Transcription Factor Activity of *C. elegans*: From Probabilistic Markov model to Gaussian process model

Muhammad Arifur Rahman
Supervisor: Prof. Neil D. Lawrence

Department of Computer Science
and
Sheffield Institute for Translational Neuroscience
The University of Sheffield

M.Rahman@dcs.shef.ac.uk

December 15, 2014

Overview

Introduction
- Motivations

Probabilistic Model with Markov property
- Data and Preprocessing
- Result analysis

GP Model
- Toward the GP model of TFA
- GP model for TFA
- Result analysis

Conclusion and Future Work
- Conclusion
- Future Work

References

Motivations

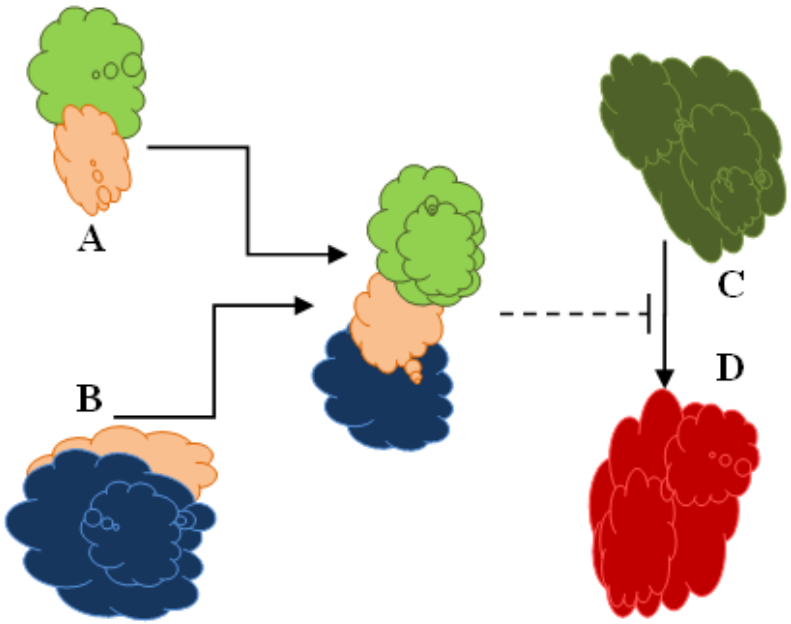


Figure: A cartoon model of protein protein interaction. Two different molecular species A and B bind to form a complex molecular. The newly formed complex hinder the rate at which molecules of species C are transformed to species D.

Motivations

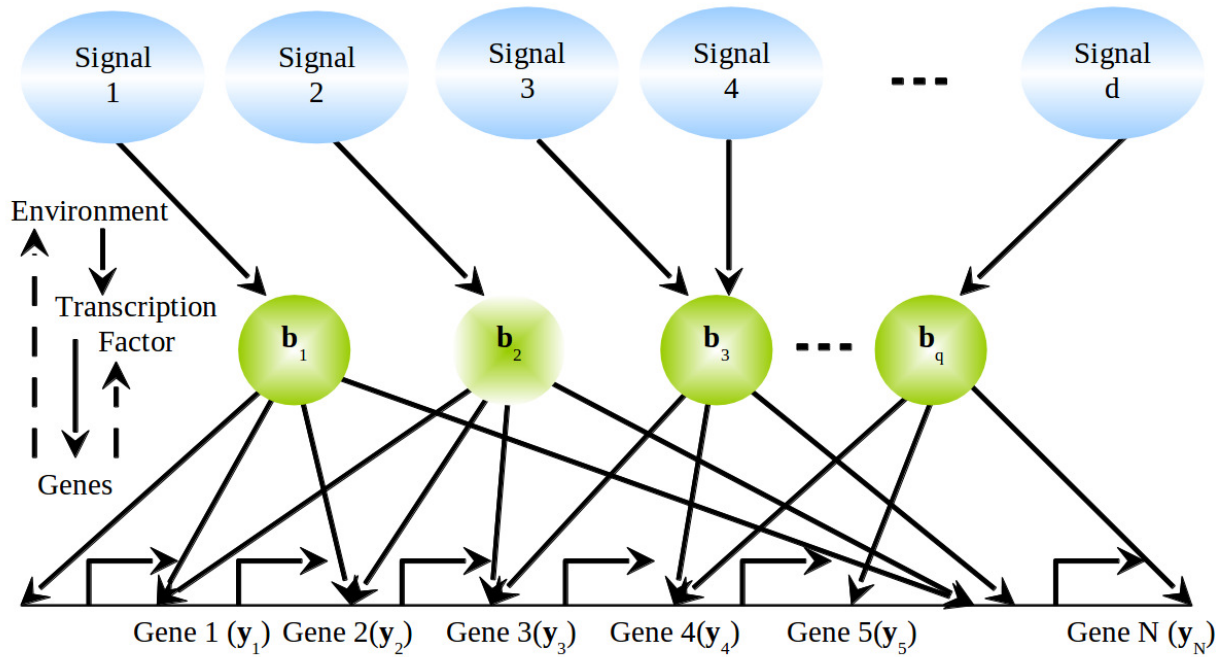


Figure: A 'cartoon' representation of mapping between environmental signal, transcription factor inside the cell the genes that they regulate.

