Visualizing the Dynamic Modelling of Gene Expressing Data

Leishi Zhang and Xiaohui Liu

Department of Information Systems and Computing, Brunel University,
Uxbridge, Middlesex, UB8 3PH, UK
{leishi.zhang, xiaohui.liu}@brunel.ac.uk

Abstract

Modelling gene expressing data could help obtain a "global" view of biological process and ultimately lead to a great improvement in the quality of human life. Kellam et al have built an effective computational framework for modelling Short, high dimensional, multivariate time series virus gene expression data [Kellam01]. However, without appropriate visualization techniques, the data mining models, though a reduced form of original data, may not fully reveal the hidden insight of the application domain. In this paper, we describe some novel visualization techniques that could be effectively applied in the time series modelling framework. Some graphical models were built in a three dimensional prototype to demonstrate the visualization effect of the modelling results.

1. Introduction

Over the last few years it has been common to use high throughput functional genomics methods to investigate multiple events in a cell or tissue that define a phenotype [D'haeseleer2000]. DNA microarrays are one such methodology that allows the simultaneous determination of mRNA abundance for many thousands of genes in a single experiment [Quacknebush01] [Lockhart2000]. The ability to study changes in transcript abundance for many genes simultaneously has been used to extract information about what uncharacterised genes do and how they are regulated [Dagum95].

Viruses have been studied extensively in molecular biology in order to understand cellular processes. Viruses also permit the study of complex multigene expression in the context of the human cell. Herpesviruses are important pathogens of both animals and human particularly in the immunocompromised host. Currently 8 human herpesviruses are known. Herpesviruses maintain an episomal genome of over 100 genes in the nucleus of infected cells. Under appropriate cellular stimulation a highly controlled cascade of viral transcription occurs with the clearly defined endpoint of new virus production. This therefore provides an excellent system for bioinformatic analysis of a transcriptome in the context of human cells. We have a DNA array of all known and putative open reading frames of human herpesvirus 8 (HHV8) available to us, consisting of 106 genes expressed at eight time points [Jenner2001].

To understand how these virus genes interact with host genes over time, the Intelligent Data Analysis Group at Brunel has developed a computational framework for modelling virus gene expression data [Kellam2002]. Since virus gene expression data are essentially a high-dimensional, short multivariate time series (MTS), the framework reduces the dimensionality of the time series before applying appropriate temporal modelling method (see section 2 for details). Although some important insights have already been revealed by applying this framework to the modelling of HHV8 data, the modelling process is hidden from biologists and the modelling outcome is primarily restricted to numeric values. This has led to the work reported in this note – the visualisation of the virus gene expression process and modelling results.

Over the last few years, various researches have been carried out on gene network discovery and visualization [Robinson97][Zapata-Rivera99]. However, most of the research work only produces static models in which interaction between genes over time might not be able to be inspected [Aoshima02]. The work in [Kellam2002] has suggested a way for modelling and visualizing the dynamic structure of gene network over time. Our work aims at applying novel and effective visualization techniques to visualize the dynamic gene network generated from this framework.

In our prototype, for the first time, 3-D animation technique is used to visualize the modelling of multivariable time series gene expression data. Focus +Context technique is also firstly applied in this domain to display the explanation model hidden behind the gene network. Four graphical models are developed using different visualization techniques (transparency, colour, distance, text, direction etc.): global view (grouping), Dynamic Bayesian Network, forecasting, and explanations, each visualizing a part of the modelling process. A database which contains the detailed information of each gene is built and linked to corresponding gene node in the models. Various mouse functions are added to the prototype to enable different user interactions, making the models not only "readable" but also "playable".

2. Background

There are three stage of the modelling strategy in Kellam's framework: correlation search, variable grouping, and short MTS modelling.

Firstly, a correlation search is performed on the gene expression data, producing a set of strong correlations between variables over different time lags. The correlation produced from the MTS data are then fed into a group algorithm producing a set of groups of variables where there are a high number of correlations between members of the same group, but a low number of correlations between members of different groups. These groups are then used as a basis for the model building process. Within this framework, a modelling paradigm has been developed to explain and forecast MTS. This makes use of Dynamic Bayesian Networks (DBNs), which offer an ideal way to perform both explanation and forecasting.

