluster of correlated diseases.

A disease-protein-protein interaction network (DPPIN) (shown in Figure 2) is built with a set of disease ICD-9-CM codes and PPI network, where each of the proteins is mapped to their EntrezIDs. In this DPPIN, the diseases are linked to the PPI network while considering the disease-gene associations. The ICD-9-CM codes and EntrezIDs are considered as nodes. The links between ICD-9-CM to EntrezID and EntrezID to EntrezID are represented by edges. MCL algorithm is then applied to this DPPIN to obtain the disease-gene clusters.

We only choose the clusters with more than one ICD-9-CM codes and discard the remaining. We also don't consider the clusters with only EntrezIDs. In the next step, only the ICD-9-CM codes within a cluster are grouped and the EntrezIDs are discarded. All possible combination of disease pairs are generated from this cluster, and the same process is followed for all the clusters. Finally, a set of disease pairs is obtained from these combinations. The ICD-9-CM code pairs if already exist in the gold standard co-morbid disease data, are discarded from this newly generated co-morbid pairs. We include these newly predicted disease pairs into the proposed knowledge graph (an example with real data is given in the supplementary Figure S3) to train the model accordingly. The workflow of the proposed approach is given in the supplementary Figure S1 and an example of applying MCL on DPPIN with real data, is given in the supplementary Figure S4.

4 Experiment

4.1 Edge Prediction

We predict the links or edges (relations) between a set of entities by using cross-validation method. Our experiment verifies the links between the gold standard co-morbid disease pairs with their corresponding biological relations.

In our experiment, we perform the 10-fold cross-validation where we select first 8-folds for training, one for testing and one for validation. To measure the performance ranking for each relation in the test set, the same substitutions of heads and tails are followed as suggested by Trouillon et al. [18] in their paper. The evaluation of test triples is done by calculating the Mean Reciprocal Rank (MRR) and the Hits at N. Elaborately, for a test triple, *(h, r, t)* we calculate the ranking of triple having h, by computing the score of $(h^{'}, r, t)$ triples for all $h^{'} \in \mathcal{E}$ and similarly, calculate the rank of a triple having t, by computing the score of $(h, r, t^{'})$ triples for all $t^{'} \in \mathcal{E}$ (where, $\mathcal{E}$ is the set of all entities). This measure is called as Mean Reciprocal Rank (MRR). MRR can be expressed as

$$MRR = \frac{1}{2 * |T|} \sum_{(h,r,t)\in T} \frac{1}{rank_h} + \frac{1}{rank_t} \quad (1)$$

Where T defines the test triples. This MRR is referred to as raw MRR. Bordes et al. [12] identifies that there is an issue in calculating the score with raw MRR, as it may give a flawed result. So they proposed a concept of filtered MRR. In filtered MRR, instead of finding the rank of an individual test triple *(h, r, t)* among all $(h^{'}, r, t)$ for all $h^{'} \in \mathcal{E}$ and $(h, r, t^{'})$ triples for all $t^{'} \in \mathcal{E}$, Bordes et al. proposed to compute the rank among triples $(h^{'}, r, t)$ only for $h^{'} \in \mathcal{E}$ such that $(h^{'}, r, t) \notin train \cup test \cup valid$. In our result we only produce the filtered MRR results with the filtered hits scores at 1, 3 and 10 values, where hits@N represent the percentage of test triples whose ranking is less than or equal to N. In the first set, our dataset only contains positive triples, so negative samples are generated using local close-world assumption similarly as proposed by Trouillon et al. [18]. For, ComplEx embedding the rank is considered as 200, λ as 0.01 and η as 10, where λ is L^2 regularization parameter and η is the number of negatives generated per positive triple. Here, we selected 100 batches per epoch as suggested by Trouillon et al. [18].

4.2 Triple Classification

Triple classification task is performed to verify whether an unseen triple fact (h, r, t) is true or false. This triple classification task is modeled as a binary classification task in usual [12], [15]. We use the OpenKE framework [46] for this triple classification task. For, the relation specific binary classification we follow the same logic proposed by Socher et al. [15].

As in the knowledge graph, all the test instances are true; the test data need to be corrupted for the classification task. So, for each of the test instances, a corresponding negative instance is made for each relation in the knowledge graph. In the proposed work, a relation specific triple is classified by a score obtained by a function f. If the accrued score for a specific triple is below than the relation-specific threshold δ_r, it is classified as a fact, otherwise, it is classified as a false fact. The value of δ_r is optimized by maximizing the classification accuracy on the validation dataset, and for different relations, different values of δ_r are set.

