Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

Mathematical Biosciences and Engineering, 16(3): 1376–1391
DOI: 10.3934/mbe.2019067
Received: 14 December 2018
Accepted: 22 January 2019
Published: 20 February 2019

Research article

GOF/LOF knowledge inference with tensor decomposition in support of high order link discovery for gene, mutation and disease

Kaiyin Zhou[1,2], Yuxing Wang[1,2], Sheng Zhang[3], Mina Gachloo[1,2], Jin-Dong Kim[4], Qi Luo[3], Kevin Bretonnel Cohen[5] and Jingbo Xia[1,2,*]

[1] College of Informatics, Huazhong Agricultural University, 430070, Wuhan, China

[2] Hubei Key Lab of Agricultural Bioinformatics, Huazhong Agricultural University, 430070, Wuhan, China

[3] College of Science, Huazhong Agricultural University, 430070, Wuhan, China

[4] Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS), Tokyo, Japan

[5] School of Medicine, University of Colorado Denver, Anschutz Medical Campus, Colorado, U.S

* **Correspondence:** Email: xiajingbo.math@gmail.com, xjb@mail.hzau.edu.cn; Tel:+86-02787288509.

Abstract: For discovery of new usage of drugs, the function type of their target genes plays an important role, and the hypothesis of "Antagonist-GOF" and "Agonist-LOF" has laid a solid foundation for supporting drug repurposing. In this research, an active gene annotation corpus was used as training data to predict the gain-of-function or loss-of-function or unknown character of each human gene after variation events. Unlike the design of(entity, predicate, entity) triples in a traditional three way tensor, a four way and a five way tensor, GMFD-/GMAFD-tensor, were designed to represent higher order links among or among part of these entities: genes(G), mutations(M), functions(F), diseases(D) and annotation labels(A). A tensor decomposition algorithm, CP decomposition, was applied to the higher order tensor and to unveil the correlation among entities. Meanwhile, a state-of-the-art baseline tensor decomposition algorithm, RESCAL, was carried on the three way tensor as a comparing method. The result showed that CP decomposition on higher order tensor performed better than RESCAL on traditional three way tensor in recovering masked data and making predictions. In addition, The four way tensor was proved to be the best format for our issue. At the end, a case study reproducing two disease-gene-drug links(Myelodysplatic Syndromes-IL2RA-Aldesleukin, Lymphoma-IL2RA-Aldesleukin) presented the feasibility of our prediction model for drug repurposing.

Keywords: drug repurposing; tensor decomposition; relation extraction

1. Introduction

Drug repurposing is to develop new uses of drugs beyond their initially approved indications. It helps discovery and identification of novel therapies for diseases, at a lower cost and in a shorter time frame, compared to traditional methods [1]. Computational methods have been proposed as a cost-effective method of doing drug repurposing. Though in most cases the computational method just discover close therapeutic drug uses which are close to their original use, there are still striking and successful attempts for new discoveries [2]. The difficulties of current computational method come from the structural complexity of high order relationships related to drug repurposing. To overcome these drawbacks, we explored the methods of using various n-way tensors to store high order relation data. By using tensor decomposition computational strategy, novel links of genes and diseases was achieved by linking all of entities together, including the mutation, gene and disease.

Tensors are multidimensional array of numerical data with a n-way structure, the theorem of which has been utilized in many machine learning domains. The first pioneering work appeared in 1927 with clear definition of tensors and decomposition theorems [3]. In order to construct knowledge graph from unstructured data, Nickel et al. [4] proposed RESCAL, which modeled relations in the triples of the form ($subject$, $predicate$, $object$). Recently, Nimishakavi et al. [5] presented a high order tensor factorization to perform higher-order relation schema induction. The same team performed the (entity-predicate-entity) schema induction by integrating side information into a RESCAL-based [4] three way tensors and achieved efficient computation result [6]. While the last two showed promising results, due to the high complexity of the problem, these recent techniques were determined not to be applicable to drug-related Omics knowledge discovery. However, a recent work from Lacroix et al. [7] suggested Canonical polyadic (CP) decomposition worked fine with the knowledge base inference.

An early work of tensor decomposition in phenotyping discovery was done by Ho et al. [8] in 2014. Limestone, a nonnegative tensor decomposition method, was introduced in this work to derive phenotype candidates from electronic health record without human supervision. After factoring the tensor, the interaction of diagnoses and medications among patients were investigated. Eventually, it confirmed that 82% of the top 50 candidates were clinically meaningful. Fang and Jonathan [9] took the first step toward modeling complex genomic and epigenomic data including mRNA, methylation, copy number variations and somatic mutations by merging into a high-order tensor. They developed a predictive model for overall survival by using CP decomposition. Taking granted that the interaction between a drug and a protein is context dependent, Taguchi et al., [10] put multi-directional data including gene and the various conditions, i.e., including diseases, patients, tissues, and time points, into a high order tensor, and carried on a Tucker decomposition to perform new drug recommendation. He identified two promising therapeutic-target genes, CYPOR and HNFA4 for cirrhosis, and suggested bezafibrate as a promising candidate drug. The result was supported by an *in silico* docking analysis.

In light of both the CP decomposition for novel link discovery and popular RESCAL based triple format predicate recognition, in this paper, tensor was applied to represent concise and clinically meaningful phenotypes, and new indications were discovered by tensor decomposition.

Specifically, four kinds of information were taken into considerations, which included genes (G), types of mutations (M) and functions of mutations (F) and diseases (D). In addition, each two of them has meaningful relationship and annotation (A) in a well annotated corpus, the Active Gene Annotation Corpus (AGAC) [11].

The aiming issue is to infer a high order relation: a *Gene* (after *Mutation*) play a GOF/LOF/Unknown *Function* in the circumstannce of *Disease*. Under the novel link discovery strategy, any new value in the updated tensor was capable of inferring new relevance among entities. The state-of-art design is to consider the selection of *gene, mutation, disease* for each entity. If using RESCAL strategy, a three way tensor contained all various entities in the first two ways and GOF/LOF/Unknown the third way. Meanwhile, a natural idea is to put *gene, mutation, disease* into three separate ways and GOF/LOF/Unknown into the forth ways. Thus a four way tensor was obtained. Similarly, if taking consideration of AGAC annotations, an additional way was added, thus a five tensor was constructed.

