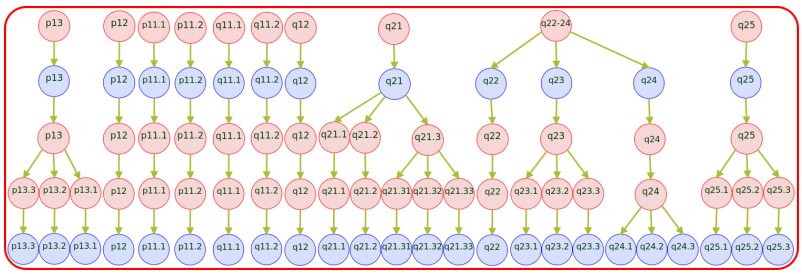
PROBABILISTIC TRANSFORMATION AND MODELLING OF MULTIRESOLUTION 0-1 DATA

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



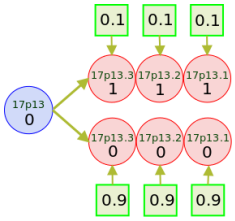
Representation of five different resolutions of chromosomal aberrations as a collection disjoint graphical models.

- Hierarchical and irregular scheme of chromosome nomenclature proposed by ISCN is such that a chromosome band in one resolution can be associated different number of chromosome bands in another resolution.
- Two different DNA Copy Number Aberrations datasets were available in resolutions 400 and 850.
- Matching samples is a problem because absence of knowledge of overlapping samples and difference in dimensionality.

METHODOLOGY

UPSAMPLING

Transforming data to finer resolution increasing dimensionality.



Upsampling is performed using Potts Model.

$$p(x = i|y = j) = \begin{cases} \frac{j}{j} & \text{with probability } 0.9 \\ & \text{with probability } 0.1 \end{cases}$$

DOWNSAMPLING

Transforming the data to coarser resolution decreasing dimensionality.

DATA PREPROCESSING

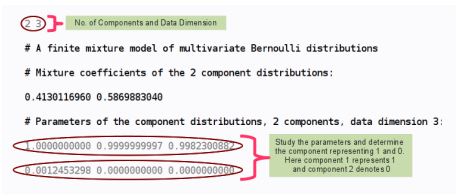
– Downsample and upsample the data.
–Calculate two Frobenius norm between datasets in two resolutions.
–Average the Frobenius norm and select the unique match for data in two resolutions.
–Merge the data where the equal number of bands in fine resolution combine to form a single band in coarse resolution

1. NAIVE BAYES CLASSIFIER

Naive Bayes Classifiers are probabilistic classifiers that assign most likely class to a given example explained by its feature vector using Bayes theorem.

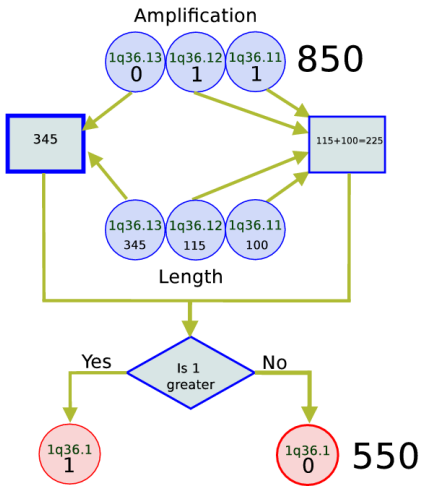
$$C_i = \arg\max_{C_i} p(C) \prod_{i=1}^{n} p(X_i|C)$$

2. CLUSTERING MIXTURE MODELS



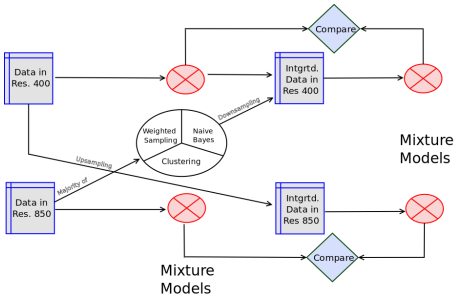
The cluster indices obtained after clustering using mixture models of finite mixtures of multivariate Bernoulli distributions are used as the class labels.

3. 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.

EXPERIMENTAL PROCEDURE

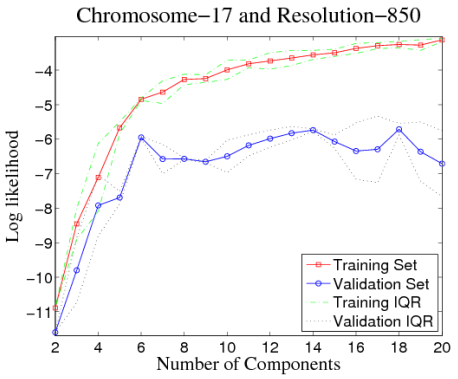


Schematic representation of experimental procedure.

MIXTURE MODELS

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

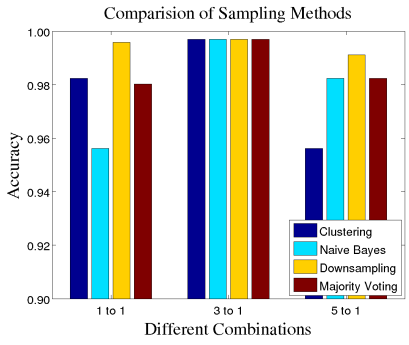
where π_j are the mixture proportions and Θ is composed of $\theta_{j1}, \theta_{j2}, \theta_{j3} \ldots \theta_{jd}$ where $j = 1, 2 \ldots J$



Example case of model selection for chromosome 17 in resolution 850. Number of components selected in this case is 6.

RESULTS

CLASSIFICATION ACCURACY

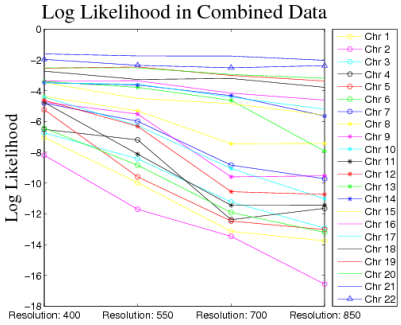


Accuracy of different classifiers.

–Data transformation methods are consistently better
–Majority voting is the second.
–Data transformation methods may overfit the datasets
–Majority voting reduces the bias
–Produces reliable estimate for the aberration patterns in coarse resolution

OVERALL RESULT

Our Focus: Train parsimonious mixture models for chromosomal aberrations in each chromosome.



Schematic representation of the experimental procedure.

RESULTS ON CHROMOSOME 17

Data Resolution	J	Likelihood in	
		Original	Resampled
Original in 400(A)	6	-3.70	-3.32
Original in 850(B)	8	-4.57	-4.66
Downsampled to 400 from B(C)	7	-3.28	-3.26
Upsampled to 850 from A(D)	8	-4.72	-4.30
Combined in 400(A+C)	6	-3.49	-3.49
Combined in 850(B+D)	6	-5.69	-5.61

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.

P. R. Adhikari, J. Hollmén. Preservation of Statistically Significant patterns in Multiresolution 0-1 data. In Tjeerd Dijkstra, Evgeni Tsivtsivadze, Elena Marchiori, and Tom Heskes, editors, Proceedings of the 5th IAPR International Conference on Pattern Recognition in Bioinformatics , Volume 6282 of Lecture Notes in Computer Science, pages 86-97 Springer-Verlag, September 2010, Nijmegen, The Netherlands.

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
HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY
ALGODAN