Increasing Power by Sharing Information from Genetic Background and Treatment in Clustering of Gene Expression Time Series

SuraZakiAlrashid
Information Technology College,
University of Babylon, Babil, Iraq
Sura_os@itnet.uobabylon.edu.iq

Muhammad ArifurRahman
Sheffield Institute for
Translational Neuroscience
University of Sheffield,
Sheffield,UK
M.Rahman@dcs.shef.ac.uk

Nabeel H. Al-Aaraji
Ministry of Higher
Education and
Scientific Research,
Baghdad, Iraq

Neil D. Lawrence
Sheffield Institute for Translational Neuroscience
University of Sheffield, Sheffield, UK
N.Lawrence@sheffield.ac.uk

Paul R. Heath
Sheffield Institute for Translational
Neuroscience University of Sheffield,
Sheffield, UK
P.Heath@sheffield.ac.uk

Abstract

Clustering of gene expression time series gives insight into which genes may be co-regulated, allowing us to discern the activity of pathways in a given microarray experiment. Of particular interest is how a given group of genes varies with different conditions or genetic background.

In this paper,this paper develops a new clustering method that allows each cluster to be parameterised according to whether the behaviour of the genes across conditions is correlated or anti-correlated. By specifying correlation between such genes,more information is gain within the cluster about how the genes interrelate. Amyotrophic lateral sclerosis (ALS) is an irreversible neurodegenerative disorder that kills the motor neurons and results in death within 2 to 3 years from the symptom onset. Speed of progression for different patients are heterogeneous with significant variability. The SOD1^{G93A} transgenic mice from different backgrounds (129Sv and C57) showed consistent phenotypic differences for disease progression. A hierarchy of Gaussian isused processes to model condition-specific and gene-specific temporal co-variances. This study demonstrated about finding some significant gene expression profiles and clusters of associated or co-regulated gene expressions together from four groups of data (SOD1^{G93A} and Ntg from 129Sv and C57 backgrounds). Our study shows the effectiveness of sharing information between replicates and different model conditions when modelling gene expression time series. Further gene enrichment score analysis and ontology pathway analysis of some specified clusters for a particular group may lead toward identifying features underlying the differential speed of disease progression.

General Terms
Design, Experimentation

Keywords *:* Gaussian process, kernel, hierarchical clusters, amyotrophic lateral sclerosis.

1. Introduction

The dynamic behaviour or analysis of time series data in particular clusters is important for exploring and understanding gene networks. In many conventional time series models, one key requirement is data with regular intervals. Gene expression experiments data with regular intervals might be less informative or may not be optimal from a statistical perspective or even may not be cost effective for various reasons. A model designed to obtain data with regular intervals may not elicit as much information as a method designed to collect pertinent special temporal features. Again, in many cases multiple biological replicates are available when the same experiments are repeated multiple times. For these cases simply considering only one experiment or taking the mean values from different replicates may not be the best solution. Interesting information might be discarded while dealing only with one data set or with their mean values.

Amyotrophic lateral sclerosis (ALS) is a diverse neurodegenerative disorder with around 10% of familial cases and the remaining sporadic. The disease is currently irreversible from onset and heterogeneous with variable severity in terms of speed of progression of the disease course. Injury and cell death of motor neurons in the brainstem, spinal cord and motor cortex are the main reasons of this relentlessly progressive disorder [Haverkamp *et al.*,1995, Ferraiuolo *et al.*, 2011; Peviani *et al.*,2010, and A. Brockington et al., 2013]. Among the familial ALS [fALS] 20% is caused by mutation in the Cu/ZnSuperoxide Dismutase1 (SOD1) gene. The median survival of this lethal disorder is less than 5 years, only 20% patients live longer than 5 years and less than 10% patients survive more than 10 years from the symptom onset [Beghi *et al.*, 2011, R. A. Saccon,et al., 2013]. The speed of disease progression is not clear from the biological basis. Even in fALS, affected members clearly show the clinical het- erogeneity in terms of site of onset, age and progression rate of the disease. [Camu *et al.*, 1999] reported the presence of potential gene modifiers and pathways that particularly affect the disease phenotype. Mutation in the SOD1 gene notably characterized the distinctive nature by intrafamilial and interfamilial variabilities in the phenotype. Many of the clinical and pathological features of human ALS can be replicated very well by transgenic mice. These murine models also mimic the human disease and show the heterogeneity in the disease progression for the clinical phenotype. These variabilities may be related with expression levels of mutant SOD1 protein or specific SOD1 mutations [Turner *et al.*, 2003].

In a previous study [Pizzasegola, 2009 ; Alrashid and Al-Aaraji, 2015]it was reported that disease progression is much faster in 129Sv mice with the survival of 129±5 days, while the C57 mouse strain can survive 180±16 days. Both the 129Sv and C57 carry the same copy numbers of human mutant SOD1 and express the same amount of mutantSOD1^{G93A} messenger RNA in the spinal cord. [Marino *et al.*, 2015] reported about the differences in protein quality control of these mouse models in terms of speed of progression of the disease course.

The aim of our paper is to specify the significantly different genes that may affect the speed of ALS progression by building a new model. Gaussian process is used and here a coregionalization principle is introduced while developing the kernel of the Bayesian hierarchical Gaussian process model. There might be some degree of temporal continuity between different replicates of various strains. So, the kernel designed considering coregionalization model will consider the shared information between the replicates and strains. We used programming language python based tool GPy[1], to develop our model. Later we optimized these models and compared them based on likelihood scores and select the best.

We investigated the clusters obtained from our model. We have calculated the enrichment scores [Huang *et al.*, 2009] for every cluster using the tool DAVID[2] (Database for Annotation, Visualization and Integrated Discovery) [Huang *et al.*, 2009] and identified clusters which have very high enrichment scores. We carried out further analysis on some clusters with high enrichment score and demonstrated some interesting characteristics in their dynamic behaviour at the four time stages (pre-symptom, onset, symptom and end-stage) of disease course. Our functional annotation clustering and pathway analysis reveal some interesting information for a group of genes which might have some functionality for the speed of propagation of ALS particularly with reference to this specific type of mouse model.

