Aalto University
School of Science

Mixture Models from Multiresolution 0-1 data

Prem Raj Adhikari[1,2], Jaakko Hollmén[1,2]

Parsimonious Modeling Research Group in
[1]Department of Information and Computer Science
Aalto University School of Science, Finland
[2]Helsinki Institute for Information Technology (HIIT)
{prem.adhikari,jaakko.hollmen}@aalto.fi
http://users.ics.aalto.fi/padhikar/
Proceedings Pages: **1–16**

Management Summary

- Motivation for the Work
- Multiresolution Data
- Mixture modelling of multiresolution data
- Summary and Conclusions

Modelling : the general perspectives

> *Modeling is like vintage wine; it matures with time.*
>
> — DECISIONCRAFT INC.
>
> *www.decisioncraft.com*

Modelling : the general perspectives

> *Modeling is like vintage wine; it matures with time.*
>
> — DECISIONCRAFT INC.
>
> *www.decisioncraft.com*

> *People may mature with time but models mature only with increasing data.*
>
> — PREM RAJ ADHIKARI
>
> *PhD Student*

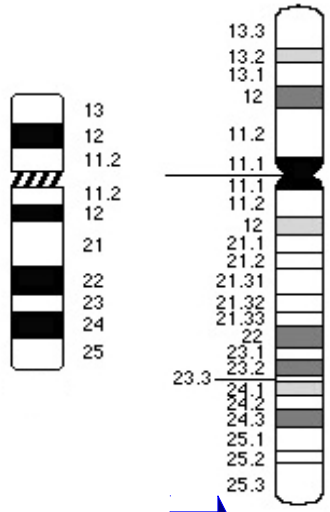
Importance of Using More Samples

The Square Root Law

$$\text{Accuracy of Information} = \sqrt{\text{Volume of Information}}$$

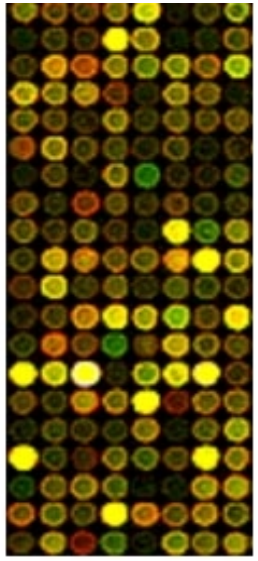
The Multiresolution data

Dimension
In hundreds



Banding

Dimension
In thousands



Microarrays

Dimension
In millions

Next Generation
Sequencing

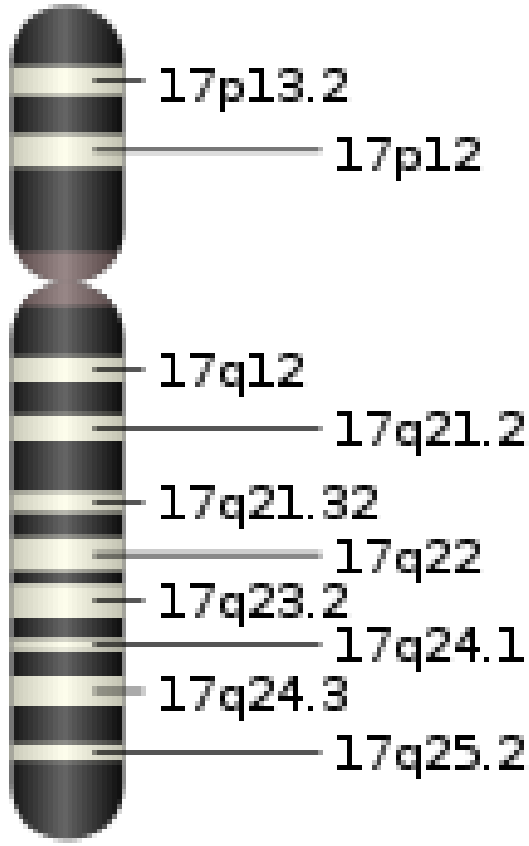
- Multiresolution data is everywhere: biology, computer vision, telecoms ...
- Older Generation Technology $\Rightarrow$ Data in Coarse Resolution
- Newer Generation Technology $\Rightarrow$ Data in Fine Resolution

Chromosomal Aberrations in Cancer

- Abnormality in the normal chromosomal content of a cell
- Different cases of DNA copy number aberrations
  - Deletion: When the copy number $<$ **2**
  - Duplication: When the copy number $>$ **2**
  - Amplification: When the copy number $\gg$ **5**
- 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)
- 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

