Analysis of **globally coherent** datasets

From enrichment and clusters to networks and mechanisms

Florian Markowetz
markowetzlab.org

"A SNP in FGFR2 promotes cancer susceptibility"

"HER2 positive breast cancers are very aggressive"

"Silencing a single gene deforms morphology"

amplification

*copy number alteration*

deletion

**SNP**

perturbation

*genome-wide*

RNA interference

*Knock-out*

pathway-specific
cancer subtypes
different risk groups

phenotype
viability
morphology
differential gene expression

perturbation
networks
phenotype

Our lab

<table>
<thead>
<tr>
<th>Natural perturbations</th>
<th>Experimental perturbations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perturbation</td>
<td>Perturbations</td>
</tr>
<tr>
<td>- copy-number alterations</td>
<td>- RNA interference</td>
</tr>
<tr>
<td>Network</td>
<td>Network</td>
</tr>
<tr>
<td>- Pathways and regulatory networks in the cell</td>
<td>- Signal transduction pathways</td>
</tr>
<tr>
<td>Phenotype</td>
<td>Phenotype</td>
</tr>
<tr>
<td>- disease subtypes, survival, etc</td>
<td>- differential gene expression downstream of pathway</td>
</tr>
<tr>
<td>Goal</td>
<td>Goal</td>
</tr>
<tr>
<td>- find regulatory hotspots to explain heterogeneity of disease</td>
<td>- Reconstruct pathway</td>
</tr>
</tbody>
</table>
What is globally coherent data?

Phenotype

DNA Intermediate phenotype

SNPs

Copy-number alterations

SNPs e.g. QTLs e.g. eQTLs proteins mRNA

metabolites

obesity survival cancer subtypes

Examples of Globally Coherent Data

METABRIC @ CRI

2000 breast tumours

Challenges of Globally Coherent Data

linking/integrating/comparing different data types

– find homogeneous sub-types of heterogeneous disease with impact on outcome
– Mechanistic explanation of observed changes (“causal models”)
– not single marker, but complete story

• methods applicable in many other integrative tasks (eg microRNA + expression)
This course is about tools

<table>
<thead>
<tr>
<th>Data</th>
<th>Pre-processing</th>
<th>Analysis</th>
<th>Follow-up</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw clay</td>
<td>Quality control</td>
<td>Sanity checks</td>
<td>Statistics</td>
<td>Machine learning et al</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Story</td>
<td>Yeah!</td>
</tr>
</tbody>
</table>

We are here!

Programme

Today

**morning**
- lectures on statistical and machine learning concepts to analyse GCDs (FM)

**afternoon**
- practical session to try them out on real data (YY, MC)

Tomorrow

**morning**
- Discussion of tools from day 1, merits and pitfalls, how best to use them for your projects (FM)

**afternoon**
- practical session continues (YY, MC)
Overview

Clustering
- Hierarchical clustering, Mixture models, Dirichlet process, Bayesian hierarchical clustering, integrative clustering

Enrichment
- hypergeometric test, Gene set enrichment analysis, rich subnetworks, HTSanalyzeR

Networks
- Schadt’s ‘causal’ networks, Bayesian networks, conditional independence
- DANCE: Differential regulation in different disease subtypes, NEMs: Nested Effects Models

Genes -> networks -> disease

clustering + enrichment = story
From data to distances

What distance measure should we use?

Distance or dissimilarity matrix

\[
D_{i,j} = \text{dist}(M_{i,:}, M_{j,:})
\]

\[
D_{j,i} = D_{i,j}
\]

\[
D_{i,i} = 0 \quad \text{for all } i
\]

Examples of distances

**Euclidean distance**

\[
\text{dist}(a, b) = \|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}
\]

**Manhattan distance**

\[
\text{dist}(a, b) = \|a - b\|_1 = \sum_i |a_i - b_i|
\]

**Cosine distance**

\[
\text{dist}(a, b) = \cos^{-1} \left( \frac{\langle a, b \rangle}{\|a\| \|b\|} \right)
\]

how is this related to correlation?

