Clustering Protein Interaction Networks Using The Matrix Factorization Method

Mohammad Shamsur RAHMAN
University of Windsor
sbintayab@gmail.com

Abstract: Protein interaction networks are an important feature in the field of bioinformatics. They reveal the characteristics of genes, living cells, molecular structures and several diseases. One of the most common natures of the protein interaction network is that all similar functional proteins are connected to each other to form a network. If protein interaction networks are grouped according to their function, it is easy to reveal the characteristics of genes, living cells etc. Many researches are investigating the clustering of protein interaction networks using several methods- matrix factorization is one of them. This survey focuses on the different approaches of the matrix factorization process for clustering of protein interaction network to reveal biological characteristics.

General Terms: Protein interaction network (PIN), Non-negative matrix factorization (nmf), Clustering.

Addition keyword: Protein complexes, Cancer.

Content

1. INTRODUCTION

The survey reviews research on several matrix factorization methods for clustering a protein interaction network (PIN) into several groups to identify the cancer type or biological processes or biological structures (cellular structures) or gene functionalities of a cell. Gene functionalities or biological processes or biological structures prediction are also required for drug and medicine development and detecting the disorder of normal behaviour of a living species.

About 57 different journal and conference papers were found which are related to the survey topics. there are only 15 papers exactly related on the topic.

The research papers for this survey were found using Google Scholar, ACM, LNCS and IEEE. Most of the papers were published in *Oxford Journal Bioinformatics*, *PLoS Computational Biology*, *Proceedings of conferences on Bioinformatics and Computational Biology*.

In the survey, ten papers are annotated. According to the annotation, it is found that all authors used non-negative matrix factorization method for predicting different biological activities such as cancer and disease detection, cell activities, gene functionality, bio-molecular structures etc. There is one interesting observation is that all authors work independently.

The authors of ten annotated papers put great contribution on the development of clustering of protein interaction networks (PINs). They proposed several matrix factorization methods for clustering a protein interaction network (PIN) into several groups according to the nature of the PIN where classical matrix factorization process fails. They proposed the method depends on the output and nature of the protein interaction network. For this reason, matrix factorization method is a reliable method for clustering the protein interaction networks.

The survey report is arranged into several sections. In the ten annotated papers, several problems are discussed. In section 2, the survey is discussed in detail which is named as Survey of Research. Concluding comments are in section 3. Finally the ten annotations are added in section 4, followed by the bibliography.

2. SURVEY OF RESEARCH

In this section, survey of research is discussed in details. In the survey, ten papers are annotated. Whole survey is classified into four sections according to the nature and characteristics of protein interaction networks and the goal of the work. In subsection 2.1, cancer related works are discussed. In subsection 2.2, how protein interaction networks help to identify the molecular structure of a cell and its identification process using matrix factorization method. Gene functionality discovery process using matrix factorization method from protein interaction networks is described in subsection, 2.3. Finally in subsection 2.4, prediction of functionality of a cell is investigated. Moreover, in each subsection, a summary is added.

2.1. Cancer and Its Type Identification

The research papers in this subsection present studies on cancer and its type identification by clustering of protein interaction networks (PINs) using several matrix factorization methods. A protein interaction network of a cancer cell is the input to the problem and the clustering of the PIN of the cancer cell is the goal to identify the cancer type. The majority of papers represent the individual process for identifying the cancer and its type by clustering of the protein interaction network.

2.1.1 *Clustering of cancer cell under different initial conditions*. It is very important to classify and identify cancer and its type for treatment. Protein interaction networks can be used to diagnose and classify the cancer types. Generally, Hierarchical Clustering (HC) and Self Organizing Maps (SOMs) are powerful methods for identifying the cancer classes from protein interaction network by clustering. According to Gao *et al.* [2005], these methods are unstable methods for producing clusters for different initial conditions. For different initial conditions, it is very important to classify the cancer and its subtypes.

The authors refer to previous work by Brunet *et al.* [2004]

For clustering or classifying the cancer and its subtypes under different initial conditions, the authors criticize Brunet *et al.* [2004]'s designed method using classical non-negative matrix factorization (NMF). Gao [2005] state that, there is a problem of working with classical NMF for classifying or clustering cancer and its subtypes- classical NMF gives no control over the sparseness of the decomposition.

Gao [2005] described a sparse non-negative matrix factorization method (SNMF) for clustering the protein interaction network of cancer into its types. According to their method, protein interaction network is represented as a matrix A. The matrix A is factorized into two matrices W and H. Here H matrix is used to determine the cluster membership. Update process of W is remained unchanged means it follows the previous method, but new method is proposed for H. Due to H is the matrix for clustering membership, sparseness of H should be controlled. For this reason, new formula is designed for H. The sparseness of H matrix is controlled by the value of λ. For the different initial conditions, the value of λ varies. These update process is continued until the Euclidean distance between A and WH is very negligible.

Gao [2005] tested their new method with three datasets which was used by Brunet *et al.* [2004]: Leukemia, Central nerve system tumors and Medulloblastoma datasets. They also compared the results with the results from Brunet *et al.* [2004]. They also investigated the biological meaning of sparseness in each cancer type.

Gao [2005] claim that SNMF outperforms NMF in the leukemia and central nerve system tumors datasets. But for the Medulloblastoma dataset, the result is not clear. In their investigation, they claim that SNMF can identify sets of proteins involved in the underlying cancer.

Gao [2005] state that their proposed SNMF improves the cancer classes discovery. For larger datasets for classifying the cancer, their proposed method can easily be applied.

2.1.2 *Identifying tumor types under local behaviour and alternative features*. Tumor and its type identification are very important for disease diagnostic process. The process of identifying a tumor type focuses on clusters or sub graphs of the protein interaction structures. Several methods or algorithms have been proposed for clustering protein interaction networks to identify the tumor and its type. According to Frigyesi [2008], the existing clustering processes focus on the dominating structures in the data, ignoring the alternative features and local behaviours. These later two factors are very important to identify the tumor and its type.

The authors refer to previous work by Lee *et al.* [1999].

Lee *et al.* [1999] propose an alternative method for clustering protein interaction network (PIN) which considers not only dominating structures but also local behaviour. This alternative method is classical non-negative matrix factorization (NMF) method. The authors find a shortcoming in the method of Lee *at al.* [1999], it cannot detect the accurate structures or clusters in global gene structure.

Frigyesi [2008] modify the classical NMF method for solving the shortcoming of previous work for identifying the tumors and their types accurately. They use Poisson Likelihood function for generating A

from H and W. Here A is the data matrix and H and W are the factorized matrices of A matrix. For noise or error calculation, new equation or formula is designed using permutation of the rows of matrix A.

Frigyesi *et al.* [2008] implemented their modified classical NMF for Time series data, CNS tumor, Leukemia data, Lung cancer data. They also compared the experimented result with other existing clustering algorithms.

The authors claim that, for time series data, their proposed NMF performs well and gives the same result as K-means gives. In CNS tumor cases, NMF can easily classify the protein interaction network except for proteins MD12 and MGlio8. In Leukemia and Lung cancer data, modified NMF solves the problem of previous classical NMF method and finds the accurate clusters.

Frigyesi *et al.* [2008] state that new classical NMF contributes to extracting relevant biological structures. Their modified classical NMF may contribute to deeper understanding of tumoigenesis and tumor behaviour. Their proposed NMF can be an attractive and useful approach for disease classification and identification.

2.1.3 *Identifying the protein complexes for detecting the cause of diseases and cancer.* Protein complexes play important role on prediction of the functionalities of cells which helps to identify the cause of disease (especially in Cancer) of living being. To find out the protein complexes, Wang [2008] address that, it is mandatory to cluster the protein interaction network into several groups according to their functionalities.

The authors do not refer to any previous work.

Wang *et al.* [2008] developed an algorithm for clustering protein interaction network into several protein complexes according to their functionalities. They use classical non-negative matrix factorization method (NMF). Before factorization, they generate the similarity matrix, S for protein interaction network using four different methods- diffusion kernel feature, shortest path based, Hamming distance based and adjacency matrix based method. According to their proposed method, the algorithm factorizes the similarity matrix, S into two non-negative matrices W and H. The factorization process continues until the Euclidean distance between S and WH is very negligible which is less than 10^{-15}. Finally they check the accuracy of the clustering process by using the modularity measure, Q which is completely designed by the authors.

They evaluated the proposed method by using a large scale protein interaction network of yeast *S. cerevisiae* which consists of 1257 proteins and 6835 interactions. They separately determined the similarity matrix using four different methods and evaluated clustering using their proposed modularity measure.

The authors claim that, for diffusion kernel feature method, NMF can accurately identify the protein complexes from protein interaction network. Whereas, shortest path based method also gives accurate result but it takes exponential time for determining the protein complexes. Other two methods cannot give accurate results but these results are near to accurate result.

The authors state that, their proposed method achieves several contributions. First of all is their proposed method automatically detects the proper number of clusters by implementing or repeating the algorithm for *k* times. In second, this method can be used in other networks- social networks. Finally, their proposed modularity measure can be used in any clustering algorithm for checking its accuracy.