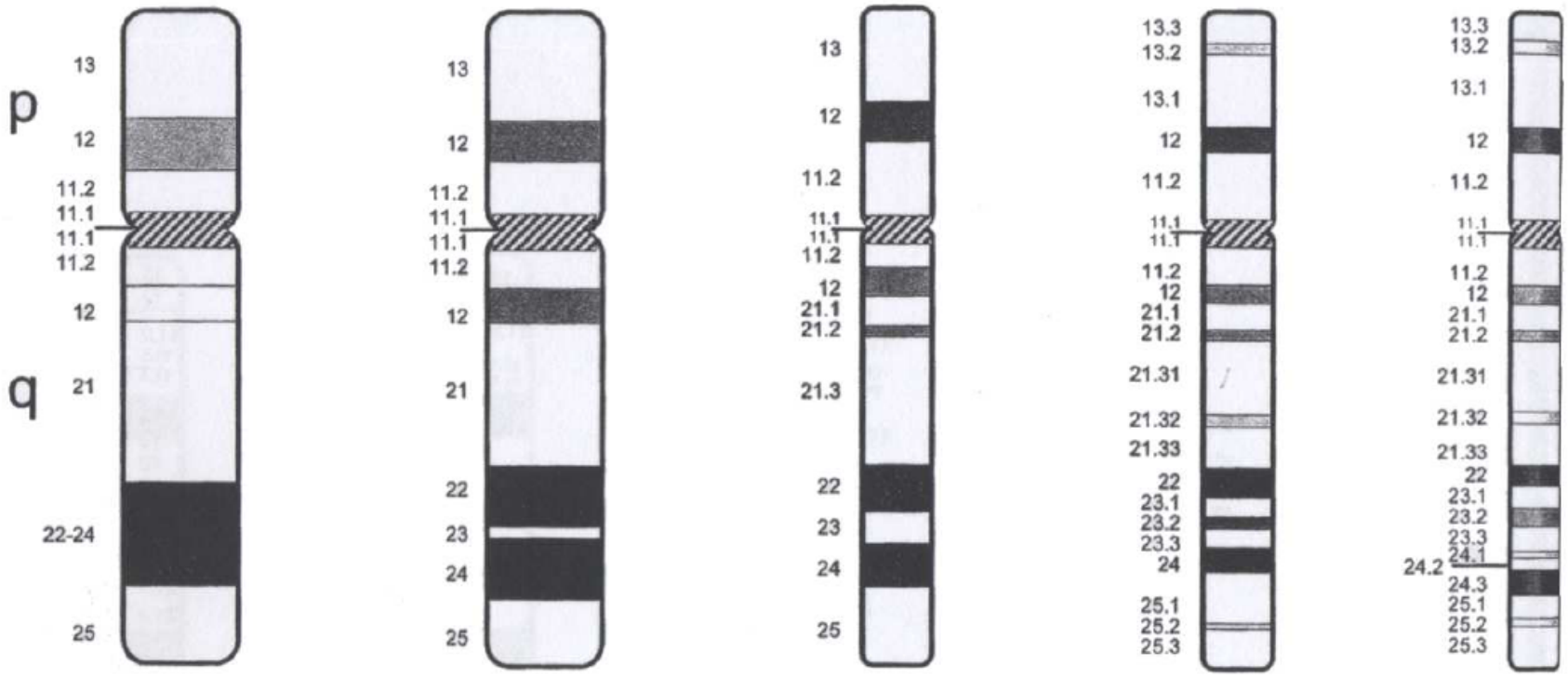
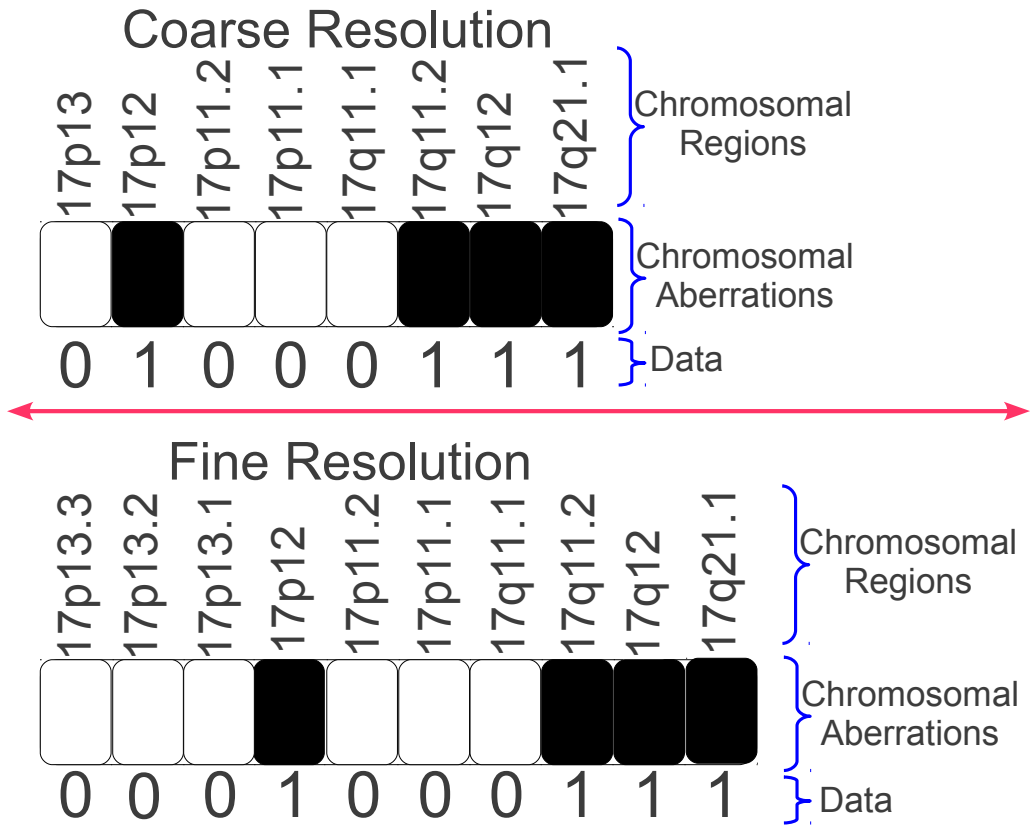


