The University of Sheffield

Transfer Report

A Gaussian Process model for Transcription Factor Activity of C. Elegans

Author:
Muhammad A. Rahman

Supervisor:
Neil D. Lawrence

A report submitted in fulfilment of the requirements for the Transfer to Doctor of Philosophy

in the

Machine Learning Research Group
Department of Computer Science

September 2014

THE UNIVERSITY OF SHEFFIELD

Abstract

Faculty of Engineering
Department of Computer Science

Transfer to Doctor of Philosophy

A Gaussian Process model for Transcription Factor Activity of C. Elegans

by Muhammad A. Rahman

In molecular biology and genetics, a transcription factor is proteins that binds to specific DNA sequences and control the flow of genetic information from DNA to mRNA. To develop models of cellular processes quantitative estimation of the regulatory relationship between transcription factors and genes is a basic requirement. But quantitative estimation is complex due to some reasons. Many of the transcription factors' activity and their own transcription level are post transcriptionally modified; very often the levels of the transcription factors' expressions are low and also contain noise. So, from the expression levels of their target genes it could be useful to infer the activity of the transcription factors. Here we are trying to develop a Gaussian process based regression to infer the exact TFAs from a combination of mRNA expression level and DNA protein binding measurement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Caenorhabditis Elegans

Caenorhabditis Elegans is a nonparasitic, soil dwelling, small nematode worm. C. Elegans and other Caenorhabditis species are found through all over the world. It can easily colonize mostly in the rotting materials with other microorganism. In 1965, Sydney Brenner introduced Caenorhabditis Elegans as a model organism to study the behaviour and development of animal Brenner [1974].

C. Elegans is a relatively new addition as a model organism but its biological characteristics and property already been studied to an extraordinary level. The anatomical characteristics and detail development of this nematode was facilitated by its simple body plan. Its an eukaryote and it shares cellular and molecular structures and control pathways with higher organism. C. Elegans is multicellular, an adult wild type consist of 959 somatic cells and among these 302 are neurons Palikaras and Tavernarakis [2013], Sulston and Horvitz [1977]. Its developmental process like embryogenesis, morphogenesis goes through a a complex process to growth to an adult. Yet monitoring of the cellular process and recording of cell division pattern is comparatively easier as its body is transparent. C Elegan's complete cell lineage at the electron microscopy level has been completed. Its already been established and this cell lineage is remarkably invariant between animal to animal. Brenner [1974], Byerly et al. [1976], Sulston et al. [1980], Wood [1988].

C. Elegans is easy to maintain in the petri dishes at the laboratory. At 25 °C C. Elegans completeits life cycle in just 2.5 days from fertilized embryos to egg-laying adult through 4 larval stages. Its typical life span is 2-3 weeks.

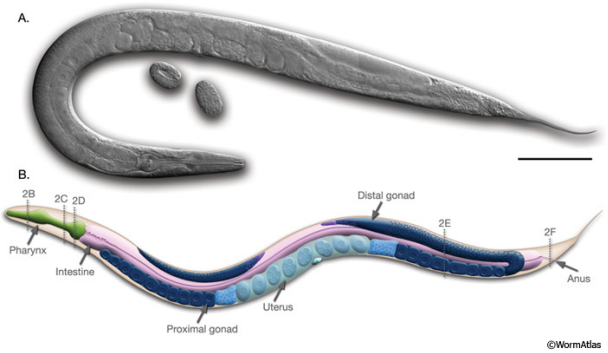

FIGURE 1.1: Anatomy of an adult hermaphrodite. A. DIC image of an adult hermaphrodite, left lateral side. Scale bar 0.1 mm. B. Schematic drawing of anatomical structures, left lateral side (Courtesy of WormAtlas http://www.wormatlas.org/hermaphrodite/introduction/IMAGES/introfig1leg.htm).

To elucidate pathways and processes relevant to human biology and disease C. Elegans is using as a vital model. There are between 20,250 and 21,700 predicted protein-coding genes in C. elegans Gerstein et al. [2010]. Using four different orthology-prediction methods Shaye and Greenwald [2011] assayed four methods to compile a list of C. elegans orthologs of human genes. a list of 7,663 unique protein-coding genes were resulted in that list and this represents 38% of the 20,250 protein-coding genes predicted in C. elegans. When human genes introduced into C.Elegans human genes replaced their homologs. On the contrary, many C. Elegans genes can function with great deal of similarity to human like mammalian genes. So, the biological insight acquire from C. Elegans may be directly applicable to more complex organism like human.

1.2 Transcription

DNA (Deoxyribonucleic acid) transcription is a process that involves the transcribing of genetic information from DNA to a complementary RNA (Ribonucleic acid). Protein is produced from the copy of DNA by the transcription process. This production of proteins and enzymes are controlled by the coding of cellular activity. Even the conversion of DNA to proteins is not straight forward. An RNA polymerase read the sequence of DNA, which produce an complementary RNA. DNA consists of four nucleotide bases named adenine (A), guanine (G), cytosine (C) and thymine (T) that are paired together (A-T and C-G) to give DNA its double helical shape. The major steps to the process of DNA transcription are-

RNA polymerase binds to DNA: In order to initiate the DNA transcription RNA polymerase and sigma factor form a holoenzyme. Transcription process starts at the

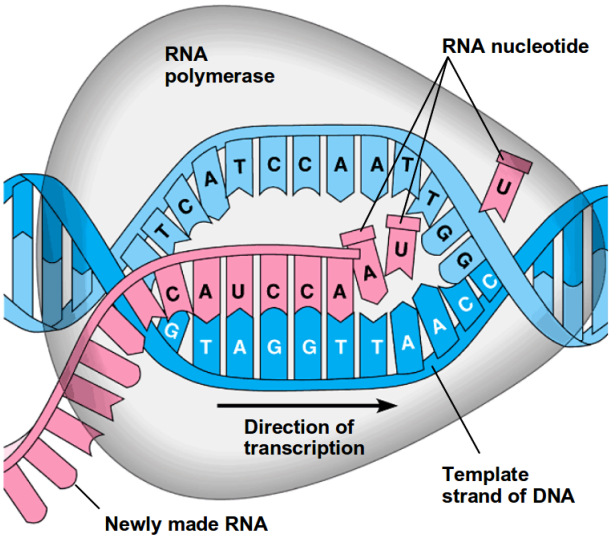

FIGURE 1.2: Transcription (Courtesy of Perason Education; Copyright 2003: Pearson Education, Inc. Publishing as Benjamin Cummings https://meyerbio1b.wikispaces.com/Transcription+and+Translation)

promoter region of a double-stranded DNA. Sigma factor can recognize the DNA and its promoter region.

Elongation: A sequence specific DNA binding factors called transcription factors then unwind the DNA strand. Elongation of the transcript then continues by the RNA polymerase and a sequence of chain is opened up. A messenger RNA (mRNA) is formed when RNA polymerase transcribe into a single stranded RNA polymer from a single strand of DNA.

Termination: RNA polymerase moves along the DNA unwinding its double helical form until it reaches the terminator sequence. At that point, RNA polymerase detaches from the DNA and releases the mRNA polymer. In this way DNA double helix is opened, transcribed and reclosed with minimum stress on the DNA molecule. At any certain time many RNA polymerase can transcribe a single DNA sequence, which can manufacture a large quantity of protein at once.

1.3 Transcription Factor

A transcription factor is a protein that binds to DNA sequences and controls the flow of genetic information coding from DNA to mRNA Karin [1990], Latchman [1997]. Transcription factors can both promote or block the transcription process and act as an activator or repressor respectively Lee and Young [2000], Nikolov and Burley [1997], Roeder [1996]. A transcription factors may contain one or more DNA-binding domains.

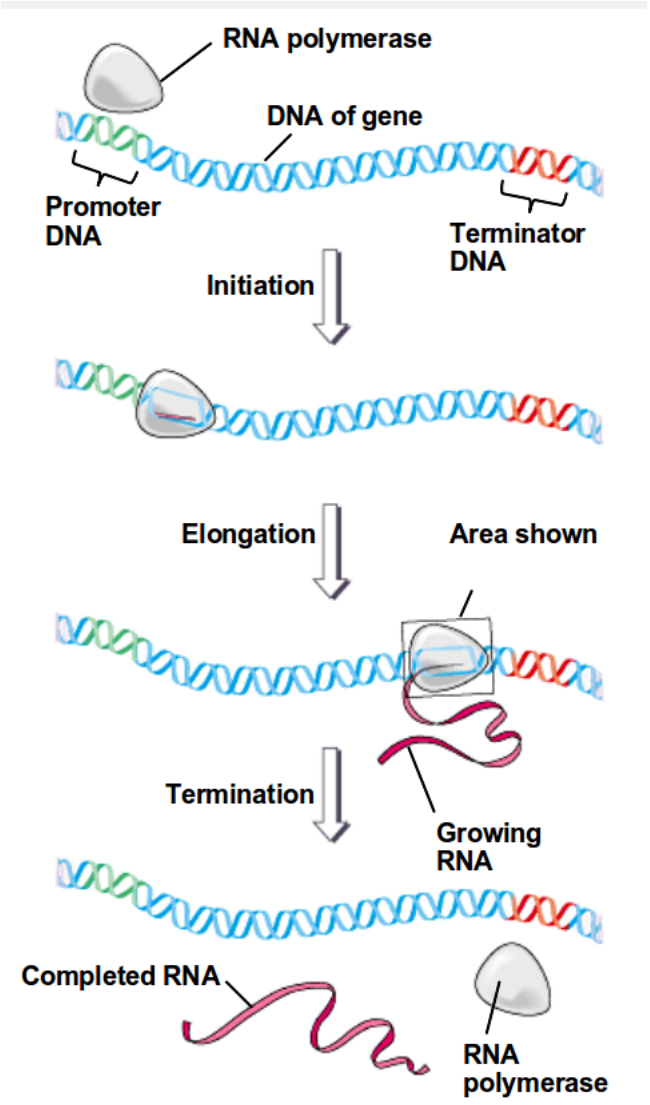

FIGURE 1.3: Transcription Process: DNA Transcribed in mRNA and mRNA Translated in to Protein (Courtesy of Perason Education; Copyright 2003: Pearson Education, Inc. Publishing as Benjamin Cummings https://meyerbio1b.wikispaces.com/Transcription+and+Translation)

These binding domains attach to specific sequences of DNA adjacent to the genes that they regulate. Though some other protein such as coactivators, deacetylases, chromatin remodelers, kinases, histone acetylases, and methylases also play crucial roles in gene regulation but due to lack of DNA-binding domains they are not classified as transcription factors Brivanlou and Darnell [2002], Mitchell and Tjian [1989], Ptashne and Gann [1997]. Figure 1.4 describes the mapping between the environmental signal, transcription factors inside the cell, and the gene that they regulate. The environmental signal activates specific transcription factor proteins. After the activation the transcription factors bind DNA to change the transcription rate (the rate at which mRNA is produced) of specific target genes. The mRNA is then translated into protein Alon [2006].

