On Models, Patterns and Prediction

Jaakko Hollmén

Helsinki Institute for Information Techhnology
Aalto University, Department of Computer Science
Espoo, Finland

e-mail: Jaakko.Hollmen@aalto.fi

Invited talk in the 5th International Workshop on New Frontiers in Mining Complex Patterns at the ECMLPKDD 2016 in Riva del Garda, Italy
September 19, 2016

Overall theme of the talk

Interaction between:

- Probability distributions
- Patterns
- Prediction

Interaction of distributions and patterns

Based on a publication by the authors:

- Jaakko Hollmén, Jouni K. Seppänen, and Heikki Mannila. Mixture models and frequent sets: combining global and local methods for 0-1 data. In Daniel Barbara and Chandrika Kamath, editors, Proceedings of the Third SIAM International Conference on Data Mining, pages 289–293. Society of Industrial and Applied Mathematics, 2003. `http://dx.doi.org/10.1137/1.9781611972733.32`

Introduction

Two Traditions of Data Mining:

- Approximating the joint distribution (global)
- Technology of fast counting (local)

We study the interaction of global and local techniques
Questions:

- How can be benefit from the combination of global and local techniques?
- Are frequent itemsets extracted from clustered data different from globally extracted frequent itemsets? How different? How to measure?
- What is the information content in such frequent set collections?

Frequent Sets and Deviation

Compare two collections of frequent sets:

- Frequent set collection $\mathcal{F}_1$
- Frequent set collection $\mathcal{F}_2$

We define a dissimilarity measure *deviation*:

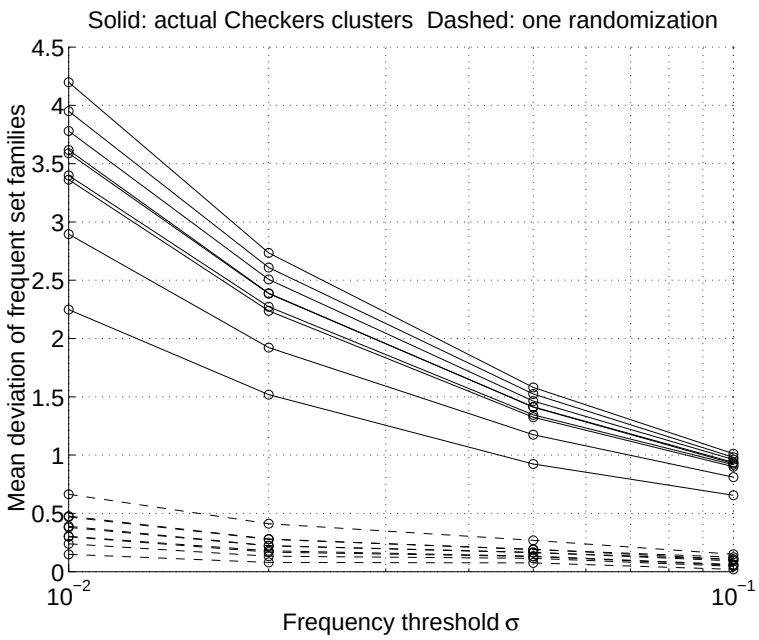
$$d(\mathcal{F}_1, \mathcal{F}_2) = \frac{1}{|\mathcal{F}_1 \cup \mathcal{F}_2|} \sum_{I \in \{\mathcal{F}_1 \cup \mathcal{F}_2\}} |f_1(I) - f_2(I)|.$$

Here, we denote by $f_j(I)$ the frequency of the set I in $\mathcal{F}_j$, or σ if $I \notin \mathcal{F}_j$. The deviation is in effect an L_1 distance where missing values are replaced by σ.

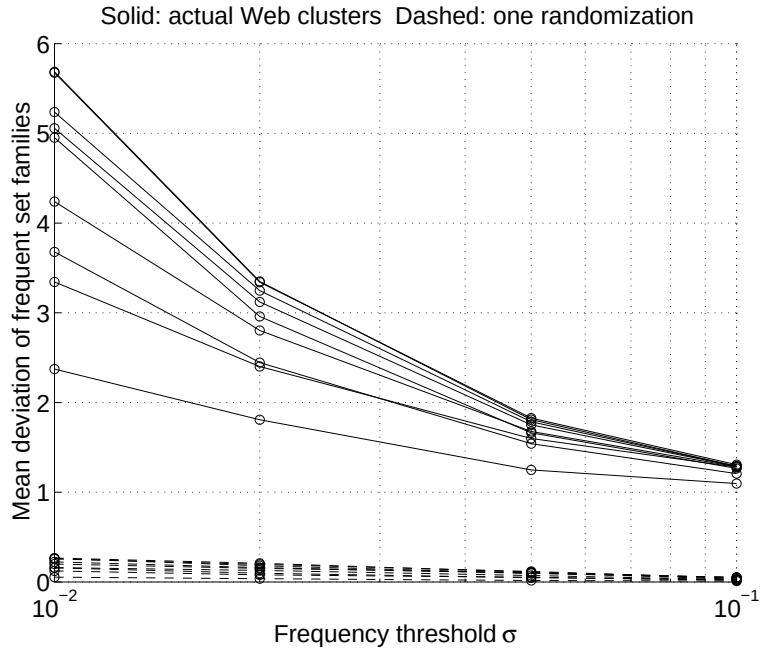
Frequent Sets in Clusters

Compare frequent sets with $d(\mathcal{F}_1, \mathcal{F}_2)/\sigma$

- Frequent set collection $\mathcal{F}_1$
- Frequent set collections from clusters $\mathcal{F}_2$



(checker)



(Web data)

Frequent sets extracted from partitioned data are markedly different

Comparing Distributions (1/2)

What is the information content in the frequent sets extracted from partitioned data? Compare distributions approximated on the basis of frequent sets.
Maximum Entropy Distribution $g(\mathbf{x})$

- satisfies frequencies of the frequents sets
- maximum entropy solution
- explicit representation with 2^d parameters
- iterative scaling algorithm

Comparing Distributions (2/2)

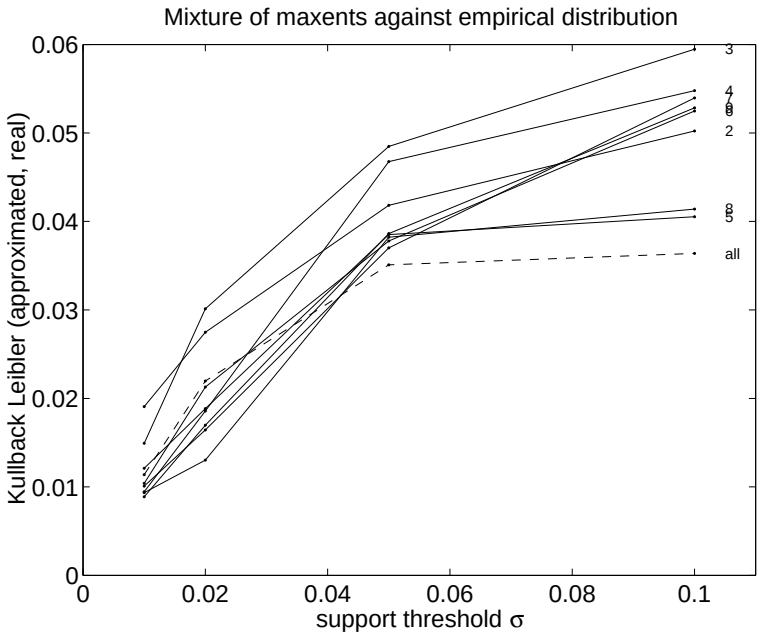
Estimate $g_j(\mathbf{x})$ from frequent sets of cluster j and mix to get a Mixture of Maximum Entropy Distributions:

$$g(x) = \sum_{j=1}^{J} \hat{P}(\mathbf{x} \in j) g_j(\mathbf{x})$$

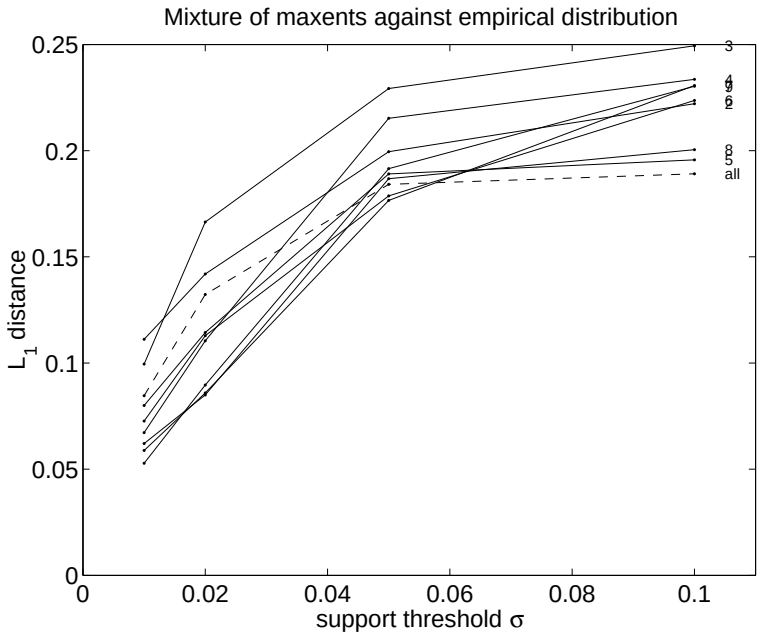
Measure the difference from the the empirical distribution $f(\mathbf{x})$ with

