



Cluster analysis of comparative genomic hybridization data

H.A. Kestler^{1,2}, A. Müller¹, F. Schwenker¹, H. Wolter³
T. Gress⁴, T. Mattfeldt³, G. Palm¹

¹Dept. of Neural Information Processing

²Dept. of Internal Medicine II

³Dept. of Pathology

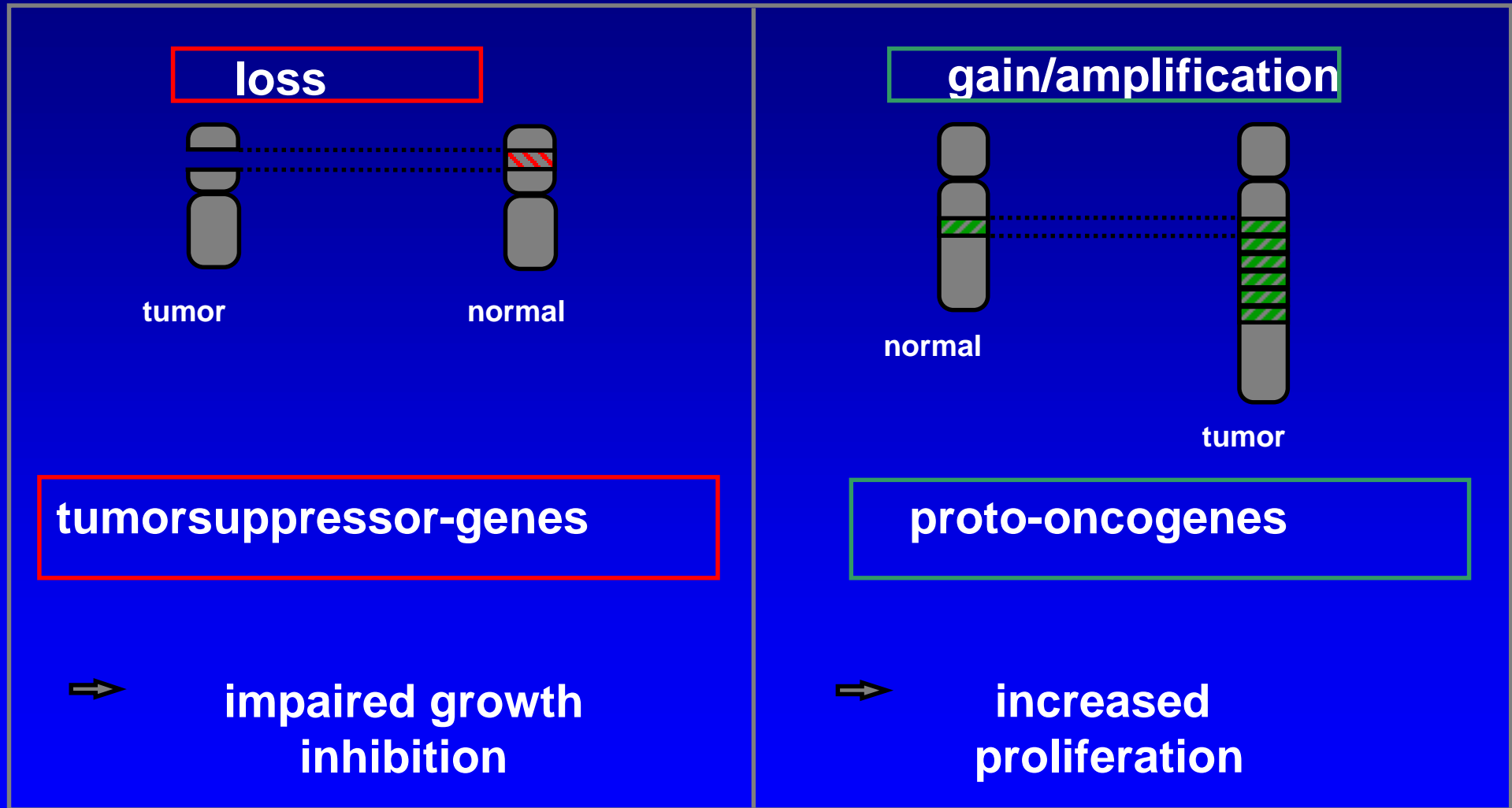
⁴Dept. of Internal Medicine I

University of Ulm / Germany

Materials and methods

Comparative Genomic Hybridization

Chromosomal aberrations

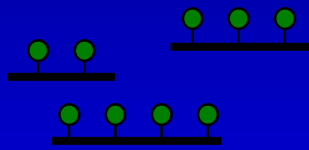


Comparative Genomic Hybridization

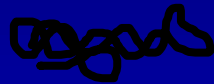
genomic DNA
tumor tissue



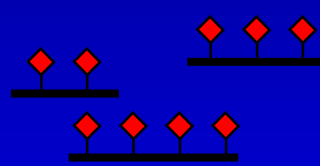
biotin labeling (●)



