A short report on: Dynamic Transcription Factor Activity of Caenorhabditis elegans

Muhammad Arifur Rahman
Supervisor: Neil D. Lawrence
Department of Computer Science
The University of Sheffield, Sheffield, UK

September 9, 2013

1 Motivation

- Caenorhabditis Elegans, a saprophytic nematode is an inhabitant of soil and leaf-litter and found in most of the parts of the world [Hope, 1999].
- Scientific reports on C. Elegans appeared in the literature for more than 100 years.
- After the genetics paper of Brenner's [Brenner, 1974] C. Elegans emerged as an important experimental model.
- Based on the bioinformatics approach used C. Elegans is 60-80% homologies with human genes. [Kaletta and Hengartner, 2006].
- Within a very short period of time (approximately 3days) it allows a large-scale (300+ of offspring) production [Hope, 1999].
- 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.
- Quantitative estimation of cellular processes 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.

2 Methodology

To determine the gene specific transcription factor activity of C. Elegans we have followed Sanguinetti's probabilistic dynamic model [Sanguinetti et al., 2006] for quantitative inference.

Let, Gene expression- $\mathbf{Y} \in \mathbb{R}^{\mathbf{N}\times\mathbf{d}}$ and connectivity matrix- $\mathbf{X} \in \mathbb{R}^{\mathbf{N}\times\mathbf{q}}$

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

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

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 $\mathbf{b}_{n(t+1)} \sim \mathcal{N}(\mathbf{b}_{nt}, \mathbf{0})$
- Second limiting case was when all the $\mathbf{b_{nt}}$ were assumed to be independent and identically distributed-
  $$\mathbf{b}_{nt} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
- [Sanguinetti et al., 2006] expected a realistic model of time series data to be somewhere in between this two extremes-
  $$\mathbf{b}_{n(t+1)} \sim \mathcal{N}(\gamma \mathbf{b}_{nt} + (1-\gamma)\boldsymbol{\mu}, (1-\gamma^2)\boldsymbol{\Sigma})$$
  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 Bayess Theorem-
  $$p(\mathbf{b_n}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{b_n})p(\mathbf{b_n})}{p(\mathbf{Y})}$$

3 Data

We have collected the data from three different sources-

Expression level: The point estimate of the expression label and the uncertainty of the expression level were extracted from the micro array data provided by Prof. Andrew Cossins using the tool puma [Pearson et al., 2009].

Transcription Factors: C.Elegans differential gene expression database (EDGEdb) [Barrasa et al., 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. It also represents the protein-DNA interactions between transcription and regulatory elements. At the initial point we have considered these 934 transcription factors for our experiment.

Connectivity between Genes and Transcription Factors: WormNet [Lee et al., 2013] is a modified Bayesian integration based probabilistic functional gene network of C. Elegans. Here the true functional linkage between genes was measured with its associated log-likelihood score. WormNet also tried to figure out the known functionality of C.Elegans and the links between different types of data from different types of organisms [Lee et al., 2008] and [Lee et al., 2010].

Evidence code of [Lee et al., 2013] represent 21 different types of relations between genes. Initially among these relations we choose the following four relations to get the connectivity between genes and transcription factors-

1. High-throughput yeast 2-hybrid assays among worm genes (CE-YH)
2. Co-expression among worm genes (CE-CX)
3. High-throughput yeast 2-hybrid assays among human genes (HS-YH)
4. Literature curated human protein physical interactions (HS-LC)

This relation leads to create a binary matrix of 0's and 1's. If there is any evidence of connectivity between any gene to a given transcription factor, then the connectivity matrix is indicated by 1. Otherwise, the value is 0. (Still we need to know about the exact relation).