- L_1 distance: $\sum_x |g(\mathbf{x}) - f(\mathbf{x})|$
- Kullback-Leibler measure:
  $E_g[\log(g/f)] = \sum_x g(\mathbf{x}) \log(g(\mathbf{x})/f(\mathbf{x}))$

Comparing Distributions



(checker, K-L)



(checker, L1)

Summary and Conclusions

We study the interaction between global and local techniques in data mining

- Combined use of frequent sets and probabilistic clustering with multivariate 0-1 data
- Define a dissimilarity measure between collections of frequent sets
  - Frequent sets extracted from clusters are markedly different from globally extracted frequent sets
- Use the frequent sets from clusters to define a mixture of maximum entropy distributions
  - Measure the difference from the empirical distribution (L_1 and K-L)

Multiresolution pattern mining

Based on the following publications:

- Prem Raj Adhikari, 2014. Probabilistic Modelling of Multiresolution Biological Data. Doctoral Dissertation, Aalto University School of Science, November 2014.
- Prem Raj Adhikari, Jaakko Hollmén, 2010. Patterns from Multiresolution 0-1 data. In Proceedings of the ACM SIGKDD Workshop on Useful Patterns (UP 2010), pp 8–16.

Multiple Resolutions: Chromosome-17

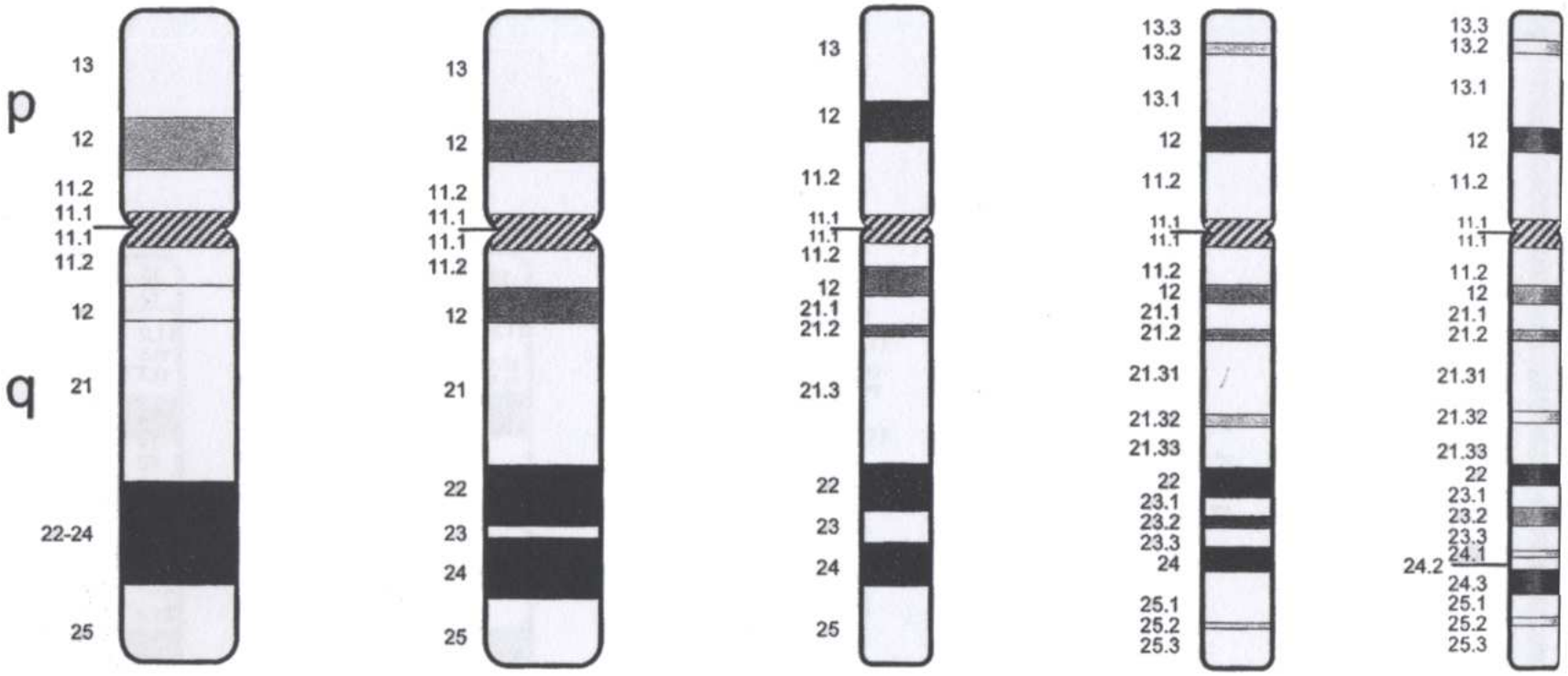
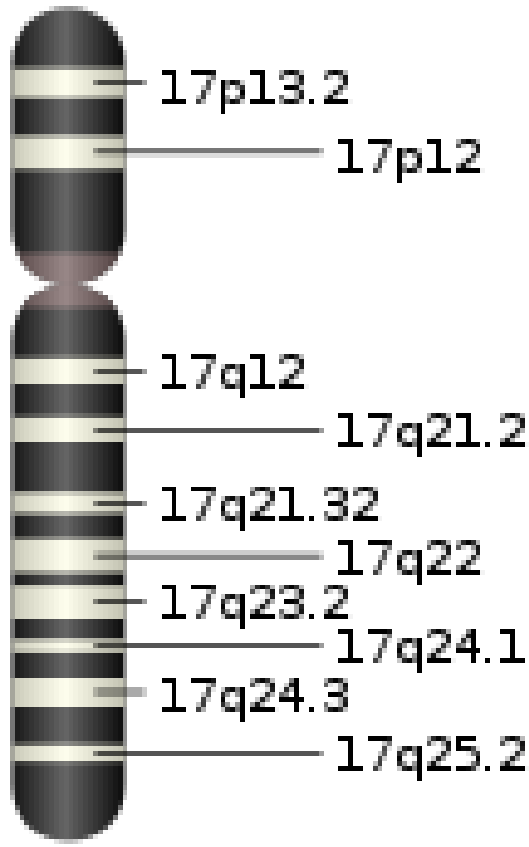


Figure: G-banding patterns for normal human chromosomes at five different levels of resolution. Source: (Shaffer et. al. 2009) . Example case in Chromosome:17.

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: Part of Chromosome-17

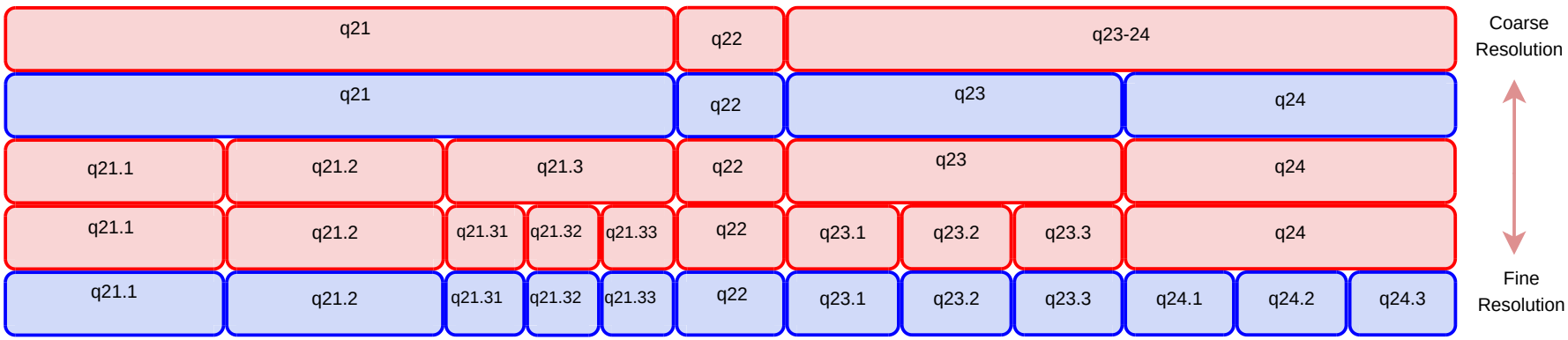


Figure: Part of chromosome 17 showing the differences in multiple resolutions.