genomic DNA
normal tissue



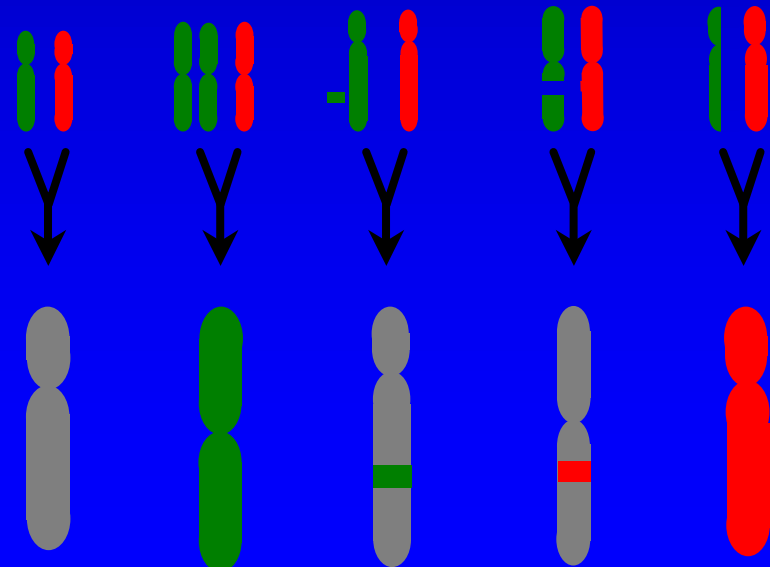
digoxigenin labeling (◆)