2.1.4. *Summary*

Year	Authors	Title of the paper	Major contribution
2005	Gao and Church	Improving molecular cancer class discovery through sparse non-negative matrix factorization	Introduces a new non-negative clustering method for clustering which improves the cancer class discovery.
2008	Frigyesi and Höglund	Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes	Improves the classical non-negative matrix factorization method for clustering time series data for classifying the tumor types.
2008	Wang, Zhang, Wang, Zhang and Chen	Clustering complex networks and biological networks by non-negative matrix factorization with various similarity measures	Improves the classical non-negative matrix factorization for automatically detect the proper number of clusters in protein interaction network for identifying the diseases.

Table-I. Major contribution in cancer and its type identification

2.2. Biological Molecular Structure Identification.

The research papers in this subsection are studies of the identification of the molecular structure of a cell by performing clustering of the protein interaction networks (PINs) using several matrix factorization methods. Same structured proteins are localized into a specific and fixed location. For predicting or searching a specific or special molecular structure of a cell, it is needed to cluster the protein interaction network according to the similar structures of the proteins. The majority of the papers investigate the individual process for predicting the cellular functionalities by clustering of the protein interaction networks.

2.2.1. *Prediction of molecular structure for gene experiment relation.* To understand the biological event- molecular structures, it is necessary to identify the local structures (molecular structure) of the protein interaction networks (PINs). For identifying the local structures, clustering technique is one of the best processes. Hierarchical clustering technique is commonly applied to cluster a protein interaction network. Puscual-Montano [2005] find that hierarchical method cannot identify the molecular structure for gene experiment relationships from clustering of protein interaction network.

Pascual-Montano [2005] do not refer any previous work.

Pascual-Montano [2005] propose bi-clustering algorithm which not only identifies the local structures but also the gene experiment relationship. According to their method, a positive data matrix V (containing protein interaction network) factorizes into three matrices W, S and H. Here S is non-smoothing matrix which controls the sparseness of the model. As well as it finds the gene experiment relationships. The matrix H gives the local structures of the PIN. The factorization process continues until the noise function returns very negligible values. The authors also develop the noise function using Poisson likelihood function. For introducing a non-smooth matrix for achieving global sparseness and identifying the gene experiment relationships, in non-negative matrix factorization method, this method is known as Non-Smooth Non-negative matrix factorization (nsNMF).

Pascual-Montano *et al.* [2005] evaluated the performance of their nsNMF by using three different datasets: soft tissue tumor dataset, gastrointestinal stromal tumor dataset, heterogeneous group of samples comprising liposarcomas, leiomyosarcomas and malignant fibrous histiocytosarcomas. They compared the result of first test case with the results of [Nielsen *et al.* 2002]. The second and third experiments were compared with existing results [Nielsen *et al.* 2002].

Pascual-Montano *et al.* [2005] claim that in their first experiment, nsNMF correctly identifies the different type of local structures which are not detected by many standard algorithms. In the second experiment, they find that, 8 gastrointestinal stromal tumor are partitioned relevantly. In third test case, Pascual-Montano *et al.* [2005] claim that, their proposed method (nsNMF) not only identifies the clustering for predicting molecular structure but also internal ranking (gene experiment relationships) of local structures of protein interaction network.

Pascual-Montano *et al.* [2005] state that, their non-smooth non-negative matrix factorization (nsNMF) method is a good alternative method for analyzing the molecular structure of protein interaction network. and discovering the process in gene experiments. It is also achieved its goals.

2.2.2. *Identifying the molecular structure according to the local property*. Biological networks (protein interaction networks) are massive and sizes are still increasing. It is very complicated to understand the structures of all molecules or proteins and their interaction which helps in the field of medicine development. According to Kim [2010] study, to understand the structures, it is necessary to cluster a biological network according to specific local properties. Each sub graph of the biological network of same local properties is known as a biological network motif (BNM).

Kim [2010] refer to previous work by Milo *et al.* [2002] and Wernicke *et al.* [2006].

The authors do not mention any shortcomings of previous works. The authors used these methods or works for comparing their proposed method.

Kim [2010] claim to improve the sparse NMF (non-negative matrix factorization) method. They generate the data matrix, A which contains the information about the biological network, from a graph representation of the biological network and a sparse constraint. The matrix A is decomposed into two non-negative matrices W and H using sparse non-negative matrix factorization method. This factorization

process continues until noise is less than 10^{-15}. The authors design a new noise equation which considers two parameters- sparseness parameter and balance parameter between correctness and sparseness for getting accurate biological network motif. Finally, the matrix *H* gives the clustering information of biological network. This NMF is known as NMF-BNM (non-negative matrix factorization- biological network motif).

Kim [2010] compared the performance of their proposed algorithm with Milo *et al.* [2002] and Wernicke *et al.*'s [2006] proposed methods. This testing was performed by two PINs (protein interaction networks) of *S. cerevisiae* (yeast) which was named as *DIP* and *Y2K*. Each of the PIN contained a good number of proteins and interactions. In their experiments, they tested the performance for 4 node and 5 node biological network motifs. They also used these PINs for testing the structural motif and biological network motif comparison for their proposed algorithm and previous works.

After comparing the performance of the NMF-BNM with previous work, Kim *et al.* [2010] claim that, except 4-node biological network motifs in *DIP* core network, NMF-BNM gives better result than others. After performing the biological network motif and structural motif comparisons by NMF-BNM and previous works, the authors claim that, ESU (enumerate subgraphs) by Wernicke *et al.* [2006] can produce only 30% of total subgraphs, whereas NMF-BNM produces almost 100%.

Kim *et al.* [2010] state that their proposed method for clustering a biological network into biological network motifs achieves several contributions. NMF-BNM successfully clusters not only biological networks but also structural networks. Their clustering algorithm clusters the network into subgraphs or subnetworks accurately, so that the understanding process of structure of the biological network is improved.

2.2.3. *Summary*

Year	Authors	Title of the paper	Major contribution
2005	Pascual-Montano, Carmona- Sáez, Pascual-Marqui, Tirado and Carazo	Two-way clustering of gene expression profiles by sparse matrix factorization	Improves the non-negative matrix factorization method for analyzing the molecular structure of protein interaction network using matrix sparseness nature.
2010	Kim, Li, Wang and Pan	Biological network motif detection and evalution	Introduction of new non-negative matrix factorization method which can cluster not only biological network but also structural networks.

Table-III. Major contribution in biological molecular structure identification

2.3. Gene Functionality Identification.

Gene function is very important to identify the cause of human or biological disorder or cause of diseases. The research papers in this subsection are studies of the functionality of the gene of a living being by performing clustering of the protein interaction networks (PINs) using several matrix factorization methods.

Similar functional genes share the same protein interaction network structured. For predicting or searching a specific or special functionality of a gene, it is needed to cluster the protein interaction network according to the similar structures of the proteins. The majority of the papers are investigating the individual process for predicting the cellular functionalities by clustering of the protein interaction networks.

2.3.1. *Identification of gene experiment relationship*. To understand gene experiment relationships, it is necessary to identify the local structures of the protein interaction networks (PINs). For identifying the local structures (protein complexes), classical one way clustering technique- hierarchical clustering is applied. Pascual-Montano [2005] identify that this method cannot identify the gene experiment relationships from protein interaction network.

Pascual-Montano [2005] do not refer any previous work.

Pascual-Montano [2005] propose bi-clustering algorithm which not only identifies the local structures but also the gene experiment relationship. According to their method, a positive data matrix V (containing protein interaction network) factorizes into three matrices W, S and H. Here S is non-smoothing matrix which controls the sparseness of the model. As well as it finds the gene experiment relationships. The matrix H gives the local structures of the PIN. The factorization process continues until the noise function returns very negligible values. The authors also develop the noise function using Poisson likelihood function. For introducing a non-smooth matrix for achieving global sparseness and identifying the gene experiment relationships, in non-negative matrix factorization method, this method is known as Non-Smooth Non-negative matrix factorization (nsNMF).

Pascual-Montano [2005] evaluated the performance of their nsNMF by using three different datasets: soft tissue tumor dataset, gastrointestinal stromal tumor dataset, heterogeneous group of samples comprising liposarcomas, leiomyosarcomas and malignant fibrous histiocytosarcomas. They compared the result of first test case with the results of [Nielsen *et al.* 2002]. The results of second and third experiment were compared with existing results [Nielsen *et al.* 2002].

Pascual-Montano [2005] claim that in their first experiment, nsNMF correctly identifies the different type of local structures which are not detected by many standard algorithms. In the second experiment, they find that, 8 gastrointestinal stromal tumor are partitioned relevantly. In third test case, the authors claim that, their proposed method (nsNMF) not only identifies the clustering but also internal ranking (gene experiment relationships) of local structures of protein interaction network.

The authors state that, their non-smooth non-negative matrix factorization (nsNMF) method is a good alternative method for analyzing the molecular structure of protein interaction network. and discovering the process in gene experiments. It is also achieved its goals.

2.3.2. *Prediction of the functionality of unknown gene.* In Gene ontology (GO), it is necessary to predict the functionalities of unknown gene. It is done by comparing the transcription factor protein. Dueck [2005] found that two same functional genes share same transcription factor protein in same location of protein interaction network of the genes. If the location of transcription factor protein of unknown

functional gene is found, it is comfortable to find the functionalities of unknown gene. Location of transcription factor protein is found by clustering of the protein interaction networks into several groups.

Dueck [2005] refer to previous work by Marcotte *et al.* [1999] and Hughes *et al.* [2000].