Multiple Resolutions: the problem

- Two different datasets are available in two different resolutions. How do you map into other resolutions such that patterns are preserved?

Changing between different resolutions

Upsampling

- Upsampling is the process of changing the representation of data to the higher or finer resolution.
- 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

Are Maximal Frequent Itemset Preserved?

Resolution 400		**Resolution 850**
Frequent Itemset	$\Rightarrow$	Frequent Itemset
{6,7,8}	$\Rightarrow$	{8,9,10,11,12,13,14}
$\Updownarrow$		$\Updownarrow$
Chromosome Bands	$\Rightarrow$	Chromosomse Bands
{17q11.2, 17q12, 17q21}	$\Rightarrow$	{17q11.2, 17q12, 17q21.1, 17q21.2, 17q21.31, 17q21.32, 17q21.33 }

Acknowledgements

Collaborative work:

- Prem Raj Adhikari, Anže Vavpetič, Jan Kralj, Nada Lavrač and Jaakko Hollmén

Based on two publications by the authors:

- Explaining Mixture Models through Semantic Pattern Mining and Banded Matrix Visualization. Proceedings of the Seventeenth International Conference on Discovery Science (DS 2014). Volume 8777 of Lecture Notes in Computer Science. Springer-Verlag. Pages 1–12, October, 2014. http://dx.doi.org/10.1007/978-3-319-11812-3_1
- Explaining Mixture Models through Semantic Pattern Mining and Banded Matrix Visualization. Machine Learning Journal, 105(1), pp. 3-39, http://dx.doi.org/10.1007/s10994-016-5550-3

Multiple Resolutions: Chromosome-17

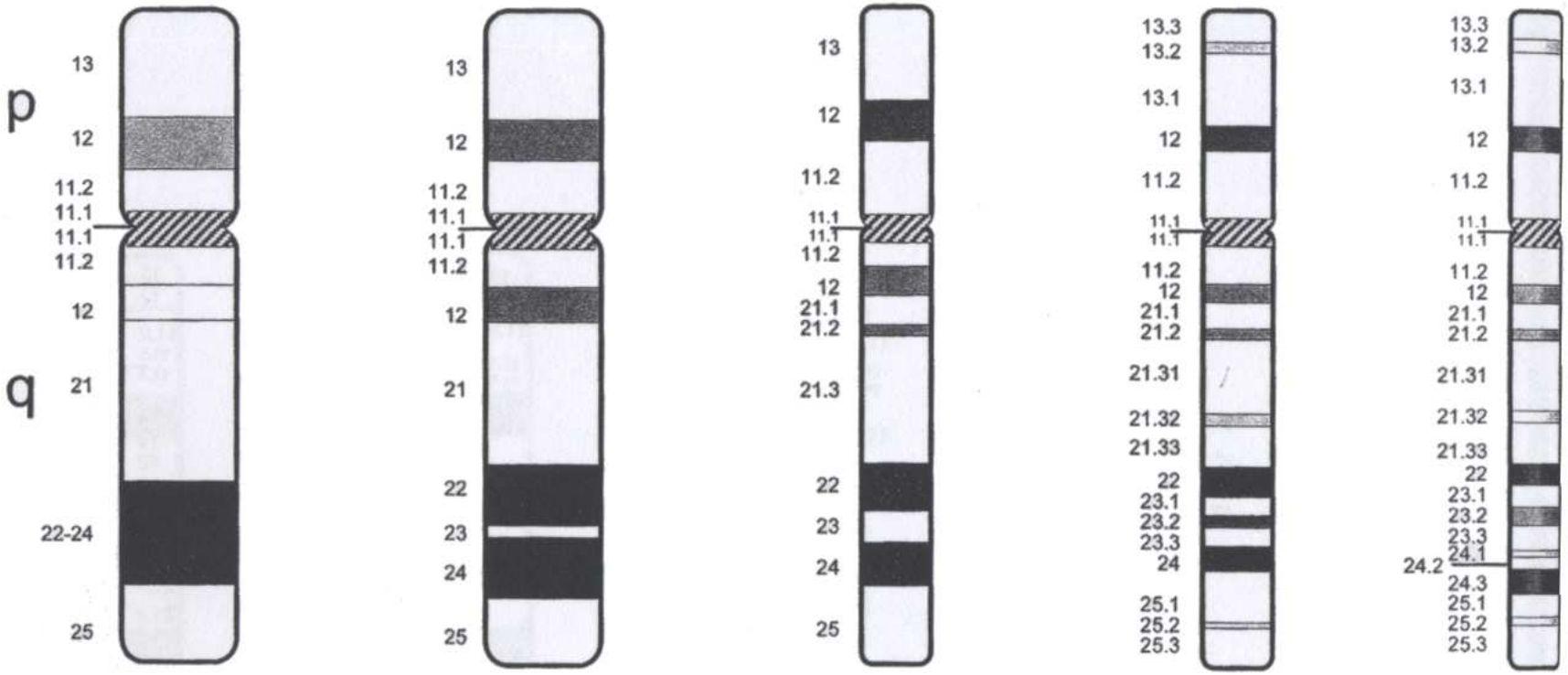
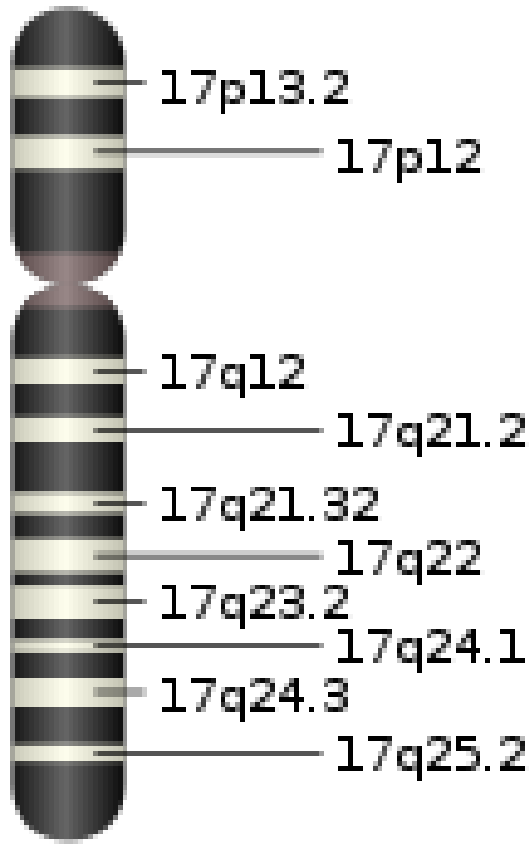


