Preservation of Statistically Significant patterns in Multiresolution 0-1 data

Prem Raj Adhikari, Jaakko Hollmén

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

PRIB 2010
September 23, 2010

Outline

1. Introduction to Chromosomal Aberrations in Multiple Resolutions
2. Sampling Data Between Different Resolutions
3. Mixture Models of Chromosomal Aberrations
4. Are Statistically significant Patterns Preserved by Models?
5. Summary and Conclusions

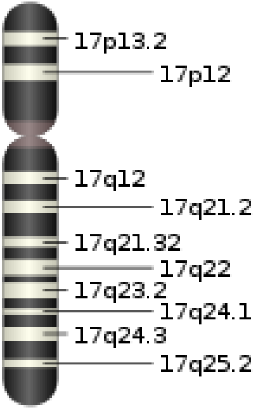
Aalto University

DNA Copy Number Aberrations

- Abnormality in the normal chromosomal content of a cell
- Different cases of DNA copy number aberrations
  - 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 five
- Why detect copy number aberrations?
- 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)
- 17q21.32: Chromosome-17, Arm-q, Region-21, Band-3 and Subband-2
- Hierarchical, irregular naming scheme



Multiple Resolutions: Chromosome-17

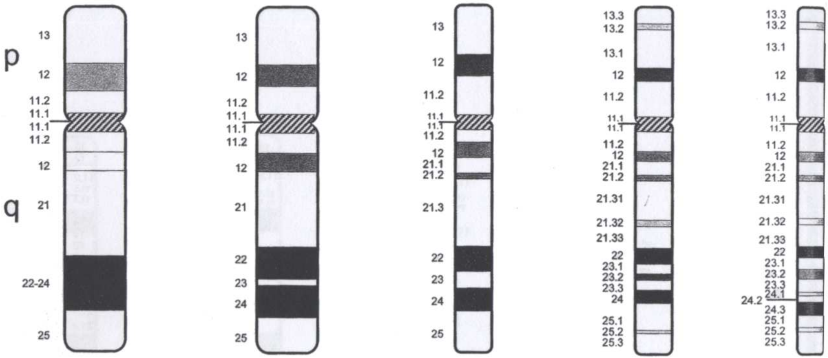


Figure: G-banding patterns for normal human chromosomes at five different levels of resolution. Source: (Shaffer et. al. 2009). Example case in Chromosome:17.

Multiple Resolutions: Chromosome-17

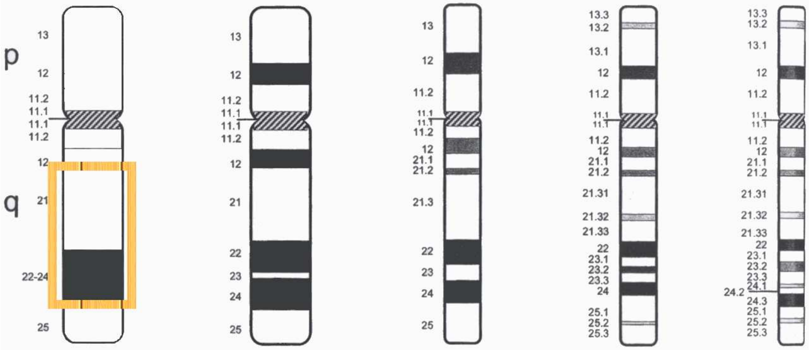
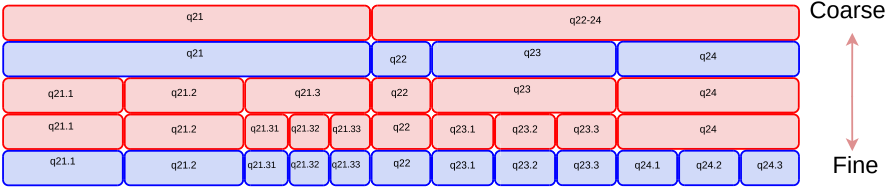


Figure: G-banding patterns for normal human chromosomes at five different levels of resolution. Source: (Shaffer et. al. 2009). Example case in Chromosome:17.

Multiple Resolutions: Part of Chromosome-17



q21 q22-24

q21 q22 q23 q24

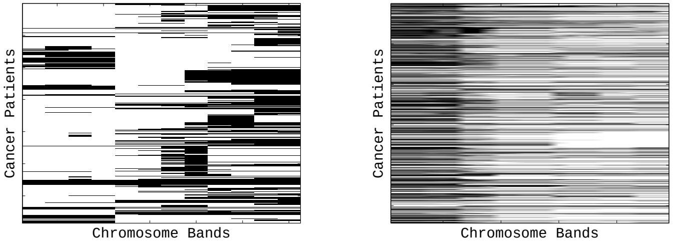
q21.1 q21.2 q21.3 q22 q23 q24

q21.1 q21.2 q21.31 q21.32 q21.33 q22 q23.1 q23.2 q23.3 q24

q21.1 q21.2 q21.31 q21.32 q21.33 q22 q23.1 q23.2 q23.3 q24.1 q24.2 q24.3

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

DNA Copy Number Amplification Dataset



(a) Resolution:400 || d=12

(b) Resolution:850 || d=24

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. Figure only shows chromosome 17.