A Dynamic Bayesian Network consists of a graphical representation of dependencies between variables over time, and conditional probability distributions between linked variables. One can query DBN in order to gain information about the relationships that exist between the variables. For example, we can enter some evidence into the network, apply an inference algorithm to propagate this evidence, and then observe the effect of this evidence has upon the distributions of other variables both forward and backward in time. This feature allows the testing of the effect of changing a biological process by manipulating key components of the system, which is of particular interest to the pharmaceutical industry in trying to identify the best target in a disease process for therapeutic intervention.

The framework explicitly manages the temporal relations between variables within each step. When applying this method to the analysis of virus gene expression data, the groups found and their corresponding networks have varying degrees of biological support, the forecasts produce good results, and some of the explanations have shown interesting biological connotations [Kellam2002]. However, the results of each stage of the data mining process are all stored in numerical form which can probably only be well understood by statisticians. How to turn these numerical data into comprehensible visualization models becomes a challenging task.

To visualize the data mining results effectively, a prototype was built using Java3D to display some sample three-dimensional graphical models. Various user interactions can be applied to the prototype, allowing user to view the grouping of gene expression data, the Dynamic Bayesian Networks of gene groups, the forecast of gene expression level together with the probabilities and the explanation models. The prototype also provides cause-and-effect insight into the operation of the network model by allowing the user to select values for feature and see the effects.

3. Design Principle

A model can be understood is a model that can be trust. While data Mining models typically generate results that were previously unknown to the user, it is important that the model visualization provides the user with sufficient levels of understanding and trust. Advanced visualization

techniques can greatly expand the range of models that can be understood by domain experts.

According to Kurt Thearling [Thearling01], three components are essential for understanding a model: representation, interaction, and integration. Representation refers to the visual form in which the model appears. A good representation displays the model in terms of visual components that are familiar to the user. Interaction refers to the ability to see the model in action in real time, and to let the user play with the model as if it were a machine. Integration provides the user context. It refers to the ability to display relationships between the model and alternate views of the data on which it is based.

Since there are many ways to graphically represent a model, the visualizations that are used should be chosen to maximise the value to the viewer. This requires that we understand the viewer's needs and design the visualization with end-user in mind. As the viewers are the experts in gene expression research area but not data mining, we must translate the model into a natural representation for them.

4. Model Visualization

4.1 Global view

The first stage of our work involves the visualization of grouping information and general information (gene ID, description) for each individual gene. To avoid an over cluttered screen, we visualize group information and gene information in different forms. The 106 genes are visualized as 3-D nodes and thrown into the information space according to their correlations. General information of individual genes is stored in a table and displayed on the right panel.