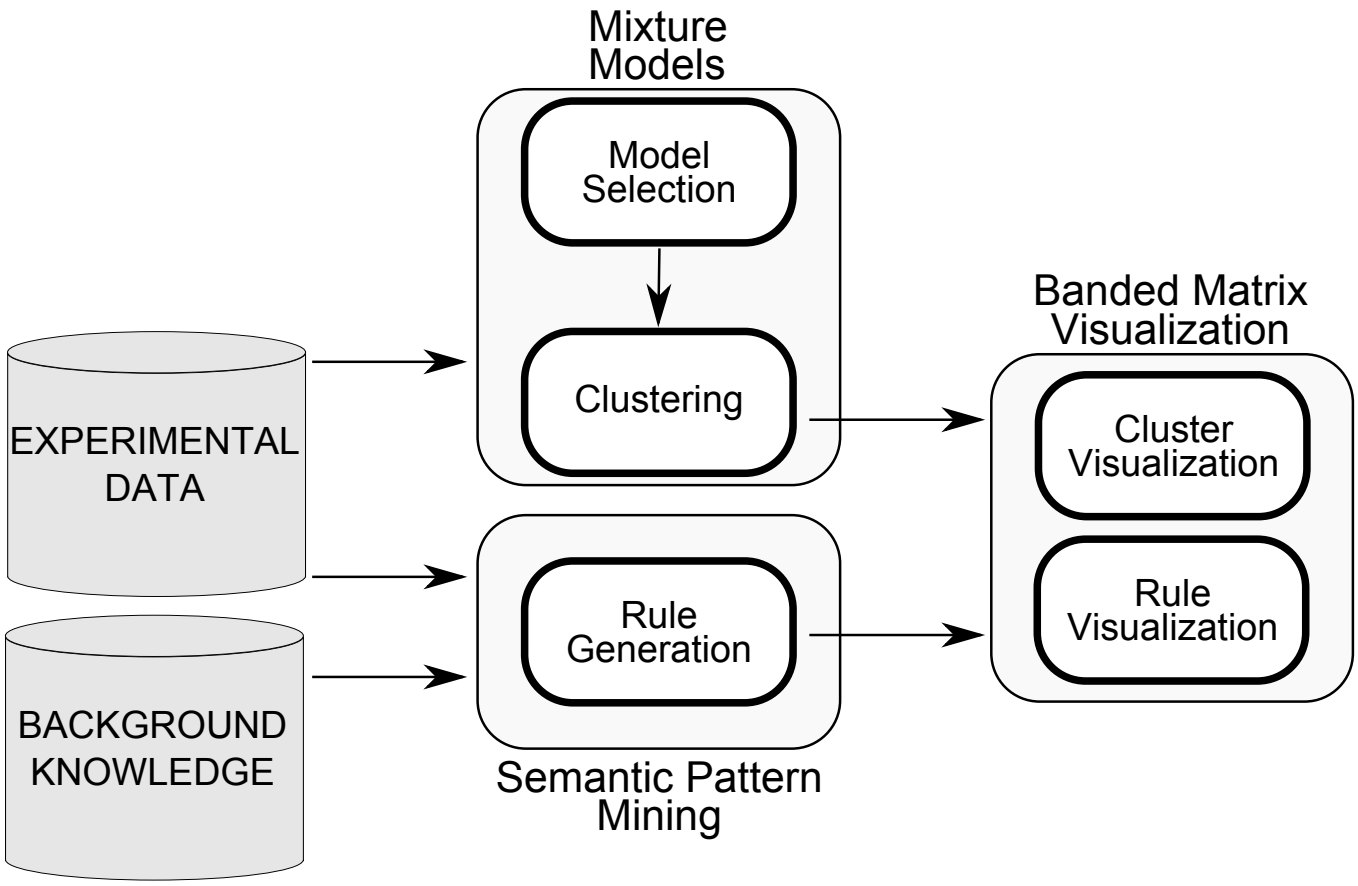
Figure: G-banding patterns for normal human chromosomes at five different levels of resolution. Source: (Shaffer et. al. 2009). Example case in Chromosome:17.

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)



Workflow for the three-part methodology



Management summary

Three-part methodology for semi-automated data analysis:

- Probabilistic clustering of 0-1 data
- Semantic pattern mining from clustered data
- Visual display of the data matrix structure (bandedness)
- Unified visual display of everything

Rest of the talk

- Mixture models and model selection
- Describe amplification data used in the study
- (Semantic) pattern mining from clustered data
- Semantic?
- Unified visual display with structured data
- Examples: visual displays and rules
- Assessment?

Mixture modeling, general

Finite Mixture model

- $p(\mathbf{x}) = \sum_{j=1}^{J} \pi_j \; p(\mathbf{x}|\theta_j)$
- Component distributions $p(\mathbf{x}|\theta_j)$
- mixing coefficients $\pi_j \geq 0, \sum_j \pi_j = 1$
- The whole is the sum of its parts

Estimation of the mixture model from data

- Framework of maximum-likelihood (ML)
- Expectation-Maximization (EM) algorithm

Mixture modeling, 0-1 data

Probability of an observed data vector $\boldsymbol{x}$:

$$p(\boldsymbol{x}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i}$$

Probability of an observed data vector $\boldsymbol{x}$:

$$p(\boldsymbol{x}|\pi_j, \boldsymbol{\Theta}) = \sum_{j=1}^{J} \pi_j p(\boldsymbol{x}|\boldsymbol{\theta}_j) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i}(1-\theta_{ji})^{1-x_i}$$

EM algorithm for the 0-1 mixture model

In the E-step, the expected values of the hidden states are estimated:

$$p(j|\boldsymbol{x}_n, \boldsymbol{\pi}^k, \boldsymbol{\Theta}^k) = \frac{\pi_j^k p(\boldsymbol{x}_n|\boldsymbol{\theta}_j^k)}{\sum_{j'=1}^{J} \pi_{j'}^k p(\boldsymbol{x}_n|\boldsymbol{\theta}_{j'}^k)}$$

In the M-step, the values of the parameters are updated:

$$\pi_j^{k+1} = \frac{1}{N} \sum_{n=1}^{N} p(j|\boldsymbol{x}_n, \boldsymbol{\pi}^k, \boldsymbol{\theta}^k),$$

$$\boldsymbol{\theta}_j^{k+1} = \frac{1}{N\pi_j^{k+1}} \sum_{n=1}^{N} p(j|\boldsymbol{x}_n, \boldsymbol{\pi}^k, \boldsymbol{\theta}^k) \boldsymbol{x}_n.$$

Example: Chromsome 1

Data: dimension of the data fixed $d = 27$
What is an appropriate complexity for the mixture model?
Model-selection problem: the number of component distributions

- J large = complex model, little data to support
- J small = simple model, more data to support

Model selection based on cross-validation

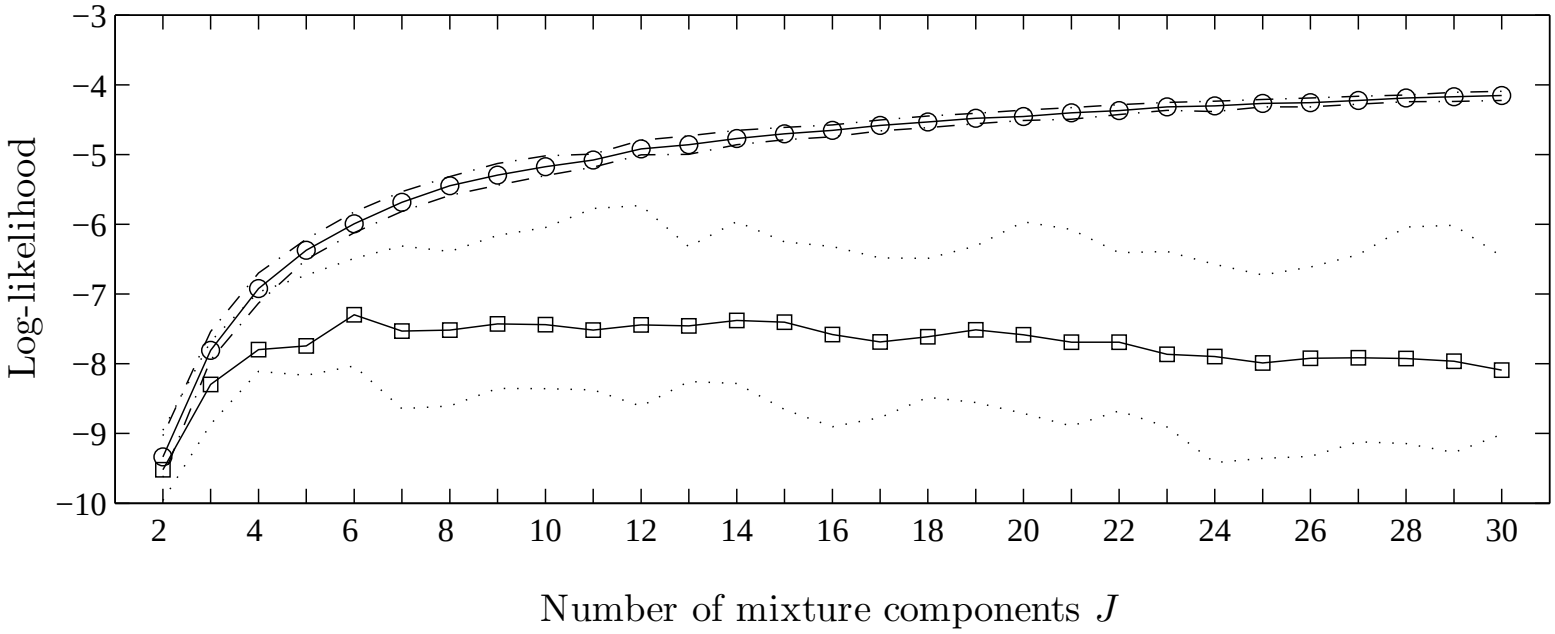
Vary the number of component distributions: $J = 2, \ldots, 30$