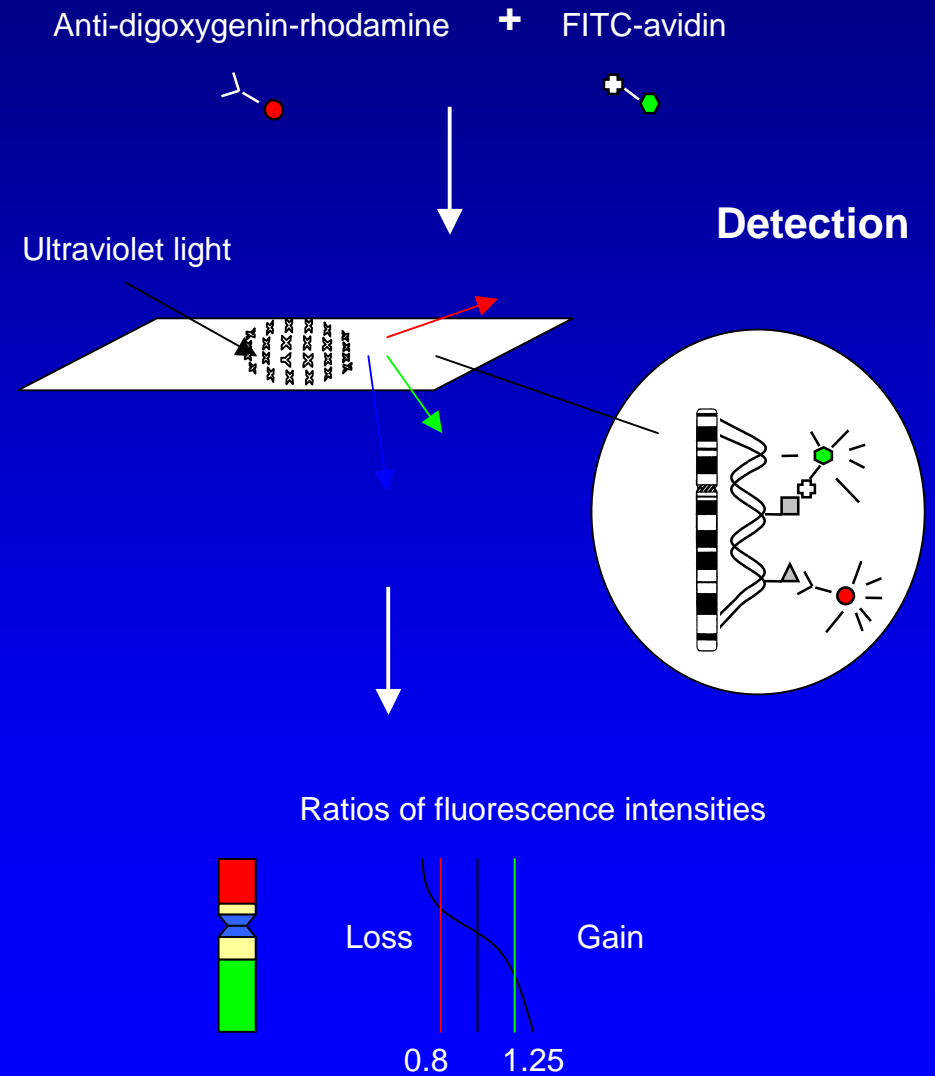
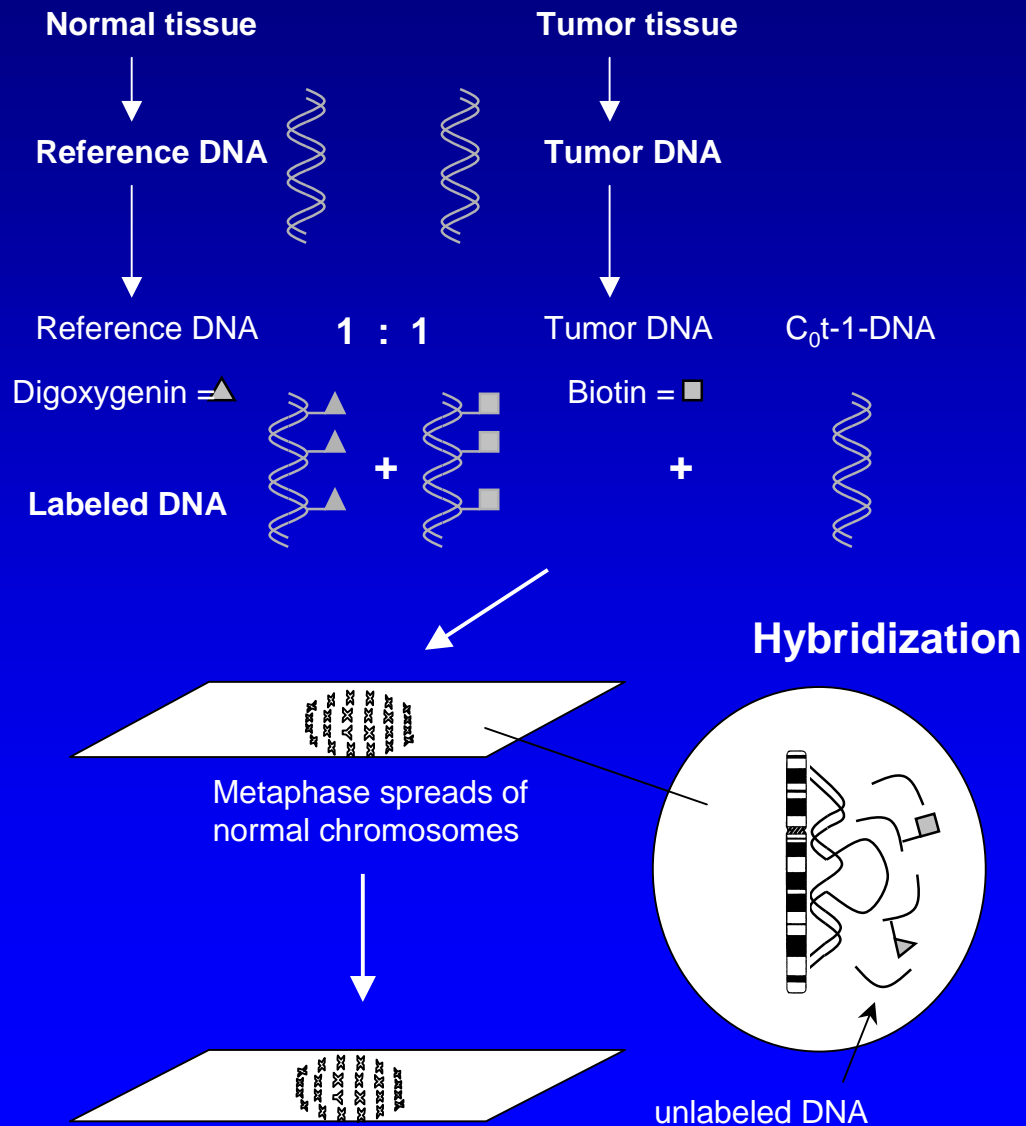
cohybridization