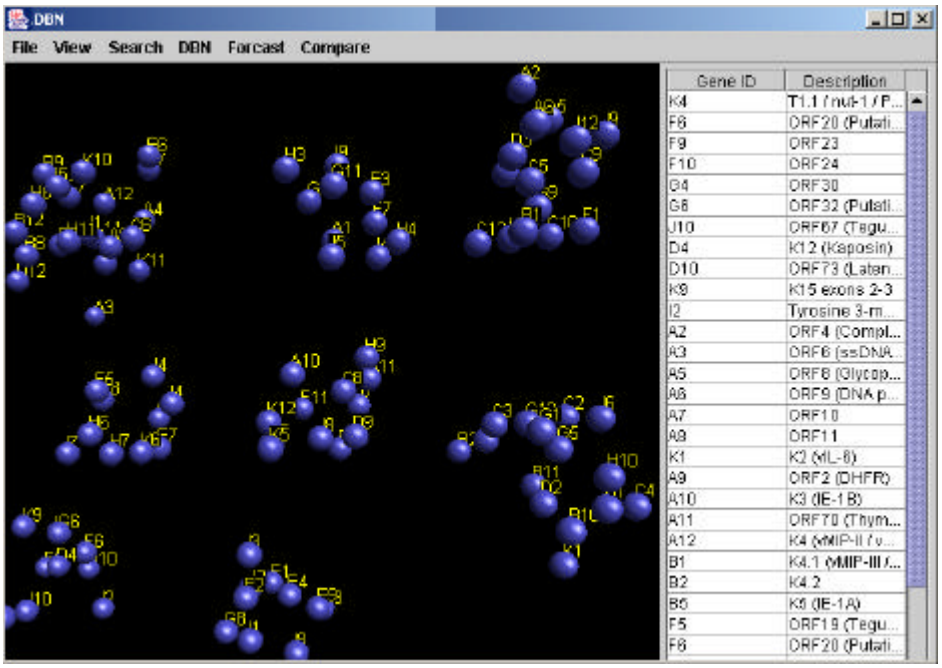

Figure 1. Global View

Both ambient light and directional light are added to the scene graph to enhance the 3-D effect. The text label above each gene is stored as raster image, which means whatever the view angle is changed to, the labels always face the viewer. This allows user change view angles freely without worrying about seeing the back of the labels.

Mouse function Orbit Behaviour is added to the 3-D scene graph, allowing user to zoom in/out or change view angles. Popup menus are allocated to each group area. User can select a particular group as a target for further inspections by click on the "group view" option in the popup menu.

A database was built to store the general information of each gene. Information in the database is displayed in the information table and linked with 3-D objects in the scene graph. Once a gene node in the 3-D scene graph is selected, the picking mouse behaviour function will return the "user data (Gene ID)" of the selected gene node and search for the corresponding datum in the database. Search results will be highlighted and scrolled to the top of the information table.

As we can see in figure 1, each gene node is represented as a small blue sphere with name on a 2-D label and allocated in a 3-D information space. Genes belong to the same group are clustered together. Through the 3-D scene graph, user can gain a global view of relationship between all gene nodes before they decide which group they are going to inspect.

4.2 Dynamic Bayesian Network

A Bayesian Network (BN) is a well-known model for performing probabilistic inference about a discrete system [Pearl88], with its dynamic counterpart additionally modelling a system over time [Dagu95]. A DBN allows the testing of the effect of changing a biological process by manipulating key components of the system. This is of particular interest to the pharmaceutical industry in trying to identify the best target in a modelled disease process for therapeutic intervention. A DBN consists of the following:

1. A set of N nodes representing the n variables in the domain at particular time slices and directed links between the nodes. Each node has a finite set of mutually exclusive states.

2. Associated to each node with a set of parents, there is an associated probability table.

To visualize five dimensions of variables (gene, time, relationships, expression level and probability) in a Dynamic Bayesian Network on a two-dimensional computer display, a 3-D model was built, and various visualization techniques were applied to the design. The essential idea is to provide as much information as possible without confusing or distracting the user with clutter and inappropriate colour schemes.

