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

Aalto University

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

Aalto University

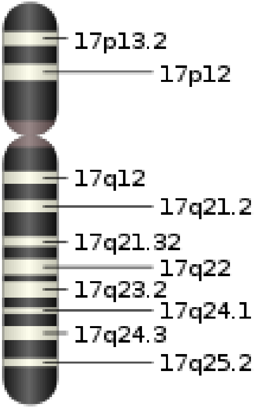
Chromosomal Aberration

- 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.

Aalto University

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)



Aalto University

Multiple Resolutions: Chromosome-17

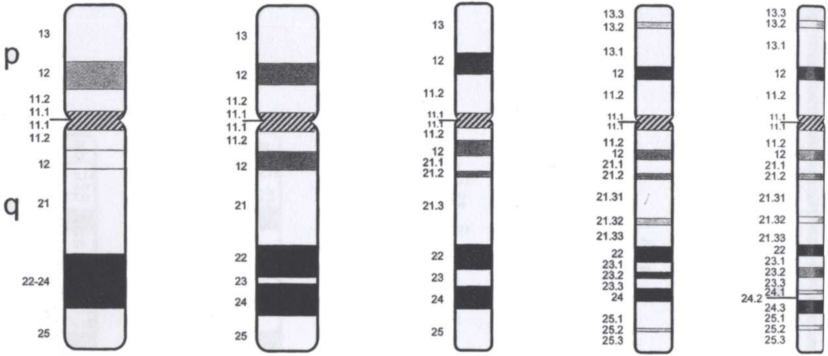


Figure: G-banding patterns for normal human chromosomes at five different levels of resolution. Source: (Shaffer et. al. 2009). Example case in Chromosome:17.

Multiple Resolutions: Part of Chromosome-17

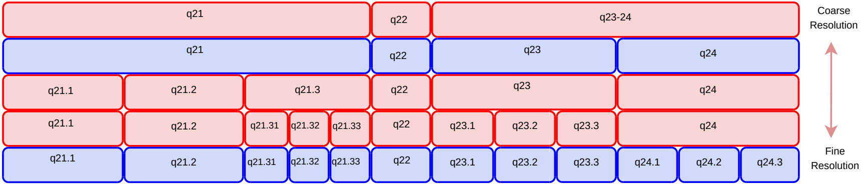


Figure: Part of chromosome 17 showing the differences in multiple resolutions.

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

(a) Resolution: 400 (b) Resolution: 850

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).

Resolution:400	Resolution:850
17p13	17p13.3
...	17p13.2
...	17p13.1

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

Aalto University

Majority Decision Downsampling method

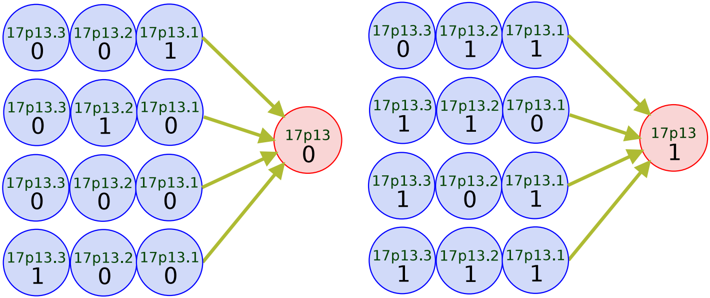


Figure: Band in lower resolution is amplified if the majority of the bands in higher resolution is amplified. In case of a tie amplification of nearest bands are taken into consideration using "golden goal" strategy until certain number of predefined steps.

OR-function Downsampling Method

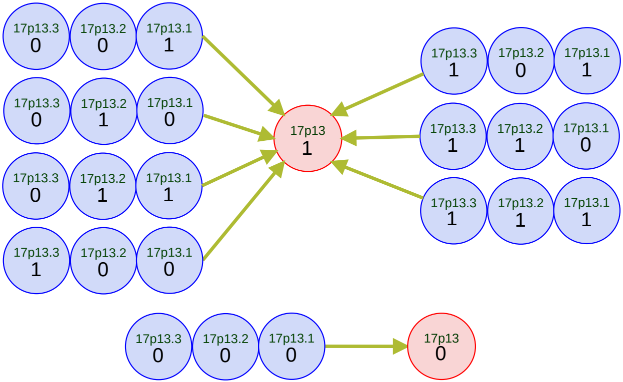


Figure: Band in lower resolution is amplified if any of the band in higher resolution is amplified.

