ISSN:2320-0790

An International Journal of Advanced Computer Technology

Discovering and Mining Links for Protein Databases

A.Immaculate Mercy, S.Padmavathi

Dept. of Computer Applications,MPC, Thanjavur

Dept. of Computer Science and Information Technology, MPC, Thanjavur, India

Abstract: This work introduces a link analysis procedure for discovering relationships in a protein database or a relational database generalizing simple correspondence analysis. It is based on extracting the links to the related protein database and malfunctioned protein database. The datasets are trained in order to find out missing interactions and the sequences related to them. Further the analysis of links proceeds by performing a random walk defining a Markov chain. The elements of interest are analysed through stochastic complementation which gives a reduced Markov chain. This reduced map is then analysed by projecting the elements of interest through Principal component analysis. Several Protein datasets are analysed using the proposed methodology, showing the usefulness of the technique for extracting relationships in relational databases or graphs.

Keywords: Link analysis, Markov chain, correspondence analysis, graph mining, principal component analysis;

I. INTRODUCTION

The process of analysing links for a protein database is a novel way of finding which protein or the protein entry has contributed towards malfunctioning. The proteins are comprised of several amino acids. These amino acids dictate the ways of pairing and interaction among different proteins. The purpose of this work is to discover and analyse the links based on the similarity or dissimilarity. Traditional methods of analysing links in data mining approaches usually assume a random sample of independent objects from a single relation which always results in the extraction of knowledge from data almost always leading to a double entry format containing the features for the sample of a population. All these research fields intend to find and exploit the links between objects which could be of various types and involved in different types of relationships. The focus of the technique has moved from over the analysis of features to the analysis of links existing between the instances in addition to the features.

This paper proposes a technique allowing discovering the relationships existing between the elements of a protein database and a malfunctioned database. For this a two-step procedure is developed. First a much smaller, reduced Markov chain only containing the elements of interest is extracted using **stochastic complementation** [13]. Then the reduced chain is analysed by projecting into the Kernel view of the graph. The element of interest for the study has been restricted to study the sequences of protein and malfunctioned protein for the species Homo sapiens.

The motivations for developing this two-step procedure are twofold. First, the computation would be cumbersome when dealing with the complete database. Second, in many cases the analyst is not interested in studying all the relationships between all the elements of the database, but only a subset of them. Therefore reducing the Markov chain by stochastic complementation allows to focus the analysis on the elements of interest and the relationships we are interested in. When dealing with a bipartite graph stochastic complementation followed by a basic diffusion graph [10] is exactly equivalent to simple correspondence analysis. On the other hand, when dealing with a star schema this two-step procedure reduces to multiple correspondence analysis.

The simple correspondence analysis aims to study the relationships between two random variables x1 and x2, having mutually exclusive, categorical outcomes denoted as attributes. An experimenter makes a series of measurements of the features x_1, x_2 on a sample of vg individuals. In a relational database this corresponds to two tables, each table corresponding to one variable and containing the set of observed attributes of the variable.

This could be modelled as a bipartite graph, where each node corresponds to an attribute and links are only defined between attributes of x_1 and x_2. The weight associated with each node corresponds to the attribute of x_1 and x_2.

The corresponding graph is built by defining one node for each individual and for each attribute while a link between an individual and an attribute is defined when the individual possess this attribute.

The experimental procedure aims to provide answers for the following questions

1. How the graph mapping provided by PCA does compares with the map projection?
2. Does the stochastic complementation provide realistic subgraph drawings?
3. How does the diffusion map and stochastic complementation compare to the dimensionality reduction technique?
4. Does the stochastic complementation accurately preserve the structural information?

The proposed methodology therefore extends the Correspondence analysis to the analysis of Protein databases. This paper has four main contributions:

1. A two-step procedure for analysing Protein databases is proposed.
2. Suggested procedure extends correspondence analysis.
3. A link feature between the Protein and their malfunctioned databases based on their dissimilarity.
4. Finally the link gets the associated malfunctioned protein structure and their related diseases for Homo sapiens.

