Proceedings of the 7th International Conference on Applied Informatics
Eger, Hungary, January 28–31, 2007. Vol. 1. pp. 129–138.

Graph clustering and visualizing methods on genomic data*

Miklós Krész, Attila Tóth

Department of Computer Science
University of Szeged, Juhász Gyula Faculty of Education
e-mail:{kresz,attila}@jgypk.u-szeged.hu

Abstract

Questions raised by DNA-chip technology pose new challenges for bioinformatics. In contrast to the information stored in static DNA databases, DNA-chip experiments provide a large amount of information about dynamic changes in the expression of several thousand genes simultaneously. It is a natural goal to exploit both of these information sources, obtaining new results and dependencies which open new horizons for bioinformatics in the branch of genomic research. Since structural relationships play an important role in modern data analysis, graph-theoretic models and algorithms are popular tools in this field. In this paper we present our experiences about graph clustering and graph visualizing methods developed in the project "Natural Language Processing, Information Extraction and Development of a Graph Based Analytic Infrastructure for Genomic Research".

Keywords: data mining, data visualization, clustering, genomic research, DNA-chip technology

1. Introduction

As a result of the dynamic development of information systems observed in the last decades, such a vast amount of data has accumulated in most databases by today that their analysis is exceedingly complicated, information seeking technologies based on these principles (relation based query, statistical analysis) are often highly time-consuming or do not give an accurate result. Artificial intelligence based, so-called data mining methods, which are capable of extracting information quickly and quite efficiently using automatic procedures from tables of considerable size, containing millions of lines, were developed in the 1980s and 90s to solve the above and similar problems.

*Supported by a grant from the Hungarian Ministry of Economy and Transport (project no.: GVOP-3.1.1-2004-05-0119/3.0).

Data mining was first used successfully in the business sector, but from the 1990s it became obvious that thanks to the development of technology, scientific experiments also produced such an enormous amount of data, that their analysis would be greatly assisted by the methods of this new field. The invention of the DNA-chip opened new horizons in genomic research. DNA-chip experiments produce an immense amount of data of the dynamic changes of the expression of thousands of genes, as opposed to the static information stored in DNA databases. Acquiring the information hidden in these data poses new challenges for bioinformatics. At the same time the development of information technology made it possible for biology-based publications and abstracts to be accessible in open databases. This naturally leads to the question of what methods can be used to facilitate the evaluation of the experimental results with the extraction of information contained in the articles.

The most vital information for genomic research is provided by the expression levels of the individual genes and proteins in different states (or conditions). Based on the above, deductions can be made about the relationships between the genes and certain proteins. The genes can be arranged into a structured data set based on the information found in the articles, where the connections between the elements can be represented in a network (graph).

Over the course of the past years the so-called biological text mining has become a dynamically developing independent field, at the same time, graph based data mining has also become the center of attention. The system called BiblioGraph Explorer was designed by combining these two methodologies. In the following, we wish to give an insight into the applications of graph based data mining by sharing the experience gained over the course of developing the data analysis and data visualization module.

2. Analysis of genomic data

Since the appearance of the microarray technology (DNA-chip) tracing the activity of any number of genes of an organism is possible, i.e. how an organism reacts to environmental effects can be detected, what is more, diseased and healthy tissues, resistant and sensitive plants can be compared. The DNA-chip is basically a large number of oligonucleotides, cDNAs, proteins or drug-like compounds placed on a chemically activated glass slide (Figure 1). The new tool has opened revolutionary horizons in functional molecular biology, making the development of quick and widely used analytical methods possible with the simultaneous observation of the expression level of different mutations of the genome, known and unknown genes and proteins.

Thus, today we must be able handle the dynamic data of genome activity as well as the descriptive, static information of genome research. The quantitative analysis of the many kinds of data is a great task for the young science of bioinformatics, since the handling of the complete database and computational toolkit of genome research, as well as that of the literary data banks is required for the recognition of

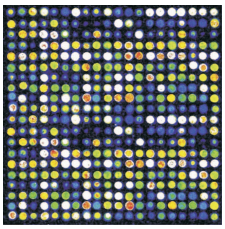

Figure 1: DNA chip

new patterns hidden in the data. Modern analysis techniques (pattern recognition, data mining) are needed to achieve the above aim.

