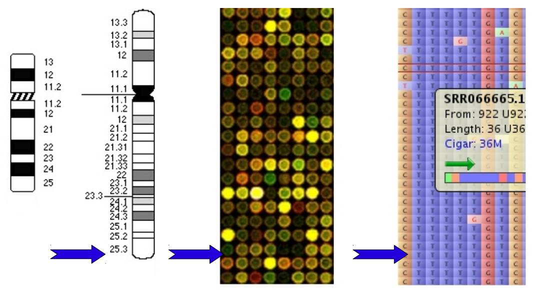
MULTIRESOLUTION MIXTURE MODELLING USING MERGING OF MIXTURE COMPONENTS

Prem Raj Adhikari[1,2] and Jaakko Hollmén[1,2] , {prem.adhikari, jaakko.hollmen}@aalto.fi

[1]Aalto University School of Science, and [2]Helsinki Institute for Information Technology,
Department of Information and Computer Science, PO Box 15400, FI-00076 Aalto, Espoo, Finland

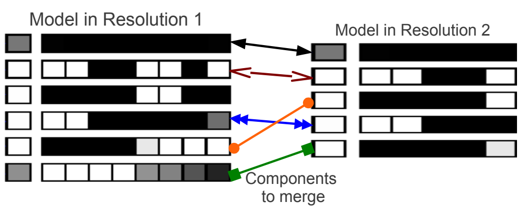


Banding   Micro-arrays   Next Generation Sequencing

MULTIRESOLUTION DATA

- Multiresolution data arise when an object or a phenomenon is described at several levels of detail
- Multiresolution data is prevalent in many application areas
  - ★ Examples include biology, computer vision
- Faster growth of multiresolution data is expected in future
- Over the years, data accumulates in multiple resolutions because
  - ★ Older Generation Technology ⇒ Data in Coarse Resolution
  - ★ Newer Generation Technology ⇒ Data in Fine Resolution
- How to analyze data in multiple resolutions i.e. dimensions?

MERGING OF MIXTURE COMPONENTS



Merge the mixture components as:

$$\pi_{merged} = \frac{\pi_{klmin,1} + \pi_{klmin,2} + \ldots + \pi_{klmin,n}}{n}$$

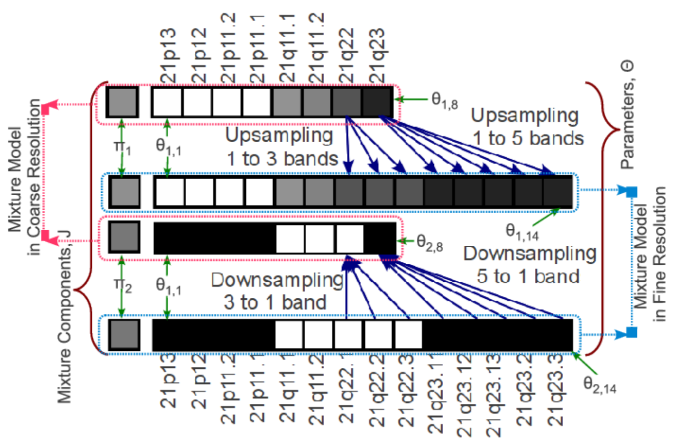
Merge the parameters according to the weight of component distributions:

$$\Theta_{merged} = \frac{\pi_{klmin,1} \times \Theta_{klmin,1} + \pi_{klmin,2} \times \Theta_{klmin,2} + \ldots + \pi_{klmin,n} \times \Theta_{klmin,n}}{\pi_{klmin,n} + \pi_{klmin,2} + \ldots + \pi_{klmin,n}}$$

Normalize the components in the model as:

$$\pi_j = \frac{\pi_j}{\sum_{j=1}^{J} \pi_j}$$

SAMPLING OF MODEL PARAMETERS



The model parameters denote the regions of chromosome. The unchanged chromosomal regions across different resolutions are not altered. The regions with changes from the coarse resolution and downsampled from the fine resolution according to the division of the chromosomal regions across different resolutions.

KULLBACK LEIBLER DIVERGENCE IN MIXTURE MODEL

In a mixture model, the KL divergence between two mixture components can be derived to

$$KL_{\theta\beta} = \sum_{i=1}^{2^d}\left[\left\{\prod_{k=1}^{d}\left(\theta_k^{X_{ik}}(1-\theta_k)^{(1-X_{ik})}\right) - \prod_{k=1}^{d}\left(\beta_k^{X_{ik}}(1-\beta_k)^{(1-X_{ik})}\right)\right\} \cdot log \prod_{k=1}^{d} \frac{\theta_k^{X_{ik}}(1-\theta_k)^{(1-X_{ik})}}{\beta_k^{X_{ik}}(1-\beta_k)^{(1-X_{ik})}}\right]$$

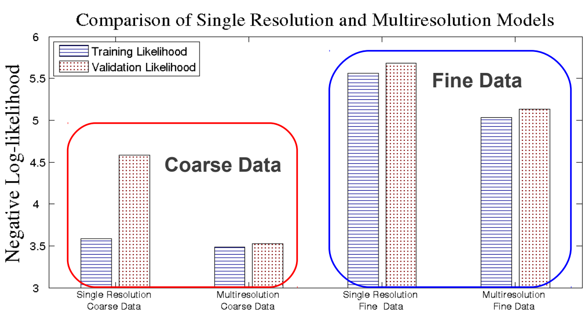
We derive data driven approximation of KL divergence in two models in different resolutions:

$$KL = \sum_{i \in X^*} \pi_\alpha \prod_{m=1}^{d}\left(\alpha_m^{X^*_{im}}(1-\alpha_m)^{(1-X^*_{im})}\right) - \sum_{i\prime \in Y^*} \pi_\beta \prod_{n=1}^{d'}\left(\beta_n^{Y^*_{i\prime n}}(1-\beta_n)^{(1-Y^*_{i\prime n})}\right)$$

APPROXIMATIONS USED

- ♦ Dropping the log-term : $log\frac{0}{0} \approx 0$
- ♦ Using only unique samples in the data instead of full state-space
- ♦ Approximating state-space by unique samples $X^* = \{x^* : x^* \in \overline{X}\}$ provides data driven approach of approximation of KL divergence

PERFORMANCE OF MULTIRESOLUTION MODELS



Multiresolution model is considerably better than single resolution model.

REFERENCES

1. P. R. Adhikari, and J. Hollmén. Multiresolution Mixture Modeling using Merging of Mixture Components. In *Proceedings of ACML 2012*, Volume 25 of Journal of Machine Learning Research - Proceedings Track, Singapore, November 4–6, 2012.

2. P. R. Adhikari, and J. Hollmén. Fast Progressive Training of Mixture Models for Model Selection. In *Proceedings of DS 2012*, Volume 7569 of Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin Heidelberg, pages 194-208, October 29-31, 2012, Lyon, France.

3. N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. SMEM Algorithm for Mixture Models. *Neural Computation*, 12(9):2109–2128, 2000.

4. S. Myllykangas, J. Tikka, T. Boöhling, S. Knuutila, and J. Hollmén. Classification of human cancers based on DNA copy number amplification modeling. *BMC Medical Genomics*, 1(15), May 2008.

5. J. Tikka, J. Hollmén, and S. Myllykangas. Mixture Modeling of DNA copy number amplification patterns in cancer. In *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN 2007)*, Volume 4507 of Lecture Notes in Computer Science, pages 972–979, 2007.

Aalto University School of Science   HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY   ALGODAN Algorithmic Data Analysis

ACKNOWLEDGEMENT

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