2. Related Work

Gene expression time series data has been used extensively over the last few decades and implemented for insilico experiments to investigate various fundamental biological processes. Among the many processes examined, some of the notable examples are cell cycle, cell signalling [Barenco *et al.*, 2006], regulatory activity [Sanguinettiet al., 2006], and developmental process [Tomancaket al., 2002]. Gaussian process has applied to gene expression time series widely with several aims and analyses, such as transcription factor target identification [Honkelaa *et al.*, 2010], inference of RNA Polymerase transcription dynamics [Maina *et al.*, 2014], and ranking differentially expressed time series [Kalaitzis and Lawrence, 2011].

Hierarchical models can significantly improve the inference in the Bayesian statistical problems [Gelman *et al.*, 2004]while dealing with multiple relatedgroups of data allowing exchange of information. Inference on the whole structure of data is always preferable thanpartial independent structure. Estimating replicate time shifts were proposed by [Liu *et al.*, 2010], where they used Gaussian process regression with uncertain measurement of mRNA expression. This method requires a large number of variables optimization.Previously, [Nget al., 2006; Medvedovic, 2004] used clustering method to model replicates using a hierarchical structure. Both of the model computes the replicate variance as multivariate Gaussian around some gene-specific mean.

In a clustering application Gaussian process regression could be useful for parsimonious temporal inference. Temporal covariance of genes within a cluster can be designed by adding a hierarchical layer, again covariance between multiple biological replicates can be constructed considering one more hierarchical layer [Hensman *et al.*, 2013]. Whilst GPs also overcome the requirement of evenly spaced time points for time expression data.

3. Methodology

3.1 Gaussian Process Definition

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution [Rasmussen and Williams, 2006].It is a continuous stochastic process and defines probability distributions for functions. It can be also viewed as a random variable indexed by a continuous variable: $f(x)$ chosen from a random function variables $f = \{f_1, f_2, f_3, \ldots, f_N\}$, with corresponding indexed inputs $X = \{x_1, x_2, x_3, \ldots, x_N\}$. In Gaussian processes, variables from these random functions are normally distributed and as a whole can be represent as a multivariate Gaussian distribution:

$$p(f|X) = N(\mu, K), \qquad (1)$$

where μ is the mean and K is covariance of Gaussian distribution $N(\mu, K)$. The Gaussian distribution is over vectors but the Gaussian process is over functions.

We need to define the mean function and covariance function for a Gaussian process prior. If $f(x)$ is a real process, a Gaussian process is completely defined by its mean function and covariance function given in equation 2 and equation 3 respectively. Usually the $m(x)$ and the covariance function $k(x,x^{'})$ are defined as

$m(x) = E[f(x)],$ (2)

$k(x,x^0) = E[(f\ (x)\ m(x))(f\ (x^0)\ m(x^0))],$(3)

where E represents the expected value. We denote the Gaussian process as

$f(x) \sim \mathcal{GP} m(x), k(x,x'))$. (4)

The covariance matrix K is constructed from the covariance function $k(x,x^{'})$ and $K_{ij} = k(x_i,x_j)$.

3.2 Gaussian Process Regression

Gaussian process regression can be done using the marginal and conditional properties of multivariate Gaussian distribution. Let’s consider that we have some observations $\boldsymbol{f}$ of a function at observation point $\boldsymbol{x}$. Now we wish to predict the values of that function at observation pointsx_*, which we are representing by f_*. Then the joint probability of f and f_*can be obtained from

$p\left(\begin{bmatrix} f \\ f_* \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} f \\ f_* \end{bmatrix} \middle| 0, \begin{bmatrix} K_{x,x} & K_{x,x_*} \\ K_{x_*,x} & K_{x_*,x_*} \end{bmatrix}\right)$ (5)

where the covariance matrix $K_{x,x}$has elements derived from the covariance function $k\ (x,x^{'})$, such that the $(i,j\ )^{th}$element of $K_{x,x}$is given by $k\ (x\ [i]\ ,x\ [i])$ The conditional property of a multivariate Gaussian is used to perform regression

the. The conditional property can be represented by

$p(f|f_*) = \mathcal{N}\left(f_* \middle| K_{x_*,x}K_{x,x}^{-1}f K_{x_*,x_*} - K_{x_*,x}K_{x,x}^{-1}K_{x_*,x_*}\right)$.(6)

In ideal case the observations f is noise free but in practice it is always corrupted with some noise. Let’s consider y is the corrupted version of f. If we consider this noise as Gaussian noise then we can write $p\ (y|f) = \mathcal{N}(y|f,\sigma^2 I)$, where$\sigma^2$is the variance of the noise and **I** is the identity matrix with appropriate size and marginalise the observation f. Then the joint probability of y and f_*can be represented by

$p\left(\begin{bmatrix} y \\ f_* \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} y \\ f_* \end{bmatrix} \middle| 0, \begin{bmatrix} K_{x,x} + \sigma^2 I & K_{x,x_*} \\ K_{x_*,x} & K_{x_*,x_*} \end{bmatrix}\right)$.(7)

Regression with Gaussian process can be seen as a Bayesian method. From the knowledge of a *prior* over a function we proceed to a *posterior* and this happens in a closed from of equation 6.