Marcotte [1999] and Hughes [2000] use the similarity of expression profile to cluster the protein interaction network (PIN) to predict the location of transcription factor protein. Their proposed method cannot reduce the noise which limits the predictive accuracy.

Dueck [2005] present an algorithm for clustering the PIN using the hidden factors, known as Probabilistic sparse matrix factorization (PSMF). According to their proposed method, X is a data matrix which contains the PIN and gene expressions. PSMF factorizes the X matrix into two matrices Y and Z using the concept of sparse matrix factorization method. Z matrix contains the clustering information of PIN and gene expressions. The factorization process continues until the error between X and $Y \cdot Z$ is very negligible.

Each element of X matrix is determined by using transcription factor protein profiles and noise of the gene. The noise is calculated by assuming the presence of the gene specific isotropic Gaussian sensor noise with variance ψ_g^2 . Free energies are also minimized for determining the element of X matrix. The process for determining X matrix is normally distributed. Due to use of probabilistic process for generating X matrix, the factorization process is known as PSMF.

The authors experimented with the PSMF for the most comprehensive mammalian gene expression dataset which contains over 40000 known transcription factor proteins. They also compared the result with other clustering algorithms- Sparse matrix factorization method (SMF), Principal Component Analysis (PCA), Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Independent Component Analysis (ICA). They further implemented the algorithm for unsupervised characterization of mRNA data.

The authors claim that, for first test, PSMF gives the same result as UPGMA. UPGMA is very expensive and almost accurate method for clustering the PIN. PSMF also gives the better result than SMF, PCA and ICA. For the unsupervised characterization of mRNA data, PSMF also gives the accurate data over the data found from Marcotte [1999] and Hughes [2000] works.

Dueck [2005] state that, normally any matrix factorization method gives more emphasis on the clustering technique, but their method gives more emphasis on accurate data matrix. For this reason their proposed method has higher statistical significance on clustering process. Another advantage is claimed by the authors about their proposed method is their method can find the secondary and higher order labels easily. The authors also claimed that, the problem of Marcotte *et al.* [1999] and Hughes *et al.* [2000] works is solved by their proposed method.

2.3.3. *Identifying differential expressed gene.* Many species in the world share same differential expressed gene which is used for the classifying the unknown species. Gene expression microarrays are a popular approach for identifying differential expressed genes between two cell types. For identifying the differential expressed gene, several hundred even few thousand typical comparisons are required. Despande [2010] found another way in protein interaction network (PIN) to identify the differentially expressed

genes. It is done by clustering PIN into several clusters according to the differential nature of the protein complexes.

No previous work is referred by Despande [2010].

The authors developed an algorithm for clustering protein interaction network into several protein complexes according to their functionalities. They use classical non-negative matrix factorization method (NMF). According to their proposed method, the algorithm factorizes the protein interaction network matrix, A into two non-negative matrices W and H. The factorization process continues until the distance between S and W, H is very negligible.

Despande [2010] implemented their proposed method for identifying the differentially expressed gene for human and mouse. They also performed a network randomization analysis in human and mouse. Also they evaluated the subnetworks (clusters of PIN) in terms of their functional coverage and relevance.

The authors claim that, their proposed algorithm successfully identifies the 255 real subnetworks of differentially expressed gene in human and mouse. For randomization analysis in human and mouse gene, the authors find that, cross-species approaches are statistically significant. In the third experiment, significant overlap is found in each subnetwork.

The authors state that, their proposed method or algorithm accurately clusters the PIN to find out the differentially expressed gene. Their method is also applicable for identifying the cross-species approach.

2.3.4. *Summary*

Year	Authors	Title of the paper	Major contribution
2005	Pascual-Montano, Carmona- Sáez, Pascual-Marqui, Tirado and Carazo	Two-way clustering of gene expression profiles by sparse matrix factorization	Improves the non-negative matrix factorization method for analyzing the gene functionality using matrix sparseness nature.
2005	Dueck, Morris, and Frey	Multi-way clustering of microarray data using probabilistic sparse matrix factorization	Improves the data matrix for accurately classifying the gene functionality.
2010	Deshpande, Sharma, Verfaillie, Hu, and Myers	A Scalable Approach for Discovering Conserved Active Subnetworks across Species	Introduction of a new non-negative matrix factorization algorithm which can accurately find out the differentially expressed gene.

Table-IV. Major contribution in gene functionality identification.

2.4. Cellular Functionality Identifications.

The research papers in this subsection describe studies on the identification of cellular functionalities of a cell by performing clustering of the protein interaction networks (PINs) using several matrix factorization methods. Interacted proteins share same characteristics or function. If the proteins of a cell are grouped according to their functionalities means protein interaction, transportation of cytoplasm etc., the functionalities of a cell can be detected easily. For this reason, it is necessary to cluster a protein interaction network to reveal the functionality of a cell. The majority of papers are investigating the individual process for predicting the cellular functionalities by clustering of the protein interaction networks.

2.4.1 *Identifying the protein complexes for detecting cellular functionality.* Protein complexes play important role on prediction of the functionalities of cells which helps to identify the cause of disease (especially in Cancer) of living being. To find out the protein complexes, Wang [2008] address that, it is mandatory to cluster the protein interaction network into several clusters according to their functionalities.

The authors do not refer to any previous work.

Wang *et al.* [2008] developed an algorithm for clustering protein interaction network into several protein complexes according to their functionalities. They use classical non-negative matrix factorization method (NMF). Before factorization, they generate the similarity matrix, S for protein interaction network using four different methods- diffusion kernel feature, shortest path based, Hamming distance based and adjacency matrix based method. According to their proposed method, the algorithm factorizes the similarity matrix, S into two non-negative matrices W and H. The factorization process continues until the Euclidean distance between S and W, H is very negligible which is less than 10^{-15}. Finally they check the accuracy of the clustering process by using the modularity measure, Q which is completely designed by the authors.

They evaluated the proposed method by using a large scale protein interaction network of yeast *S. cerevisiae* which consists of 1257 proteins and 6835 interactions. They separately determined the similarity matrix using four different methods and evaluated clustering using their proposed modularity measure.

The authors claim that, for diffusion kernel feature method, NMF can accurately identify the protein complexes from protein interaction network. Whereas, shortest path based method also gives accurate result but it takes exponential time for determining the protein complexes. Other two methods cannot give accurate results but these results are near to accurate result.

The authors (Wang *et al.* [2008]) state that, their proposed method achieves several contributions. First of all is their proposed method automatically detects the proper number of clusters by implementing or repeating the algorithm for k times. In second, this method can be used in other networks- social networks. Finally, their proposed modularity measure can be used in any clustering algorithm for checking its accuracy.

2.4.2 *Predicting the functionality of a cell by find the overlapping modules of protein interaction network.* Direct decomposition or clustering of protein interaction networks (PIN) into overlapping

modules or communities is the most important for finding the cellular functionalities. Normally clustering of PIN is being operated on the principle of connectivity of proteins. Qi [2009] identify that types of the activation of the condition is not accepted during the clustering process.

The authors refer to previous work by Murali *et al.* [2008].

Murali *et al.* [2008] convert the condition sensitive PIN into response networks according to the conditions. From response networks, Murali *et al.* [2008] cluster the network into several modules or network legos. The authors find several shortcomings of Murali *et al.* [2008]'s work. At first they compute the number of possible network legos. After that, algorithm clusters the network according to the number of legos. This number is not fixed. But in real world, it is fixed. Second short coming is- if reverse engineering is applied to the legos for converting into response networks, sometimes exact response networks are not found.

Qi *et al.* [2009] propose a new non negative matrix factorization method for clustering condition sensitive PIN to solve the shortcomings of previous work. In their method, they factorize a matrix $\boldsymbol{R}$ representing the response networks into two non-negative matrices: a lego matrix $\boldsymbol{L}$ and a loading matrix $\boldsymbol{Z}$. Here R is $N \times D$ matrix where N is the number of response network and D is the number of interaction in global PIN. Each row represents the response networks and column represents the interaction and its weights. L is $K \times D$ matrix where K is the number of legos. Z is a $N \times K$ matrix. Each column of Z matrix represents the number of legos and rows represents the response networks. Z matrix contains only 0s and 1s.

Initially, the value of K is chosen to be infinite or large enough which is considered as infinite. For this reason, the factorization method is known as the Infinite non-negative matrix factorization (INMF) method. The value of K is reduced after each update process until any column of Z matrix contains zero valued columns. Update of L is performed by using Gaussian prior distribution functions and Z is by Bayesian nonparametric prior distribution function. Update process is continuing until noise is very negligible where noise is calculated by using Gaussian likelihood functions. Finally, each column of rows of lego matrix, L represents which interaction belongs to which legos or modules.

Qi [2009] performed a comparison between INMF, classical NMF (non-negative matrix factorization) and ILGM (Infinite Linear Gaussian Model) in a condition sensitive PIN for different number of response networks. As well as they implemented the INMF in Human Functional Genomic Data (HFGD) and Cancer Map Data (CMD) and compared the result with Murali *et al.* [2008] method. They also applied reverse engineering from legos to response networks.