C. elegans and Transcription Process

- Eukaryotic cells $\sim$ 1000
- Neurons $\sim$ 300
- Genes $\sim$ 15,139
- TF $\sim$ 940

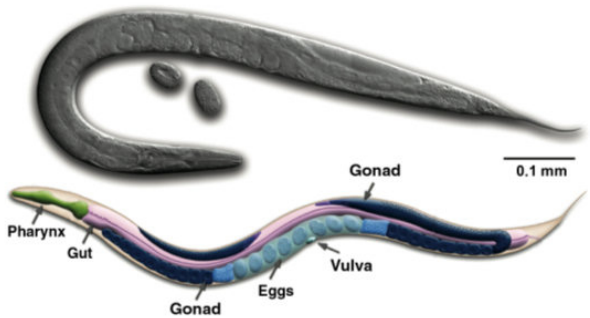
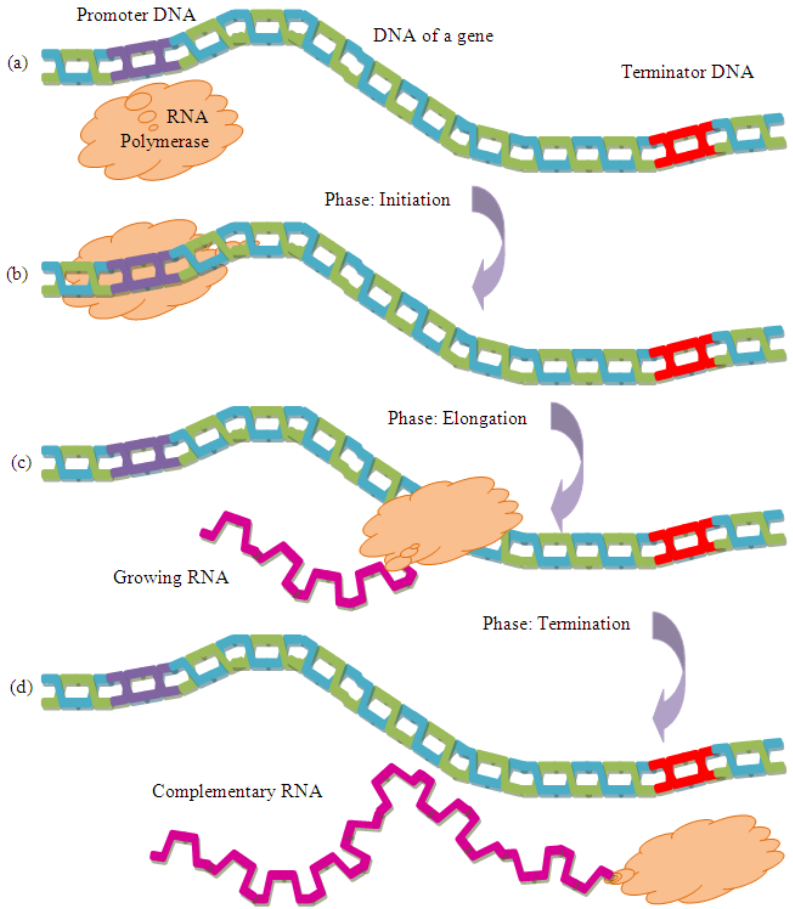


Figure: *C. elegans*



Key Research Questions (Initial stage)

- Using existing mechanistic model can we step forward to find out the transcription factor activities of multicellular eukaryote (from a unicellular microorganism)?

Probabilistic Model with Markov property

Our basic approach follows a dynamic model that extends the Linear Regression Model of [Liao, 2003] and Probabilistic Model of [Sanguinetti, 2006] to model the distribution of each transcription factor acting on each gene.

Let, Gene expression- $\mathbf{Y} \in \Re^{N \times d}$; Connectivity matrix- $\mathbf{X} \in \Re^{N \times q}$
TFAs can be obtained by regressing the gene expressions using the connectivity information, giving the following linear model-

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

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 $d \times q$ matrix $\mathbf{B_n}$ models the gene specific TFAs. TFA for gene n at any time $(t + 1)$ is-

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

for $t = 1, ..., (d - 1)$ and $\mathbf{b}_{n1} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Data Set: Gene Expression and Transcription Factors

Gene Expression level: The point estimate of the expression level and the uncertainty of the expression level were extracted from the micro array data using the tool puma [Pearson,R. et al. 2009].

Transcription Factors: *C. elegans* differential gene expression database (EDGEdb) [Barrasa, 2007] is the storage and retrieval of protein-DNA interactions. EDGEdb contains the sequence information of *C. elegans*'s 934 transcription factor and their DNA binding domains.

Data Set: Connectivity between Genes and TF

Evidence code of Wormnet [1] and [Lee, 2013] represent different (21) types of relations between genes. We choose the following four relations to get the connectivity between genes and transcription factors-