Generally speaking, a high order tensor achieved a more precise higher relationship among entities, while it also led to a higher sparsity in tensor cells that decrease the link discovery reliability.

Therefore, in order to explorer the trade-off between the dimension of tensor and the effectiveness of knowledge discovery in terms with tensor decomposition, all three-way, four-way and five-way tensors were used for tensor-form data structure. Here, a traditional three way tensor was built which suited RESCAL algorithm. Instead of creating excessive amount arrays to store the relation of them, a four-way tensor (denoted as GMFD-tensor) was also designed to represent the high order relations of the multiple tuple informations including gene, mutation, function and disease. Meanwhile, a five-way tensor was designed to use enriched annotation information.

In this research, CP decomposition was used for four way and five way tensors. Meanwhile, a state-of-art RESCAL baseline method was carried on to the three-way tensor as well. Comparison results of the experiments showed that the four way GMFD tensor decomposition was the best candidate for novel high order link discovery in this circumstances. Finally, a case study was performed in this research from which a novel drug-gene-disease link ”(Lymphoma, Myelodysplatic Syndromes)-IL2RA-Aldesleukin” was reproduced through the algorithm.

2. Material and Method

2.1. OMIM text data collection and the AGAC corpus

This part showed our data collecton and introduced training set, AGAC corpus.

2.1.1. OMIM text data collection

The text data were manually collected from Online Mendelian Inheritance in Man (OMIM) (`https://www.omim.org/`), which is a public database of bibliographic information about human genes and genetic disorders. One of the most valuable features of OMIM is the list of ”Allelic Variants” for a given gene. There are general criteria for selecting specific mutations of a given gene, including: the first or first few disease-related mutations to be identified in the given gene; any mutation with a particularly high frequency; mutations of historical interest.

In total, there are 1,178 OMIM entries curated manually. Each OMIM entry includes a full text summery of a genetic phenotype and/or gene and has many links to other genetic resources such as DNA and protein sequence, PubMed references, mutation Databases and known mendelian disorders.

Each collected text recorded a mutation in a gene regulated a biological function and induced a disease. According to the effect of a mutation of a gene to its function, down-regulation or up-regulation, the effect was classified into loss of function (LOF) or gain of function (GOF). An automatic classifier was developed for the purpose, based on AGAC.

The data size and samples of relevant genes and mutations are shown in Table 1.

Table 1. Data format and data size.

Text ID	Gene (MIMID)	Mutation	Disease	Function
1	B2M(109700)	ALA11PRO	Immunodeficiency	LOF
2	ACADSB(600301)	IVS3DS, A-G	2-Methylbutyryl-CoA Dehydrogenase Deficiency	LOF
3	PIK3R1(171833)	1-BP INS, 1906C	Short Syndrome	LOF
4	RYR1(180901)	GLY2434ARG	Malignant Hyperthermia	GOF
5	SCN1A(182389)	GLN1489LYS	Migraine	GOF
6	SIGMAR1(601978)	IVS1DS, G-T	Spinal Muscular Atrophy Distal	Unknown
7	COLEC11(612502)	3-BP DEL, 648CTC	3Mc Syndrome	Unknown
1178	197	198	307	3

2.1.2. Active Gene Annotation (AGAC) Corpus for GOF/LOF knowledge discovery

In order to fill in the cell value in the tensor as shown in the subsequent section, a corpus was needed for GOF/LOF/Unknown predication. For this purpose, an active gene annotation corpus (AGAC) and built-in classifiers [11] were used for this research.

AGAC was a OMIM text based corpus with abundant fine-grained annotations. The annotations in AGAC were centered on gene mutation events, which includes the mutations, the functions effected and how they were effected. Actually, AGAC involved two types of annotation labels: biology concept trigger labels, root regulatory trigger labels. Among them, bio-concept trigger labels includes: 1. Variation, 2. Molecular Physiological Activity, 3. Interaction, 4.Pathway, and 5. Cell physiological activity. Meanwhile, three regulatory trigger labels were designed: 6. Positive Regulation 7. Negative Regulation 8. Regulation

2.1.3. AGAC-based GOF/LOF/Unknown classifiers

For each text block labeled with given gene and mutation, a three-class text classification was carried on to predict *function* either being GOF, LOF, or Unknown . Some existed experiments in the text classification showed that

AGAC annotation labels improved the prediction accuracy of GOF and LOF, thus support the reasonability of the corpus design. The classifier was built by using Bidirectional Long Short-Term Memory Network (Bi-LSTM). In the first non-artificial-features engineering network, Bi-LSTM was used to extract contextual features, and the Softmax layer was utilized to decide which category the text belongs. In the second features engineering network, in addition to using the contextual features of Bi-LSTM encoded, we count each type of label in separate samples, and encode it into a 12-dimensional vector, after that we concatenate artificial and contextual features together, then Softmax was used for final input layer.

2.2. *Knowledge format of high order relationship, and fundamentals of tensor decomposition for novel knowledge discovery*

2.2.1. Knowledge format of high order relationship in the support of GOF-/LOF-/Unknown- gene discovery

Taking into account the so called ”high order” knowledge representation of *function* of targeted *gene* after *mutation* for a given *disease*, the relationship (*gene, mutation, function, disease*) contained three entities and one predicate, i.e.,*function=GOF/LOF/Unknown*. For semantic representation of the above relationship, all triple and multiple n-tuples were considered .

(i) Triple, an indirect way. As all known, the most popularized knowledge form in current knowledge graph circumstance was the triple form of (it entity, predicate, entity), as used in RESCAL algorithm. Henceforth, a natural analogue of knowledge representative is to set up a high order relationship via triples combinations. For instance, a pair of (*gene, function, disease*) and (*gene, function, mutation*) formed the ”high order” relationship: (*gene, mutation, function, disease*).