After performing the comparison between INMF, classical NMF and ILGM for different response networks of a PIN, the authors claim that, their proposed INMF gives better result in comparison to other methods. Sometime, classical NMF does not perform, where INMF performs accurately. After experimenting HFGD and CMD by INMF and comparing result, the authors find and claim that INMF give same result which solves the shortcoming of Murali *et al.* [2008] proposed method. According to their claim, reverse engineering is successful for the INMF method.

Authors state that their proposed algorithm makes the several contributions: a. Successful implementation of factorization methods for clustering condition sensitive PIN by using the concept of network legos, b. INMF automatically learns the dimension of the factorized matrices, c. INMF solves the problems of classical NMF for clustering condition sensitive protein interaction network, d. After performing INMF, if reverse engineering is applied, always original response networks are found and e. It solves the shortcoming of the method proposed by Murali *et al.* [2008] proposed method.

2.4.3 *Identifying the proteomic functionalities of a cell.* Protein interaction networks (PINs) play key roles in the biological processes including cell cycle control, differentiation, protein folding, signalling, transcription, translation and transport. To understand and identify the biological process, it is essential to find out the protein complexes in protein interaction network. Protein complexes are groups of protein that densely interact with each another. Tu [2010] identify that, protein complex has dynamic nature which is difficult to accurately predict the protein complexes from PIN.

The authors do not refer to previous work.

Tu [2010] proposed a new method for identifying the protein complexes by clustering the PIN using a matrix factorization method. Their proposed method is known as binary matrix factorization method (BMF). According to their method, the PIN is converted into the symmetric adjacency matrix $X_{n,N}$ where $x_{i,j} \in \{0, 1\}$. After generating X matrix, proposed method factorized into two matrices A and Y using Bayesian Ying-Yang harmony learning function (BYY). BYY function can automatically predict the number of clusters. Finally, the matrix, Y gives the clustering information. In their method, they also modify the evaluation criteria- sensitivity, positive predictive value, accuracy and separation.

Tu [2010] had performed an experiment on protein interaction network from the MIPS database by their proposed method. They also implemented the algorithm on the altered graphs by randomly adding and deleting edges. Finally they compared the result with the results of two existing methods: Markov Cluster Algorithm (MCL) and the Spectral Clustering (SC) for clustering PIN.

In the first case, the authors claim that, BMF gives more accurate clustering information than MCL and SC algorithm which is 0.91. In second experiment, it is claimed that, BMF is better than MCL which is better than SC.

Tu [2010] state that their proposed BMF BYY method can accurately cluster the protein interaction network without prior knowledge of the number of clusters. Secondly, they also claim that, BMF BYY does not depend on any parameter or threshold like MCL or SC.

2.4.4 *Predicting the cohesive nature of a cell.* In protein interaction network, interacted proteins share common functionalities. But they do not belong to cohesive nature which is a special characteristic of protein responsible for causing disorder of Golgi body of any living cell. Pinker [2010] identify that, cohesive set of proteins are highly connected but sparsely scattered in the network which is difficult to cluster the protein interaction networks according to the cohesive nature.

The authors refer to previous work by Sharan *et al.* [2007]

Sharan *et al.* [2007] use the direct methods to identify the cohesive set of proteins. The identified shortcomings of the Sharan *et al.* [2007]'s proposed method by Pinker [2010] is- some cohesive set of proteins are remain undefined in the network

Pinkert *et al.* [2010] use the classical non-negative matrix factorization (NMF) method to find out the cohesive set of protein by using concept of clustering of protein interaction network. Before performing classical NMF, whole protein interaction network is converted into an image graph using adjacency and weight matrices of protein interaction network (PIN). After conversion of non diagonal image graph from PIN, classical NMF is implemented to find out the cohesive set of proteins.

Pinker [2010] tested with their proposed algorithm on HPRD protein interaction network containing more than 8500 nodes. They also compared the result of their method with diagonal image graphs.

The authors claim that, their proposed method successfully identifies the eight cohesive sets of proteins by converting the PIN into non-diagonal image graph. In second experiments, it is found that their proposed non-diagonal image graph is better suited to diagonal image graph for HPRD PIN.

The authors state that, their proposed method can successfully separate the PIN (protein interaction network) into cohesive group of nodes with more internal than external connection.

2.4.5 *Summary.*

Year	Authors	Title of the paper	Major contribution
2008	Wang, Zhang, Wang, Zhang and Chen	Clustering complex networks and biological networks by non-negative matrix factorization with various similarity measures	Improves the classical non-negative matrix factorization for automatically detect the proper number of clusters in protein interaction network for identifying the functionality of a living cell.
2008	Qi, Ding, Rivera, and Murali,	Learning Network Legos by Infinite Non-Negative Matrix Factorization	Improves the clustering process using network legos and solves the problem of classical non-negative matrix factorization method for clustering the condition sensitive protein interaction network for identifying the cellular functionality.
2010	Tu, Chen, and Xu	A binary matrix factorization algorithm for protein complex prediction	Introduction of a new non-negative matrix factorization method which can cluster protein interaction network accurately without prior knowledge on the number of clusters.

2010	Pinkert, Schultz, and Reichardt	Protein Interaction Networks- More Than Mere Modules	Introduction of new non-negative matrix factorization method for clustering protein interaction network according to cohesive nature of a living cell.

Table-II. Major contribution in cellular functionality identifications

3. CONCLUDING COMMENTS

The survey has reviewed ten papers. All of the papers are related to the clustering of protein interaction networks by using several matrix factorization methods for revealing the biological functions of a living being. The authors of the ten papers used matrix factorization method for predicting the different biological functionalities such as cancer and different disease detections, living cell functionality, molecular structure of a cell, gene prediction and discovering its functionalities etc. by performing clustering process. They are successfully predicted the different biological functionalities using matrix factorization methods.

In each of ten papers, the authors make some major contributions in predicting the biological functionalities. The major contributions of the ten annotated papers are shown in a tabular form in table-V.

Year	Authors	Title	Paper referred to	Major contribution
2005	Dueck et al.	Multi-way clustering of microarray data using probabilistic sparse matrix factorization	Marcotte et al., Hughes et al.	Improves the data matrix for accurately classifying the gene functionality.
2005	Gao et al.	Improving molecular cancer class discovery through sparse non-negative matrix factorization	Brunet et al.	Introduces a new non-negative clustering method for clustering which improves the cancer class discovery.
2005	Pascual-Montano et al.	Two-way clustering of gene expression profiles by sparse matrix factorization	No previous work	Improves the non-negative matrix factorization method for analyzing the gene functionality and molecular

				structures using matrix sparseness nature.
2008	Frigyesi et al.	Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes	Lee et al.	Improves the classical non-negative matrix factorization method for clustering time series data for classifying the tumor types.
2008	Qi et al.	Learning Network Legos by Infinite Non-Negative Matrix Factorization	Murali et al.	Improves the clustering process using network legos and solves the problem of classical non-negative matrix factorization method for clustering the condition sensitive protein interaction network for identifying the cellular functionality.
2008	Wang et al.	Clustering complex networks and biological networks by non-negative matrix factorization with various similarity measures	No previous work	Improves the classical non-negative matrix factorization for automatically detect the proper number of clusters in protein interaction network for identifying the functionality of a living cell.
2010	Despande et al.	A Scalable Approach for Discovering Conserved Active Subnetworks across Species	No previous work	Introduction of a new non-negative matrix factorization algorithm which can accurately find out the differentially expressed gene.
2010	Kim et al.	Biological network motif detection and evalution	Milo et al., Wenicke et al.	Introduction of new non-negative matrix factorization method which can cluster not only biological network but also structural networks.
2010	Pinkert et al.	Protein Interaction Networks- More Than Mere Modules	Sharan et al.	Introduction of new non-negative matrix factorization method for clustering protein interaction network according to cohesive nature of a living cell.
2010	Tu et al.	A binary matrix factorization algorithm for	No previous work	Introduction of a new non-negative matrix factorization method which

		protein complex prediction		can cluster protein interaction network accurately without prior knowledge on the number of clusters.

In concluding the survey, it was found that the matrix factorization method is one of the best methods to cluster any protein interaction network. Most of the authors simply modify the update process of factorized matrices by maintaining the main concept of matrix factorization. In concluding the survey it was also found that probabilistic approach is one of strong mechanism which is used in matrix factorization for accurate clustering of a protein interaction network. On the other hand, most of the authors put more emphasis on the theoretical approaches rather than biological parameters. Besides, some methods suffer high computational time.

In the survey, only six authors discuss future expansion of their works. Dueck *et al.* [2005] use only basic factors to cluster a protein interaction network of profile and adult tissues. They do not consider 50 hidden factors for clustering. Using these 50 hidden factors for clustering can be used as future extension of their work. They also mention that, their method can be improved for all tissues as future works.

Gao [2005] indicate that in their method more emphasis is given to theoretical approaches rather than biological. As future work, their proposed method can be improved by giving more emphasis on the biological approaches to discover cancer classes.

Qi [2009] state that their proposed method takes large amount of time for clustering large protein interaction network (PIN), because of the high time complexity. To reduce the time complexity of their method, they indicate two methods as future work of their method. One method is to introduce parallelism in INMF and another is to amend bi-clustering process in their method.

Despande[2010] also find that, in real time, the algorithm takes several days to run. For improving the efficiency of cluster discovery, more formal selection criteria for the parameter may be used as future work.

