Clustering Gene Expression Data using Continuous Markov Models

Adele H. Marshall
Statistics & OR Research Division
Queen's University of Belfast
Northern Ireland BT7 1NN
a.h.marshall@qub.ac.uk

Roy Sterritt
School of Computing and Mathematics
University of Ulster at Jordanstown
Northern Ireland BT37 0QB
r.sterritt@ulster.ac.uk

Abstract

The Coxian phase-type distribution is a special type of Markov model that represents a process as consisting of phases or states which change as time progresses. It is this phase-type representation of the process that makes the Coxian phase-type distributions appealing to use by easing the complexity of the system and in some cases further reducing the amount of cumbersome numerical calculations required. The main aim of this paper is to introduce a new clustering technique based on the Coxian phase-type distribution. The modelling technique is still formed using a continuous Markov model but in addition allows the modelling of similar characteristics within each cluster. Such a technique will provide insights into the data by identifying clusters whereby each cluster may be described as behaving in a certain way or cluster members behaving in similar ways. The clustering technique in this paper may be used to model the behaviour of gene expression data over continuous time.

1. Introduction

A considerable amount of research has previously been carried out on clustering genes on the basis of their expression patterns. Various statistical techniques have been developed using for example Euclidean and mutual information cluster analyses to represent discretised gene expression sequences. Such work is valuable in assisting with the identification of genes that behave in similar ways or alternatively allows the grouping of similar responses for example to growth conditions, mutations and drugs.

The main aim of this paper is to introduce a statistical technique that may be used to investigate expression data to enable the identification of unknown clusters and further highlight any common characteristics that may exist within such clusters. Such cluster analysis of genes is based on their behaviour over continuous time.

2. Statistical Models

In statistical theory, Markov models are often used to represent stochastic processes. The theory of general Markov models has been considered by Bartholomew [1]. The model assumes a probabilistic behaviour of objects moving around the system and therefore is useful at providing a realistic representation of real world situations.

2.1. Phase-Type Distributions

Phase-type distributions describe the time to absorption of a finite Markov chain in continuous time, when there is a single absorbing state and the stochastic process starts in a transient state [2]. The models describe duration until an event occurs in terms of a process consisting of a sequence of latent phases - the states of a latent Markov model. In fact the assumptions of the distributions state that the 1,…,n states are all transient, so absorption into the state n+1, from any initial state is certain.

There are many examples in the literature where phase-type distributions are being used, originally within the applied probability domain but now also as a tool for data analysis. Applications are wide ranging from calculating the expected load of mobile phone networks [3], to analysing the duration of stay of elderly patients in hospital [4]. Previous research [5] used this model to find a suitable distribution for the duration of stay of a group of geriatric patients in hospital. They found that the phase-type distributions were ideal for measuring the survival times, the lengths of stay of patients in hospital and showed how it was also possible to consider other variables that may influence duration.

Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference Workshops (CSBW'05)
0-7695-2442-7/05 $20.00 © 2005 IEEE

For example, the length of time a patient spends in hospital can be thought of as a series of transitions through phases such as: acute illness, intervention, recovery, or discharge. As such, phase-type distributions can represent the diverse nature of the patient lengths of stay where there may be a large variation in the amount of time patients spend in hospital.

By using matrix notation to represent phase-type distributions, the calculations become more manageable replacing the somewhat cumbersome numerical integrations that can occur in many situations. In addition, the distributions have the ability to describe detailed information about the behaviour of the stochastic models while also allowing the *lack of memory* property to exist. Another benefit in using phase-type distributions is that in most instances the models can be generalised to include almost all continuous distributions [2] such as the exponential (one phase), the Erlang and mixed exponential distributions.

The phase-type distribution relates directly to the statistical approach known as survival analysis.