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

Table 1: Genes regulated by multiple TF

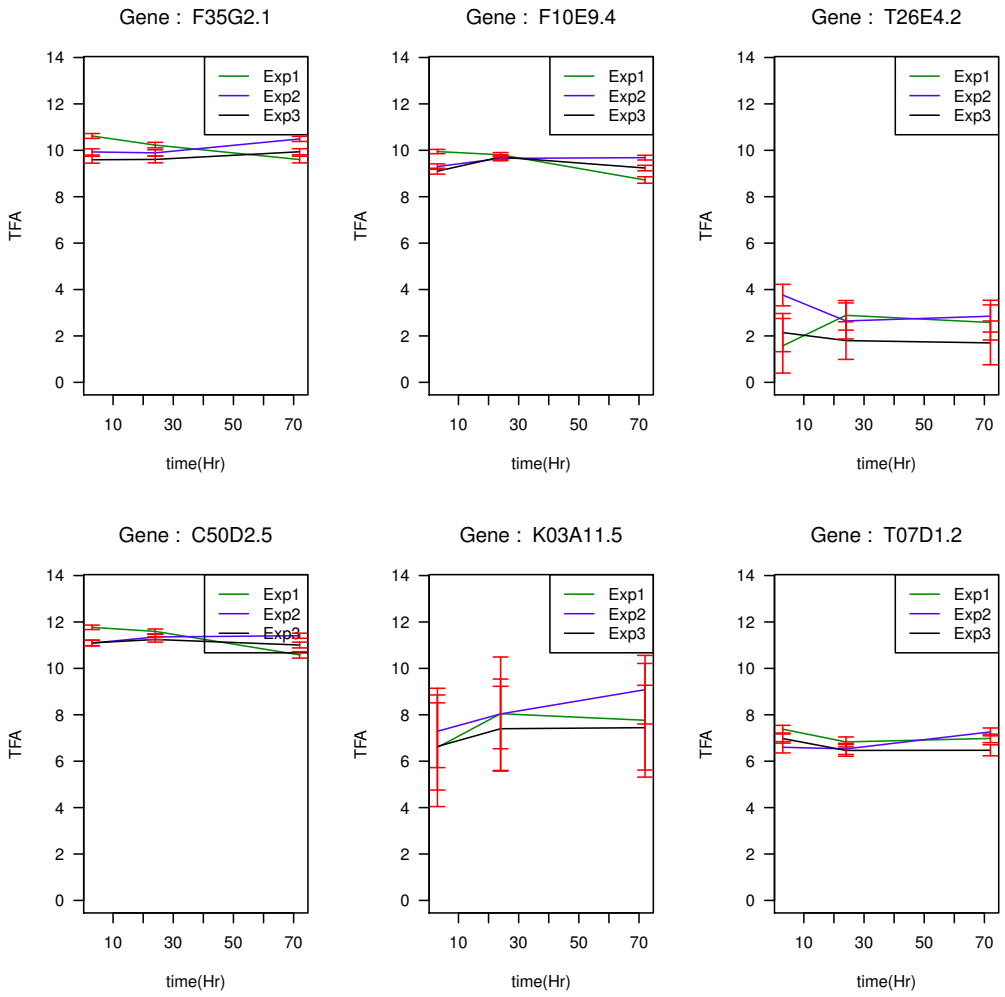

Figure 1: Gene specific TFA of T20B12.8

4 Results

Table 1 on page 4 describe some evidence of genes regulated by multiple transcription factors. Gene C44B12.5 is 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. In some cases, the error margin is quite high but for some case it is considerably low. This is a random example, if we are interested about any particular gene then it could

be possible to find their regulators and corresponding activity.