Like Gao *et al.* [2005], Kim *et al.* [2010] also give more emphasis on topological parameters to find the desire number of subgraphs. As the future extension of their work, biological information should be taken more than topological parameters.

The method of Tu *et al.* [2010] also suffers from higher computing cost. They do not mention any specific way to improve the speed of their method. They also mention that their work may be extended in future to find out the non-overlapping clusters.

4. ANNOTATIONS

4.1. Dueck *et al.* 2005

Bibliographic entry. DUECK D., MORRIS, Q. D., AND FREY, B. J., 2005, Multi-way clustering of microarray data using probabilistic sparse matrix factorization, In *Oxford Journals Bioinformatics*, 21, 1, 144-151.

Problem. In Gene ontology (GO), it is necessary to predict the functionalities of unknown gene. It is done by comparing the transcription factor protein. Two same functional genes share same transcription factor protein in same location of protein interaction network of the genes. If the location of transcription factor protein of unknown functional gene is found, it is comfortable to find the functionalities of unknown gene. Location of transcription factor protein is found by clustering of the protein interaction networks into several groups.

Previous work. The authors refer to previous work by Marcotte *et al.* [1999] and Hughes *et al.* [2000].

Shortcomings of previous work. Marcotte *et al.* [1999] and Hughes *et al.* [2000] use the similarity of expression profile to cluster the protein interaction network (PIN) to predict the location of transcription factor protein. Their proposed method cannot reduce the noise which limits the predictive accuracy.

New idea/ algorithm. Dueck *et al.* [2005] design an algorithm for clustering the PIN using the hidden factors, known as Probabilistic sparse matrix factorization (PSMF). According to their proposed method, X is a data matrix which contains the PIN and gene expressions. PSMF factorizes the X matrix into two matrices Y and Z using the concept of sparse matrix factorization method. Z matrix contains the clustering information of PIN and gene expressions. The factorization process continues until the error between X and $Y \cdot Z$ is very negligible.

Each element of X matrix is determined by using transcription factor protein profiles and noise of the gene. The noise is calculated by assuming the presence of the gene specific isotropic Gaussian sensor noise with variance ψ_g^2 . Free energies are also minimized for determining the element of X matrix. The process for determining X matrix is normally distributed. Due to use of probabilistic process for generating X matrix, the factorization process is known as PSMF.

Experimented conducted. The authors [Dueck *et al.* [2005]] experimented the PSMF for the most comprehensive mammalian gene expression dataset which contains over 40000 known transcription factor proteins. They also compared the result with other clustering algorithms- Sparse matrix factorization method (SMF), Principal Component Analysis (PCA), Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Independent Component Analysis (ICA). They further implemented the algorithm for unsupervised characterization of mRNA data.

Results. The authors claim that, for first test, PSMF gives the same result as UPGMA. UPGMA is very expensive and almost accurate method for clustering the PIN. PSMF also gives the better result than SMF,

PCA and ICA. For the unsupervised characterization of mRNA data, PSMF also gives the accurate data over the data found from Marcotte *et al.* [1999] and Hughes *et al.* [2000] works.

Claims made by the authors. Dueck *et al.* [2005] state that, normally any matrix factorization method gives more emphasis on the clustering technique, but their method gives more emphasis on accurate data matrix. For this reason their proposed method has higher statistical significance on clustering process. Another advantage is claimed by the authors about their proposed method is their method can find the secondary and higher order labels easily. The authors also claimed that, the problem of Marcotte *et al.* [1999] and Hughes *et al.* [2000] works is solved by their proposed method.

Citations. [Kim *et al.* 2007; Li *et al.* 2007; Schreiber *et al.* 2007; Li *et al.* 2008; Mejía-Roa *et al.* 2008; Barash *et al.* 2010; Chua *et al.* 2011]

4.2. Despande *et al.* 2010

Bibliographic entry. DESHPANDE, R., SHARMA, S., VERFAILLIE, C. M., HU, W.-S., AND MYERS, C. L., 2010, A Scalable Approach for Discovering Conserved Active Subnetworks across Species, In *PLoS Computational Biology*, 6, 12, 1-18.

Problem. Many species in the world share same differential expressed gene which is used for the classifying the unknown species. Gene expression microarrays are a popular approach for identifying differential expressed genes between two cell types. For identifying the differential expressed gene, several hundred even few thousand typical comparisons are required. But another way is found significantly in protein interaction network (PIN) to identify the differentially expressed genes. It is done by clustering PIN into several clusters according to the differential nature of the protein complexes.

Previous work. No previous work is referred by the authors.

Shortcomings of previous work. No shortcoming of previous because of absence of previous work.

New idea/ algorithm. The authors developed an algorithm for clustering protein interaction network into several protein complexes according to their functionalities. They use classical non-negative matrix factorization method (NMF). According to their proposed method, the algorithm factorizes the protein interaction network matrix, A into two non-negative matrices W and H. The factorization process continues until the distance between S and W, H is very negligible.

Experimented conducted. The authors implemented their proposed method for identifying the differentially expressed gene for human and mouse. They also performed a network randomization analysis in human and mouse. Also they evaluated the subnetworks (clusters of PIN) in terms of their functional coverage and relevance.

Results. The authors claim that, their proposed algorithm successfully identifies the 255 real subnetworks of differentially expressed gene in human and mouse. For randomization analysis in human and mouse gene, the authors find that, cross-species approaches are statistically significant. In the third experiment, significant overlap is found in each subnetwork.

Claims made by the authors. The authors state that, their proposed method or algorithm accurately clusters the PIN to find out the differentially expressed gene. Their method is also applicable for identifying the cross-species approach.

Citations. [Hyoung *et al.* 2011; Jeon *et al.* 2011]

4.3. Frigyesi *et al.* 2008

Bibliographic entry. FRIGYESI, A., AND HÖGLUND, M., 2008, Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes, In *Cancer Informatics*, 6, 275-292.

Problem. The process of identifying a tumor type focuses on clusters or sub graphs of the protein interaction structures. Several methods or algorithms have been proposed for clustering protein interaction networks to identify the tumor and its type. These clustering processes focus on the dominating structures in the data, ignoring the alternative features and local behaviours. These later two factors are very important to identify the tumor and its type.

Previous work. The authors refer to previous work by Lee *et al.* [1999].

Shortcomings of previous work. Lee *et al.* [1999] propose an alternative method for clustering protein interaction network (PIN) which considers not only dominating structures but also local behaviour. This alternative method is classical non-negative matrix factorization (NMF) method. Frigyesi *et al.* [2008] find a shortcoming in the method of Lee *at al.* [1999], it cannot detect the accurate structures or clusters in global gene structure.

New idea/ algorithm. Frigyesi *et al.* [2008] modify the classical NMF method for solving the shortcoming of previous work for identifying the tumors and their types accurately. They use Poisson Likelihood function for generating A from H and W. Here A is the data matrix and H and W are the matrix factorization of A matrix. For noise or error calculation, new equation or formula is designed using permutation of the rows of matrix A.

Experimented conducted. Frigyesi *et al.* [2008] implemented their modified classical NMF for Time series data, CNS tumor, Leukemia data, Lung cancer data. They also compared the experimented result with other existing clustering algorithms.

Results. The authors claim that, for time series data, their proposed NMF performs well and gives the same result as K-means gives. In CNS tumor cases, NMF can easily classify the protein interaction network except for proteins MD12 and MGlio8. In Leukemia and Lung cancer data, modified NMF solves the problem of previous classical NMF method and finds the accurate clusters.

Claims made by the authors. Frigyesi *et al.* [2008] state that new classical NMF contributes to extracting relevant biological structures. Their modified classical NMF may contribute to deeper understanding of tumoigenesis and tumor behaviour. Their proposed NMF can be an attractive and useful approach for disease classification and identification.

Citations. [Gaujoaux *et al.* 2010]

4.4. Gao *et al.* 2005

Bibliographic entry. GAO, Y., AND CHURCH, G., 2005, Improving molecular cancer class discovery through sparse non-negative matrix factorization, In *Oxford Journals Bioinformatics*, 21, 21, 3970-3975.

Problem. Accurate classification of cancer types is very important for treatment. Protein interaction networks can be used to diagnose and classify the cancer types. Generally, hierarchical clustering (HC) and Self organizing maps (SOMs) are powerful methods for identifying the cancer classes. These methods are unstable methods for producing clusters for different initial conditions. For different initial conditions, it is very important to classify the cancer and its subtypes.

Previous work. The authors refer to previous work by Brunet *et al.* [2004]

Shortcomings of previous work. For clustering or classifying the cancer and its subtypes for different initial conditions, Brunet *et al.* [2004] designed a new method using classical non-negative matrix factorization (NMF). Gao *et al.* [2005] state that, there is a problem of working with classical NMF for classifying or clustering cancer and its subtypes- classical NMF gives no control over the sparseness of the decomposition.

New idea/ algorithm. Gao *et al.* [2005] described a sparse non-negative matrix factorization method (SNMF) for clustering the protein interaction network of cancer into its types. According to their method, protein interaction network is represented as a matrix A. The matrix A is factorized into two matrices W and H. Here H matrix is used to determine the cluster membership. Update process of W is remained unchanged means it follows the previous method, but new method is proposed for H. Due to H is the matrix for clustering membership, sparseness of H should be controlled. For this reason, new formula is designed for H. The sparseness of H matrix is controlled by the value of λ. For the different initial conditions, the value of λ varies. These update process is continued until the Euclidean distance between A and WH is very negligible.