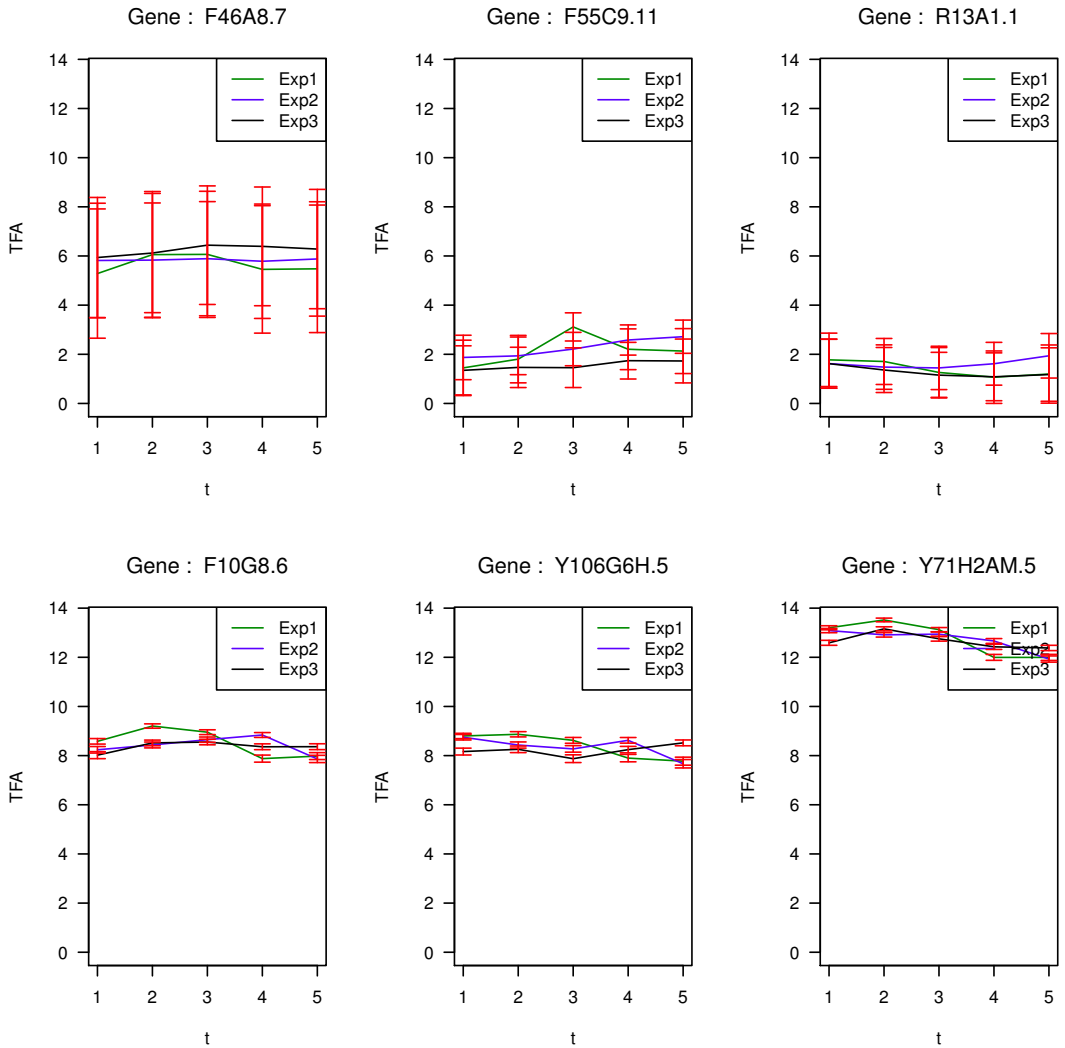
- High-throughput yeast 2-hybrid assays among worm genes
- Co-expression among worm genes
- High-throughput yeast 2-hybrid assays among human genes
- Literature curated human protein physical interactions

These relations leads to create a binary matrix of 0's and 1's.

$$x_{i,j} = \begin{cases} 1, & \text{if transcription factor } j \text{ can bind gene } i \\ 0, & \text{otherwise} \end{cases}$$

[1] http://www.functionalnet.org/wormnet/about.html

Gene Specific TFA of hmg-4[2] (Sequence: T20B12.8)



[2]High Mobility Group- 4 is required for locomotion and larval development

Ranking Differentially expressed gene expressions and correlation between them

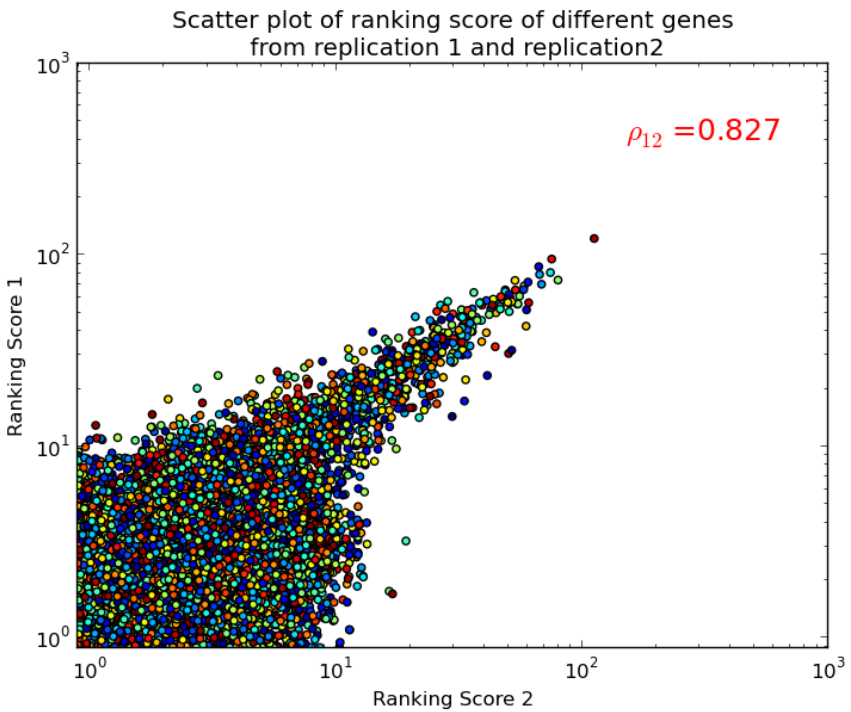
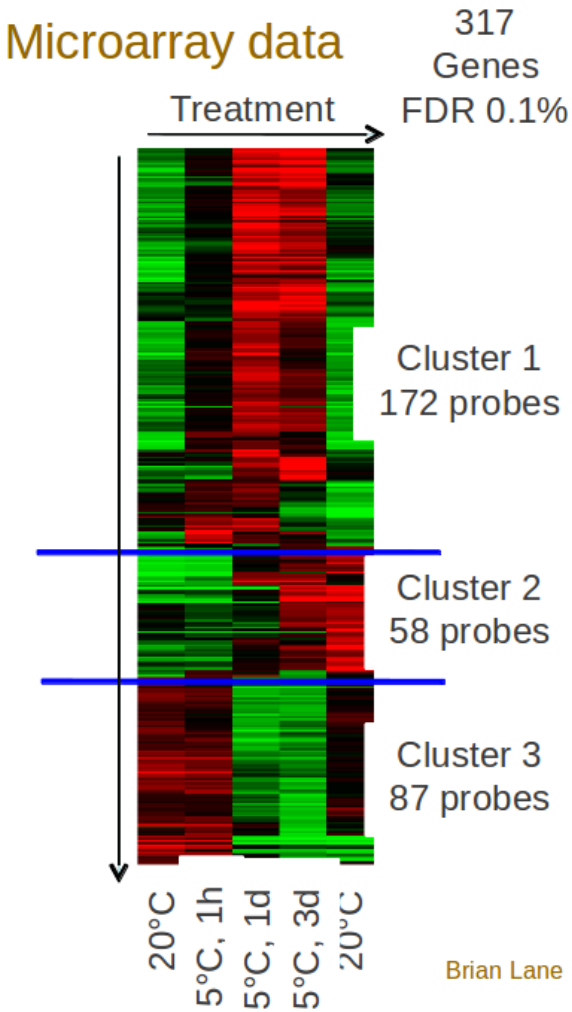