(ii) Four-tuple, a direct way. It is a natural extension of triple to design a quartic tuple directly: (*gene, mutation, function, disease*). Being a direct design of high order relationship, this quartic-tuple was then filled in a four way tensor.

(iii) Five-tuple, an enriched way. Taking into consideration of enriched information in the n-tuple relationships, semantic annotations obtained from AGAC corpus were added into the 5-tuple. Thus a (*gene, mutation, GOF/LOF/Unknown, disease, annotation*) tuple was considered in an enriched way.

2.2.2. Fundamentals of tensor and tensor decomposition for novel knowledge discovery

Fundamentals of tensors and CP decomposition were referred to references [12] [13]. Here, tensor is an extension of the vector outer product concept $X = ab^T := a \circledcirc b$. For a general tensor product of N vectors, one produces a Rank-one tensor $\mathcal{X}_0 \in \mathbb{R}^{I_1\times\cdots\times I_N}$: $\mathcal{X}_0 := a^{(1)} \circledcirc a^{(2)} \circledcirc \cdots \circledcirc a^{(N)}$ with $x_{i_1 i_2 \cdots i_N} = a^{(1)}_{i_1} a^{(2)}_{i_2} \cdots a^{(N)}_{i_N}$, where $i_1 = 1, 2, \cdots, I_1$; $\cdots$; $i_N = 1, 2, \cdots, I_N$. This is a straightforward definition, and so, a rank-one matrix can therefore be written as $\mathcal{X}_0 = a \circledcirc b$ and a rank-one 3-way tensor as $\mathcal{X}_0 = a \circledcirc b \circledcirc c$.

For a N-way tensor $\mathcal{X} \in \mathbb{R}^{I_1\times\cdots\times I_N}$

which conveyed relationship info, the purpose of tensor decomposition was to find a low rank approximate of the original tensor. By lowing the rank, both the sparsity issue of the original tensor was relieved and the novel links were discovered in the newly appeared nonzero cells. For instance, the classic CP decomposition was to compute a low rank approximate tensor $\hat{\mathcal{X}}$ which was a sum of Rank-one tensors: $\min\limits_{\hat{\mathcal{X}}} \|\mathcal{X} - \hat{\mathcal{X}}\|$ with $\hat{\mathcal{X}} = \sum_{r=1}^{R} \lambda_r a_r^{(1)} \circledcirc a_r^{(2)} \circledcirc \cdots \circledcirc a_r^{(N)} := [[\lambda, A^{(1)}, A^{(2)}, \cdots, A^{(N)}]]$. where $\lambda_r \in \mathbb{R}^R$, $A^{(n)} \in \mathbb{R}^{I_n\times R}$ for $n = 1, 2 \cdots N$, and $a_r^{(k)}$ was the r-th column vector of matrix $A^{(k)}$.

2.3. *Three way, four way and five way tensor designs in the support of novel link discovery.*

2.3.1. State-of-art RESCAL-based link discovery strategy for a three way tensor

As mentioned in the last section, we regard the *gene, mutation* and *disease* as entity, and *function=GOF/LOF/Unknown* as predicate. The triplet class is (Entity, Predicate, Entity), and an example

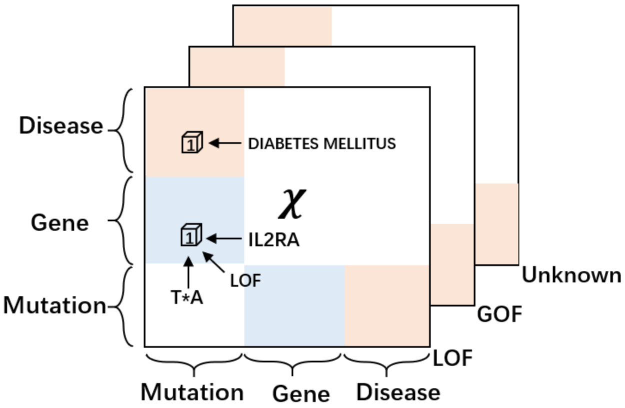

Figure 1. Structure of three way tensor used in RESCAL.

is (IL2RA, LOF, lymphedema).

Assuming there are G genes, M mutations, $F(=3)$ types of functions, and D kinds of diseases, a three-way tensor $\mathcal{X}^{(3)} \in \mathbb{R}^{n \times n \times 3}$ was defined, where $n(= G + M + D)$ is the amount of the entities. Here,

$$\mathcal{X}^{(3)}_{ijf} = \begin{cases} 1, \text{ if the 3-tuple } (entity_i, function_f, entity_j) \text{ existed} \\ 0, \text{ otherwize} \end{cases}, \tag{2.1}$$

The structure of the three way tensor met the requirement for the implementation of RESCAL. As shown in Figure 1, entities of *gene, mutation, disease* were located in the horizontal and vertical axes, while *function=GOF/LOF/Unknown* was located in the frontal axis. A tensor cell $\mathcal{X}^{(3)}_{ijf} = 1$ denoted the fact that there existed a relation (i-th entity, f-th predicate, j-th entity). More precisely, they employed the following rank-r factorization, where each slice $\mathcal{X}_f$ is factorized as $\mathcal{X}^{(3)}_f \approx \tilde{\mathcal{X}}^{(3)}_f := A\mathcal{R}_f A^T$, for $f = 1, 2, 3$. Here, $A = (A_1, \cdots, A_r)$ is a $n \times r$ matrix that contains the latent-component representation of the entities in the domain and R_f is an asymmetric $r \times r$ matrix that models the interactions of the latent components in the f-th predicate.

2.3.2. A four way GMFD-tensor structure

Though RESCAL was widely used in many relation discovery applications, it was not straightforward in the current case when there are more than three various kind of entities, and there are only three different predicates, i.e., *GOF, LOF, Unknown*. Thus, RESCAL was regarded as the first baseline method. Similarly, though the GMAFD-tensor was designed in a straightforward way, the high order structure made the tensor sparse and redundant. Alternatively, for reducing the size of tensor and killing the sparsity, only gene, mutation, function and disease links were considered in a four way tensor, i.e., Gene-Mutation-Function-Disease(GMFD) tensor $\mathcal{X}^{(4)} \in \mathbb{R}^{G \times M \times F \times D}$.