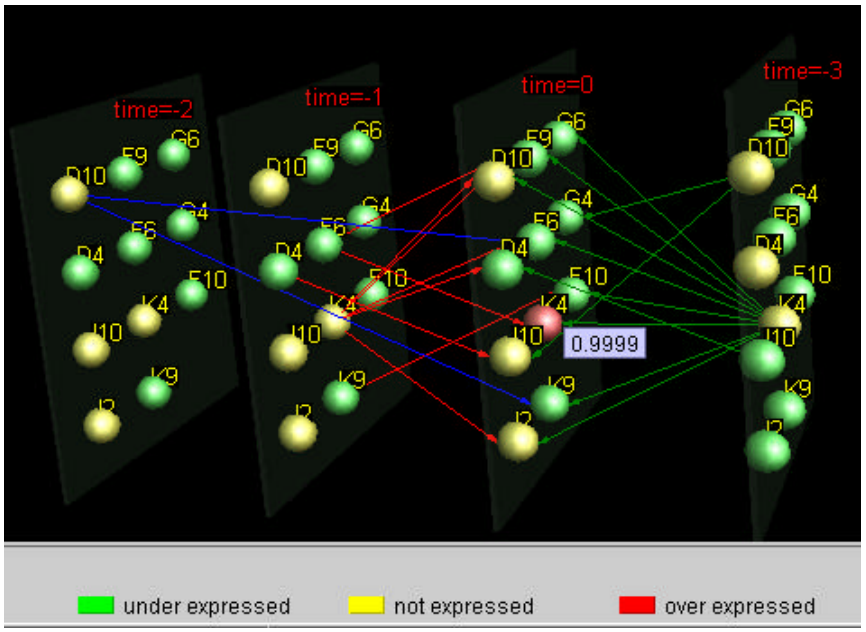

Figure 2. Dynamic Bayesian Network

- Expression level

Three colours are used to represent gene expression levels (red for over expressed, yellow for not expressed and green for under expressed). This visualization method can reduce user's cognitive load by reloading user's pre conceptions of colours in most DNA microarray images.

- Time

A series of half-transparent walls are paralleled in the 3-D space, each representing a time slice in the time series. This effect is visualized by setting transparency attribute (both transparency mode and transparency) of each wall object.

- Gene

Group of genes are represented as a group of small coloured spheres and are embedded in the half-transparent walls. As the expression level of each gene node in the DBN might change according to the changing of evidence, a switch is set to control the colour of each gene node.

- Relationship

Arrows are used to visualize links between genes from different time points. Each arrow consists of two geometry objects: a cone and a cylinder. Switches are added to control the colour and visibility of the arrow.

- Probability

Probability of each gene is stored in a tool tip for the gene node, i.e., when user puts the mouse nearby a particular gene node in the 3-D scene graph, the probability of the gene will appear under the gene node.

Various mouse functions are added to allow user interactions such as zoom in or out the information space, change view angles or drag a particular time point out in order to gain a clearer view of a particular relation.

Figure 2 shows a Dynamic Bayesian Network with three time lags. As we can see from the scene graph, the half-transparent time walls are almost invisible but still give out a strong indication of time changes. The first time point "time=-3" is dragged to the right side of the scene, providing a clearer view of the network structure. Colours are used to visualize expression levels and arrows are used to represent relationships. The probability is displayed in the tool tip under the selected gene node. Five dimensions of variables in the Dynamic Bayesian Network are visualized in a simple and natural way.

4.3 Forecasting

Instead of traditional line graph visualization technique [Unwin03], 3-D animation is used to visualize the multivariate time series generated during forecasting stage (see figure 3).

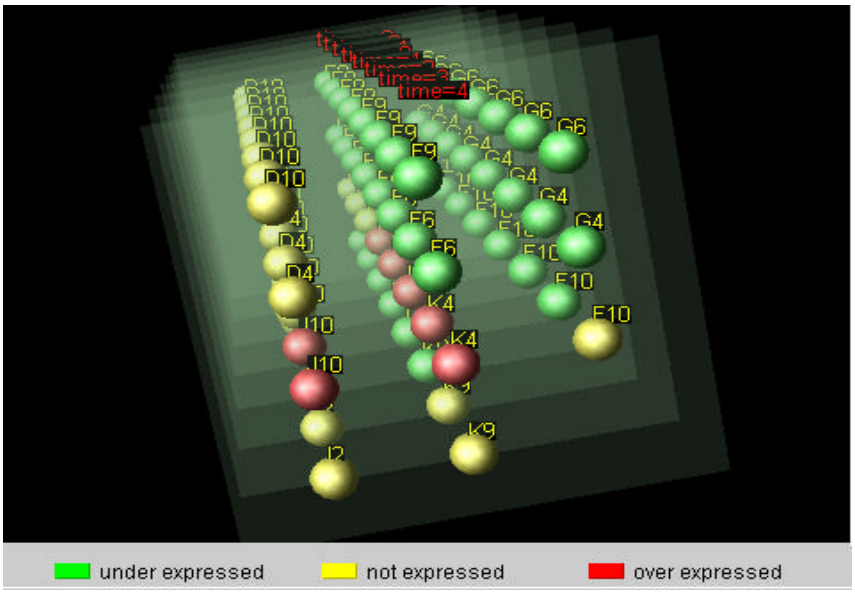

Figure 3. Forecasting

An interpolator object, together with an Alpha object is created to manipulate the parameter of a scene graph object to create the time-based animation. After loading first 5 values of each gene, the next 3 values of each gene will be predicted. The prediction result for each time point will fly through the display in time order. This visualization avoids an overlapped display and provides a vivid view of how gene expression level changes over time.

