Effectiveness of Social Networks for Studying Biological Agents and Identifying Cancer Biomarkers

Ghada Naji, Mohamad Nagi, Abdallah M. ElSheikh, Shang Gao, Keivan Kianmehr, Tansel Özyer, Jon Rokne, Douglas Demetrick, Mick Ridley, and Reda Alhajj

Abstract Social networks form phenomena that exist and evolve; they are dynamic. These phenomena have been realized and studied by the anthropology and sociology research communities since 1930. However, the recent rapid development in information technology and the internet has increased the interest in social networks and as a model they have been adapted to more applications and domains. Though researchers first studied social networks of humans, for our study described in this chapter we argue that genes and proteins act collaboratively and exist in communities analogous to humans, animals, insects, etc. They complement each other and collectively achieve specific tasks where some would have major roles appearing upfront and others may play minor background roles. However, molecules turn into aggressive actors when their internal structure is augmented; consequently, they may deviate from their target, change camp, and disturb other molecules leading to disaster. Such mutations may be uncontrolled and unintentionally occur

G. Naji
Department of Biology, Faculty of Sciences III, Lebanese University, Tripoli, Lebanon

M. Nagi · M. Ridley
Department of Computing, School of Computing Informatics and Media, University of Bradford, Bradford, UK

A.M. ElSheikh · S. Gao · K. Kianmehr · T. Özyer · J. Rokne
Computer Science Department, University of Calgary, Calgary, AB, Canada

D. Demetrick
Departments of Pathology, Oncology and Biochemistry & Molecular Biology, University of Calgary, Calgary, AB, Canada

R. Alhajj (✉)
Computer Science Department, University of Calgary, Calgary, AB, Canada
and
Department of Computer Science, Global University, Beirut, Lebanon
and
Department of Information Technology, Hellenic American University, NH, USA
e-mail: alhajj@ucalgary.ca

U.K. Wiil (ed.), *Counterterrorism and Open Source Intelligence*, Lecture Notes in Social Networks 2, DOI 10.1007/978-3-7091-0388-3_15, © Springer-Verlag/Wien 2011

inside a body, or they may be intentional and controlled by humans to serve one of two purposes, treatment or bioterrorism. In other words, mutation in the molecules (genes) can lead to a change in behavior. This may lead to good or bad effect, e.g., recovery from illness or diseases that may severely affect the body causing disability or death. Once mutated outside the body, molecules may turn into harmful biological weapons of mass destruction. The latter process does not require sophisticated equipment and hence is extremely dangerous with the uprising global terrorism activities. Bioterrorism is therefore a serious concern for humanity. One could say that mutated biological agents outside the body once misused could be way more dangerous than mutated molecules within the body. In this chapter, we will elaborate on bioterrorism and its consequences; we will also propose a model to study social networks of genes within the body leading to the identification of disease biomarkers.

1 Introduction

We witness the establishment of social communities as part of our daily life. Since the dawn of humans on earth they have tended to come together and socialize leading to social communities that evolved into the currently known nations. The existing communities could dynamically change by having new members joining and some of the existing members leaving. For example, it is very common for people to move from one country to another and even from one location to another within the same country. People may change political parties. Employees may change employers and even departments within the same organization. This natural phenomenon has been studied by researchers in anthropology and sociology from early in the twentieth century. Social network research today is multidisciplinary requiring expertise from anthropology, sociology [22], behavioral science, psychology, statistics, mathematics, computer science, etc. Finding a balance between these domains of knowledge is by itself challenging and requires significant effort. We further argue that social network methodology is rich enough to successfully serve a variety of applications, including web mining [33, 51], web services, personalized search, patient monitoring, biological networks [38], disease biomarker detection, outlier detection, team work, wildlife monitoring, document and text analysis, database design and partitioning, traffic monitoring, homeland security, among others [25, 30, 46, 50, 52, 57, 58]. The analysis of social networks leads to valuable discoveries that may have essential social and economic impact. From the social perspective, the discoveries may highlight terrorism groups [25, 52, 57], common hobbies, family relationships, social functions, occupations, friendship, disease biomarkers, etc. From the economic perspective, the analysis may lead to certain target customer groups, the development of drugs, exceptional weather conditions, unusual trends in the stock market, etc.

For long time, the social network methodology was dependent on manual processes and hence concentrated on small networks involving mostly humans

[4, 9–11, 30, 46, 47, 58, 65]. The main theme was to analyze human interactions in specific environments in order to discover key persons, groups, etc. However, the recent development in information technology and the wide spread of the world wide web have highly influenced and shaped the research in social networks [33, 53, 64]. People are joining online social networks and socialize on the web. There is a clear shift from real to virtual life. Furthermore, researchers in different fields have started to realize the effectiveness of social network models in a variety of new applications.

1.1 The Social Network Model

The simplest model of a social network is a set of actors linked by certain type of relationship [11]. Actors (interchangeably called individuals) such as words, pieces of code, items in stock, routers in a network, employees, physicians, patients, trends in weather behavior, animals, humans, insects, genes, proteins, drugs, web pages, etc., can be grouped together into communities. Graphs form most attractive representation of social networks. The edges in a graph may represent social interactions, organizational structures, physical proximity, or even more abstract interactions such as hyperlinks or similarity, among others. The study of social networks has been extensively realized in the research community as we see more conferences emerging and more journals starting dedicated to the computing aspects of mining and analysis of social networks. This will then further interest in the field from more researchers who are willing to contribute to social network research.

If it is possible, though challenging to figure out social communities of abstract items like pieces of code or physical objects like flowers. Realizing the social communities of dynamic and moving objects like stars, humans and animals, e.g., [25] might require less effort. The actors in one community share something which distinguishes their community from the other existing communities. The overall structure of communities is in general hierarchical. For instance, humanity forms a single huge community sharing minimal characteristics of being human. This single huge community is composed of subcommunities that have further distinguishing characteristics which recursively split into subcommunities by considering further characteristics until we end up having each individual family as a community. In the same manner in a cell all the gene products form one large community which could be distributed into smaller communities by considering more features of the gene products. One gene product may belong to two or more communities and it would have a specific role within each community. Finally, social communities are mostly dynamic (whether modeled as stationary or dynamic actors) [9]; and hence the analysis of evolving communities is more demanding to cope with the dynamics of the changing network [9]. Data mining techniques have been effectively used to analyze and study social networks. The social network modeling framework described in this chapter is heavily based on data mining techniques.

1.2 Effectiveness of Data Mining

Data mining techniques are attractive for studying and analyzing huge amounts of data from different points of view and for summarizing the data into useful information through which knowledge hidden within the data can be extracted. In general, data mining techniques are used for social network analysis where people who have similar social profiles are grouped together in social communities [47]; for instance, people may be grouped into small communities based on their income and exchanged emails. Data mining also involves the process of finding correlations in the data. For instance, data mining is used by companies to target their promotions to certain individuals based on the individuals' purchase history; it can also be used in market prediction where it is possible to predict the market status based on the history. One of the data mining techniques successful for such applications is clustering, which has been extensively studied in the literature [41].

Clustering is simply the process of grouping a given set of instances into classes/clusters such that instances in each class are similar and instances in different classes are dissimilar. Clustering may lead to different grouping patterns based on the set of features considered in the process. For instance, people may be classified differently based on any combination of age, sex, weight, nationality, address, etc.

Clustering may be supervised when the classes are known in advance; it is unsupervised when the classes are yet to be determined. The former is called classification and involves two major steps: model construction by using training data to build a model which is tested using remaining unseen data instances in order to determine its accuracy. On the other hand, clustering as unsupervised learning process does not rely on predefined classes and training examples while classifying the data objects. It is the duty of the clustering approach to decide on the classes. Clustering has been successfully applied to different domains, including customer classification, web data, financial data, spatial data, explanatory pattern-analysis and classification, grouping, decision making, document retrieval, gene expression data analysis, image segmentation, among others. Objects are clustered based on some similarity measures, namely homogeneity and separateness. Determining the number of clusters is a problem common to most of the existing algorithms. Usually the number of clusters is user defined. The best number of clusters can be defined by running the algorithm for different values of the number of clusters, and then choose the most appropriate value based on majority voting of some validation indexes. We further argue that even clustering algorithms that claim to run without the number of clusters predefined (like DBScan) still require specifying some parameters that guide the clustering process; tuning and deciding on the latter parameters is equally challenging as deciding on the best number of clusters. In this study, we use k-means clustering algorithm, which requires the number of clusters as input. We decide on the number of clusters by considering the result from applying frequent pattern mining as described in the sequel.

1.3 Realizing Molecule Interactions as Social Network

In this chapter, we concentrate on social networks of molecules with particular emphasis on their mutations, which we define as a significant change in their natural function, since careful attention has to be given to the case of beneficial mutations and precautions have to be taken to stop/avoid/prevent the consequences of bad mutations. We first cover bioterrorism and then describe how the social network model fits well for identifying disease biomarkers. We mainly concentrate on social communities of gene products. A gene can be defined as a string of nucleotides from the alphabet $\{A, T, G, C\}$, where each letter represents a chemical compound. Further details of genetics are out of the scope of this chapter. These genes are the functional segments of DNA that encode mRNA which generates the proteins. Genes form the basic unit for heredity and genetic material transformation. Surprisingly, recent research has revealed that less that 5% of the whole DNA sequence in human involves functional genes. The number of genes differs among organisms; for example, a human genome encodes approximately 20,500 genes, and yeast has around 6,000 genes.