To construct the covariance function still we need to consider the hyperparameters. The most efficient and commonly used selection technique for hyperparameters in Gaussian process is maximum likelihood. If we consider all the hyperparametersα, $\sigma^2$and $l$in to a vector θ, then we can use gradient methods to optimize p (y| θ) with respect to θ. The Log likelihood is given by

$p(y|\theta) = -\frac{D}{2}\log 2\pi - \frac{1}{2} \times \log\left|K_{x,x} + \sigma^2 I\right| - \frac{1}{2} y^T \left[K_{x,x} + \sigma^2 I\right]^{-1} y$(8)

We can have the log maximum likelihood by

$\theta_{max} = argmax\left(p(y|\theta)\right)$.(9)

3.3 Hierarchical Gaussian Process

Our gene expression time series came from four different strain and there are four biological replicates. So for every individual gene wecanincorporate these in a hierarchical fashion. Let y_{nx}denotes gene expression of n^{th}gene in the r^{th}biological replicates and i^{th}biological strain. Measurements were made at four different times and collected into a vector x_{nir}. The data for i^{th}strain's n^{th}gene is denoted by$Y_n = \{y_{nr}\}_{r=1}^{N_n}andX_n = \{x_{nr}\}_{r=1}^{N_n}$.

$$g_n(x) \sim \mathcal{GP}\left(0, k_g(x, x')\right) \qquad (10)$$

$$r_{ni}(x) \sim \mathcal{GP}\left(g_n, k_e(x, x')\right) \qquad (11)$$

$$f_{nir}(x) \sim \mathcal{GP}\left(r_{ni}, k_f(x, x')\right) \qquad (12)$$

For the input dataset $X_n$and hyperparametersθwe can calculate the likelihood by

$$p(Y_n|X_n, \theta) = \mathcal{GP}(\hat{Y}_n|0, \Sigma_n), \qquad (13)$$

and $\hat{Y}_n = \left[Y_{n,1}^T, Y_{n,2}^T, \ldots Y_{n,N_n}^T\right]^T$and θ represents the hyperparamters forthe covariance function $k_g, k_e$and k_f. The structure of the covariance matrix $\Sigma_n$for two genes n hyperparameters for $n$and $n'$are given by otherwise.

$$\Sigma[n, n'] = \begin{cases} \Sigma_n + k_h(x_n, x_{n'}), & if\ n = n' \\ k_h(x_n, x_{n'}) & otherwise. \end{cases} \qquad (14)$$

While designing different kernels k we have used Coregionalization model.

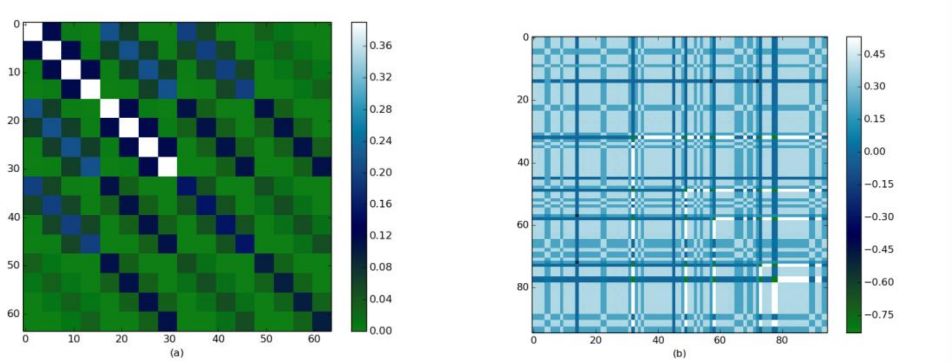

Figure 1:Simple representation of kernel- (a). Coregionalization kernel in the input space (b). kernel after optimization considering only 100 genes.

Here we constructed a hierarchical GP [11] based model to analyse the gene expression time series data collected from four mouse models with different genetic background (129Sv and C57 with transgenic and non-transgenic). We also considered their replications (four in our case) and build a covariance matrix based on their shared information and the time points were pre-symptom, onset, symptom and end stage of the disease course.

3.4 Kernel Design with Coregionalization

Gaussian process models have been used already to capture structure in the data arising from temporal correlation. Our innovation is to realise that there is actually additional correlation structure relating to the genetic background of the organism (in our case, the mice strains) and the status as control/experiment (in our case the presence or absence of the SOD1 mutation). By acknowledging such structure in thecovariance matrix, we can increase the power of our method. Standard approaches force each of these conditions to be fully independent. Our model allows the correlation structure to be learned.

Our formalism for introducing correlations across conditions and strains is the coregionalization principle [Alvarez and Lawrence, 2011] that originates in geostatistics [Wackernagel, et al., 2003].

Coregionalization matrices allow us to share the information between the replicates and strains. In machine learning language, this approach is some- times known as 'multi-task learning' where eachcondition and strain is assumed to be a different task. However, in statistical terms it is simply a multi-variate regression or amultiple output model. An appropriate general model that can capture the dependencies between all the data points and conditions is known as the linear model of coregionalization (LMC) is a model where output is a linear combination of independent random functions. (A detail explanation of the coregionalization model is available at [Alvarez and Lawrence, 2011; Alvarez *et al.*, 2012]). If we can consider our problem with a set of D output functions for $x \in \mathbb{R}^p$input domain, then output function $\{f_d(x)\}_{d=1}^{D}$of LMC can be expressed as

$$f_d(x) = \sum_{q=1}^{Q} a_{d,q} u_q(x) \qquad (15)$$

Here the interpretation is that $\{u_q^i(x)\}_{i=1}^{R_q}, i = 1 \ldots, R_q$are a set of functions that each share the same covariance function (one can think of them as some form of underlying latent processes that determine system behaviour). The parameters $a_{d,q}$represent the relationship between a given latent function, q and an observed condition and or strain. If we consider there can be several different covariance functions associated with separate latent sets then equation 15 is expressed as

$$f_d(x) = \sum_{q=1}^{Q} \sum_{i=1}^{R_q} a_{d,q}^i u_q^i(x) \qquad (16)$$

and the cross covariance function between $f_{d'}(x)$in terms of the function $u_q^i(x)$ is given by

$$cov[f_d(x), f_{d'}(x')] = \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \sum_{i=1}^{R_q} \sum_{i'=1}^{R_q} a_{d,q}^i a_{d'q'}^{i'} cov[u_q^i(x), u_{q'}^{i'}(x')]. \qquad (17)$$

For the so-called homotopic case [Alvarez and Lawerence, 2011,Wackernagel, 2003] the covariance matrix for the joint process f can be rewritten as a sum of Kroneckerproducts,finally, we can write the covariance as

$$K_{f,f} = \sum_{q=1}^{Q} \mathrm{A}_q \mathrm{A}_q^T \otimes K_q = \sum_{q=1}^{Q} B_q \otimes K_q \qquad (18)$$

where ⊗ represents Kronecker product, $\mathrm{A}_q \in \mathbb{R}^{D \times R_q}$and $B_q$is the *coregionalization matrix*. The positive semi-definite covariance functions of the latent processes, $k_q(x, x^{'})$ can be chosen from wide range of covariance functions. Here we used a combination of exponentiated quadratic kernel (also known as squared exponential or RBF kernel) to describe the properties of the function which underlay each cluster. We used a white noise kernel in additive form to deal with the noise of the process. The experimental conditions of acquisition of gene expression measurements cannot be ideally controlled, so the measurements could be corrupted by noise, incorporated either at the biological origin or introduced in the measurement process. Figure 1 shows a simple representation of the coregionalization kernel in the input space and the representation of an optimized kernel where we considered only 100 genes.