Taking concern that CP decomposition was computationally efficient upon a four way tensor, CPD for GMFD-tensor was assumed to be the best design for the current research.

Here

$$\mathcal{X}^{(4)}_{gmfd} = \begin{cases} 1, \text{ if the high order } (gene, mutation, function, disease) \text{ existed} \\ 0, \text{ otherwize} \end{cases}, \tag{2.2}$$

when the g-th gene played f-th function after the m-th mutation events based on the retrieved texts of d-th disease.

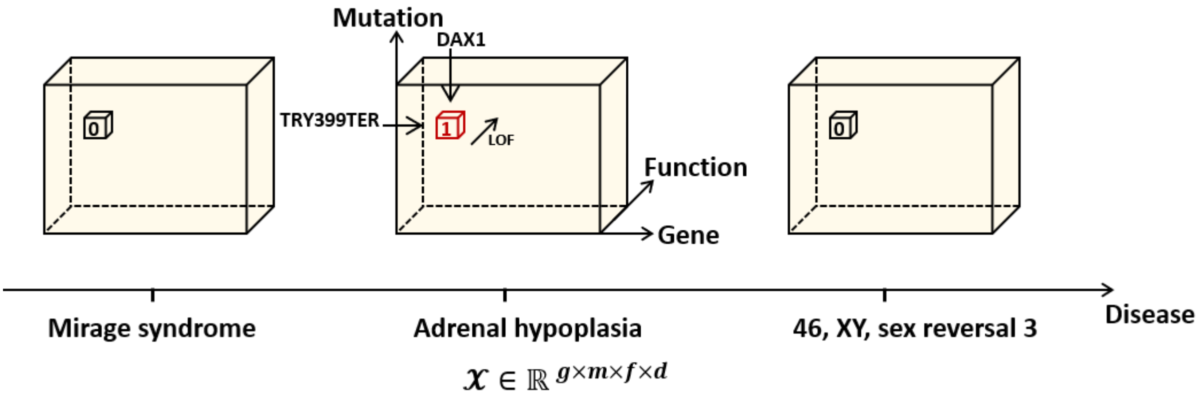

Figure 2. Structure of the GMFD-tensor.

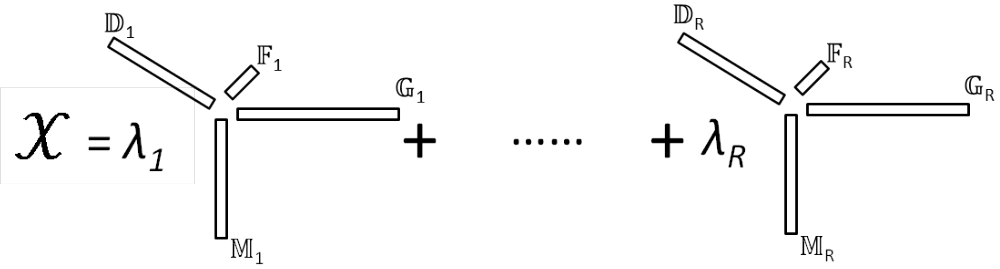

Figure 3. CP Decomposition of GMFD-tensor.

As shown in Figure 2, there is a mutation, TRY399TER, in DAX1 gene which leads to a LOF function in Adrenal hypoplasia, then the value of the cell in tensor is one.

For the knowledge inference, the CP decomposition is to compute the sum of series of rank-one tensors:

$$\mathcal{X}^{(4)} \approx \tilde{\mathcal{X}^{(4)}} = \sum_{r=1}^{R} \lambda_r \mathbb{G}_r \circledcirc \mathbb{M}_r \circledcirc \mathbb{F}_r \circledcirc \mathbb{D}_r, \tag{2.3}$$

with $\mathcal{X}_{gmfd} = \sum_{r=1}^{R} \lambda_r \mathbb{G}_{rg} \mathbb{M}_{rm} \mathbb{F}_{rf} \mathbb{D}_{rd}$, where $g = 1, 2, \cdots, G$, $m = 1, 2, \cdots, M$, $f = 1, 2, \cdots, F$, and $d = 1, 2, \cdots, D$. As shown in figure 3, a CP decomposition aims to find a low rank approximation of given tensor, and so as to achieve a reliable tensor completion. As an analogue of matrix completion for a given sparse matrix, the newly appeared non-zero valued cell in a given tensor correspondes the novel knowledge inference.

2.3.3. An enriched five way tensor structure

Assume there are G genes, M mutations, F types of functions, D kinds of diseases, and A kinds of annotations in AGAC corpus for the related texts, the data curated from OMIM were put into a five-way tensor, thus formed the Gene-Mutation-Annotation-Function-Disease (GMAFD) tensor $\mathcal{X}^{(5)} \in$

$\mathbb{R}^{G\times M\times A\times F\times D}$. Here,

$$\mathcal{X}^{(5)}_{gmafd} = \begin{cases} \#\{a - annotation\}, \text{ for a given tensor } \mathcal{X}^{(4)} \in \mathbb{R}^{G\times M\times F\times D} \\ 0, \text{ otherwize} \end{cases}, \tag{2.4}$$

if the g-th gene played f-th function after the m-th mutation events based on the retrieved texts of d-th disease, and there are #{a-annotation} anotation for each a-th annotation labels. The structure of the four way tensor is shown in figure 4.