2.2. Survival Analysis

Survival analysis is a statistical methodology which models duration until a particular event occurs. The time (in months, weeks, or days) is measured from the beginning of the follow-up of an individual until the event occurs [6]. This is commonly known as survival time as it measures the length of time that an individual survives. The event, often referred to as a failure, is usually a negative individual experience such as the occurrence of death, disease incidence or relapse from remission ([7], [8]).

Kleinbaum [6] discusses examples of survival analysis such as predicting the length of time in remission for leukaemia patients, heart transplant patients' time until death and the time taken by subjects to complete specified tasks in a psychological trial. These are all considered to be survival analysis problems due to the nature of the data where there is an outcome variable of time until an event occurs. Other examples, discussed by Collett [9], include the prognosis of women with breast cancer where survival analysis is used to predict the survival prospects of breast cancer patients who may be developing secondary tumours.

2.2.1. Censored Data. A special feature of survival data is the possibility that the exact survival times of some individuals may be unknown and not observed for the full time to failure [8]. This was initially a focus in clinical trials when at the close of a trial, or at the current time in the database, patients may have survived without ever experiencing the event. Alternatively the survival time of the patient may not be known due to the patient withdrawing from the study or being lost to follow-up during the study period. Such an incomplete observation of failure time is called *censoring* [6]. The measure of survival for these individuals is then referred to as the censored survival time measured from time of entry to the study until the last recorded time when the individual is in the study. To identify the censored data, a new variable is introduced, δ where δ=(0,1) random variable such that $\delta = 1$ if failure and $\delta = 0$ if censored.

This kind of censored data is referred to in the literature as right censored as the censoring has occurred after the individual has entered into the study, that is, to the right of the last known survival [9]. Left censored data occurs when the actual survival time for an individual is less than that observed, where the survival time is incomplete at the left side of the follow-up period.

2.2.2. Functions of Survival Analysis. In summarising survival analysis, there are three functions of central interest namely;

- the survivor function, denoted by S(t);
- the probability density function, *p.d.f.*, denoted by f(t), and,
- the hazard function, denoted by h(t).

These three functions are mathematically equivalent, that is, if one of them is given, the other two can be derived [10].

Let T denote the random variable for a person's survival time. As T is a measure of time, the possible values of T will include all non-negative numbers. The actual survival time of an individual may then be defined as t. The survivor function is defined as the probability that an individual survives longer than t.

$$S(t)=P(\text{an individual survives longer than } t) = P(T > t). \quad (1)$$

Due to the generality of phase-type distributions, the estimation of parameters for the survival function and *p.d.f.* may become difficult. To overcome this problem the following Coxian phase-type distributions were introduced.

3. Coxian Phase-Type Distributions

Coxian phase-type distributions [11] are a special type of phase-type distribution that describes the probability P(t) that the process is still active at time t [4] where the transient states (or phases) of the model are ordered. The process begins in the first phase but is restricted in the sense that it must either progress through the phases sequentially or enter into the absorbing state (the terminating event - phase n+1). In the case of medicine, transitions through the ordered transient states could correspond to various stages in the patients disease progression for example diagnosis, assessment, rehabilitation and long-stay care where patients eventually discharge, transfer or die.

A Coxian phase-type distribution $\{X(t); t \geq 0\}$ may be defined as a Markov chain in continuous time with states $\{1, 2, ...n, n + 1\}$, X(0) =1, and for i= 1, 2, ...n-1

$$\text{prob}\{X(t+\delta t) = i+1 \,|\, X(t) = i\} = \lambda_i \delta t + o(\delta t) \qquad (2)$$

and for i= 1, 2, ...n

$$\text{prob}\{X(t+\delta t) = n+1 \,|\, X(t) = i\} = \mu_i \delta t + o(\delta t). \qquad (3)$$

Here states $\{1, 2, ...n\}$ are latent (transient) states of the process and state n + 1 is the absorbing state. The transition from state i to state (i+1) through the ordered transient states is represented by λ_i while the transition from state i to the absorbing state (n+1) is denoted by μ_i.