Figure: Scatter plot of ranking score of different genes from replication 1 and replication 2 and Pearson's correlation

Examples of Genes regulated by multiple TF

Gene Name/Sequence	Regulators activity
zip-8 (F23F12.9)	K10D2.3 $= 3.083311 \pm 0.239013$,
pitp-1 (M01F1.7)	F38A5.13 $= 0.714906468 \pm 0.2790247$, T24C4.7 $= 0.006439146 \pm 0.1516683$
ssp-33 (R08A2.3)	T19B10.11 $= 0.5793706446 \pm 1.1397625$ W02D7.6 $= 0.5793706446 \pm 1.1397625$ D1014.8 $= 0.0004742555 \pm 0.0438619$
C38D4.1	ZC513.6 $= 0.9681545 \pm 0.4944999$ T04C10.4 $= -0.4815358 \pm 1.0099097$ M01E11.5 $= -0.5252536 \pm 0.4952643$ Y11D7A.12 $= -6.3354545 \pm 1.4231987$

Active TF in different clusters of Microarray Data



Cluster Name	Active TF
1 - Chill upregulated	6
2 - Chill late upregulated	245
3 - Chill downregulated genes	128
4 - Others	203

Table: Number of Transcription Factor active in different Clusters

Key Research Questions (Secondary stage)

- Using existing mechanistic model can we step forward to find out the transcription factor activities of multicellular eukaryote (from a unicellular microorganism)?

  Yes; Our *R* based *chipDyno* tool is ready!

- Can we develop a robust data driven dynamic system for gene specific transcription factor activities using Gaussian process over a mechanistic model?

Gaussian Process (Definition)

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution[3]. It is a continuous stochastic process and defines probability distributions for functions.
If $f(\mathbf{x})$ is a real process, a Gaussian process is completely defined by its mean function and covariance function given by-

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{1}$$

The mean function $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x}')$ are defined as-

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \tag{2}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))], \tag{3}$$

where $\mathbb{E}$ represents the expected value.

[3] Rasmussen and Williams

Toward the GP model of TFA

In the earlier probabilistic approach gene specific TFAs was-

$$\mathbf{b}_{n(t+1)} \sim \mathcal{N}(\gamma \mathbf{b}_{nt} + (1-\gamma)\boldsymbol{\mu}, (1-\gamma^2)\boldsymbol{\Sigma}) \tag{4}$$

For a discrete time variable k and a one-dimensional process with the property with μ and s are scalar-

$$u_{k+1} \sim \mathcal{N}\left(\gamma u_k + (1-\gamma)\,\mu, (1-\gamma^2)s\right), \tag{5}$$

Assume that u_k's are actually values u_{t_k} from a continuous process $u(t)$ and $t_k = kDt$. A good candidate for this kind of model is the mean-reverting $Ornstein - Uhlenbeck$ model-

$$du = -\lambda\,(u-\mu)\,dt + q^{1/2}dB, \tag{6}$$

where λ is mean reversion rate, B is standard Brownian motion and q is volatility. This equation can now be solved on the time instants t_k and the result is a recursion

$$u(t_k) = au(t_{k-1}) + b\mu + w_{k-1};\, where,\, w_{k-1} \sim \mathcal{N}(0, c) \tag{7}$$

Toward the GP model of TFA(Cont..)

We can now match the coefficients:

$$a = \exp(-\lambda Dt) = \gamma \tag{8}$$

$$b = 1 - \exp(-\lambda Dt) = 1 - \gamma \tag{9}$$

$$c = [q/(2\lambda)][1 - \exp(-2\lambda Dt)] = (1 - \gamma^2)s \tag{10}$$

We can now recall that the (stationary) covariance function of the Ornstein-Uhlenbeck process we get-

$$k_u(t, t') = s\gamma^{|t-t'|} \tag{11}$$

The original vector valued $\mathbf{b}$, is separable, then the covariance function is obtained by formally replacing s with $\mathbf{\Sigma}$ everywhere-

$$\mathbf{K}_b(t, t') = \mathbf{\Sigma}\boldsymbol{\gamma}^{|t-t'|} \tag{12}$$

Thus is equivalent to considering the vector process-

$$\mathbf{db} = -\lambda(\mathbf{b} - \boldsymbol{\mu})\mathbf{dt} + Q^{1/2}\mathbf{dB}. \tag{13}$$

GP model for TFA

Gene expression is $\mathbf{Y} \in \Re^{n \times T}$ and the unobserved corresponding TFA is $\mathbf{F} \in \Re^{q \times T}$. Basic assumptions-

- TFA are in time series, they are likely to be temporally smooth.
- TF are potentially correlated with one another.

Correlation Between Transcription Factors: The correlation between different TF is covariance matrix, $\mathbf{\Sigma}$ with $q \times q$ dimensionality.

Temporal Smoothness: The TFs' activities is temporally smooth, and drawn from an underlying GP with covariance $\mathbf{K}_t$.

Intrinsic Coregionalization Model: We assume that the joint process across all q TFA 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.

$$\mathbf{K}_f = \mathbf{K}_t \otimes \mathbf{\Sigma} \tag{14}$$

GP model for TFA(cont..)