Linkage: distances to clusters

\[
\text{dist}(C_1, C_2) = \max \{ \text{dist}(i, j) : i \in C_1, j \in C_2 \}
\]

\[
\text{complete linkage}
\]

\[
= \min \{ \text{dist}(i, j) : i \in C_1, j \in C_2 \}
\]

\[
\text{single linkage}
\]

\[
= \text{mean} \{ \text{dist}(i, j) : i \in C_1, j \in C_2 \}
\]

\[
\text{average linkage}
\]

Distances between individual genes

\[D(3,4) = ???\]

\[D(2.3,4) = ???
\]
Hierarchical clustering

Ingredients: data matrix, distance function, linkage function

Dendrogram

Normal mixture models

All mixture components are Normal, but with different means and covariances.

\[ p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \]

Special case: k-means

All mixture components are Normal with different means, but same simple covariance matrices.

\[ p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \]
How many clusters?

- Too few 😞
- Too many :( 

Key questions in clustering

- How many clusters?
- How to integrate many data types?
  - E.g., to find concomitant patterns in gene expression and copy-number
- How to identify important features?
  - Which genes/alterations drive the clustering

Generating mixture data

A fixed collection of parameter sets

\[ G = \sum_{k=1}^{K} \pi_k \delta_{\mu_k, x_k} \]

Randomly select one parameter set

\[ \theta_i | G \sim G \]

Given the parameter set, sample the data

\[ x_i | \theta_i \sim N(\cdot | \theta_i) \]
Dirichlet process clustering

Not a fixed collection of parameters
Parameter sets are sampled
One is chosen…
… to generate a data point

Dirichlet process: Chinese restaurant

Tables = parameter set
Customers = data points
Customer either (i) chooses already occupied table or (ii) opens new table.
Probability increases with number of people already at the table (-> clustering!)

Dirichlet process mixtures (DPMs)

• Mixture models + Dirichlet process = models with (countably) infinite mixtures
• Big advantage:
  – estimate of how many clusters are in the data
  – not limited to pre-defined number
• How do we find the best model for our data?
  – popular (as always): Markov Chain Monte Carlo
• Uh oh … but that will be taking a reeeaaaaally long time for large genomic datasets.
**Bayesian Hierarchical Clustering**

- Like hierarchical clustering, but using marginal likelihood of DPM instead of (ad hoc) distance.
- Bayesian hypothesis testing decides which clusters to merge.
- **Advantage**: estimates optimal number and size of clusters in the data.

**BHC in practical session**

**Key questions in clustering**

- How many clusters?
- How to integrate many data types?
- How to identify important features?
**iCluster: basic model**

- (k-means) mixture model
- written as latent variable model
- sparsity constraints

For one data type:

\[ X = WZ + \varepsilon \]

**iCluster: integrative clustering**

For many data types: 1, 2, ..., m

\[ X_1 = W_1 Z + \varepsilon_1 \]
\[ X_2 = W_2 Z + \varepsilon_2 \]
\[ \ldots \]
\[ X_m = W_m Z + \varepsilon_m \]

Different coefficients for different data types, the same clustering matrix for all data types.

**iCluster: Sparsity**

For many data types: 1, 2, ..., m

\[ X_1 = W_1 Z + \varepsilon_1 \]
\[ X_2 = W_2 Z + \varepsilon_2 \]
\[ \ldots \]
\[ X_m = W_m Z + \varepsilon_m \]

Implemented by Lasso penalty during estimation: Sum of absolute value of entries must be small.

Sparsity assumption: The W-matrices are sparsely populated, i.e., most entries = 0.
Clustering: PROs and CONs

PRO
- Standard analysis, almost always applicable
- Global first overview
- Can identify strong trends and patterns in the data

NEG
- Often applied in situations where other methods would be more appropriate (e.g. supervised analysis)

Overview

Clustering
- Hierarchical clustering, Mixture models, Dirichlet process, Bayesian hierarchical clustering, integrative clustering

Enrichment
- Hypergeometric test, Gene set enrichment analysis, rich subnetworks, HTSanalyzeR

Networks
- Schadt’s ‘causal’ networks, Bayesian networks, conditional independence
- DANCE: Differential regulation in different disease subtypes, NEMs: Nested Effects Models

Gene Ontology (GO)

www.geneontology.org
Over-representation analysis

Hyper-geometric test

Hyper-geometric distribution

Hyper-geometric test: example
Gene Set Enrichment Analysis (GSEA)
Subramanian et al. (2005)

**Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**

**GSEA: construction**

\[ P_{hit}(S, i) = \frac{\text{phenotype before } i}{\sum \text{all phenotype}} \quad \text{if } p = 1 \]

\[ P_{miss}(S, i) = \frac{\text{Nr. non-hits before } i}{\text{Nr. all non-hits}} \]

The ES is the maximum deviation from zero of \( P_{hit} - P_{miss} \)
**GSEA: examples**

Table 1. P value comparison of gene sets by using original and new methods

<table>
<thead>
<tr>
<th>Gene set</th>
<th>Original method nominal P value</th>
<th>New method nominal P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: drX inactive</td>
<td>0.007</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>2: vrlb pathway</td>
<td>0.51</td>
<td>0.035</td>
</tr>
<tr>
<td>3: nkt pathway</td>
<td>0.023</td>
<td>0.54</td>
</tr>
</tbody>
</table>

**Differential GSEA**

**PROs and CONs**

**Result:**

- p-values
- (hyper-geometric or GSEA)

**Advantages:**
- standard analysis
- comprehensive first overview
- "unbiased" and "hypothesis-free"

**Disadvantages:**
- "unbiased" and "hypothesis-free"
- relies on known gene sets
- can not uncover new pathways
- pathway = "unconnected" gene set
- soon: more gene sets than genes!
Sub-networks rich in hits

Sub-networks with highly correlated phenotypes

Which networks?

Networks from large-scale experiments

Networks from analyzing the experimental literature

Networks from probabilistic data integration
Overview

Clustering
- Hierarchical clustering, Mixture models, Dirichlet process, Bayesian hierarchical clustering, integrative clustering

Enrichment
- Hypergeometric test, Gene set enrichment analysis, rich subnetworks, HTSanalyzeR

Networks
- Schadt’s ‘causal’ networks, Bayesian networks, conditional independence
- DANCE: Differential regulation in different disease subtypes, NEMs: Nested Effects Models

Schadt’s “causal” networks

Really really really complicated model
Bayesian network: Definition

A Bayesian network for \( X = (X_1, \ldots, X_n) \) consists of:

1. A directed acyclic graph (DAG) with vertices corresponding to \( X_1, \ldots, X_n \)
2. for each node a conditional probability distribution (LPD) \( P(X_i | Pa_i) \)
3. Each LPD is parameterized by a vector \( \theta_i \)

There is nothing really Bayesian about Bayesian networks so far.

Joint distribution

\[
P(X_1, \ldots, X_n)
\]

"Naive" chain rule: for every \( (X_1, \ldots, X_n) \) the joint distribution can be factorized as

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
\]

... but sparser by using conditional independence relations encoded in the DAG:

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa_i)
\]

Conditional independence

[Diagram showing conditional independence]
Single family Bayesian network

Directed acyclic graph defines families: gene and regulators.

Relation of parents to child is described by conditional distribution

\[
P(\text{child} | \text{parents})
\]

\[
\begin{array}{c|ccc}
& \text{home} & \text{work} \\
\text{home} & .1 & .8 & .1 \\
\text{work} & .1 & .8 & .1 \\
\end{array}
\]

Which model is the right one?
The one that fits the data best!

Which model is the right one?
The one that fits the data best!

Which model is the right one?
The one that fits the data best!

\[
P(L, R, C) = P(L) P(R | L) P(C | R)
\]

\[
P(L, R, C) = P(L) P(R | L) P(C | R)
\]

\[
P(L, R, C) = P(L) P(R | L) P(C | R)
\]

Model selection: Akaike Information Criterion (AIC)
- Find parameters that maximize the likelihood for each model
- Subtract number of parameters to penalize complex models
- Select model with highest AIC value

Score-based model selection

Conditional independence tests

- L is correlated with R and C
- C is independent of L given R

(= partial cor. between C and L equals 0)

Shrinkage tests in GeneNet package
Assumptions of “causal” networks

- **Inference from observational data** based on conditional independence
- In general: **direction of edges not determined** (partial correlation still only a correlation)
- In Schadt’s scenario: genomic locus serves as anchor to direct edges
- This assumption needs to be checked, e.g. in cancer genomic alterations are not fixed but evolutionary selected.