The Coxian phase-type distribution, illustrated in Figure 1, is defined as having a transition matrix **Q** of the following form, where the λ_i's and μ_i's are from Cox and Miller's [12] further developed theory of Markov chains defined by (2) and (3).

$$\mathbf{Q} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \dots & 0 & 0 \\ . & . & . & & . & . \\ . & . & . & & . & . \\ . & . & . & & . & . \\ 0 & 0 & 0 & \dots & -(\lambda_{n-1}+\mu_{n-1}) & \lambda_{n-1} \\ 0 & 0 & 0 & \dots & 0 & -\mu_n \end{pmatrix}, \qquad (4)$$

Figure 1 is an illustration of a phase-type distribution where λ_i represents the transition from phase i to phase (i+1) and μ_i the transition from phase i to the absorbing phase (n+1).

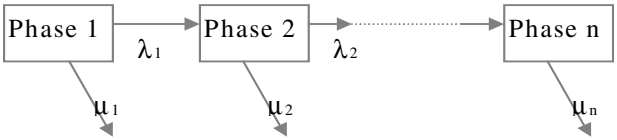

Figure 1. Illustration of phase-type distributions

The system represents the continuous variable as notional partitions of time referred to as stages or phases of the system. An object will start in phase 1 of the system, stay there for a time period until it leaves the system or completely at a rate of μ_1, or continue to stay in the process by moving into the next stage or phase, phase 2. The movement from phase 1 to phase 2 of survival time is represented by the rate, λ_1. The final transition is when the object leaves the system completely thus reaching the absorbing state of the process.

The survival probability that X(t) = 1,2,…,n is given by

$$\mathbf{S}(t) = \mathbf{p} \exp\{\mathbf{Q}t\}\mathbf{1} \qquad (5)$$

where

$$\mathbf{p} = (1\ 0\ 0\ ...0\ 0\ ), \qquad (6)$$

and **1** is a column vector of 1's and **Q** is the transition matrix as previously shown in (4). The *probability density function (p.d.f.)* of T then follows by differentiation:

$$f(t) = \mathbf{p} \exp\{\mathbf{Q}t\}\mathbf{q} \qquad (7)$$

where

$$\mathbf{q} = -\mathbf{Q}\mathbf{1} = (\mu_1\ \mu_2 .. \mu_n)^T. \qquad (8)$$

Previous work [13] has used this model to find a suitable distribution for representing the duration of stay of a group of geriatric patients in hospital.

4. Coxian Phase-Type Clustering (PhC)

Clustering is a statistical process that partitions a data set into homogeneous groups or clusters of objects such that objects in the same cluster are more similar among themselves than to those in other clusters [14]. Given a dataset the clustering algorithm separates the objects or elements into clusters where objects within a cluster are alike in nature, that elements from different clusters have low similarity to each other and clusters are non-overlapping.

There are two main forms of clustering algorithm available namely, those that have supervised learning and those that are unsupervised.

This paper introduces a new clustering technique called Coxian Phase-type Clustering (PhC). PhC uses the Coxian phase-type distribution to cluster data according to an underlying continuous survival time. In doing so, the phases in the survival distribution are mapped onto clusters in the data set and common characteristics identified within each cluster using

standard statistical techniques.

4.1. Motivating Example

Preliminary work on this technique ([5],[13]) clustered patients in a data set according to their continuous survival times (length of stay) in a UK hospital. The data set was found to be best represented by 3 clusters of patient mapped from 3 survival phases. The three clusters of patient were interpreted as those in acute care in hospital, those in rehabilitation and those in long stay. On investigation of patients, those within each cluster possessed common characteristics and displayed similar behaviours. For example, those patients in the acute cluster mainly comprised of those patients who are discharged home or in such an acute state on arrival to hospital that they die after a short period of time. The second phase considered those patients who have a much longer duration of stay in hospital with quite distinguishable characteristics compared to the other 2 clusters in particular with respect to their admission reason to hospital, their dependency score and outcome on departure from hospital. This cluster mainly comprised of patients who eventually transferred to another form of care, such as nursing home care.