Assume that the jth gene's expression at the ith time point is given by-

$$\mathbf{y}_{i,j} = \mathbf{S}\mathbf{f}_{:,i} + \boldsymbol{\epsilon}_i \tag{15}$$

where, Gaussian noise- $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathbf{S}$ connectivity matrix. From standard properties of multivariate Gaussian distributions-

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{16}$$

$$\mathbf{K} = \mathbf{K}_t \otimes \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top + \sigma^2\mathbf{I}. \tag{17}$$

The likelihood of a multivariate Gaussian is:

$$L = -\frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\mathbf{y}^\top\mathbf{K}^{-1}\mathbf{y} = (L_q + L_u + Const) \tag{18}$$

$$L_q = -\frac{1}{2}\log|\mathbf{K}_t \otimes \boldsymbol{\Lambda}\mathbf{V}^\top\boldsymbol{\Sigma}\mathbf{V}\boldsymbol{\Lambda} + \sigma^2\mathbf{I}| - \frac{1}{2}\hat{\mathbf{y}}_q^\top\left[\mathbf{K}_t \otimes \boldsymbol{\Lambda}\mathbf{V}^\top\boldsymbol{\Sigma}\mathbf{V}\boldsymbol{\Lambda} + \sigma^2\mathbf{I}\right]^{-1}\hat{\mathbf{y}}_q \tag{19}$$

$$L_u = -\frac{T(n-q)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\hat{\mathbf{y}}_u^\top\hat{\mathbf{y}}_u \tag{20}$$

GP model for TFA: Making Prediction

Using Kronecker product, Rotation, SVD and considering noise we can rewrite the Equation 16 as:

$$\mathbf{y_q} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{t,t} \otimes \mathbf{\Lambda V}^T \mathbf{\Sigma V \Lambda} + \sigma^2 \mathbf{I}\right) \tag{21}$$

To make predictions about the test data we need the conditional distribution- $p(\mathbf{f}_\star|\mathbf{y})$. This conditional distribution is also Gaussian-

$$\mathbf{f}_\star \sim \mathcal{N}\left(\boldsymbol{\mu}_F, \mathbf{C}_F\right) \tag{22}$$

The mean of the posterior distribution of Equation 22 is:

$$\boldsymbol{\mu}_F = \mathbf{K}_{t_\star,t} \otimes \mathbf{\Sigma V \Lambda} \left[\mathbf{K}_{t,t} \otimes \mathbf{\Lambda V}^T \mathbf{\Sigma V \Lambda} + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{y}_q \tag{23}$$

Covariance of the posterior distribution of Equation 22 given by:

$$\begin{aligned} \mathbf{C}_F = \mathbf{K}_{t_\star,t_\star} \otimes \mathbf{\Sigma} - \mathbf{K}_{t_\star,t} \\ \otimes \mathbf{\Sigma V \Lambda} \left[\mathbf{K}_{t,t} \otimes \mathbf{\Lambda V}^T \mathbf{\Sigma V \Lambda} + \sigma^2 \mathbf{I}\right]^{-1} \left[\mathbf{K}_{t_\star,t} \otimes \mathbf{\Lambda V}^T \mathbf{\Sigma}\right] \end{aligned} \tag{24}$$

Covariance Matrix and TFA

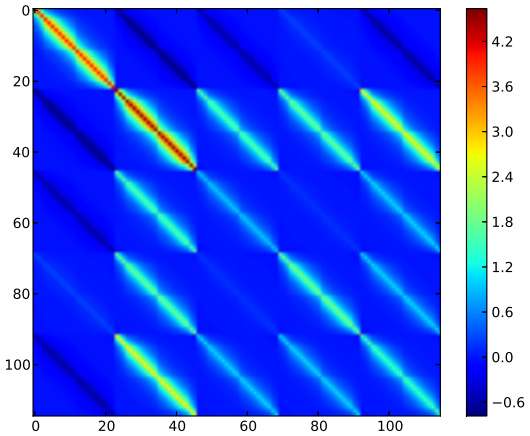


Figure: Kernel of Intrinsic Coregionalization model $\mathbf{K}_f$ considering 5 Transcription factors where covariance matrix $\mathbf{\Sigma}$ of (*Equation*17) was constructed using Ornstein-Uhlenbeck kernel and White kernel in additive form

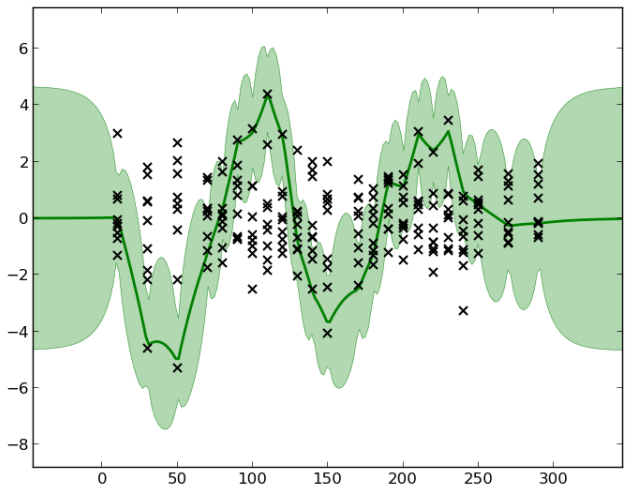


Figure: Transcription factor activity of ACE2 shaded area represents 95% confidence interval