3.5 Clustering

Our aim was to discover groups of genes that were exhibiting the same functionalbehavior across times and conditions. Our coregionalization approach allows us to cluster these sub groups through a mixture of Gaussian process models: each component is a function over time, genetic background and condition.

Partitioning genes into clusters can be done by some using our Gaussian process prior over the functions and a Dirichlet process prior for the mixing coefficients. This can be achieved through Gibbs sampling [Dunson *et al.*, 2010], but this can be slow in practice. A potentially improved model was proposed by [Hensman *et al.,* 2013],

where they consider the structure of covariance across the gene and separately across replicates. The use a variation lower bound for model inference. Each gene is placed in an individual cluster and later merged with a greedy selection process to maximize the log marginal likelihood of time series data. Hyperparameters are optimized when no merges are

possible to improve the overall marginal likelihood. Then expectation maximization algorithm is used with new covariance function

4. Dataset and Results

Microarray Data Analysis:

We used the Affymetrix data from [Nardoet al., 2013]. In this experiment spinal cord tissues were obtained from C57 and 129Sv transgenic SOD1^{G93A} mice and age-matched non-transgenic littermates at the presymptomatic, the early symptomatic (onset) stage, symptomatic and end stage. The transcription profiles of laser captured motor neurons isolated from the lumbar ventral spinal cords of the rapid progressor (129Sv SOD1^{G93A}), slow progress (C57 SOD1^{G93A})

mice at four stages of the disease (presymptomatic, onset, symptomatic, end stage) and respective non-transgenic littermates were generated using the murine GeneChip Mouse Genome 430 2.0 Plus (Affy MOE4302). We used Bioconductorpacakge Puma [Pearson *et al.,* 2009] to extract the point estimates of gene expression levels from the GeneChipAffymetrix data.

Select differentially expressed genes:

All the gene expression time series data extracted from Affymetrix data might not be differentially expressed and filtering out the requisite genes is obvious. Considering the temporal nature of data using Gaussian process [Kalaitzis *et al.*, 2011, Alrashid and Alarajii, 2015] can be used to analyse the time series gene expression and filter the quiet or inactive genes from the differentiallyexpressed ones. In addition, identifying genes that have a good signal- to-noise ratio (SNR) is also used to filter down the total number of genes that need further analysis. We can rank the genes by the ratio of the mean replicate-wise variance to the variance of the replicate-wise means. In our analysis, we used a combination of both of the approach. First, we made the initial ranking of the gene expressions (45, 037 genes for our case) using method of [Kalaitzis *et al.*, 2011] and then we use the SNR to choose 10, 000 genes for further analysis. Before the filtering the gene expression levels of each replicates were normalized to zero-mean over all the samples.

Cluster analysis:

In the previous analysis stage, we derived 10,000 genefrom the total of 45,037 probe sets which were more dynamically differentially expressed. We applied our own hierarchical Gaussian process cluster model on these genes and collected the results for further experiments. Figure 2 shows a small part of our result. For any individual graph, along x-axis the four time stages are pre-symptom, onset, symptom and end-stage. We have used four different colours (yellow, red, green and blue) to separate four mouse strains (129Sv SOD1,129Sv Ntg,C57 Ntg and C57 SOD1 respectively). Any individual cluster contains a number of genes which might be biologically associated or co-regulated and we mention the number of the genes belong to that cluster at the corner of the plot. In the plot, a solid line rep- resents posterior mean function and shaded area represents 95% confidence interval. We have found a total of 203 different clusters with a variety of number of genes. Many of the clusters indicated different dynamic behaviour of the gene set. Many of the clusters were attractive for further analysis but that is beyond the scope this study. We included some examples in the Figure 3. We have limited our consideration to the clusters where the strain 129Sv SOD1 (yellow color in our representation) has different characteristics and focussed our consideration for further analysis.

Enrichment score analysis:

A typical biological process is regulated with a group of genes. If we apply a high throughput screen technology then the co-functioning genes are very much more likely to appear together with a higher potential (or enrichment) score. These logical reasons instigate the analysis of a gene list or group of genes moving from individual gene oriented view. The enrichment score is a quantitative measure derived from some well-known statistical methods like Binomial probability, hyper-geometric distribution, Chi-square, Fisher’s exact test. In a previous study, [Huanget al., 2009] reported about 68 Bioinformatics tools to compute the enrichment score and grouped them in three major categories. DAVID [Huanget al., 2009] is a widely used tool developed based on Fisher’s Exact and extensively used for singular enrichment analysis (SEA) and modular enrichment analysis (MEA). We used DAVID on our clusters of genes to calculate the enrichment score for individual clusters. Figure 4 shows the result. Whilst in an analysis a group of genes with an enrichment score of 1.3 can be considered as a threshold value to decide whether this list of genes is enriched or not, here for our 203 clusters we have found at least 15 clusters have an enrichment score of 2.

Pathway analysis:.

