Modelling and Analysis in Bioinformatics: Inferring gene regulation

Antti Honkela

11 October 2016

Last week: Simulation of gene expression

Let X be the system state, i.e. numbers of each molecule species
Reactions:

$$\alpha_{j1} X_{i_{j1}} + \cdots + \alpha_{jn} X_{i_{jn}} \to \beta_{j1} X_{o_{j1}} + \cdots + \beta_{jn} X_{o_{jn}}, \quad j = 1, \ldots, M$$

Let c be the known reaction rates.

Last week: Simulation of gene expression

Let X be the system state, i.e. numbers of each molecule species
Reactions:

$$\alpha_{j1}X_{i_{j1}} + \cdots + \alpha_{jn}X_{i_{jn}} \to \beta_{j1}X_{o_{j1}} + \cdots + \beta_{jn}X_{o_{jn}}, \quad j = 1, \ldots, M$$

Let c be the known reaction rates.

- Stochastic simulation algorithm (SSA; Gillespie, 1977)

Last week: Simulation of gene expression

Let X be the system state, i.e. numbers of each molecule species
Reactions:

$$\alpha_{j1} X_{i_{j1}} + \cdots + \alpha_{jn} X_{i_{jn}} \to \beta_{j1} X_{o_{j1}} + \cdots + \beta_{jn} X_{o_{jn}}, \quad j = 1, \ldots, M$$

Let c be the known reaction rates.

- Stochastic simulation algorithm (SSA; Gillespie, 1977)
  1. Sample time until next reaction (τ)
  2. Sample which reaction (depends on rates), simulate it

Last week: Simulation of gene expression

Let X be the system state, i.e. numbers of each molecule species
Reactions:

$$\alpha_{j1}X_{i_{j1}} + \cdots + \alpha_{jn}X_{i_{jn}} \to \beta_{j1}X_{o_{j1}} + \cdots + \beta_{jn}X_{o_{jn}}, \quad j = 1, \ldots, M$$

Let c be the known reaction rates.

- Stochastic simulation algorithm (SSA; Gillespie, 1977)
  1. Sample time until next reaction (τ)
  2. Sample which reaction (depends on rates), simulate it
- Deterministic differential equation model

Last week: Simulation of gene expression

Let X be the system state, i.e. numbers of each molecule species
Reactions:

$$\alpha_{j1} X_{i_{j1}} + \cdots + \alpha_{jn} X_{i_{jn}} \to \beta_{j1} X_{o_{j1}} + \cdots + \beta_{jn} X_{o_{jn}}, \quad j = 1, \ldots, M$$

Let c be the known reaction rates.

- Stochastic simulation algorithm (SSA; Gillespie, 1977)
  1. Sample time until next reaction (τ)
  2. Sample which reaction (depends on rates), simulate it
- Deterministic differential equation model
- Stochastic differential equation model

Warmup for this week

- How to infer gene regulatory relationships?
  i.e.
- Given a set of observed variables, how to find ones that are related?

Warmup for this week

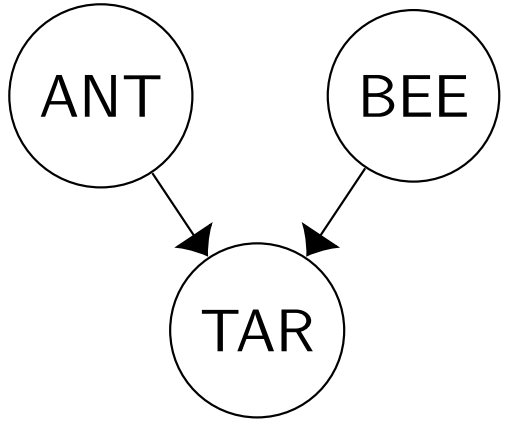
- How to infer gene regulatory relationships?
  i.e.
- Given a set of observed variables, how to find ones that are related?
- ... that are *causally* related?

