Big Biological Data (Analyses)
- NGS data

Alexander Platzer – February 2015
Schedule

- 1001 Genomes Project
- Dimension reduction
- GWAS with transposable elements
- Indel pattern search
- Transcriptional Enhancement with Natural Antisense Transcripts
1001 Genomes Project

• More than 1000 full-genome sequenced Arabidopsis thaliana

• Related data (same accessions) from/with other projects:
  – RNAseq
  – Methylation
  – Phenotypes
1001 Genomes Project
1001 Genomes Project

1001 genomes - TE-copy numbers
Hierarchical clustering

country
- UK
- GER
- SWE_S
- AZE
- USA
- ESP
- SWE_N
- GEO
1001 genomes - TE-copy numbers
Hierarchical clustering

country
- UK
- GER
- SWE_S
- AZE
- USA
- ESP
- SWE_N
- GEO

Alexander Platzer
Visualization of SNPs with t-SNE

Dimension reduction
Problem definition - specific

- A bunch of binary/numerical variables (a lot of SNPs)
- One nominal label (phenotype)

![Matrix of binary/numerical variables](image1.png)

![Heatmap of SNP clusters](image2.png)

Alexander Platzer
Problem definition - general

a huge amount of

numerical

ordinal

nominal
dimensions

2 or 3 numerical dimensions
(for plotting)

+ maybe 1-some
nominal effect
dimensions
Problem definition - general

A huge amount of numerical, ordinal and nominal dimensions lead to 2 or 3 numerical dimensions.

... there is a subfield ‘dimension reduction’ for this.

Beside the widely used PCA, the alternatives are e.g.: Sammon mapping, Isomap, Locally Linear Embedding, Classical multidimensional scaling, Laplacian Eigenmap, m-SNE, t-SNE, ... .
Visualizations

PCA

data from:

---

t-SNE
Visualizations

data selection from The International HapMap Consortium.

Integrating common and rare genetic variation in diverse human populations
Visualizations

PCA

PC1

PC2

data selection from
A map of rice genome variation reveals the origin of cultivated rice.
Nature 490: 497-501

Alexander Platzer
Measurements - Question

- how well is the data structured?
- how much (correct) insight can be obtained from it?
Measurements - Structuredness - As classification problem

Transformed data / diagrams

<table>
<thead>
<tr>
<th>Coordinates</th>
<th>Classification Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.0717, 15.357</td>
<td>Burma</td>
</tr>
<tr>
<td>-9.5674, 14.878</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>15.314, -12.017</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>14.705, -8.3615</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>3.1356, -34.54</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>15.119, -9.0675</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>15.634, -9.2958</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>27.437, 11.955</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>-0.34996, -33.382</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>-28.565, -1.9802</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>-0.30935, -35.71</td>
<td>Bangladesh</td>
</tr>
</tbody>
</table>

C4.5, PART, Perceptron, Naive Bayes
## Measurements - As classification problem

<table>
<thead>
<tr>
<th></th>
<th>1001 genomes project</th>
<th>RegMap</th>
<th>hapmap3 r2</th>
<th>hapmap3 r3</th>
<th>Rice</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x cross-val.%</td>
<td>PCA</td>
<td>t-SNE</td>
<td>PCA</td>
<td>t-SNE</td>
<td>PCA</td>
</tr>
<tr>
<td>C4.5</td>
<td>55.6</td>
<td>72.7</td>
<td>79.2</td>
<td>89.7</td>
<td>72.9</td>
</tr>
<tr>
<td>PART</td>
<td>60.6</td>
<td>76.8</td>
<td>77.6</td>
<td>89.1</td>
<td>72.7</td>
</tr>
<tr>
<td>Perceptron</td>
<td>67.7</td>
<td>76.8</td>
<td>80.7</td>
<td>85.8</td>
<td>70.3</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>62.6</td>
<td>75.8</td>
<td>75.2</td>
<td>80.3</td>
<td>74.6</td>
</tr>
<tr>
<td>Mean diff.</td>
<td>13.9</td>
<td>8.1</td>
<td>15.8</td>
<td></td>
<td>13.5</td>
</tr>
<tr>
<td>St.dev.</td>
<td>3.6</td>
<td>3.4</td>
<td>2.6</td>
<td></td>
<td>1.2</td>
</tr>
</tbody>
</table>

