Visualization of Protein-Protein Interaction Network for Knowledge Discovery

Weizhong Zhu, Xia Lin, Xiaohua Hu, Bahrad A. Sokhansanj

***Abstract*— A new visualization tool, called "Visual Concept Explorer (VCE)", was developed to visualize concept relationships in bio-medical literature. VCE integrates Pathfinder Network Scaling and Kohonen Self-organizing Feature Map Algorithm for visual mapping. As a case study, VCE was applied to visualize a chromatin protein-protein interaction (PPI) network. The mapping results demonstrated that VCE could explore the semantic structure and latent domain knowledge hidden in protein-protein interaction data sets generated from bio-medical literature.**

***Index Terms*—knowledge discovery, protein-protein interaction network, information visualization**

I. Introduction

RECENT advances in proteomics technologies, such as the yeast two-hybrid system [1], make it possible to rapidly produce large data sets of protein-protein interaction. New protein complexes according to protein interaction data from literature mining may be feasibly generated and provide higher level of functionality. However, for very large protein interaction data set from mining with multi-dimensions, the latent functional relationships among proteins are too complicate to identify even for domain experts. In such a case, information visualization will be a necessary addition to exploration of mining results generated by data mining techniques such as data dimension deduction and similarity clustering.

Visual Concept Explorer (VCE) is designed to explore the semantic structure of bio-medical literature and inherent relationships between concepts through co-occurrence analysis of terms, for instance, protein MESH terms. Word co-occurrence analysis is a content analysis technique that can be used to identify the strength of associations between words based on their co-occurrence in the same document [2]. Words that appear together often will have strength closer to 1, and words that never appear together, strength of 0. According to different strengths between protein pairs, VCE classifies proteins into a different graphical structure. In VCE, the classification process is based on two models, Kohonen Self-organizing Map model and Pathfinder Network model. Kohonen Self-organizing Map [3] is one of the major artificial neural network models for data analysis and pattern recognition. The resultant Kohonen maps are organized in such a way that similar data are mapped onto the same neuron or to neighbor neurons. For PPI networks, the arrangement of the clusters on the maps may reflect certain general biological functions of genes. Pathfinder Network Scaling is originally developed by cognitive psychologist to capture salient relationships between concepts [4]. In prior studies, Pathfinder Network Scaling was used to extract underlying patterns in the similarity matrix and to present them spatially in a class of "pathfinder networks" by Chen [5] .The approach was implemented in the Star-Walker system that displays citation networks as a set union of all possible minimum spanning trees. In this study, Pathfinder Network Scaling was extended to explore biological pathways of PPI networks.

In this article, we first introduce the Kohonen Self-organizing Map algorithm and Pathfinder Network Scaling algorithm. Then we present mapping results of using the two algorithms for a chromatin PPI network. Finally we discuss how these two proximity cluster algorithms can assist users in cell biology domains in carrying out the knowledge discovery process.

Manuscript received Jun 2, 2005.

Weizhong Zhu is with College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104 USA (e-mail: wz32@ drexel.edu).

Xia Lin is with College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104 USA (e-mail: xlin@ cis.drexel.edu, Lin's work is supported by IBM Shared University Research Grant).

Xiaohua Hu is with College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104 USA (e-mail: thu@ cis.drexel.edu, Hu's work is supported partially by the NSF Career grant IIS-0448023 and PA Dept of Health Grants #239667).

Bahrad A. Sokhansanj is with Department of Bio-Science & Bio-Technology, Drexel University, Philadelphia, PA, 19104 USA (e-mail: bahrad.sokhansanj@ drexel.edu).

II. Algorithm and Implemetation

A. Co-occurrence Space Generation of the Chromatin PPI Network from Literature Mining

About 380 chromatin proteins were selected by domain expert as input for an information extraction system called SPIE [7] and protein-protein interactions were extracted from all the related literature from MEDLINE up to Jun, 2004.

0-7803-9017-2/05/$20.00 ©2005 IEEE

Totally 1217 protein pairs with interaction frequencies and differentiated semantic interaction types were achieved by the experiments of SPIE. Ignoring the semantic interaction type and adding up the interaction frequencies of the same protein pairs, a co-occurrence space for a PPI network was generated with distinguished 575 protein pairs. It is represented as a whole 380 x 380 non-normalized co-occurrence matrix that has 575 non-zero entries with 380 different proteins.

