PROGRESSION ANALYSIS OF DISEASE APPLIED TO BREAST CANCER RESEARCH
Rachel Jeitziner
20.06.2015
INTRODUCTION
INTRODUCTION

- During the past few decades, science has made tremendous progress, and the latest technological tools allow assaying tens of thousands of genes simultaneously.
- Hence, we get large volumes of data to search for cancer biomarkers.
- How can we analyze such tons of data?
HOW CAN WE ANALYZE SUCH TONS OF DATA?

- **Or** how can we extract some qualitative signal form this data?
- **And** extract information that is robust in the presence of noise and errors.
- **And** still catching the important features of the data.
Solution... Topological Data Analysis!

- Reduce the complexity of the data.
- Extract some qualitative information.
- Can describe the notion of closeness from one special "point" to a "set" (for example a gene expressions from one person to the gene expressions of many others, specially a diseased genome to a group of microarrays from healthy tissues).
SOLUTION... PAD

- The method is called Progression analysis of disease (PAD) and is composed of two parts
  - Disease specific genomic analysis (DSGA)
  - and Mapper.
INTERESTING ARTICLES-DSGA AND PDA


PROGRESSION ANALYSIS OF DISEASE METHOD
**Mapper**

- One has to choose a « magnitude function » called $f: X \rightarrow Z$, where $Z$ is a reference space. Here, $Z$ is

- and a « distance function » $d$, which is here the euclidean distance in $\mathbb{R}^3$.
- $x \sim x'$ if $d(x,x') < \varepsilon$.
- This is not yet an equivalence relation, but becomes one under transitive closure, i.e., if $x \sim x_0$ and $x_0 \sim x_1$, then we set $x \sim x_1$.
- The algorithm consists in creating "clusters", by taking the equivalence classes.

Gunnar et al., *Extracting insights of complex data using topology*, 2013
**Mapper**

- Choose a covering \( U = \{U_\alpha\} \) of \( Z \).
- One applies this clustering to \( X_\alpha = f^{-1}(U_\alpha) \).
- Take equivalence classes on \( X_\alpha \).
- Hence, we create for every \((\alpha, c)\) a vertex.
- \( \{(\alpha_0, c_0), (\alpha_1, c_1), \ldots, (\alpha_k, c_k)\} \) spans a \( k \)-simplex if and only if the clusters \( X_{(\alpha_0, c_0), \ldots, X_{(\alpha_k, c_k)}} \) have a non empty intersection.
DISEASE-SPECIFIC GENOMIC ANALYSIS

- Normal like tissue microarray: $N_1, \ldots, N_S$, where $S$ is the number of normal samples ($S$ has to be smaller than the number of genes $n$),

- **GROUP A**

- Diseased tissue microarray: $T_1, \ldots, T_R$, where $R$ is the number of diseased samples.

- **GROUP B**
DISEASE-SPECIFIC GENOMIC ANALYSIS

- We define a subspace \( \mathcal{N} \) which represents the normal tissue data

\[ T = Nc.T + Dc.T \]

- Now, we have to reduce the size \( n \) of genes and every statistical method of size reduction can be used.
METHOD

- **GROUP A**: \( N_1, \ldots, N_S \)
- **GROUP B**: \( T_1, \ldots, T_R \)

\( T = N_c.T + D_c.T \). Hence, one generates: \( D_c.T_1, \ldots, D_c.T_R \).

The magnitude function \( f: X \rightarrow Z \), is defined as

\[
f(Dc.T) = \left( \sum |g_r|^l \right)^{k/l}
\]

The distance is the correlation distance

\[
d(u, v) = 1 - \frac{(u - \text{mean}(u)) \cdot (v - \text{mean}(v))}{\|u - \text{mean}(u)\| \|v - \text{mean}(v)\|}.
\]

Cover \( K \) intervals with \( T\% \) overlap.
TOY EXAMPLE: PCA vs MAPPER

- $u_+ = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, u_- = \begin{pmatrix} -1 \\ -2 \\ -3 \end{pmatrix}, u_2 = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$

- Distance matrix:
  \[
  \begin{pmatrix}
  0 & 2 & 0 \\
  2 & 0 & 2 \\
  0 & 2 & 0 
  \end{pmatrix}
  \]

- 5 intervals / 66% overlap
FIRST APPLICATION OF PAD
BACKGROUND ON RANKL

- Receptor activator of nuclear factor kB ligand (RANKL) has been shown to be an important protein in mouse mammary gland development and mammary carcinogenesis (M. Beleut et al.)
- Has it the same role in humans?
CONTEXT OF THE EXPERIMENT

Add Vehicle

Stimulate with RANKL

Same patient
M ODERATED t-TEST ADJUSTED WITH THE B ENJAMINI-H OCHERBERG METHOD

No significant gene!
TOPOLOGY?

- Use the DSGA method to find significantly different genes specific to RANKL.
- “Healthy State model” is the control group (GROUP A)
- “Tumour component” is the group where RANKL was added (GROUP B).
Sample 1: 1nmol/l Sample 2: 20nmol/l Sample 3: 18nmol/l
Sample 4: 2nmol/l Sample 5: 1nmol/l Sample 6: 0.7nmol/l
SECOND APPLICATION OF PAD
HORMONES IMPLICATED IN BREAST CANCER

- Epidemiological studies showed that there is an increase of the breast cancer risk of 1.24 for women taking oral contraceptives.
- There is no significant risk anymore 10 years (or more) after stopping use.

The idea of oral contraceptives is to block ovulation by introducing a low level of estrogen. This has the consequence that the luteal phase is abrogated. However, we do not know the action of progestins, which are meant to be replacing the progesterone peak.

MAPPER-ANALYSIS

Close to the HSM

Far away from the HSM

Contraceptives 2

Luteal 10

Contraceptives 4

Contraceptives 11, Luteal 12

Luteal 16

Contraceptives 11, Luteal 12, Luteal 16

Luteal 16

Contraceptives 11, Luteal 16

Contraceptives 13

Luteal 8

Luteal 8, Contraceptives 13

Luteal 10

Contraceptives 6, Luteal 7

Contraceptives 18

Luteal 7

Luteal 7
STANDARD TOOLS, DENDROGRAM
STANDARD TOOLS, DSGA OUTPUT

Cluster Dendrogram

PAD colour

Height

Normal_10.Dis

Normal_4.Dis

Normal_6.Dis

Normal_7.Dis

Normal_16.Dis

Normal_18.Dis

Normal_13.Dis

Normal_11.Dis

Normal_12.Dis

Normal_8.Dis

Normal_2.Dis
PRINCIPAL COMPONENT ANALYSIS