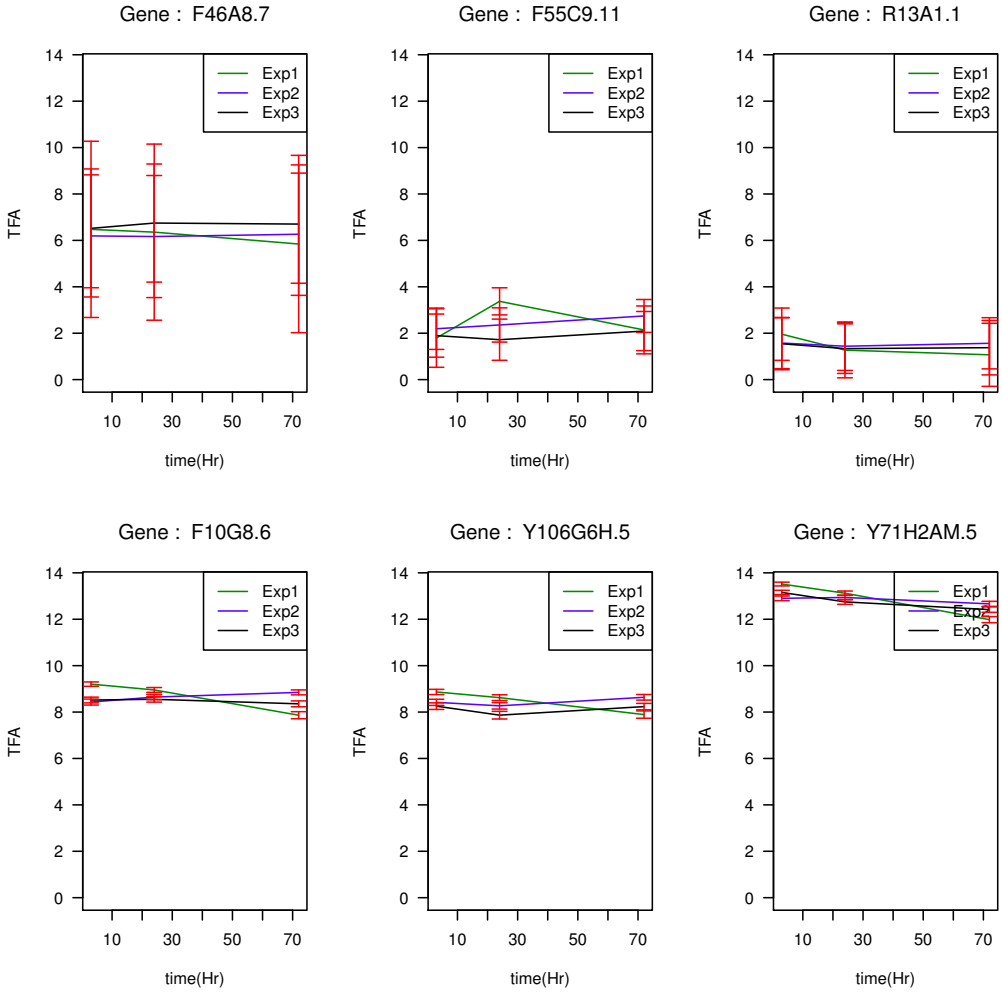

Figure 2: Gene specific TFA of F55D12.4

Figure 1 shows the transcription factor activity of T20B12.8 on F35G2.1, F10E9.4, T26E4.2, C50D2.5, K03A11.5 and T07D1.2. For this experiment the connectivity between genes and transcription factors were obtained from high-throughput yeast 2-hybrid assays among worm genes [Lee et al., 2013]. Here we have only 3 time points 3 hour, 24 hour and 72 hour and the experiments were done at 5 °C. All the experiments were replicated 3 times with

the same environmental conditions. The line graph shows the transcription factor activity with is associated error bars.

Figure 2 shows the transcription factor activity of F55D12.4 on F46A8.7 F55C9.11, R13A1.1, F10G8.6, Y106G6H.5 and Y71H2AM.5. Here the connectivity between genes and transcription factors were obtained from literature curated human protein physical interactions [Lee et al., 2013]. Here the experimental time points are 3 hour, 24 hour and 72 hour, and the experiments were done at 5 °C.

5 Conclusions

We have built a R based tool using the probabilistic approach of [Sanguinetti et al., 2006]. Using our newly developed 'chipDynoR' tool we have tried to analyze the gene specific dynamic transcription factor activity of C.Elegans. Still we are looking for expert comments, which might be helpful to conclude our findings.

6 Key Questions

- Is there any specific gene (or set of genes) which dynamic TFA could be interesting?
- Are we interested to figure out any specific regulator's activity for a certain gene?
- Is there any specific transcription factor which activity could be appealing?
- Which relation of [Lee et al., 2013] is the right one to obtain the connectivity between genes and transcription factors?
- We have only considered the common data from [Lee et al., 2013], [Barrasa et al., 2007] and given data (from Andrew Cossins's lab). Is it OK?
- For annotation and mapping between genes ID and ORF we have used bioconductor package [Carlson,]. Is it OK?
-

References

[Barrasa et al., 2007] Barrasa, M. I., Vaglio, P., Cavasino, F., Jacotot, L., and Walhout, A. J. (2007). Edgedb: a transcription factor-dna interaction database for the analysis of c. elegans differential gene expression. *BMC Genomics*. 8:21, doi:10.1186/1471-2164-8-21.

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

[Carlson,] Carlson, M. *celegans.db: Affymetrix celegans annotation data (chip celegans).* R package version 2.9.0.

[Hope, 1999] Hope, I. A. (1999). Background on caenorhabditis elegans. *C. elegans: A Practical Approach. NY: Oxford University Press*, 55(1):1–15.

[Kaletta and Hengartner, 2006] Kaletta, T. and Hengartner, M. O. (2006). Finding function in novel targets: C-elegans as a model organism. *Nat. Rev. Drug Discovery*, 5:387398.

[Lee et al., 2008] Lee, I., Lehner, B., Crombie, C., Wang, W., Fraser, A. G., and Marcotte, E. M. (2008). 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 et al., 2013] Lee, I., Lehner, B., Crombie, C., Wang, W., Fraser, A. G., and Marcotte, E. M. (2013). Probabilistic functional gene network of caenorhabditis elegans. `http://www.functionalnet.org/wormnet/about.html`. [Online; accessed 05-September-2013].

[Lee et al., 2010] Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A. G., and Marcotte, E. M. (2010). Predicting genetic modifier loci using functional gene networks. *Genome Research*, pages 1143–1153. 20(8).

[Pearson et al., 2009] Pearson, R., Liu, X., Sanguinetti, G., Milo, M., Lawrence, N., and Rattray, M. (2009). puma: a bioconductor package for propagating uncertainty in microarray analysis. *BMC Bioinformatics*. 10:211.

[Sanguinetti et al., 2006] 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*, 22(14):1753–1759.