The third and final cluster was extremely useful in highlighting particular patients with very long, extreme lengths of stay in hospital. In fact, such patients are referred to in the UK National Health Service (NHS) as *bed blockers* as they should have been discharged from hospital a long time previous but instead are still using hospital beds blocking them from use by other patients. Some of the common characteristics of patients in this cluster include patient gender as female and age between 57 and 64 years (where the age range for the full data set is between 42-105 years).

As such, the study on patient length of stay demonstrated how the proposed clustering technique can provide insights into the data by identifying clusters whereby each cluster may be described as behaving in a certain way or cluster members behaving in similar ways. The paper concluded that the identification of patient clusters and common characteristics within such, offered huge potential for hospital managers and clinicians who had valuable insight into the overall management and bed allocation of the hospital wards.

4.2. Fitting PhCs

The clustering technique is unsupervised in that the number of clusters can be determined by the fit of the data. Alternatively, if the number of clusters is previously known, the algorithm can be simplified to partition the data for that specified number of clusters.

In order to perform PhC, the Coxian phase-type distribution has to be fitted to the continuous time variable. The parameters of which can be estimated using the maximum likelihood function and the Nelder-Mead simplex algorithm [15]. The Nelder-Mead algorithm is a non-gradient approach which uses a simplex formed by a set of (n+1) mutually equidistant points in n dimensional space. The method compares the values of the function at the (n+1) vertices using the simplex which it then guides towards the optimum point during the iterative process. The three basic operations used to direct the simplex are reflection, expansion and contraction. The approach is considered a very robust, powerful technique.

Table 1. Coxian phase-type clustering algorithm

Input: A dataset containing m objects, and Corresponding survival times t_i each for i^{th} object.

Output: A set of c clusters with associated likelihood value

Method:

1 Initialise variables c=0, likelihoodc_1=0
2 Repeat
3 likelihood=0
4 c=c+1
5 define **p**, **Q**, and **q** for PhC with c clusters
6 for i = 1,…, m
7 read survival time t_i
8 likelihood
= likelihood + log(**p**×(exp(**Q**×t_i))×**q**)
9 end
10 estimate λ_i and μ_i parameters using Nelder Mead algorithm and likelihood value
11 compare likelihood values for c clusters and c-1 clusters
12 likelihoodc_1=likelihood
13 Until no change in Likelihood or insignificant change in likelihood

The PhC Algorithm, Table 1, begins by fitting a Coxian phase-type distribution with one phase, (when c=1) corresponding to the exponential distribution, to the continuous survival data. This is implemented and parameters estimated using the Nelder Mead Algorithm along with the following likelihood function

$$L = \sum_{i}^{n} \log\left(\mathbf{p}\exp\{\mathbf{Q}t_i\}\mathbf{q}\right). \quad (9)$$

The fitting process continues as a sequential procedure whereby a series of Coxian phase-type distributions are fitted and assessed. This is performed by repeatedly adding an additional phase to the distribution (c=c+1), fitting the data, estimating the parameters for the distribution, calculating the likelihood of the new model and repeating the process until c phases are tried with the process terminating when there is very little or no improvement made to the fit from adding an additional phase. Such an approach may employ a series of likelihood ratio tests [8] to compare model likelihoods to assess whether there is significant improvement by adding an additional phase.

The PhC Algorithm is implemented using the MATLAB [16] mathematical software package to perform the likelihood ratio tests which determine the most suitable number of clusters for the data set.