Transcription factor activity of different TF

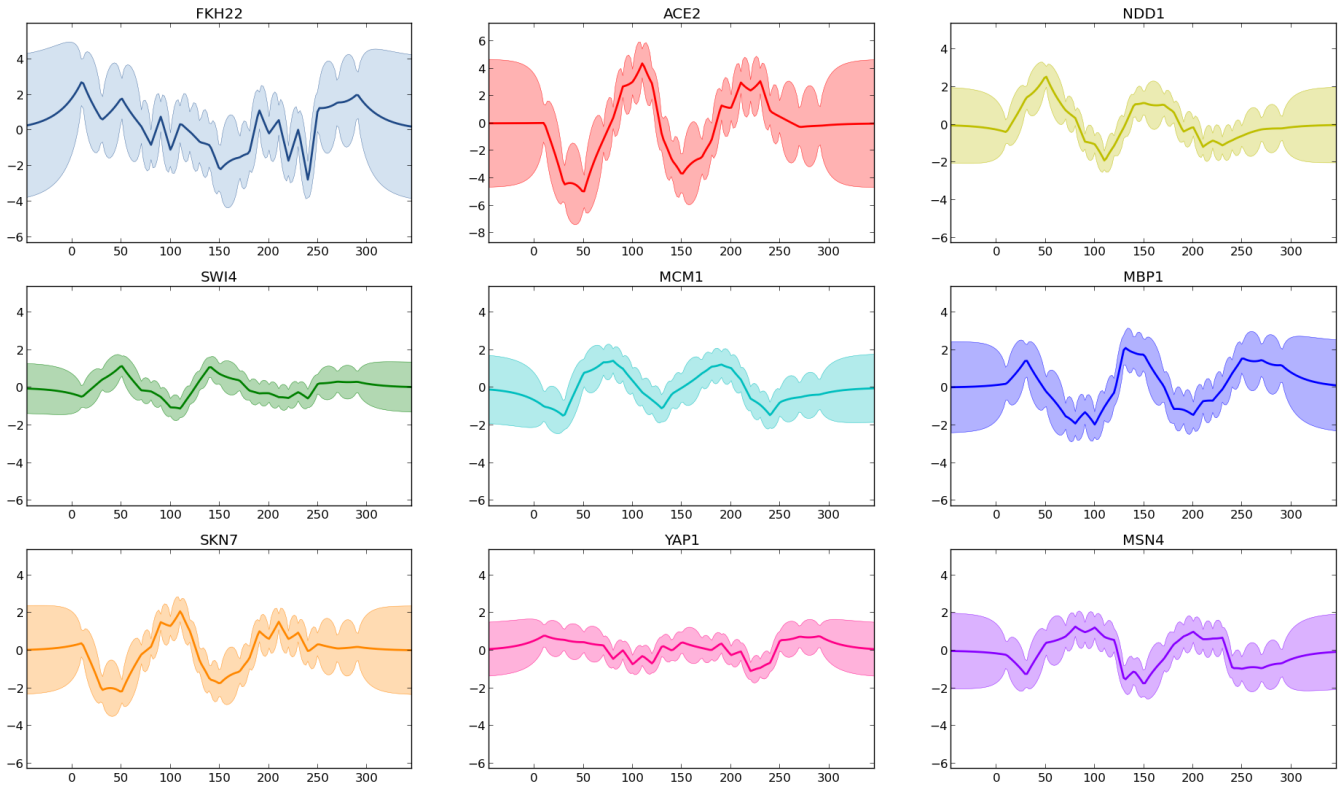


Figure: Transcription factor activity of different TF using Ornstein-Uhlenbeck kernel and White kernel in additive form

Conclusion

- Our *R* based tool *chipDyno* integrate the connectivity information between genes and transcription factors, and microarray data and infer the TFA.
- Earlier the model was developed for a unicellular microorganism (yeast) but we were successful to determine the gene specific TFA for *C. elegans*, a multicellular eukaryote.
- We were also successful to filter out the quiet genes from the differentially expressed genes for *C. elegans*.
- Our GP model will overcome the restriction of parametric model and temporal sampling of equally spaced time intervals.
- 38% of the protein-coding genes orthologs of human genes; The biological insight acquire from *C. elegans* may be directly applicable to more complex organism like human..

Future Work

- Transgenic mice can express human SOD1 mutation and replicates different histopathological and clinical features of Motor Neurone Disease (MND). But we didn't found any evidence which analyse MND considering the genetic background on different phenotype. We will use our GP model to infer the TFA on gene expression data obtained from different murine models and try to find out some insights.
- It is assumed that gene involved in the same biological process will be expressed with a similarity sharing underlying time series. Again it is very common to have multiple biological replicates of the gene expression. Just taking average of the replicates surely lead toward discarding insight. Using this GP based hierarchical clustering analysis we will find some robust clusters for the gene expression data of *C. elegans*.

Acknowledgements

- **Prof. Andrew Cossins**, Institute of Integrative Biology, University of Liverpool for the data set and valuable suggestions.
- **Dr. Simo Särkkä**, Academy Research Fellow, Department of Biomedical Engineering and Computational Science, Aalto University.
- **puma** : The Bioconductor package.
- **GPy** : The Gaussian process framework.
- **Ministry of Science and Information Technology, Bangladesh** for funding the scholarship.

References

Liao, J.C. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. (2003)

Network component analysis: reconstruction of regulatory signals in biological systems.

Proc Natl Acad Sci U S A. 2003 Dec 23;100(26):15522-7. Epub 2003 Dec 12.

Liu,X. et al. (2005)

A tractable probabilistic model for affymetrix probe-level analysis across multiple chips.

Bioinformatics, 21, 36373644.

Nachman,I. et al. (2004)

Inferring quantitative models of regulatory networks from expression data.

Bioinformatics, 20, i248i256.

Sanguinetti G, Rattray M., and Lawrence N.D. (2006)

A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription

Bioinformatics, Oxford University Press, Vol. 22 no. 14, pages 17531759

Motivations(Cont..)

- In molecular biology and genetics, a transcription factor is a protein that binds to specific DNA sequence.
- Transcription factors control the flow (or transcription) 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.
- It is difficult for a number of reasons: transcription factors expression levels are often low and noisy, and many transcription factors are post- transcriptionally regulated.
- So, from the expression levels of their target genes it is functional to infer the activity of the transcription factors.

Caenorhabditis elegans

Some basic features-

- Sydney Brenner (1927 -) established C. Elegans as a model organism to study genetics and cell development.
- Adults are 1mm long.
- They can be grown on agar plates with lawn of bacteria.
- They have a short generation time- 3 days from egg-laying to adulthood.