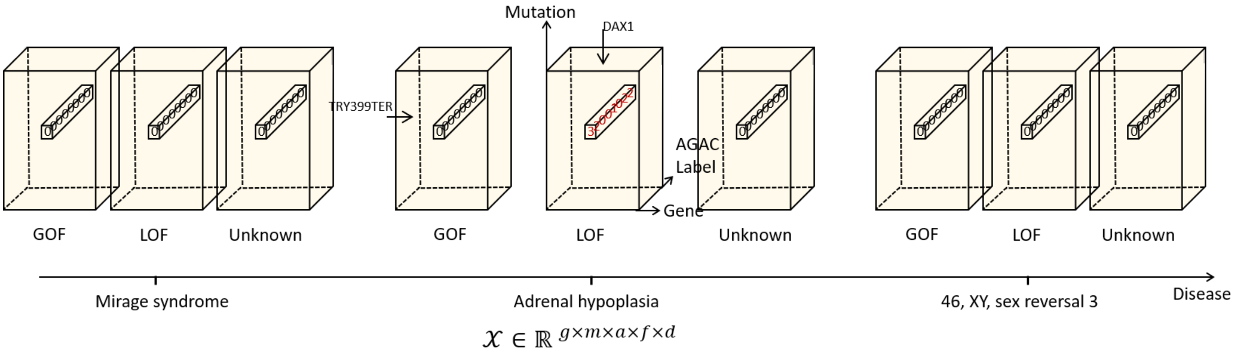

Figure 4. Strucutre of the GMAFD-tensor.

The CP decomposition for the 5-dimension tensor
$\mathcal{X}^{(5)}$ is

$$\mathcal{X}^{(5)} \approx \tilde{\mathcal{X}^{(5)}} = \sum_{r=1}^{R} \lambda_r \mathbb{G}_r \circledcirc \mathbb{M}_r \circledcirc \mathbb{A}_r \circledcirc \mathbb{F}_r \circledcirc \mathbb{D}_r, \tag{2.5}$$

2.4. Evaluation criteria for novel link discovery via tensors

The idea of tensor decomposition here was to compute an approximation tensor $\tilde{\mathcal{X}}$ of original tensor $\mathcal{X}$ and minimize the difference. The evaluation metris used in this research were designed and listed as below:

(i) Recall, precision and F-score. These was traditional metrics for evaluating the reproducibility of n-tuple knowledge retrieval from the approximated tensor. For example, a cell $\mathcal{X}_{gfd} = 1$ in a three way tensor represented a relationship of $(gene_g, function_f, disease_d)$, and a cell $\tilde{\mathcal{X}}_{gfd}$ in the approximate tensor inferred a recalled entry, thus recall rate, precision and F-score were computed

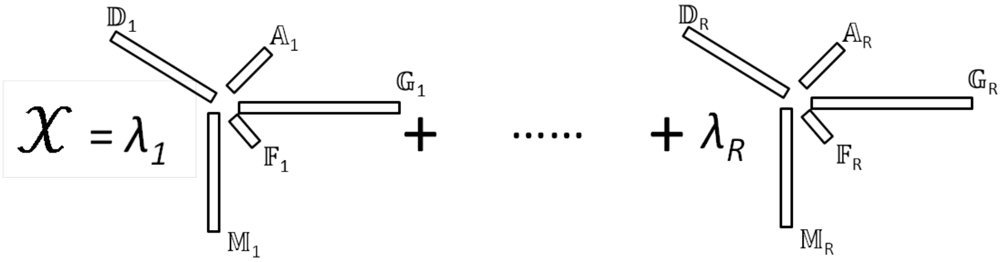

Figure 5. CP Decomposition of GMAFD-tensor.

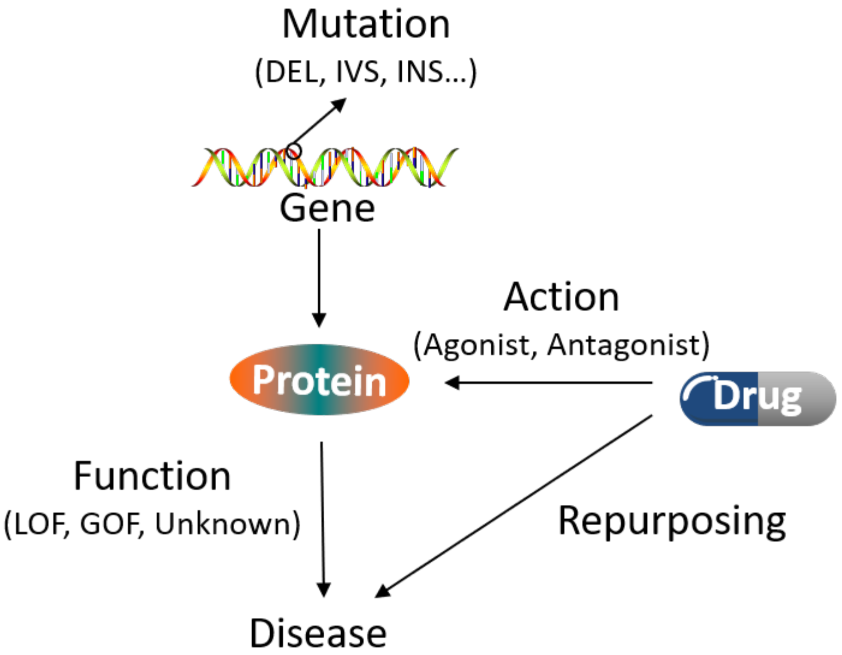

Figure 6. The relationships between tensor dimensions.

in a classic way, i.e., $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$ and $F - score = \frac{2Precision*Recall}{Precision+Recall} = \frac{2TP}{2TP+FP+FN}$.

(ii) Approximation evaluation (AE). In this case, the accuracy of the tensor approximation was computed via $AE = 1 - \frac{\|\tilde{X}-X\|_F^2}{\|X\|_F^2}$.

(iii) Mask recall rate (MRR). In this evaluation, 20% cells with nonzero values in $\mathcal{X}$ were masked, and then an approximated tensor was computed. The corresponding 20% cells in $\tilde{\mathcal{X}}$ were observed to be nonzero or not, and the recall rate was computed as MR. The greater the MR achieves, the higher the chance that the existed ”high order” relationships reproduced by the new tensor.

(iv) Jumping rate (JR). In the circumstance of novel link discovery in our research, the percentage of new link over existed link was denoted as JR. Taking an example from the three way tensor, for given function and disease, $function_{given}$, $disease_{given}$, a novel gene appeared in the triple $(gene_{novel}, function_{fixed}, disease_{fixed})$, and the percentage of jumping rate was computed by $JR = \frac{\#\{(gene_{novel}, function_{fixed}, disease_{fixed})\}}{\#\{(gene_{all}, function_{fixed}, disease_{fixed})\}}$. The higher the value, the more capable the new tensor produces novel link.