Figure : G-banding patterns for normal human chromosomes at five different levels of resolution. Source: (Shaffer et. al. 2009). Example case in Chromosome:17.

Multiresolution Data in Cancer Genomics



Finite Mixture Modeling 0–1 Data

- Why Mixture Models?
- Cancer is a heterogeneous collection of several diseases and mixture models are well known for their ability to model heterogeneity

  $P(x) = \sum_{j=1}^{J} \pi_j P(x|\theta_j) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i} (1 - \theta_{ji})^{1-x_i}$

- Mixture models are probabilistic and clustering capabilities
- Mixture models can be easily learned using Expectation Maximization (EM) algorithm
- Open–source BernoulliMix program package to learn mixture models available from the authors
  http://users.ics.aalto.fi/jhollmen/BernoulliMix/

Mixture Modeling of Multiresolution 0–1 Data

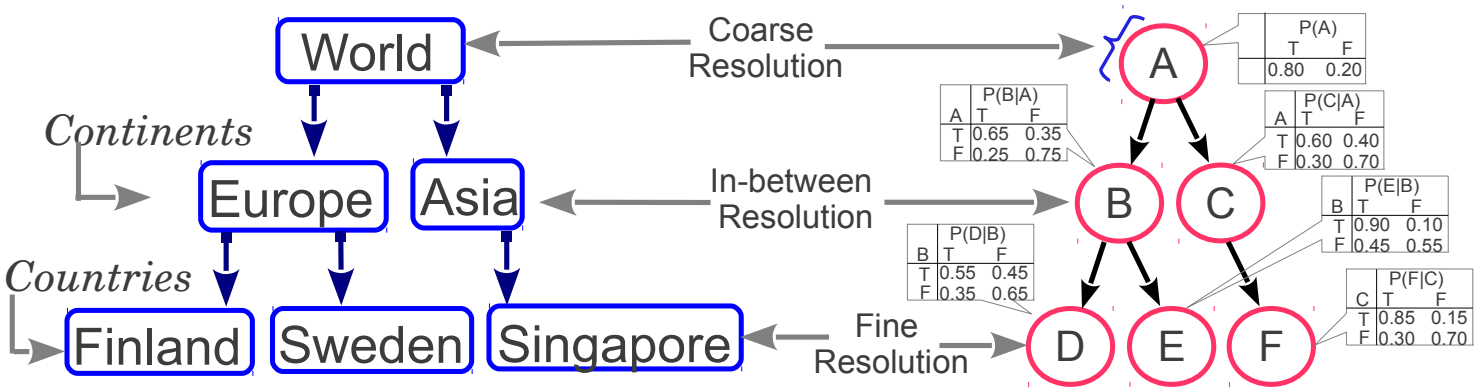
- Mixture models generally cannot model multiresolution data

$$P(x) = \sum_{j=1}^{J} \pi_j P(x|\theta_j) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i} (1 - \theta_{ji})^{1-x_i}$$

i is different for each resolution and requires different models for each resolution