The result of gene expression experiments is a complex data mass, which can only be systemized and appropriately assessed with time-consuming literature review and previous knowledge. Thus, one of the longest portion of biological research is literature review, which can be accelerated with text based data mining methods. The analysis of the extracted information with data mining methods not only accelerates research, it also makes the exploration of new connections possible. The case study reviewed in this paper used the abstracts found in the widely accessible MedLine database for information extraction. Articles [1] and [2] outline the developments made over the course of the project of text mining processing of MedLine abstracts. In the following, we will show the results obtained by the help of graph based data mining and visualization methods.

3. Building the graph structure

The feasibility and efficiency of data mining procedures highly depends on the structure of the analytical infrastructure containing the data to be analyzed and the relationships between them. In connection with problems arising in practice, it can be frequently observed that data as nodes form a complex network, in which the edges represent the relations specified by the analytical points of view. Graphs with edges labelled by either the type of the relationships between the data or weights expressing the strength of the relationships can be used for the formal description of the analytical infrastructure of the above complex networks.

With text mining methods one gets a structured "contents page" of the MedLine database that is defined by the gene expression data gained from experiments. In the case of genomic research, the accurate exploration of the relationships between the genes gives the basis for recognizing the principles. The relationships between genes were analyzed by text based data mining on three levels. Level one relationship meant two genes were found in the same abstract. Whether or not the two genes were found in the same sentence constituted level two. At the same time it was important that data extracted from the literature be available to use

in complex analyses, based on how the different genes affect each other in different states. These states may form the information categories of the structural table produced by the abstracts, e.g. a given tissue, treatment or disease. Thus, on level three, the parameters of the relationships between the nodes are provided by how the expression level of the investigated genes changes in a given state (e.g. both increase). Based on level three, one can conclude whether the information found in the given abstract refers to the interaction of the two genes. Therefore, in the graph forming the base of the analytical infrastructure, the genes are the nodes, and two genes are connected by an edge on a certain level if they can be found together in at least one abstract, sentence, or if they interact with one another. The degree of the relationship is determined with the help of the above aspects.

In comparison with other methods of investigating gene networks, new analytical viewpoints are created by another genomic category of the graph structure, namely functional groups. A functional group is made up of the necessary genes for the specified special function. A gene can belong to any number of functional groups, which can furthermore be structured into a hierarchy. In a particular relationship the gene is in connection with all the descendents of the hierarchy within the functional group, thus, always the highest possible level of functional groups is given when stating the relationship between the gene and the functional group. Connection through functional groups provides another parameter in addition to the extracted data form the literary abstracts for the determination of the strength of the relationship between the genes.

In summary: Different levels of extracted information from the MedLine database (same abstract, same sentence, interaction), together with connection through functional groups determine the strength of the relationship of two genes in the structure, which is shown in the graph by the weight of the edge connecting the nodes. Hence, the task is analyzing a structured data set represented by a weighted graph.

4. Data mining and visualization

Data mining in reality is a collective term comprising procedures and technologies which are all capable of efficiently searching vast data bases for the connections between these data. However, since data gained by text based data mining may contain errors (e.g. due to the great number of synonyms, analysis is an extremely complex task), the integration of methods capable of automatic conclusions into the system described in the present case study did not seem practical. Thus, our aim was to apply and develop procedures (also in connection with data mining) which facilitate the interpretation of results.

Automatic segmentation (clustering) and visualization form the base of the above methodology. The objective of automatic segmentation is to group the genes that have a strong relationship. Since the data are arranged in a weighted graph, the nodes of the graphs must be grouped into disjoint clusters in such a way that the subgraph induced by a cluster is “dense” (the total weight of the edges is large

compared to the number of nodes), while the edges running between the clusters determine a "sparse" graph.

Visualization offers an opportunity to present the "logics" of the investigated graph structure, i.e. one sees a simplified portrayal of the relationships. This image highlights the connections that are important in a certain respect, thus facilitating the interpretation of the results. However, clustering is of great significance in the course of visualization as well, since relations become chaotic in a large graph, unless the nodes are put into cluster groups. Because of these reasons it was expected of the developed clustering procedures to be integrable into visualization.