Pathway analysis allows us to gain an insight of the underlying biology of the differentially expressed genes. Path- way analysis can reduce the complexity and increase the explanatory power where high-throughput sequencing and gene profiling are used to investigate whether a gene or a list of gene have any roles for a phenotype or a given phenomena. It is also used for the analysis of gene ontology, physical interaction networks, inference of pathways from expression and sequence data, and further comparisons. In a given condition it can identify the pathway by correlating information with a pathway knowledge base. We identified some clusters (which were deemed interesting in the earlier stages) and performed gene ontology enrichment analysis (one example is given at Table 1) and pathway analysis on individual clusters. We identified one of our clusters (cluster197; Figure 3) which was selected at the earlier stage for further analysis and has a relatively high enrichment score (2.16) which is related with ALS. In previous study [Brockingtonet al., 2013] reported about SOD1 related genes and ALS. One of the SOD1 related genes, Derlin 1, can accumulate with other misfolded proteins and cause the neuron death and belongs to our chosen cluster. Figure 5 shows the pathway analysis that we have found for one of our cluster using tool DAVID. We have also found some other genes from the same cluster are responsible for neuron death and related with some other neural disorder like parkinson.

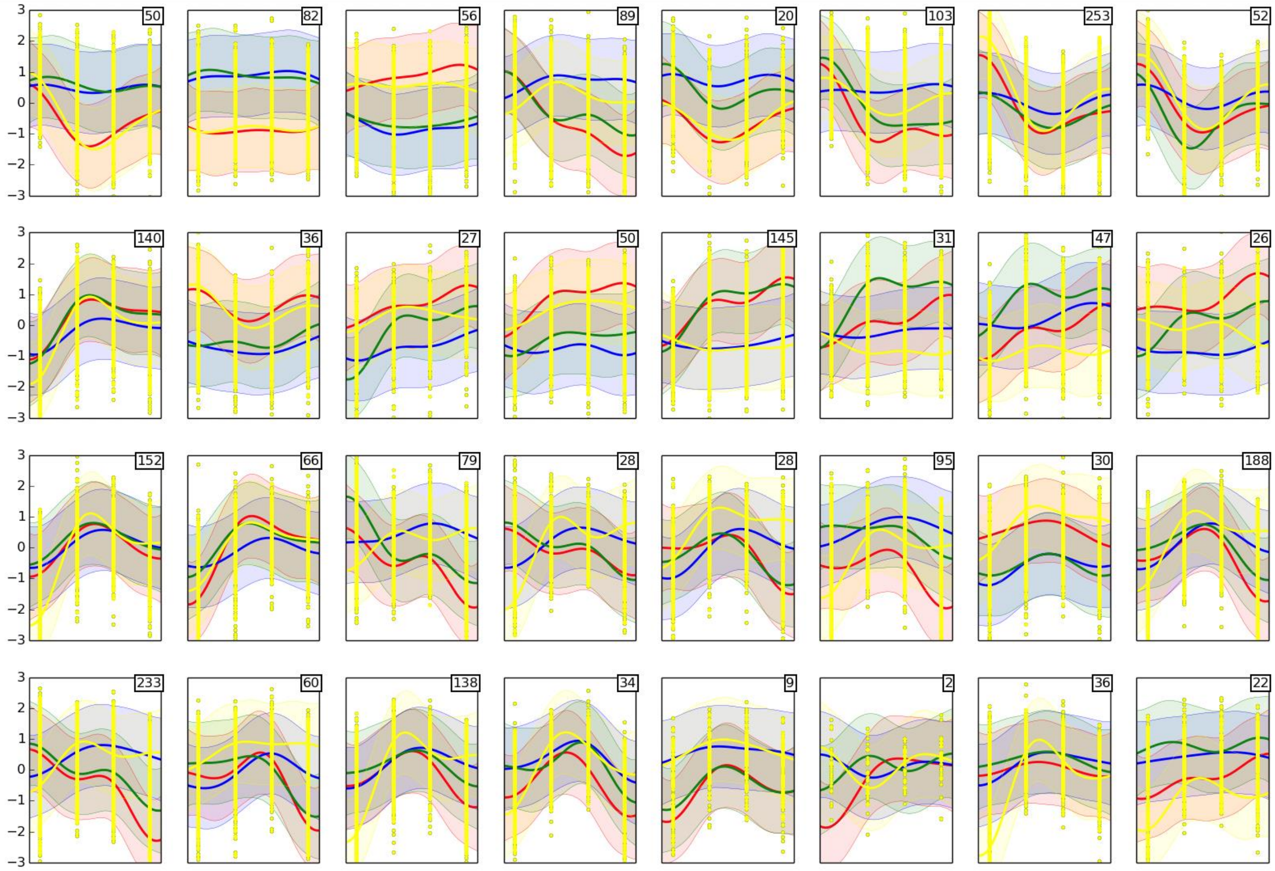

Figure 2: Clustering genes expressions using hierarchy of Gaussian processes. Some representative clusters from the 203 clusters generated (top to bottom, left to right: cluster 01 to 08, 16 to 23, 31 to 38 and 46 to 53). Along x-axis the four time stages are pre-symptom, onset, symptom and end-stage (all the data points together formed like solid yellow vertical lines). Four different colours yellow, red, green and blue are representing four mouse strains 129Sv SOD1,129Sv Ntg, C57 Ntg and C57 SOD1 respectively. Number at the corner indicates number of genes belong to this cluster. Solid line represents a posterior mean function and shaded area represents 95% confidence interval.