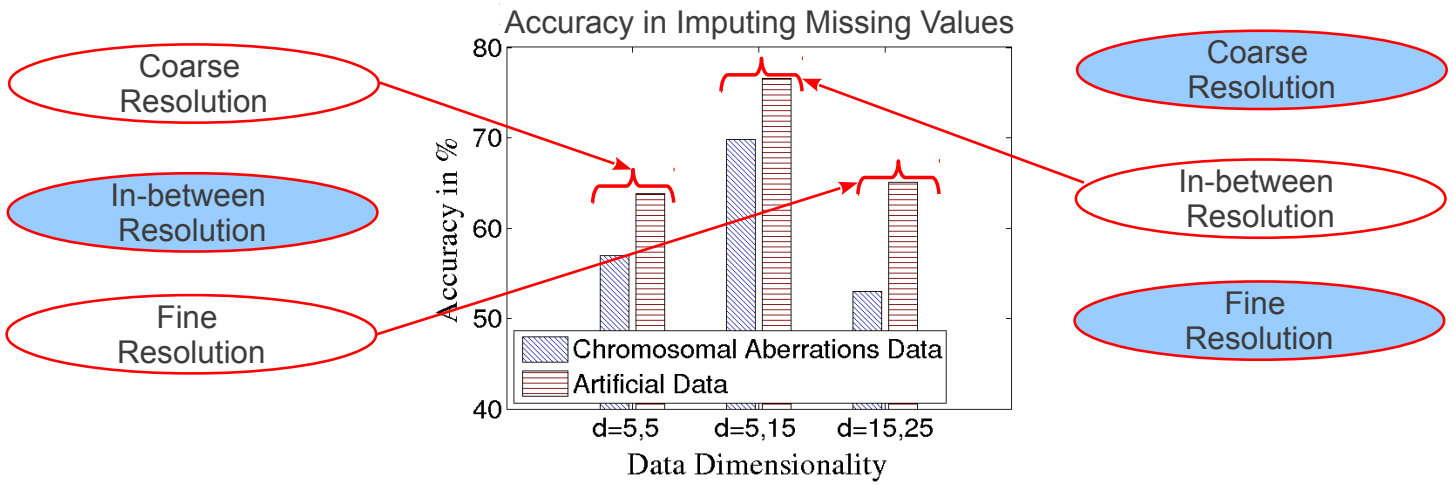
- Only mixture modeling solution to multiresolution data is to model each resolution separately
- Multiresolution data can be transformed to single resolution for mixture modeling (Adhikari & Hollmén, 2010)
- Model the multiresolution data by modelling the interactions between the models in different resolutions (Adhikari & Hollmén, 2012)