To evaluate the relation specific performance of our proposed method, we use precision, recall and F1 score as follows.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2927310, IEEE/ACM Transactions on Computational Biology and Bioinformatics

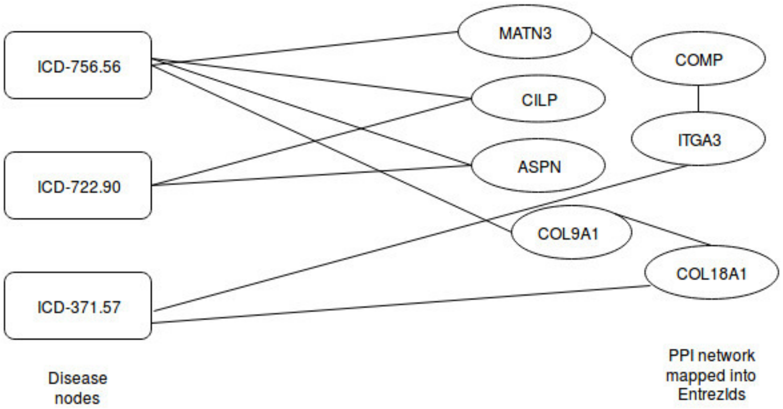

Fig. 2: Sub-Graph of Disease-Protein-Protein-Interaction-Network

TABLE 1: Filtered Mean Reciprocal Rank (MRR) and Hit@N on Each Relation

Relation name	MRR	Hit@1	Hit@3	Hit@10
_ has_ function	0.761	0.728	0.785	0.816
_ has_ interaction	0.778	0.747	0.807	0.814
_ is_ treated_ by	0.887	0.848	0.926	0.943
_ disrupts	0.555	0.468	0.598	0.728
_ is_ function_ of	0.765	0.732	0.789	0.818
_ has_ side_ effects	0.815	0.788	0.828	0.858
_ side_ effect_ of	0.805	0.775	0.821	0.849
_ is_ caused_ by	0.816	0.767	0.857	0.889
_ has_ disease_ phenotype	0.199	0.141	0.209	0.313
_ involves	0.433	0.334	0.474	0.623
_ has_ gene_ phenotype	0.852	0.820	0.879	0.896
_ associated_to*	**0.985**	**0.977**	**0.992**	**0.999**
_ treats	0.898	0.857	0.936	0.955
_ causes	0.817	0.768	0.857	0.890
_ is_ gene_ phenotype_ of	0.857	0.824	0.884	0.901

* defines disease-disease relation

$$precision = \frac{tp}{tp + fp} \tag{2}$$

$$recall = \frac{tp}{tp + fn} \tag{3}$$

$$F1\,score = 2 * \frac{precision * recall}{precision + recall} \tag{4}$$

Where, tp, fp and fn stand for true positive, false positive and false negative, respectively.

5 Results

5.1 Edge Prediction on Known Co-morbid diseases and Triple Classification including newly generated Co-morbid disease pairs

In this paper, the prediction of an edge between two entities gives the result as shown in Table 1, where the relation _ *associated_ to*, defines the comorbidity between disease pairs and the other relations define the background information regarding disease and its associated genes. For example, the relation _ *treats* defines the specific drugs used to treat certain diseases. Similarly, _ *has_ gene_ phenotype* and _ *has_ function* define the genes to phenotypes and genes to their corresponding relations, respectively. In our edge prediction result, the filtered MRR (Mean Reciprocal Rank) score of disease-disease associations gives a very good score (as higher the MRR scores, better the results) and also with that, the Hits score at ten also gives a pretty high score. The relation between entities is determined by a certain threshold value, and it is chosen according to (maximizing) the classification accuracy on the validation set. In Table 1, it is reflected that only the relation, _ *has_ disease_ phenotype* gives lesser MRR score than the others.