Experimented conducted. Gao *et al.* [2005] tested their new method with three datasets which was used by Brunet *et al.* [2004]: Leukemia, Central nerve system tumors and Medulloblastoma datasets. They also compared the results with the results from Brunet *et al.* [2004]. They also investigated the biological meaning of sparseness in each cancer type.

Results. Gao *et al.* [2005] claim that SNMF outperforms NMF in the leukemia and central nerve system tumors datasets. But for the Medulloblastoma dataset, the result is not clear. In their investigation, they claim that SNMF can identify sets of proteins involved in the underlying cancer.

Claims made by the authors. Gao *et al.* [2005] state that their proposed SNMF improves the cancer classes discovery. For larger datasets for classifying the cancer, their proposed method can easily be applied.

Citations. [Kim *et al.* 2007; Li *et al.* 2007; Fogel *et al.* 2007; Ochs 2010; Zhang *et al.* 2011]

4.5. Kim *et al.* 2010

Bibliographic entry. KIM, W., LI, M., WANG, J., AND PAN, Y., 2010, Biological network motif detection and evaluation, In *Proceedings of The 2010 International Conference on Bioinformatics and Computational Biology*, Las Vegas, NV, USA, 5, 1-13.

Problem. Biological networks (protein interaction networks) are massive and sizes are still increasing. It is very complicated to understand the structures of all molecules or proteins and their interaction. To understand the structures, it is necessary to cluster a biological network according to specific local properties. Each sub graph of the biological network of same local properties is known as a biological network motif (BNM).

Previous work. The authors refer to previous work by Milo *et al.* [2002] and Wernicke *et al.* [2006].

Shortcomings of previous work. The authors do not mention any shortcomings of previous works. The authors used these methods or works for comparing their proposed method.

New idea/ algorithm. Kim *et al.* [2010] claim to improve the sparse NMF (non-negative matrix factorization) method. They generate the data matrix, A which contains the information about the biological network, from a graph representation of the biological network and a sparse constraint. The matrix A is decomposed into two non-negative matrices W and H using sparse non-negative matrix factorization method. This factorization process continues until noise is less than 10^{-15}. The authors design a new noise equation which considers two parameters- sparseness parameter and balance parameter between correctness and sparseness for getting accurate biological network motif. Finally, the matrix H gives the clustering information of biological network. This NMF is known as NMF-BNM (non-negative matrix factorization- biological network motif).

Experimented conducted. Kim *et al.* [2010] compared the performance of their proposed algorithm with Milo *et al.* [2002] and Wernicke *et al.*'s [2006] proposed methods. This testing was performed by two PINs (protein interaction networks) of *S. cerevisiae* (yeast) which was named as *DIP* and *Y2K*. Each of the PIN contained a good number of proteins and interactions. In their experiments, they tested the performance for 4 node and 5 node biological network motifs. They also used these PINs for testing the structural motif and biological network motif comparison for their proposed algorithm and previous works.

Results. After comparing the performance of the NMF-BNM with previous work, Kim *et al.* [2010] claim that, except 4-node biological network motifs in *DIP* core network, NMF-BNM gives better result than others. After performing the biological network motif and structural motif comparisons by NMF-BNM and previous works, the authors claim that, ESU (enumerate subgraphs) by Wernicke *et al.* [2006] can produce only 30% of total subgraphs, whereas NMF-BNM produces almost 100%.

Claims made by the authors. Kim *et al.* [2010] state that their proposed method for clustering a biological network into biological network motifs achieves followings-

- NMF-BNM successfully clusters not only biological networks but also structural networks.

- Their clustering algorithm clusters the network into subgraphs or subnetworks accurately, so that the understanding process of structure of the biological network is improved.

Citations. [Zhang *et al.* 2011a]

4.6. Pascual-Montano *et al.* 2005

Bibliographic entry. PASCUAL-MONTANO, A., CARMONA-SÁEZ, P., PASCUAL-MARQUI, R. D., TIRADO, F., AND CARAZO, J.M., 2005, Two-way clustering of gene expression profiles by sparse matrix factorization, In *Proceedings of IEEE Computational Systems Bioinformatics Conference Workshop*, 5-6.

Problem. To understand the biological event- molecular structures, it is necessary to identify the local structures of the protein interaction networks (PINs). For identifying the local structures (protein complexes), classical one way clustering technique- hierarchical clustering is applied. But this method cannot identify the gene experiment relationships from protein interaction network.

Previous work. Pascual-Montano *et al.* do not refer any previous work.

Shortcomings of previous work. Due to no previous work, there is no shortcoming of previous work.

New idea/ algorithm. Pascual-Montano *et al.* [2005] propose bi-clustering algorithm which not only identifies the local structures but also the gene experiment relationship. According to their method, a positive data matrix V (containing protein interaction network) factorizes into three matrices W, S and H. Here S is non-smoothing matrix which controls the sparseness of the model. As well as it finds the gene experiment relationships. The matrix H gives the local structures of the PIN. The factorization process continues until the noise function returns very negligible values. The authors also develop the noise function using Poisson likelihood function. For introducing a non-smooth matrix for achieving global sparseness and identifying the gene experiment relationships, in non-negative matrix factorization method, this method is known as Non-Smooth Non-negative matrix factorization (nsNMF).

Experimented conducted. Pascual-Montano *et al.* [2005] evaluated the performance of their nsNMF by using three different datasets: soft tissue tumor dataset, gastrointestinal stromal tumor dataset, heterogeneous group of samples comprising liposarcomas, leiomyosarcomas and malignant fibrous histiocytosarcomas. They compared the result of first test case with the results of [Nielsen *et al.* 2002]. The second and third experiments were compared with existing results [Nielsen *et al.* 2002].

Results. Pascual-Montano *et al.* [2005] claim that in their first experiment, nsNMF correctly identifies the different type of local structures which are not detected by many standard algorithms. In the second experiment, they find that, 8 gastrointestinal stromal tumor are partitioned relevantly. In third test case, Pascual-Montano *et al.* [2005] claim that, their proposed method (nsNMF) not only identifies the clustering but also internal ranking (gene experiment relationships) of local structures of protein interaction network.

Claims made by the authors. Pascual-Montano *et al.* [2005] state that, their non-smooth non-negative matrix factorization (nsNMF) method is a good alternative method for analyzing the molecular structure of protein interaction network. and discovering the process in gene experiments. It is also achieved its goals.

Citations. There are no specific references to this paper by other researchers in this survey.

4.7. Pinkert *et al.* 2010

Bibliographic entry. PINKERT, S., SCHULTZ, J., AND REICHARDT, J., 2010, Protein Interaction Networks-More Than Mere Modules, In *PLoS Computational Biology*, 6, 1, 1371-1383. *Problem.*

Problem. In protein interaction network, interacted proteins share common functionalities. But they do not belong to cohesive nature. Cohesive nature is required to identify the cancer causing proteins. Cohesive set of proteins are highly connected but sparsely scattered in the network. Different clustering process is required to identify the cohesive set of proteins.

Previous work. The authors refer to previous work by Sharan *et al.* [2007]

Shortcomings of previous work. Sharan *et al.* [2007] use the direct methods to identify the cohesive set of proteins. The shortcomings of the Sharan *et al.* [2007]'s proposed method is- some cohesive set of proteins are remain undefined in the network

New idea/ algorithm. Pinkert *et al.* [2010] use the classical non-negative matrix factorization (NMF) method to find out the cohesive set of protein by using concept of clustering of protein interaction network. Before performing classical NMF, whole protein interaction network is converted into an image graph using adjacency and weight matrices of protein interaction network. After conversion of non diagonal image graph from PIN, classical NMF is implemented to find out the cohesive set of proteins.

Experimented conducted. The authors experimented with their proposed algorithm on HPRD protein interaction network containing more than 8500 nodes. They also compared the result of their method with diagonal image graphs.

Results. The authors claim that, their proposed method successfully identifies the eight cohesive sets of proteins by converting the PIN into non-diagonal image graph. In second experiments, it is found that their proposed non-diagonal image graph is better suited to diagonal image graph for HPRD PIN.

Claims made by the authors. The authors state that, their proposed method can successfully separate the PIN (protein interaction network) into cohesive group of nodes with more internal than external connection.

Citations. [Lei *et al.* 2011; Kaltenbach *et al.* 2012]

4.8. Qi *et al.* 2009

Bibliographic entry. QI, Y., DING, N., RIVERA, C. G., AND MURALI, T. M., 2009, Learning Network Legos by Infinite Non-Negative Matrix Factorization, In *World Congress of Pain Clinicians – Proceedings*, 1-12.

Problem. Direct decomposition or clustering of protein interaction networks (PIN) into overlapping modules or communities is the most important for finding the cellular functionalities. Normally clustering of PIN is being operated on the principle of connectivity of proteins. During the clustering, it is not

considered under which condition interaction is activated or not. So, it is important to design a clustering or decomposition algorithm which considers the interaction condition.

Previous work. The authors refer to previous work by Murali *et al.* [2008].

