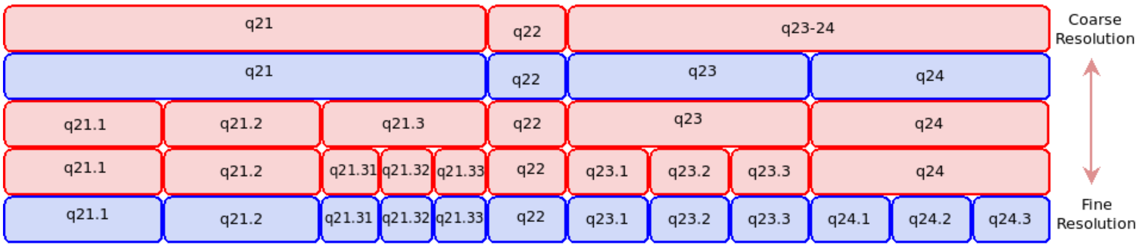
Aalto University
School of Science
and Technology

HELSINKI
INSTITUTE FOR
INFORMATION
TECHNOLOGY

AALTO UNIVERSITY
SCHOOL OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION AND COMPUTER SCIENCE
KONEMIEHENTIE-2, 02150 ESPOO
FINLAND

MIXTURE MODELLING OF MULTIRESOLUTION 0-1 DATA

Prem Raj Adhikari (prem.adhikari@tkk.fi) and Jaakko Hollmén (jaakko.hollmen@tkk.fi)



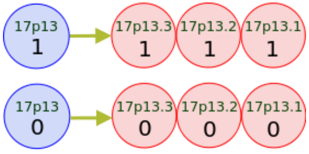
Division of bands in different resolutions; an example case for a part of chromosome 17.

Biological data are available in different resolutions. Computational algorithms can handle only specific resolution of the data. So, data needs to be upsampled and downsampled to different resolutions.

SCALING RESOLUTIONS

UPSCALING

Changing the data resolution to finer resolution

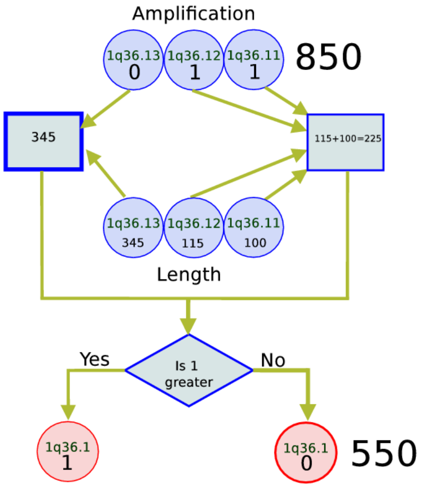


Duplicate copies of similar cytogenetic bands is made in the higher resolution

DOWNSCALING

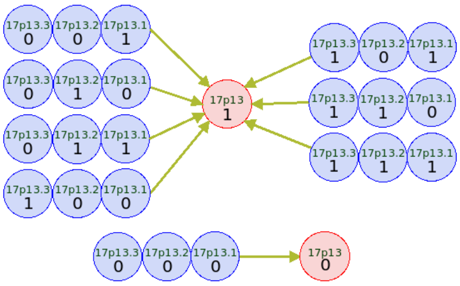
Changing the data resolution to coarser resolution.

1. Weighted Downscaling



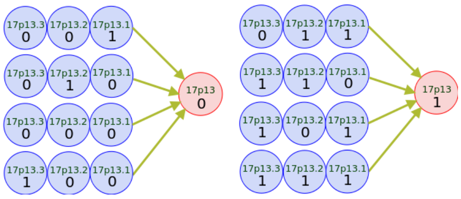
Cytogenetic band in lower resolution is amplified if total length of the amplified bands in higher resolution is greater than the total length of unamplified bands, otherwise it not amplified.

2. Or-Function Downscaling



Cytogenetic band in lower resolution is amplified if any of the bands in higher resolution are amplified, otherwise not.