The database is considered as a graph where the nodes correspond to the elements contained in the tables and the links correspond to the relation between the tables[14]. The databases considered for the work is the Protein database and their corresponding malfunctioned databases. A link analysis between the protein and the malfunctioned databases are found through the means of stochastic complementation where only the proteins of interest are taken into account. Since the sequences are large varying from one protein entry to the other a reduced Markov chain is obtained which works with the sequence of interest. Each and every sequence has varied length and also varies from species to species.

II. ANALYZING RELATIONS THROUGH STOCHASTIC COMPLEMENTATION

The database for the work is converted into a graph. This graph is assumed that it does not contain any self - loops. If the graph is not connected then there is no relationship at all between the different components and the analysis has to be performed separately on each of them. Partitioning a graph into connected components is from its adjacency matrix. Based on the adjacency matrix the Laplacian matrix is defined. From this graph a natural random walk through the graph is performed by the way of associating a state to each node.

2.1 Markov Chain:

The database for the work is the original protein database and malfunctioned database. First the self- interacting proteins are removed and then they are converted into a graph. The nodes of the graph indicate the names of the proteins.

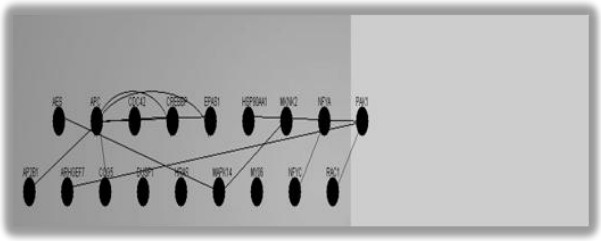

Fig :1 Conversion of database into graph based on the interaction set.

The graph has a set of states S= {S_1, S_2, S_3…S_r}.The process starts in one of these states and moves successively from one state to another. Each node is called a step. If the chain is currently in state S_i, then it moves to state S_j at the next step with a probability denoted by p_{ij}. This probability does not depend upon which states the chain was in before the current.

The probabilities p_{ij} are called transition probabilities. The process can remain in the state it is in and this occurs with probability p_{ii}. An initial probability distribution defined on S, specifies the starting state. A markov chain starts in a state chosen by the probability distribution on the set of states known as probability vector.

The first thing that needs to be done in the construction of a reduced Markov chain is that partitioning the graph into connected components from its adjacency matrix.

2.1.1 Adjacency Matrix: The adjacency matrix provides a means of representing which node of a graph is adjacent to which other vertices. In this work the adjacency matrix is built based on the interaction set. If the specified protein does not have an adjacent node (i.e. Interaction) it is considered to have its value as 0 else as 1.

	AES	AP2B1	APC	ARHGEF7
AES	0	0	0	0
AP2B1	0	0	1	0
APC	0	1	0	0
ARHGEF7	0	0	0	0

Table: 1 Adjacency matrix.

2.1.2 Laplacian matrix: Laplacian Matrix is a matrix representation of a graph which can be used to calculate the number of spanning trees for a given graph .Given a simple graphs with "n" vertices its Laplacian matrix

$L:=(l_{ij})^{nxn}$ is defined as

$$L=D-A$$

Where "D" is the degree of the matrix and "A" is the adjacency matrix. The random walk normalized Laplacian is defined as

$$T: =D^{-1} A.$$

D^{-1} is the diagonal Matrix and "T" is the transition matrix of a standard random walk of the given graph.

	AES	AP2B1	APC	ARHGEF7
AES	1	-1	-1	-1
AP2B1	-1	1	-1	-1
APC	-1	-1	4	0
ARHGEF7	0	0	0	1

Table: 2 Laplacian Matrix

The Laplacian matrix [1],[4] that is obtained is partitioned so that we arrive at two new states S_1 and S_2. That is during any random walk on the original chain only the states belonging to S_1 are recorded i.e. according to this work only the proteins which have proper interactions and which in turn has proper adjacency is maintained. The other proteins are recorded in state S_2.

Thus we arrive at a new graph which is reduced but still further reduction is required to attain an irreducible graph. From the graph that is obtained a natural random walk of the chain is performed by the way of associating a state to each node. Each element represents a state of the Markov chain describing the sequence of states that was visited. A random variable s(t) contains the current state of the Markov chain at time step t: if the random walker is in state i at time t, then s(t) =i.