- 5-fold crossvalidation repeated 10 times
- 50 partitions of data into a training set and validation set

Train the model fifty times and calculate likelihoods

- 50 likelihood values for the training set
- 50 likelihood values for the validation set
- Computational effort: train a mixture model 1450 times

Model selection based on cross-validation



- Choose the number of components $J = 6$ and train the final model with all the data $\Rightarrow$ Plausible, localized amplification patterns

Mixture modeling "block" ready

- Automatic (?) model selection
- Soft clustering: probabilities (no thanks)
- Hard clustering: data partitions (yes, please!)
- No need to modify the subsequent blocks

Available as an open-source software:

- `http://users.ics.aalto.fi/jhollmen/BernoulliMix/`
- Now: Materials

DNA copy number amplification data

Bibliomics survey from scientific articles of chromosomal comparative genomic hybridization (CGH) studies:

- 838 journal articles
- period of 10 years between 1992 and 2002

DNA copy number amplifications recorded

- 4590 patients with DNA copy number amplifications
- 393 chromosomal regions
- data matrix has 4590 rows and 393 columns
- cancer type for every patient recorded

DNA copy number amplification data

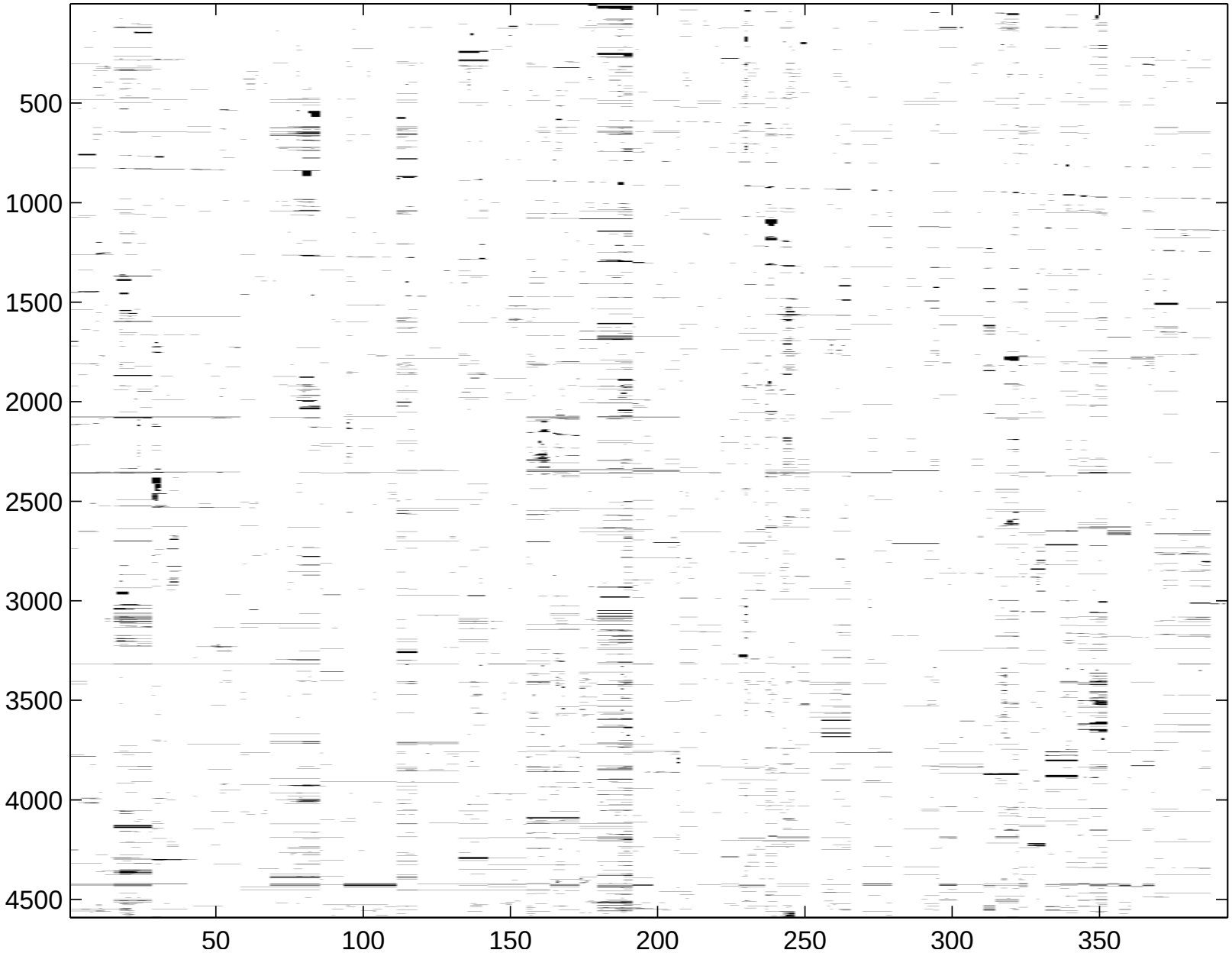
Data matrix: $X = (x_{ij})$, $i = 1, \ldots, 4590$, $j = 1, \ldots, 393$

- $x_{ij} = 1$, if DNA copy number amplification present
- $x_{ij} = 0$, if no amplification present

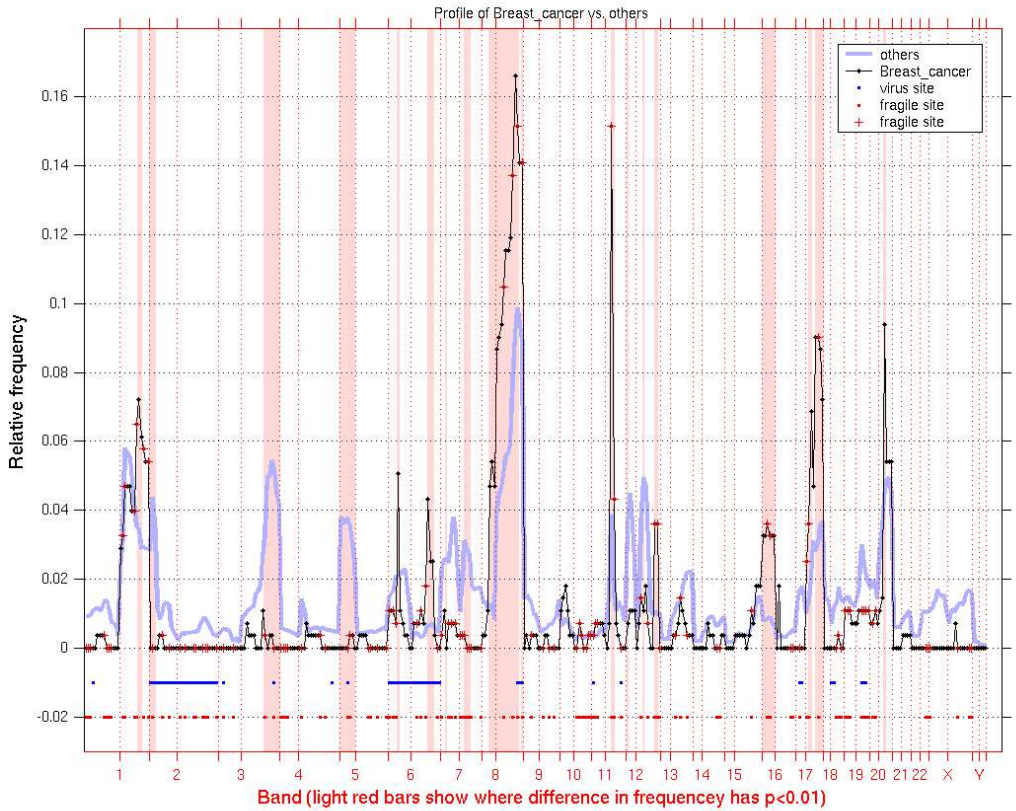
DNA copy number amplifications recorded

- chromosomal regions: 1p36.3, 1p36.2, 1p36.1, ...
- cancer types: Acute lymphoid leukemia, Acute myeloid leukemia, Adrenocortical carcinoma, B-cell lymphoma, Barrett´s adenocarcinoma, ...

DNA copy number amplification data



Profiles of DNA copy number amplification



- Prevalence of an amplification with reference to the rest of the data (time series context!)