Learning goals for this week

- To understand reasons for difference of correlation and causality
- To recognise basic regulatory network inference methods
- To apply simple methods for network inference from data

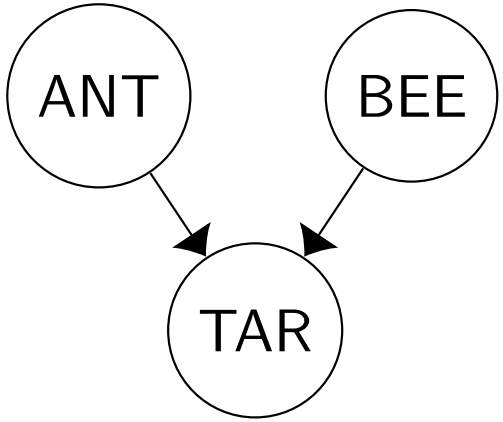
What is a gene regulatory network?

- How should the arcs be interpreted?



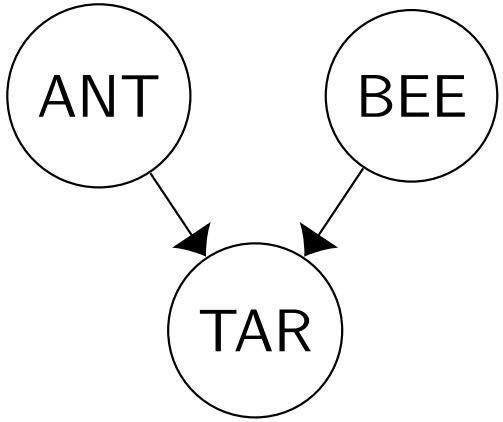
What is a gene regulatory network?

- How should the arcs be interpreted?
  - Interaction active somewhere?



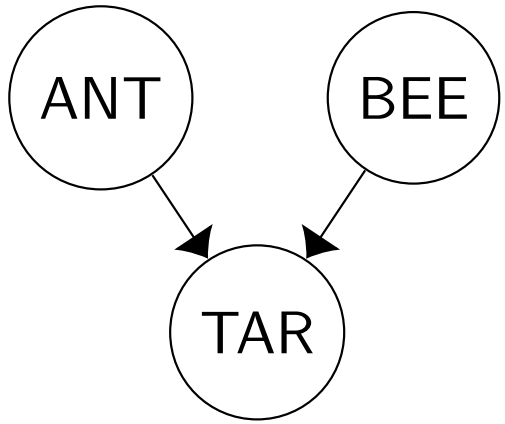
What is a gene regulatory network?

- How should the arcs be interpreted?
  - Interaction active somewhere?
  - Interaction active at a given condition?



What is a gene regulatory network?

- How should the arcs be interpreted?
  - Interaction active somewhere?
  - Interaction active at a given condition?
  - Rate-limiting interaction at a given condition?