We argue that genes, RNA and proteins form communities. For instance, it is known that some proteins collectively function to maintain cellular protein conformation during stressful proteotoxic insults. Signaling pathways such as that regulating cell division can be thought of as a community. Likewise, investigating whole microbial communities instead of individual micro-organisms could guide scientists to answer fundamental questions such as how ecosystems respond to climate change or pollution. Actually, all the interactions between different organisms within an ecosystem must be taken into consideration in order to assess the environmental impact on microbial communities. This is not possible by following traditional approaches such as those which examine changes in gene expression of individual microbial cells. It is more realistic to investigate and analyze the gene expression of a whole community at once.

We describe a model for identifying disease biomarkers by mining and analysis of social communities of genes and/or gene products. The discoveries from the study described here will support and extend our previous findings related to biomarkers [6, 7]. The social network of gene expressions is watched over time in order to study the behavior of gene products: how they change camps to assume different functions, and how they affect other molecules within the cell. We employ different data snapshots related to the same set of gene products and collected from the same patients at different time points. A social network of the gene products is derived based on each snapshot. The different social networks are analyzed to identify and compare the expressed genes. The outcome will lead to highlight the benefit or thread caused by the analyzed genes.

A social network is derived by employing three perspectives. The first perspective is frequent pattern mining where gene expression and samples represent items and transactions, respectively. We utilize maximal-frequent sets of genes. In our discussion of the machine learning applications, the set of genes is frequent when it

satisfies a minimum frequency threshold by having the genes concurrently expressed in a certain number of samples. It is closed if none of its supersets has the same frequency in the analyzed snapshot; it is maximal-closed if in addition to being closed, none of its supersets is frequent. The second perspective is the k-means clustering algorithm where the number of clusters k is set to the number of derived maximal-closed frequent sets of genes. The coexistence of two genes in the frequent patterns is used to decide on whether to have them connected in the social network, where we use the term "gene" simply as a label for the "gene product". When the two genes exist in the same cluster, the weight of the connection is set to the number of frequent patterns hosting the two genes concurrently; otherwise, the weight is set to the reciprocal of the latter value. The weight is adjusted further based on a third perspective which considers the data snapshot directly as discussed later in Sect. 2. The constructed social networks from all the data snapshots are to be investigated further to derive social communities that lead to better identification of biomarkers, where we define "biomarker" to be a "biological property", and not in the conventional sense of a biological predictor of clinical behavior. Though our results reported in this chapter are related to cancer biomarkers, the approach is general enough to be applied to identify biomarkers of other diseases should the corresponding data be available for modeling and analysis. The latter study has been left as future work. Finally, support vector machines (SVMs) are used to classify the given samples based on the identified biomarkers.

The remaining parts of this paper are organized as follows. Section 2 is an overview of the social networks methodology. Bioterrorism is briefly discussed in Sect. 3. Section 4 summarizes the literature on disease biomarkers with emphasis on cancer biomarkers. Section 5 presents the proposed framework to identify social communities of genes and their analysis. Section 6 reports the experimental results. Section 7 is conclusions and future work.

2 Basic Methodology for Social Network Analysis

The simplest model of a social network involves a homogeneous set of actors which are connected based on certain criteria. For instance, in pharmacology the actors could be drugs, and two actors are connected if it is possible for them to appear together in the same prescription. In a university environment the actors could be students and two actors are connected if they are enrolled together in at least 3 courses during the current semester. Analyzing the first network will lead to communities of drugs that are used together and analysis of the second network will lead to communities of students who study together during the current semester. The identified students may then have the potential to be enrolled in the same courses in the future. The latter piece of information may be valuable in developing a recommendation system for students to select courses such that a set of courses is recommended to students who belong to the same community. The links may reflect either binary relationships (a missing link indicates no relationship) or weighted

connections to indicate the strength or degree of the relationship (which may be negative or positive). A graph that represents a social network is generally called a *sociogram*.

It is also possible to have more than one set of actors. The most common trend is to have two disjoined sets of actors like mRNA and proteins. Then, two actors can be connected if and only if they do not belong to the same set and they satisfy the criteria employed to derive the links, like a mRNA and a protein may be connected to show that the mRNA translates into the protein. This model is better represented using bipartite graph. For instance, drugs and diseases may be two sets of actors such that a link between a drug and a disease indicates the usage of the drug for treating the disease. Another example involves students and courses as two sets of actors such that a student is linked to every course he/she is enrolled in. The first network could be analyzed to discover the most important drugs used in treating most of the diseases. The second network leads to valuable information like identifying the most crowded courses.

The number of actors' groups in the model specifies the degree of the mode for the social network. The two versions described above are known as one-mode and two-mode social networks. These are the two most common settings. A two-mode social network is constructed for the third perspective that we have considered for enriching the social networks of genes tackled in this study. The two mode network is derived directly from the data snapshot where the two sets of actors are the genes and the sample. A link is present in the social network between gene g and sample s if and only if g is expressed in s.

It is possible to derive two one-mode networks from a two-mode network by applying a process known as folding, which operates directly on the adjacency matrix that corresponds to the social network. Assume the adjacency matrix for the two-mode network of genes and samples is constructed so that rows represent genes and columns represent samples. Multiplying the adjacency matrix by its transpose will produce a new adjacency matrix for a one-mode social network where the actors are the genes. The links between the actors in the latter network reflect the influence of the samples in the original two-mode network. On the other hand, multiplying the transpose by the original adjacency matrix will lead to a one-mode network where the actors are the samples. For our study, we are interested in the one-mode network of genes. The outcome from this perspective will influence the social network derived by considering the frequent pattern mining and clustering.

Once a model is constructed, social network analysis can be applied to knowledge discovery in the model. The analysis is classified into two categories, individual centric and group centric. The former analysis starts the from a single actor as a key player in the network and studies its neighborhood. The latter on the other hand considers the whole group at once, and studies the interactions within the group as a whole. The choice of which approach to follow depends on whether the interest is in studying the whole group at once or on identifying individual leaders and use them to influence the whole group. For instance, we may study the correlation between different terms in a book by analyzing how frequent the terms occur together in the same chapter of the book, and we may even fine-tune

the analysis to consider sections or paragraphs instead of chapters. We may also study the social communities of actors in a play by analyzing how often they come together in the same scene or sketch.

Formally, a social network is represented as a graph with weights, denoted by $G = (V, E, W)$, where V is the set of actors in the network, E is the set of edges connecting the actors to indicate relationship, and W is a $|V| \times |V|$ matrix of real values representing the weights of the different links between the $|V|$ actors. Normally, W is neglected in a social network with binary relationships between the actors. For totally connected graphs, $E = (V \times V)$, but in general $E \subseteq (V \times V)$. Once the graph is constructed, different metrics are employed in the analysis for knowledge discovery. The most commonly used metrics include density, centrality [19, 31], and cliques' identification (where every node is connected to every other node in a clique). Finding complete cliques may not be possible in real life; thus we try to find the maximally connected groups which are very close to form a clique.

Density is measured as the ratio of the number of edges in E over the total number of edges in a complete graph (which is $|V| \times (|V| - 1)$ for a complete directed graph). Density gives an indication about cohesion. Density may also be applied to subgroups by considering subgraphs instead of the whole graph. Within a subgraph, density measures in the subgraph the ratio of the actual connections to all possible connections. Measuring the density between two groups (subgraphs) is also possible.

Centrality generally refers to the importance of individual actors in a given group. Centrality may be also measured in terms of degree, betweenness [18, 40, 60], closeness and eigenvector. Degree is a simple measure realized for actor a as the number of actors connected to a divided by the total number of actors minus one (i.e., $\frac{degree(a)}{|V|-1}$); degree is distinguished as in-degree and out-degree of each node in directed graphs. In case graph G is weighted, any of the three values in-degree weight, out-degree weight, or gain in weight could be used to measure the degree of centrality. For actor a, and by considering each other actor i, the latter three weighted measures are computed as:

$$in\_degree(a) = \sum_i (W_{ia}). \tag{1}$$

$$out\_degree(a) = \sum_i (W_{ai}). \tag{2}$$

$$gain\_in\_weight(a) = \sum_i (W_{ia} - W_{ia}). \tag{3}$$

Anthonisse [8] and Freeman [35] independently introduced betweenness as a measure of centrality for the analysis of social networks which only considers the shortest paths in the graph. It refers to how a given actor could be considered as the hub of the network and this is determined by the number of shortest paths that pass via the given actor. In other words, other actors do not have direct link and must

communicate via the given actor. Formally, let $d_{i,j}$ be the shortest distance between two actors i and j, $\sigma_{i,j} = \sum_{i}^{|V|} \sum_{j}^{|V|} (d_{i,j})$ is the set of shortest paths between all pairs of actors i and j in the social network, and $\sigma_{i,a,j} = \sum_{i}^{|V|} \sum_{j}^{|V|} (d_{i,a} + d_{a,j})$ is the set of shortest paths that pass via actor a and connect i to j, i.e., $\sigma_{i,a,j} \subseteq \sigma_{i,j}$; the betweenness of actor a is computed as:

$$Betweenness(a) = \frac{\sigma_{i,a,j}}{\sigma_{i,j}}. \tag{4}$$

Betweenness may be also measured using Bavelas–Leavitt index [12], which is the reciprocal of (4). The Bavelas–Leavitt index of centrality for a given actor a, denoted $BL(a)$, is therefore computed as:

$$BL(a) = \frac{\sigma_{i,j}}{\sigma_{i,a,j}}. \tag{5}$$

By considering connected actors (whether directly or indirectly), closeness can be regarded as a measure of how long it will take information to spread from a given actor to other reachable actors in the network. It is the ratio of the number of links that an actor must follow to visit each of the other reachable actors in the network and the actor is considered more central when its ratio of closeness is closer to 1, i.e., it is directly connected to all reachable actors. Once individual groups are identified, it is possible to study their overlapping members in order to figure out the interactions between the different communities within the network. Having interrelated communities is preferred to isolated ones. This is equivalent to fuzzy clustering which is a more natural way of expressing membership for most real-world applications.