Shortcomings of previous work. Murali *et al.* [2008] convert the condition sensitive PIN into response networks according to the conditions. From response networks, Murali *et al.* [2008] cluster the network into several modules or network legos. The authors find several shortcomings of Murali *et al.* [2008]'s work. At first they compute the number of possible network legos. After that, algorithm clusters the network according to the number of legos. This number is not fixed. But in real world, it is fixed. Second short coming is- if reverse engineering is applied to the legos for converting into response networks, sometimes exact response networks are not found.

New idea/ algorithm. Qi *et al.* [2009] propose a new non negative matrix factorization method for clustering condition sensitive PIN to solve the shortcomings of previous work. In their method, they factorize a matrix $\boldsymbol{R}$ representing the response networks into two non-negative matrices: a lego matrix $\boldsymbol{L}$ and a loading matrix $\boldsymbol{Z}$. Here R is $N \times D$ matrix where N is the number of response network and D is the number of interaction in global PIN. Each row represents the response networks and column represents the interaction and its weights. L is $K \times D$ matrix where K is the number of legos. Z is a $N \times K$ matrix. Each column of Z matrix represents the number of legos and rows represents the response networks. Z matrix contains only 0s and 1s.

Initially, the value of K is chosen to be infinite or large enough which is considered as infinite. For this reason, the factorization method is known as the Infinite non-negative matrix factorization (INMF) method. The value of K is reduced after each update process until any column of Z matrix contains zero valued columns. Update of L is performed by using Gaussian prior distribution functions and Z is by Bayesian nonparametric prior distribution function. Update process is continuing until noise is very negligible where noise is calculated by using Gaussian likelihood functions. Finally, each column of rows of lego matrix, L represents which interaction belongs to which legos or modules.

Experimented conducted. Qi *et al.* [2009] performed a comparison between INMF, classical NMF (non-negative matrix factorization) and ILGM (Infinite Linear Gaussian Model) in a condition sensitive PIN for different number of response networks. As well as they implemented the INMF in Human Functional Genomic Data (HFGD) and Cancer Map Data (CMD) and compared the result with Murali *et al.* [2008] method. They also applied reverse engineering from legos to response networks.

Results. After performing the comparison between INMF, classical NMF and ILGM for different response networks of a PIN, the authors claim that, their proposed INMF gives better result in comparison to other methods. Sometime, classical NMF does not perform, where INMF performs accurately. After experimenting HFGD and CMD by INMF and comparing result, the authors find and claim that INMF give same result which solves the shortcoming of Murali *et al.* [2008] proposed method. According to their claim, reverse engineering is successful for the INMF method.

Claims made by the authors. Authors state that their proposed algorithm makes the following contributions-

- Successful implementation of factorization methods for clustering condition sensitive PIN by using the concept of network legos.
- INMF automatically learns the dimension of the factorized matrices.
- INMF solves the problems of classical NMF for clustering condition sensitive protein interaction network.
- After performing INMF, if reverse engineering is applied, always original response networks are found.
- It solves the shortcoming of the method proposed by Murali *et al.* [2008] proposed method.

Citations. There are no specific references to this paper by other researchers in this survey.

4.9. Tu *et al.* 2010

Bibliographic entry. TU, S., CHEN, R., AND XU, L., 2010, A binary matrix factorization algorithm for protein complex prediction, In *Proceedings of International Workshop on Computational Proteomics, 2010*, Hong Kong, China, 9, 18-25.

Problem. Protein interaction networks (PINs) play key roles in the biological processes including cell cycle control, differentiation, protein folding, signalling, transcription, translation and transport. To understand and identify the biological process, it is essential to find out the protein complexes in protein interaction network. Protein complexes are groups of protein that densely interact with each another. Protein complex has dynamic nature which is difficult to accurately predict the protein complexes from PIN.

Previous work. The authors do not refer to previous work.

Shortcomings of previous work. Due to no previous work, there is no shortcoming.

New idea/ algorithm. Tu *et al.* [2010] proposed a new method for identifying the protein complexes by clustering the PIN using a matrix factorization method. Their proposed method is known as binary matrix factorization method (BMF). According to their method, the PIN is converted into the symmetric adjacency matrix $X_{n,N}$ where $x_{i,j} \in \{0, 1\}$. After generating X matrix, proposed method factorized into two matrices A and Y using Bayesian Ying-Yang harmony learning function (BYY). BYY function can automatically predict the number of clusters. Finally, the matrix, Y gives the clustering information. In their method, they also modify the evaluation criteria- sensitivity, positive predictive value, accuracy and separation.

Experimented conducted. Tu *et al.* [2010] had performed an experiment on protein interaction network from the MIPS database by their proposed method. They also implemented the algorithm on the

altered graphs by randomly adding and deleting edges. Finally they compared the result with the results of two existing methods: Markov Cluster Algorithm (MCL) and the Spectral Clustering (SC) for clustering PIN.

Results. In the first case, the authors claim that, BMF gives more accurate clustering information than MCL and SC algorithm which is 0.91. In second experiment, it is claimed that, BMF is better than MCL which is better than SC.

Claims made by the authors. Tu *et al.* [2010] state that their proposed BMF BYY method can accurately cluster the protein interaction network without prior knowledge of the number of clusters. Secondly, they also claim that, BMF BYY does not depend on any parameter or threshold like MCL or SC.

Citations. There are no specific references to this paper by other researchers in this survey.

4.10. Wang *et al.* 2008

Bibliographic entry. WANG, R.-S., ZHANG, S., WANG, Y., ZHANG, X.-S., AND CHEN, L., 2008, Clustering complex networks and biological networks by non-negative matrix factorization with various similarity measures, In *Neurocomputing*, 72, 134-141.

Problem. Protein complexes play important role on prediction of the functionalities of cells which helps to identify the cause of disease of living being. To find out the protein complexes, it is mandatory to cluster the protein interaction network into several clusters according to their functionalities.

Previous work. The authors do not refer to any previous work.

Shortcomings of previous work. No shortcoming of previous because of non-presence of previous work.

New idea/ algorithm. Wang *et al.* [2008] developed an algorithm for clustering protein interaction network into several protein complexes according to their functionalities. They use classical non-negative matrix factorization method (NMF). Before factorization, they generate the similarity matrix, S for protein interaction network using four different methods- diffusion kernel feature, shortest path based, Hamming distance based and adjacency matrix based method. According to their proposed method, the algorithm factorizes the similarity matrix, S into two non-negative matrices W and H. The factorization process continues until the Euclidean distance between S and W, H is very negligible which is less than 10^{-15}. Finally they check the accuracy of the clustering process by using the modularity measure, Q which is completely designed by the authors.

Experimented conducted. They evaluated the proposed method by using a large scale protein interaction network of yeast *S. cerevisiae* which consists of 1257 proteins and 6835 interactions. They separately determined the similarity matrix using four different methods and evaluated clustering using their proposed modularity measure.

Results. The authors claim that, for diffusion kernel feature method, NMF can accurately identify the protein complexes from protein interaction network. Whereas, shortest path based method also gives

accurate result but it takes exponential time for determining the protein complexes. Other two methods cannot give accurate results but these results are near to accurate result.

Claims made by the authors. The authors (Wang *et al.* [2008]) state that their proposed method achieves following factors-

- Their proposed method automatically detects the proper number of clusters by implementing or repeating the algorithm for *k* times.
- This method can be used in other networks- social networks.
- Modularity measure can be used in any clustering algorithm for checking its accuracy.

Citations. [Wang *et al.* 2010]

REFERENCES

[BARASH 2010] BARASH,Y., BLENCOWE, B.J. AND FREY, B.J., 2010, Model-based detection of alternative splicing signals, In *Oxford Journal of Bioinformatics*, 26 **(12)**, i325-i333.

[BRUNET 2004] BRUNET, J-P., TAMAYO,P., GOLUN,T.R., AND MESIROV, J.P., 2004, Metagenes and molecular pattern discovery using matrix factorization. In *Proceedings of Natural Academic Science*, USA, 101(12), 4164–4169.

[CHANGOYEN 2006] CHAGOYEN, M., CARMONA-SUEZ, P., GIL, C., CARAZO, J. M., AND PASCUAL-MONTANO, A., 2006, A literature based similarity metric for biological processes, In *BMC Bioinformatics*, 6, 363-375.

[CHUA 2011] CHUA, H.N., AND ROTH, F.P., 2011, Discovering the Targets of Drugs Via Computational System Biology, in *Journal of Biological Chemistry*, 286, 23653-23658.

[DESHPANDE 2010] DESHPANDE, R., SHARMA, S., VERFAILLIE, C. M., HU, W.-S., AND MYERS, C. L., 2010, A Scalable Approach for Discovering Conserved Active Subnetworks across Species, In *PLoS Computational Biology*, 6(12), 1-18.

[DUECK 2005] DUECK D., MORRIS, Q. D., AND FREY, B. J., 2005, Multi-way clustering of microarray data using probabilistic sparse matrix factorization, In *Oxford Journals Bioinformatics*, 21(1), 144-151.

[FOGEL 2007] FOGEL, P., YOUNG, S., S., HAWKINS, D., M., AND LADIRAC, N., 2007, Inferential, robust non-negative matrix factorization analysis of microarray data, In *Oxford Journal of Bioinformatics*, 23(1), 44-49.

[FRIGYES 2008] FRIGYESI, A., AND HÖGLUND, M., 2008, Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes, In *Cancer Informatics*, 6, 275-292.

