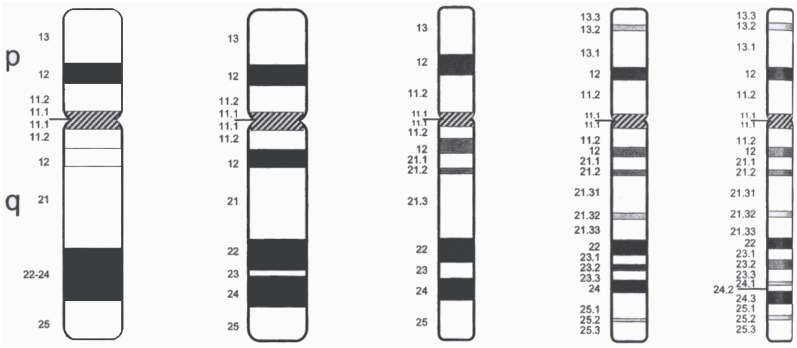
# PRESERVATION OF STATISTICALLY SIGNIFICANT PATTERNS IN MULTIRESOLUTION 0-1 DATA

Prem Raj Adhikari ( prem.adhikari@tkk.fi ) and Jaakko Hollmén ( jaakko.hollmen@tkk.fi )
Aalto University School of Science and Technology, Espoo, Finland



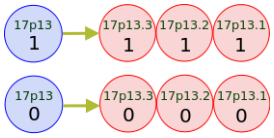

G-banding patterns for normal human chromosomes at five different levels of resolution. Source: (Shaffer et. al. 2009). Example case in Chromosome:17. Division of regions is not consistent and different for different regions.

- Biological experiments performed with high throughput and high resolutions techniques often produce data in multiple resolutions.
- International System for Human Cytogenetic Nomenclature(ISCN) has defined five different resolutions of the chromosome bands: 300, 400, 550, 700 and 850.
- Same chromosome is divided into different regions in different resolution.
- Typically computational algorithms work with a single resolution of data.

## THEORETICAL FRAMEWORK

### UPSAMPLING

Transforming the data resolution to finer resolution. Dimensionality of data increases.

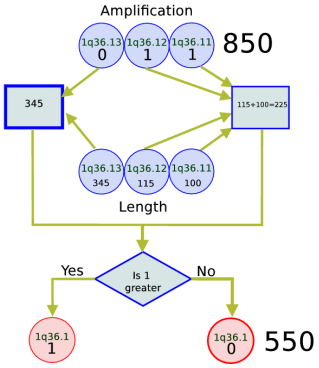



Duplicate copies of similar cytogenetic bands is made in the finer resolution.

### DOWNSAMPLING

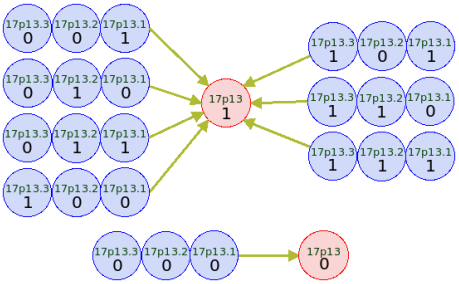
Transforming the data resolution to coarser resolution. The dimensionality of the data decreases.

#### 1. WEIGHTED DOWNSAMPLING



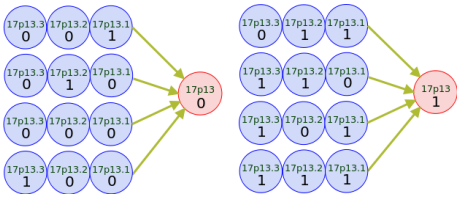

Cytogenetic band in coarse resolution is amplified if total length of the amplified bands is greater than that of unamplified bands in fine resolution.

#### 2. OR-FUNCTION DOWNSAMPLING




Cytogenetic band in coarse resolution is amplified if any of the bands in fine resolution is amplified.

#### 3. MAJORITY DECISION DOWNSAMPLING



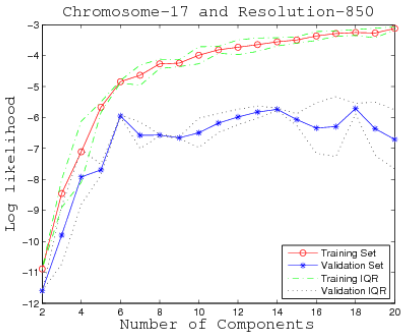

Cytogenetic band in coarse resolution is amplified if majority of the bands in fine resolution are amplified.