The overall triple classification accuracy considering all the relations is 0.9317. As our present work predicts the co-morbid disease pairs, it only produces the triple classification result of _ *associated_ to* relation. To investigate the performance of our proposed method, we have chosen the disease co-morbidity prediction results of three recent well-established methods. Namely, the methods proposed by Maddouri [30], PCID [47] and comoR [48]. The R package of the comoR method implemented three algorithms as, Co-morbidityPath, Co-morbidityDO, and Co-morbidityOMIM. If the pair of diseases share common pathways according to the KEGG database, Co-morbidityPath algorithm predicted them as prevalent disease pairs. Similarly, Co-morbidityDO and Co-morbidityOMIM predicted coexisting diseases based on common phenotypes (DO database) and sharing genes (OMIM database), respectively. As shown by Moni et al. [48], these algorithms outperform the earlier approaches [49], [50], [51], it is enough to use these methods as the first benchmark result for comparison. The other approach proposed by He et al. [47] excels the comoR method. Finally, Maddouri [30] proposed a knowledge graph-based method, which outperforms all the previous approaches. All of these methods are well defined and justified for *state-of-the-art* comparison with our proposed method.

The performance of the proposed work with the other prediction methods are shown in the Table 2.

In our method, we consider total of 645 diseases as gold standard data and cross-validate the whole dataset for getting the best hyper-parameters according to specific relations. Here, for each test instance, we calculate the score and check if its value is below than a threshold δ_r. If the obtained score is below than the δ_r, it is predicted as true otherwise false. The threshold δ_r is certainly calculated for each relation by maximizing classification accuracy on the

TABLE 2: The Benchmark Results of Different Methods

Method	Precision	Recall	F1 Score
Proposed Method	**0.977**	**0.984**	**0.981**
Maddouri [30]	0.918	0.9258	0.918
PCID [47]	0.761	0.803	0.778
Co-morbidityPath [48]	0.538	0.519	0.521
Co-morbidityDO [48]	0.522	0.481	0.496
Co-morbidityOMIM [48]	0.415	0.352	0.387

TABLE 3: Predicted Co-morbid pairs found in literature

Disease1 (ICD-9-CM)	Disease2 (ICD-9-CM)	PMID
586	530.81	11887839
205.9	372.3	27168715
151.6	240	23844325
084.6	322	24748325
151.6	336.9	10879773
577	528.8	24348612
192	435.8	15051699
286	176.4	20606415
586	099.3	21722341

validation set. As the gold dataset used here are well studied and forming a dense subgraph, the model predicts well on the dataset. Apart from that after including the MCL obtained dataset to train the model, interestingly we again get good novel predictions on the test set.

5.2 Newly Predicted Co-morbid Disease Pairs

Apart from producing the comparison with the gold standard dataset, our method also predicts some new co-morbid disease pairs. Later, to validate the predictions we further queried the PubMed database to search if those have been reported in the literature. Among all of the new prediction interestingly some have been found in the literature (see the supplementary Table S1). Some of the pairs also show common sharing of pathways and some with common genes. For, instance gastric cancer (ICD-9-CM: 151.6) and goiter (ICD-9-CM: 240) are predicted to be co-morbid in our proposed method. It has been shown that iodine deficiency is more frequent cases in gastric adenocarcinoma (stomach cancer) patients. So, goiter frequency is more prevalent with gastric cancer than usual urine iodine level [52]. Our model also predicts the co-morbid relationship between renal failure (ICD-9-CM: 586) and reactive arthritis (ICD-9-CM: 099.3). In literature, it has been found that the patients with reactive arthritis also present with IgA nephropathy or amyloid-A nephropathy [53]. This result verifies the association between renal disease and reactive arthritis.

Table 3 shows effective evidence from the literature that validates our newly predicted prevalent disease pairs, which certainly implies the effectiveness of our proposed method.

Further, we analyze those predictions and find that the correlated disease pairs share common pathways or genes or even common pathways and genes among them. A few also share common PPIs among them. For instance, gastric cancer (ICD-9-CM: 151.6) and paraganglion neoplasm (ICD-9-CM: 237.3) share interaction between KIT and PDGFRA [54]. Another co-morbid disease pairs, renal failure (ICD-9-CM: 586) and autoimmune lymphoproliferative syndrome both are related to IL-10 and BCL-2 [55], [56], [57].