Clustering integrated into visualization can basically be realized by two main principles. One notion creates the clusters first and places them optimally in space in the course of the visualization process. It also attempts to visualize the nodes based on their relationships with one another within the clusters. (see [3]) The other approach tries to apply metrics for both clustering and visualization, which can be easily transformed into each other. We chose the latter approach, thus, the combinatorial methods based on graph structure clustering were not even examined, only procedures which define an appropriate metrics on the graph were tested. However, there is either no visualization application for the metrics based graph clustering procedures found in the literature (e.g. see [7]), or the realization is connected to the given metrics ([4]). Since the different queries may result in graphs with different structures, our concept was to realize the analytics with the combination of several methods, and to enable the user to choose from a number of clustering procedures. This meant using several metrics, which made it necessary for visualization to contain a general metrics-preserving mapping instead of using distinct methods in each case for the placement of the nodes.

When visualizing graphs, one must define what drawing conventions should take into consideration. In the present situation it is natural to expect the distance between the nodes to symbolize the degree of dissimilarity, as the graph nodes represent genes. Since the realization of visualization is metrics-preserving, the scale defined on the set of the nodes of the graphs must express the dissimilarity between the genes.

In summary: We defined metrics on the graphs, which represent the dissimilarity between the nodes, i.e. symmetrical distances were created, which are defined on any node-pair, and triangle inequality applies to them. By applying a metrics-preserving mapping, one arrives at the 3D placing of the nodes. Applying a clustering procedure with the same metrics, visualization will be consistent with segmentation.

4.1. Metrics

The created distance function came about as the weighted combination of the different metrics. These metrics are described below.

Euclidean distance: In a vector space, the length of the difference vector of two vectors is applied as distance function, i.e. the square root of the scalar product of the difference vector with itself. In the case of graphs it means that it is ap-

plied to the rowvectors of the adjacency matrix to express the dissimilarity of the lines. The longer the difference vector, the greater the distance between the given nodes (genes) (represented by those lines). The literature does not recommend this method among the metrics used for graph clustering ([5]), yet it is adequate for identifying strong relationships between the nodes, therefore, this, too, is a part of our combined metrics.

$$D(i,j) = \sqrt{\sum_{k \neq i,j} ((a(i,k) - a(j,k))^2}$$

where n represents the number of nodes, $a(i,j)$ is the (i,j)th element of the adjacency matrix, i.e. the weight of the edge connecting nodes i and j.

Metrics based on random walks: Random walks are frequently used for determining the strength of relationships in weighted graphs. Web search engines are a good example for the application of this concept. The underlying principle is that if the information has a greater probability of flowing between two nodes, their relationship is stronger. These methods can also be used for identifying the so-called weak relationships, unlike the classical method described above. The starting point of this procedure is a random walk on the nodes of the graph, the so-called Brownian motion. The considered random walks assume that in the case of making a step traversing node v, the probability of it happening on a given edge e incident with v is calculated by the ratio of the weight of e and the total weight of edges incident with v. These values are stored in the so-called probability transition matrix. The base value is determined by the expected value of the number of steps of the random walk between nodes ([7]):

$$B(i,j) = \sum_{l=1}^{n} (\tfrac{1}{I - A(j)})_{il}$$

where I is the identity matrix, $A(j)$ is obtained from the probability transition matrix by making column j constant zero. This mapping does not meet the requirements of the metrics (neither that of symmetry, nor that of triangle inequality), however, a transformation similar to the Euclidean metrics provides us with a real metrics ([8]):

$$D(i,j) = \frac{\sqrt{\sum_{k \neq i,j} ((B(i,k) - B(j,k))^2}}{n-2}$$

Diffusion metrics: The diffusion distance ([6]) between two nodes is defined with the help of the basic metrics which expresses the probability of getting from one certain node to another one after exactly t steps. Next, Euclidean metrics is applied to these base values by norming the components of the difference vector with respect to the weigthed degree of the given node. Choosing parameter t suitably is of great importance in this method, our test showed that the values between 3 and 5 are optimal.

$$D(i,j) = \sqrt{\sum_{k=1}^{n} \frac{(P_{ik}^t - P_{jk}^t)^2}{d(k)}}$$

where P^t is the tth power of the probability transition matrix, $d(k)$ is the total weight of edges incident with node k (weighted degree of node k).