- Number of eukaryotic cells ~ 1000
- Number of neurons ~ 300
- Number of genes $\sim 15{,}139$ [a]
- Number of Transcription factors ~ 940 [b]

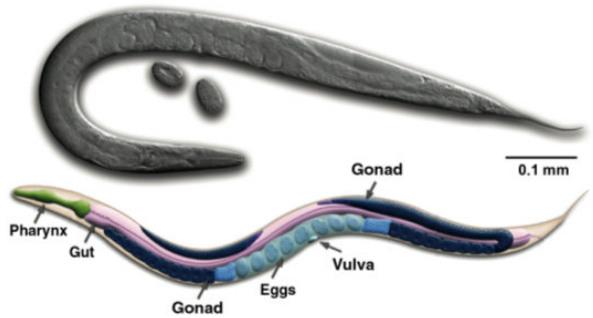


Figure: *C. Elegans*

[a] http://www.functionalnet.org/wormnet/about.html
[b] http://edgedb.umassmed.edu/TFFilesListingAction.do

Transcription

Some basic features-

1. The information in DNA is not directly converted into proteins, but must first be copied into RNA
2. Transcription produces genetic messages in the form of mRNA.
3. During transcription, a DNA sequence is read by an RNA polymerase.
4. Transcription produces a complementary, anti-parallel RNA strand

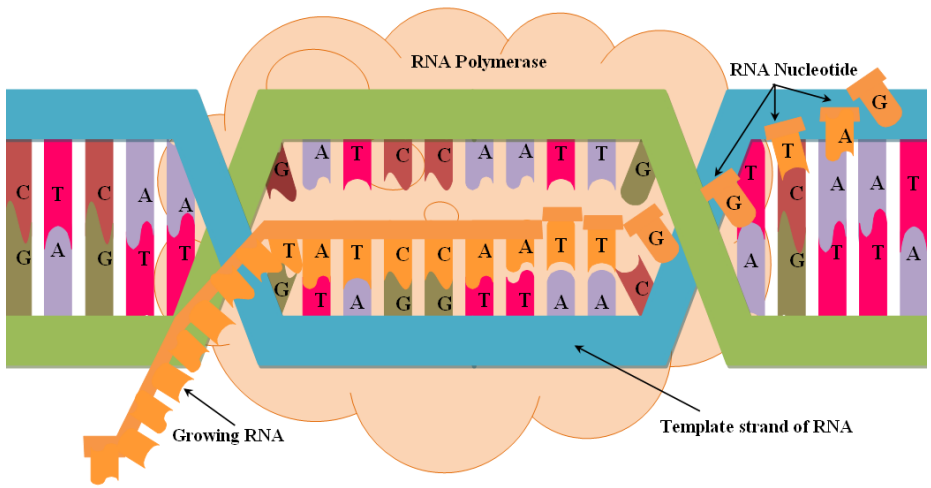


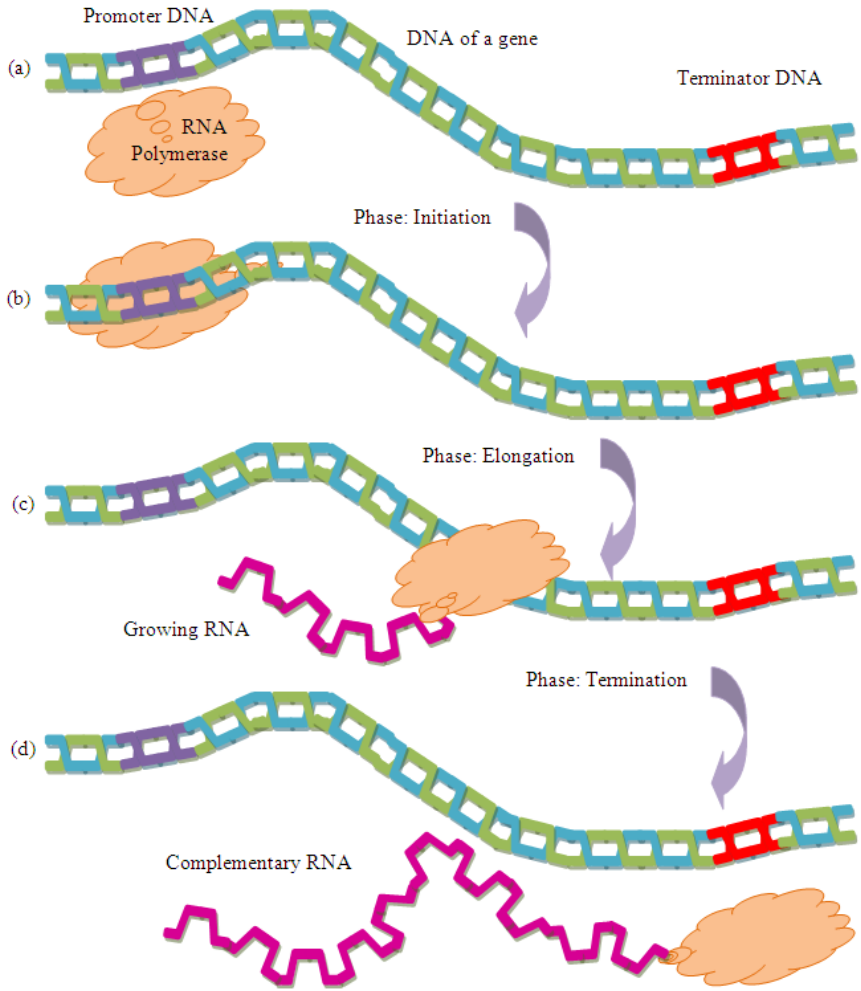
Figure: Transcription

Transcription Process

Initiation : Transcription process starts at the promoter region of a double-stranded DNA.

Elongation : A sequence specific DNA binding factors called transcription factors then unwind the DNA strand

Termination : Unwinding DNA's double helical form RNA polymerase reaches the terminator sequence. Here, RNA polymerase releases the mRNA polymer.



Probabilistic Model(cont..)

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 experimental data set consists of replication of condition i.e. all the $\mathbf{b}_{nt}$ assumed to be identical, so that-
  $\mathbf{b}_{n1} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{b}_{n(t+1)} \sim \mathcal{N}(\mathbf{b}_{nt}, \mathbf{0})$
- Second limiting case was when all the $\mathbf{b_{nt}}$ were assumed to be independent and identically distributed-

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

Probabilistic Model(cont..)