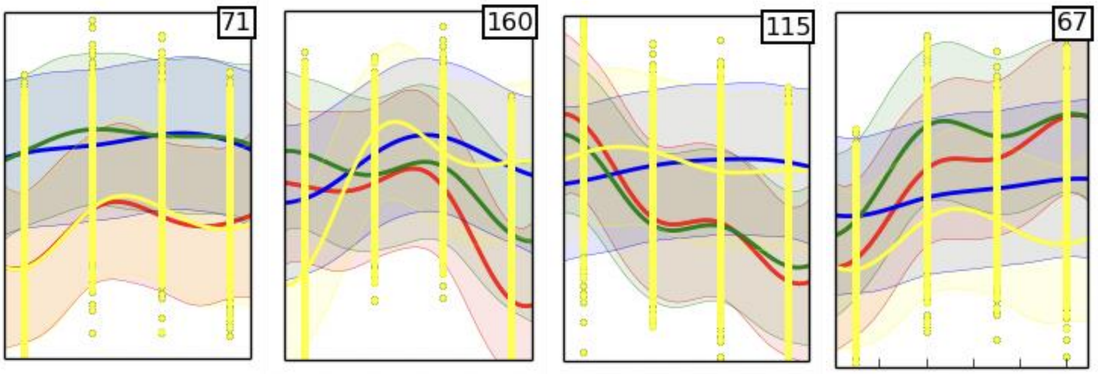

Figure 3: Along x-axis of each individual figure four time stages are pre-symptom, onset, symptom and end-stage. Four different colours yellow, red, green and blue are representing four mouse strains 129Sv SOD1, 129Sv Ntg, C57 Ntg and C57 SOD1 respectively. Examples of clusters where genes from different phenotypic background have different behaviour in time series expression. We used a simple numbering system to represent our clusters and here we are presenting (Figure left to right) cluster119, cluster154, cluster14 and cluster197. Cluster119 showing the clear separation between transgenic group (129Sv SOD1 and C57 SOD1) with non-transgenic mouse model(129Sv Ntg and C57 Ntg), while cluster154 separating mouse C57 from mouse 129sv. Cluster14 and cluster197 showing the different characteristics of 129Sv SOD1 from other three models where it is increasing sharply or becoming very low respectively after the end stage.

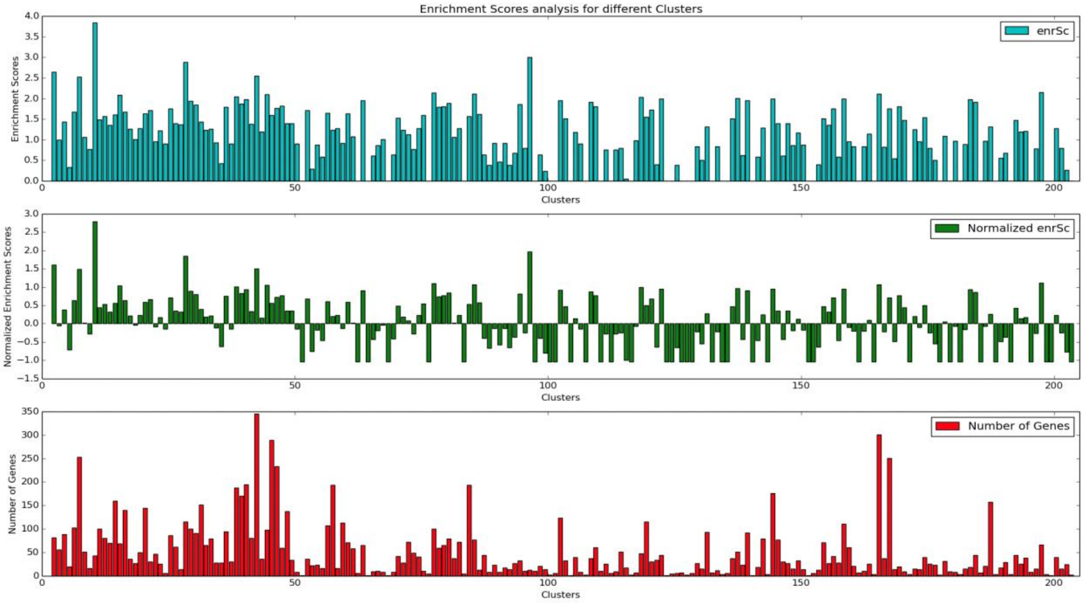

Figure 4: Enrichment scores analysis. Figure in the top shows the enrichment score for different clusters where x axis is the cluster number and y axis shows the enrichment score of that cluster. Figure located at the middle shows the normalized score. While the bottom figure shows the number of genes belongs to any specific cluster.