2.5. High order novel link discovery for gene, mutation and disease in support of drug repurposing

A tensor decomposition based high order novel link discovery is carried on to infer new GOF and LOF to mutated genes.

The strategy of high order novel link discovery is represented in figure 6, while the algorithm is introduced as below:

[GOF/LOF/Unknown knowledge inference and drug repurposing]

Step 1 Collect OMIM data and curate high order relationship *(gene, mutation, function, disease)* manually;

Step 2 Fill-in the cells in three-way, GMFD- or GMAFD-tensor with the known high order relationships;

Step 3 Factor these tensors with CP or RESCAL decomposition algorithm and obtain low-rank approximation tensors;

Step 4 Extract nonzero cells with values greater than a preset threshold, and a novel link of *(gene, mutation, function, disease)* is found;

Step 5 Match an antagonist(/agonist) drug with its target gene with LOF(/GOF) function. After adding the drug information in the high order relationship, a new correlation for drug-disease pair is discovered.

3. Result

As suggested in the method section, CP tensor decomposition for a four way tensor was regarded as a good trade-off for both tensor size and data sparsity. In this section, CPD for GMFD tensor decomposition was compared with baseline method.

3.1. Performance of GOF/LOF/Unknown prediction based on the AGAC corpus annotations

Three topic classifiers were designed and the purpose was to test the effectiveness of the used features. Here, BiLSTM was regarded an efficient neural network which encoded context information accurately and thus was used for GOF/LOF/Unknown classification. For an additional comparison, a variant BiLSTM was designed by integrating AGAC labels as hidden layers input.

As shown in Table 2, the performance of BiLSTM achieved 0.387, 0.349, 0.317 in precision, recall and F-score, respectively, while BiLSTM-tags increased these metrics to 0.571, 0.534 and 0.546. This result showed that to utilize AGAC annotations labels enhanced the performance of topic classifier.

Besides, AGAC labels were used as the only feature in a traditional classifier based on Support Vector Machine (SVM), and a dramatic performance improvement were obtained, as shown in Table 2. The F-score was updated to 0.841, and the result fully showed that the use of AGAC trigger words and labels were key factors for text classification in GOF/LOF/Unknown recognition.

All of these results were calculated under the leave-one-out cross-validation (LOOCV), and the Macro-F employed as the metric in a three-class classification model.

Table 2. Performance of GOF/LOF/Unknown text classifiers with or without AGAC features in a LOOCV evaluation.

Classifier	Features	**Precision**	**Recall**	**Macro F-score**
BiLSTM	Lexical features	0.387	0.349	0.317
BiLSTM-tags	Lexical and AGAC features	0.571	0.534	0.546
SVM	AGAC features	0.832	0.851	0.841

The result showed that the annotations offered by AGAC corpus provided meaningful enriched information for GOF/LOF gene prediction.

3.2. *Comparison of three way, four way, or five way tensor decomposition in terms of RESCAL or CP decomposition*

Table 3. Comparison of three tensor decomposition methods.

	Precision	Recall	F-value	AE	MRR(%)	JR(%)
RESCAL for three way tensor	0.299	0.998	0.460	0.972	3.8	0
CPD for GMFD-tensor	0.508	0.298	0.376	0.436	5.1	61.5
CPD for GMAFD-tensor	0.0˜	0.0˜	0.0˜	0.043	0.0	99.4

A thorough comparison among all n-way tensor decomposition was carried on so as to better quantify the effectiveness. Generally, metrics including precision, recall, F-value and AE were used to measure the accuracy of the reconstructed tensor vs. the original tensor. Meanwhile, MRR was to evaluate the ability of recover masked cells. Furthermore, JR was a metric to represent an ability for algorithm to produce novel high order relationships. In the experiment, the cells with value being lower than threshold value $\theta = 0.01$ were removed. Furthermore, the remained cells were considered to represent reliable links among gene, mutation, disease and function. The full results were listed in table 3.

After the implementation of CPD for GMFD-tensor, 691 high order relations, i.e., (Gene, mutation, function, disease), were obtained. Among them, 351 out of 691 ones were from the original tensor. The rest 340 cells were new links to address novel high order relations. In this case, $n = 1178$ referred to the amount of entities. $TP = 351$ denoted the True Positive. Thus, the precision, recall and F-value obtained were 0.508, 0.298 and 0.376, respectively. In addition, 209 out of 340 new cells contained newly-linked gene-disease pairs. Thus the jumping rate metric equaled to 61.5%, and this value represented the ability of novel link prediction of CPD for GMFD-tensor. Furthermore, the knowledge recovery rate was evaluated by a masking strategy. Here, we randomly masked 234 cells in the original tensor $\mathcal{X}$, and then observe the percentage of the reproduced nonzero cells in the new tensor $\tilde{\mathcal{X}}$. After computing the approximated tensor, 12 out of 234 cells being observed, and got Mask recall rate 5.1%.

In CPD for GMAFD-tensor, the precision, recall and F-value were all extremely close to 0, thus, AE = 0.043 was much less than it in CPD for GMFD-tensor and MRR was 0 while JR was 99.4% which indicated that CPD for GMAFD-tensor was able to produce a good body of new links but the results were not reliable.

As to RESCAL for three way tensor, except precision (=0.299) was lower, recall (=0.998), F-value (=0.460), AE (=0.972) were all better than CPD for GMFD-tensor. However, MRR (=3.8%) and JR (=0) hinted that this method was not strong enough for credible novel links prediction issue.

Results in the above comparison evidenced that CPD for GMFD-tensor achieved the best trade-off between the F-score-related metrics and JR-related metrics. A sufficiently high F-score ensured a high percentage of knowledge reproducibility of high order link, and a sufficiently high JR value led to a promising novel knowledge discovery. Taking granted the GOF and LOF information was effective integrated in the novel link discovery, the novel high order relations were more reliable. This trade-off is of utmost importance in the case of drug repurposing, when novel and reliable gene-disease pairs were highly expected. Hence, CPD for GMFD-tensor was regarded as the best model for equipping both reliability and predictive ability.