Transposable Elements
TE-locate

TE-Locate: A Tool to Locate and Group Transposable Element Occurrences Using Next-generation Sequencing Data

Input:

• (Read pairs)
• A reference genome
• Annotated TE elements therein
• SAM (dependency: A aligner)
• Optional: Mapping from annotated TEs to a higher hierarchical level)
TE – Hierarchy
(as in The Gypsy Database (GyDB) of Mobile Genetic Elements)

• Superfamily (e.g. LTR/Gypsy)
  • Systems (e.g. LTR retroelements)
    • Families (Ty1/Copia)
      • Elements (Hydra1-1)
        • Annotated reference loci (AT1TE09970)
TE-Locate: Method

Ref. Seq.  TE1  TE1  TE2 (additional to reference)  TE3

TE-locate

New accession seq. (with called TEs)

Alexander Platzer
Results: GWAS of copy numbers

Sources:
- New SNP release
- TE superfamily copy number as phenotype

e.g.:
Results: GWAS of copy numbers
Dependency of GWASes
Indel patterns
Indels pattern search

• +/- 30 bp are extracted
• from the loci of called indels (Q30) and
• randomly from the reference (TAIR10)

• Classified with C5.0

• as all rules are starting with the bp just before the event, this is declared as BP1 (the BP2, BP3, ....), the pattern strings are always starting with BP1
Indels patterns – MSH2\(^{-}\) vs Swedish

**MSH2\(^{-}\) (MA line)**
- 186 indels vs. 2000 random
- zero ratio = 91.5%
- PCC = 97.7%

- **Rule 1**
  - 0.923 (48%)
  - NAAAAAA

- **Rule 2**
  - 0.871 (53%)
  - NTTTTTT

**Population (part of 1001 genomes)**
- 1031959 indels vs. 1000000 random
- zero ratio = 50.8%
- PCC = 67.2%

- **Rule 1**
  - 0.874 (50.0%)
  - NNAAAAA

- **Rule 4**
  - 0.776 (70.1%)
  - BP2 = T
  - BP4 = T
  - BP8 = T

- **Rule 5**
  - 0.682 (11.6%)
  - TAA

- **Rule 6**
  - 0.678 (6.5%)
  - NGGG

- **Rule 7**
  - 0.668 (6.5%)
  - NCCC

- **Rule 8**
  - 0.597 (2.0%)
  - NGAGA

- **Rule 9**
  - 0.560 (277.3%)
  - BP3 = A

- **Rule 10**
  - 0.597 (276.3%)
  - BP3 = T
Indels patterns

- Class 2 ... 'MSH2'-alike rules: 1, 2 and 5
- Class 1 ... [CCC|GGG]
- Class 0 ... the remainder

Several patterns, but ~two main classes:
- As|Ts
- Gs|Cs
MA-lines - Indels - Main classes

- Swedish
- CVI
- yeast
- MAindividuals
- MATissues
- yeast ~ (MSH2-)
- Ecoli mutL-

Alexander Platzer

~real populations

increased mutation accumulation MSH2-
MLH in Arabidopsis

doi: 10.1111/j.1365-313X.2007.03145.x

An Arabidopsis MLH1 mutant exhibits reproductive defects and reveals a dual role for this gene in mitotic recombination

Éric Dion†, Liangliang Li†,* Martine Jean and François Belzile*
Département de phytologie, 1243 Pavillon C.-E. Marchand, Université Laval, Québec, QC, G1K 7P4, Canada

Received 12 February 2007; revised 20 March 2007; accepted 26 March 2007.
*For correspondence (fax +1 418 656 7176; e-mail fbelzile@rsvs.ulaval.ca).
†These authors contributed equally to this work.
*Present address: Medicago Inc., 1020 route de l’Église, Bureau 600, Québec, QC, G1V 3V9, Canada.