4.2. Dimension reduction methods

One of the most popular methods for approximating metric preserving mappings is the class of the dimension reduction methods, which project a mass M into an n-dimensional Euclidean space in such a way that the spatial positions of the nodes approximately preserve the distance defined on the mass. The best known methods are Locally Linear Embedding, Principal Component Analysis, and Multidimensional Scaling. Our team decided to use Multidimensional Scaling (MDS), which proved to be the most stable numerically among the above methods. Multidimensional Scaling is a collection of statistics based techniques, where the objective function is the error function determined by the square sum of the original distances and the distances induced by the mapping:

$$E_M = \sum_{k \neq l} (d(k,l) - d'(k,l))^2$$

where $d(k,l)$ is the original distance between nodes k an l, $d'(k,l)$ is the Euclidean distance after projecting the nodes. The aim is to minimize the above error function, i.e. to determine a mapping for which the above value is minimal.

4.3. Clustering

By determining the appropriate metrics, classical clustering methods can be used on the nodes of the graph. Clustering can be essentially done in one of two ways, these are the following: partitional and hierarchical methods. Partitional methods group data into k clusters, where k is given beforehand, while hierarchical methods create the hierarchical decomposition of the data. A drawback of partitional methods is that the value of k must be determined in advance, however, it gives a good estimate as to the optimum for a given k. In the case of hierarchical methods, clustering is refined by either merging two clusters or dividing a given cluster, i.e. we either start out with one cluster and divide the selected cluster (divisive methods), or start out with n clusters and merge two clusters at every step (agglomerative methods). Hierarchical clustering results in a more efficient implementation, but if we get to a certain branch on the hierarchic tree, there is no turning back. Usually agglomerative hierarchical clustering is used on graphs, we, however, developed partitioning methods as well, by taking advantage of the metric space.

In the case of partitional methods, parameter k had to be tested with several possible values. The employed methods were built on the k-mean and k-medoid procedures. In the case of the k-mean procedure similarity is measured against the center of the clusters (mean of the cluster elements), while in the case of the k-medoid procedure it is measured against the medoid of the cluster elements. In both cases, at the beginning of the method k elements are selected as center

or medoid of the distinct clusters, then the remaining elements are grouped into the closest cluster. Next, the new centers and medoids are determined. This is continued until the composition of the clusters changes.

In the case of hierarchical methods, agglomerative procedures were employed. In this instance the main question is how the distance between the clusters is determined, and what method is used to choose between the clusters to be merged. The distances between the closest, the farthest and the medoid nodes, as well as the average distance were analyzed as cluster distance. Comparisons showed no major differences in our tests, thus, the optimal choice was the simplest one, the distance between the closest nodes. For cluster selection, we analyzed according to several indices, but no significant difference was found. Hence, taking into consideration of efficiency as well, the optimal choice proved to be that of the closest clusters.

4.4. Evaluation

In the case of both methodologies it was an important aspect that mathematical evaluation be possible beside the analysis from the biological meaning. To investigate this, various indices were used, finally modularity ([5]) proved to be optimal based on the biological analysis and efficiency. The formula of modularity can be efficiently calculated and shows how the ratio of the sum of the edge weights within the clusters relates to a random clustering.

$$Q = \sum_i e_{ii} - \sum_{ijk} e_{ij} e_{ki}$$

where e_{ij} is the fraction of the weight sum of the edges in the graph that connect nodes in clusters i to those in cluster j in the graph.

A higher value means a better quality of clustering. However, beside modularity, the number of clusters is also to be considered when customizing the algorithms.

5. Application

The application called BiblioGraph Explorer was developed making use of text mining and graph based data mining methods. The Functional Genomics Laboratory of the Biological Research Center (BRC), Hungarian Academy of Sciences (DNA-chip experiments, biological data analysis), the University of Szeged (USZ) Informatics Institute (text mining), the Department of Computer Science, USZ, Faculty of Education (graph based data mining and visualization), and Data Explorer Inc. (system development) took part in the project. The application of the system is illustrated with the following test result.