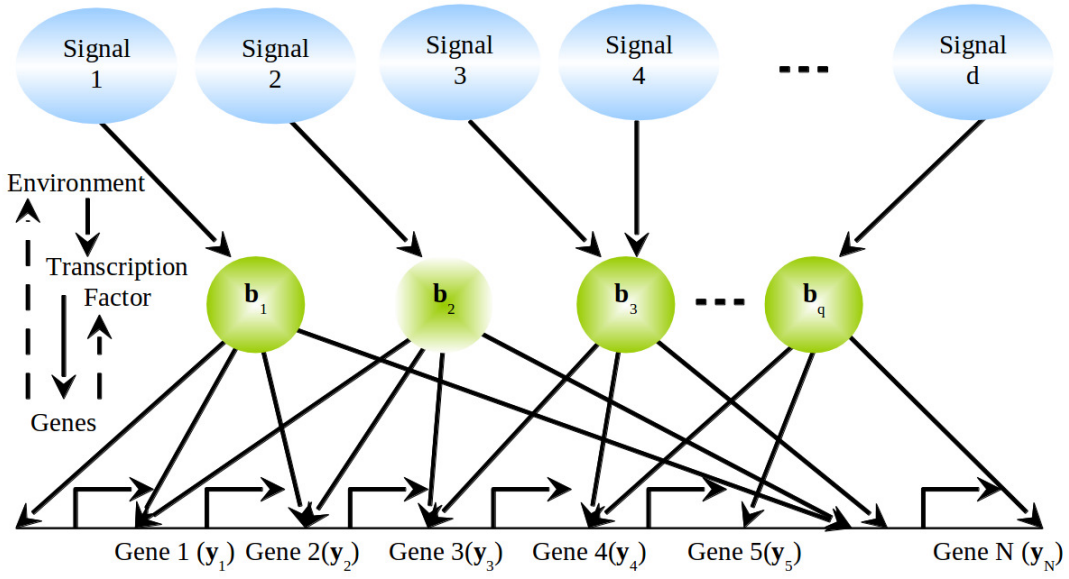

FIGURE 1.4: The mapping between environmental signal, transcription factor inside the cell the genes that they regulate.

1.4 Key Research Questions

Key Research question-

- Can we develop a robust data driven dynamic system for gene specific transcription factor activities using Gaussian Process over a mechanistic model?
- Can we step forward to find out the transcription factor activities from a unicellular microorganism to multicellular eukaryote?
- Is it possible to build some replicate specific robust clusters of the gene expression data for C Elegans?

Chapter 2

Dynamic Modelling of TFA-Literature Review

2.1 Literature Review

In modern molecular biology the biological systems like cells are treated as a complex systems. The usual conception of the complex system is a very large number of simple but identical elements interact to generate the complex behaviour. But the actual behaviour of biological systems are different from this conception. A vast number of functionally different and multifunctional group of elements act with each other selectively, perhaps nonlinearly to generate coherent instead of complex behaviour. Mostly functions of biological systems depends on a combination of the network and specific elements involved.

Development of molecular biology has discovered a large number of biological facts like sequencing genome, protein properties etc.. But to explain the biological systems only these are not sufficient. Study of cell tissues, organs, organisms etc. are also the systems of components to consider and their specific interaction which is defined by the evolution could be more supportive to reach the prime goal of biology. Though advancement in more accurate quantitative experimental approach will continue but the detail functional insights of biological systems may not give the exact results from purely intuitive basis due to the intrinsic complexity of biological systems. A proper combination of experimental and computational approaches is more likely to solve this problem.

In modern molecular biology the organizational and functional activity of gene regulatory network is a key experimental and computational challenge.

Living cells contains thousands of genes. These genes codes for one or more protein. Expression of these genes are regulated by many of these protein through a very complex regulatory pathway. Usually this regulation occurred to accommodate the change of the environment, as well as through the cell cycle of the development process. Ong et al. [2002] modelled the regulatory pathway in E.coli from the time series gene expression microarray data by modelling causality, feedback loops or hidden variables using a Dynamic Bayesian network and tried to gain the insight of regulatory pathway. By analysing gene expression data Friedman et al. [2000] were the first to determine the properties of transcriptional program for Baker's yeast using Bayesian network.

Many of the resent studies already established the fact that the gene function of regulatory network depends on qualitative as well as quantitative aspects of the organization of the network like high throughput data, including genomic sequence, expression profiles and transcription factor.

Among them one of the major challenges is to quantitative measurement and analysis of the mechanisms regulating mRNA transcription. Though using high throughput techniques it is comparatively easier to measure the output of transcription, but it is experimentally very complicated to measure the protein concentration levels of transcription factors and chemical affinity to the genes. Very often transcription factors are post-transcriptionally modified. So, the actual protein concentration levels and binding affinities could be an unreliable proxy of the mRNA expression levels of transcription factors Sanguinetti et al. [2006].

Due to the advancement of the experimental technique lot of interest in recent years has been growing to infer information about regulatory activity from target genes. Now biologists can acquire the information about the structure of the transcriptional regulatory network. Lee et al. [2002] determined the transcriptional regulatory network of yeast using chromatin immunoprecipitation(ChIP). They tried to figure out how yeast transcriptional regulators bind to promoter sequences across the genome. By calculating a confidence value (P value) and setting up specific threshold they considered the protein-DNA interactions and artificially imposes a binding or not binding binary decision for each of the protein- DNA pair.

Xie et al. [2005] used motif conservation information for higher organisms like human, dog, rat and mouse. For promoter analysis they considered a number of network motif (also known as transcription factor binding sites) and some new motifs. These type of data termed as connectivity data Liao et al. [2003] provide information about whether a certain transcription factor can bind the promoter region of a gene or not.

In recent years most methods aim to infer a matrix of transcription factor activities (TFAs). These TFAs are sum up in a single number at a certain experimental point to find the concentration of the transcription factor and its binding affinity to its target genes. Many of the researcher used different ways or algorithm to find out these TFAs. For example, Liao et al. [2003] developed a data decomposition technique with dimension reduction and introduced 'network component analysis'. This method takes account of the connectivity information by imposing algebraic constraints on the factors. They argued that classical statistical methods such as principal component analysis and independent component analysis ignore the underlying network structure while computing low dimensional or hidden representation of a high-dimensional data sets like DNA microarray.

Alter and Golub [2004] used dimension reduction technique (SVD) to figure out TFAs and also the correlation between DNA replication initiation and RNA transcription during the yeast cell cycle. Using multivariate regression and backward variable selection to identify active transcription factors Gao et al. [2004] targeted the same; Boulesteix and Strimmer [2005] used partial least squares (PLS) regression to infer the true TFAs from a combination of tRNA expression and DNA protein binding measurement. A major drawback of the above mentioned methods is that transcription factor activities do not contain any information about the strength of the regulatory interactions between the transcription factor and its different target genes. But it is expected that depending on the experimental conditions transcription factor activities can vary from gene to gene. Even it is also expected that different transcription factors may bind the same gene. In most of the cases, realistic information about the intervals may not be true as they were not based on fully probabilistic model. Moreover, false positives are always a problem for connectivity data, typically a large portion of Chip data suffers form it Boulesteix and Strimmer [2005]. Furthermore, due to the various cellular process or changes in environmental conditions the structure of the regulatory network of the cell can change considerably. Using regression-based methods it is difficult to track these changes. Nachman et al. [2004] build a probabilistic model, using the basic framework of dynamic Bayesian networks using discrete random variables for protein concentrations and binding affinities. Though the model was more realistic but the computational complexity for genome-wide analysis can be expensive.

We will propose a dynamic model that extends the linear regression model of Liao et al. [2003] and probabilistic model of Sanguinetti et al. [2006] to model the distribution of each transcription factor acting on each gene. We will try to model the temporal changes in the gene-specific TFAs from timeseries gene expression data using Gaussian Process. The covariance structure of the transcription factors will be shared among all

genes. This approach will lead to a manageable parameter space and figure out useful information about the correlation of TFAs.

Initially to build our model we will use two datasets: the classical yeast cell cycle dataset of Spellman et al. [1998] and the yeast metabolic cycle dataset of Tu et al. [2005]. Both of the data set was used to study the above mentioned models. So, these data will be a source of useful comparisons. In both cases the connectivity data will be Chip data (Lee et al. [2002]; Harbison et al. [2004]). Finally we will use the data set of Caenorhabditis elegans (C. elegans) to obtain a deeper insight. C. elegant was used to build the probabilistic functional gene network Network.

2.2 TO DO Regulator density and network motif

Lee et al. [2002] page 800

Chapter 3

Probabilistic Dynamic Modelling

3.1 Probabilistic Modelling

We have developed our *R* based tools *chipDyno* based on the probabilistic approach of Sanguinetti et al. [2006]. First we will give a brief introduction of that approach then we will present results.

The logged gene expression measurements are collected in a design matrix , $\mathbf{Y} \in \mathbb{R}^{\mathbf{N}\times\mathbf{d}}$ where N is the number of genes and d the number of experiments. The connectivity measurements are collected in a binary matrix $\mathbf{X} \in \mathbb{R}^{\mathbf{N}\times\mathbf{q}}$, where q is the number of transcription factors; element (i, j) of $\mathbf{X}$ is one if transcription factor j can bind gene i, zero otherwise.

Based on Sanguinetti et al. [2006] TFAs can be obtained by regressing the gene expressions using the connectivity information, giving the following linear model-

$$\mathbf{y_n} = \mathbf{B_n x_n} + \boldsymbol{\epsilon_n} \tag{3.1}$$

Here $n = 1, ..., N$ indexes the gene, $\mathbf{y_n} = \mathbf{Y}(n, :)^T$, $\mathbf{x_n} = \mathbf{X}(n, :)^T$ and $\boldsymbol{\epsilon_n}$ is an error term. The matrix $\mathbf{B_n}$ has d rows and q columns, and models the gene specific TFAs.

Two plausible assumptions for TFA-

- Firstly, Gene specific TFA $\mathbf{b}_{nt}$ at time t depends solely on the gene specific TFA at time $(t-1)$.
- Secondly, it was assumed that the prior distribution to be a stationary in time.

Two limiting case- The first limiting case when all the $\mathbf{b}_{nt}$ assumed to be identical, so that-

$$\mathbf{b}_{n1} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), and\ \mathbf{b}_{n(t+1)} \sim \mathcal{N}(\mathbf{b}_{nt}, \mathbf{0}) \tag{3.2}$$

Second limiting case was when all the $\mathbf{b_{nt}}$ were assumed to be independent and identically distributed-

$$\mathbf{b}_{nt} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.3}$$

Sanguinetti et al. [2006] expected a realistic model of time series data to be somewhere in between this two extremes-

$$\mathbf{b}_{n(t+1)} \sim \mathcal{N}(\gamma \mathbf{b}_{nt} + (1-\gamma)\boldsymbol{\mu}, (1-\gamma^2)\boldsymbol{\Sigma}) \tag{3.4}$$

for $t = 1, ..., (d-1)$ and $\mathbf{b}_{n1} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ Where γ is a parameter measuring the degree of temporal continuity of the TFAs.

Likelihood function-

$$p\left(\mathbf{Y}|\mathbf{B}, \mathbf{X}\right) = \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathbf{B}_n, \mathbf{x}_n\right) \tag{3.5}$$

TFAs can be estimated a posteriori using Bayes's Theorem-

$$p\left(\mathbf{b}_n|\mathbf{Y}\right) = \frac{p\left(\mathbf{Y}|\mathbf{b}_n\right) p\left(\mathbf{b}_n\right)}{p\left(\mathbf{Y}\right)} \tag{3.6}$$

3.2 Datasets

Sanguinetti et al. [2006] has done their experiments on yeast's cell cycle data of Spellman et al. [1998] which is a unicellular microorganism. One of our research key question was can we step forward to find out the transcription factor activities from a unicellular microorganism to multicellular eukaryote. C Elegans is a established multicellular eukaryotic model organism. To find out the TFA of C Elegans basically we had to work with three type of datasets. 1. Time series data. 2. Transcription Factors. 3. Connectivity information between genes and transcription factors.