B. *Kohonen Self-organizing Map*

Kohonen Self-organizing Map is an artificial neural network used to project multi-dimensional data on to a 2-D representation space. By unsupervised competitive learning, a data classification can be viewed as a mapping neuron grid in which neurons establish predefined neighborhood relationships in input. The learning process goes through as followings:

Step1: Select

$$P(t) = \{P_1(t), P_2(t), ...P_N(t)\}$$

randomly as an input vector at time t and

$$W^K(t) = \{W^K{}_1(t), W^K{}_2(t), ...W^K{}_N(t)\}$$

as a weight vector for neuron K at time t.

Step2: Find the winning neuron $W^W(t)$ whose weights are closest to the input vector in the N-dimensional space so that

$$\| W^W(t) - P(t) \| = \min_K \| W^K(t) - P(t) \| \quad (1)$$

Step3: Adjust the weights of the winning neuron and the neurons close to the winning neuron in the 2-D representation space based on

$$W^K(t+1) = W^K(t) + \alpha(t)k(t)(W^K(t) - P(t)) \quad (2)$$

where

$$\alpha(t)\,(0 \le \alpha(t) \le 1)$$

is a time decreasing function that makes the map converge and $k(t)$ is a neighborhood adaptation function that shrinks the neighborhood of a neuron gradually over time .

In this application of Kohonen Self-organizing Map in VCE, Gaussian functions were adapted to describe and unify $\alpha(t)$ and $k(t)$ within the formula [8]:

$$\alpha(t)k(t) = A_1 \exp(-t / A_2) \exp(-tD(K,W) / A_3) \quad (3)$$

where $D(K,W)$ is the Euclidian distance between the neuron K and the winning neuron W, the first Gaussian function controls the weight update speed and the second one defines the neighborhood shrinkage.

This unsupervised learning process requires no category information that accompanies within the training patterns. This learning strategy makes it possible to discover new knowledge and unexpected patterns hidden in the large proximity data set.

In this PPI Network study, we selected a subset of the whole protein co-occurrence space, a 110 x 110 matrix, with 110 proteins which interact with other proteins the most frequently and are showed in Fig. (1).

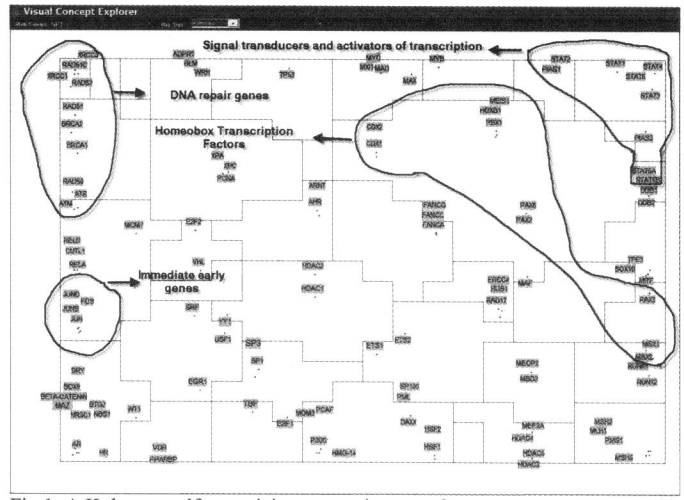

Fig.1. A Kohonen self-organizing semantic map of a co-occurrence space with 110 chromatin proteins is automatically generated by VCE after 5000 training cycles.

The map was presented to a domain expert for review. He confirmed that the proteins were explicitly mapped into different regions that reflect certain gene functionality. For instance, regions circled in Fig.(1) could be described respectively as homeobox transcription factors [9], signal transducers and activators of transcription [10], immediate early genes [11][12], and DNA repair genes [19].

C. *Pathfinder Network Scaling*

Pathfinder Network Scaling [14] relies on the triangle inequality to eliminate redundant links. Given two paths in a network that connect two nodes, the path is preserved that has a greater weight defined via the Minkowski metric. It is assumed that the path with the greater weight better captures the interrelationship between the two nodes and that the alternative path with less weight is redundant and should be pruned from the network.