metaphase-spreads of
normal chromosomes



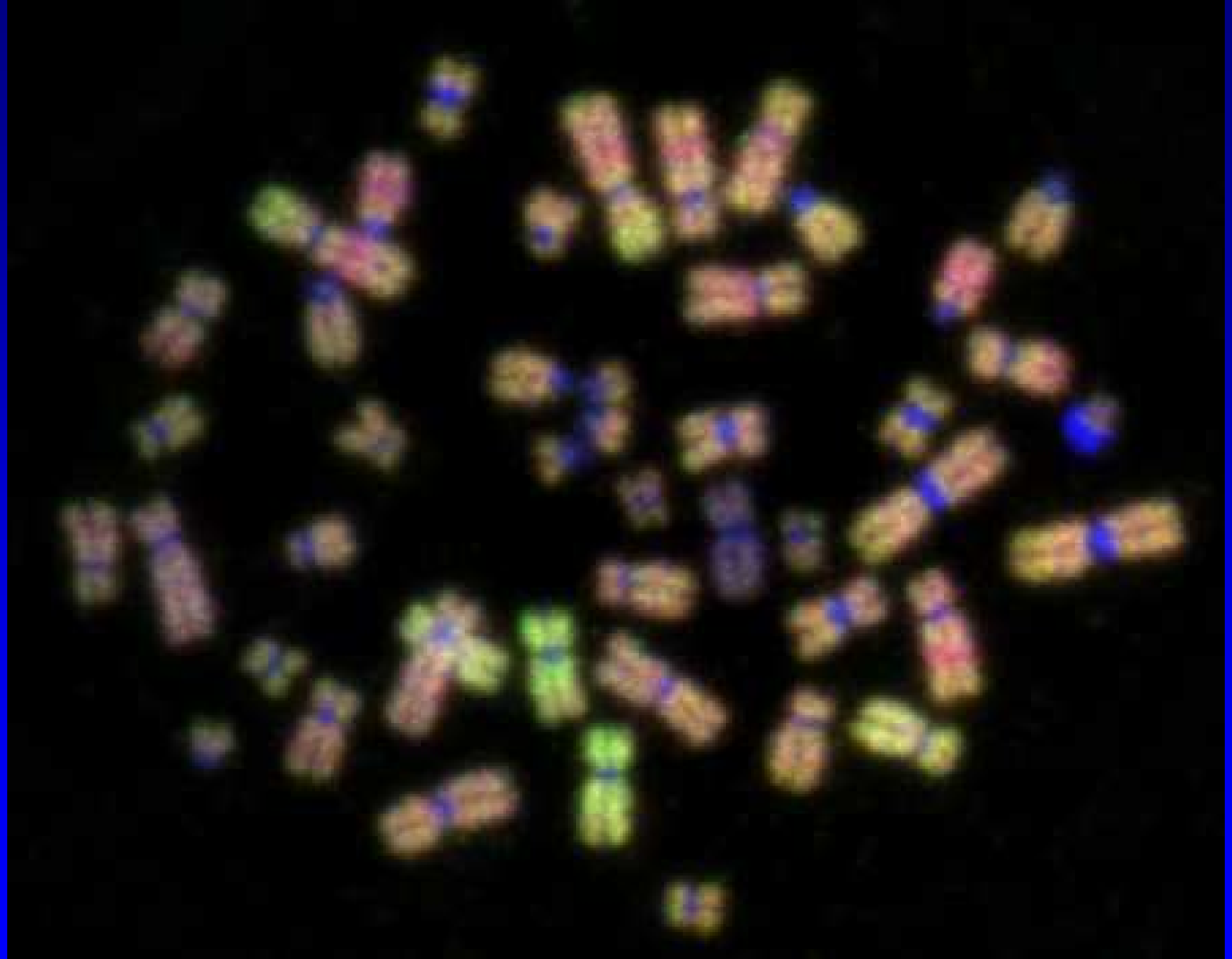
Comparative genomic hybridization (details)