Clinical relevance of amplification patterns

Amplification patterns have clinical importance:

- 2p in neuroblastoma
- 17p in osteosarcoma
- 18q in lymphoma
- 1q and 8 in Ewing's sarcoma

Experiments with other data sets

Demonstrate the validity of the approach for other data sets:

- Cities data set describes the most liveable cities in the world according to Mercer ranking
- NY Daily data set desribes the crawled news items along with their sentiment scores
- Tweets data set is a collection of tweets with different features where the original task is to identify sports related tweets
- Stumble Upon data set consists of training data set used in the Kaggle competition

Semantic Pattern Mining

Hedwig system

- Rule induction by specialization
- first-order logical expressions
- Supports ontologies (next slide)
- Example: `Cluster3(X)` $\leftarrow$ `1q43-44(X)` $\wedge$ `1q12(X)`

Available as an open-source software:

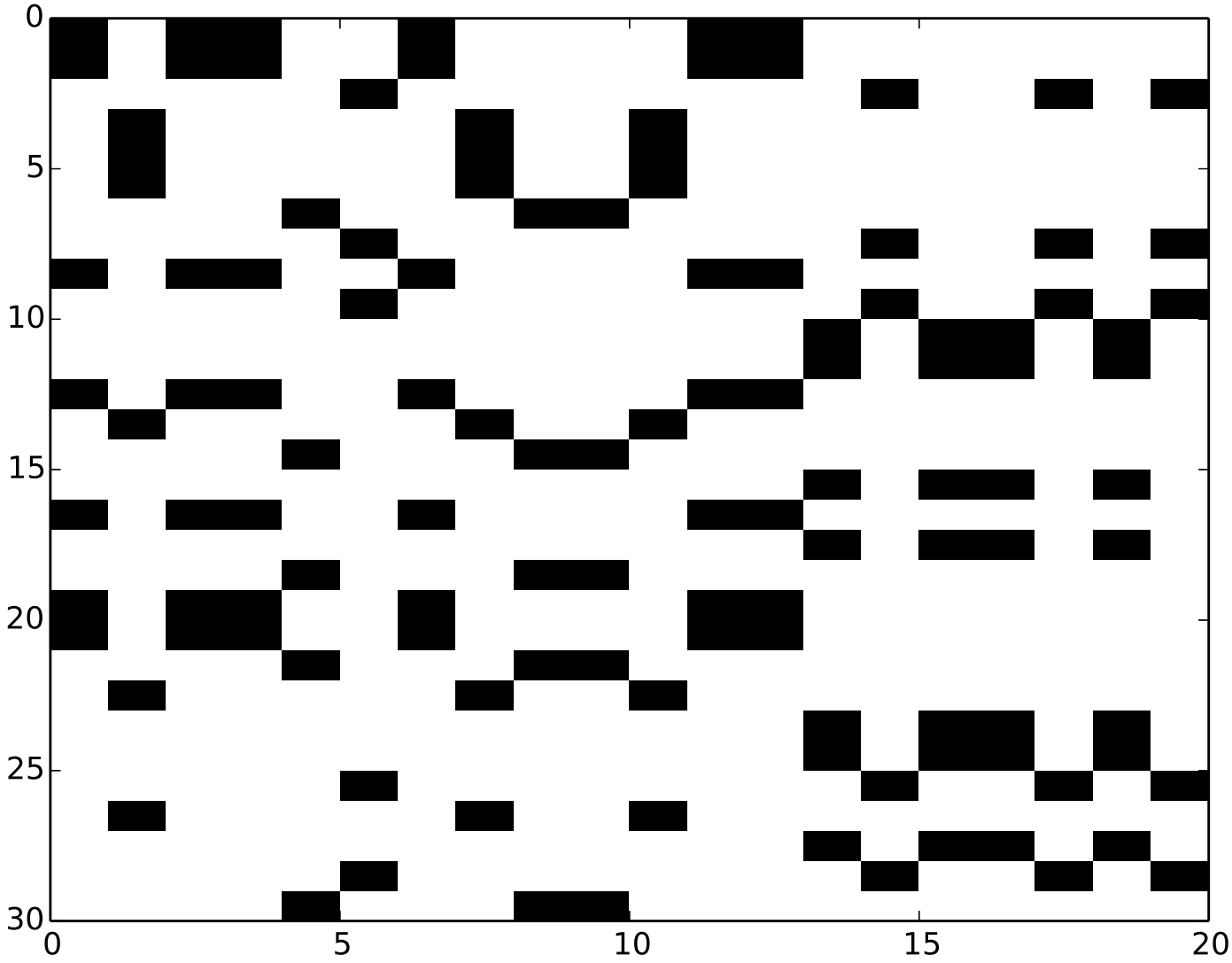
- `https://github.com/anzev/hedwig`

Ontology and semantic pattern mining

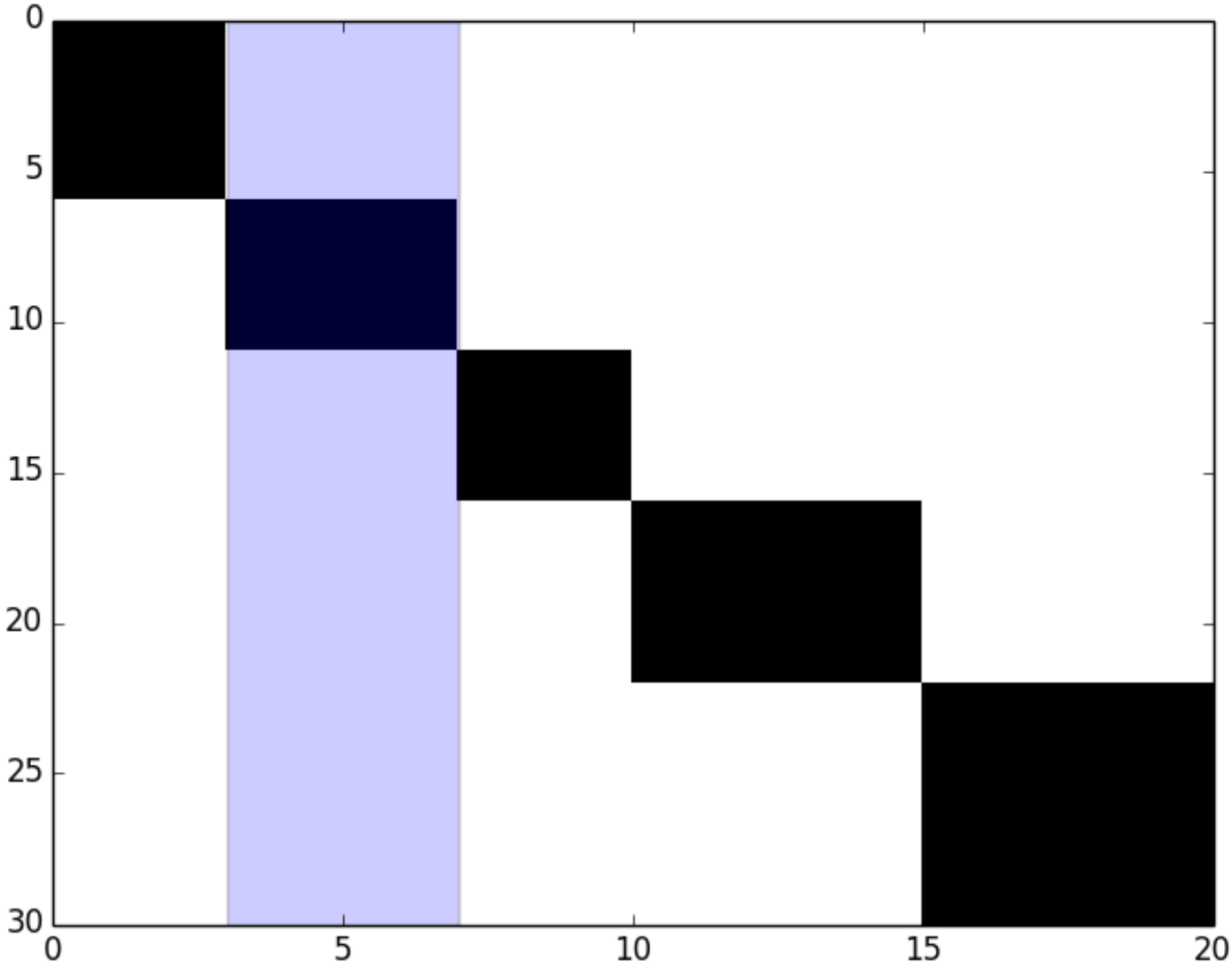
Extraction of semantic patterns (rules) using an ontology of different resolutions of the multiresolution data
Example:

- Riva del Garda *is part of* Italy
- We are in Riva del Garda, We are in Italy
- Genomic region 1q21.1 is part of chromosome 1
- Genomic region 1q21.1 is part of chromosome 1q
- Genomic region 1q21.1 is part of chromosome 1q21
- January 2 is part of week 1 (temporal domain)

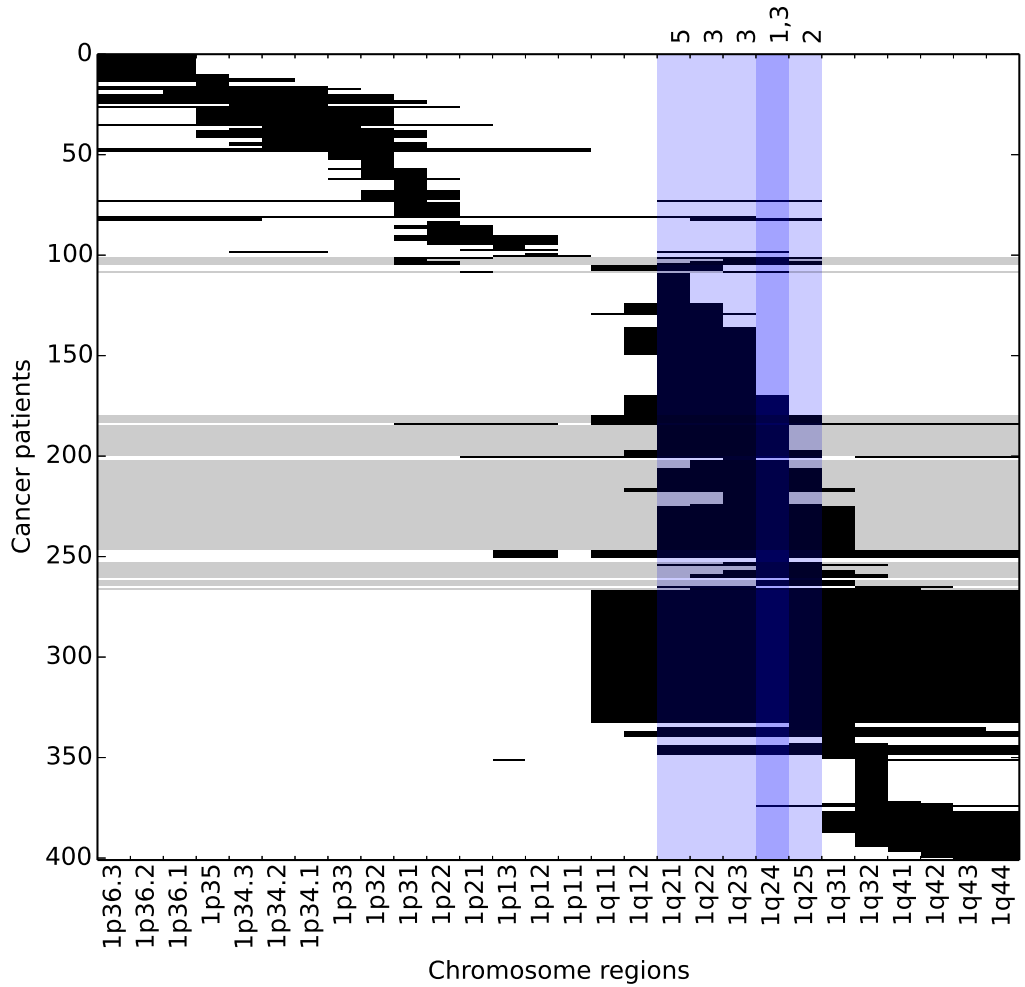
Structural visualization of 0-1 data matrices



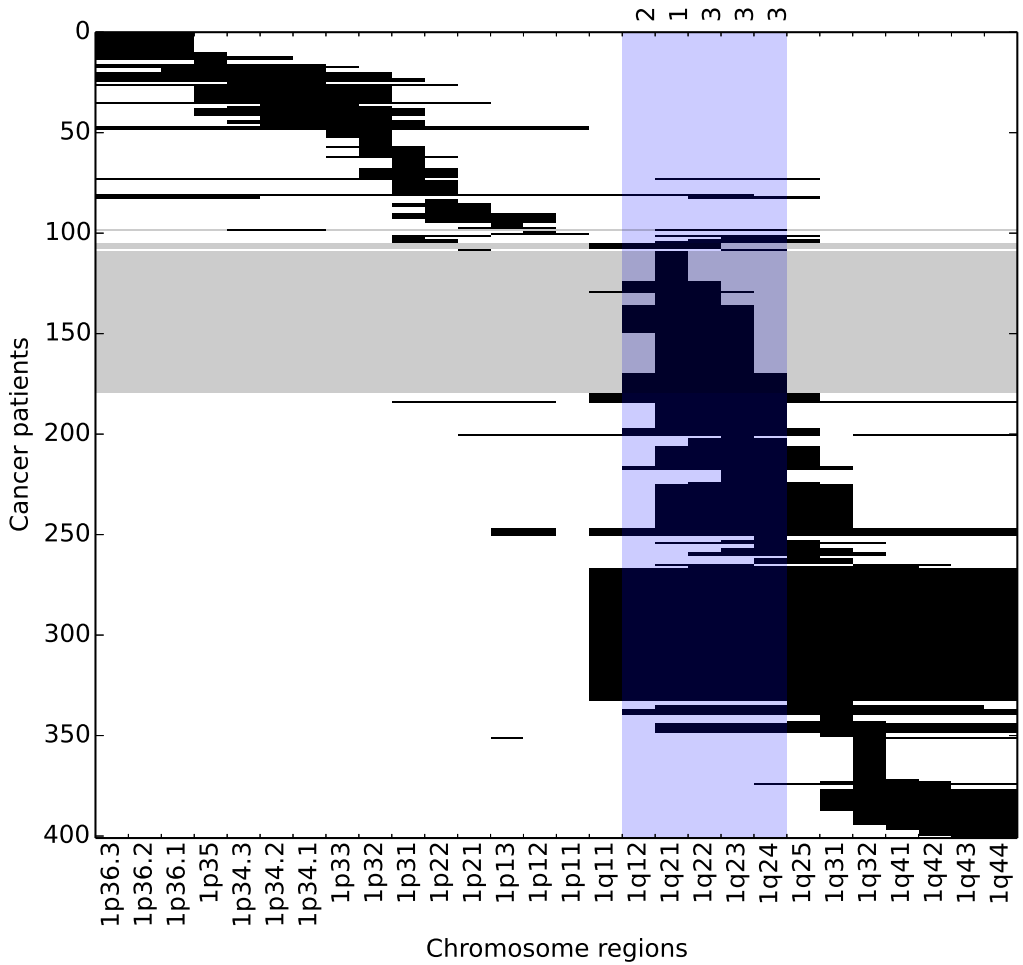
Structural visualization of 0-1 data matrices