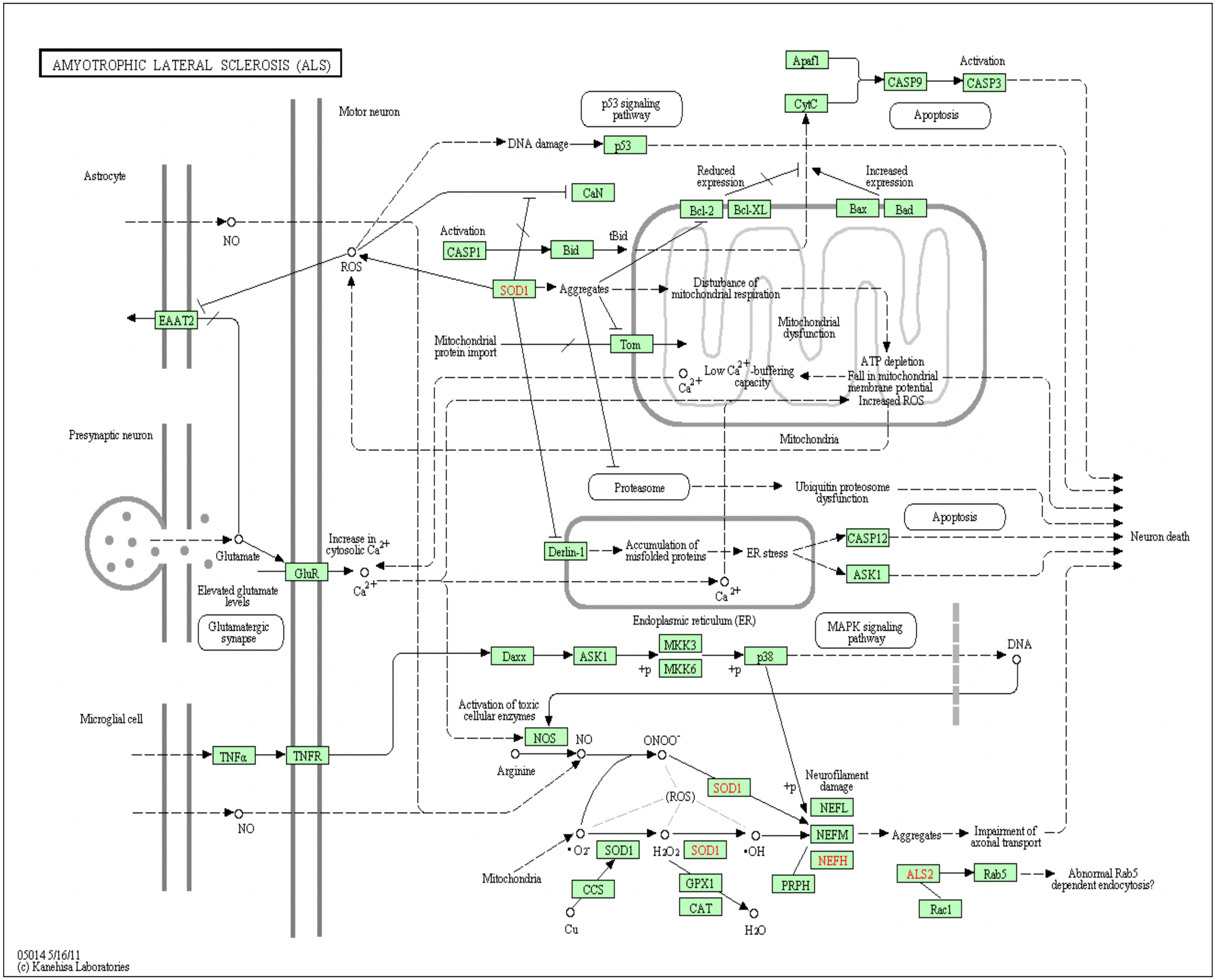

Figure 5:Pathway analysis: KEGG Pathway

Table 1: Functional Annotation Clustering using David.

Annotation Cluster 1	Enrichment Score: 2.16	Count	P Value	Benjamini
GOTERM CC FAT	organelle inner membrane	9	3.0E-5	4.5E-3
GOTERM CC FAT	mitochondrial inner membrane	8	1.6E-4	1.2E-2
GOTERM CC FAT	organelle membrane	12	2.8E-4	1.4E-2
GOTERM CC FAT	mitochondrial membrane	8	6.1E-4	2.3E-2
GOTERM CC FAT	mitochondrial envelope	8	8.7E-4	2.6E-2
GOTERM CC FAT	organelle envelope	9	1.2E-3	3.1E-2
GOTERM CC FAT	envelope	9	1.3E-3	2.7E-2
SP PIR KEYWORDS	mitochondrion inner membrane	5	3.8E-3	4.2E-1
GOTERM CC FAT	mitochondrial part	8	4.6E-3	8.3E-2
SP PIR KEYWORDS	mitochondrion	9	5.8E-3	3.5E-1
GOTERM CC FAT	mitochondrial membrane part	3	1.6E-2	1.9E-1
GOTERM BP FAT	transmembrane transport	6	3.1E-2	8.8E-1
GOTERM MF FAT	hydrogen ion transmembrane transporter activity	3	3.3E-2	9.9E-1
GOTERM MF FAT	monovalent inorganic cation transmembrane transporter activity	3	3.6E-2	9.4E-1
GOTERM MF FAT	inorganic cation transmembrane transporter activity	3	7.1E-2	8.9E-1
SP PIR KEYWORDS	transit peptide	5	7.4E-2	5.8E-1
GOTERM CC FAT	mitochondrion	10	7.6E-2	5.5E-1
UP SEQ FEATURE	transit peptide:Mitochondrion	5	1.0E-1	1.0E0
KEGG PATHWAY	Oxidative phosphorylation	3	1.0E-1	8.6E-1
GOTERM BP FAT	generation of precursor metabolites and energy	3	2.6E-1	9.9E-1

5. Conclusions

We have performed genome-wide analysis to cluster genes systematically and analyse the rationale behind the variation in the speed of propagation for ALS. Our particular innovation was to include the condition and genetic background of the organisms within the underlying functional component of our clusters. This ensured that sub-groups where the underlying expression behaved similarly were more likely to cluster together. The hierarchical Gaussian process we used considers multiple replicates. For validation, we have used a widely acceptable gene ontology and functional annotation tool to validate our clusters and their characteristics obtained from our model. We found a number of clusters are highly enriched. Characteristics curve and enrichment scores analyse helped us to narrow down our search and lead toward finding the lists of genes or clusters which could be involved in the speed of disease propagation. Our pathway analysis found a gene which is known to be involved in the disease process. Here we started with whole genome set and ended with a single gene. This finding leads us to conclude that the model we have developed based on Gaussian process can cluster the genes successfully and they are very much informative. These clusters can be useful for further analysis. Even the model we have developed using hierarchical Gaussian process will be useful to investigate other biological activity where clustering is required.

6. Acknowledgments

We would like to thank Giovanni Nardo, Raffaele Iennaco, NicoloFusi, Marianna Marino, Maria C. Trolese, Laura Ferraiuolo, Pamela J. Shaw and Caterina Bendotti (authors of [Nardo *et al.*, 2013]) who shared their data and allowed us to do further analysis. We would also like to acknowledge Ministry of Science and Information & Communication Technology, Bangladesh for funding Muhammad Arifur Rahman, we would also like to acknowledge Ministry of higher Education and Scientific Research, Iraq for funding SuraZaki Al-Rashid.

7. References

Alvarez M.A. and N.D. Lawrence. Computationally e Kent convolved multiple output Gaussian processes. Journal of Machine Learning Research, (12):1459–1500, June 2011.

Alvarez, M.A; L. Rosasco, and N.D. Lawrence. Kernels for vector-valued functions: A review. Foundations and Trends in Machine Learning, 4(3), June 2012.

Barenco, M.; D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubankcorresponding. Ranked prediction of p53 targets using hidden variable dynamic modeling. Genome Biology, 7(3):R25, March 2006.

Beghi, E.; A. Chio, P. Couratier, J. Esteban, O. Hardiman, G. Logroscino, A. Millul, D. Mitchell, P.-M. Preux, E. Pupillo, Z. Stevic, R. Swingler, B. J. Traynor, L. H. V. den Berg, J. H. Veldink, and S. Zoccolella. The epidemiology and treatment of als: focus on the heterogeneity of the disease and critical appraisal of therapeutic trials. Amyotrophic Lateral Sclerosis, 12(1):1–10, January 2011.