Multiresolution Mixture Components



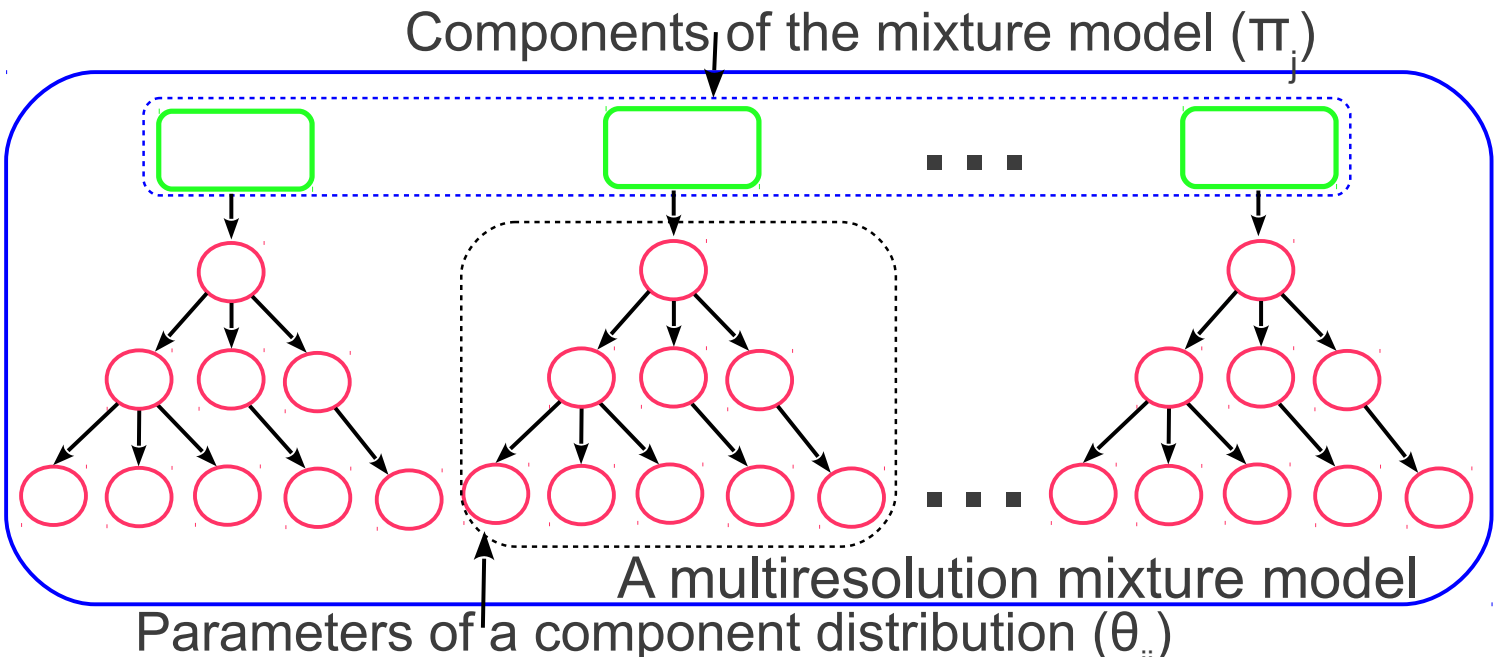
- Domain ontology provides information about relationships between features in different resolutions
- We can create a tree structure where features in the coarse resolution form the root and features in the fine resolution leaves of the tree

Bayesian Networks to Impute Missing Resolutions



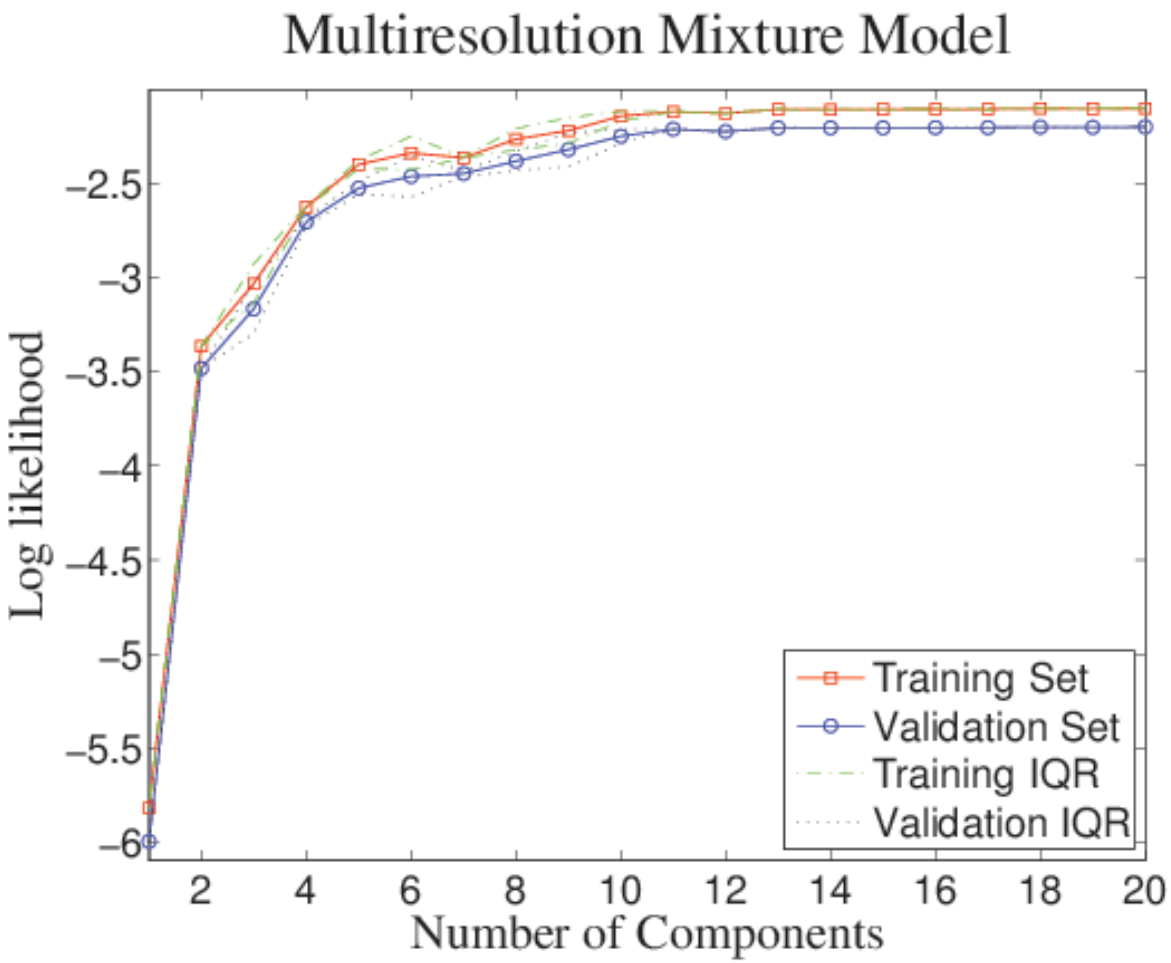
- Bayesian networks can be used to impute missing resolutions using marginal inference.
- For a joint distribution P(A,B,C) and an evidence B=true, marginal inference calculation is:
  $P(A \mid B = true) \propto \sum_C P(A, B = true, C).$