TABLE 4: Pathway Patterns in Predicted Co-morbid Disease Pairs with literature validation

Disease (ICD-9-CM)	Co-morbid Diseases (ICD-9-CM)	Sharing Pathway
586	443.81	Membrane Trafficking
	410.9	
586	141.9	GPCR Signaling Pathway
	031.9	
151.6	237.3	Pyruvate Metabolisms & TCA Cycle
	428.9	
577	528.8	Metabolism
	536.3	
586	305	Cytokine Signaling in Immune System
	429	

5.3 Pathway Patterns in Predicted Co-morbid Disease Pairs

The molecular pathway information often seems essential for the analysis of disease co-morbidity and their hidden biological mechanisms. So, we further investigate the common sharing pathways between the newly predicted prevalent disease pairs. Interestingly in our findings, we get some disease pairs with literature validation for their overlapping pathways. For instance, the presence of mutations leading to protein trafficking defect can be defined as a mechanism of pathogenesis seen in renal failure and peripheral angiopathy [58]. Other study showed that protein trafficking is a dynamic, ongoing occurrence and critically impacts on channel function, which is highly relevant for heart failure [59]. So, the collective sharing of similar pathway among the diseases, renal failure (ICD-9-CM: 586), peripheral angiopathy (ICD-9-CM: 443.81) and heart failure (ICD-9-CM: 410.9) justifies their coexistence. Interleukin-6 (IL-6) is a pleiotropic cytokine that regulates the immune and inflammatory response of the body. Recently, the IL-6 signaling pathway and the interplay of IL-6 to renal-resident cells are explained by Hua et al. [60]. They demonstrated that, in nephrotoxin induced acute kidney disease, IL-6 expression was dramatically enhanced in the kidney, primarily in renal tubular epithelial cells, and strongly correlated with the damage of the organ. Th2 cytokine responses are mostly visible in the most severe forms of myocarditis (ICD-9-CM: 429) where eosinophils are prominent [61]. Other study showed the up-regulation of the cytokine signaling pathway in alcohol dependence disease (ICD-9-CM: 305) [62]. Some of the literature validated co-morbid disease pairs with common sharing pathway are shown in the Table 4.

6 Discussion and Conclusion

In this paper, we propose a novel tensor factorization based computational approach for co-morbid disease prediction. Our approach has several advantages over existing prevalent disease prediction approaches. First, we integrate all disease-gene relations with other biological and biomedical entities to build a knowledge graph, which integrates multiple relations. Second, instead of using any automatic reasoning toolkit for knowledge graph completion, we use

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2927310, IEEE/ACM Transactions on Computational Biology and Bioinformatics

a knowledge graph embedding approach which itself has taken care of all binary relational properties (reflexive, symmetric, transitive) in the graph. Third, we do not use any DeepWalk [63] and Word2Vec [64] models externally for generating the feature vectors from the knowledge graph. Our proposed model generates a random vector for entities and relations in the knowledge graph and updates it accordingly with the learning of the model by minimizing the loss function. We use the stochastic gradient descent method with AdaGrad optimizer (as it gives the best performance) for learning the model. For, dealing with anti-symmetric relations we first introduce complex vector embedding and hermitian dot product in a biological knowledge graph.

Other previous approach namely, PCID was based on a very small set of well-studied diseases that fails to predict many new possible disease pairs. In contrast to that, to predict new prevalent disease pairs apart from the gold dataset, we apply an MCL based method on disease-protein-protein interaction network (DPPIN). Later, we combine these newly generated disease pairs in the knowledge graph and predicted new prevalent diseases with ComplEx embedding model.

Our proposed method of coexisting disease prediction, outperforms all the earlier well-established techniques, and it is produced in the Table 2. The presented approach not only predicts the co-morbid disease pairs on the gold dataset with impressive performance but also predicts new co-morbid disease pairs, some of which later validated with medical literature. Thus the promising aspect of our method implies that any possible disease pairs can be predicted with our approach from any biological knowledge graph.

In this method, the tensor decomposition is done with complex-valued latent factors. For all real square matrices, this decomposition exists and represented as a real part of normal matrices. In ComplEx embedding the result extends to a set of real square matrices which are referred to as tensors and this task is done through the knowledge graph completion while handling the symmetric and asymmetric relations. The method shows that biological relations can also be approximated as the real part of low-rank standard matrices. Besides that, the effectiveness of this robust embedding method also introduces the concept of linear algebra in biological knowledge graphs.