**Conflict/Ties** In case of a tie amplification of nearest bands are taken into consideration using "golden goal" strategy until certain number of predefined steps. If tie can not be concluded with "golden goal" strategy then the band is coarse resolution is deemed as amplified.

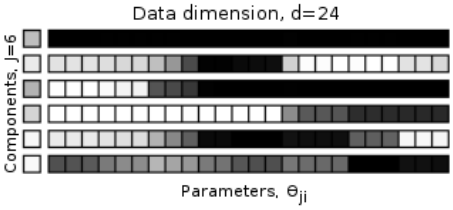
### MIXTURE MODELS

$$p(\mathcal{D}|\Theta) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i} (1-\theta_{ji})^{1-x_i}$$

where $\pi_j$ are the mixture proportions and $\Theta$ is composed of $\theta_{j1}, \theta_{j2}, \theta_{j3} \ldots \theta_{jd}$ where $j = 1, 2 \ldots J$



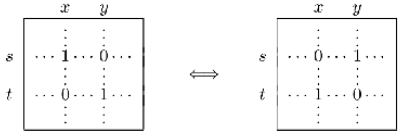

Example case of model selection for chromosome-17 in resolution 850. Number of components selected in this case is 8.



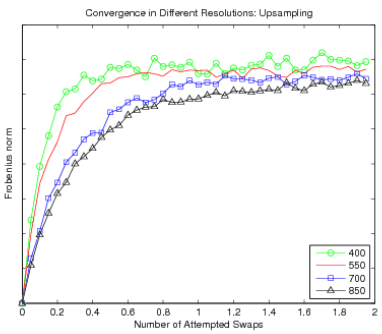

Visualization of one of the final model trained for chromosome-17 in resolution: 850.

## EXPERIMENTS AND RESULTS

### SWAP RANDOMIZATION



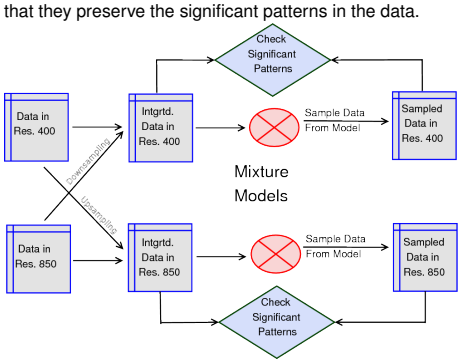

Schematic representation of a swap in a 0-1 matrix: Source: Gionis et.al. 2007.




Convergence of 0-1 swap for different resolutions. Number of attempted swaps is multiple of $10^5$.

### EXPERIMENTAL PROCEDURE

**Our Focus:** Generate maximally simple or compact (parsimonious) models for chromosomal aberrations such that they preserve the significant patterns in the data.




Schematic representation of the experimental procedure.

### EXAMPLE RESULTS

| Original: Resolution 400 | | Sampled: 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} |

### REFERENCES

P. R. Adhikari, J. Hollmén. Patterns from Multiresolution 0-1 data. In Proceedings of the ACM SIGKDD Workshop on Useful Patterns (Washington, DC, July 25 - 25, 2010). UP '10. ACM, New York, NY, 8-16, 2010.

J. Tikka, J. Hollmén and S. Myllykangas, Mixture modeling of DNA copy number amplification patterns in cancer, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4507 LNCS, pp. 972-979, 2007.

S. Myllykangas, J. Himberg, T. Böhling, B. Nagy, J. Hollmén and S. Knuutila, DNA copy number amplification profiling of human neoplasms, Oncogene, 25 (55), pp. 7324-7332, 2006.

A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. ACM Transactions on Knowledge Discovery from Data, 1(3):14, 2007.

L.G. Shaffer and N. Tommerup. ISCN 2009: An International System for Human Cytogenetic Nomenclature(2005) Recommendations of the International Standing Committee on Human Cytogenetic Nomenclature. Karger, 2009.

Aalto University School of Science and Technology

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

ALGODAN