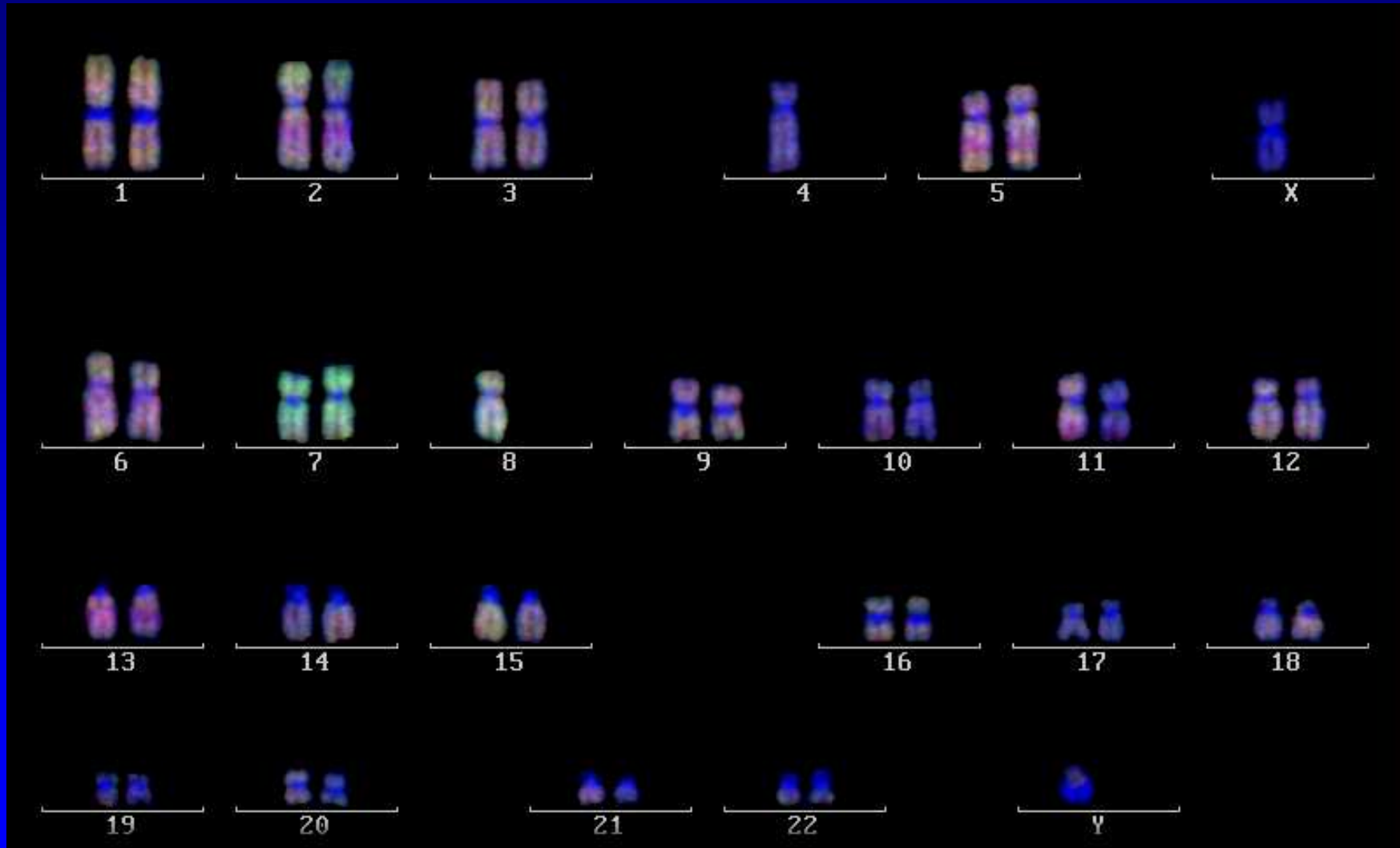
CGH – metaphase spread

CGH metaphase spread.

Hybridized with test-DNA, detected with FITC (green) and reference DNA detected with rhodamine (red).