Changing between different resolutions

Upsampling

- Upsampling is the process of changing the representation of data to the finer resolution. The dimensionality of data increases
- 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 : From fine to coarse resolution

Downsampling

- Downsampling is the process of changing the representation of the data to the coarser resolution.
- How to map the |0|0|1| or |1|0|1| to |0| or |1|? Or more generally how to map different combination of the regions, bands and sub-bands to 1 region or band?

1. Majority Decision Downsampling
2. OR-function Downsampling
3. Weighted Downsampling

Mixture Modelling of DNA Copy Number Aberrations

- Cancer is a collection of heterogeneous diseases
- Data is 0-1 data: presence or absence of chromosomal aberrations
- 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 program package (Freely available at http://users.ics.tkk.fi/~jhollmen/BernoulliMix/

Model Selection: # of Components

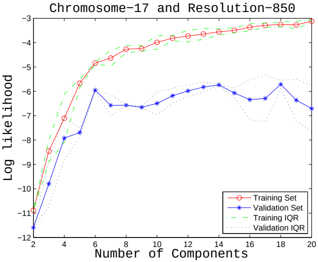


Figure: 10-fold Cross-Validation repeated 50 times and *J* varied between 2-20. Example case: Chromosome: 17 resolution : 850. Similar approach to (J. Tikka et. al, 2007) and (J. Hollmén, 2007).

Visualization of Mixture Model

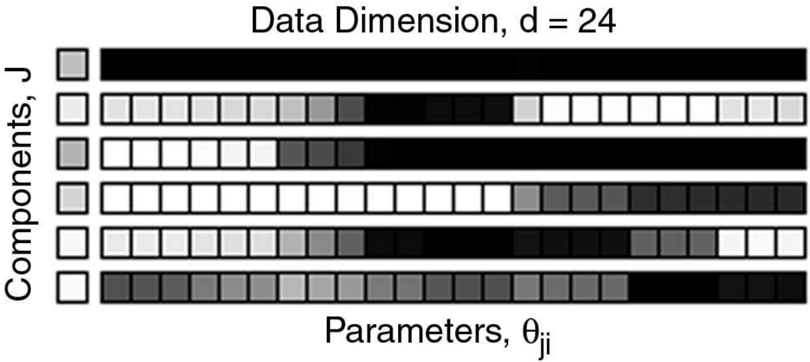


Figure: A visualization of one of the final model trained for chromosome-17 in resolution: 850.

Statistical Significance Testing

- Data, $\mathcal{D}$, belongs to the class of Empirical Distribution (PDF is unknown)
- Fix a Null distribution and sample the data, $\mathcal{D}_i$, from Null distribution using MCMC (Besag et. al. 1989)
- Randomization strategy for 0-1 matrix according to Gionis et. al. (2007) using swaps
- Observe the test statistic, $\mathcal{A}$, on the orginal data and the data sampled from null distribution
- Empirical Monte Carlo p-value

$$\tilde{p} = \frac{1}{n+1} \left(\sum_{i=1}^{n} I(\mathcal{A}(\mathcal{D}_i) \geq \mathcal{A}(\mathcal{D})) + 1 \right)$$

Experimental Procedure

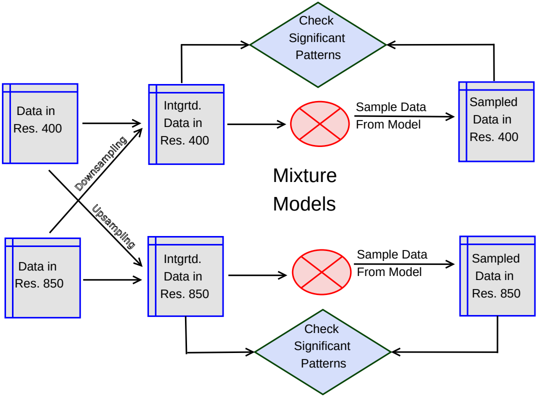


Figure: The experimental procedure.

Are Frequent Patterns Preserved?

Original Data: Resolution 400		**Sampled from Model: Resolution 400**
Before Database Integration		
Frequent Itemset	⇒	Frequent Itemset
{9,10}, {11,12}	⇒	{9,10}, {11,12}
After Database Integration		
Frequent Itemset	⇒	Frequent Itemset
{ 5, 7}, { 5, 12}, Subset of cardinality 2 of {8,9,10,11,12}	⇒	{ 5, 7},{ 5, 12},{ 7, 12}, Subset of cardinality 2 of {8,9,10,11,12}

Summary and Conclusions

- Downsampling and upsampling to work with various resolutions of data useful for database integration
- Mixture models of 0-1 data in different resolutions
- Statistical significance testing using randomization
- Patterns are preserved in our modelling approach