Eigenvector centrality measures the importance of an actor in a social network. It is computed based on the adjacency matrix A, where entry $A_{i,j} = 1$ if i and j represent adjacent (directly connected) actors in the social network and $A_{i,j} = 0$ otherwise. On the other hand, $A_{i,j} = W_{i,j}$ for weighted graphs. The eigenvector centrality measure for actor a, denoted e_a, is computed as:

$$e_a = \frac{1}{\lambda} \sum_{j=1, j \neq a}^{|V|} (e_j) = \frac{1}{\lambda} \sum_{j=1}^{|V|} (A_{i,j} \times e_j), \tag{6}$$

where λ is a constant called the eigenvalue when (6) is written in vector notation as $A\,e = \lambda e$. If there are n actors in the social network, then the adjacency matrix is of size $n \times n$. For such matrices, there are n eigenvalues and we are interested in the largest eigenvalue. The eigenvector e of the latter eigenvalue is used for the eigenvector centrality in the (6) measure for actor a.

For the study described in this chapter, we compute the number of times two genes occur together in the same maximal-closed frequent set of genes and the outcome is supported by the results from applying two other perspectives, namely

the outcome from k-means clustering and the outcome from folding the two-mode network of genes and samples. The framework is described next in Sect. 5.

3 Bioterrorism

Bioterrorism involves the aggressive and planned release of biological agents like viruses, bacteria, or other germs with the target to severely affect humans, animals, resources or the economy by causing illness or death in people, animals, or plants. The utilized biological agents are typically found in nature, but it is possible that they could be mutated or engineered to stimulate and/or increase their ability to spread into the environment and to cause damage as well as to make them resistant to current medicines.

Mutated biological agents can spread through the environment, including air, water, or food. Some agents used in bioterrorism, like the smallpox virus [1, 32, 42, 44, 49, 55, 56] can spread from person to person and other agents, like anthrax [24,34], can not. In his study Mayr [55] emphasized smallpox. He described many pox viruses and how they can be potential threats to human and animal life. The complications involved with post-vaccinal impairments are also discussed. Agosto [1] discussed the smallpox and the Variola viruses and he also elaborated on how these viruses may turn into epidemics causing panic and social unrest.

For a long time, humans have used in conflicts biological agents ranging from naive simple to recent sophisticated agents. For instance, it is claimed that back in the history the Assyrians poisoned the wells of their enemies with rye ergot. In the recent history, people used biological weapons to severely affect their enemies in a trial to go for quick victory. However, they never considered the consequences of having the effect of the biological weapons staying in the region after it is controlled.

Biological weapons were heavily used during World War I and World Ware II, e.g., [23]. However, there are tremendous efforts since 1970s to prevent the development of biological weapons. Unfortunately, the development of biological weapons is attractive to less developed countries as well as to terrorists [68] because biological agents are relatively easy and inexpensive to obtain or produce. They can be easily disseminated and they can cause widespread fear and panic beyond the actual physical damage they can cause.

The development in biological weapons has been much faster than the development of the medication to treat the affected casualities. This has led to epidemics in the targeted regions. However, over time, biological warfare became more complex and countries began to develop weapons which were much more effective on the targeted group, and much less likely to cause infection to the wrong party. One significant enhancement in development of biological weapons was the use of anthrax. Finally, for long time, bioterrorism has been used to target individuals and groups as well as the economy. For instance, there has been recently several incidences of delivering letters laced with infectious anthrax.

As far as the economy is concerned, there is no clear evidence whether the source of the infecting viruses is from the nature or manmade. For instance, the recent spread in viruses like the swine flu (influenza), bird flu and SARS has raised major concerns worldwide as people started speculations who could be the planner and what could be the goal. Major serious concerns could be raised after closer look and deeper investigation of the areas hit by the virus and this may lead to the belief that it is not a coincidence. One can imagine the SARS hitting in China where the economy is booming. The H1N1 hit in Mexico, one of the tourism attractions in North America. The foot-and-mouth disease virus affected the economy of UK in 2001 and 2007, in addition to a number of other countries since then though to a less extent. Finally, Blancou and Pearson [16] discussed the consequences of bioterrorism on the economy.

It is not possible at all to control the development and spread of biological weapons. Students studying genetics engineering are attractive targets for involvement in producing biological weapons. Accordingly, it is important to carefully select these students and to educate them how to use their knowledge and expertise solely to the benefit of the humanity. They should be armed with the knowledge that will guard them from being deceived by people and organizations which are mainly terrorists who may misuse their expertise and knowledge. As long as there are conflicts and competitions, people who feel themselves weaker and cornered like terrorists will preferentially choose easy to produce, transport and use harmful weapons, that is biological weapons. As long as there exist people who are willing to kill others, it will not be possible to eliminate biological weapons. Terrorists will always look for maximizing causalities and will not hesitate in using biological weapons once they get the ability to acquire them. In other words, ongoing efforts to use biological warfare has been more apparent in small radical organizations attempting to create fear in the eyes of large groups. Some efforts have only been partially effective in creating fear, due to the lack of visibility associated with modern biological weapon used by small organizations.

As long as we are not able to globally identify and get ride of all terrorists, it might be more feasible to educate people in two main directions. First, people should be warned not to be deceived by terrorists who try brainwash them and to turn them into dangerous individuals. Terrorists try to approach people emotionally by using different factors including poverty, political and ethnic discrimination, and religious speeches falsely manipulated to suit the propaganda of the terrorists. Thus, it is possible to avoid losing the fight against terrorism by concentrating on effective ways to handle the factors mostly misutilized by terrorists, by opening new job opportunities that will eliminate poverty, by appropriately dealing with all ethnic groups to address their concerns and meet their reasonable expectations and by spreading the right and true understanding of religion. Misunderstandings of religion has always created problems and conflicts. It is important to watch out for unknowledgeable scholars who play with the emotions of the youth and motivate them to turn into terrorist candidates. This issue is going out of control where it is easy for anyone to declare himself as a scholar and start spreading poison into the minds of the youth. Preventing this would be much easier than recovering

and cleaning up the mess created by false propaganda. However international coordination and collaboration will be needed to effectively tackle the problem, especially with the widespread of the internet based propaganda and TV channels. Equally important is to have people feel that they are equally treated and not discriminated against. Preferential treatment to certain groups is one of the main items in the propaganda of terrorists. Second, people should be educated how to protect themselves from biological weapons and how to be treated in case they are affected. The latter issue has been addressed to some extent by different groups. For instance, Bronze [21] describes potential vaccines and pharmaceutical strategies for either prevention or treatment of established infections. Blank et al. [17] discussed how to learn lessons from the anthrax attacks of 2001 in order to react better to future bioterrorism attacks. The authors also analyzed the effectiveness of the distribution of antibiotics before and after an attack. Binder et al. [15] commented on the importance of using medicine and science as well as public knowledge to defend against bioterrorism.

4 Related Work on Identifying Disease Biomarkers

Identifying disease biomarkers in general and cancer biomarkers in particular is an interesting research problem that has received the attention of a number of research groups who tackled the problem of determining the best biomarkers for different types of diseases like cancer, including leukemia, bladder, lung, prostate, liver, breast, etc. Devarajan [28] reported some biomarkers that could help in the early prediction of acute kidney injury. Sahab et al. [67] presented a good overview of biomarkers for different diseases including heart, rheumatoid arthritis, asthma and cystic fibrosis, in addition to several cancer biomarkers like prostate, breast, ovarian and lung.

Leukemia is one type of cancer which has been extensively studied in the literature. Golub et al. [39] may be considered as the first group who tried to distinguish between acute myelogenous leukemia (AML) and acute lymphocytic leukemia (ALL) based on gene expression data. They used a model of self-organizing maps in combination with a weighted voting scheme. They obtained a strong prediction of 29/34 samples in the test data using 50 genes. Furey et al. [36] applied SVMs to the AML/ALL data. Significant genes are derived based on a score calculated from the mean and standard deviation of each gene type. Tests are performed for the 25, 250, 500, and 1,000 top ranked genes. At least two test samples are misclassified in all SVM tests. Guyon et al. [75] also applied SVM, but with a different feature selection method called recursive feature elimination. For each iteration, the weights associated with the genes are recalculated by a SVM test, and genes with the smallest weights are removed. On 8 and 16 genes, the classification error on the test set is zero. Jinyan and Wong [48] used new feature selection method called emerging patterns. When they applied their method on the

AML/ALL data, they were able to identify one gene (zyxin), which was able to classify 31/34 of the samples.