In spite of the significant performance benefits with the *state-of-the-art*, there is still space for improvement of the work. In our method, we only integrate data from well known available databases. As all the entity relations are not present in these databases, our knowledge graph suffers from some missing biological information which may contribute further for a better understanding of the prevalent disease pairs. Next, we can integrate more gene molecular level information that will also help to produce intimate relations between different diseases. The model takes a considerable amount of time for training. So, it can be further modified to a better way of generating more informative negatives to obtain good performance in relatively reduced training time. Finally, we hope that the proposed work will be useful for the prediction of disease pairs with their drug indications and their corresponding side-effects which may further help to contribute to the research of drug repositioning.

Appendix A

Additional file 1: Supplementary document, available online. The file contains the workflow of the proposed approach, link for availability of data and codes, supplementary Figs. S1-S4 and supplementary Table S1, available online.

References

[1] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 601–610.

[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 2008, pp. 1247–1250.

[3] C. Fellbaum, "A semantic network of english verbs," *WordNet: An electronic lexical database*, vol. 3, pp. 153–178, 1998.

[4] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[5] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 697–706.

[6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell, "Toward an architecture for never-ending language learning." in *AAAI*, vol. 5. Atlanta, 2010, p. 3.

[7] T. Katayama, M. D. Wilkinson, K. F. Aoki-Kinoshita, S. Kawashima, Y. Yamamoto, A. Yamaguchi, S. Okamoto, S. Kawano, J.-D. Kim, Y. Wang *et al.*, "Biohackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains," *Journal of biomedical semantics*, vol. 5, no. 1, p. 5, 2014.

[8] L. De Raedt, A. Kimmig, and H. Toivonen, "Problog: A probabilistic prolog and its application in link discovery," 2007.

[9] P. Domingos, S. Kok, D. Lowd, H. Poon, M. Richardson, and P. Singla, "Markov logic," in *Probabilistic inductive logic programming*. Springer, 2008, pp. 92–117.

[10] N. Lao and W. W. Cohen, "Fast query execution for retrieval models based on path-constrained random walks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 881–888.

[11] M. Nickel, V. Tresp, and H.-P. Kriegel, "Factorizing yago: scalable machine learning for linked data," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 271–280.

[12] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in neural information processing systems*, 2013, pp. 2787–2795.

[13] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *International Conference on Machine Learning*, 2016, pp. 2071–2080.

[14] Y. Lin, Z. Liu, H. Luan, M. Sun, S. Rao, and S. Liu, "Modeling relation paths for representation learning of knowledge bases," *arXiv preprint arXiv:1506.00379*, 2015.

[15] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in neural information processing systems*, 2013, pp. 926–934.

[16] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Journal of Mathematics and Physics*, vol. 6, no. 1-4, pp. 164–189, 1927.

[17] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," *arXiv preprint arXiv:1412.6575*, 2014.

[18] T. Trouillon, C. R. Dance, É. Gaussier, J. Welbl, S. Riedel, and G. Bouchard, "Knowledge graph completion via complex tensor factorization," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4735–4772, 2017.

[19] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse, "Relations in biomedical ontologies," *Genome biology*, vol. 6, no. 5, p. R46, 2005.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2927310, IEEE/ACM Transactions on Computational Biology and Bioinformatics

[20] D. Wood, M. Zaidman, L. Ruth, and M. Hausenblas, *Linked Data*. Manning Publications Co., 2014.

[21] U. Consortium, "Uniprot: a hub for protein information," *Nucleic acids research*, vol. 43, no. D1, pp. D204–D212, 2014.

[22] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2rdf: towards a mashup to build bioinformatics knowledge systems," *Journal of biomedical informatics*, vol. 41, no. 5, pp. 706–716, 2008.

[23] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi *et al.*, "The ebi rdf platform: linked open data for the life sciences," *Bioinformatics*, vol. 30, no. 9, pp. 1338–1339, 2014.

[24] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker *et al.*, "Pubchem substance and compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2015.

[25] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?" *International journal of human-computer studies*, vol. 43, no. 5-6, pp. 907–928, 1995.

[26] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases," *Scientific reports*, vol. 5, p. 10888, 2015.

[27] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler, "Owl 2: The next step for owl," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 4, pp. 309–322, 2008.

[28] I. Horrocks, "Obo flat file format syntax and semantics and mapping to owl web ontology language," *University of Manchester*, 2007.

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[30] O. Maddouri, "Deep learning on biological knowledge graphs," Ph.D. dissertation, Hamad Bin Khalifa University (Qatar), 2017.