The following formula is derived in order to represent the length of stay in terms of k phases. Let π_i be the probability that an object leaves the system from Ph_i. This can be calculated by taking the probability density formula for each phase or state as follows. For example when i=1, the *p.d.f.* is

$$f(t) = p\exp\{Qt\} = \mu_1 e^{-(\lambda_1+\mu_1)t} \tag{10}$$

and the probability that the object leaves the system from phase 1 is

$$\pi_1 = \int_0^\infty \mu_1 e^{-(\lambda_1+\mu_1)t}dt = \frac{\mu_1}{\lambda_1+\mu_1}. \tag{11}$$

Similarly,

$$\pi_2 = \int_0^\infty \mu_2 e^{-(\lambda_2+\mu_2)t}dt \int_0^\infty \lambda_1 e^{-(\lambda_1+\mu_1)t}dt \tag{12}$$

$$= \left(\frac{\lambda_1}{\lambda_1+\mu_1}\right)\left(\frac{\mu_2}{\lambda_2+\mu_2}\right),$$

.

.

and

$$\pi_k = \left(\frac{\lambda_1}{\lambda_1+\mu_1}\right)\left(\frac{\lambda_2}{\lambda_2+\mu_2}\right).....\left(\frac{\lambda_{k-1}}{\lambda_{k-1}+\mu_{k-1}}\right). \tag{13}$$

A data set of individuals or objects may then be divided into clusters according to their survival time, where the clusters are represented by c_k determined by the following equation:

$$c_k = \left\{x^{(j)} : m\sum_{i=1}^{k-1}\pi_i < j \le m\sum_{i=1}^{k}\pi_i\right\}, \tag{14}$$

for k = 1, …, c where $x^{(1)}$, … , $x^{(m)}$ represents the ordered survival time for each object and m represents the number of patients or objects in the data set. The characteristics within each cluster may then be examined to determine commonalities.

4.2.1. PhC for Censored Data. One feature considered by survival analysis is the inclusion of missing or censored data. This problem can be overcome for simplified cases of the coxian phase-type model by estimating survival using the EM (Expectation - Maximisation) algorithm [17]. For example, the EM algorithm can be used for data where say the outcome is typically missing since for some individuals for example when modeling time until relapse, the patients may not experience a relapse by the time the data collection ceases. For some models such incompleteness may be straightforwardly taken account of by incorporating appropriate terms into the likelihood functions. The likelihoods may then be maximised using the EM algorithm [17]. Use of the EM algorithm with phase-type distributions has been described by Aalen [18].

The approach is illustrated using the following two cluster model. Let $x_1,\ldots x_p$ be continuous survival time for say, patients who leave hospital alive. Let $y_1,\ldots y_q$ be the continuous survival time for those patients that are known to die whilst in hospital. Let $z_1,\ldots z_r$ be the continuous survival for those patients who have not yet left hospital, that is, those cases where outcome is unknown, the censored data. The likelihood function for patient survival is then:

$$L = \prod_{i=1}^{p}(a\lambda_1 e^{-\lambda_1 x_i})\prod_{i=1}^{q}(b\lambda_2 e^{-\lambda_2 y_i})\prod_{i=1}^{r}(a\lambda_1 e^{-\lambda_1 z_i} + b\lambda_2 e^{-\lambda_2 z_i}) \tag{15}$$

The parameters a, b, λ_1 and λ_2 are then fitted to maximise the likelihood. Using the EM algorithm [17], this maximisation is achieved by replacing the incomplete data by its expectation in the expressions for maximum likelihood estimators for complete data. The following iterative equations are derived by differentiating the likelihood of the model and solving for the parameters a, λ_1, and λ_2

$$a^{(n+1)} = \frac{p + a^{(n)}r}{p+q+r} \tag{16}$$

$$\lambda_1^{(n+1)} = \frac{p + a^{(n)}r}{\sum_{i=1}^{p} x_i + a^{(n)}\sum_{i=1}^{r}(z_i + (\lambda_1^{(n)})^{-1})}$$

$$\lambda_2^{(n+1)} = \frac{q + (1-a^{(n)})r}{\sum_{i=1}^{q} y_i + (1-a^{(n)})(\sum_{i=1}^{r}(z_i + (\lambda_2^{(n)})^{-1})}$$