User can inspect further details in the whole time series by changing view angles, dragging a time point out and reset the view focus. The model also provides cause-and-effect insight into the operation of the network model by allowing the user to select values of gene expression level and see the effects.

4.4 Explanation

Focus +Context visualization is initially designed to help user quickly understand the nature of connections within a web locality [Mukherjea97]. We apply this technique in the explanation model design because a Focus +Context view can display the details of a particular node, plus the neighbouring nodes that are directly linked to the focal node.

Once a gene node is clicked, only the parents of the selected gene and their relationship in the Dynamic Bayesian Network will remain in colour. The probability will be displayed at the bottom of each coloured gene as a reference. All other genes will become half-transparent and all other relationships will become invisible. This is done by switching the appearance of gene nodes and relationships in the network when a gene node is selected. It aims at producing a simple and clear view, which only focuses on the explanation model hiding behind the Dynamic Bayesian Network.

Figure 4 illustrates one such explanation in the modelling of HHV8 data. First of all, gene C7 was observed "over expressed", but was not expressed one time slice earlier. These observations may be explained by the fact that gene H8 and gene B12 were over expressed two time slices earlier. They were in turn affected by the activities of gene A7, B6 and B12 in the previous time slices.

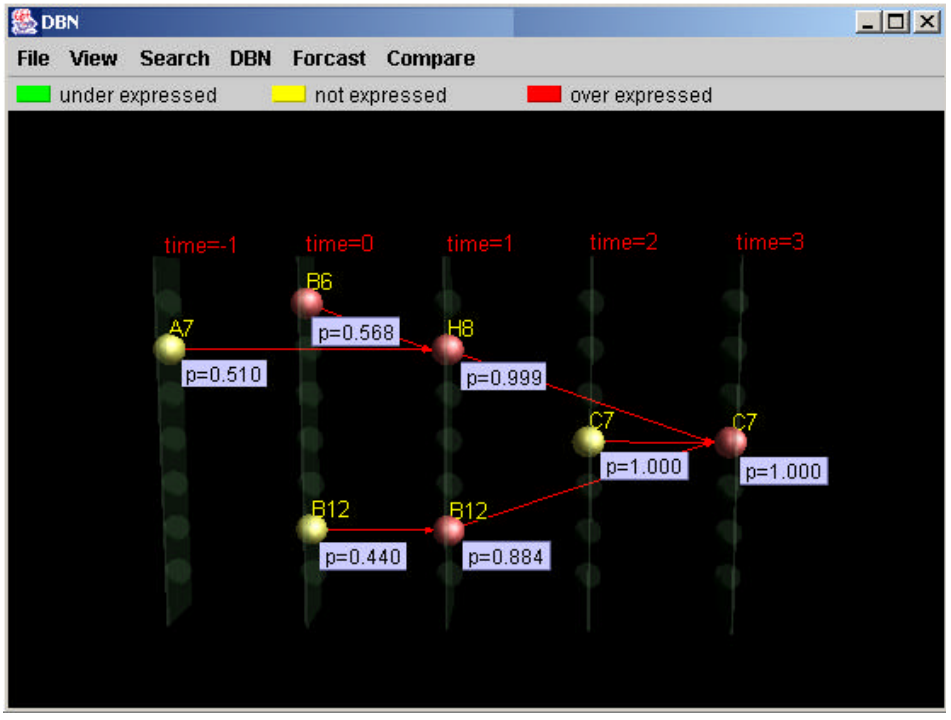

Figure 4. Explanation

5. Conclusion

A picture is worth a thousand words. Visualisation of the virus gene expression process and modelling results can greatly expand the range of models that can be understood by biologist. Here we suggested some visualization techniques that could be applied in each step of the strategy. A prototype was built to demonstrate the visualization effect of the modelling results.