Two parameters r and q influence the topology of a pathfinder network. The r-parameter influences the weight of a path based on the Minkowski metric. The weight of a path P with k links, W (P), is determined by weights w_1, w_2.....w_k of each individual link as follows [6]:

$$W(P) = (\sum_{i=1}^{k} w_i^r)^{1/r} \quad (4)$$

For r=1, the path weight is the sum of the link weights

along the path; for r=2, the path weight is the same as Euclidean distance; and for r=∞, the path weight is the same as the maximum weight associated with any link along the path.

The q-parameter defines the number of links in alternative paths up to which the triangle inequality must be maintained

$$w_{n_1,n_k} = (\sum_{i=1}^{k-1} w^r_{n_i,n_{i+1}})^{1/r} \quad \forall k \le q \quad (5)$$

A network of N nodes can have a maximum path length of q=N-1. With q=N-1 the triangle inequality is maintained throughout the entire network.

Protein-protein interaction networks are typically visualized in PFNETs where the nodes represent proteins and the edges represent the interactions. The positions of nodes in PFNETs are computed using a spring embedding algorithm described by Kamada and Kawai [15]. The FDP (Force Directed Placement) layout views nodes as physical bodies and edges as springs connected to the nodes providing forces between them. Nodes move according to the forces on them until a local energy minimum is achieved. FDP could be used to sort randomly placed nodes into a desirable layout that satisfies the aesthetics for visual presentation, such as symmetry, non-overlapping and so on. The length of the edges represents how relevant the two proteins are. The shorter the length is, the more relevant the two proteins are.

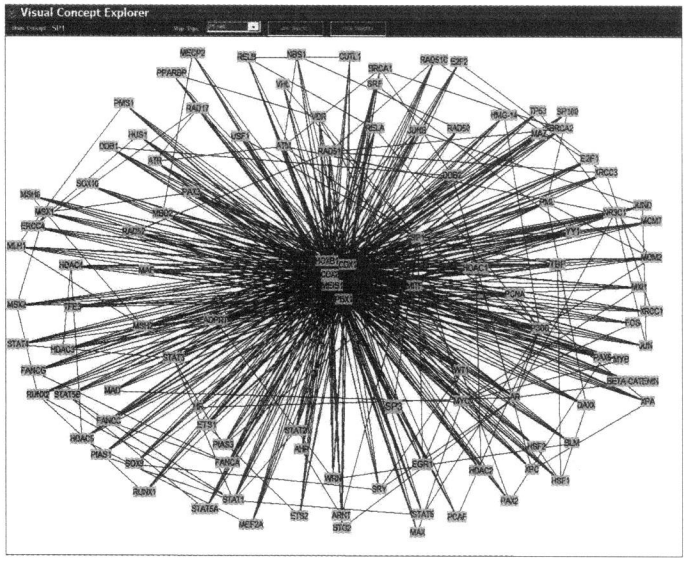

Fig.2 A PFNET of a co-occurrence space that includes 110 chromatin proteins is automatically generated by VCE with r=∞ and q =109.

Fig. (2) shows a PFNET of a PPI network using the same data set as that in Fig. (1). Based on expert feedback, we selected the PPI network with 110 protein nodes and 645 edges that was generated by using the parameter values r = ∞ and q = 109. The sparsely lattice-like graphs in this network reveal the complex biological pathways among the proteins. As other recent empirical studies have reported [16] [17], the PPI network has the scale free property characterized by a power law degree distribution. That is, the PPI network has many nodes but only few high degree nodes. In the center of Fig. (2), proteins PBX1, HOXB1, MESI1, CDX1 and CDX2 are located as a "hub" of the network and other proteins scatter around them.

III. Discussions

Kohonen Self-organizing Map clusters proteins into different regions according to semantic gene functions and the PFNET of PPI network explores what kind of routines these proteins followed to interact with each other. Combining information presented in the two types of map view, it is easy for biologists to understand and construct an interaction framework of PPI networks. For instance, PBX1, HOXB1 and MESI1 belong to the region identified as homeobox transcription factors. As a common sense in cell biology, these transcription factors control communications between DNA and mRNA and interact with many other proteins. So these proteins are located in the PFNET as the core. After all, in this level, Kohonen Self-organizing Mapping and Pathfinder Network Scaling do depict part of the domain knowledge in cell biology that has published in the literature of MEDLINE.