The random walk is defined by following single-step transition probabilities of jumping from any state i to an adjacent state j=s(t+1). The transition probabilities only depend on the current state and not on the past ones. Since the graph is completely connected, the Markov chain is irreducible, that is every state can be reached from any other state. This graph provides the state probability distribution

$X(t)=[x_1(t),x_2(t),\ldots x_n(t)]$ at time T being the matrix transpose . The probability distribution at each state at time t when starting from state I at time t=0.

Since the Markov chain represents a random walk on graph G, the transition matrix is simply $P=D^{-1}A$.If the adjacency matrix A is symmetric , the Markov chain is reversible and the steady state vector,π, is simply proportional to the degree of each state d. All the Eigen values of the transition matrix are real.

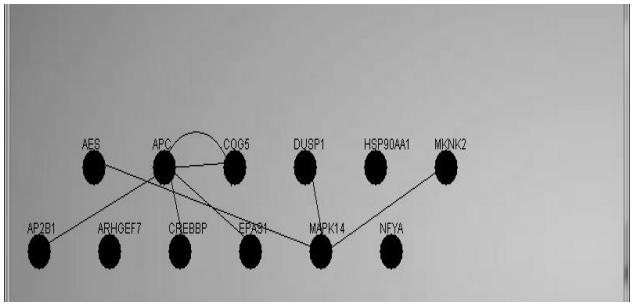

Fig. 2: Reduced graph

2.2 Computing a reduced Markov chain by stochastic complementation

A reduced Markov chain can be computed from the original chain. First, the set of states is partitioned into two subsets, S_1- corresponding to the nodes of interest to be analysed and S_2- Corresponding to the remaining nodes to be hidden. The number of states in S_1 and S_2 are denoted by n_1 and n_2. The number of states in S_1 and S_2 specified as $n_2>>n_1$. Thus the transition matrix is repartitioned into

$$P= \begin{array}{c} \\ S_1 \\ S_2 \end{array} \begin{array}{c} \begin{array}{cc} S_1 & S_2 \end{array} \\ \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \end{array}$$

The main idea is to censor the useless elements by masking them during the random walk. During any random walk on the original chain only the states belonging to the S_1 are recorded. In the perspective of the work taken into study the Proteins of proper interactions alone are taken into the state S_1 and the remaining proteins which contribute towards malfunction are to be hidden which is kept in the second state S_2. The resulting Markov chain that is obtained is the stochastic complement of the original chain. Thus, performing a stochastic complementation allows focusing on the elements representing the features of interest. The reduced chain inherits all the characteristics from the original chain; it simply censors the malfunctioned proteins. After obtaining the stochastic chain the chain is again partitioned as

$$P_c=P_{11}+P_{12}\,(I-P_{22})^{-1}\,P_{21}$$

	AES	AP2B1	APC	ARHGEF7
AES	0.0	0.0	0.0	0.0
AP2B1	0.0	0.0	1.0	0.0
APC	0.0	0.25	0.0	0.0
ARHGEF7	0.0	0.0	0.0	0.0

Table: 3 Transition matrix

It is observed that the matrix is stochastic, that is the sum of elements of each row is equal to 1. This corresponds to a valid transition matrix. If the initial chain is periodic, the reduced becomes aperiodic by stochastic complementation.

III. ANALYZING THE REDUCED MARKOV CHAIN

Once a reduced Markov chain containing only the nodes of interest has been obtained a low dimensional space of the graph is achieved which shows the proximity between the nodes.

3.1 Simple Correspondence Analysis: The simple correspondence analysis[6] aims to study the relationships between two proteins which are taken in a random manner namely x1 and x2. The features are measured based on the sequence pattern of the proteins in comparison with the interaction set sequence. The frequencies of the links are recorded. This is modelled as a bipartite graph, where each node corresponds to an attribute and links are only defined between attributes of one protein and the other protein.

The strength of the links is once again evaluated based on the Eigen values and the Eigen vectors.