Outline

Motivation

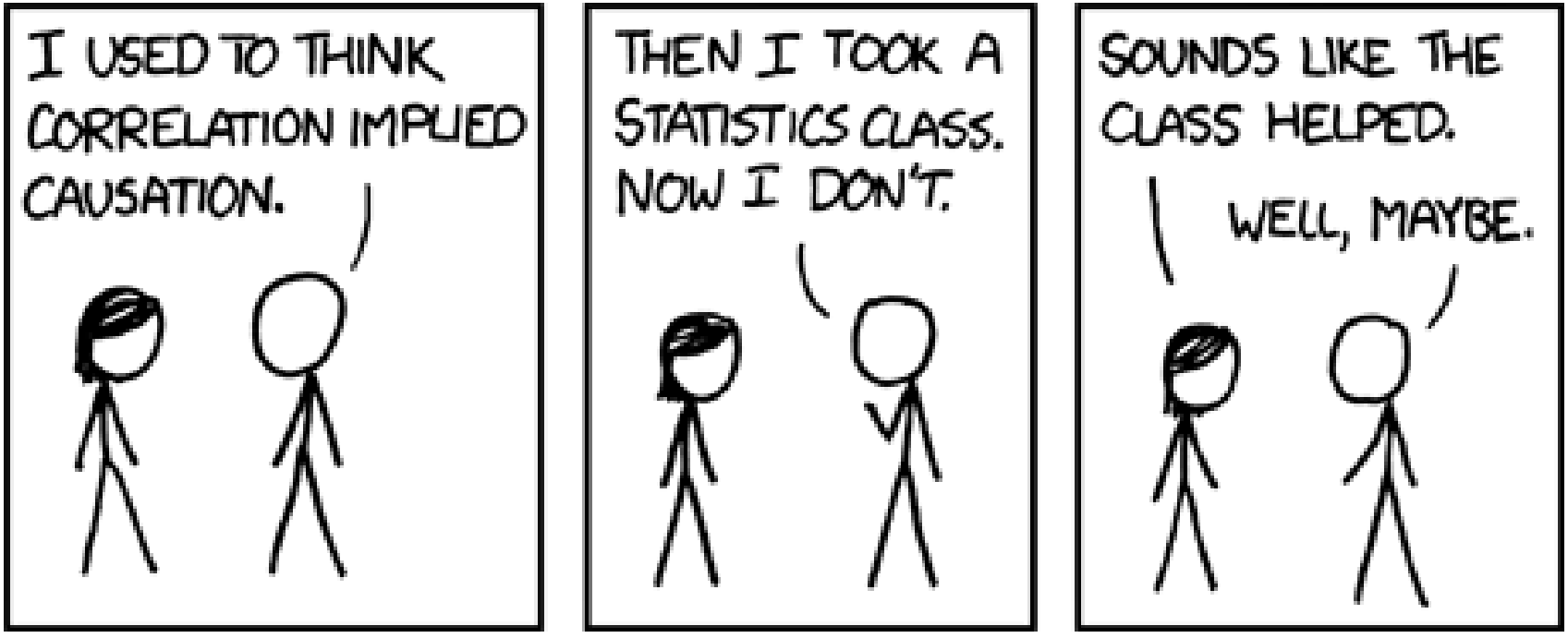
Causal inference

Experimental methods for regulatory network inference

Computational methods for regulatory network inference

Conclusion and next steps

Outline



Correlation and causation

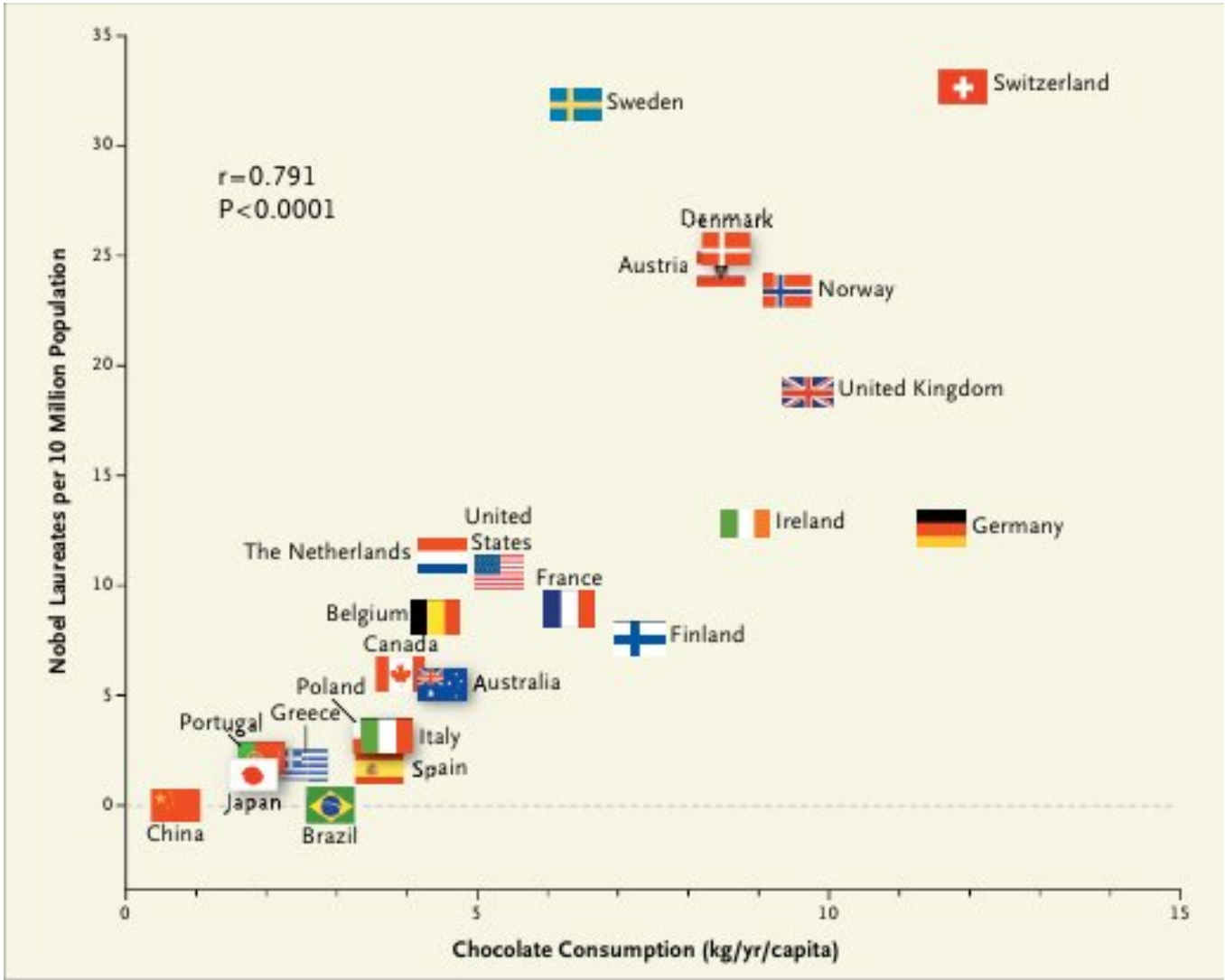


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Image from Messerli (2012), doi:10.1056/NEJMon1211064

Correlation and causation examples

- Drownings and ice cream sales in the summer
- Number of pirates and global average temperature
- Diet, health and lifestyle
- ...

Linear dependence

- As we saw, *correlation does not imply causation*.
- Assume two random variables X and Y related by

$$X = a \cdot Y + \epsilon$$

- We can equivalently solve

$$Y = (X - \epsilon)/a$$

- Similar reasoning also applies with non-linear dependencies
- In general not possible to tell which of two variables causes which

Direct and indirect links

- In general, we would like to distinguish between direct and indirect links
- Assume three random variables X, Y and Z related by

$$Y = b \cdot Z + \epsilon_Y$$
$$X = a \cdot Y + \epsilon_X$$

- X will also depend on Z:

$$X = a \cdot b \cdot Z + a \cdot \epsilon_Y + \epsilon_X$$

- Similar reasoning also applies with non-linear dependencies
- Additional assumptions needed to separate direct and indirect links

How to infer causality

- Interventions
- Randomisation
  - Double blinding
- Clever experimental design

Addiction. 2015 Sep 11. doi: 10.1111/add.13152. [Epub ahead of print]

Drinking and mortality: long-term follow-up of drinking-discordant twin pairs.

Sipilä P[1], Rose RJ[2], Kaprio J[1,3,4].

Author information

Abstract

AIMS: To determine if associations of alcohol consumption with all-cause mortality replicate in discordant monozygotic twin comparisons that control for familial and genetic confounds.

DESIGN: A 30-year prospective follow-up.

SETTINGS: Population-based Older Finnish Twin Cohort.

PARTICIPANTS: Same-sex twins, aged 24-60 years at the end of 1981, without overt co-morbidities, completed questionnaires in 1975 and 1981 with response rates of 89% and 84%. 15,607 twins were available for mortality follow-up from the date of returned 1981 questionnaires to December 31, 2011. 14,787 twins with complete information were analysed.