Tseng et al. [72] studied collectively potential diagnostic biomarkers for four cancer types, namely liver, prostate, lung and bladder. They identified 99 distinct multi-cancer biomarkers in the comparison of all three tissues in liver and prostate and 44 in the comparison of normal versus tumor in liver, prostate and lung. They also found out that bladder samples have different list of biomarkers from the other three cancer types. Shariat et al. [69] studied whether the assessment of five previously characterized bladder cancer biomarkers (p53, pRB, p21, p27, and cyclin E1) could improve the ability to predict disease recurrence and cancer-specific survival after radical cystectomy in patients with pTa-3N0M0 urothelial carcinoma of the bladder.

Wagner et al. [74] discuss several classification-based approaches to finding protein biomarker candidates using protein profiles obtained via mass spectrometry; they assess the statistical significance of the discovered biomarkers; the target is to link the biomarkers to given disease states, and thus to narrow the search for biomarker candidates.

Sorensen and Orntoft [70] covered advances in biomarker discovery for prostate cancer by microarray profiling of mRNA and microRNA expression. The authors discussed limitations in the application of microarray-based expression profiling for identification of prostate cancer biomarkers and hence the strong need for promising biomarkers to enable more accurate detection of prostate cancer, improve prediction of tumor aggressiveness and facilitate discovery of new therapeutic targets for tailored medicine.

Toure and Basu [73] applied a neural network to cancer classification where 10 genes were used for classification purposes. The neural network was able to fully separate the two classes during the training phase. However, the classification of the test set samples did not achieve high accuracy since 15 samples were misclassified. In another work, Li et al. [54] applied GA/KNN method to the same data set where 50 genes were used for classification. GA/KNN correctly classified all training set samples, and all but one of the 34 test samples. Bijlani et al. [14] used independently consistent expression discriminator (ICED) for feature extraction. They could select 3 AML distinctors and 13 ALL distinctors, which were able to classify the training set without any errors; but one sample was misclassified in the test data. Zhang and Ke [79] used the 50 genes reported by Golub et al.[39] and applied SVM and central support vector machine (CSVM) for classification. Two misclassifications occurred using SVM, but no errors were reported when CSVM was used.

As the colon data set is concerned, Bicciato et al.[13] used Autoassociative Neural network model for classification. No sample was misclassified during the training session, while seven samples received wrong classification during the testing phase. In another study [76], Wang et al. used SVM for classification, 4 samples were misclassified. Tusher et al.[73] used two novel classification models. The first was combination of optimally selected self-organizing map, followed by fuzzy C-means clustering. And the second was the use of pair-wise fisher's linear discriminant. In the former model, 12% of the samples were misclassified and in the

latter model 18% of the samples were incorrectly classified. Several other studies investigated the colon data set, e.g., [14]. At least four misclassifications occurred in most of the studies carried out. Many other groups have demonstrated the power of fuzzy logic in microarray analysis, e.g., [66, 75]; but the main problem of how to choose the fuzziness parameter m has to be further investigated. Usually 2 is the preferred value for m. However, Dembele and Kastner [26] have shown that 2 is generally not appropriate value for all data sets. Our group has already reported interesting results regarding the better choice of the values of m [6, 7].

5 Identifying Social Communities of Genes by Frequent Pattern Mining, K-means and Network Folding

In this section, we describe our approach to analyzing gene expression data that has been employed to generate the social communities of genes, which are further analyzed to identify expression of key genes as disease biomarkers and we concentrate on cancer biomarkers for the data sets used in the testing. Given a set of actors which are the expressed genes in this particular study, our target is to find the links between them in order to establish the social network of the genes. This is possible by investigating the relationship between the different genes. Our approach is based on a unified framework that combines three perspectives, namely frequent pattern mining (maximal-closed sets of genes in particular), k-means clustering and folding of the two-mode social network that connects genes to samples. First, the frequent pattern mining process is employed to produce the initial adjacency matrix where two genes are linked to reflect the number of maximal frequent sets of genes that contain the two genes concurrently. Second, k-means clustering and the folding process are separately applied to adjust the weight from the result from the first perspective. Finally, we compute the average of all the values in the matrix. Based on the latter average we produce an adjacency matrix for all the genes by accepting two genes as adjacent if and only if their correlation value in the adjacency matrix is larger than the average. The adjacency matrix helps in finding communities of genes and the central gene within each community where the latter gene is considered as a disease biomarker.

5.1 Frequent Pattern Mining

Frequent pattern mining is one of the most attractive data mining techniques described in the literature [2,3]. It was initially developed as the first of two steps for market basket analysis and later it has been adapted to different interesting applications. In general, frequent pattern mining investigates the correlation between items within the transactions in a given database. So, any problem that can be modeled in terms of transactions and items could be classified among the applications of the frequent pattern mining framework. Fortunately, the social network of genes

problem investigated in this work could benefit from frequent pattern mining by considering the samples as transactions and the genes correspond to the set of items.

To understand the frequent pattern mining model, we provide a brief coverage in the rest of this section. Consider a database of transactions, such that each transaction contains a set of items, the frequent pattern mining process determines sets of items that satisfy a minimum support threshold specified mostly by a domain expert where t support of a set X is the percentage of transactions that contain all items in X. It is also possible to derive the minimum support threshold in an automated way by considering characteristics of the data. However, this is outside the scope of the work described in this chapter.

Market basket analysis is one of the first applications of frequent pattern mining [3]. Organizations that deal with transactional data aim at using the analysis outcome to decide on better marketing strategies, to design better promotional activities, to make better product shelving decisions, and above all to use these as a tool to gain competitive advantage.

Agrawal et al. [2] first introduced frequent pattern mining in 1993, and since then it is one of the topics most frequenly investigated by researchers in the data mining arena. During the past two decades, several research groups have provided solutions to this problem in many different ways, e.g., [37,45,62,63]. The time and space scalability of the developed approaches greatly vary based on their techniques to mine the investigated databases. They mainly differ in the number of database scans, and hence the time consumed by the mining process, as well as in the data structures they use, which are mostly main memory resident. All the latter mentioned performance related issues are outside the scope of this chapter because the main target is to determine the maximal-closed frequent sets of items regardless of the performance of the approach to be utilized.

A very naive and brute force approach for finding all frequent patterns from a particular database is to generate all possible patterns from the database, and then check the corresponding frequency of each pattern against the database. The problem with this approach is that there can be 2^n (where n is the number of items, genes in our study) candidate patterns to be checked, and it is not computationally or space efficient to determine the frequency of such huge number of patterns.

Over the past two decades, researchers in the area have come up with numerous frequent pattern mining algorithms in an attempt to efficiently solve the problem, e.g., [37,45,62,63]. Covering all frequent pattern mining approaches is outside the scope of this chapter; interested readers may refer to the literature for comprehensive coverage. In this chapter, the focus is on describing the general process of finding maximal-closed patterns which are the target of our study as one possible toos for producing the social network of genes.

5.2 Finding Frequent Sets of Expressed Genes

There are several algorithms for finding frequent patterns in a dataset. We will present the Apriori algorithm here which has been accepted as the first algorithm

developed for frequent pattern mining. It makes multiple passes over the data to find frequent sets of items of all lengths. In the first pass, it counts the support (frequency) of individual items and determines which ones of them are frequent, i.e., satisfy the minimum support constraint. In each subsequent pass, it uses the frequent sets of items generated from the previous pass to produce the new potentially candidate frequent sets of items, and counts the support of these candidate sets of items to find the ones that are indeed frequent.

The Apriori algorithm generates candidate sets of items in the current pass by only considering frequent sets of items from the previous pass. The intuition behind this is based on what is known as the Apriori-heuristic, which states that a set of items may be frequent if all its subset sets of items are frequent. This can be done by a self-join of the frequent sets of items of length k, say L_k with itself and then pruning from the result any set of items which has all of its subsets not included in L_k. This process results in generating a much smaller number of candidate sets of items. Therefore, candidate generation consists of two steps: the join step and the pruning step. After the pruning step, the remaining candidates are checked by scanning the database to determine their frequencies. This process is recursively repeated until it is not possible to construct more frequent sets of items.

The Apriori algorithm is level wise in nature. It requires multiple database scans to find the support count for a potentially large candidate set of items and this can be very time consuming. Moreover, the Apriori algorithm requires generating a large number of candidate sets of items at each level, especially for levels two and three. These can also be considered as CPU and memory intensive tasks.

There are several other Apriori-like algorithms, such as DHP [63], DCP [61], and DCI [62], which mainly focus on improving the performance of mining by reducing the candidate generation and/or by introducing special data structures that reduce the time for counting the support of candidates. On the other hand, algorithms like DIC [20] and CARMA [45] try to improve the performance by reducing the number of database scans.