Moreover, are there some assumptions that could be made from these maps to conduct biologists' future research works? Swanson's approach [18] provides a way to think and discover hypotheses from the vast mount of latent connections. Swanson's ABC model states that given two premises that A causes B and that B causes C, the assumption would be whether A causes C. If the answer is positive, the causal relation is transitive. Modified to a PPI network, the approach could be represented in such a form: protein A interacts with protein B and protein B interacts with protein C, so that the question to ask is whether protein A interacts with protein C.

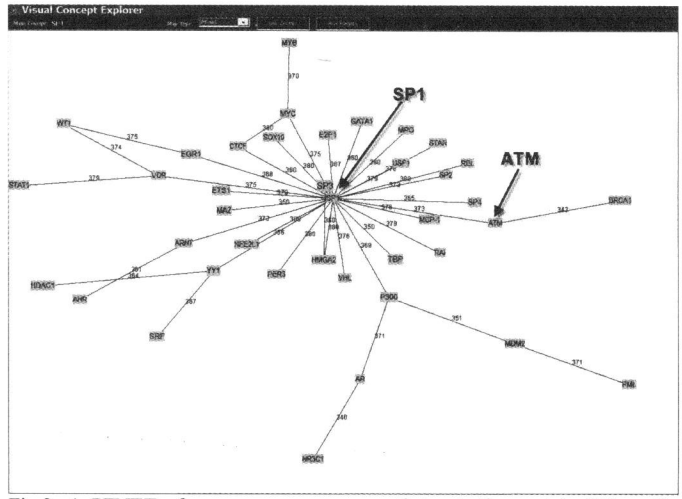

Fig.3. A PFNET of a co-occurrence space that sets the protein SP1 as the center is automatically generated by VCE with r=∞ and q =38. The space includes 38 proteins that have been identified to interact with SP1 from the literature mining.

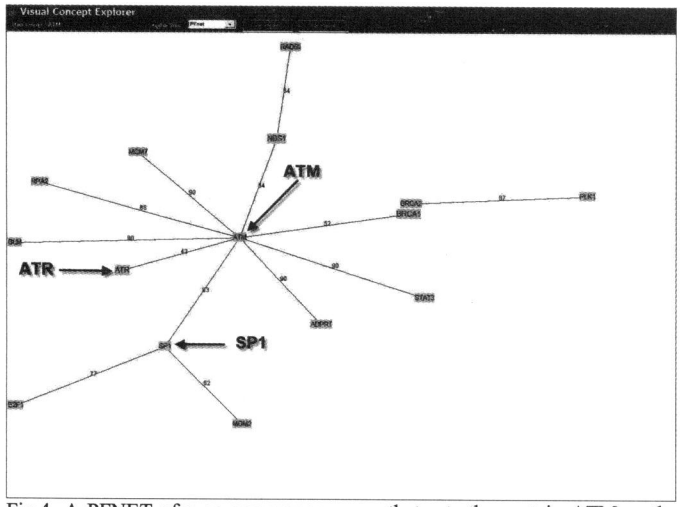

Fig.4. A PFNET of a co-occurrence space that sets the protein ATM as the center is automatically generated by VCE with $r=\infty$ and q =14. The space includes 14 proteins that have been identified to interact with ATM from the literature mining.

For instance, Fig. (3) shows a co-occurrence space of protein SP1 with a PFNET and other 38 proteins that have been known to interact with SP1 from the literature mining and Fig. (4) shows a co-occurrence space of protein ATM and other 14 proteins that have been known to interact with ATM from the literature mining. On one hand, Fig. (3) and (4) depict that protein ATR associates with protein ATM and protein ATM interacts with protein SP1. On the other hand, in Fig. (1) protein ATM and protein ATR locate in the same region – DNA repair genes, which predicts ATM and ATR have similar functionality (DNA damage sensors [13]). So we could suggest that studying the interaction between ATR and SP1 might be a promising research direction. But in the complex biological world, such transitive properties may not always be there. These hypotheses need domain expert to validate experimentally.

IV. Conclusion and Future Works

The results of Co-occurrence analysis on PPI networks by VCE are promising. To some extent, mapping by PFNET differentiates proteins from different biological pathways, and mapping by Kohonen Self-organizing Feature Map conveys diverse biological functions. They also reveal some possibilities for future research on cell biology. Currently, the VCE tool can be used as an initial step for gene interaction exploration. More needs to be done to make the tool more effective and powerful. Some of the issues we will address in our future work include:

(1) How to integrate ontology, such as gene ontology, within VCE and make the evaluation of the system automatic or semi-automatic? In this study, we got aids from a domain expert to identify the regions of gene functions and protein interaction pathways. But domain experts have their limitation in the validation because each domain expert is only familiar with 20 to 30 proteins, but a PPI network is often composed of hundreds to thousands of proteins.