0.1524	**-0.1244**	**0.1549**	**0.1098**
0.1524	**-0.1244**	**0.1549**	**0.1098**
0.0772	**0.1577**	**0.024**	**0.2796**
0.1524	**-0.1167**	**0.0588**	**0.1098**

Table: 4 Eigen vector

3.2 *Principal Component Analysis:* This presents a way of identifying patterns in data and expressing the data in such a way to highlight their similarities and differences. The dimensions could be reduced [11].

For a protein database the sequence for every protein is different and finding the difference and similarity for every properly interacted data set and malfunctioned set plays a major role. In order to reduce the dimensions on the interaction set the mean of the sequence is subtracted from the data dimension. All the x values have x' subtracted and all the y values have y' subtracted from them. This produces the data set whose mean is zero. The Eigen vectors and Eigen values are calculated for the covariance matrix that is obtained. The Eigen vector with the highest Eigen value is the principal component of the data set [1],[8].

After the Eigen vectors are found from the covariance matrix, the next step is to order them by Eigen value from highest to lowest. This gives the components in order of significance once the components that are chosen, the transpose of a vector is taken and it is multiplied with the left of the original data set which is transposed.

State1
AES
AP2B1
APC
ARHGEF7
COG5
CREBBP
DUSP1
EPAS1
HSP90AA1
MAPK14
MKNK2
NFYA
NFYC
PAK1
RAC1

Table: 6 Partitioned matrix

3.3 Kernel View of the Map: A variant of the basic diffusion model is introduced. It has been assumed that initial adjacency matrix is symmetric. This extension presents several advantages in comparison with the original basic diffusion map.

1. The kernel version of the diffusion [9] is applicable to directed graphs while the original model is restricted to undirected graphs.
2. The extended model induces a valid kernel on the graph.
3. The resulting matrix is symmetric positive. The spectral decomposition can thus be computed on a symmetric positive definite
4. The resulting matrix represents the Euclidean space.

	AES	***AP2B1***	***APC***	***ARHGEF7***
AES	0	0	0	0
AP2B1	0	0	1	0
APC	0	0.2	0	0
ARHGEF7	0	0	0	0

Table : 5 Repartitioned Matrix

This kernel technique will be referred to as the diffusion map or the KDM PCA. The matrix KDM is the natural kernel associated to the squared diffusion map distances. It is observed that the matrix is symmetric positive semi definite and contains inner products in a Euclidean space. This Euclidean space brings out the real interaction sets which have the proper sequence of the proteins and the other protein that does not contribute to the proper functioning are eliminated.

Performing principal component analysis aims to change the coordinate system by adding new axes in the direction of maximal variances. This method suffices to compute the first Eigen values of KDM and to consider that these Eigen vectors are multiplied by the square root of the corresponding Eigen values. These values are the

coordinates of the nodes in the principal component space spanned by the Eigen vectors.

It can be shown that the initial graph is undirected; this PCA based on the kernel matrix KDM is similar to the diffusion map. The resulting Kernel matrix can be centered by HKDMH with H=I-(eeT/n), where e is a column vector all of whose elements are "1",H is called the centering matrix .

AES
AP2B1
APC
ARHGEF7
COG5
CREBBP
DUSP1
EPAS1
HSP90AA1
MAPK14
MKNK2
NFYA

Table: 7 Final states

Based on the final states a diffused graph is obtained which gives us the projection of proteins which contribute towards the proper functioning of the protein sets [2],[3].

The transformation from the original graph to the final graph gives us an idea of dimensionality reduction when working large and sparse data sets. The idea behind the reduction is that results could be easily drawn where the proteins which have true interactions alone are taken into consideration whereas the rest of the interactions results in malfunction. It could be easily deduced from the fact that based upon the dissimilarity we can find out the malfunctioned protein sets.