3.2.1 Time series data

The Affymetrix CEL data files came from the Prof. Andrew Cossins's lab, Institute of Integrative Biology, University of Liverpool (Proper reference required!). 15 Affymetrix single color GeneChip data on point estimate of expression level came without estimates of uncertainty level. To extract this data we use package *puma* The experimental data had 5 different time points. Apart from the temperature rest of the environmental

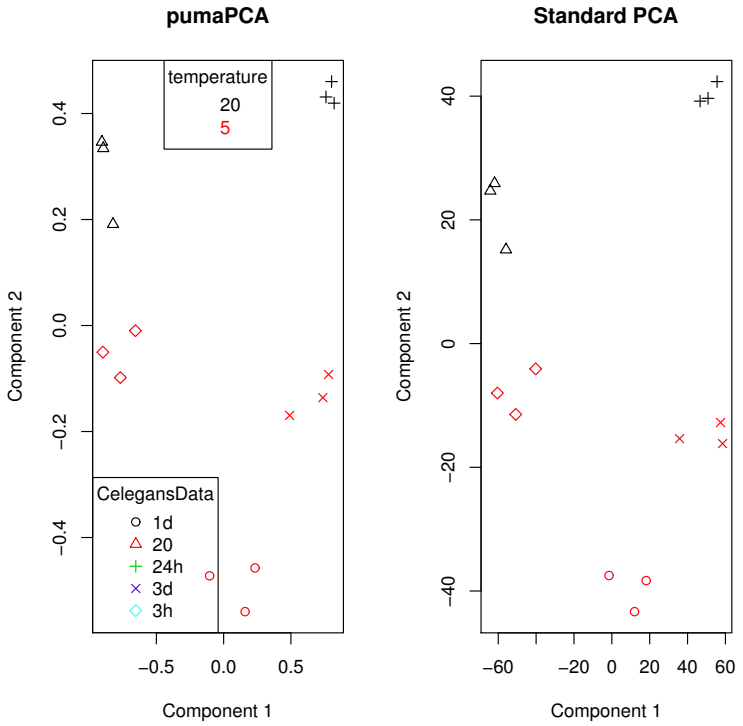

FIGURE 3.1: Principal component analysis of time series data

condition was same for the consistent result. The experimental data was collected within one day of its adulthood at the temperature 20 °C. To measure the gene response to chill exposure then the temperature was reduced to 5 °C and experiments were done after one hour, then after 1 day (24 hour), then again after 3 days (72 hours) and for final experiments the temperature was set back to 20 °C and data were collected within one day of rise of temperature. All the experiments were repeated two more times. i.e. we have 3 independent replicates of similar experiments. Figure 3.1 shows the PCA analysis of the time series data.

3.2.2 Transcription Factors

From different data source we found different number/list of transcription factor. Inmaculada et al. [2007] build a database named C. Elegans differential gene expression database (EDGEdb) which contains the sequence information about 934 predicted transcription factors and their DNA binding domains. Initially we took these 934 transcription factors for our baseline experimental setup but we also kept the opening to deal with any number of transcription factor depending on the requirement/ update of the sequence information of transcription factors.

3.2.3 Connectivity information

WormNet [2014] is a gene network of protein-encoding genes for C. Elegans based on based on probabilistic function and modified Bayesian integration. They have considered 15,139 genes and 999,367 linkages between genes associated with a log-likelihood score (LLS). These measured scores represents a true functional linkage between a pair of genes Lee et al. [2007]. The linkage between two genes were measured based on the following evidence codes-

- CE-CC Co-citation of worm gene
- CE-CX Co-expression among worm genes
- CE-GN Gene neighbourhoods of bacterial and archaeal orthologs of worm genes
- CE-GT Worm genetic interactions
- CE-LC Literature curated worm protein physical interactions
- CE-PG Co-inheritance of bacterial and archaeal orthologs of worm genes
- CE-YH High-throughput yeast 2-hybrid assays among worm genes
- DM-PI Fly protein physical interactions
- HS-CC Co-citation of human genes
- HS-CX Co-expression among human genes
- HS-DC Co-occurrence of domains among human proteins
- HS-LC Literature curated human protein physical interactions
- HS-MS human protein complexes from affinity purification/mass spectrometry
- HS-YH High-throughput yeast 2-hybrid assays among human genes
- SC-CC Co-citation of yeast genes
- SC-CX Co-expression among yeast genes
- SC-DC Co-occurrence of domains among yeast proteins
- SC-GT Yeast genetic interactions
- SC-LC Literature curated yeast protein physical interactions
- SC-MS Yeast protein complexes from affinity purification/mass spectrometry

- SC-TS Yeast protein interactions inferred from tertiary structures of complexes

We have constructed the connectivity matrix between genes and associated transcription factors from the gene to gene linkage and log-likelihood scores. We choose Co-expression among worm genes (CE-CX), High-throughput yeast 2-hybrid assays among worm genes (CE-YH), Literature curated human protein physical interactions (HS-LC) and High-throughput yeast 2-hybrid assays among human genes (HS-YH) to start our experiments. But if needed we can consider any of the evidance to reconstruct the connectivity matrix. From the gene list we have picked the protein-coding genes (i.e. transcription factors) and later binarized it. If there is an associated LLS value between a gene and a transcription factor we set the value '1' and '0' otherwise.

3.3 Result Analysis

Based on Sanguinetti et al. [2006] we have developed a *R* based tool *chipDyno* for the identification of quantitative prediction of regulatory activities of the gene specific TFA through posterior estimation. For C Elegans gene specific TFA is quite a new experiments. So we didn't find any other result to consider a baseline and compare our result.

There are number of gene for C Elegans is 15,139 WormNet [2014] and according Inmaculada et al. [2007] number of transcription factors are 934. All the network motif, i.e. autoregulation, multi-component loop, feedforward loop, single input, multi-input motif, regulator chain were visible for transcription factor activity. So it was a mammoth task to choose all the transcription factors and show their activity. Rather we choose some random transcription factor and tried to find out its activity on different genes.

As a random sample we choose transcription factor ZK370.2 and tried to find its activity on different genes. Figure 3.2 shows that ZK370.2 can act on C37C3.2, Y105E8B.3, Y45F10B.3, C34F11.3, F26E4.6 and T24G10.2. In the dataset we had three replication of same experimental setup and it outcome. We also tried to run our experiments three times and collect the result individually. Later we plot all the outcome together. From the outcome of our experimental result we can say that for some cases (i.e. F26E4.6 and T24G10.2) the result is very flat and doesn't looks so informative but some of the results looks really impressive (i.e. C37C3.2 and Y45F10B.3). Perhaps these are the genes which regulates significantly by this transcription factor. For some cases the error bar is quite high. False positive could be an issue here. The magnitude of TFA also differs from one to another. We picked another random transcription factor T20B12.8.3. Figure 3.3 shows its activity on different genes.

3.3.1 Gene with multiple regulators

For the case of multi input motif a single gene could be regulated by multiple transcription factor. Our developed tool can determine the posteriori of the relative weight for the different transcription factors regulating the genes. Table 3.1 shows some examples. Gene C44B12.5 can be regulated by transcription factor Y116A8C.35 and F33A8.3. While gene Y105E8B.3 is regulated by T20B12.8, F33A8.3, Y116A8C.35, F11A10.2 and C16A3.7. Though for some cases the expression level is quite low and noise margin is significantly high but we can rank these gene using Kalaitzis and Lawrence [2011].

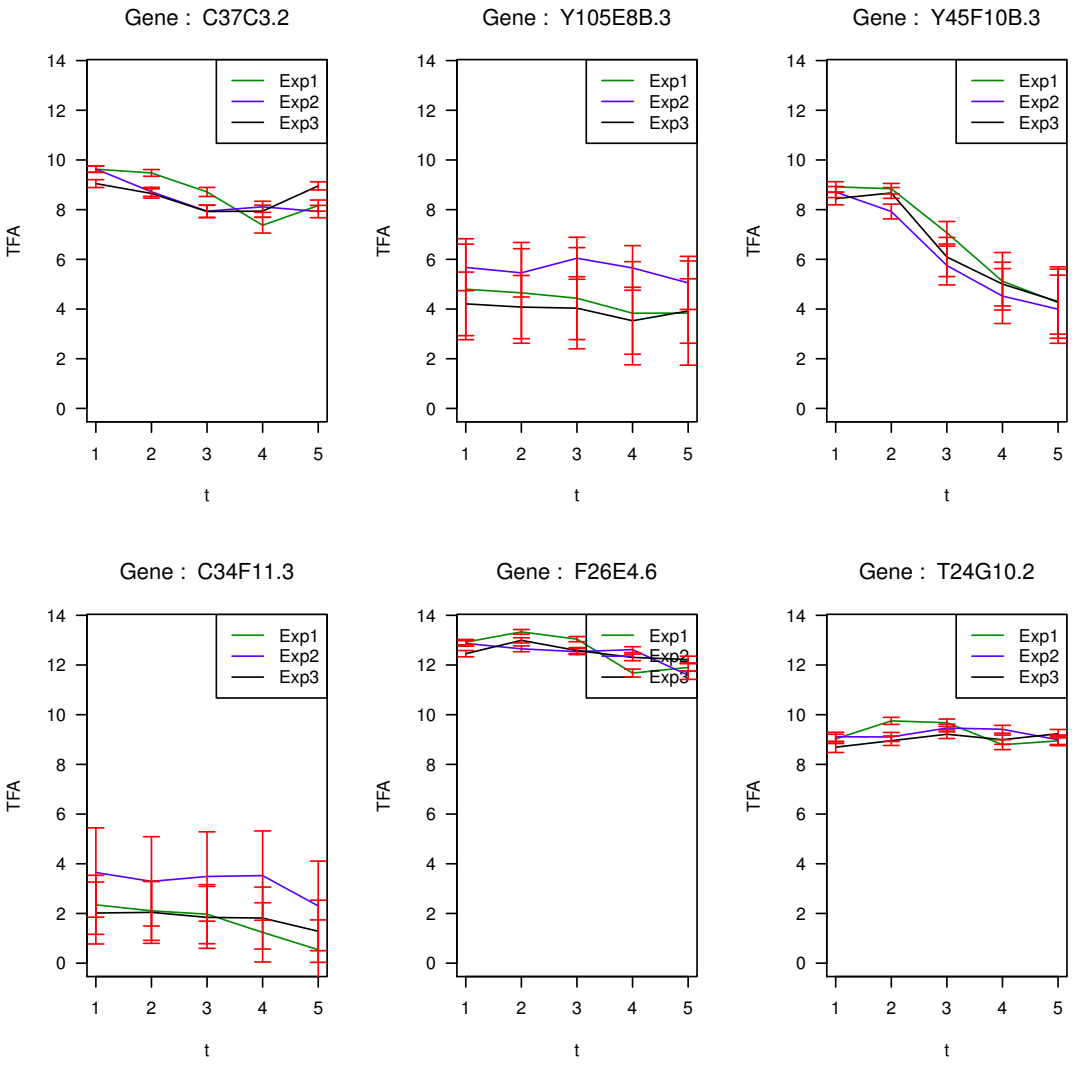

FIGURE 3.2: Gene Specific transcription factor activity of ZK370.2

3.3.2 Different clusters and corresponding active TF

Andrew Cossins [citation required!] made some clusters based on different properties of the genes and its subsequent activities of the cell properties. Here are the basic clusters and some of the phenotype properties-

Cluster 1 - Chill upregulated basically related with cell morphogenesis, cell growth, regulation of cell size, electron transport regulation of cell growth, generation of precursor metabolites and energy, anatomical structure morphogenesis, cellular metabolic process, proteolysis, etc.