Summary

The eukaryotic DNA mismatch repair (MMR) system contributes to maintaining genome integrity and DNA sequence fidelity in at least two important ways: by correcting errors arising during DNA replication, and also by preventing recombination events between divergent sequences. This study aimed to investigate the role of one key MMR gene in recombination. We obtained a mutant line in which the AtMLH1 gene has been disrupted by the insertion of a T-DNA within the coding region. Transcript analysis indicated that no full-length transcript was produced in mutant plants. The loss of a functional AtMLH1 gene led to a significant reduction in fertility in both homozygotes and heterozygotes, and we observed a strong bias against transmission of the mutant
Transcriptional Enhancement with Natural Antisense Transcripts
Transcriptional Enhancement with Natural Antisense Transcripts

The vast majority of known NATs are lowering mRNA expression and/or translation.

Beside that there are two strongly validated examples of enhancing translation.

- PHO1-2 in rice
- Uchl1 in mouse

More examples? Classification and prediction? Mechanism?
somehow conserved sequence of cisNAT of PHO1-2 non-overlapping part (combined with Arabidopsis thaliana)
Summary

- 1000´s fully sequenced genomes
- Dimension reduction methods
- GWAS with transposable elements
- Sequence pattern search
- Transcriptional Enhancement with Natural Antisense Transcripts
Acknowledgements

Quan Long
Irina Druzhinina
Laurens van der Maaten
Ivo Hofacker
t-SNE

constraints:
• Dimension reduction for not more than 3 dimensions
• Local linearity (violated by highly-varying manifolds)
• Non-convexity of the t-SNE cost function
t-SNE - Parameters

- **Initial dimensions**: The number of dimensions the data is reduced in advance (with PCA), because t-SNE works only up to quadratic matrices; this can be set to the maximum (= number of data records = number of individuals)

- **perplexity**: As defined in information theory, it is a measure how confused a model might be; can be interpreted as smoothing of the data and/or as expected noise in the data, or seen as the number of neighbors taken into account at looking to the distances -> lower than the number of individuals, range: 5 to #individuals/2
t-SNE - Perplexity

The parameter is used to normalize the conditional probability distributions to this perplexity.

\[ 2^{H(p,q)} \quad \text{with} \quad H(p,q) = -\sum_x p(x) \log q(x). \]
t-SNE - Algorithm

**Algorithm 1**: Simple version of t-Distributed Stochastic Neighbor Embedding.

**Data**: data set $X = \{x_1, x_2, ..., x_n\}$,

- cost function parameters: perplexity $Perp$,
- optimization parameters: number of iterations $T$, learning rate $\eta$, momentum $\alpha(t)$.

**Result**: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, ..., y_n\}$.

**begin**

- compute pairwise affinities $p_{ji}$ with perplexity $Perp$ (using Equation 1)

$$p_{ji} = \frac{p_{ji} + p_{ij}}{2n}$$

- sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, ..., y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$

**for** $t=1$ to $T$ **do**

- compute low-dimensional affinities $q_{ij}$ (using Equation 4)

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

- compute gradient $\frac{\partial C}{\partial y_i}$ (using Equation 5)

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j) \left(1 + \|y_i - y_j\|^2\right)^{-1}$$

- set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\partial C}{\partial y} + \alpha(t) \left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$

**end**

**end**

momentum : 0.5 -> 0.8

init. learning rate : 500 (then adaptive)
Measurements – Structuredness - As a clustering result

transformed data / diagrams

Coordinates

as clustering result

<table>
<thead>
<tr>
<th>distance</th>
<th>label1</th>
<th>label2</th>
</tr>
</thead>
</table>

Dunn’s Validity Index, Silhouette Validation Method

Alexander Platzer
# Measurements - Indices of cluster validity

<table>
<thead>
<tr>
<th>Data</th>
<th>Dunn’s Validity Index</th>
<th>Silhouette Validation Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCA</td>
<td>t-SNE</td>
</tr>
<tr>
<td>1001 genomes</td>
<td>0.52 (0.09)</td>
<td>0.61 (0.07)</td>
</tr>
<tr>
<td>RegMap</td>
<td>0.50 (0.06)</td>
<td>0.50 (0.04)</td>
</tr>
<tr>
<td>Hapmap3R2</td>
<td>0.16 (0.01)</td>
<td>0.25 (0.02)</td>
</tr>
<tr>
<td>Hapmap3R3</td>
<td>0.16 (0.01)</td>
<td>0.35 (0.01)</td>
</tr>
<tr>
<td>Rice</td>
<td>0.06 (0.07)</td>
<td>0.10 (0.10)</td>
</tr>
</tbody>
</table>
Backup