In this two term case, b can easily be calculated by subtracting a from 1. Here the complete estimators of the probabilities a and b are just the proportions who leave hospital alive and dead respectively while the complete estimators of the exponential parameters are given by the number of discharges of the appropriate type (alive or dead) divided by the total time spent in the hospital. The EM iterations then estimate, for example, the number of live discharges in the incomplete data by the number of incomplete observations r multiplied by the probability a of these ending in a death. The expected time spent in hospital by a patient censored at z who eventually leaves hospital alive is $z + \lambda_1^{-1}$; similar expressions may be obtained for patients who die in hospital. Dempster et al. [17] show that the iterative scheme converges to the maximum likelihood solution but may be slow. MATLAB [16] software may be used to speed up the process.

5. Clustering Gene Expression Data

A fundamental progression in gene expression data analysis, was the application of techniques capable to identifying subsets of genes that possess similar expression patterns based on the analysis of gene expression data. Such techniques emerged in the form of clustering algorithms [19].

There has been tremendous research and development in clustering gene expression data, both in terms of standard statistical clustering techniques currently in existence and the emergence of new methods specifically focused on gene expression data. Clustering can help identify groups of genes that have similar expression patterns under various conditions or across different tissue samples [20].

The application of current statistical techniques include the widely known and well accepted methods of survival analyses mainly the Kaplan Meier (KM) technique, log-rank test and Cox Model ([6],[8]). Such applications involve the modeling of phenotypes, the observable or measurable traits of an individual as produced by its genotype and the environment.

The approach relates gene expression profiles to survival phenotypes by first grouping tumour samples into several clusters based on gene expression patterns across many genes, and then to use the Kaplan-Meier (KM) curves or the log-rank test to indicate whether there is a difference in survival time among different tumour groups [19]. Alternatively the Cox model has been used to model survival outcome based on clusters of gene expressions across different samples [21].

However, the Cox's model intrinsically assumes proportional hazards, that is the instantaneous or immediate potential per unit time of failure, given that the survival up to time t is proportional [7].
Li and Gui [21] also use Cox's model to develop an extension of the partial least squares method to construct predictive components which model survival. Similar to principle component analysis (PCA), the method examines the relationship among the variables and identifies linear combinations of the original variables as predictors. In addition to PCA, the method makes use of the response variable in constructing the latent components. However, the technique cannot accurately handle continuous survival distributions. As is the case with the cluster analysis system introduced by Eisen et al. [20] for analysing genome-wide expression data from DNA microarray hydridization.

Meltzer et al. [22] comment that new algorithmic approaches are required to accommodate the complexity of underlying rules of genome function emphasizing the particular importance in approaches which help identify critical genes and pathways which are essential to tumour growth and survival.

5.1. PhC of Gene Expression Data

As stated by Golub et al. [23], one of the most promising applications of gene expression analysis is the classification of tissue types to their gene expression profiles. This paper proposes the PhC technique as a method capable of doing just that. The application of the Coxian phase-type clustering technique to microarray data has the potential of greatly assisting the prediction of various clinical phenotypes based on the gene expression profile. This is similar in nature to the example discussed earlier in this paper where a continuous survival time is clustered according to some further information concerning the objects in the data set. When considering microarray data, the survival time could be for example the time a patient survives after treatment or the time to cancer relapse. The clusters will then be represented by linking gene expression profiles to the survival variable. Identification of the clusters is comparable to identifying different streams of behaviour within the heterogeneous data and relating these to survival phenotype. For example the different clusters could be considered as groups of patients or genes who are at varying degrees of risk to a negative experience such as time until cancer recurs in the body or time t until death. These degrees of risk are related to phenotype.

Due to the continuous nature of the Coxian phase-type distribution and its ease of representation as a

survival function, the resulting PhC model could be used to estimate hazard functions for the different cluster of patient. Such developments will facilitate the modeling of future samples thus permitting the prediction of patient survival according to clinically relevant high and low risk groups.