As we can see from the prototype, a natural representation of the models was designed by applying various 3D visualization techniques. The 3D gene distribution model provides a global view of the whole gene groups. The half-transparent wall design gives out clear indication of time changes without hiding other objects from the view. The forecast model shows a cause-and-effect insight into the Dynamic Bayesian Network. The explanation model highlights the phenomena of interest in the network. For the first time, 3-D animation and Focus +Context technique is used to visualize multivariate time series gene expression data modelling. Various user interactions such as changing view angel, selecting values for feature, choosing explanation target are added to the models, providing the user with sufficient understanding and trust.

The work can be further improved by adding efficient algorithms to scale the variables and allocate 3-D objects (gene nodes, time walls) automatically as the current prototype only includes the graphical visualization of the modelling. Another challenge would be to fully integrate visualization with the virus gene expression data mining programming environment. Finally, an appropriate evaluation strategy will need to be developed to assess the aspect of visualization work on the comprehension of the modelling process and results by biologists.

Acknowledgement

We would like to thank Dr Allan Tucker for helpful discussions and Dr Paul Kellam for providing the data.

Reference

[Aoshima02] K., Aoshima, M., kawa, S., Tanaka1, "A Visualization Tool for Gene Network Discovery - G.NET", Genome Informatics 13: 445-446 (2002)

[Dagum95] P. Dagum, P., A. Galper, E. Horvitz, A. Seiver, " Uncertain reasoning and forecasting", International Journal of Forecasting 11(1): 73-87(1995)

[D'haeseleer2000] D'haeseleer, S Liang and R Somogyi "Genetic Network Inference: from Co-Expression Clustering to Reverse Engineering", Bioinformatics, 16:707-726(2000)

[Jenner2001] Jenner, R.G., Alba, M.M., Boshoff, C., Kellam, P "Kaposi's Sarcoma-Associated Herpesvirus Latent and Lytic Gene Expression as Revealed by DNA Arrays", Journal of Virology 75(2): 891-902(2001)

[Keim02] D. Keim, "Information Visualization and Visual Data Mining", IEEE Transactions On Visualization and Computer Graphics, 7(1), 1-8(2002)

[Kellam02] P Kellam, X Liu, N Martin, C Orengo, S Swift and A Tucker, "A framework for Modelling Virus Gene Expression Data", Intelligent Data Analysis, 6:265-279 (2002)

[Lockhart2000] B. Lockhart and E. Winzeler, Genomics, Gene Expression and DNA Arrays, Nature 405: 827-836(2000)

[Mukherjea97] S, Mukherjea, Y, Hara, "Focus +Context view of World-wide Web nodes", Proceedings of Hypertext 97 (Southampton, UK), ACM Press, 187-196

[Pearl88] Pearl, J., "Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference", Morgan Kaufmann Publishers, San Mateo, CA (1988)

[Quacknebush01] Quacknebush, J, "Computational genetics: computational analysis of microarray data", Nature Reviews Genetics, 2: 418-427(2001).

[Robinson97] Alan J. Robinson and Tomas P. Flores "Novel Techniques for Visualising Biological Information" Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, 241-249(1997)

[Tang02] C. Tang, L. Zhang, A. Zhang, "Interactive Visualization and Analysis for Gene Expression Data", 35th Annual Hawaii International Conference on System Science HICSS'02, 6: 143(2002)

[Thearling01] Kurt Thearling, Barry Becker, Dennis DeCoste, Bill Mawby, Michel Pilote, and Dan Sommerfield "Visualizing Data Mining Models", Published in Information Visualization in Data Mining and Knowledge Discovery, edited by Usama Fayyad, Georges Grinstein, and Andreas Wierse. Morgan Kaufman, 205-222(2001)

[Unwin03] Antony Unwin, Graham Wills, "Exploring Time Series Graphically", Articficial Computing & Graphics Newsletter, 2:13-15 (2003)

[Zapata-Rivera99] J.D., Zapata-Rivera, E., Neufeld, J., Greer, "Visualization of Bayesian Belief Networks", IEEE Visualization 1999 Late Breaking Hot Topics Proceeding. San Francisco, CA. 85-88(1999)