Overview

**Clustering**
- Hierarchical clustering, Mixture models, Dirichlet process, Bayesian hierarchical clustering, integrative clustering

**Enrichment**
- Hypergeometric test, Gene set enrichment analysis, rich subnetworks, HTSanalyzeR

**Networks**
- Schadt’s ‘causal’ networks, Bayesian networks, conditional independence
- DANCE: Differential regulation in different disease subtypes, NEMs: Nested Effects Models

Copy number alterations measured by genomic markers
Automated quantitative cellularity correction

Impact of CNA on expression

\[ Y_i = \begin{cases} \text{Baseline} & \text{Yi} \\ \text{cis-effect} & x_i \beta_i + \epsilon_i \\ \text{cis- and trans-effects} & x_i \beta_i + \sum_{j \neq i} x_j \beta_j + \epsilon'_i \end{cases} \]

\[ \text{score}_j = - \ln \left( \frac{\sigma_{\text{with}}^2}{\sigma_{\text{without}}^2} \right) \]

Differential regulation

Subtype A, eg ER+ breast cancer

Subtype B, eg ER- breast cancer
Differential regulation

\[
\begin{align*}
Y_1 &= X_1B^r + X_1B^d + \epsilon_1 \\
Y_2 &= X_2B^r + \epsilon_2
\end{align*}
\]

Gene Expression

Copy-number

Reference "network"

Differential "network"

\[
\begin{bmatrix}
Y_1 \\
Y_2
\end{bmatrix} =
\begin{bmatrix}
X_1 & X_1 & B^r & B^d \\
X_2 & 0 & &
\end{bmatrix}
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2
\end{bmatrix}
\]
solved by Lasso

Hotspots: ER+ versus ER-

concomitant CN signatures only

21q22.3 Amp with few cis-changes

6q22.31 WISP3/PPAC/PPD Wnt-1 inducible protein

Data from Chin et al, 07

“The end of the screen is the beginning of the experiment”

Boutros and Ahringer 2008
Anatomy of the NFκB pathway

Step 1

Step 2

Knock-down Known pathway members New RNAi Hits

Compare expression phenotypes by NEMs

Nested Effects Models

Nested effect models: subset relations similarity

Other statistical methods: similarity


Nested Effects Models

INPUT
1. Set of candidate pathway genes
2. High-dimensional phenotypic profile, e.g. microarray

OUTPUT Graph explaining the phenotypes

Phenotypic profiles Inferred pathway

Gene perturbations

A B C D E F G H

Effects
Summary

Clustering
- Hierarchical clustering, Mixture models, Dirichlet process, Bayesian hierarchical clustering, integrative clustering

Enrichment
- Hypergeometric test, Gene set enrichment analysis, rich subnetworks, HTSanalyzeR

Networks
- Schadt’s ‘causal’ networks, Bayesian networks, conditional independence
- DANCE: Differential regulation in different disease subtypes, NEMs: Nested Effects Models

Network biology

- Better algorithms
- Better questions

Today

Tomorrow ??
We do network analysis
what network?

“The network is a model”
Model of what?
mechanistic model or predictive?

What analysis?

More than (just) a cluster?

Is the network the solution?
Would another representation be clearer?

The’ or ‘A’?

Shameless self-promotion

How to Understand the Cell by Breaking It: Network Analysis of Gene Perturbation Screens

Software

- HTSanalyzeR provides an integrated interface to enrichment and network analysis.
- DANCE quantifies the impact of genomic alterations on gene expression and compares it between tumour sub-types.
- lol contains various optimization methods for matrix-to-matrix Lasso inference.
- nem infers Nested Effects Models from data.

http://www.markowetzlab.org/software/
Some *take home* messages

• Large chunks of network analysis are based on **clustering and enrichment**.
• Most of the rest is based on **conditional independence** and (sparse) **regression**.
• Big networks tend to be hairy. Avoid the **hairball** by asking more focused questions.
• Network analysis is even (more?) useful when targeting a **single pathway**.

Practical session

**Clustering**
• Hierarchical clustering, **BHC**: Bayesian hierarchical clustering, **iCluster**: integrative clustering

**Enrichment**
• Hypergeometric test, Gene set enrichment analysis, **BioNet**: rich subnetworks, **HTSanalyzeR**

**Networks**
• Schadt’s ‘causal’ networks, conditional independence
• **DANCE**: Differential regulation in different disease sub-types

the team

CANCER RESEARCH UK

UNIVERSITY OF CAMBRIDGE

Hutchison Whampoa
Analysis of globally coherent datasets

Thank you!

Florian Markowetz
markowetzlab.org