It is without question the practical implications of being able to make predictions of this nature. Certainly the ability to predict patient survival opens up a wealth of opportunities in the healthcare domain.

Another application is that of pharmacogenomics, the study of whole genomes or substantial numbers of genes in order to assess drug responses. For example, pharmocogenomics could assist in the identification of large-scale differences in the patterns of gene expression in response to chemical compounds or to classify cancers for predicting different patient's survival or response to drugs, varying levels of drugs or other treatments.

In summary, the PhC technique provides opportunity to analyse gene expression data for many important applications for example to classify cancers or to assess drug responses in pharamcogenomics and drug discovery. Previous studies such as Golub et al. [23] have shown how gene expression data can distinguish between similar cancer types thus assisting with diagnosis and treatment. Additionally the PhC technique can be used to model the progression of disease as consisting of various states or clusters of health and disease, associated with which is a list of common characteristics for each cluster. Thus highlighting distinguishing characteristics for the different clusters and the features that identify an element as belonging to one particular cluster as opposed to any other. Such information can facilitate the prediction of future disease status and survival at the molecular level.

When considering time to remission of cancer patients, the large variability in the data often results in censored survival phenotypes. As previously discussed, there is also the possibility of extending the PhC procedure for censored survival times by utilizing the EM algorithm. The construction of such mutually uncorrelated components, based on microarray gene expression data, provides an opportunity to investigate the objects within the clusters to capture common characteristics and similar behaviours. These may then be used for future inference, for example once a cluster model is determined for the survival times and common characteristics identified, it may be used as a means of predicting the outcome and survival times of future patients.

Alternative techniques can enforce additional assumptions for example, in the Cox's proportional hazards model, the hazard functions are assumed to be in proportion to each other which is not always the case for gene expression data. Whereas other clustering algorithms require the specification of the number of clusters in advance or that the data is normally distributed.

Most algorithms produce poor partitions in presence of outliers while PhC can correctly reveal the structure of data and identify outliers simultaneously. Such versatility of the PhC models is partly due to the nature of the Coxian phase-type distribution capturing the structure of the data, which may be skewed, may contain outliers and may have censoring.

Indeed the nature of the PhC technique lends itself nicely to a plethora of applications within gene expression modelling.

6. Summary and Further Work

This paper introduces a statistical procedure, the Coxain Phase-type Clustering technique (PhC) as a new method of separating data into groups according to the continuous survival of the objects in the data set. In doing so each cluster is considered to only contain objects that are alike as possible, that is, the objects within a cluster display similar characteristics. Objects that do not belong to the same cluster are expected to have significantly different characteristics. Such a clustering algorithm can be utilized to model gene expression data.

There is a growing wealth of research focused on clustering genes on the basis of their expression patterns. This has been both in terms of applying standard statistical clustering techniques currently in existence and the emergence of new methods with specific attention on gene expression data. However, there are drawbacks and in some cases additional assumptions associated with these techniques. As such the PhC model is introduced as an alternative method for investigating gene expression data to enable the identification of unknown clusters and then further highlight any common characteristics that may exist within such clusters.

The PhC model represents the continuous survival variable as a Coxian phase-type distribution, a special type of Markov model that represents a process as consisting of phases or states which change as time progresses. These phases are then the basis for the clusters of data. Such a technique has potential to be applied to many areas of gene expression modeling for

instance in prediction of cancer survival times, in drug development and responsiveness to treatments and pharmacogenomics.

The technique, unlike its competitors, also has fewer assumption enforced upon it, such as the nature of the underlying distribution of survival, the restriction of having to specify in advance the number of clusters, and the limitations that some methods place on the type of data used. The PhC model can accommodate data that contains outliers, includes censoring and most importantly is specifically tailored for continuous survival times. One drawback, however is that fitting the PhC to huge data sets can be time consuming and as such further research is currently underway to improve such a fitting process.