Structure of Mixture Model



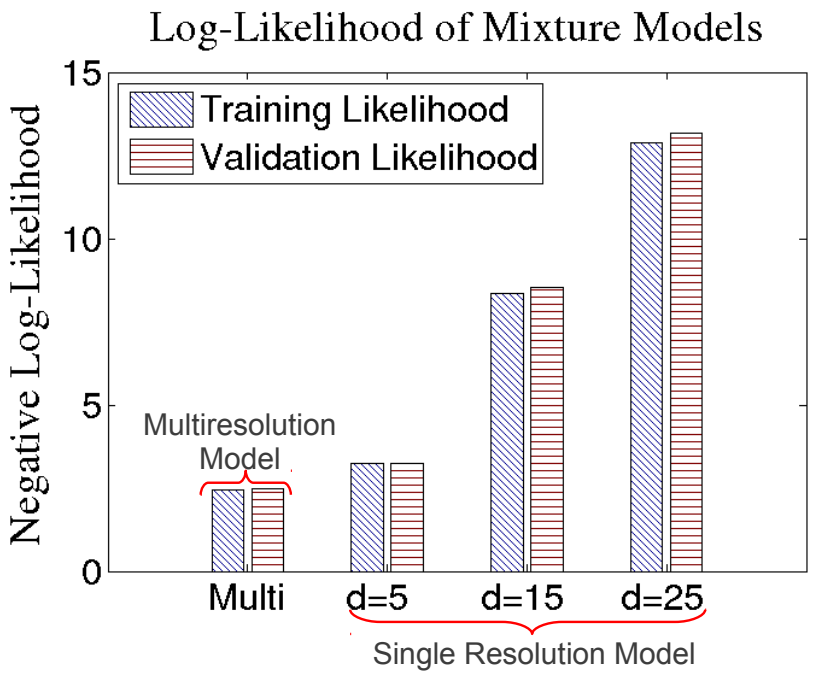
- The components of the mixture model are Bayesian networks themselves
- Now, the problem is to learn the parameters $\boldsymbol{\Theta} = \{J, \{\pi_j, \theta_j\}_{j=1}^{J}\}$

Model Selection in Mixture Model



P. R. Adhikari, J. Hollmén, DS'2012, **Fast Progressive Training of Mixture Models for Model Selection.**

Multiresolution Mixture Model Results



- The Y-axis shows the negative log likelihood, therefore, the shorter the bar, better the result
- The multiresolution model outperforms single resolution models

Summary and Conclusions

- Mixture Modeling of Multiresolution 0–1 Data in three ways:
  - Data Transformation
  - Merging of mixture components
  - Bayesian network as component distributions
- Experiments on multiresolution chromosomal datasets
- Experiments show that multiresolution models outperform single resolution models

Questions, Comments, Feedback and Acknowledgment

? ! ”

The work is funded by Helsinki Doctoral Programme in Computer Science–Advanced Computing and Intelligent Systems (Hecse)