Aalto University

Length Weighted Downsampling Method

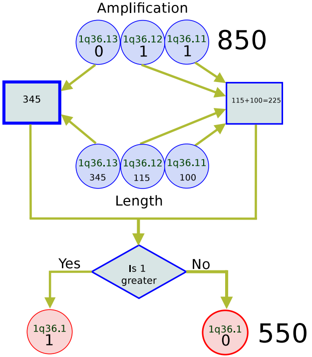


Figure: Length of the bands are considered in this case. Band in lower resolution is amplified if the length of amplified band is greater.

Comparison of Downsampling Methods

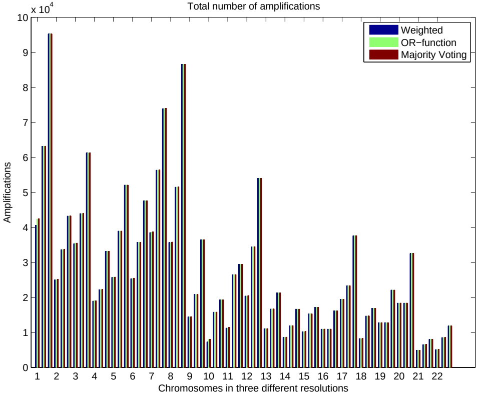


Figure: Total number of amplifications produced by three different downsampling methods

Aalto University

Comparison of Downsampling Methods

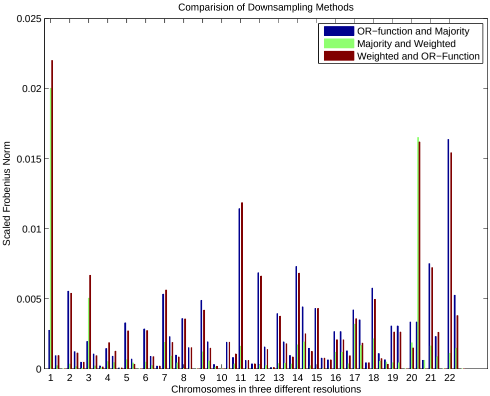


Figure: Comparison of three different downsampling methods : The difference measure used is scaled Forbenius norm.

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

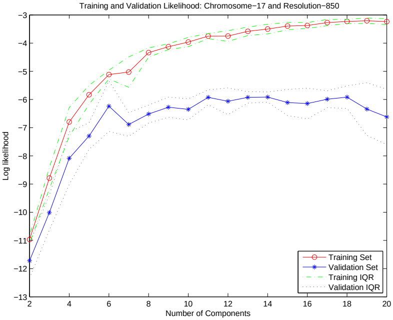


Figure: 10-fold Cross-Validation repeated 50 times and J varied between 2-20. Example case: Chromosome: 17 resolution : 850. Similar approach to (J. Tikka et. al, 2007) and (J. Hollmén, 2007).

Visualization of Mixture Model

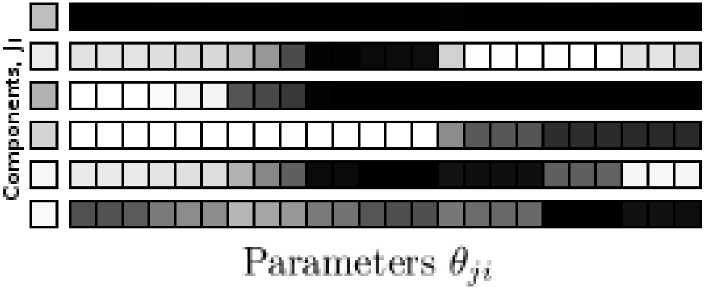


Figure: A visualization of one of the final model trained for chromosome-17 in resolution: 850.

Results of Mixture Models

Data Resolution	J	Likelihood
Original in 393	8	-3.39
Original in 850	8	-4.75
Downsampled to 393	6	-3.41
Upsampled to 850	6	-5.23
Combined in 393	7	-3.36
Combined in 850	7	-5.11

Table: Results of experiments on chromosome-17. J denotes the selected number of component distributions.

Aalto University

Are Maximal Frequent Itemset Preserved?

Resolution 400		Resolution 850
Frequent Itemset	$\Rightarrow$	Frequent Itemset
{6,7,8}	$\Rightarrow$	{8,9,10,11,12,13,14}
$\Updownarrow$		$\Updownarrow$
Chromosome Bands	$\Rightarrow$	Chromosomse Bands
{17q11.2, 17q12, 17q21}	$\Rightarrow$	{17q11.2, 17q12, 17q21.1, 17q21.2, 17q21.31, 17q21.32, 17q21.33 }

Aalto University

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

Aalto University