Currently, covariates of the data can be incorporated into the various clusters in the PhC model, however it is possible to develop this further. One approach currently being developed is the formulation of a new method using Bayesian network theory to model a network structure of covariates to exist within each cluster of gene expression data.

References

[1] D.J. Bartholomew, *Stochastic Models for Social Processes*, 3rd edn, London, Wiley, 1982.

[2] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach.* John Hopkins University Press, 1981.

[3] P. Fazekas, S. Imre and M. Telek, "Modeling and analysis of broadband cellular networks with multimedia connections", *Telecommunications Systems*, 2002, 19 (3-4), pp. 263 - 288.

[4] M.J. Faddy, "Examples of fitting structured phase-type distributions", *Applied Stochastic Models and Data Analysis*, 1994, 10, pp. 247 - 255.

[5] A.H. Marshall, and S.I. McClean, "Using Coxian Phase-Type Distributions to Identify Patient Characteristics for Duration of Stay in Hospital", *Health Care Management Science Journal*, 2004, 7(4) pp. 285 - 289.

[6] D.G. Kleinbaum, *Survival Analysis - A Self Learning Text*, 2nd edn, New York, Springer-Verlag, 1997.

[7] J.D. Kalbfleisch and R.L. Prentice, *The Statistical Analysis of Failure Time Data*, New York, Wiley, 1980.

[8] D.R. Cox and D. Oakes, *Analysis of Survival Data*, Chapman and Hall, 1984.

[9] D. Collett, *Modelling Survival Data in Medical Research*, Chapman and Hall, 1984.

[10] E.T. Lee, *Statistical Methods for Survival Data Analysis*, Belmont, CA: Life-time Learning Publications, 1980.

[11] D.R. Cox, "A use of complex probabilities in the theory of stochastic processes", *Camb. Phil. Soc.* 1951, pp. 313 - 319.

[12] D.R. Cox and H.D. Miller, *The Theory of Stochastic Processes*. London, Methuen, 1965.

[13] A.H. Marshall, *Bayesian Belief Networks Using Conditional Phase-type Distributions,* DPhil Thesis, University of Ulster, 2001.

[14] B. Everitt, S. Landau, and M. Leese, *Cluster Analysis*, Hodder Arnold Press, 4th edition, 2001.

[15] J.A. Nelder, and R., Mead, "A simplex method for function minimization", *Computer Jnl*, 1965, 7, pp. 308-313.

[16] MATLAB, *MATLAB Reference Guide*, MathsWorks Inc., Natick, Massachusetts, 1992.

[17] A., Dempster, D. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society* - Series B, Methodological, 1977, 39, pp. 1 - 38.

[18] O.O. Aalen, "Phase-type Distributions in Survival Analysis", *Scandinavian Journal of Statistics*, 1995, 4, pp. 447 - 463.

[19] H. Li, K. Zhang, T. Jiang, "Minimum Entropy Clustering and Applications to Gene Expression Analysis", *Proc. of IEEE Computational Systems Bioinformatics Conference*, IEEE CS Press, 2004, pp. 142-151.

[20] M.B. Eisen, P.T. Spellman, P.O. Browndagger, and D. Botstein, "Cluster Analysis and Display of Genome-wide Expression Patterns". *Proc. National Academy of Sciences of USA*, 1998, 95(25), pp. 14863-14686.

[21] H. Li and J. Gui, "Partial Cox Regression Analysis for High-dimensional Microarray Gene Expression Data", *Bioinformatics*, 20(1), 2004, pp. i208-i215.

[22] P. Meltzer, S. Davis, K. Baird and Y. Chen, "Are we There Yet? Genomic Profiling and Mechanism in Cancer Research", *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, IEEE CS Press, 2004, pp. 6.

[23] T. Golub, D. Slonim, P. Tamayo, C.M. Huard, J.M. Caasenbeek, H. Collier, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, E. Lander, "*Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*", Science 1999, 286; pp. 531 – 537.