Patterns from Multiresolution 0-1 data

Prem Raj Adhikari, Jaakko Hollmén

Aalto University School of Science and Technology
Department of Information and Computer Science
Espoo, Finland

KDD 2010 Workshop on Useful Patterns
July 25, 2010
Outline

1. Introduction to Chromosomal Aberration in Multiple Resolution
2. Sampling data between different resolutions
3. Comparison of Different Downsampling methods
4. Mixture Models of Amplification Patterns in Cancer
5. Maximal Frequent Itemset
6. Summary and Conclusions
Chromosomal Aberration

- Disruptions in the normal chromosomal content of a cell
- A major cause of genetic conditions in humans
- Chromosomal aberration cause cancers and other diseases
- The complex case of Copy numbers
  - Deletion is the case when the copy number is less than two
  - Duplication is the case when the copy number is more than two
  - Amplification is the case when the copy number increases more than 5.
- Why detect copy numbers?
- DNA copy number aberrations are hallmarks of cancer.
Chromosome Nomenclature

- International System for Human Cytogenetic Nomenclature (ISCN)
- Short arm locations are labeled p (petit)
- Long arms q (queue)
- 17p13.2: chromosome 17, the arm p, region(band) 13, subregion(subband) 2
- Hierarchical, irregular naming scheme; cumbersome for scripting/manual
Multiple Resolutions: Chromosome-17

Figure: G-banding patterns for normal human chromosomes at five different levels of resolution. Source: (Shaffer et. al. 2009). Example case in Chromosome:17.
**Multiple Resolutions: Part of Chromosome-17**

<table>
<thead>
<tr>
<th>Coarse Resolution</th>
<th>Fine Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>q21</td>
<td>q21.1</td>
</tr>
<tr>
<td>q21.2</td>
<td>q21.2</td>
</tr>
<tr>
<td>q21.3</td>
<td>q21.3</td>
</tr>
<tr>
<td>q22</td>
<td>q22</td>
</tr>
<tr>
<td>q23</td>
<td>q23.1</td>
</tr>
<tr>
<td>q23.2</td>
<td>q23.2</td>
</tr>
<tr>
<td>q23.3</td>
<td>q23.3</td>
</tr>
<tr>
<td>q24</td>
<td>q24.1</td>
</tr>
<tr>
<td>q24.2</td>
<td>q24.2</td>
</tr>
<tr>
<td>q24.3</td>
<td>q24.3</td>
</tr>
</tbody>
</table>

**Figure:** Part of chromosome 17 showing the differences in multiple resolutions.
Multiple Resolutions: the problem

Problem

Two different datasets are available in two different resolutions. How do you map into other resolutions such that patterns are preserved?
DNA copy number amplification Dataset

**Figure:** Collected by bibliomics survey of 838 journal articles during 1992-2002 in (S. Myllykangas et. al. 2006 and 2008). 4590 samples in resolution 400(left panel) and different dataset in resolution 850(Right Panel). Sparse and spatially dependent matrix. Available from the authors.
Changing between different resolutions

Upsampling

- Upsampling is the process of changing the representation of data to the higher or finer resolution.
- Simple transformation table involving chromosome bands was used to upsample data from the resolution 400 to different finer resolutions.
- The transformation table were chromosome specific and resolution specific (88 tables for 5 resolutions).

<table>
<thead>
<tr>
<th>Resolution:400</th>
<th>Resolution:850</th>
</tr>
</thead>
<tbody>
<tr>
<td>17p13</td>
<td>17p13.3</td>
</tr>
<tr>
<td>...</td>
<td>17p13.2</td>
</tr>
<tr>
<td>...</td>
<td>17p13.1</td>
</tr>
</tbody>
</table>
Downsampling to different resolutions

Downsampling is the process of changing the representation of the data to the lower or coarser resolution.

### Complexity

How to map the $|0|0|1|$ or $|1|0|1|$ to $|0|$ or $|1|$?

1. Majority Decision Downsampling
2. OR-function Downsampling
3. Weighted Downsampling
Figure: Band in lower resolution is amplified if the majority of the bands in higher resolution is amplified. In case of a tie amplification of nearest bands are taken into consideration using “golden goal” strategy until certain number of predefined steps.
Figure: Band in lower resolution is amplified if any of the band in higher resolution is amplified.
Length Weighted Downsampling Method

**Figure:** Length of the bands are considered in this case. Band in lower resolution is amplified if the length of amplified band is greater.
Comparison of Downsampling Methods

Figure: Total number of amplifications produced by three different downsampling methods
Comparison of Downsampling Methods

**Figure:** Comparison of three different downsampling methods: The difference measure used is scaled Frobenius norm.
Mixture Modelling of Cancer

- Cancer is a collection of heterogeneous diseases
- Finite Mixture Modelling of Multivariate Bernoulli Distribution
  \[ P(x) = \sum_{j=1}^{J} \pi_j P(x|\theta_j) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i}(1 - \theta_{ji})^{1-x_i} \]
- EM algorithm to train the Mixture Models using BernoulliMix
**Model Selection: # of Components**

![Log likelihood vs Number of Components](image)

**Figure:** 10-fold Cross-Validation repeated 50 times and $J$ varied between 2-20. Example case: Chromosome: 17 resolution: 850. Similar approach to (J. Tikka et. al, 2007) and (J. Hollmén, 2007).
Visualization of Mixture Model

Figure: A visualization of one of the final model trained for chromosome-17 in resolution: 850.
## Results of Mixture Models

<table>
<thead>
<tr>
<th>Data Resolution</th>
<th>J</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original in 393</td>
<td>8</td>
<td>-3.39</td>
</tr>
<tr>
<td>Original in 850</td>
<td>8</td>
<td>-4.75</td>
</tr>
<tr>
<td>Downsampled to 393</td>
<td>6</td>
<td>-3.41</td>
</tr>
<tr>
<td>Upsampled to 850</td>
<td>6</td>
<td>-5.23</td>
</tr>
<tr>
<td>Combined in 393</td>
<td>7</td>
<td>-3.36</td>
</tr>
<tr>
<td>Combined in 850</td>
<td>7</td>
<td>-5.11</td>
</tr>
</tbody>
</table>

**Table:** Results of experiments on chromosome-17. J denotes the selected number of component distributions.
### Are Maximal Frequent Itemset Preserved?

<table>
<thead>
<tr>
<th>Resolution 400</th>
<th>Resolution 850</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frequent Itemset</strong></td>
<td><strong>Frequent Itemset</strong></td>
</tr>
<tr>
<td><code>{6,7,8}</code></td>
<td><code>{8,9,10,11,12,13,14}</code></td>
</tr>
<tr>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
<tr>
<td><strong>Chromosome Bands</strong></td>
<td><strong>Chromosomse Bands</strong></td>
</tr>
<tr>
<td><code>{17q11.2, 17q12, 17q21}</code></td>
<td><code>{17q11.2, 17q12, 17q21.1, 17q21.2, 17q21.31, 17q21.32, 17q21.33}</code></td>
</tr>
</tbody>
</table>
Something to take home about

- Downsampling and upsampling to work with various resolutions of data useful for database integration
- Mixture models of 0-1 data in different resolutions
- Effect of Resolution and Sample Size on likelihood and number of components