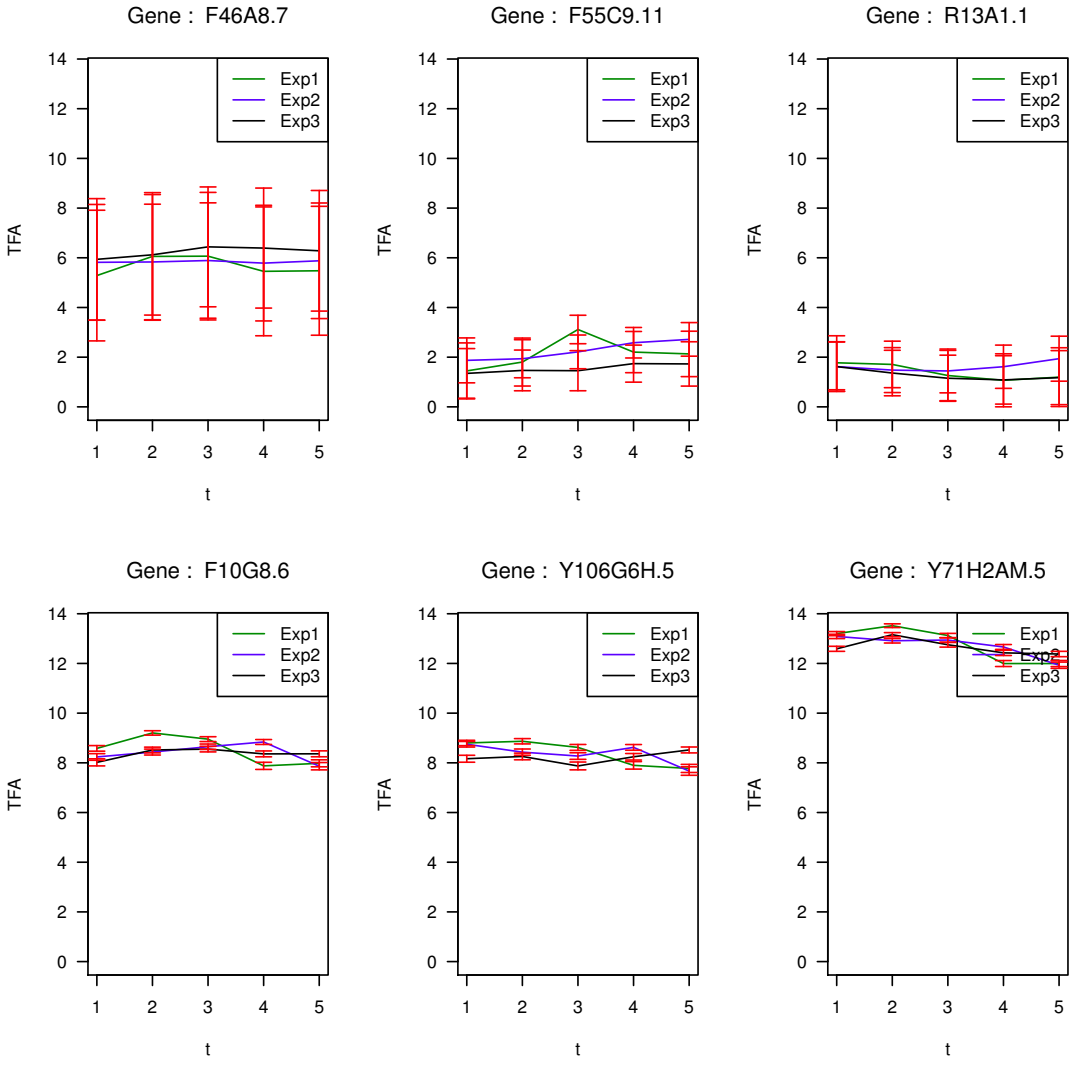

FIGURE 3.3: Gene Specific transcription factor activity of T20B12.8.3

Gene Name	Regulators activity
C44B12.5	Y116A8C.35 = 1.719797 ± 3.493205, F33A8.3 = 1.415785 ± 3.492985
Y105E8B.3	Y54G2A.1 = 0.07157665 ± 1.2222137 F33D11.12 = 0.03861905 ± 0.7252534 ZK370.2 = −1.20157055 ± 2.0318513
Y105E8B.3	T20B12.8 = 0.25474933 ± 2.5665869 F33A8.3 = 0.11619828 ± 3.5107742 Y116A8C.35 = 0.03289664 ± 3.8071374 F11A10.2 = 0.03016348 ± 1.7737585 C16A3.7 = 0.01883489± 0.9431105

TABLE 3.1: Genes regulated by multiple TF

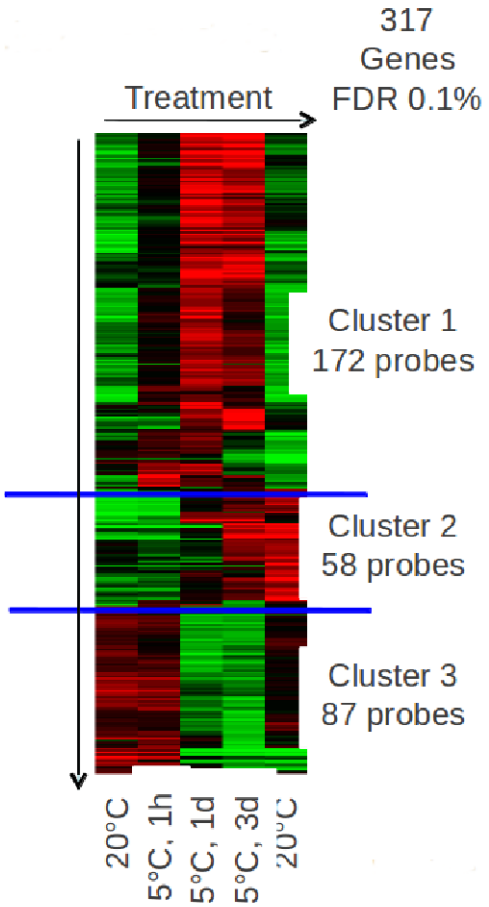

FIGURE 3.4: Clustering of TF

Clusters	Active TF
1. Chill upregulated	**6**
2. Chill late upregulated	**245**
3. Chill downregulated	**128**
4. Others	**203**

TABLE 3.2: Active TF on different clusters

Cluster 2 - Chill late upregulated related with chromosome organization and biogenesis, DNA packaging, chromatin architecture chromatin modification, negative regulation of developmental process, chromatin remodeling, regulation of developmental process, DNA metabolic process larval development (sensu Nematoda), organelle organization and biogenesis, post-embryonic development etc.

Cluster 3 - Chill downregulated genes related with amino acid and derivative metabolic process, carboxylic acid metabolic process, organic acid metabolic process, fatty acid metabolic process, amino acid metabolic process, monocarboxylic acid metabolic process, etc. Rest of the genes were placed in the group 'Others'.

Figure 3.4 shows heeat map generated from DNA microarray data to reflect the gene expression values at different temperature and their basic clusters[Ref required! taken Andrew Cossins presentation]. Based on the above clusters we have tried to find out the the transcription factors active for different clusters. We have managed to managed to find out the active transcription factor for each clusters and Table 3.2 shows the numbers.

3.4 Ranking Differentially expressed gene expressions

Kalaitzis and Lawrence [2011] analysed the time series gene expression and tried to filter the quiet or inactive genes from the differentially expressed genes. They have devoloped the model considering the temporal nature of data using Gaussian process. We have used this model to rank our time series gene expression and ranked the differentially expressed gene expressions. We tried to rank the three replicates of our data seperately and later determine the Pearson correlation between ranking score of different samples. Figure 3.5 shows the Pearson correlation between different ranking scores. The correlation coffecient for all three relations (between sample 1 and sample 2, sample 2 and sample 3 and sample 3 and sample 1) were quite high. Which indicates the similarity of differentially expressed genes and quiet genes of different samples or replication of time series data. So, if required, based on these ranking we can easily filter out some of the quiet genes and keep the other genes for further experiments.

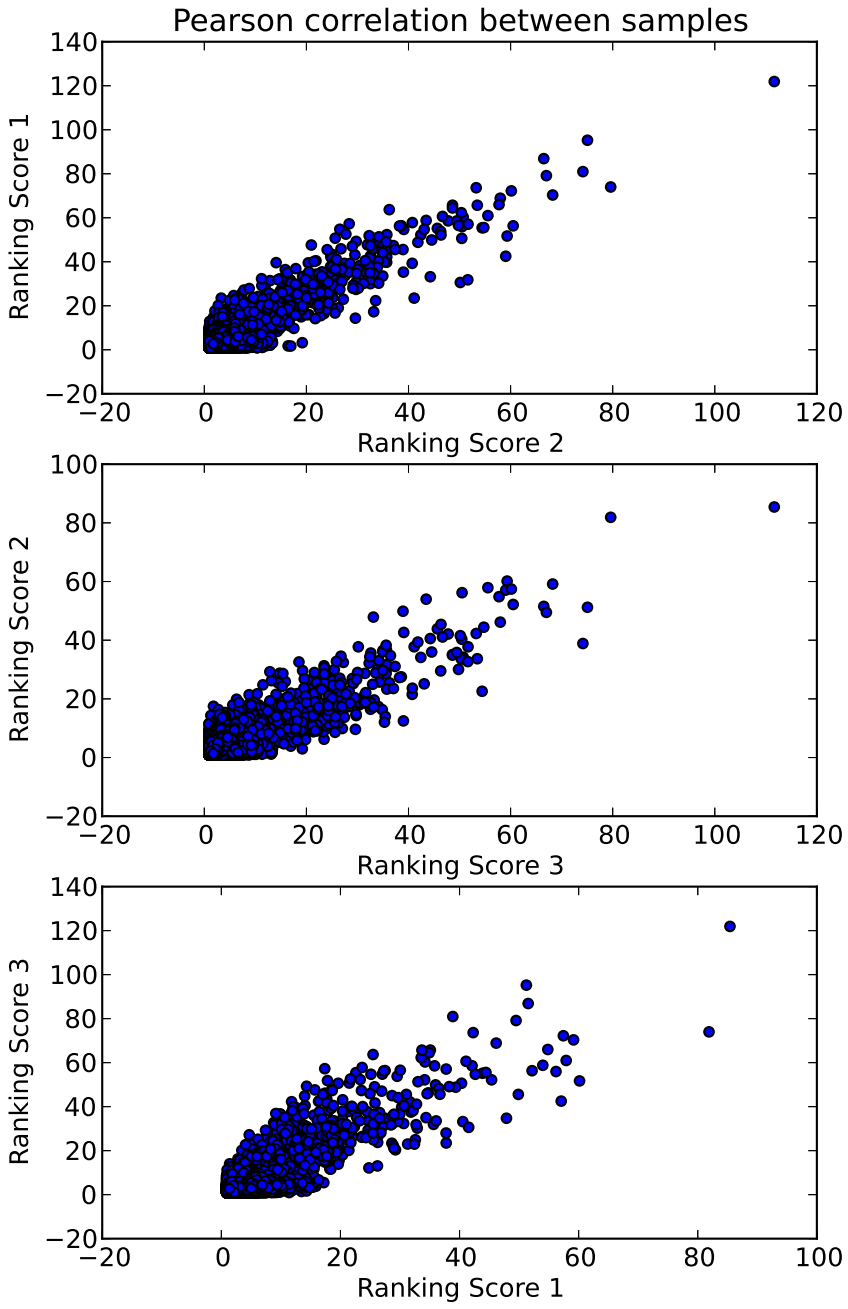

FIGURE 3.5: Pearson's correlation between different ranking scores

Chapter 4

Gaussian Process Regression

4.1 Brief history of Gaussian Process

The Gaussian processes is one of the most simple and widely used families of stochastic processes for modeling dependent data observed over time, or space, or time and space together. As a general setting, Gaussian process of many types have been studied and incorporated in research for centuries. The Wiener process (e.g. Papoulis [1991]) (one of the best known Lévy processes) is a particular type of Gaussian process. The story of using Gaussian process is still a long one. Probably Gaussian process were first used for time series prediction, Kolmogorov [1941], Wiener [1949] worked in this domain for this purpose date backs to the 1940's.