Brockington, A.; K. Ning, P. R. Heath, E. Wood, J. Kirby, N. Fusi, N. Lawrence, S. B. Wharton, P.G. Ince, and P.J. Shaw. Unravelling the enigma of selective vulnerability in neurodegeneration: motor neurons resistant to degeneration in als show distinct gene expression characteristics and decreased susceptibility to excitotoxicity. ActaNeuropathol, 125(1):95–109, January 2013.

Camu, W.; J. Khoris, B. Moulard, F. Salachas, V. Briolotti, G. Rouleau, and V. Meininger. Genetics of familial als and consequences for diagnosis. frenchals research group. J NeurolSci, 165:s21–s26, January 1999.

Dunson. D.D. Nonparametric bayes applications to biostatistics. Bayesian Nonparametrics, Cambridge University press, 2010.

Ferraiuolo, L.; J. Kirby, A. J. Grierson, M. Sendtner, and P. J. Shaw. Molecular pathways of motor neuron injury in amyotrophic lateral sclerosis. Nat Rev Neurol, 7(11):616–630, November 2011.

Gelman, A.; J.B. Carlin, H.S. Stern, and D.B. Rubin. Bayesian Data Analysis. Chapman and Hall / CRC, 2004.

Haverkamp, L.J.; V. Appel, and S. H. Appel. Natural history of amyotrophic lateral sclerosis in a database population validation of a scoring system and a model for survival prediction. Brain, 118:707–719, 1995.

Hensman, J.; N.D. Lawrence, and M. Rattray. Hierarchical bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. BMC Bioinformatics, 14(252), August 2013.

Honkelaa, A.; C. Girardotb, H. Gustafsonb, Y.-H. Liub, E. E. M. Furlongb, N. D. Lawrencec, and M. Rattrayc. Model-based method for transcription factor target identification with limited data. Proceedings of the National Academy of Sciences, 107(17):7793–7798, 2010.

Huang, D.W. ; B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research, 37(1):1–13, 2009. D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature Protoc., 4(1):44–57, 2009.

Kalaitzis A.A. and N.D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through gaussian process regression. BMC Bioinformatics, 12(180), May 2011.

Liu, Q. ; K.K. Lin, B. Andersen, P. Smyth, and A. Ihler. Estimating replicate time shifts using gaussian process regression. Bioinformatics, 26(6):770–776, March 2010.

M.Medvedovic, K. Yeung, and R. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. Bioinformatics, 20(8), May 2004.

Maina, C.W. ; A. Honkela, F. Matarese, K. Grote, H. G. Stunnenberg, G. Reid, N. D. Lawrence, and M. Rattray. Inference of rna polymerase ii transcription dynamics from chromatin immunoprecipitation time course data. PLoSComputBiol, 10(5), May 2014.

Marino, M. ; S. Papa, V. Crippa, G. Nardo, M. Peviani, C. Cheroni, , M. C. Trolese, E. Lauranzano, V. Bonetto, A. Poletti, S. DeBiasi, L. Ferraiuolo, P. J. Shaw, and C. Bendotti. Differences in protein quality control correlate with phenotype variability in 2 mouse models of familial amyotrophic lateral sclerosis. Neurobiology of Aging, 36:492–504, January 2015.

Nardo, G.; R. Iennaco, N. Fusi, P. R. Heath, M. Marino, M. C. Trolese, L. Ferraiuolo, N. Lawrence, P. J. Shaw, and C. Bendotti. Transcriptomic indices of fast and slow disease progression in two mouse models of amyotrophic lateral sclerosis. Brain A journal of Neurology, 136(11), September 2013.

Ng, S.K. ; G. J. McLachlan, K. Wang, L. B.-T. Jones, and S. W. Ng. A mixture model with random-effects components for clustering correlated gene-expression profiles. Bioinformatics, 22(14):1745–1752, April 2006.

Pearson, R.D. ; X. Liu, G. Sanguinetti, M. Milo, N. D. Lawrence, and M. Rattray. Puma: a Bioconductor package for propagating uncertainty in microarray analysis. BMC Bioinformatics, 10(211), 2009.

Peviani, M.; I. Caron, C. Pizzasegola, F. Gensano, M. Tortarolo, and C. Bendotti. Unraveling the complexity of amyotrophic lateral sclerosis: recent advances from the transgenic mutant sod1 mice. CNS NeurolDisord Drug Targets, 9(4):491–503, August 2010.

Pizzasegola, C.; I. Caron, C. Daleno, A. Ronchi, C. Minoia, M. T. Carrı, and C. Bendotti. Treatment with lithium carbonate does not improve disease progression in two different strains of sod1 mutant mice. Amyotrophic Lateral Sclerosis, 10(4):221–228, 2009.

Rasmussen C. E. and C. K. Williams. Gaussian Processes for Machine Learning. 2006.

Saccon, R.A. ; R. K. A. Bunton-Stasyshyn, E. M. Fisher, and P. Fratta. Is sod1 loss of function involved in amyotrophic lateral sclerosis? Brain A journal of Neurology, 136(pt 8):2342–58, August 2013.

Sanguinetti, G. ; M. Rattray, and N. D. Lawrence1. A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. Bioinformatics, Oxford University Press, 22(14):1753–1759, 2006.

Sura Z. Alrashid and Nabeel H. Al-Aaraji. Bayesian Models with Coregionalization to Model Gene Expression Time Series for Mouse Model for Speed Progression of ALS Disease. European Journal of Scientific Research, Volume 132 Issue 1, 2015.

Tomancak, P.; A. Beaton, R. Weiszmann, E. K. abdShengQiang Shu, S. E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. E. Celniker, and G. M. Rubin. Systematic determination of patterns of gene expression during drosophila embryogenesis. Genome Biology, 3(12):research 0088.1 – 0088.14, December 2002.

Turner B.J. and K. Talbot. Transgenics, toxicity and therapeutics in rodent models of mutant sod1-mediated familial als. ProgNeurobiol, 85(1):94–134, May 2008.

Wackernagel. H. Multivariate Geostatistics An Introduction with Applications. Springer Science and Business Media, 2003.