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

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

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-

$$p(\mathbf{Y}|\mathbf{B}, \mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{y_n}|\mathbf{B_n}, \mathbf{x_n})$$

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

$$p(\mathbf{b_n}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{b_n})p(\mathbf{b_n})}{p(\mathbf{Y})}$$

How the clusters were made?

Cluster 1 - Chill upregulated: 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

Cluster 2 - Chill late upregulated: 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

Cluster 3 - Chill downregulated genes: 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

Covariance function and Kernels Representation

Exponentiated Quadratic covariance function

$$K_{EQ}(r) = a^2 \exp\left(-\frac{r^2}{2l^2}\right) \quad (25)$$

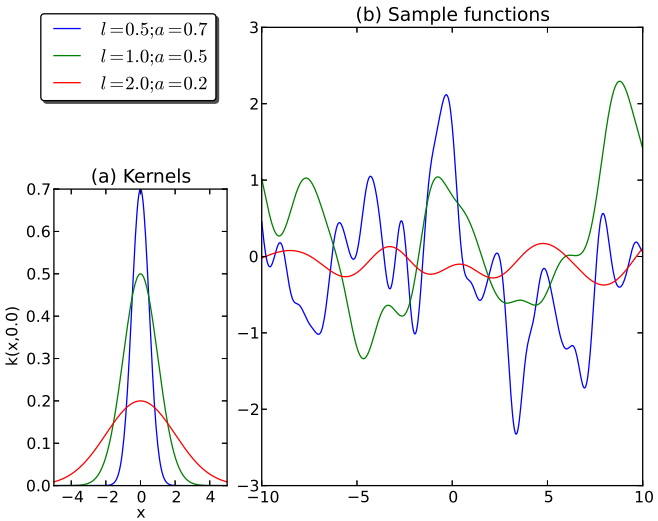


Figure: Exponentiated Quadratic kernel and sample functions

The Ornstein-Uhlenbeck covariance function

$$K_{\nu=1/2}(r) = \exp\left(-\frac{r}{l}\right) \quad (26)$$

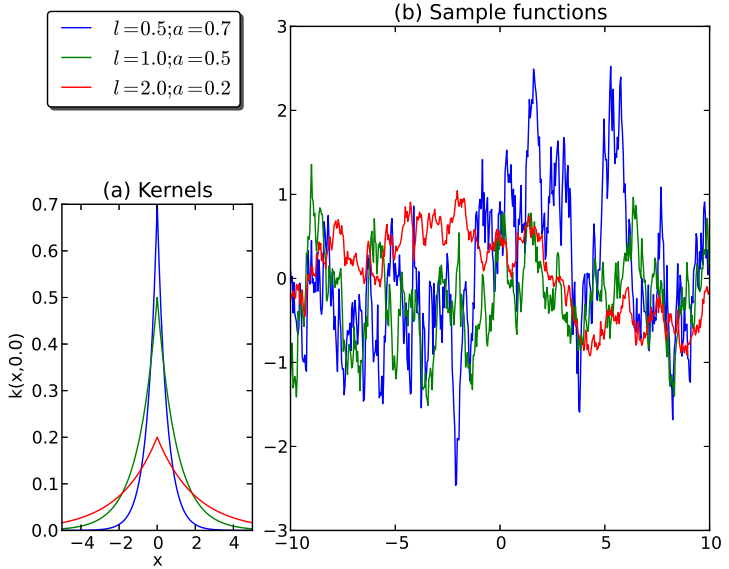


Figure: The OU kernel and random sample functions

Kernels Representation

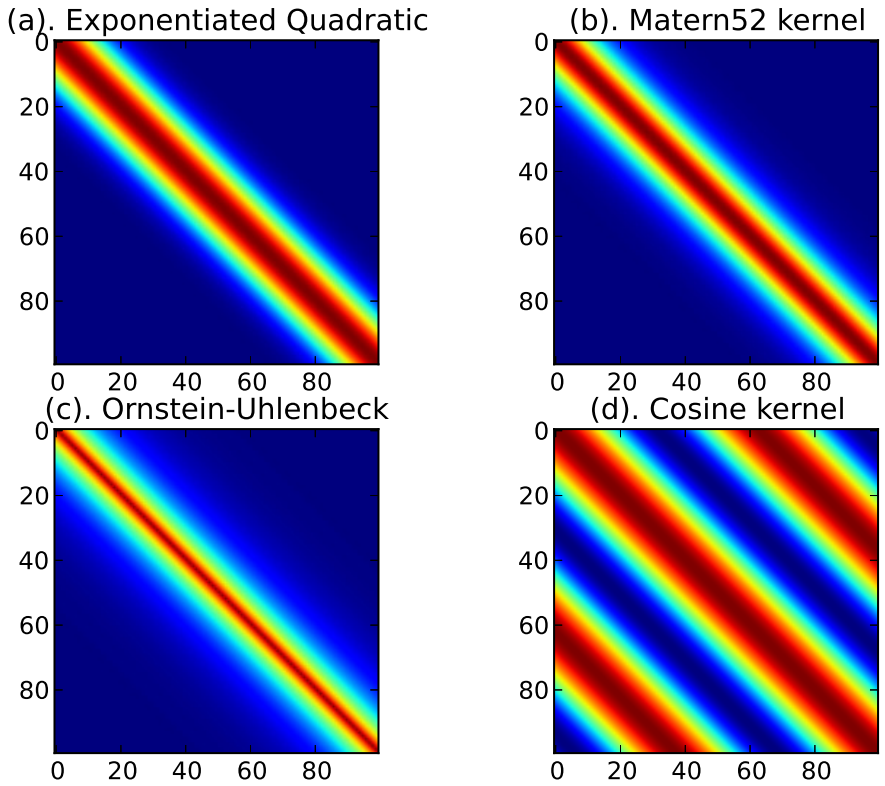


Figure: Representation of some basic kernels (a). Exponentiated Quadratic kernel, (b). Matérn52 kernel (c). Ornstein-Uhlenbeck kernel (d). Cosine kernel