Since the 1970's Gaussian process have been widely adopted in the field of meteorology and geostatistics. Around that time Gaussian process regression was named as kriging and used by Matheron [1973] for prediction in geostatistics. O'Hagan [1978] used Gaussian process in the field of statistics for multivariate input regression problem. For general purpose function approximators Bishop [1995] used neural networks, Neal [1996] showed the link between Gaussian process and neural networks and in the machine learning context Williams and Rasmussen [1996] first described Gaussian process regression.

Over the last two decades Gaussian process in machine learning has turned to a major interest and much work has been done. Rasmussen and Williams [2006] is the must read book on Gaussian process for machine learning and most of the discussed in this chapter can be found there in detail form.

4.2 The regression problem

Machine learning problem can be roughly categorized into three basic classes. Supervised learning: inferring a function from labeled training data, unsupervised learning: to find hidden structure of unlabeled data and reinforcement learning: take action by maximizing the cumulative reward. MacKay [2003], Bishop [2006] describes the concepts in detail. Supervised learning may be further sub-categorized in two fundamental tasks: regression and classification. Regression problem deals with estimating the relationship among some dependent variables with some independent variables, whereas classification identifies the desired discrete output levels.

Regression is the task of making some prediction of a continuous output variable at a desired input, based on a training input output data set. The input data can be any type of object or real valued features located in $\mathbb{R}^D$ which have some predictability for an unobserved location.

By definition of regression, it is obvious that there will be some inference based on a function mapping the outputs from a set of given inputs, because by inferring a function we can predict the response for a desired input. In the case of Bayesian inference, a prior distribution over function is required. Then the model go through some training process and update the prior, based on the training data set $\mathcal{D}$ constructed with N input vectors, such as $\{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{X} \equiv \{\mathbf{x}_n\}_{n=1}^N$, $\mathbf{x}_n \in \mathbb{R}^D$ are the training inputs and $\mathbf{y} \equiv \{y_n\}_{n=1}^N$, $\mathbf{y}_i \in \mathbb{R}$ are the training outputs. Now a key question arises, how can we consider a distribution over an infinite dimensional object as a function?

Although using plain and simple statistics regression problem can be solved, but to model a more complex and specific learning task with improved reliability and robustness Gaussian process is a better selection. Gaussian process models can be used for regression model having an object featuring infinite dimensionality. Even at present Gaussian process have been advanced beyond the regression model and now using for classification cite, unsupervised learningcite, reinforcement learning cite and many more. TODO

We assume the outputs considered at the training level may contain some noise and observed from the underlying mapping function $f(\mathbf{x})$. The objective of the regression problem is to construct $f(\mathbf{x})$ from the data $\mathcal{D}$. This task is ill-defined and dealing with noisy data leads to an exercise in reasoning under uncertainty. Hence, a single estimate of $f(\mathbf{x})$ clearly could be misleading, rather a probability distribution over likely functions could be much more appealing. A regression model based on Gaussian process is a fully probabilistic Bayesian model, and definitely will serve our purpose. In contrast with other regression model, here we will get the opportunity to choose the best estimate of

$f(\mathbf{x})$. If we consider a probability distribution on functions $p(f)$ as the Bayesian prior for regression, then from data Bayesian inference can be used to make predictions:

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f)p(f)}{p(\mathcal{D})} \tag{4.1}$$

The dynamic activity of transcription factors can be viewed as a regression task.

4.3 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 kind of continuous stochastic process and defines probability distribution for functions. It can be also viewed as a random variables indexed by a continuous variable: $f(\mathbf{x})$ chosen from a random function variables $\mathbf{f} = \{f_1, f_2, f_3, ..., f_N\}$, with corresponding indexed inputs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_N\}$. In Gaussian process, variables from these random functions are normally distributed and as a whole can be represent as a multivariate Gaussian distribution:

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}) \tag{4.2}$$

where $\boldsymbol{\mu}$ is the mean and $\mathbf{K}$ is covariance of Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$ 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 $f(\mathbf{x})$ is a real process, a Gaussian process is completely defined by its mean function and covariance function given in 4.3 and 4.4 respectively. Usually the $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x}')$ are defined as-

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{4.3}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \tag{4.4}$$

where $\mathbb{E}$ represents the expected value. We denote the Gaussian process as-

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{4.5}$$

The covariance matrix $\mathbf{K}$ is constructed from the covariance function $k(\mathbf{x}, \mathbf{x}')$ and $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

4.4 GP: Covariances

Though for convenience, we often define the mean of the prior of the GP as zero but the posterior mean of the GP $p(f|\mathcal{D})$ obtained from the GP regression is not a zero mean process.

Based on our problem we are free to design our covariance function. The mandatory requirement of a covariance matrix is symmetric positive semi-definite. So as long as the covariance function generates symmetric positive semi-definite matrix, we can use that function for a Gaussian process. Smoothness, periodicity, amplitude, lengthscale etc. are the basic properties while choosing a Gaussian Process covariance function. It is very crucial to choose a appropriate function for further Gaussian Process Modelling. One of the main goal of this thesis is to develop a covariance function which will solve our problem, hopefully more robust and flexible way. Here first we will discuss about some of the very well known and widely used covariance functions. The in detail description will be found at Rasmussen and Williams [2006].

4.4.1 Exponentiated Quadratic covariance function

Exponentiated Quadratic covariance is the most widely used covariance function for Gaussian process. This is also known as squared exponential (SE) covariance or radial basis function (RBF). The exponentiated quadratic has become the de-facto default kernel for Gaussian process and has the following form.

$$K_{EQ}(r) = a^2 \exp\left(-\frac{r^2}{2l^2}\right) \tag{4.6}$$

where $r = \|\mathbf{x} - \mathbf{x}'\|$. Here $\|\mathbf{x} - \mathbf{x}'\|$ is invariant to translation and rotation. So, Exponentiated Quadratic covariance is stationary, as well as isotropic. Here the parameter for output variance a and lengthscale parameter l govern the property of sample functions and commonly known as hyperparameters. Parameter a determines the typical amplitude, i.e. average distance of the function away from the mean. l control the lengthscale, i.e. the length of the wiggles of the function.

Fighure 4.1 represents the kernel and random sample functions drawn from Gaussian process using Exponentiated Quadratic covariance with different lengthscale and amplitude hyperparameter. The random function was generated for a given input range by drawing a sample from the multivariate Gaussian using equation 4.2 with zero mean. The smoothness of the sample function depends on the equation 4.6. Function variable located closer in the input space are highly correlated, whereas function variable located

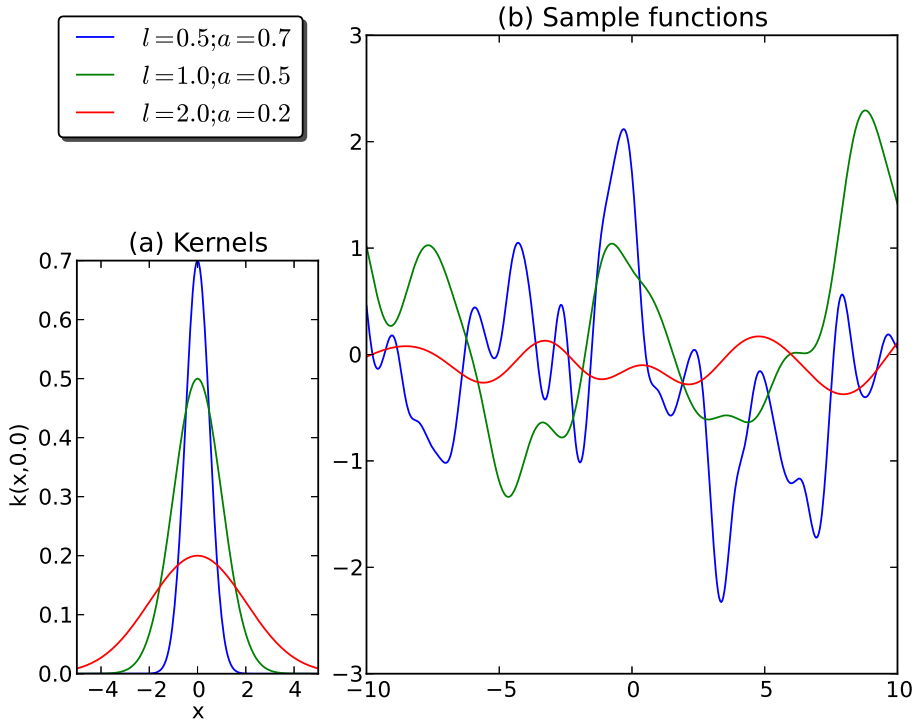

FIGURE 4.1: Exponentiated Quadratic kernel and sample functions

at distance are loosely correlated or even uncorrelated. Exponentiated Quadratic covariance might be too smooth to perform any realistic regression task. Depending on the basic nature of the function other covariance function could be interesting.

4.4.2 Rational Quadratic covariance function

Rational Quadratic covariance function is equivalent to adding together multiple exponentiated quadratic covariance function having different lengthscale. Gaussian process prior kernel function expect smooth function with many lengthscale. Here the parameter α can control the relative weights for lengthscale variations. Exponentiated quadratic covariance function can be viewed as a special case of rational quadratic covariance function. If $\alpha \to \infty$, then both of the functions become identical.

$$K_{RQ}(r) = a^2 \left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha} \tag{4.7}$$

where $r = \|\mathbf{x} - \mathbf{x}'\|$. Figure 4.2 (a) shows the kernels and (b) shows three different random sample functions drawn with different setting of hyperparameters a and l.