(2) How to represent hierarchical relationships between concepts or latent domain knowledge? In this study, Kohonen Self-organizing Map only reflects these gene functions of the regions in one layer and couldn't give many hints to inter-relationships between these regions.

In addition, we will also address the issue of system performance and visualization effect. PFNET requires a relative high computational cost, especially for high scalable PPI networks. How to use PFNETs to represent huge amounts of data sets is a big challenge. All these developments will be explored in the near future.

References

[1] Uetz, P., Giot, L., Cagney, G. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, nature, 403, 623-627, 2000

[2] Callon, M., Law, J., & Rip, A. (1986) Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World (Macmillan, London).

[3] Kohonen, T. (1995). Self-Organizing Maps. Berlin: Spring-verlag.

[4] R. W. Schvaneveldt, F. T. Durso, and D. W. Dearholt. "Network structures in proximity data," in The Psychology of Learning and Motivation, 24, G. Bower, Ed. New York: Academic, 1989, pp. 249–284.

[5] Chen, C., & Carr, L. (1999) Trailblazing the literature of hypertext: Author co-citation analysis (1989-1998). In Proceedings of the 10th ACM Conference on Hypertext, pp. 51-60.

[6] Chen, C., Kuljis, J., Paul, R. J. (2001) Visualizing latent domain knowledge. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 31(4), 518 - 529.

[7] X Hu, Illhoi Yoo, Il-Yeol Song, Ming Song, Jianchao Han, Mark Lechner. Extracting and Mining Protein-Protein Interaction Network from Biomedical Literature. Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology.

[8] X. Lin, G. Marchionini, and D. Soergel. A self-organizing semantic map for information retrieval, In: Proc. of the 14th International ACM/SIGIR Conference on Research and Development in Information Retrieval, pp. 262-269, ACM Press, New York, 1991.

[9] Clemente Cillo, Monica Cantile, Antonio Faiella, Edoardo Boncinelli. Homeobox genes in normal and malignant cells, Journal of Cellular Physiology, Volume 188, Issue 2 , Pages 161 – 169, Published Online: 11 Jun 2001.

[10] Halupa A, Bailey ML, Huang K, Iscove NN, Levy DE, Barber DL. A novel role for STAT1 in regulating murine erythropoiesis: deletion of STAT1 results in overall reduction of erythroid progenitors and alters their distribution. Blood. 2005 Jan 15;105(2):552-61. Epub 2004 Jun 22.

[11] Yaseen NR, Park J, Kerppola T, Curran T, Sharma S. A central role for FOS in human B- and T-cell NFAT (nuclear factor of activated T cells): an acidic region is required for in vitro assembly. Mol Cell Biol. 1994 Oct; 14(10):6886-95.

[12] Hanlin Wangl, Mark Birkenbach and John Hart. Expression of Jun family members in human colorectal adenocarcinoma, Carcinogenesis, Vol. 21, No. 7, 1313-1317, July 2000

[13] Ariumi Y, Turelli P, Masutani M, Trono D. DNA damage sensors ATM, ATR, DNA-PKcs, and PARP-1 are dispensable for human immunodeficiency virus type 1 integration. Journal of Virology, March 2005, p. 2973-2978, Vol. 79, No. 5

[14] Schvaneveldt, R. (Ed.) (1990). Pathfinder Associative Networks: Studies in Knowledge Organization. Norwood, NJ: Ablex.

[15] Kamada, T. and Kawai, S. (1989). An Algorithm for Drawing General Undirected Graphs. Information Processing Letters, vol. 31, no. 1: 7-15.

[16] G. Bader and C. Hogue. An automated method for finding molecular complexes in large protein interaction networks. BMC Informatics, 4(1):2, 2003.

[17] Barabasi, A.-L. & Albert, R. Emergence of scaling in random networks Science 286, 509-512 (1999).

[18] Swanson, DR.. Fish-oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30(1), 7-18, 1986b.
[19] Richard D. Wood, Michael Mitchell, John Sgouros, and Tomas Lindahl. Human DNA Repair Genes. Science Feb 16 2001: 1284-1289.