5.3 Finding Maximal-closed Frequent Itemsets

Redundancy is the main problem with keeping all the frequent sets of genes as described in Sect. 5.2. The number of frequent sets of genes is directly related to how the genes are co-expressed across the samples. As the co-expression across the samples increases, the number of frequent sets of genes increases. However, many of the frequent sets of genes may share the same frequency and noticing this would help in minimizing the number of frequent sets of genes to maintain by keeping only maximal-closed frequent sets of genes. In other words, we reduce the number of frequent sets of genes by only concentrating on maximal-closed frequent sets of genes.

A frequent set of genes is said to be closed if and only if its support is different from the support of all its frequent supersets. A frequent set of genes is maximal-closed if and only if it is closed and none of its supersets is frequent. Based on

this, we keep frequent sets of genes of maximum size. To illustrate the process, consider five samples and six genes. For each sample, we list the expressed genes as follows: $s_1 = \{g_1, g_3, g_4\}$, $s_2 = \{g_2, g_3, g_5\}$, $s_3 = \{g_1, g_2, g_3, g_5\}$, $s_4 = \{g_2, g_5\}$, and $s_6 = \{g_3, g_5, g_6\}$. Assume the minimum support threshold is 2, i.e., a set of genes is said to be frequent if it is supported by at least two samples. Excluding the singleton frequent sets of genes, we could enumerate the following five frequent sets of genes: $\{g_1, g_3\}$, $\{g_2, g_3\}$,$\{g_2, g_5\}$, $\{g_3, g_5\}$, and $\{g_2, g_3, g_5\}$. Out of these sets of genes only four are closed frequent, namely $\{g_1, g_3\}$, $\{g_2, g_5\}$, $\{g_3, g_5\}$ and $\{g_2, g_3, g_5\}$. Finally, only $\{g_1, g_3\}$ and $\{g_2, g_3, g_5\}$ are maximal-closed frequent sets of genes. From this result, it is obvious that g_3 is the most important gene; it has the highest closeness, betweenness and degree centralities. Further, the three genes, g_2, g_3 and g_5 form a clique.

The identified maximal-closed frequent sets of genes will be sufficient for determining the frequency of each set of genes that are mostly co-expressed. This allows us to concentrate only on sets of genes that either have different support, or once have same support they do not totally overlap, i.e., none of them subsumes the other.

5.4 Constructing Social Network of Genes and Identifying Biomarkers

The outcome from the proposed three pronged approach is a rich source of information for constructing a social network. Given the n maximal-closed frequent sets of genes, say $MC_1, MC_2, \ldots, MC_n$ and the utilized m genes, say $g_1, g_2, \ldots, g_m$, we construct a matrix $M = m \times m$ to include one row and one column per gene. Entries in matrix M are computed by considering each frequent set of genes MC_k $(k = 1, n)$ and increment $M(i, j)$ by 1 if the pair of genes g_i and g_j exist inside the maximal-closed frequent set of genes MC_k. In other words, $M(i, j) = r$, for all $1 \leq i \leq m$ and $1 \leq j \leq m$, where $0 \leq r \leq n$ is the number of maximal-closed frequent sets of genes in which the pair of genes (g_i, g_j) coexist. It is obvious that $M(i, i) = n$ for all $1 \leq i \leq m$.

To produce a more robust social network of the genes, a second perspective is applied. This perspective works directly on the original data which consists of the samples. A two-mode network of genes and samples is produced. Gene g_i is connected to sample s_j if and only if g_i is expressed in sample s_j. Then, we apply folding on the produced two-mode social network to derive a one-mode social network that covers only the genes. As mentioned above, the one-mode network is produced by multiplying the adjacency matrix of the two-mode network by its transpose. Two genes g_i and g_j are linked in the one-mode network to reflect the strength of having g_i and g_j co-expressed in the same samples. In other words, the one-mode social network of genes is a kind of weighted graph; the weight of each link reflects the degree of co-expressiveness of the two connected genes. The weight

of the link connecting genes g_i and g_j is added to the value in entry (i, j) in matrix M produced by the first perspective.

The third perspective is the k-means clustering algorithm which is used to produce n clusters of the m genes. The clustering result is reflected onto matrix M by considering the following strategy. For every two genes g_i and g_j, if g_i and g_j coexist in the same cluster then the entry $M(i, j)$ is maintained, otherwise (if g_i and g_j do not exist in the same cluster then $M(i, j)$ is replaced by $\frac{1}{M(i,j)}$ as a kind of punishment for the two genes. The basic idea behind this strategy is simple: if two genes are related then they should exist in the same cluster. Actually the test results confirm the validity of this strategy because we realized that genes that do exist together in the same cluster when they coexist in large number of maximal-closed frequent sets of genes.

After all entries in M are determined, we compute the average, say A_v of all the values in M as follows: $A_v = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} M(i,j)}{m^2}$. Based on the comparison of each entry in M with A_v, we normalize every entry $M(i, j)$ to $\frac{M(i,j)}{A_v}$ and then we set $M(i, j) = 0$ if and only if $M(i, j) < 1$. The revised matrix M represents the adjacency matrix of the actual social network where there exist an edge between genes g_i and g_j if and only if $M(i, j) > 0$. After M is transformed into adjacency matrix, genes are clustered into communities by considering the overlap of nonzero values in the corresponding rows. Each gene joins the community where it has more overlap. Then, we determine the central gene within each community by considering the degree of centrality which is determined by computing two values for each gene g in a given community:

1. The weighted degree centrality of gene g, denoted D_g is the sum of the values in the row of g in matrix M divided by the number of genes in the same community, say n_c, $D_g = \frac{\sum_{j=1}^{m} M(g,j)}{n_c}$.
2. The un-weighted degree of centrality is the number of non-zero entries in row g, denoted z_g.

Based on the values of D and z computed for each community, we find the most central gene g within each community as the gene that has high values for D_g and z_g. For this, we sort each of the two lists of values in descending order where the list of z values is given higher priority in the analysis because the values in list D are weighted and hence do not reflect the actual number of neighbor genes. The latter values are considered more seriously to differentiate genes that have closer z values. Central genes within the communities are analyzed further as the biomarkers.

6 Test Results

We have conducted two types of experiments. The first set of experiments is intended to illustrate the validity of our main argument that genes do socialize and form dynamic communities. In the second set of experiments, we directly applied the three-pronged approach to identify some potential diagnostic cancer biomarkers.

6.1 Illustrating the Dynamic Behavior of Genes

To illustrate the argument that gene products do socialize and change functional camps, we report the network of the yeast gene expression data is described in [27]. The authors carried out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration and they also studied genes whose expression was affected by deletion of the transcriptional co-repressor TUP1 or over-expression of the transcriptional activator YAP1. Using the same data described in [27] social networks are constructed at times 9.5, 13.5 and 20.5 h with 500 genes. The corresponding networks are shown in Figs. 1, 2, and 3, respectively. To explicitly shown the change in connections, we also plotted part of each of the three networks by zooming in to shown certain genes where the three zoomed parts labeled with names of genes are shown in Figs. 4–6, respectively. These figures explicitly demonstrate the changes in the network; and hence the change in the correlation between the expressed genes. This verifies our argument that gene products do socialize and change camps. However, more in depth biological analysis in the wet-lab may be needed to carefully study the communities of genes leading to a solid verification of our claim.