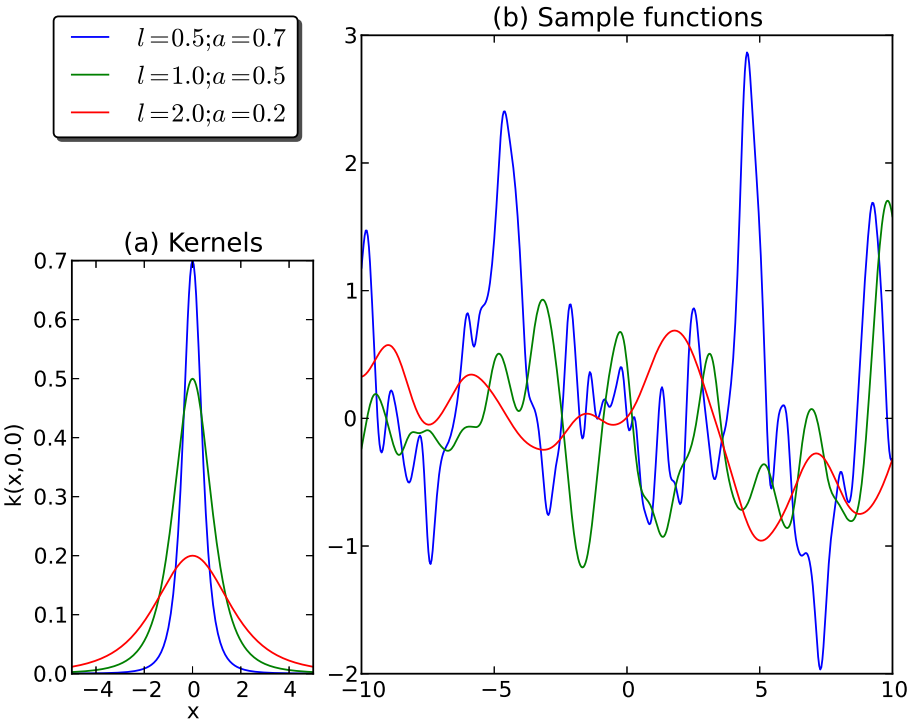

FIGURE 4.2: Rational Quadratic kernel and random sample functions

4.4.3 The Matérn covariance function

The Matérn class of covariance function are given by equation 4.8-

$$K_{Mat}(r) = a^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} r}{l} \right) \tag{4.8}$$

where a, l, ν are positive hyperparameter, K_{ν} is a modified Bessel function and $\Gamma(.)$ is the Gamma function. Hyperparameter ν controls the roughness of the function and as like Exponentiated quadratic covariance function the parameters a and l controls the amplitude and lengthscale respectively. Though for $\nu \to \infty$ we can obtain the exponentiated quadratic kernel, but for finite value of ν the sample functions are significantly rough. The simpler form of Matérn covariance function is obtained when ν is half integer: $\nu = p + 1/2$, where p is a non-negative integer.The covariance function can be expressed as a product of an exponential and a polynomial of order p. Abramowitz and Stegun [1965] derived the general expression as follows-

$$K_{\nu=p+1/2}(r) = \exp\left(-\frac{\sqrt{2\nu} r}{l} \right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+i)!}{i!\,(p-i)!} \left(\frac{\sqrt{8\nu} r}{l} \right)^{p-i} \tag{4.9}$$

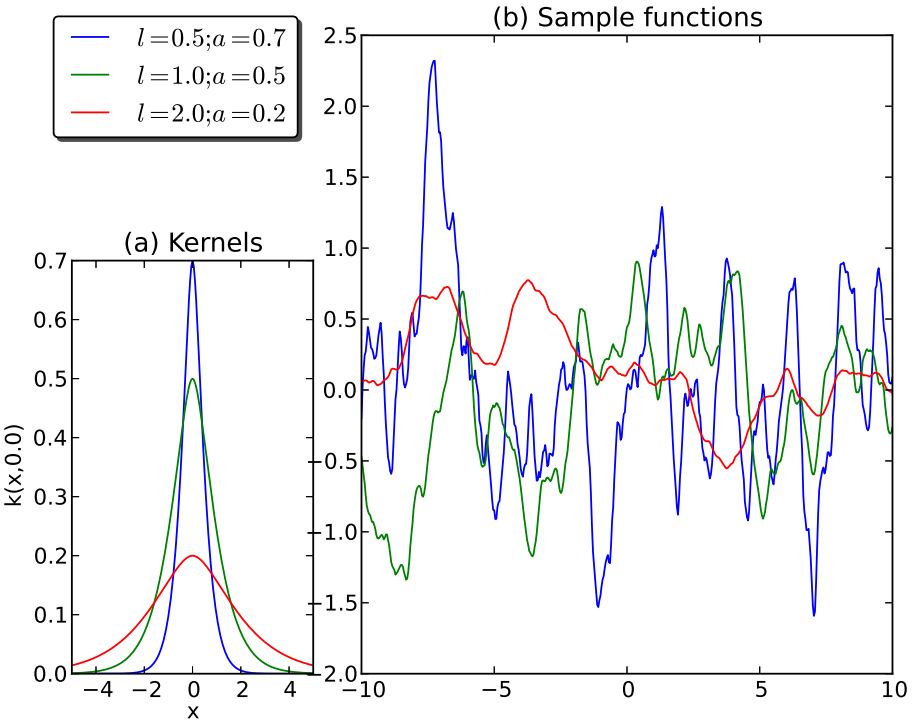

FIGURE 4.3: The Matérn32 kernel and random sample functions

The most interesting cases for machine learning are $\nu = 3/2$ and $\nu = 5/2$, for which we get the following equations respectively-

$$K_{\nu=3/2}(r) = \left(1 + \frac{\sqrt{3}r}{l}\right) \exp\left(-\frac{\sqrt{3}r}{l}\right) \tag{4.10}$$

$$K_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}r}{l}\right) \tag{4.11}$$

4.4.4 The Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck process Uhlenbeck and Ornstein [1930] is a special case of Matérn class covariance functions. The Ornstein-Uhlenbeck process was developed as a mathematical model of the velocity of a particle moving with Brownian motion. The OU process can be found setting up $\nu = 1/2$ and expressed as equation 4.12. Figure4.4 shows the sample functions form the OU process having the exactly same amplitude parameter$a$ and lengthscale parameter l.

$$K_{\nu=1/2}(r) = \exp\left(-\frac{r}{l}\right) \tag{4.12}$$

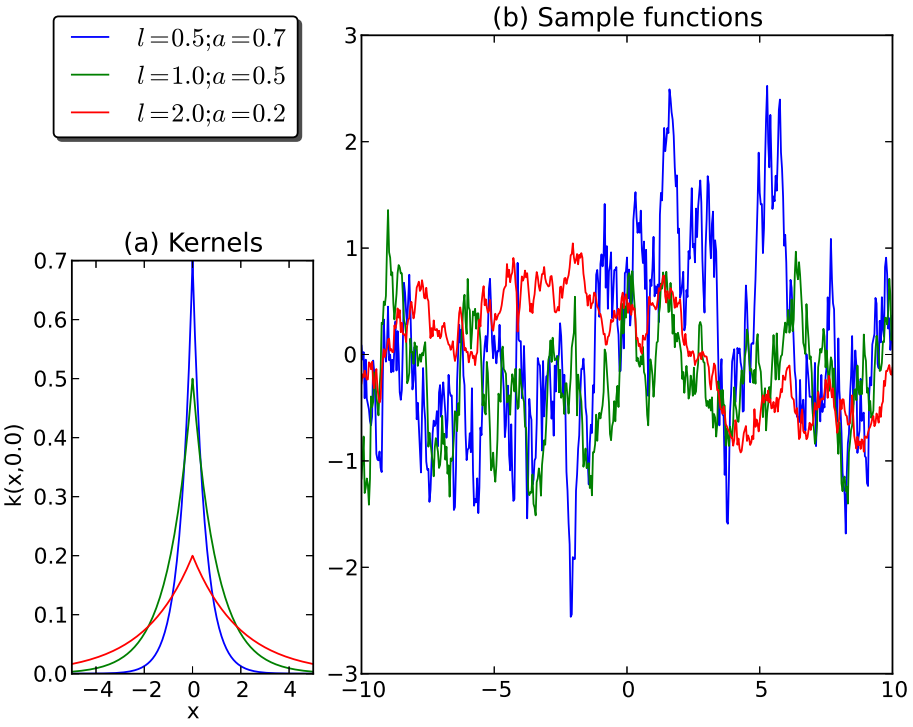

FIGURE 4.4: The OU kernel and random sample functions

4.5 Gaussian process regression

Gaussian Process regression can be done using the marginal and conditional properties of multivariate Gaussian distribution. Lets consider that we have some observations $\mathbf{f}$ of a function at observation point $\mathbf{x}$. Now we wish to predict the values of that function at observation points $\mathbf{x}_\star$, which we are representing by $\mathbf{f}_\star$. Then the joint probability of $\mathbf{f}$ and $\mathbf{f}_\star$ can be obtained from equation 4.13-

$$p\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{f}_\star\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{f}_\star\end{bmatrix}\middle|\mathbf{0}, \begin{bmatrix}\mathbf{K}_{\mathbf{x},\mathbf{x}} & \mathbf{K}_{\mathbf{x},\mathbf{x}_\star}\\ \mathbf{K}_{\mathbf{x}_\star,\mathbf{x}} & \mathbf{K}_{\mathbf{x}_\star,\mathbf{x}_\star}\end{bmatrix}\right) \tag{4.13}$$

where the covariance matrix $\mathbf{K}_{\mathbf{x},\mathbf{x}}$ has elements derived from the covariance function $k(x,x')$, such that the $(i,j)^{th}$ element of $\mathbf{K}_{\mathbf{x},\mathbf{x}}$ is given by $k(\mathbf{x}[i],\mathbf{x}[i])$ The conditional property of a multivariate Gaussian is used to perform regression the. The conditional property is can be represented by the equation 4.14:

$$p(\mathbf{f}|\mathbf{f}_\star) = \mathcal{N}\left(\mathbf{f}_\star\middle|\mathbf{K}_{\mathbf{x}_\star,\mathbf{x}}\mathbf{K}_{\mathbf{x},\mathbf{x}}^{-1}\mathbf{f}, \mathbf{K}_{\mathbf{x}_\star,\mathbf{x}_\star} - \mathbf{K}_{\mathbf{x}_\star,\mathbf{x}}\mathbf{K}_{\mathbf{x},\mathbf{x}}^{-1}\mathbf{K}_{\mathbf{x},\mathbf{x}_\star}\right) \tag{4.14}$$

Though in ideal case observation $\mathbf{f}$ is noise free but in practice it is always corrupted with some noise. Lets consider $\mathbf{y}$ is the correlated version of $\mathbf{f}$. If we consider this noise is Gaussian noise then we can write $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I})$, where σ^2 is the variance of the noise and $\mathbf{I}$ is the identity matrix with appropriate size and marginalise the observation $\mathbf{f}$. Then the joint probability of $\mathbf{y}$ and $\mathbf{f}_\star$ can be represented by the equation 4.15

$$p\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{f}_\star\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{f}_\star\end{bmatrix}\middle|\mathbf{0}, \begin{bmatrix}\mathbf{K}_{\mathbf{x},\mathbf{x}} + \sigma^2\mathbf{I} & \mathbf{K}_{\mathbf{x},\mathbf{x}_\star}\\ \mathbf{K}_{\mathbf{x}_\star,\mathbf{x}} & \mathbf{K}_{\mathbf{x}_\star,\mathbf{x}_\star}\end{bmatrix}\right) \tag{4.15}$$

Regression with Gaussian process is 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 4.14. To construct the covariance function still we need to consider the hyperparameters and optimize those. The most efficient and commonly used optimization technique for hyperparameters can be done using maximum likelihood. If we consider all the hyperparameters α, σ^2 and l in to a vector $\boldsymbol{\theta}$, then we can use gradient methods to optimize $p(\mathbf{y}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The Log maximum likelihood is given by

$$p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{D}{2}log2\pi - \frac{1}{2} \times log\left|\mathbf{K}_{\mathbf{x},\mathbf{x}} + \sigma^2\mathbf{I}\right| - \frac{1}{2}\mathbf{y}^T\left[\mathbf{K}_{\mathbf{x},\mathbf{x}} + \sigma^2\mathbf{I}\right]^{-1}\mathbf{y} \tag{4.16}$$