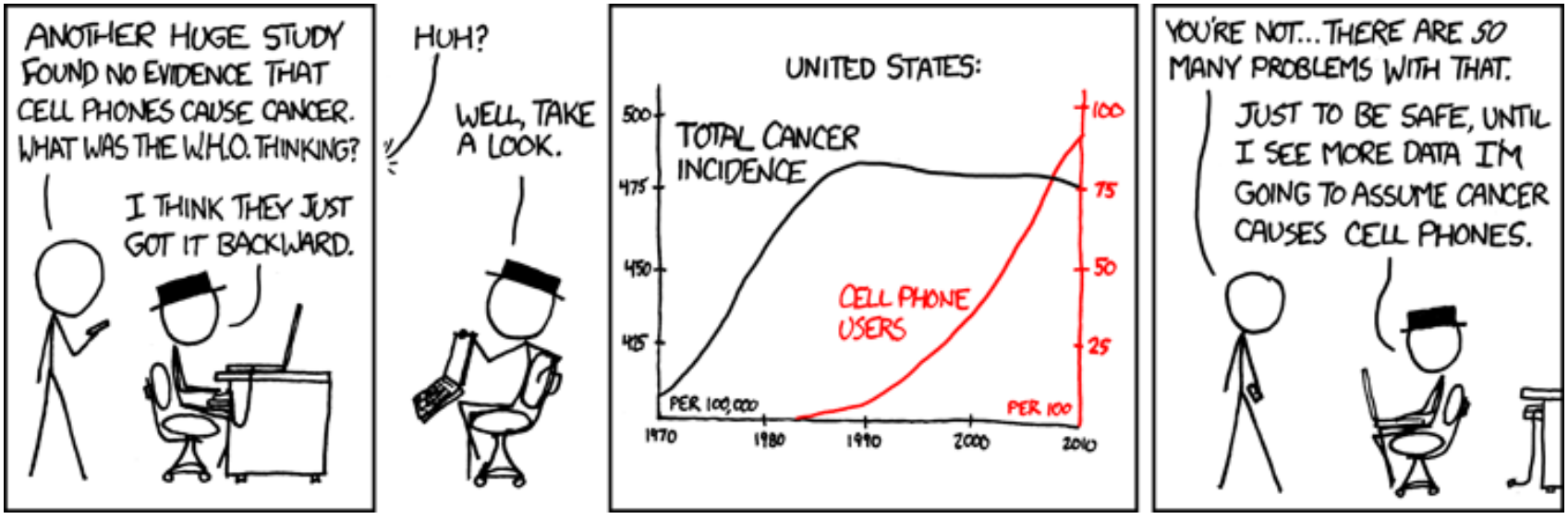
MEASUREMENTS: Self-reported monthly alcohol consumption, heavy drinking occasions (HDO), and alcohol-induced blackouts. Adjustments for age, gender, marital and smoking status, physical activity, obesity, education and social class.

FINDINGS: Among twins as individuals, high levels of monthly alcohol consumption (≥259 grams/month) associated with earlier mortality (HR = 1.63, 95% confidence interval (CI) = 1.47-1.81). That association replicated in comparisons of all informatively drinking-discordant twin pairs (HR = 1.91, 95% CI = 1.49-2.45) and within discordant monozygotic (MZ) twin pairs (HR = 2.24, 95% CI = 1.31-3.85), with comparable effect size. Smaller samples of MZ twins discordant for HDO and blackouts limited power; a significant association with mortality was found for multiple blackouts (HR = 2.82, 95% CI = 1.30-6.08), but not for HDO.

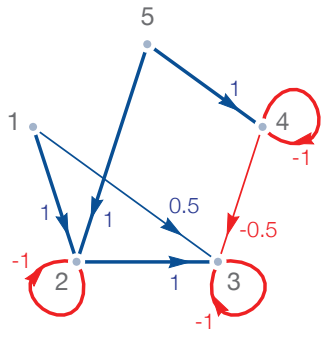
CONCLUSIONS: The associations of high levels of monthly alcohol consumption and alcohol-induced blackouts with increased all-cause mortality among Finnish twins cannot be explained by familial or genetic confounds; the explanation appears to be causal.