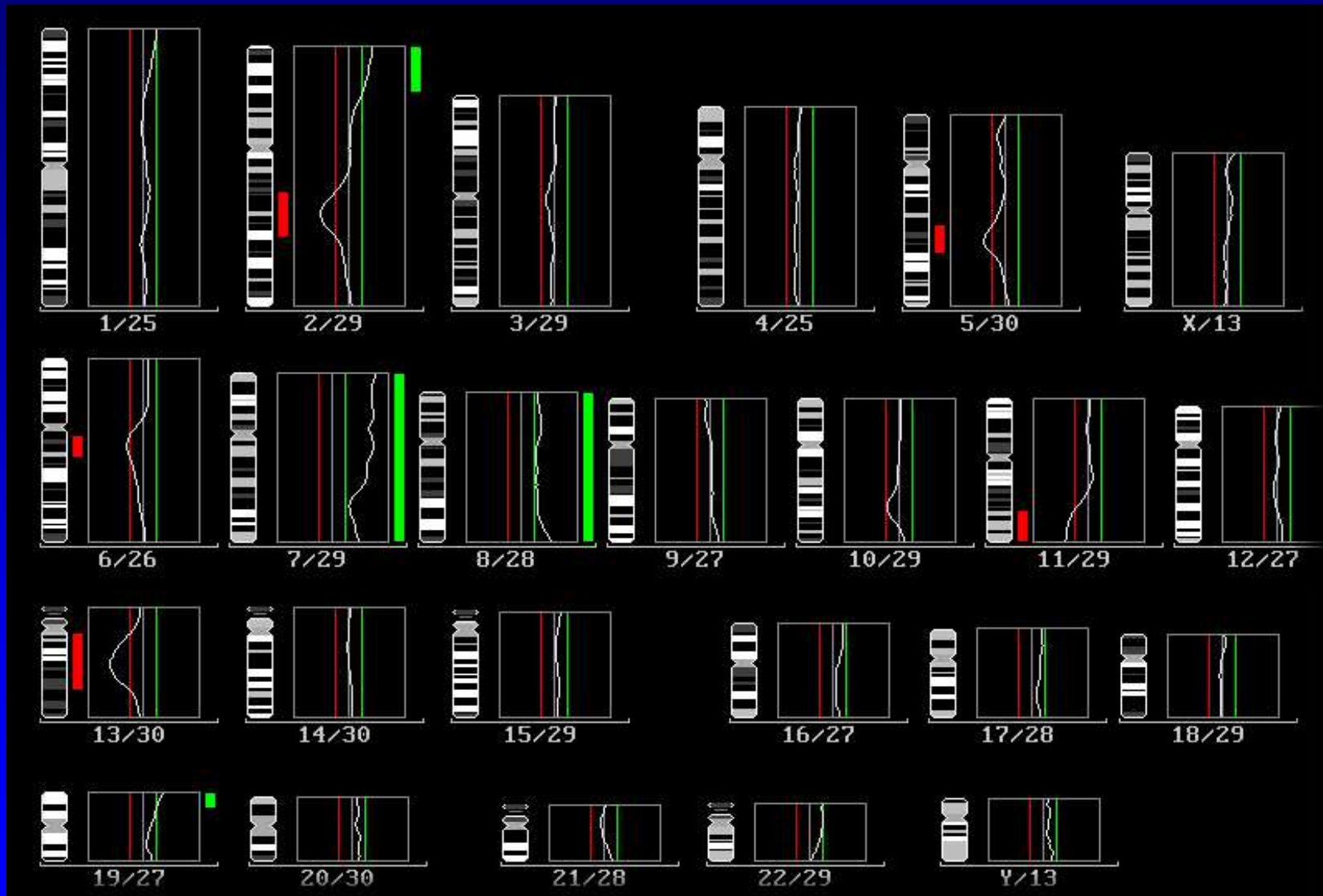


CGH - karyogram



CGH karyogram from the same metaphase (previous figure).