[GAO 2005] GAO, Y., AND CHURCH, G., 2005, Improving molecular cancer class discovery through sparse non-negative matrix factorization, In *Oxford Journals Bioinformatics*, 21(21), 3970-3975.

[GAUJOAUX 2010] GAUJOAUX, R., AND SEOIGHE, C, 2010, A flexible R package for nonnegative matrix factorization, In *BMC Bioinformatics*, 11, 367-375.

[GUINNEY 2010] GUINNEY, J., FEBBO, P., MAGGIONI, M., AND MUKHERJEE, S., 2010, Multiscale factor models for molecular networks, In *Joint Statistical Meeting, Canada,* August, 574-590.

[HUGHES 2000] HUGHES,T.R., MARTON,M.J., JONES,A.R., ROBERTS,C.J., STOUGHTON,R., ARMOUR,C.D., BENNETT,H.A., COFFEY,E., DAI,H., AND HE,Y.D., 2000, Functional discovery via a compendium of expression profiles. In *Cell*, 102, 109–126.

[HYOUNG 2011] HYOUNG, K.,K., THU, V.T., HEO, H-J, KIM, N., HAN, J., 2011, Cardiac proteomic responses to ischemia--reperfusion injury and ischemic preconditioning, In *Expert Review of Proteomics*, 8(2), 241-261.

[JEON 2011] JEON, J., JEONG, J.,H., BAEK, J-H., KOO, H-J., PARK, W-H., YANG, J-S., YU, M-H., KIM, S., AND PAK, Y., K., 2011, Network Clustering Revealed the Systemic Alterations of Mitochondrial Protein Expression, In *PLoS Computational Biology*, 7(6), i1002093-i10020104.

[JIAO 2011] JIAO, Q.-J., ZHANG, Y.-K., LI, L.-N., AND SHEN, H.-B., 2011, BinTree Seeking: A Novel Aproach to Mine Both Bi-Sparse and Cohesive Modules in Protein Interaction Networks, In *PLoS Computational Biology*, 6(11), 646-657.

[KIM 2007] KIM, H., AND PARK, H., 2007, Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, In *Oxford Journal of Bioinformatics*, 23(12), 1495-1502.

[KIM 2010] KIM, W., LI, M., WANG, J., AND PAN, Y., 2010, Biological network motif detection and evalution, In *Proceedings of The 2010 International Conference on Bioinformatics and Computational Biology*, Las Vegas, NV, USA, 5, 1-13.

[KIM 2011] KIM, Y., KIM, T.-K., KIM, Y., YOO, J., YOU, S., LEE, I., CARLSON, G., HOOD, L., CHOI, S., AND HWANG, D., 2011, Principal network analysis: identification of subnetworks representing major dynamics using gene expression data, In *Oxford Journals Bioinformatics*, 27(3), 391-398.

[KALTENBACH 2012] KALTENBACH, H-M., STELLING, J., 2012, Modular Analysis of Biological network, In *Advances in System Biology*, 736(1), 3-17.

[LEE 1999] LEE, D.D. AND SEUNG, S., 1999, Learning the parts of objects by non-negative matrix factorization. In *Nature*, 401, 788–791.

[LEI 2011] LEI, X., 2011, Codimensional matrix pairing perspective of BYY harmony learning: hierarchy of bilinear systems, joint decomposition of data-covariance, and applications of network biology, In *FRONTIERS OF ELECTRICAL AND ELECTRONIC ENGINEERING IN CHINA*, 6(1), 86-119.

[LI 2007] LI, H., SUN, Y., AND ZHAN, M., 2007, The discovery of transcriptional modules by a two-stage matrix decomposition approach, In *Oxford Journal of Bioinformatics*, 23 (4), 473-479.

[LI 2008] LI, H., AND ZHAN, M., 2008, Unraveling transcriptional regulatory programs by integrative analysis of microarray and transcription factor binding data, In *Oxford Journal of Bioinformatics*, 24 **(17)**, 1874-1880.

[MARCOTTE 1999] MARCOTTE,E.M., PELLEGRINI,M., NG,H.L., RICE,D.W., YEATES,T.O. AND EISENBERG,D., 1999, Detecting protein function and protein– protein interactions from genome sequences. In *Science*, 285, 751–753.

[MEJÍA-ROA 2008] MEJÍA-ROA, E., CARMONA-SAEZ, P., NOGALES, R., VICENTE, C., VÁZQUEZ, M., YANG, X.Y., GARCÍA, C., TIRADO, F., AND PASCUAL-MONTANO, A., 2008, bioNMF: a web-based tool for nonnegative matrix factorization in biology, **In** *Oxford Journals of Nucleic Acid Research*, 36(2), W523-W528.

[MILO 2002] MILO, R., SHEN-ORR, S, ITZKOVITZ, S, KASHTAN, N, CHKLOVSKII, D, AND ALON, U, 2002, Network Motifs: Simple Building Blocks of Complex Networks. In *Journal of Science*, 298 (5594), 824-827.

[MÜLLAR 2008] MÜLLAR, F. J., LAURENT, L. C., KOSTKA, D., ULITSKY, I., WILLIAMS, R., LU, C., PARK, I.-H., RAO, M. S., SHAMIR, R., SCHWARTZ, P. H., SCHIMDT, N. O., AND LORING, J. F., 2008, Regulatory networks define phenotypic classes of human stem cell lines, In *Nature*, 455, 401-406.

[MURALI 2008] MURALI, T.M., AND RIVERA, C.G., 2008, Network Legos:Building Blocks of Cellular Wiring Diagram, In *Journal of Computational Biology*, 15, 829-843.

[NEWMAN 2010] NEWMAN, A. M., AND COPPER, J. B., 2010, AutoSOME: a clustering method for identifying gene expression modules without priori knowledge of cluster number, In *BMC Bioinformatics*, 11, 1-15.

[OCHS 2010] OCHS, M., 2010, Knowledge-based data analysis comes of age, In *Oxford Journal of Briefings in Bioinformatics*, 11(1), 30-39.

[PASCUAL-MONTANO 2005] PASCUAL-MONTANO, A., CARMONA-SÁEZ, P., PASCUAL-MARQUI, R. D., TIRADO, F., AND CARAZO, J.M., 2005, Two-way clustering of gene expression profiles by sparse matrix factorization, In *Proceedings of IEEE Computational Systems Bioinformatics Conference Workshop*, 5-6.

[PINKERT 2010] PINKERT, S., SCHULTZ, J., AND REICHARDT, J., 2010, Protein Interaction Networks-More Than Mere Modules, In *PLoS Computational Biology*, 6(1), 1371-1383.

[QI 2009] QI, Y., DING, N., RIVERA, C. G., AND MURALI, T. M., 2009, Learning Network Legos by Infinite Non-Negative Matrix Factorization, In *World Congress of Pain Clinicians – Proceedings*, 1-12.

[SCHREIBER 2007] SCHREIBER, A.W., AND BRAUMANN, U, 2007, A framework for gene expression analysis, In *Oxford Journal of Bioinformatics*, 23 (2), 191-197.

[SHARAN 2007] SHARAN, R., ULITSKY, I., SHAMIR, R., 2007, Network-based prediction of protein function. In *Molecular Systems Biology*, 3, 88-97

[TU 2011] TU, S., CHEN, R., AND XU, L., 2011, A binary matrix factorization algorithm for protein complex prediction, In *Proceedings of International Workshop on Computational Proteomics, 2010*, Hong Kong, China, 9, 18-25.

[WANG 2008] WANG, R.-S., ZHANG, S., WANG, Y., ZHANG, X.-S., AND CHEN, L., 2008, Clustering complex networks and biological networks by non-negative matrix factorization with various similarity measures, In *Neurocomputing*, 72, 134-141.

[WANG 2010] WANG. J., LI, M., DENG, Y., AND PAN, Y., 2010, Recent advances in clustering methods for protein interaction networks, In *BMC Genomics,* 11, 10-28.

[WERNICKE 2006] WERNICKE, S.,2006, Efficient Detection of Network Motifs. In *IEEE/ACM Trans Computational Biology Bioinformatics*, 3(4), 347-359.

[YANG 2005] YANG, C. F., YE, M., AND ZHAO, J., 2005, Clustering Based on Nonnegative Sparse Matrix Factorization, In *Proceedings of International Conference on Natural Computation 2005*, Berlin Heidelberg, Germany, 1, 557–563.

[YU 2010] YU, S., TRANCHEVENT, L.-C., MOOR, B. D., AND MOREAU, Y., 2010, Gene prioritization and clustering by multi-view text mining, In *BMC Bioinformatics*, 11, 28-49.

[ZAREI 2009] ZAREI, M., IZADI, D., AND SAMANI, K. A., 2009, Detecting overlapping community structure of networks based vertex-vertex correlations, In *Journal of Statistical Mechanics: Theory and Experiment*, 1013-1029.

[ZHANG 2011] ZHANG, S., LI, Q., LIU, J., ZHOU, X, J, 2011, A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules, In *Oxford Journal of Bioinformatics*, 27(13), i401-i409.

[ZHANG 2011a] ZHANG, K., LIU, Y., YANG, J.,Y., ARABNIA, H.R., NIEMIERKO, A., GHAFOOR, A., LI, W., AND DENG, Y., 2011, From genes to networks: in systematic points of view, In *BMC Systems Biology*, 5(3), 509-511.