3.3. Case study of gene IL2RA and drug Aldesleukin: an application of tensor decomposition to drug repurposing

Table 4. Input data of IL2RA collected from OMIM.

Gene	Mutation	Disease	Function
IL2RA	DELETION	Immunodeficiency	LOF
IL2RA	C→A	Diabetes Mellitus	LOF
IL2RA	T→A	Diabetes Mellitus	LOF
IL2RA	INSERT	Immunodeficiency	LOF
IL2RA	C→T	Immunodeficiency	LOF
IL2RA	SER→ASN	Immunodeficiency	LOF
IL2RA	TYR→SER	Immunodeficiency	LOF

A case study is presented in this section to show cells in a decomposed tensor which indicate new links between IL2RA and other mutation-disease-function entities.

The data collected from OMIM consisted of 1178 nonzero cells, which corresponded to various entities including 197 genes, 198 mutations, and 307 diseases. Among them, IL2RA is a human gene which encodes interleukin-2 receptor alpha chain (also called CD25). IL2RA is widely expressed in various immune cells, including B-cell neoplasms, acute nonlymphocytic leukemias, neuroblastomas, mastocytosis and tumor infiltrating lymphocytes. It functions as the receptor for HTLV-1 and is consequently expressed on neoplastic cells in adult T cell lymphoma/leukemia.

Seven IL2RA related records was included, as shown in Table 4. For instance, a Cytosine to Adenine base pair change led to a LOF in IL2RA, and this caused diabetes mellitus, while a DELETION in IL2RA led to another LOF and it caused Immunodeficiency.

After applying CP decomposition in the GMFD-tensor, a tensor completion was performed with a low-rank tensor approximation. In the decomposed tensor, the newly added nonzero cells revealed novel links among entities of corresponding axis.

As shown in table 5, four of the cells with the highest scores, both of which were higher than 0.14, indicated that LOF mutation in IL2RA related to lymphedema and myelodysplastic syndrome. These predictions just matched to the effects of IL2RA. Then, the drug Aldesleukin targeted to IL2RA was found in Drugbank, and it act as an agonist. Aldesleukin was recorded in Drugbank that had the efficiency to lymphoma and myelodysplastic. The logic of this case study is shown in Figure 7.

This case study proved that predictions by tensor decomposition were credible, and it was feasible to be applied in drug repurposing.

In addition, there is a common finding in CP decomposition predictions. The predicted diseases of a gene and the recorded diseases of it in OMIM usually belonged to the same category. For instance, the diseases of MYL3 gene were recorded as cardiomyopathy in OMIM, and it's the data that was put into tensor. After tensor decomposition, LOF mutation in MYL3 was predicted to related to lain distal my-

Table 5. Newly output data in decomposed GMFD-tensor revealing novel links among IL2RA, mutation, function, and disease.

Score	Gene	Mutation	Disease	Function
0.1952	IL2RA	DELETION	LYMPHEDEMA	LOF
0.1839	IL2RA	INSERT	LYMPHEDEMA	LOF
0.1569	IL2RA	DELETION	MYELODYSPLASTIC SYNDROME	LOF
0.1478	IL2RA	INSERT	MYELODYSPLASTIC SYNDROME	LOF
0.0930	IL2RA	ARG→TRP	LYMPHEDEMA	LOF
0.0930	IL2RA	IVS	LYMPHEDEMA	LOF
0.0930	IL2RA	PRO→LEU	LYMPHEDEMA	LOF
0.0747	IL2RA	PRO→LEU	MYELODYSPLASTIC SYNDROME	LOF
0.0747	IL2RA	ARG→TRP	MYELODYSPLASTIC SYNDROME	LOF
0.0747	IL2RA	IVS	MYELODYSPLASTIC SYNDROME	LOF
0.0711	IL2RA	C→T	LYMPHEDEMA	LOF
0.0711	IL2RA	TYR→SER	LYMPHEDEMA	LOF
0.0711	IL2RA	SER→ASN	LYMPHEDEMA	LOF
0.0571	IL2RA	TYR→SER	MYELODYSPLASTIC SYNDROME	LOF
0.0571	IL2RA	C→T	MYELODYSPLASTIC SYNDROME	LOF
0.0571	IL2RA	SER→ASN	MYELODYSPLASTIC SYNDROME	LOF
0.0319	IL2RA	DELETION	LYMPHEDEMA	unknown
0.0300	IL2RA	INSERT	LYMPHEDEMA	unknown
0.0256	IL2RA	DELETION	MYELODYSPLASTIC SYNDROME	unknown
0.0241	IL2RA	INSERT	MYELODYSPLASTIC SYNDROME	unknown
0.0197	IL2RA	ARG→LEU	LYMPHEDEMA	LOF
0.0197	IL2RA	ARG→TER	LYMPHEDEMA	LOF
0.0198	IL2RA	CYS→ARG	LYMPHEDEMA	LOF
0.0159	IL2RA	ARG→TER	MYELODYSPLASTIC SYNDROME	LOF
0.0159	IL2RA	CYS→ARG	MYELODYSPLASTIC SYNDROME	LOF
0.0159	IL2RA	ARG→LEU	MYELODYSPLASTIC SYNDROME	LOF
0.0151	IL2RA	ARG→TRP	LYMPHEDEMA	unknown
0.0151	IL2RA	IVS	LYMPHEDEMA	unknown
0.0151	IL2RA	PRO→LEU	LYMPHEDEMA	unknown
0.0122	IL2RA	PRO→LEU	MYELODYSPLASTIC SYNDROME	unknown
0.01221	IL2RA	ARG→TRP	MYELODYSPLASTIC SYNDROME	unknown
0.0122	IL2RA	IVS	MYELODYSPLASTIC SYNDROME	unknown
0.0116	IL2RA	C→T	LYMPHEDEMA	unknown
0.0116	IL2RA	TYR→SER	LYMPHEDEMA	unknown
0.0116	IL2RA	SER→ASN	LYMPHEDEMA	unknown

opathy, left ventricular noncompaction, and myopathy. Both of them were myopathy diseases or heart disease. This finding showed that the predictions by tensor decomposition fellow some meaningful biological regularities.