CGH - ideograms

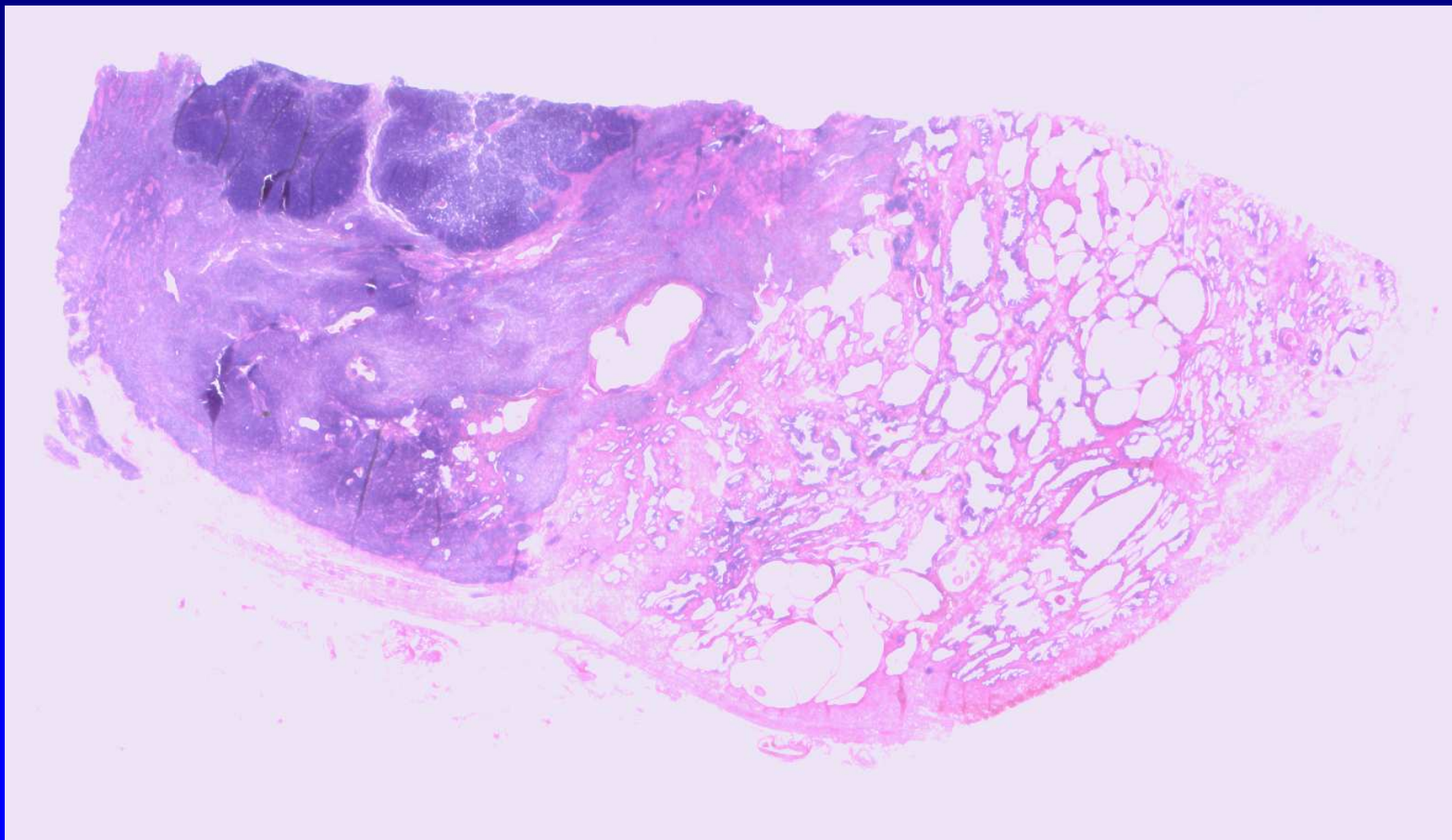


Ideograms with green/red ratio profiles of 15 metaphase cells.