The researchers of the BRC Functional Genomics Laboratory investigated what gene expression changes occur in tumor cells treated with polysaturated fatty acids. The information from this experiment is very valuable in the field of fatty acid research, as several studies deal with the anti-tumor effect of these substances, but the pharmacodynamics is not known in detail. Using DNA-chip technology it has been observed that the expression of a number of genes changed. The relationship

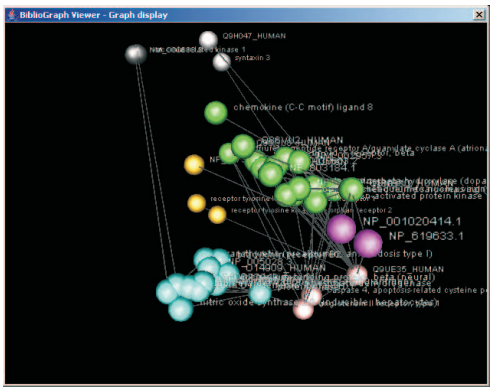

Figure 2: BiblioGraph Explorer

between the genes was determined by the BiblioGraph Explorer software. The results were checked without the software with the employment of the PubMed biological publications database. The abstracts of the publications in which the studied genes appeared were found in every instance. The illustration created by this software of the graph structure (the relationships of the genes) built on this basis can be seen in Figure 2. The clusters are seen in different colors. By shedding light on the relationship system we get closer to understanding the changes in gene expression.

The following could be concluded from the figure and the clustering. The joint appearance of genes showing a change defined six separate groups, all of which could be characterized biologically, the two largest groups being genes connected with the cell cycle (top right), and genes connected with inflammation (bottom left).

It is seen from the clustering results that gene Q9UE35, which has a special role in the cell cycle as well as other mechanisms, is found in the center (middle left). A connection can also be detected between this gene and inflammation-specific genes. The recognition of the above is also new to the inhibiting effect of polyunsaturated fatty acids on cell growth. The above example illustrates that with the help of the application not only can the relationships be represented, but new connections can also be recognized.

6. Summary

Nowadays graph based data mining is one of the most efficient tools of knowledge discovery by exploring the structural connections in vast, open databases. Due to the rapid increase of results from biological experiments, this modern technology is beginning to become an important part of bioinformatics. In the case

study outlined in this paper it was shown how data mining and visualization methods integrated into software can be developed based on the data acquired from the results of DNA-chip experiments with the help of the information extracted from the MedLine database. These new methods open the door to the discovery of new functional genomic connections.

References

[1] Busa-Fekete, R., Kocsor, A., Extracting Human Protein Information from MEDLINE Using a Full Sentence Parser, *Acta Cybernetica*, to appear.

[2] Csendes, D., Alexin, Z., Busa-Fekete, R., Kovács, K., New, Linguistics-based, Ontology-enabled Approaches, *in Biological Information Management, in the Proceedings of the e-Challenges 2006 Conference*, October 25-27, Barcelona, Spain, (2006), 1352–1359.

[3] Herman, I., Melancon, G., Marschall, M. S., Graph Visualization and Navigation in Information Visualization: a Survey, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 6 (1), (2000), 23–42.

[4] Lafon, S., Lee, A. B., Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28 (9), (2006), 1393–1403.

[5] Newman, M. E. J., Detecting community structure in networks, *Eur. Phys. J. B* Vol. 38, (2004), 321–330.

[6] Pons, P., Latapy, M., Computing Communities in Large Networks Using Random Walks, *Lecture Notes in Computer Science* Vol. 3733, (2005), 284–293.

[7] Zhou, H., Network landscape from a Brownian particle's perspective, *Phys. Rev. E* Vol. 67 041908, (2003).

[8] Zhou, H., Distance, dissimilarity index and network community structure, *Phys. Rev. E* Vol. 67, 061901 (2003).

Miklós Krész, Attila Tóth
Department of Computer Science
University of Szeged, Juhász Gyula Faculty of Education
Boldogasszony sgt. 6
H-6725 Szeged
Hungary