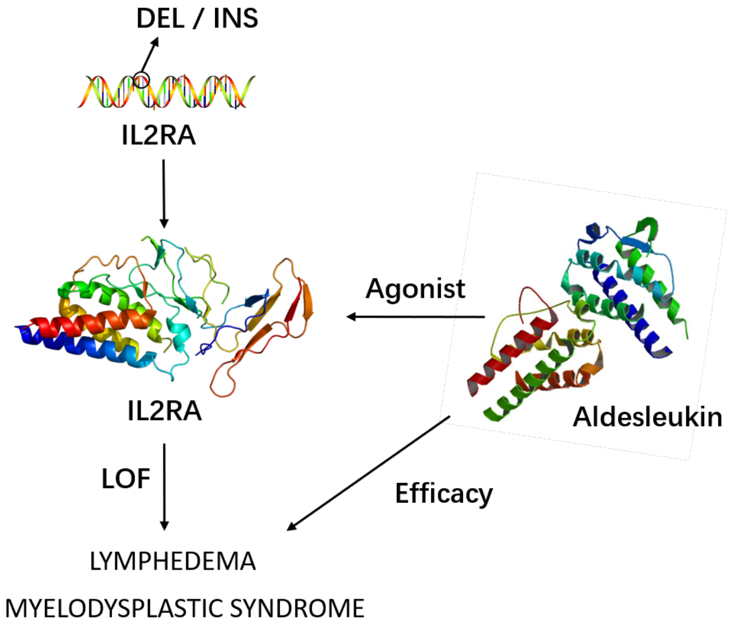

Figure 7. An case study example for the novel drug indication discovery by using novel link discovery in tensor decomposition.

4. Conclusion

The undergoing research of tensor decomposition unveiled potential application of novel knowledge inference for drug repurposing.

As shown in Figure 7 of the case study research, genes with diseases were associated by the function of mutation, which were LOF, GOF and Unknown, and this association was predicted by tensor decomposition. The other association between drugs and genes were the actions of drugs, which were agonist and antagonist, and the information of this part were searched from a drug database Drugbank. Then, by using genes as connections, drugs and its new indications were inferred, and drug repurposing was accomplished.

In this research, multiple way data representation for drug-related items gene, mutation, function and disease were achieved by using abundant tensor structure. The cell valuing strategies for both GMFD- and GMAFD- tensors are straightforward. Furthermore, in both cases, AGAC corpus was fully used. For GMFD-tensor, AGAC corpus was utilized as a training data to predict the text describing given gene and mutation carry LOF or GOF semantics. Meanwhile, as an extensive version, GMAFD-tensor incorporated trigger labels information from AGAC, and raised the order of the tensor.

Regarding the instructive case study on novel link discovery for ”(Lymphoma, Myelodysplatic Syndromes)-IL2RA-Aldesleukin”, as well as the poor performance of cp decomposition for GMAFD-tensor, the research results fully showed that it is a promising trend to design more sophisticated tensor decomposition methods in the future, which suits the data structure of multi-Omics data and be potential and effective computational way for drug repurposing.

Acknowledgement

The authors would like to express high gratitude for CHIP 2018 reviewers who brought nice comments and suggestions to the early version of this work. This work is funded by the Fundamental Research Funds for the Central Universities of China (Project No. 2662018PY096).

Conflict of interests

The authors declare no conflict interests.

References

1. M. Simsek, B. Meijer, A. A. van Bodegraven, N. K de Boer, and J. J. M. Chris, Finding hidden treasures in old drugs: the challenges and importance of licensing generics, *Drug Discov. Today*, **23**(2018), 17–21.
2. N C. Baker, S. Ekins, A. J. Williams and A. Tropsha, A bibliometric review of drug repurposing, *Drug Discov. Today*, (2018).
3. F. L. Hitchcock, The expression of a tensor or a polyadic as a sum of products, *J. Math. Phys.*, **6** (1927), 164–189.
4. N. Maximilian, V. Tresp and H. P. Kriegel, A three-Way model for collective learning on multi-relational data, *ICML*, **11** (2011), 809–816.
5. N. Madhav, M. Gupta and P. Talukdar, Higher-order relation schema induction using tensor factorization with back-off and aggregation, *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, **1** (2018), 1575–1584.
6. N. Madhav, U. S. Saini and P. Talukdar, Relation schema induction using tensor factorization with side information, *arXiv preprint*, arXiv:1605.04227 (2016).
7. L. Timoth*é*e, N. Usunier and G. Obozinski, Canonical tensor decomposition for knowledge base completion, *arXiv preprint*, arXiv:1806.07297 (2018).
8. J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin and J Sun, Limestone: High-throughput candidate phenotype generation via tensor factorization, *J. Biomed. Inform.*, **52** (2014), 199–211.
9. J. Fang and W. Jonathan, Tightly integrated genomic and epigenomic data mining using tensor decomposition, *Bioinformatics*, (2018), 1–7.
10. Y. H. Taguchi, Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets, *Sci. Rep.*, **7** (2017), 13733.
11. Y. Wang, X. Yao, K. Zhou, X. Qin, J. D. Kim, K. B Cohen and J. Xia, Guideline design of an active gene annotation corpus for the purpose of drug repurposing, *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics(CISP-BMEI 2018)*, Oct, 2018, Beijing. (Accepted)

12. T. G. Kolda and W. B. Bader, Tensor decompositions and applications. *SIAM Rev.*, **51** (2009), 455–500.

13. R. Stephan, O. Shchur and S. Günnemann, Introduction to tensor decompositions and their applications in machine learning, *arXiv preprint*, **1711**(2017),10781.

14. L. Hao, S. Liang, J. Ye and Z. Xu, TensorD: A tensor decomposition library in TensorFlow, *Neurocomputing*, **318**(2018), 196–200.

AIMS Press

© 2019 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)