Data Science and me

Guido Sanguinetti

ANC- School of Informatics, University of Edinburgh

October 7, 2014

Positional statement

- I was trained as a physicist/ mathematician
- Emphasis on Science in Data Science
- I'm unconvinced by statements that large-scale data gathering will eliminate the need for theory (i.e. hypothesis driven research), except perhaps in some engineering applications.
- However, science also produces vast amounts of data
- Statistical models and machine learning techniques are increasingly central in turning data into knowledge.

1. Data science and biology

2. Dynamics of transcriptional regulation (A. Ocone)

3. Epigenetics (G. Schweikert/ T. Mayo/ D. Benveniste)

Biology in a slide

- Living organisms contain a heritable blueprint of their biochemical capabilities in each cell, the **genome**
- The fundamental units in the genome, genes, are **transcribed** into an intermediate polymer (*mRNA*) and then **translated** into proteins
- Proteins are molecular machines that carry out most of the important functions of life
- All cells have the same genome; the differences are established by how the two key dynamical processes of transcription and translation are regulated

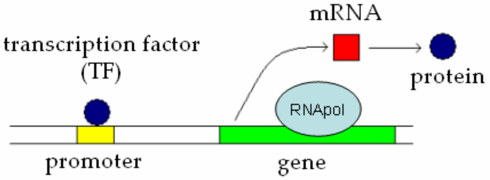
Systems Biology

- Since late 90s, biologists have been able to measure various biochemical components of cells in a high-throughput fashion
- Also, more precise microscopy-based measurements give time-resolved measurements at single cells
- Each measurement is a noisy readout of one facet of a (set of) complex biological processes
- Interpretable statistical models are (probably) the only way to integrate these disparate data in one coherent mechanistic picture
- Specifically, I work with probabilistic latent variable models (key difference: the latent variables and parameters have physical meanings)

Systems Biology

- Since late 90s, biologists have been able to measure various biochemical components of cells in a high-throughput fashion
- Also, more precise microscopy-based measurements give time-resolved measurements at single cells
- Each measurement is a noisy readout of one facet of a (set of) complex biological processes
- Interpretable statistical models are (probably) the only way to integrate these disparate data in one coherent mechanistic picture
- Specifically, I work with probabilistic latent variable models (key difference: the latent variables and parameters have physical meanings)

Transcription as a hybrid system

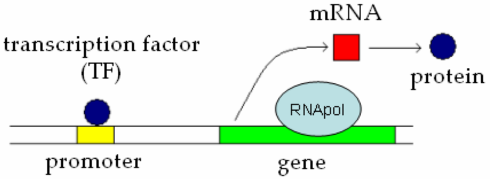


- Given the promoter state, the proteins obey the following dynamical model of transcription (linear SDE)

$$dx(t) = (A\mu(t) + b - \lambda x(t))\, dt + \sigma dW \qquad (1)$$

- The promoter state μ is a stochastic switching process *telegraph process*
- If we observe the proteins at discrete times, can we infer the system trajectory and parameters? Can we generalise to complex networks?

Transcription as a hybrid system



- Given the promoter state, the proteins obey the following dynamical model of transcription (linear SDE)

$$dx(t) = (A\mu(t) + b - \lambda x(t))\, dt + \sigma dW \tag{1}$$

- The promoter state μ is a stochastic switching process *telegraph process*
- If we observe the proteins at discrete times, can we infer the system trajectory and parameters? Can we generalise to complex networks?

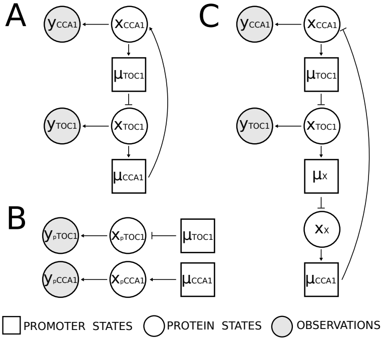
Technical ingredients

- Latent variable models: the trajectory of the system (in continuous time) is a latent variable, as well as the kinetic parameters
- Approximate Bayesian inference: we compute a variational approximation to the intractable joint posterior using optimisation of a Kullback-Leibler divergence
- Dynamical systems: we developed a novel algorithm for solving the optimisation problem based on the classic forward-backward recursions

Application: Ostreococcus tauri's circadian clock

- Circadian clocks are genetic circuits that enable organisms to adapt to light/ darkness cycles
- Andrew Millar at SBS pre-eminent expert on circadian clocks in plants
- O.tauri is a picoalga described as smallest free living eukaryote; possible model for a minimal clock (two genes)?
- Data consists of luciferase time series with different photoperiods from reporter constructs

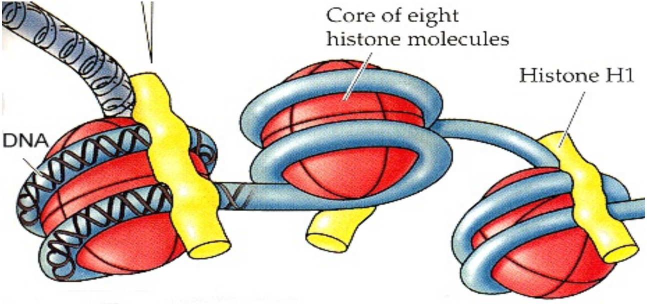
Comparison of two models



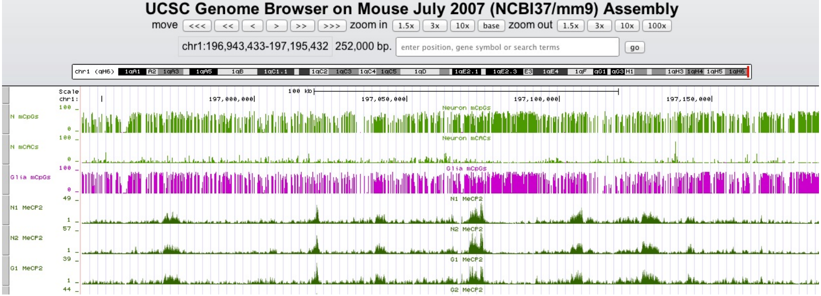
Different models were estimated from the available data showing that a three-gene network is in fact better supported by evidence than the postulated minimal two-gene network.

Epigenetics

Genetics and transcription cannot be all; spatial organisation of chromosomes plays a role. This is determined by chemical modifications to DNA and histones.



Epigenetics: what the data looks like



UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly

chr1:196,943,433-197,195,432 252,000 bp.

Each row is a tiny fraction of a next-generation sequencing experiment's data. Each row ≥1GB of data. How do we determine relationships between the rows?

Lines of attack

- Identifying statistically significant differences between the rows is already difficult: some success adapting a kernel method, *Maximum Mean Discrepancy* (Gretton et al 2008), to sequencing data (Schweikert et al, BMC Genomics 2013, Mayo et al, under review)
- Predictive models are useful: e.g., given a hypothesis that the green rows are mechanistically determined by the pink rows, we should be able to train a fairly accurate regression model
- Recent success in predicting histone modifications from binding of transcription factor proteins (Benveniste et al, PNAS 2014)
- Technical challenges: large size of the data sets, large number of covariates, inhomogeneities along chromosomes (latent variables?)