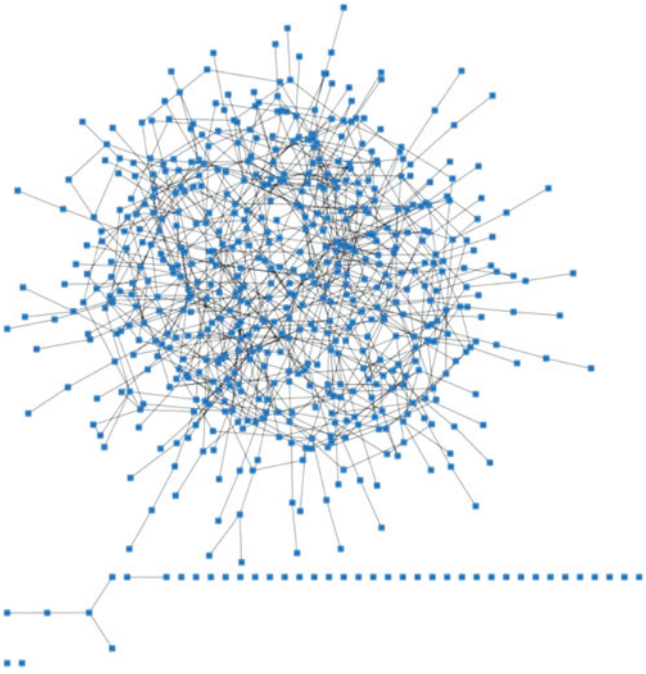

Fig. 1 The complete network of yeast at time 9.5 h

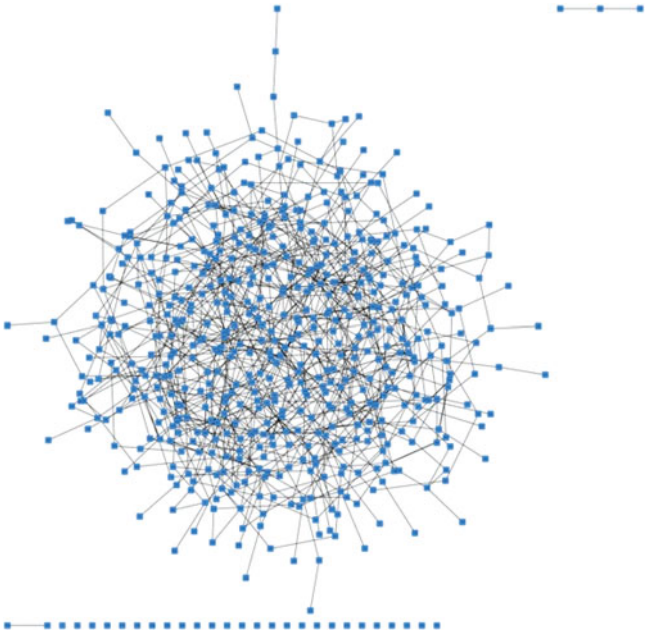

Fig. 2 The complete network of yeast at time 13.5 h

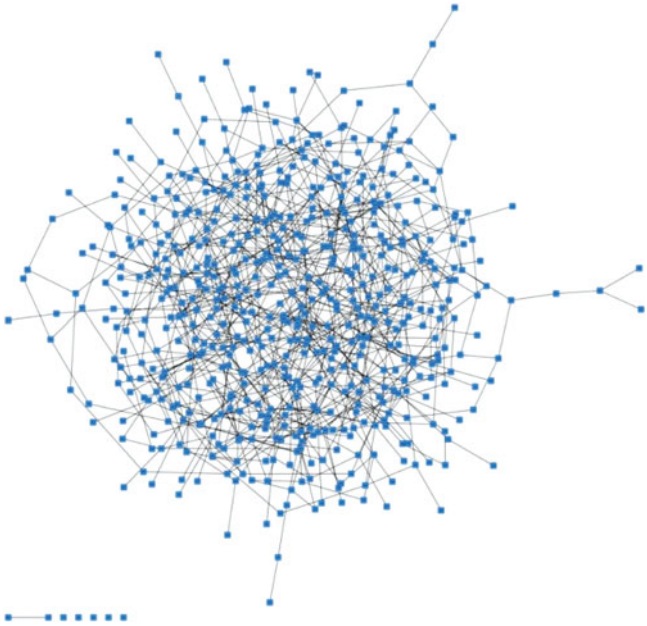

Fig. 3 The complete network of yeast at time 20.5 h

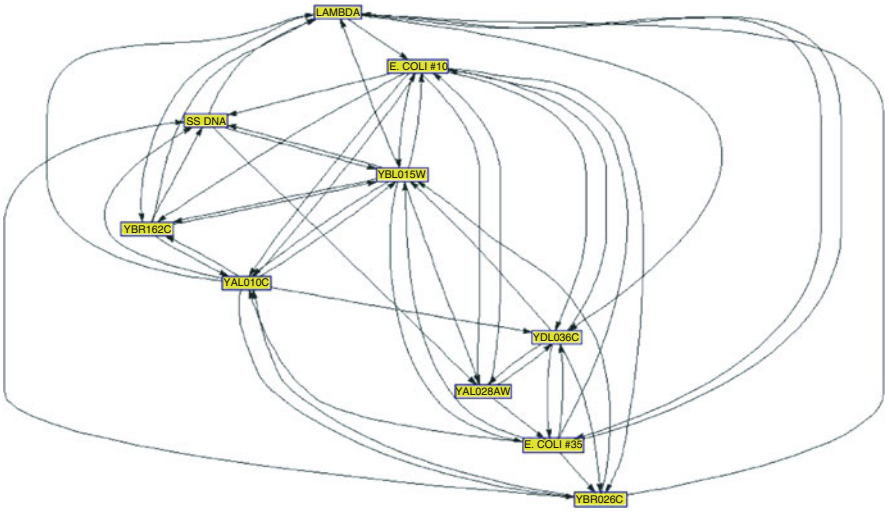

Fig. 4 Part of the network of yeast at time 9.5 h

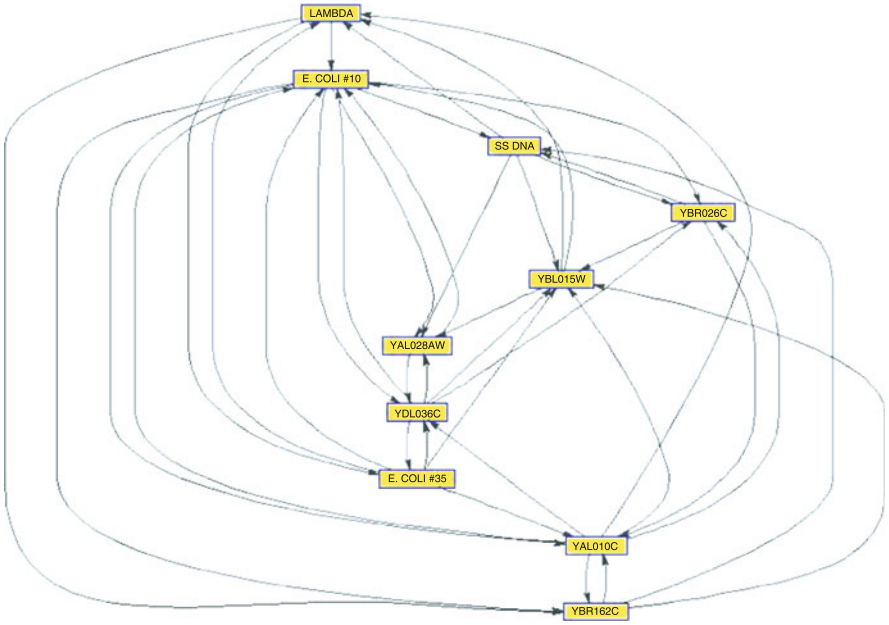

Fig. 5 Part of the network of yeast at time 13.5 h

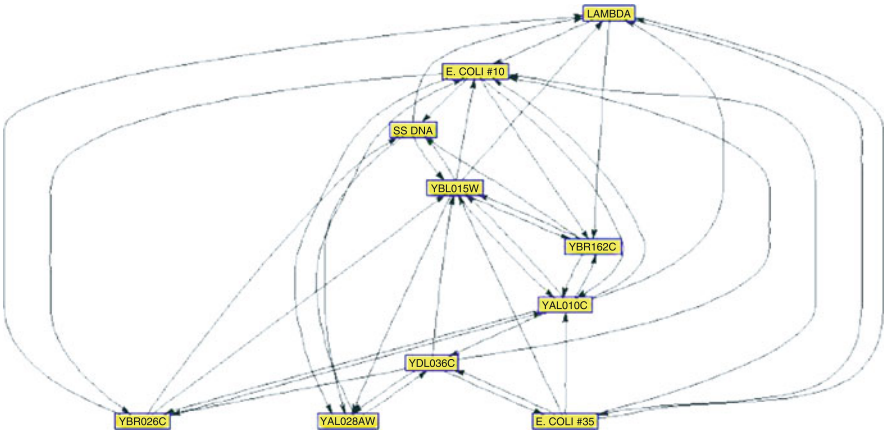

Fig. 6 Part of the network of yeast at time 20.5 h

6.2 Illustrating the Proposed Social Network Construction Framework

In this section, we discuss the conducted experiments. We highlight the results of our approach and evaluate its effectiveness and applicability. We report the results from the comparison of our approach with other existing methods which investigated the same cancer classification problem. We have also analyzed the results both in terms of accuracy and biological significance.

Data preprocessing has been conducted using matlab 7.0. To derive maximal-closed frequent sets of genes, we first used the CloSpan algorithm originally proposed by Yan et al. [78] and then applied on the result a postprocessing step to find the maximal-closed sets of genes. The clustering has been conducted using the k-means implementation in Matlab. Gene selection has been performed using t2test in matlab. For classification, we have used LIBSVM package implemented in matlab. LIBSVM is a free library for classification and regression available online at (http://www.csie.ntu.edu.tw/~cjlin/libsvm/).