KEYWORDS: Alcohol drinking; alcoholic intoxication; binge drinking; causality; confounding factors; follow-up studies; mortality; twins

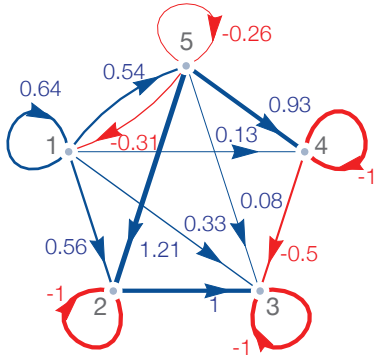
PMID: 26359785 [PubMed - as supplied by publisher]



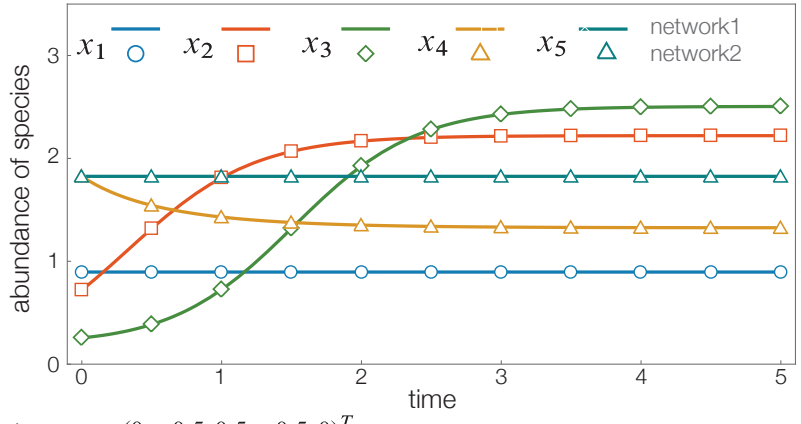
Two networks producing identical output



network 1



network 2



reaction rates: $r = (0, -0.5, 0.5, -0.5, 0)^T$

initial conditions: $x(0) = (0.895349, 0.72093, 0.255814, 1.82558, 1.82558)^T$

Image from Angulo *et al.* (2015), arXiv:1508.03559

What can help

- Combining data sets from different modalities
- Diverse data, perturbations
- Prior information, e.g. sparsity
- . . . ?

Outline

Typical data types

1. TF knockouts
2. Expression data
3. TF binding data
4. (Chromatin accessibility data)
5. (Chromatin 3D structure data)

TF = transcription factor (regulator gene)
chromatin = DNA packaging

TF knockout data

- Experimental intervention to disable a gene
- Measure gene expression afterwards
- Challenge: dramatic perturbation
  - Example fruit fly genes: *eyeless*, *tinman*
  - Are we still studying the same network?

Expression data

- Data under diverse conditions
- Time series are very helpful
  - Otherwise difficult to identify dynamical parameters that may confound the network
- Experimental design—measurements are expensive!

TF binding data

- Useful for establishing a mechanism
- But:
  - Not all regulators bind directly to DNA (could bind via other TFs)
  - How to map enhancers to genes?

Outline

Regulatory network inference methods

- Modelling various data sets
  - Modelling knockouts
  - Dynamical models of time series
  - Prior knowledge from TF binding
- Regression methods
- Correlation and mutual information based approaches

Modelling knockouts

- Typical workflow: compare the expression of a gene before and after knockout or other perturbation
- Can also check the sign of change
- Combine data from multiple experiments to work out the network

Outline

Linear dynamical systems

$$\frac{d\mathbf{X}(t)}{dt} = \mathbf{A}\mathbf{X}(t)$$

- Here $\mathbf{A}$ is a matrix of regulatory links
- Why:
  - Simple representation
  - Efficient inference algorithms
- Why not:
  - Unrealistic

Two regulators

$$0 + 0 = 0$$
$$1 + 0 = 1$$
$$0 + 1 = 1$$
$$1 + 1 = 2$$

Limit behaviour

- Consider a 1D linear differential equation:

$$\frac{dx(t)}{dt} = ax(t)$$

- This has a solution (check!):

$$x(t) = Ce^{at}$$

Limit behaviour

- Consider a 1D linear differential equation:

$$\frac{dx(t)}{dt} = ax(t)$$

- This has a solution (check!):

$$x(t) = Ce^{at}$$

- Assume $x(0) > 0$. As $t \to \infty$, either

$$\begin{cases} x(t) \to \infty & \text{If } a > 0 \\ x(t) \to 0 & \text{If } a < 0 \end{cases}$$