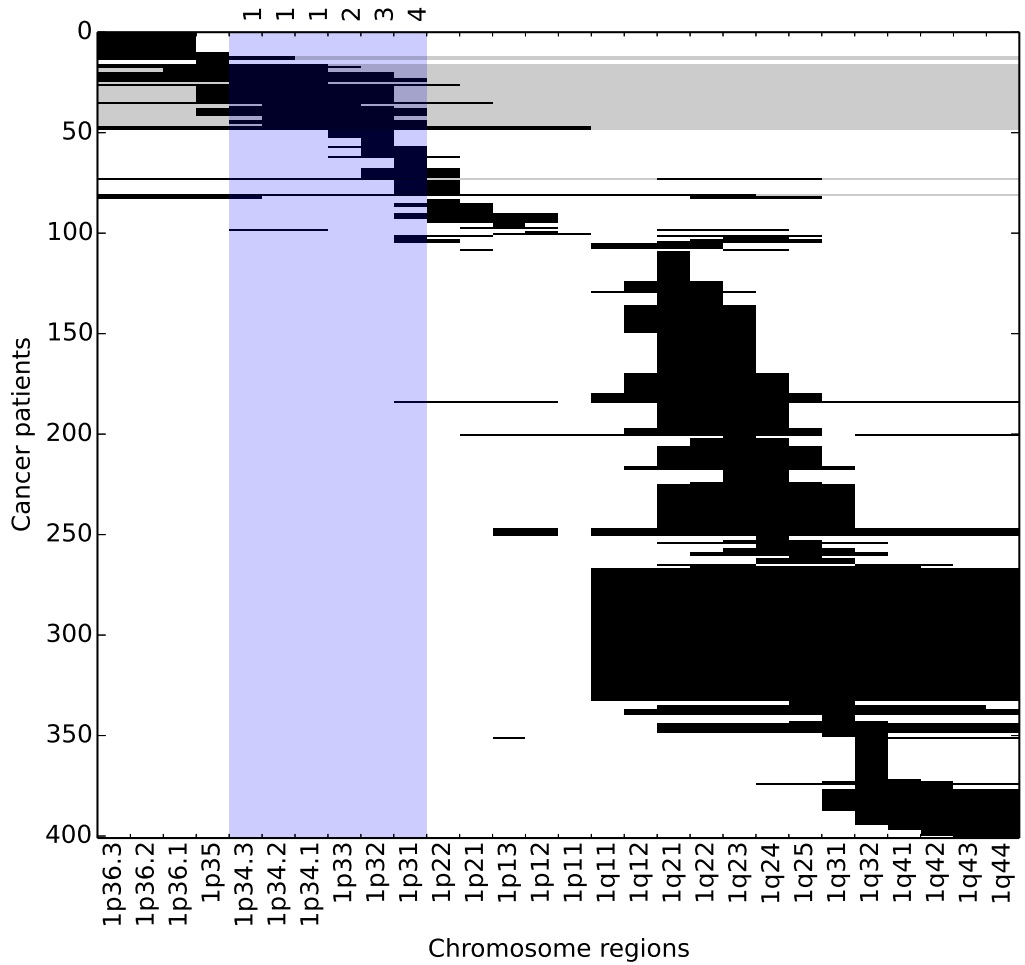
Visual overlay: clusters and rules (Cluster 4)



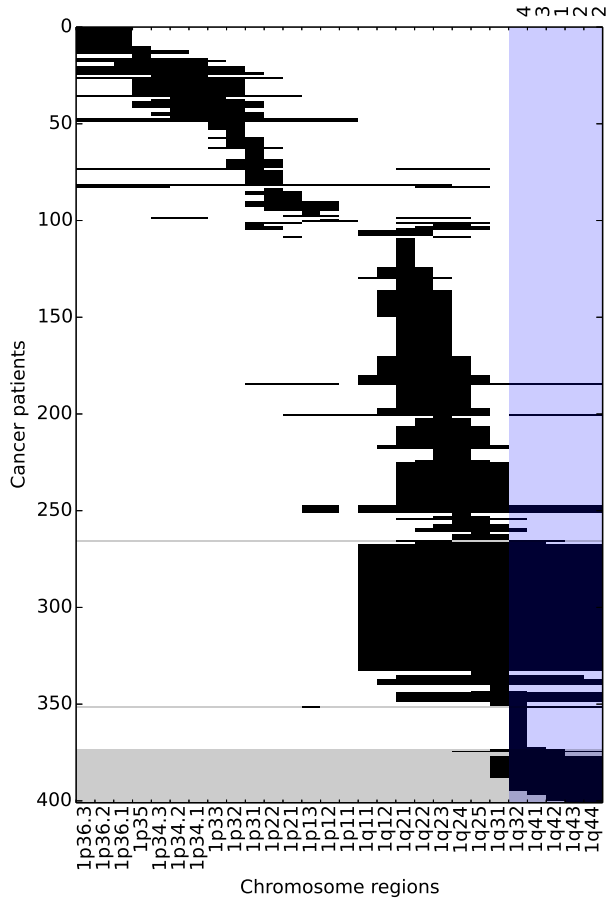
Visual overlay: clusters and rules (Cluster 5)



Visual overlay: clusters and rules (Cluster 6)



Visual overlay: clusters and rules (Cluster 1)

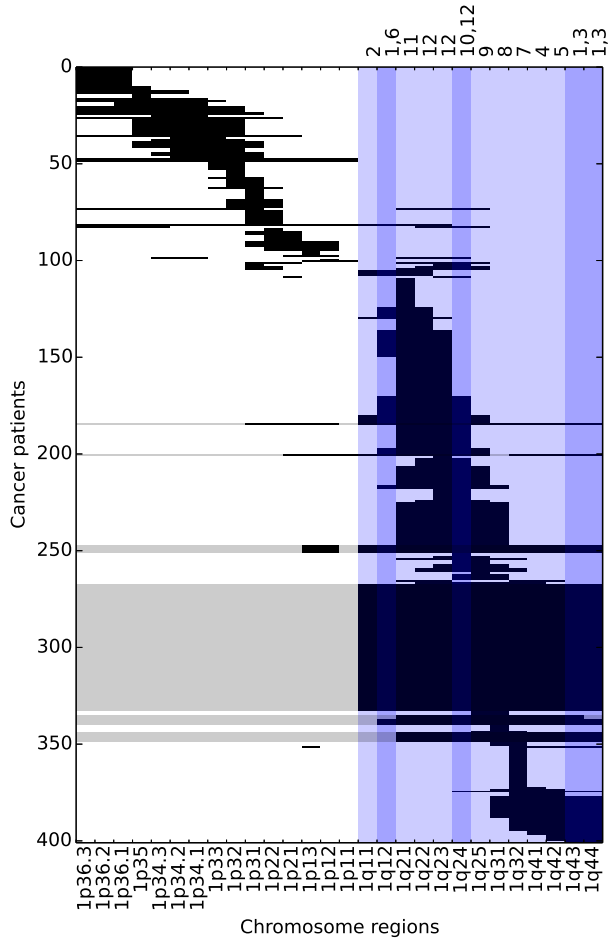


Semantic patterns extracted from cluster 1

#	Rules for cluster 1	TP	FP	Precision	Lift	p-value
1	`Cluster1(X)` ← 1q43–44(X)	26	88	0.23	3.09	0.000
2	`Cluster1(X)` ← 1q41(X)	26	90	0.22	3.04	0.000
3	`Cluster1(X)` ← 1q32(X)	24	116	0.17	2.33	0.000
4	`Cluster1(X)` ← HotspotSite(X)	30	280	0.10	1.31	0.000
5	`Cluster1(X)` ← FragileSite(X)	30	317	0.09	1.17	0.002

Table: Rules induced for cluster 1 of the chromosome 1 data set.

Visual overlay: clusters and rules (Cluster 3)



Extracted rules from cluster 3 of the chromosomal data

#	Rules for cluster 3	TP	FP	Precision	Lift	p-value
1	Cluster3(X) ← 1q43--44(X) 1q12(X)	81	0	1.00	4.62	0.000
2	Cluster3(X) ← 1q11(X)	78	9	0.90	4.15	0.000
3	Cluster3(X) ← 1q43--44(X)	88	26	0.77	3.57	0.000
4	Cluster3(X) ← 1q41(X)	88	28	0.76	3.51	0.000
5	Cluster3(X) ← 1q12(X)	81	43	0.65	3.02	0.000
6	Cluster3(X) ← 1q32(X)	88	52	0.63	2.91	0.000
7	Cluster3(X) ← 1q31(X)	87	54	0.62	2.85	0.000
8	Cluster3(X) ← 1q25(X)	88	64	0.58	2.68	0.000
9	Cluster3(X) ← 1q24(X)	88	97	0.48	2.20	0.000
10	Cluster3(X) ← 1q21(X)	88	134	0.40	1.83	0.000
11	Cluster3(X) ← 1q22--24(X)	88	149	0.37	1.72	0.000
12	Cluster3(X) ← HotspotSite(X)	88	222	0.28	1.31	0.000
13	Cluster3(X) ← CancerSite(X)	88	245	0.26	1.22	0.000
14	Cluster3(X) ← FragileSite(X)	88	259	0.25	1.17	0.000

Table: Rules induced for cluster 3 of the chromosome 1 data set.

Description: assessment

- Predictive models, prediction error
- Data understanding, ???
- Solution: A/B testing ???
- Information systems: create and test framework
- What role does generalization have in description?
- Can you describe one, given data set and generalize well?

Summary and Conclusions

- Three-part methodology: pieces of research knitted together to form a semi-automated workflow
- Clustering "produces" class labels, rule descriptions from clusters (classes)
- Visual display of everything
- Assessment on data understanding remains an open problem

Author information

- Jaakko Hollmén, Aalto University, Department of Computer Science, Finland
- Publications: `http://users.ics.aalto.fi/jhollmen/`