In this work we have used two cancer data sets:

1. Acute myeloid leukemia (AML)/Acute lymphocytic leukemia (ALL) taken from [39]
2. Colon data set from [13]

The AML/ALL data contains 7,130 genes and 73 patient samples. We split this data as follows: 58 sample for training and 15 for testing. The colon cancer data studied has 2,000 genes and 62 patient samples with 40 tumor and 22 normal. Samples were split as follows, 48 samples were used for training and 14 samples were used for testing. As a normalization step, the intensity values have been normalized such

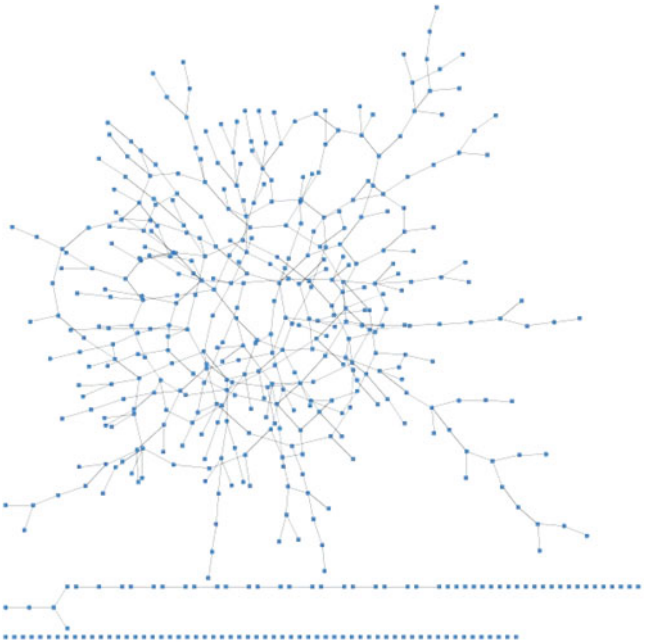

Fig. 7 The network of the lukemia data

that the overall intensities for each chip are equivalent. This is done by fitting a linear regression model using the intensities of all genes in both the first sample (baseline) and each of the other samples. The inverse of the "slope" of the linear regression line becomes the (multiplicative) re-scaling factor for the current sample. This is done for every chip (sample) in the data set, except the baseline which gets a re-scaling factor of one. A missing value means that the spot was not identified. Missing values were predicted according by to k-nearest neighbors strategy; we replace a missing value by the weighted average of the five nearest neighbors; we compute the weighted average by first dividing each of the five nearest neighbors by its distance from the missing value and the computed the average of the five produced values.

We applied the three-pronged approach using the training part from each of the two data sets. First, the frequent pattern mining perspective produced 150 and 62 maximal closed sets of genes for the AML/ALL and colon data, respectively. Second, we applied k-means clustering on each of the two data sets by setting the values of k to the reported number of maximal-closed frequent sets of genes. Third, we applied folding on the two mode social networks of the AML/ALL and colon data. Finally, we combined the results from the three perspectives as described in Sect. 5.4. The network produced from the lukemia data is shown in Fig. 7.

The analysis of the social network constructed for the Lukemia data set revealed six communities of genes. Then the most central gene within each community was

Table 1 Results from other works compared with our results where the errors in the training and testing set were provided and the number of features used for classification. **NA** stands for Not applied in the work

	[39]	[36]	[75]	[48]	[14]	[79]	[5]	Our method
Errors in training set	2	2	0	NA	0	0	0	0
Errors in test set	5	2–4	0	3	1	0	0	0
Number of features used for classification	50	25–1000	8–16	1	16	50	8	6

reported leading to six biomarker genes were a gene is accepted as central when it has high degree centrality and high closeness centrality within its community. We then used the discovered six genes (as data representatives) to build the SVM. The reported results of 100% accuracy (for the training set) and 100% cross validation (for the test set) illustrate the effectiveness of the proposed approach as robust framework for detecting biomarkers. As reported in Table 1, comparing our results to those mentioned in the literature it can be easily seen that the proposed framework is more effective and robust.

For the colon data set, the social network construction and analysis process reported 12 communities. Again we identified the most central gene in each community. The latter 12 genes were then used in building the SVM classifier and the reported results are 85% accuracy and 94% cross validation. Finally, the network for the colon data is shown in Fig. 8. At the end, comparing all the networks plotted in this chapter would lead to stronger support for the argument that genes do socialize and act collaboratively within the cell. In fact every cell contains the same genes and only certain genes are active in each cell type.

7 Summary and Conclusions

We argue that biological molecules within the cell form communities and act collaboratively within those communities to achieve certain goals. Unfortunately, optimal methods that could closely evaluate the interactions between the genes and lead to the extraction of the actual social networks are still lacking. In this work, we tried to construct the social network of gene products by employing a three-pronged approach leading to a robust framework. For visualization, the reader may think of the simple model of two communities of genes within each cell–one community containing expressed genes and the other community containing unexpressed genes for each particular cellular response to a stimulus. The community of expressed genes could be further divided based on the particular cellular goal of expressed genes (cell proliferation, metabolism, synthesis of a particular cellular product, etc.). The aim of our methods is to define the representative genes by analysis of the social communities. Our analysis allows us to find weighted links between

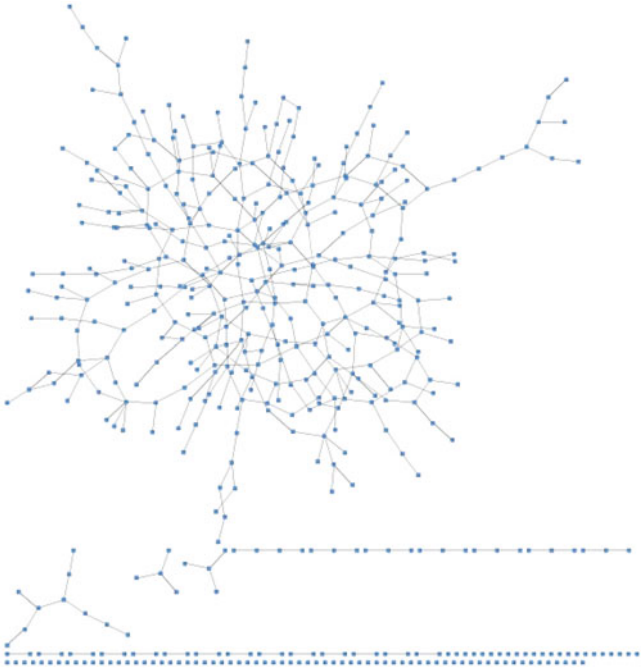

Fig. 8 The network of the colon data

genes based on their co-occurrence in the outcome from the three perspectives employed, namely frequent pattern mining, clustering and folding of the two-mode network. The outputs from the analysis of the social communities reported the most promising biomarker genes. After demonstrating its ability to identify potential cancer diagnostic biomarkers, the same methodology described in this paper can be applied to identify potential biomarkers of other diseases. Also, we are considering social communities of proteins as well as social communities that result from the interactions between genes and proteins. The latter networks are more challenging to study; these are at the center of our current research efforts.

References

1. Agosto, J.: Confronting bioterrorism: Epidemiologic, clinical, and preventive aspects of smallpox. Salud Publica de Mexico, pp. 298–309 (2003)
2. Agrawal, R., Imieliski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 207–216. Washington, D.C., May 1993
3. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the International Conference on Very Large Data Bases, pp. 487–499. San Francisco, CA, (1994)

4. Albert, R., Barabosi, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–97 (2002)
5. Alshalalfa, M., Özyer, T., Alhajj, R., Rokne, J.: Discovering cancer biomarkers: From DNA to communities of genes. Int. J. NVO **8**(1/2), 158–172 (2011)
6. Alshalalfa, M., Alhajj, R.: Cancer class prediction: Two stage clustering approach to identify informative genes. Intell. Data Anal. **13**(4) (2009)
7. Alshalalfa, M., Alhajj, R., Rokne, J.: Identifying disease-related biomarkers by studying social networks of genes. In: Lim, C.P., Jain, L.C. (eds.) New Directions in Decision Support Systems: Methodologies and Applications. Springer, Berlin (2009)
8. Anthonisse, J.M.: The rush in a directed graph. Technical Report BN9/71, Stichting Mahtematisch Centrum, Amsterdam, Oct 1971
9. Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: Membership, growth, and evolution. Proceedings of the ACM KDD (2006)
10. Barabosi, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
11. Baumes, J., Goldberg, M., Magdon-Ismail, M., Wallace, W.: Discovering hidden groups in communication networks. In: Proceedings of NSF/NIJ Symposium on Intelligence and Security Informatics. (2004)
12. Bavelas, A.A.: A Mathematical model for group structures. Hum. Organ. **7**, 16–30 (1948)
13. Bicciato, S., Pandin, M., Didon, G., Di Bello, C.: Pattern identification and classification in gene expression data using an autoassociative neural network model. Biotechnol. Bioeng. **81**(5), 594–606 (2002)
14. Bijlani, R., Cheng, Y., Pearce, D.A., Brooks, A.I., Ogihara, M.: Prediction of biologically significant components from microarray data: Independently consisitent expression discriminator(ICED). Bioinformatics **19**, 62–70 (2003)
15. Binder, P., et al.: Medical management of biological warfare and bioterrorism: Place of the immunoprevention and the immunotherapy. Comp. Immunol. Microbiol. Infect. Dis. **26**(5–6), 401–421 (2003)
16. Blancou, J., Pearson, J.E.: Bioterrorism and infectious animal diseases. Comp. Immunol. Microbiol. Infect. Dis. **26**(5–6), 431–443 (2003)
17. Blank, S., Moskin, L.C., Zucker, J.R.: An Ounce of Prevention is a Ton of Work: Mass Antibiotic Prophylaxis for Anthrax, New York City, 2001. (Policy Review), Emerg. Infect. Dis., **9**(6), 615–612 (2003)
18. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**(2), 163–177 (2001)
19. Brandes, U., Pich, C.: Centrality estimation in large networks. Int. J. Bifurcat. Chaos **17**(7), 2303–2318 (2007)
20. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 255–264. Tucson, Arizona, May 1997
21. Bronze, M.S.: Preventive and therapeutic approaches to viral agents of bioterrorism. Drug Discov. Today 740–745 (2003)
22. Carley, K., Prietula, M. (eds.): Computational Organization Theory. Lawrence Erlbaum associates, Hillsdale, NJ (1994)
23. Christopher, G.W., et al., Biological warfare: A historical perspective. JAMA **278**(5), 412 (1997)
24. Cieslak TJ, Eitzen EM Jr. Clinical and Epidemiologic principles of anthrax, Emerg Infect Dis. **5**(4):552–5. Jul-Aug 1999
25. Croft, D.P., James, R., Thomas, P., Hathaway, C., Mawdsley, D., Laland, K., Krause, J.: Social structure and co-operative interactions in a wild population of guppies (poecilia reticulata). Behav. Ecol. Sociobiol. **59**(5), 644–650 (2006)
26. Dembele, D., Kastner, P.: Fuzzy c-means method for clustering microarray data. Bioinformatics **19**, 973–980 (2003)