Limit behaviour

$$\frac{d\mathbf{X}(t)}{dt} = \mathbf{A}\mathbf{X}(t)$$

- In higher dimensions possibilities are only slightly more complex:
  - $|\mathbf{X}(t)| \to \infty$
  - $|\mathbf{X}(t)| \to 0$
  - $\mathbf{X}(t)$ approaches a harmonic oscillator (sine curve)
- The behaviour depends on the eigenvalues of $\mathbf{A}$

Non-linear dynamical systems

- More complex models
- Why:
  - Realistic model
- Why not:
  - More difficult to learn: more choices, more parameters
  - Less efficient algorithms

Granger causality

- Assume a linear discrete-time dynamical model between two variables $x_1(t)$ and $x_2(t)$:

$$x_1(t) = \sum_{j=1}^{p} a_{11,j} x_1(t-j) + \sum_{j=1}^{p} a_{12,j} x_2(t-j)$$

$$x_2(t) = \sum_{j=1}^{p} a_{22,j} x_2(t-j) + \sum_{j=1}^{p} a_{21,j} x_1(t-j)$$

- If a model with $a_{12} \neq 0$ explains x_1 better, it is said that x_2 Granger-causes x_1.
- This provides evidence that x_2 may cause x_1
- But: Granger causality $\neq$ causality

Outline

Correlation and covariance matrices

- For two random variables X and Y we define

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] \quad \text{(covariance)}$$

$$\rho_{XY} = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Cov}(X, X)\mathrm{Cov}(Y, Y)}} \quad \text{(correlation)}$$

- For more variables we can collect these to a matrix such as

$$\mathrm{Cov}[(X_1, X_2, X_3)] =$$

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \mathrm{Cov}(X_1, X_3) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \mathrm{Cov}(X_2, X_3) \\ \mathrm{Cov}(X_3, X_1) & \mathrm{Cov}(X_3, X_2) & \mathrm{Cov}(X_3, X_3) \end{pmatrix}$$

Gaussian Markov random field

- Undirected graphical model, edges denote dependence
- Gaussian marginals $\Rightarrow$ multivariate Gaussian joint distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Theorem: $\{i, j\} \notin E \;\Rightarrow\; (\boldsymbol{\Sigma}^{-1})_{i,j} = 0$
- In words: the *precision matrix* $\boldsymbol{\Sigma}^{-1}$ is sparse with non-zero elements corresponding exactly to edges in the dependency graph

Graphical lasso

- Previous sparsity property suggests model structure learning algorithms that promote such sparsity
- Given observations X, we aim to estimate the precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ by minimising

$$-\log p(X|\mathbf{\Theta}) + \alpha \cdot \mathrm{pen}(\mathbf{\Theta}),$$

  where the first term is negative log-likelihood and the second term is penalty that encourages sparsity
- Ideally $\mathrm{pen}(\mathbf{\Theta}) = \#\{(i,j)|\theta_{ij} \neq 0\}$ but computationally difficult
- *Graphical Lasso*: $\mathrm{pen}(\mathbf{\Theta}) = \sum_{i,j} |\theta_{ij}|$
  - Efficient convex optimisation + sparsity
  - May not be suitable for gene expression data

Limitations of static interaction models

- Self-interactions?
- Loops resolved through time?
- Directionality difficult or impossible to resolve

Putting it all together

Four step plan to network inference:

1. Using structural assumptions like sparsity that effect global network properties
2. Generating and using informative priors on network structure that effect single edges
3. TF activity estimation, and
4. Using guided/intelligent TF-TF interaction terms

Source: Richard Bonneau and Tarmo Äijö
Biophysically motivated regulatory network inference: progress and prospects
bioRxiv 051847; doi: http://dx.doi.org/10.1101/051847

Outline

Learning goals for this week

- To understand reasons for difference of correlation and causality
- To recognise basic regulatory network inference methods
- To apply simple methods for network inference from data

Tasks for the study circle on Thursday

Paper:

- O. Heinävaara, J. Leppä-aho, J. Corander and A. Honkela. On the inconsistency of ℓ_1-penalised sparse precision matrix estimation. arXiv:1603.02532 [cs.LG]

Task for all:

- Read the paper to form an overview of the topic.
- You will not need to understand all the mathematical details.

Next week: guest lectures

See the course website for details!