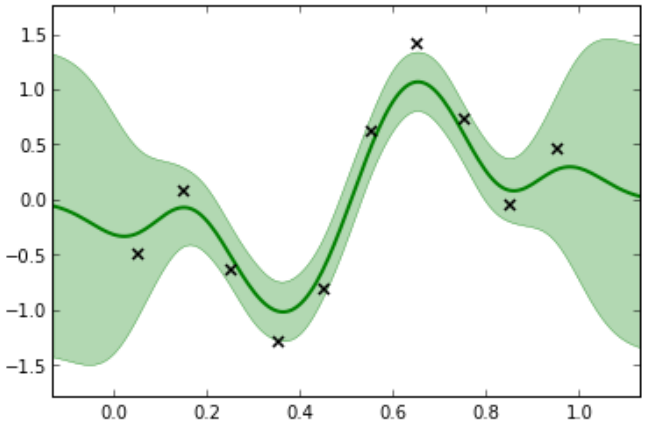

FIGURE 4.5: Simple example of regression using Gaussian Process

Chapter 5

Gaussian Process Model of Gene Expressions

In this chapter we design a covariance function for reconstructing transcription factor activities given gene expression profiles and a connectivity matrix (binding data) between genes and transcription factors. Our modelling framework builds on ideas in Sanguinetti et al. [2006] who used a linear-Gaussian statespace modelling framework to infer the transcription factor activity of a group of genes.

We note that the linear Gaussian model is equivalent to a Gaussian process with a particular covariance function. We therefore build a model directly from the Gaussian process perspective to achieve the same effect. We introduce a computational trick, based on judicious application of singluar value decomposition, to enable us to efficiently fit the Gaussian process in a reduced 'TF activity' space.

First we load in the classic Spellman et al. [1998] Yeast Cell Cycle data set. The cdc15 time series data has 23 time points. We can load this gene expression data in with GPy.

Time series of synchronized yeast cells from the CDC-15 experiment of Spellman et al. [1998]. Two colour spotted cDNA array data set of a series of experiments to identify which genes in Yeast are cell cycle regulated. We can make a simple helper function to plot genes from the data set (which are provided as a pandas array).

Our second data set is from ChiP-chip experiments performed on yeast by Lee et al. [2002]. These give us the binding information between transcription factors and genes. In this notebook we are going to try and combine this binding information with the gene expression information to infer transcription factor activities.

5.1 Model for Transcription Factor Activities

We are working with *log* expression levels in a matrix $\mathbf{Y} \in \Re^{n \times T}$ and we will assume a linear (additive) model giving the relationship between the expression level of the gene and the corresponding transcription factor activity which are unobserved, but we represent by a matrix $\mathbf{F} \in \Re^{q \times T}$. Our basic assumption is as follows. Transcription factors are in time series, so they are likely to be temporally smooth. Further we assume that the transcription factors are potentially correlated with one another (to account for transcription factors that operate in unison).

5.1.1 Correlation Between Transcription Factors

If there are q transcription factors then correlation between different transcription factors is encoded in a covariance matrix, $\mathbf{\Sigma}$ which is $q \times q$ in dimensionality.

5.1.2 Temporal Smoothness

Further we assume that the log of the transcription factors' activities is temporally smooth, and drawn from an underlying Gaussian process with covariance $\mathbf{K}_t$.

5.1.3 Intrinsic Coregionalization Model

We assume that the joint process across all q transcription factor activities and across all time points is well represented by an intrinsic model of coregionalization where the covariance is given by the Kronecker product of these terms.

$$\mathbf{K}_f = \mathbf{K}_t \otimes \mathbf{\Sigma}$$

This is known as an intrinsic coregionalization model Wackernagel [2003]. Alvarez et al. [2012] presented the machine learning orientated review of these methods. The matrix $\mathbf{\Sigma}$ is known as the coregionalization matrix.

5.2 Relation to Gene Expressions

We now assume that the jth gene's expression is given by the product of the transcription factors that bind to that gene. Because we are working in log space, that implies a log linear relationship. At the ith time point, the log of the jth gene's expression, $\mathbf{y}_{i,j}$ is

linearly related to the log of the transcription factor activities at the corresponding time point, $\mathbf{f}_{i,:}$. This relationship is given by the binding information from $\mathbf{S}$. We then assume that there is some corrupting Gaussian noise to give us the final observation.

$$\mathbf{y}_{i,j} = \mathbf{S}\mathbf{f}_{:,i} + \boldsymbol{\epsilon}_i$$

where the Gaussian noise is sampled from

$$\epsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}).$$

5.3 Gaussian Process Model of Gene Expression

We consider a vector operator which takes all the separate time series in $\mathbf{Y}$ and stacks the time series to form a new vector $n \times T$ length vector $\mathbf{y}$. A similar operation is applied to form a $q \times T$ length vector $\mathbf{f}$. Using Kronecker products we can now represent the relationship between $\mathbf{y}$ and $\mathbf{f}$ as follows: Standard properties of multivariate Gaussian distributions tell us that

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}),$$

where

$$\mathbf{K} = \mathbf{K}_t \otimes \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top + \sigma^2\mathbf{I}.$$

This results in a covariance function that is of size n by T where n is number of genes and T is number of time points. However, we can get a drastic reduction in the size of the covariance function by considering the singular value decomposition of $\mathbf{S}$. The matrix $\mathbf{S}$ is n by q matrix, where q is the number of transcription factors. It contains a 1 if a given transcription factor binds to a given gene, and zero otherwise.

$$L = -\frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\mathbf{y}^\top\mathbf{K}^{-1}\mathbf{y}$$

In the worst case, because the vector $\mathbf{y}$ contains $T \times n$ points (T time points for each of n genes) we are faced with $O(T^3n^3)$ computational complexity. We are going to use a rotation trick to help.

5.4 The Main Computational Trick

5.4.1 Rotating the Basis of a Multivariate Gaussian

For any multivariate Gaussian you can rotate the data set and compute a new roated covariance which is valid for the rotated data set. Mathematically this works by first inserting $\mathbf{R}\mathbf{R}^\top$ into the likelihood at three points as follows:

$$L = -\frac{1}{2}\log|\mathbf{K}\mathbf{R}^\top\mathbf{R}| - \frac{1}{2}\mathbf{y}^\top\mathbf{R}^\top\mathbf{R}\mathbf{K}^{-1}\mathbf{R}^\top\mathbf{R}\mathbf{y} + \text{const}$$

The rules of determinants and a transformation of the data allows us to rewrite the likelihood as

$$L = -\frac{1}{2}\log|\mathbf{R}^\top\mathbf{K}\mathbf{R}| - \frac{1}{2}\hat{\mathbf{y}}^\top\left[\mathbf{R}^\top\mathbf{K}\mathbf{R}\right]^{-1}\hat{\mathbf{y}} + \text{const}$$

where we have introduced the rotated data: $\hat{\mathbf{y}} = \mathbf{R}\mathbf{y}$. Geometrically what this says is that if we want to maintain the same likelihood, then when we rotate our data set by $\mathbf{R}$ we need to rotate either side of the covariance matrix by $\mathbf{R}$, which makes perfect sense when we recall the properties of the multivariate Gaussian.

5.4.2 A Kronecker Rotation

In this notebook we are using a particular structure of covariance which involves a Kronecker product. The rotation we consider will be a Kronecker rotation Stegle et al. [2011]. We are going to try and take advantage of the fact that the matrix $\mathbf{S}$ is square meaning that $\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top$ is not full rank (it has rank of most q, but is size $n \times n$, and we expect number of transcription factors q to be less than number of genes n).

When ranks are involved, it is always a good idea to look at singular value decompositions (SVDs). The SVD of $\mathbf{S}$ is given by:

$$\mathbf{S} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{V}^\top$$

where $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$, $\boldsymbol{\Lambda}$ is a diagonal matrix of positive values, $\mathbf{Q}$ is a matrix of size $n \times q$: it matches the dimensionality of $\mathbf{S}$, but we have $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$. Note that because it is not square, $\mathbf{Q}$ is not in itself a rotation matrix. However it could be seen as the first q columns of an n dimensional rotation matrix (assuming n is larger than q, i.e. there are more genes than transcription factors).

If we call the $n - q$ missing columns of this rotation matrix $\mathbf{U}$ then we have a valid rotation matrix $\mathbf{R} = \begin{bmatrix}\mathbf{Q} & \mathbf{U}\end{bmatrix}$. Although this rotation matrix is only rotating across the

n dimensions of the genes, not the additional dimensions across time. In other words we are choosing $\mathbf{K}_t$ to be unrotated. To represent this properly for our covariance we need to set $\mathbf{R} = \mathbf{I} \otimes \begin{bmatrix}\mathbf{Q} & \mathbf{U}\end{bmatrix}$. This gives us a structure that when applied to a covariance of the form $\mathbf{K}_t \otimes \mathbf{K}_n$ it will rotate $\mathbf{K}_n$ whilst leaving $\mathbf{K}_t$ untouched.

When we apply this rotation matrix to $\mathbf{K}$ we have to consider two terms, the rotation of $\mathbf{K}_t \otimes \mathbf{S}\mathbf{\Sigma}\mathbf{S}^\top$, and the rotation of $\sigma^2\mathbf{I}$.

Rotating the latter is easy, because it is just the identity multiplied by a scalar so it remains unchanged

$$\mathbf{R}^\top \mathbf{I}\sigma^2 \mathbf{R} = \mathbf{I}\sigma^2$$

The former is slightly more involved, for that term we have

$$\left[\mathbf{I} \otimes \begin{bmatrix}\mathbf{Q} & \mathbf{U}\end{bmatrix}^\top\right] \mathbf{K}_t \otimes \mathbf{S}\mathbf{\Sigma}\mathbf{S}^\top \left[\mathbf{I} \otimes \begin{bmatrix}\mathbf{Q} & \mathbf{U}\end{bmatrix}\right] = \mathbf{K}_t \otimes \begin{bmatrix}\mathbf{Q} & \mathbf{U}\end{bmatrix}^\top \mathbf{S}\mathbf{\Sigma}\mathbf{S}^\top \begin{bmatrix}\mathbf{Q} & \mathbf{U}\end{bmatrix}.$$

Since $\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{V}^\top$ then we have

$$\begin{bmatrix}\mathbf{Q} & \mathbf{U}\end{bmatrix}^\top \mathbf{S}\mathbf{\Sigma}\mathbf{S}^\top \begin{bmatrix}\mathbf{Q} & \mathbf{U}\end{bmatrix} = \begin{bmatrix}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}\mathbf{V}\mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}.$$

This prompts us to split our vector $\hat{\mathbf{y}}$ into a q dimensional vector $\hat{\mathbf{y}}_u = \mathbf{U}^\top\mathbf{y}$ and an $n - q$ dimensional vector $\hat{\mathbf{y}}_q = \mathbf{Q}^\top\mathbf{y}$. The Gaussian likelihood can be written as

$$L = L_u + L_q + \text{const}$$

where

$$L_q = -\frac{1}{2}\log|\mathbf{K}_t \otimes \mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}\mathbf{V}\mathbf{\Lambda} + \sigma^2\mathbf{I}| - \frac{1}{2}\hat{\mathbf{y}}_q^\top \left[\mathbf{K}_t \otimes \mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}\mathbf{V}\mathbf{\Lambda} + \sigma^2\mathbf{I}\right]^{-1} \hat{\mathbf{y}}_q$$

and

$$L_u = -\frac{T(n-q)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\hat{\mathbf{y}}_u^\top\hat{\mathbf{y}}_u$$

Strictly speaking we should fit these models jointly, but for the purposes of illustration we will firstly use a simple procedure. Firstly, we fit the noise variance σ^2 on $\hat{\mathbf{y}}_u$ alone using L_u. Once this is done, fix the value of σ^2 in L_q and optimize with respect to the other parameters.