27. DeRisi, J.L., Iyer, V.R., Brown, P.O.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278**(5338), 680–686 (1997)
28. Devarajan, P.: Novel biomarkers for the early prediction of acute kidney injury. Cancer Ther. **3**, 477–488 (2005)
29. Diestel, R.: Graph Theory, 2nd edn. Graduate Texts in Mathematics. Springer, Berlin (2000)
30. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, CA (2001)
31. Everett, M.G., Borgatti, S.P.: The centrality of groups and classes. J. Math. Sociol. **23**(3), 181–201 (1999)
32. Ferguson, N.M., et al.: Planning for smallpox outbreaks. Nature **425** (2003)
33. Flake, G.W., Lawrence, S., Giles, C.L.: Efficient identification of web communities. Proceedings of ACM International Conference on Knowledge Discovery and Data Mining, pp.150–160 (2000)
34. Forrester, M., Stanley, S.: Calls about anthrax to the Texas Poison Center Network in relation to the anthrax bioterrorism attack in 2001. Vet. Hum. Toxicol. 247–248 (2003)
35. Freeman, L.C.: A set of measures of centrality based upon betweenness. Sociometry **40**, 35–41 (1977)
36. Furey T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics **16**(10), 906–14 (2000)
37. Ganti, V., Gehrke, J., Ramakrishnan, R.: Demon: Mining and monitoring evolving data. IEEE Trans. Knowl. Data Eng. **13**(1), 50–63 (2001)
38. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
39. Golub, T.R., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lender, E.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**, 531–537 (1999)
40. Gould, R.V.: Measures of betweenness in non-symmetric networks. Soc. Networks **9**, 277–282 (1987)
41. Grabmeier, J., Rudolph, A.: Techniques of cluster algorithms in data mining. Data. Min. Knowl. Discov. **6**, pp.303–360 (2003)
42. Grais, R.F., Ellis, J.H., Glass, G.E.: Forecasting the geographical spread of smallpox case by air travel. Epidemiol. Infect. **131**, 849–857 (2003)
75. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)
44. Halloran, E., Longini Jr, I.M., Nizam, A., Yang, Y.: Containing Bioterrorist smallpox. Science **298** (2002)
45. Hidber, C.: Online association rule mining. In: Proceedings of ACM SIGMOD international conference on Management of data, pp. 145–156, Philadelphia, Pennsylvania (1999)
46. Janssen, M.A., Jager, W.: Simulating market dynamics: Interactions between consumer psychology and social networks. Artif. Life **9**, 343–356 (2003)
47. Jensen, D., Neville, J.: Data mining in social networks. In: Proceedings of the Symposium on Dynamic Social Network Modeling and Analysis (2002)
48. Jinyan, L., Wong, L.: identifying good diagnosis gene group from gene expression profile using the concept of emerging patterns. Bioinformatics **18**, 725–734 (2002)
49. Kaplan, E.H., Craft, D.L., Wein, L.M.: Emergency Response to a smallpox attack: The case for mass vaccination. PNAS **100**(7) (2003)
50. Kianmehr, K. and Alhajj, R.: Calling Communities Analysis and Identification Using Machine Learning Techniques. Expert. Syst. Appl. **36**(3), 6218–6226 (2009)
51. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)

52. Klerks, P.: The Network Paradigm Applied to Criminal Organisations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. CONNECTIONS **24**(3), 53–65 (2001)
53. Lawrence, S., Giles, C.L.: Accessibility of information on the web. Nature **400**, 107–109 (1999)
54. Li, L., Pedersen, L.G., Darden, T.A., Weinberg, C.R.: Class prediction and discovery based on gene expression data. Iostatistics Branch and Lab of Structural Biology, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina (2000)
55. Mayr, A.: Smallpox vaccination and bioterrorism with pox viruses. Comp. Immunol. Microbiol. Infect. Dis. **26**(5–6), 423–430 (2003)
56. Meltzer, M.L., et al.: Modeling potential responses to smallpox as a bioterrorist weapon. Emerg. Infect. Dis. **7**(6) (2001)
57. Memon, N., Larsen, H.L.: Structural Analysis and Mathematical Methods for Destabilizing Terrorist Networks. Proceedings of the International Conference on Advanced Data Mining Applications, Springer-Verlag Lecture Notes in Artificial Intelligence (LNAI 4093), pp. 1037–1048 (2006)
58. Menczer, F.: Evolution of document networks. Proc. Natl. Acad. Sci. USA **101**, 5261–5265 (2004)
59. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA **98**, 404–409 (2001)
60. Newman, M.E.J.: A measure of betweenness centrality based on random Walks. Soc. Networks **27**, 39–54 (2005)
61. Orlando, S., Palmerini, P., Perego, R.: Enhancing the apriori algorithm for frequent set counting. In: Proceedings of ACM International Conference on Data Warehousing and Knowledge Discovery, pp. 71–82, London, UK (2001)
62. Orlando, S., Palmerini, P., Perego, R., Silvestri, F.: Adaptive and resource aware mining of frequent sets. In: Proceedings of IEEE International Conference on Data Mining, p. 338, Washington, DC (2002)
63. Park, J.S., Chen, M.S., Yu, P.S.: Using a hash-based method with transaction trimming for mining association rules. IEEE Trans. Knowl. Data Eng. **9**(5), 813–825 (1997)
64. Pennock, D.M., Flake, G.W., et al.: Winners don't take all: Characterizing the competition for links on the web. Proc. Natl. Acad. Sci. USA **99**(8), 5207–5211 (2002)
65. Powell, W.W., White, D.R., et al.: Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. Am. J. Sociol. **110**(4), 1132–1205 (2005)
66. Ressom, H., Reynolds, R., Varghese, R.S.: Increasing the efficiency of fuzzy logic-based gene expression data analysis. Physiol. Genomics **13**, 107–117 (2003)
67. Sahab, Z.J., Semaan, S.M., Sang, Q.-X.A.: Methodology and Applications of Disease Biomarker Identification in Human Serum. Biomark Insights **2**, 21–43 (2007)
68. Stern, J.E.: Will Terrorists Turn to Poison? Orbis **37**(3), 393–410 (1993)
69. Shariat, S.F., et al.: Multiple biomarkers improve prediction of bladder cancer recurrence and mortality in patients undergoing cystectomy. Cancer **112**(2), 315–25 (2008)
70. Sorensen, K.D., Orntoft, T.F.: Discovery of Prostate Cancer Biomarkers by Microarray Gene Expression Profiling. Expert Rev Mol Diagn. **10**(1), 49–64 (2010)
73. Toure, A., Basu, M.: Application of neural network to gene expression data for cancer classification. In: Proceedings of IEEE International Joint Conference on Neural Networks, pp. 583–587 (2001)
72. Tseng, G.C., et al.: Investigating Multi-cancer Biomarkers and Their Cross-predictability in the Expression Profiles of Multiple Cancer Types. Biomarker Insights **4**, 57–79 (2009)
73. Tusher, V.G., Tibshirani, R., Chu, G.: Significant analysis of microarrays applied to the ionizing radiation response. PNAS **98**(9), 5116–5121 (2001)
74. Wagner, M., et al.: Computational protein biomarker prediction: a case study for prostate cancer. BMC Bioinformatics **5**(26) (2004). doi:10.11861471-2105-5-26
75. Woolf, P.J., Wang, Y.: A fuzzy logic approach to analyzing gene expression data. Physiol. Genomics **3**, 9–15 (2000)

76. Wang, J., Hellem, T., Jonassen, I., Myklebost, O., Hovig, E.: Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. BMC Bioinformatics **4**, 60–72 (2003)
77. Xu, J.J., Chen, H.: CrimeNet Explorer: A framework for criminal network knowledge discovery. ACM Trans. Inform. Syst. **23**(2), 201–226 (2005)
78. Yan, X., Han, J., Afshar, R.: CloSpan: Mining Closed Sequential Patterns in Large Datasets. Proc. of 2003 SIAM Int. Conf. Data Mining (SDM' 03) (2003)
79. Zhang, X., Ke, H.: ALL/AML cancer classification by gene expression data using SVM and CSVM approach. Genomics informatics **11**, 237–239 (2000)