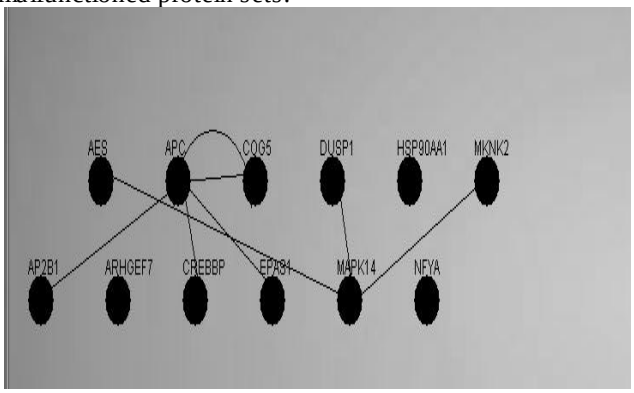

Fig. 3: Final graph

IV. DISCUSSION OF THE RESULTS

The two step procedure provides an embedding in a low dimensional subspace from which useful information can be extracted. The results shows that highly related proteins are close to each other based on their links (sequence) based on their interaction. The stochastic complementation reasonably preserves proximity information when combined with the PCA. For the diffusion map this is normal, since both the stochastic complementation and the diffusion map distance are based on the Markov chain. On the contrary stochastic complementation should not be combined with a Laplacian Eigen map because the resulting mapping would not be accurate.

V. CONCLUSION

This work introduced the discovering and mining of links for protein databases allowing analysing the relationships. The database is viewed as a graph, where the nodes correspond to the elements contained in the tables and the links between the interaction sets corresponds to the links in the tables. This work proposes to use stochastic complementation for extracting a subgraph containing the elements of interest from the original graph.. Also this work introduces a kernel-based extension of the basic diffusion map for displaying and analysing the reduced subgraph.

The stochastic complementation reduces the original graph and allows focusing the analysis on the proteins of interest, without having to define a state of the Markov chain for each protein of the database. A limitation of this work is that when trying to work with MYSQL it provides a limitation with the number of data sets to be considered for the work. The time taken for converting the table to a graph happens in O(n)2 time which also is a limitation when working with large data sets. A recommendation for this work could be obtained through Big Data Analytics with the mining concepts.

REFERENCES

[1]. M.Belkin and P.Niyogi ,*"Laplacian EigenMaps for Dimensionality reduction and Data Representation",* Neural Computation , Vol 15,pp.1373-1396,2006

[2]. I.Borg and P.Groenen, *Modern Multidimensional Scaling :Theory and Applications,* springer,1997.

[3]. T.Cox and M.Cox , *Multidimensional Scaling*, second ed. Chapman and Hall 2001.

[4]. X.Geng, D-C.Zhan and Z-H Zhou, *"Supervised Nonlinear Dimensionality Reduction for Visualization and Classification"IEEE Tras.Systems, Man and Cybernetics,* Part B: Cybernetics, vol.35, no.6, pp 1098-1107, Dec 2005.

[5]. J.Gower and D.Hand, *Biplots*, Chapman & Hall , 1995.

[6]. M.J.Greenacre, *Theory and Applications of Correspondence analysis* . Academic Press, 1984.

[7]. K.M Hall, *"An R-Dimensional Quadratic Placement Algorithm ,"*Management science, vol.17, no.8,pp.219-229, 1970.

[8]. J.Lee and M.Verleysen, *Nonlinear Dimensionality Reduction*, Springer , 2007.

[9]. B.Nadler,S.Lafon, R.Coifman and I.Keverekidis, *"Diffusion Maps, Spectral Clustering and Reaction coordinates of Dynamical systems",* Applied and Computational Harmonic Analysis, Vol.21, pp.113-127,2006.

[10]. P.Pons and M.Laptapy, *"Computing Communities in Large Networks Using Random Walks,"* Int'l Symposium of Computer and Information sciences, pp.191-218,2006.

[11]. M.Telwal, *Link Analysis: An Information Science Approach*. Elsevier , 2004.

[12]. L.Yen, F.Fouss, C.Decaestecker and M.Saerens, *"Graph nodes Clustering Based on the Commute –Time Kernel* (PAKDD 07'), 2007

[13]. C.D. Meyer, "*Stochastic Complementation, Uncoupling Markov Chains, and the Theory of Nearly Reducible systems,"* SIAM Rev.,vol. 31, no. 2, pp. 240-272, 1989.

[14]. F. Geerts, H. Mannila, and E. Terzi, "Relational Link-Based Ranking," *Proc. 30th Very Large Data Bases Conf. (VLDB)*, pp. 552-563, 2004.