[31] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, p. 25, 2000.

[32] S. Köhler, S. C. Doelken, C. J. Mungall, S. Bauer, H. V. Firth, I. Bailleul-Forestier, G. C. Black, D. L. Brown, M. Brudno, J. Campbell *et al.*, "The human phenotype ontology project: linking molecular biology and disease through phenotype data," *Nucleic acids research*, vol. 42, no. D1, pp. D966–D974, 2013.

[33] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant *et al.*, "Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," *Nucleic acids research*, vol. 43, no. D1, pp. D1071–D1078, 2014.

[34] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork *et al.*, "The string database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D561–D568, 2010.

[35] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular systems biology*, vol. 6, no. 1, p. 343, 2010.

[36] J. Oyston, "Online mendelian inheritance in man," *Anesthesiology: The Journal of the American Society of Anesthesiologists*, vol. 89, no. 3, pp. 811–812, 1998.

[37] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The genetic association database," *Nature genetics*, vol. 36, no. 5, p. 431, 2004.

[38] A. P. Davis, T. C. Wiegers, P. M. Roberts, B. L. King, J. M. Lay, K. Lennon-Hopkins, D. Sciaky, R. Johnson, H. Keating, N. Greene *et al.*, "A ctd–pfizer collaboration: manual curation of 88 000 scientific articles text mined for drug–disease and drug–phenotype interactions," *Database*, vol. 2013, 2013.

[39] A. Fabregat, F. Korninger, G. Viteri, K. Sidiropoulos, P. Marin-Garcia, P. Ping, G. Wu, L. Stein, P. DEustachio, and H. Hermjakob, "Reactome graph database: Efficient access to complex pathway data," *PLoS computational biology*, vol. 14, no. 1, p. e1005968, 2018.

[40] K. Sidiropoulos, G. Viteri, C. Sevilla, S. Jupe, M. Webber, M. Orlic-Milacic, B. Jassal, B. May, V. Shamovsky, C. Duenas *et al.*, "Reactome enhanced pathway visualization," *Bioinformatics*, vol. 33, no. 21, pp. 3461–3467, 2017.

[41] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.

[42] J. Park, D.-S. Lee, N. A. Christakis, and A.-L. Barabási, "The impact of cellular networks on disease comorbidity," *Molecular systems biology*, vol. 5, no. 1, p. 262, 2009.

[43] M. Alshahrani, M. A. Khan, O. Maddouri, A. R. Kinjo, N. Queralt-Rosinach, and R. Hoehndorf, "Neuro-symbolic representation learning on biological knowledge graphs," *Bioinformatics*, vol. 33, no. 17, pp. 2723–2730, 2017.

[44] Y. Ko, M. Cho, J.-S. Lee, and J. Kim, "Identification of disease comorbidity through hidden molecular mechanisms," *Scientific reports*, vol. 6, p. 39433, 2016.

[45] S. M. Van Dongen, "Graph clustering by flow simulation," Ph.D. dissertation, 2000.

[46] X. Han, S. Cao, X. Lv, Y. Lin, Z. Liu, M. Sun, and J. Li, "Openke: An open toolkit for knowledge embedding," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 139–144.

[47] F. He, G. Zhu, Y.-Y. Wang, X.-M. Zhao, and D.-S. Huang, "Pcid: A novel approach for predicting disease comorbidity by integrating multi-scale data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 14, no. 3, pp. 678–686, 2017.

[48] M. A. Moni and P. Liò, "comor: a software for disease comorbidity risk assessment," *Journal of clinical bioinformatics*, vol. 4, no. 1, p. 8, 2014.

[49] H. Paik, H.-S. Heo, H.-j. Ban, and S. B. Cho, "Unraveling human protein interaction networks underlying co-occurrences of diseases and pathological conditions," *Journal of translational medicine*, vol. 12, no. 1, p. 99, 2014.

[50] S. Park, J.-S. Yang, Y.-E. Shin, J. Park, S. K. Jang, and S. Kim, "Protein localization as a principal feature of the etiology and comorbidity of genetic diseases," *Molecular systems biology*, vol. 7, no. 1, p. 494, 2011.

[51] S. Park, J.-S. Yang, J. Kim, Y.-E. Shin, J. Hwang, J. Park, S. K. Jang, and S. Kim, "Evolutionary history of human disease genes reveals phenotypic connections and comorbidity among genetic diseases," *Scientific reports*, vol. 2, p. 757, 2012.