With the current design the model is switching off the temporal correlation. The next step in the analysis will be to reimplement the same model as described by Sanguinetti et al. [2006] and recover their results. That will involve using an Ornstein Uhlbeck covariance and joint maximisation of the likelihoood of L_u and L_q.

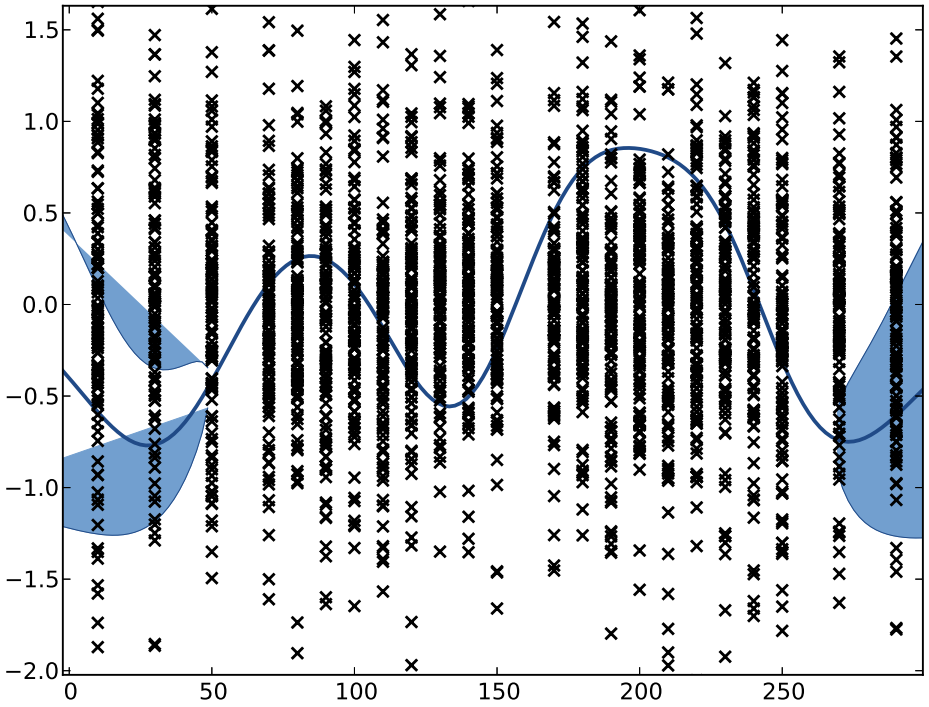

FIGURE 5.1: Gene Specific transcription factor activity of YER124C

Chapter 6

Conclusion and Future work

We have developed a tool based on programming language *R* named *chipDyno* using the model of Sanguinetti et al. [2006] which integrate the connectivity information between genes and transcription factors, and micro array data. The probabilistic nature of the model can determine the significant regulations in a given experimental condition.

Earlier the model was developed for a unicellular microorganism (yeast) but we have successfully manage to determine the gene specific transcription factor activity for C Elegans, a multicellular eukaryote. We were also successful to filter out the quiet genes from the differentially expressed genes.

To elucidate pathways and processes relevant to human biology and disease C. Elegans is been using as a vital model. Different orthology-prediction methods Shaye and Greenwald [2011] are using to compile a list of C. elegans orthologs of human genes. Already a list of 7,663 unique protein-coding genes were resulted in that list and this represents 38% of the 20,250 protein-coding genes predicted in C. elegans. When human genes introduced into C.Elegans human genes replaced their homologous. On the contrary, many C. Elegans genes can function with great deal of similarity to human like mammalian genes. So, the biological insight acquire from C. Elegans may be directly applicable to more complex organism like human.

Lots of computational approaches on gene expression data for time series analysis are not well suited where time points are irregularly spaced. Even in commonly used state-space model time points must occur at regular intervals. On the other side gene expression experiments with regular samples may not be cost effective or optimal from the perspective of statistics. It is expected that models with irregular time points might be more informative if the time points are selected considering some temporal features. Gaussian Process is not restricted to equally spaced time series data. Already Gaussian Process

regression have been successfully applied to overcome this issue and analyse time series data Kalaitzis and Lawrence [2011]. So our expected model will overcome the restriction of temporal sampling of equally spaced time intervals.

6.1 Future Work

Sanguinetti et al. [2006] model to infer the transcription factor activity is a linear-Gaussian state-space model. We believe that this linear Gaussian model is equivalent to Gaussian process with a specific covariance function. We have developed a model directly from Gaussian process to achieve the same goal. We are quite close to develop a valid covariance function for reconstructing transcription factor activities given gene expression profile and binding information between genes and transcription factors. Here we will introduce a computational trick using singular value decomposition and intrinsic coregionalization model. We believe this method will enable us to efficiently fit the Gaussian process in a reduced transcription factor activity space.

Clustering of gene expression time series is another major interest of the research to get the view of groups of co-regulated or associated genes. It is assumed that gene involved in the same biological process will be expressed with a similarity sharing underlying time series. Cossins et al. [2007] did some additional cluster analysis (not published yet!) based on some phenotype properties. Again it is very common to have multiple biological replicates of the gene expression time series data. Just taking average of the replicates surely lead toward discarding insight. Recently Hensman et al. [2013] used a hierarchy of Gaussian Process to model a gene specific and replicate specific temporal covariance. They also used this model for clustering application. Using this Gaussian Process based hierarchical clustering analysis of Hensman et al. [2013] we will try to find some robust clusters for the gene expression data of C Elegans. Once if we can do so, it will easily lead us to find out the active transcription factors related with these clusters and their subsequent dynamic behavior as well.

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. Dover, New York.

Alon, U. (2006). An introduction to systems biology: design principles of biological circuits. *CRC Press*.

Alter, O. and Golub, G. H. (2004). Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between dna replication and rna transcription. *Proceedings of the National Academy of Sciences USA*, 101(47):16577–16582.

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

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

Boulesteix, A.-L. and Strimmer, K. (2005). Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. *Theoretical Biology and Medical Modelling*, 2(23).

Brenner, S. (1974). The genetics of caenorhabditis elegans. *Genetics*, 77:71–94.

Brivanlou, A. H. and Darnell, J. E. (2002). Transcription signal transduction and the control of gene expression. *Science*, 295(5556):819–818.

Byerly, L., Cassada, R., and Russell, R. (1976). The life cycle of the nematode caenorhabditis elegans. i. wild-type growth and reproduction. *Dev Biol*, 51(1):23–33.

Cossins, A. R., Murray, P., Hayward, S. A. L., Govan, G. G., and Gracey, A. Y. (2007). An explicit test of the phospholipid saturation hypothesis of acquired cold tolerance in caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 104(13):5489–5494.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *J. Comput. Biol.*, 7(3/4):601–620.

Gao, F., Foat, B. C., and Bussemaker, H. J. (2004). Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data. *BMC Bioinformatics*, 5(31).

Gerstein, M. B., Lu, Z. J., Nostrand, E. L. V., and Cheng, C. (2010). Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, 330.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo1, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104.

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

Inmaculada, B. M., Philippe, V., Fabien, C., Laurent, J., and Albertha, W. (2007). Edgedb: a transcription factor-dna interaction database for the analysis of c. elegans differential gene expression. *BMC Genomics*, 8(1):21.

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

Karin, M. (1990). Too many transcription factors: positive and negative interactions. *New Biol.*, 2(2):126–131.

Kolmogorov, A. N. (1941). Interpolation und extrapolation von stationären zufälligen folgen. *Bull. Acad. Sci. (Nauk) U.R.S.S. Ser. Math.*, 5:3–14.

Latchman, D. S. (1997). Transcription factors: an overview. *Int. J. Biochem. Cell Biol.*, 29(12):1305–1312.

Lee, I., Lehner, B., Crombie, C., Wang, W., Fraser, A. G., and Marcotte, E. M. (2007). A single network comprising the majority of genes accurately predicts the phenotypic effects of gene perturbation in c. elegans. *Nature Genetics*, 40:181–188.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert,

T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804.

Lee, T. I. and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, 34:77–137.

Liao, J. C., Boscolo, R., Yang, Y.-L., Tran, L. M., Sabatti, C., and Roychowdhury, V. P. (2003). Network component analysis: Reconstruction of regulatory signals in biological systems. *PNAS*, 100(26).

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, 5:439–468.

Mitchell, P. J. and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific dna binding proteins. *Science*, 245(4916):371–378.

Nachman, I., Regev, A., and Friedman, N. (2004). Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20(1).

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics No. 118, New York: Springer-Verlag.

Nikolov, D. B. and Burley, S. K. (1997). Rna polymerase ii transcription initiation: A structural view. *Proc. Natl. Acad. Sci. U.S.A.*, 94(1):15–22.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society B*, 40(1):1–42.

Ong, I. M., Glasner, J. D., and Page, D. (2002). Modelling regulatory pathways in e. coli from time series expression profiles. *Oxford Univ Press*, 18(1):S241–S248.

Palikaras, K. and Tavernarakis, N. (2013). Caenorhabditis elegans (nematode). (1):404–408.

Papoulis, A. (1991). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Companies, 3rd edition.

Ptashne, M. and Gann, A. (1997). Transcriptional activation by recruitment. *Nature*, 386:569–577.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Processes for Machine Learning*.

Roeder, R. G. (1996). The role of general initiation factors in transcription by rna polymerase ii". *Trends Biochem. Sci.*, 21(9):327–335.

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

Shaye, D. D. and Greenwald, I. (2011). Ortholist: A compendium of c. elegans genes with human orthologs. *PLoS ONE*, 9(1).

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297.

Stegle, O., Lippert, C., Mooij, J. M., Lawrence, N. D., and Borgwardt, K. M. (2011). Efficient inference in matrix-variate gaussian models with \iid observation noise. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 630–638. Curran Associates, Inc.

Sulston, J. E. and Horvitz, H. (1977). Post-embryonic cell lineage of the nematode, caenorhabditis elegans. *Developmental Biology*, 56(1):110–156.

Sulston, J. E., Schierenberg, E., j. G. White, and Thomson, J. N. (1980). The embryonic cell lineage of the nematode caenorhabditis elegans. *Developmental Biology*, 100(1):64–119.

Tu, B. P., Kudlicki, A., Rowicka, M., and McKnight, S. L. (2005). Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science*, 310(5751):1152–1158.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of the brownian motion. *Phys. Rev.*, 36:823–841.

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

Wiener, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series.*

Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. *Advances in Neural Information Processing Systems 8, MIT Press*, pages 1–7.

Wood, W. B. (1988). The nematode c. elegans. *Cold Spring Harbor Laboratory Press, New York*, pages 1–16.

WormNet (2014). Wormnet.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals. *Nature*, (434):338–345.