3. Majority Decision Downscaling

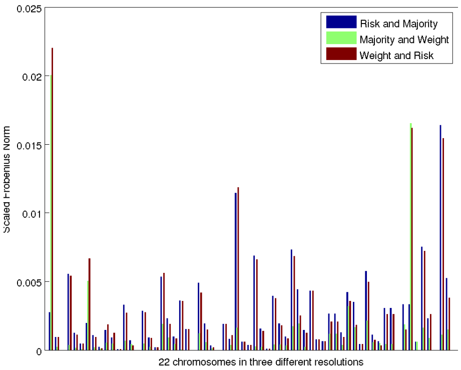


Cytogenetic band in lower resolution is amplified if majority of the bands in higher resolution are amplified

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 the amplification of lower resolution can not be concluded with "golden goal" strategy then the band is lower resolution is deemed as amplified.

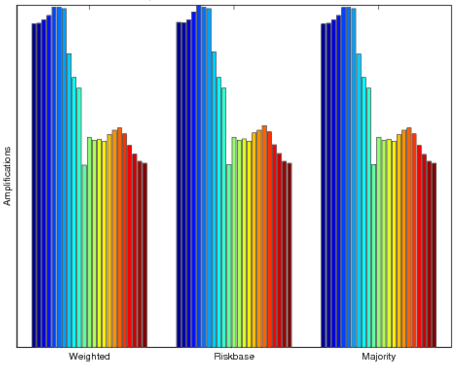
COMPARISION

1. Scaled Frobenius Norm



The scaled frobenius norm between the results of different downsampling techniques. The differences are negligible.

2. # of Amplifications Produced



The number of amplifications produced by different downsampling techniques are also fairly similar. This is an example case for chromosome-6 and resolution-393

MIXTURE MODELLING

STRATIFIED CROSS VALIDATION

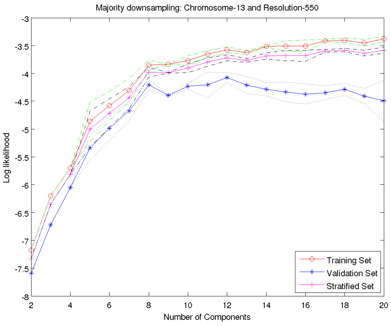
Problem: For small datasets with very few unique rows, cross-validation can suffer from the problem of *"unfortunate-split"*

```
Algorithm 1 Stratified Cross Validation
Input: Dataset $\mathcal{D}$
Output: A testset $\mathcal{T}$
$\mathcal{U} \leftarrow$ the unique rows in the data
$\mathcal{U}_n \leftarrow$ # of rows in $\mathcal{U}$
$\mathcal{K} \leftarrow$ Dynamically adapted uniqueness constant
$\mu_{iter} \leftarrow 0$ and $j = 0$
for $i = 1$ to $\mathcal{U}_n$ do
    $\mu_c \leftarrow$ number of copies of unique row $i$ in data
    $\mu_{iter} \leftarrow \mu_{iter} + \mu_c$
    while $\mu_{iter} \geq \mathcal{K}$ do
        Copy the unique row to the test set $\mathcal{T}_j \leftarrow \mathcal{U}_i$
        $\mu_{iter} = \mu_{iter} - \mathcal{K}$
        $j = j + 1$
    end while
end for
return $\mathcal{T}$
```

MULTIVARIATE BERNOULLI

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

where π_j are the mixture proportions and Θ is composed of $\theta_1, \theta_2, \theta_3 \ldots \theta_d$. Mixture modelling approach is similar to [1] and data from [2]



Example case of model selection for chromosome-13 in resolution 550. Number of components selected in this case is 8.

EXAMPLE RESULTS

Chromosome-1		
Resolution	**Components**	**Likelihood**
300	4	-0.7028
400	7	-0.8981
550	10	-1.5252
700	10	-1.7346
850	13	-1.6269

REFERENCES

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

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