[52] M. Tabaeizadeh, V. Haghpanah, A. Keshtkar, S. Semnani, G. Roshandel, K. Adabi, R. Heshmat, D. Rohani, A. Kia, E. Hatami *et al.*, "Goiter frequency is more strongly associated with gastric adenocarcinoma than urine iodine level," *Journal of gastric cancer*, vol. 13, no. 2, pp. 106–110, 2013.

[53] H.-J. Anders and V. Vielhauer, "Renal co-morbidity in patients with rheumatic diseases," *Arthritis Research & Therapy*, vol. 13, no. 2, p. 222, 2011.

[54] T. Nishida, J.-Y. Blay, S. Hirota, Y. Kitagawa, and Y.-K. Kang, "The standard diagnosis, treatment, and follow-up of gastrointestinal stromal tumors based on guidelines," *Gastric Cancer*, vol. 19, no. 1, pp. 3–14, 2016.

[55] O. Niss, A. Sholl, J. J. Bleesing, and D. A. Hildeman, "Il-10/janus kinase/signal transducer and activator of transcription 3 signaling dysregulates bim expression in autoimmune lymphoproliferative syndrome," *Journal of Allergy and Clinical Immunology*, vol. 135, no. 3, pp. 762–770, 2015.

[56] G. Gobe, X.-J. Zhang, D. A. Willgoss, E. Schoch, N. A. Hogg, and Z. H. Endre, "Relationship between expression of bcl-2 genes and growth factors in ischemic acute renal failure in the rat," *Journal of the American Society of Nephrology*, vol. 11, no. 3, pp. 454–467, 2000.

[57] I. Sinuani, I. Beberashvili, Z. Averbukh, and J. Sandbank, "Role of il-10 in the progression of kidney disease," *World journal of transplantation*, vol. 3, no. 4, p. 91, 2013.

[58] C. Schaeffer, A. Creatore, and L. Rampoldi, "Protein trafficking defects in inherited kidney diseases," *Nephrology Dialysis Transplantation*, vol. 29, no. suppl_4, pp. iv33–iv44, 2014.

[59] S. Xiao and R. M. Shaw, "Cardiomyocyte protein trafficking: Relevance to heart disease and opportunities for therapeutic intervention," *Trends in cardiovascular medicine*, vol. 25, no. 5, pp. 379–389, 2015.

[60] H. Su, C.-T. Lei, and C. Zhang, "Interleukin-6 signaling pathway and its role in kidney disease: an update," *Frontiers in immunology*, vol. 8, p. 405, 2017.

[61] N. R. Rose, "Critical cytokine pathways to cardiac inflammation," *Journal of Interferon & Cytokine Research*, vol. 31, no. 10, pp. 705–710, 2011.

[62] R. D. Beech, J. Qu, J. J. Leffert, A. Lin, K. A. Hong, J. Hansen, S. Umlauf, S. Mane, H. Zhao, and R. Sinha, "Altered expression

of cytokine signaling pathway genes in peripheral blood cells of alcohol dependent subjects: preliminary findings," *Alcoholism: Clinical and Experimental Research*, vol. 36, no. 9, pp. 1487–1496, 2012.

[63] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.

[64] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

Saikat Biswas received his Master degree in Computer Science and Engineering from Jadavpur University, Kolkata. He is a doctoral student at Advanced Technology Development Centre, IIT Kharagpur. His research focuses on data mining and computational biology.

Pabitra Mitra received his Ph.D. degree in Computer Science from Indian Statistical Institute, Kolkata. He is a professor at the Department of Computer Science and Engineering, IIT Kharagpur. His research focuses on machine learning, pattern recognition, data mining, information retrieval, and computational biology. He is a recipient of prestigious Indian National Academy of Engineering Young Engineer Award and Royal Society UK Indian Science Network Award. He is a member of the Indian Unit for Pattern Recognition and Artificial Intelligence. He is a senior member of IEEE.

Krothapalli Sreenivasa Rao recieved his Ph.D. degree in Computer Science and Engineering from IIT Madras. He is a professor at the Department of Computer Science and Engineering, IIT Kharagpur. His research focuses on signal processing, speech processing, machine learning, pattern recognition, and biomedical engineering. He is an editorial board member of several journals. He is a senior member of IEEE.