Materials and methods

Prostate CGH data

Example of prostate cancer

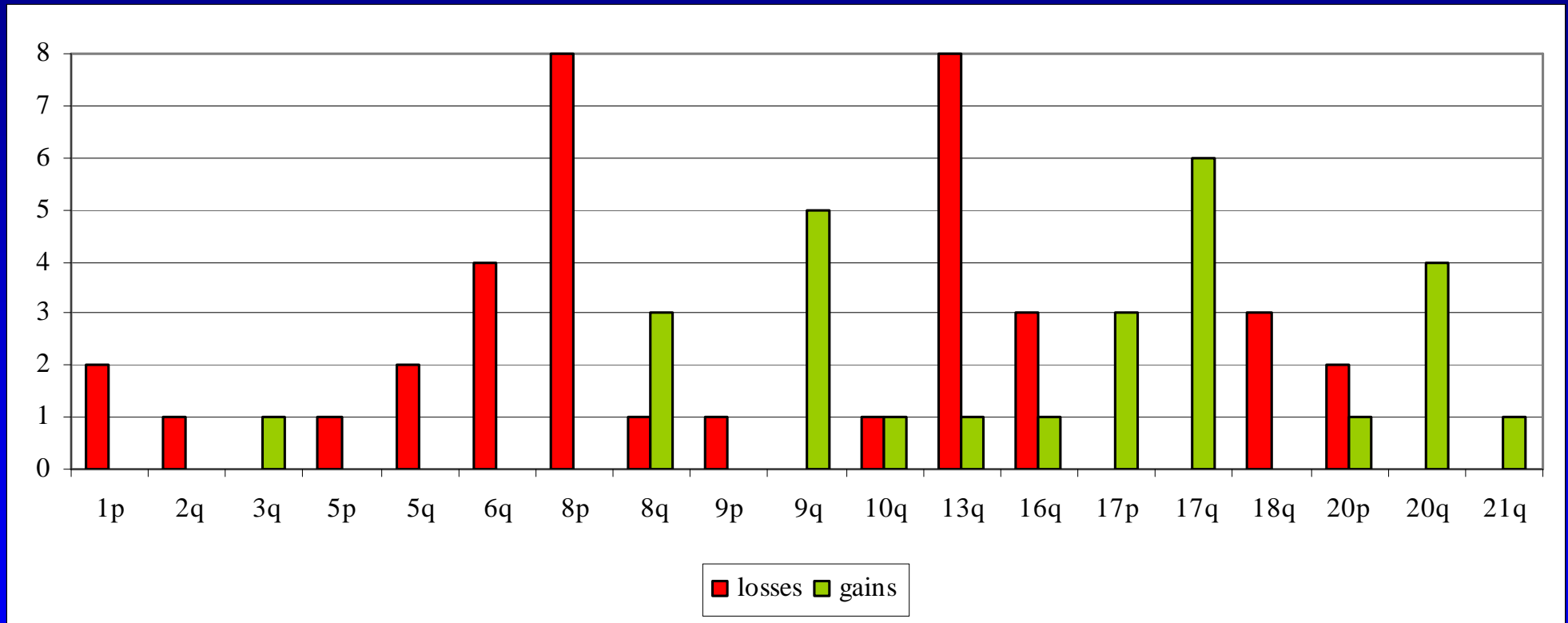


Study group

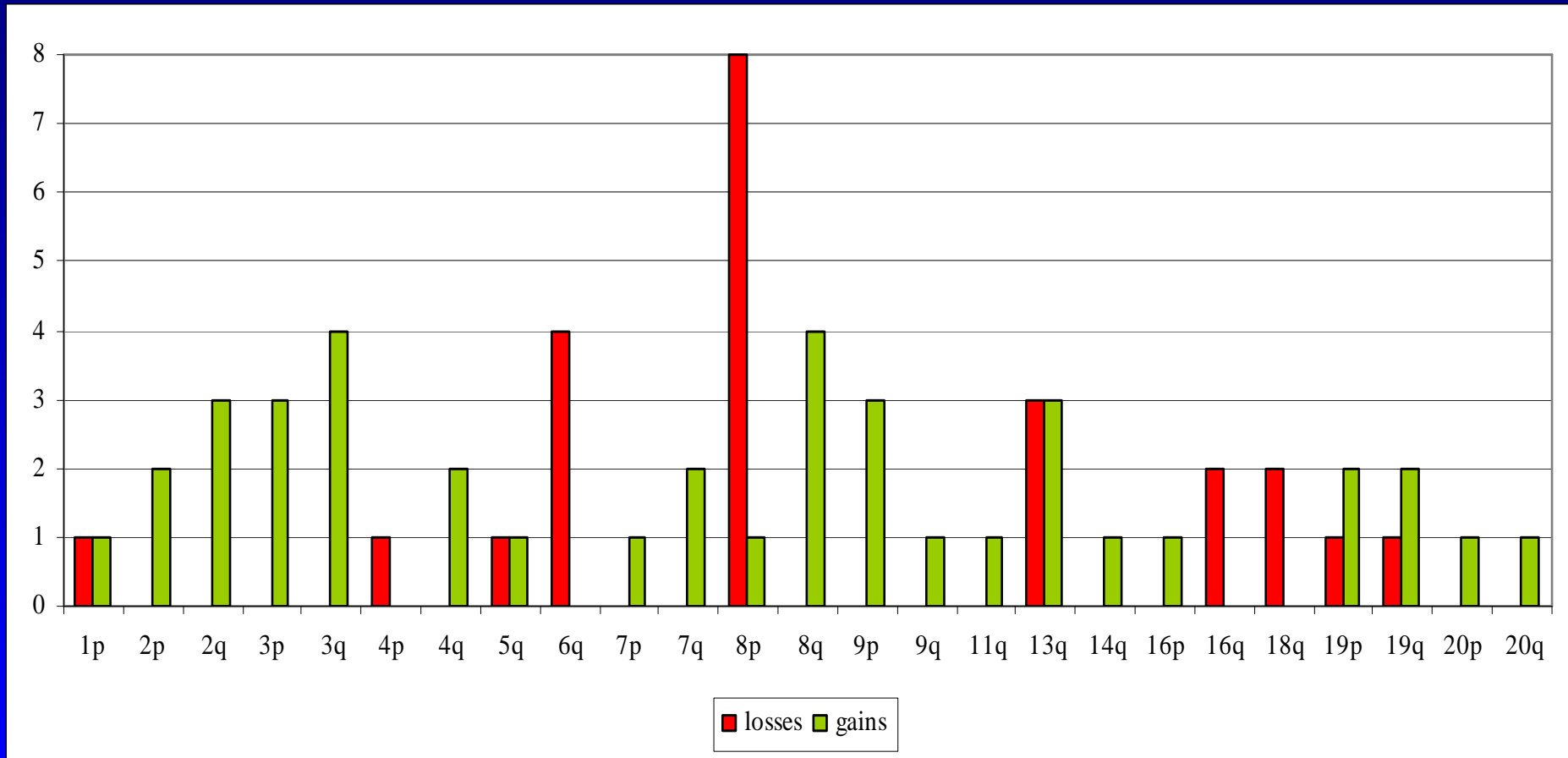
- a) 40 recently obtained primary uncultured prostate carcinomas with the pTNM classification of pT2N0, i.e. tumor restricted to the prostate gland and no regional lymph nodes affected.**

- b) 20 cases with prostate carcinoma stage pT2N0 in whom radical prostatectomy with pelvic lymphadenectomy were selected from the archives of the Urological Department of the University of Ulm. Of these 10 cases with tumor progression and acceptable CGH were selected for the study. The remaining 10 cases without tumor progression were matched by age and duration of follow-up.**

Losses and gains (40 cases without follow-up)



Losses and gains (20 cases with follow-up)



Data

Features: ratios of **green** and **red** intensities within a chromosome arm (30)

Feature values:
0 ⇒ **loss**
1 ⇒ **no change**
2 ⇒ **gain**

Feature vector:

1	0	2	1																2	1	0	1
1p	2p	2q	3p																	21q			

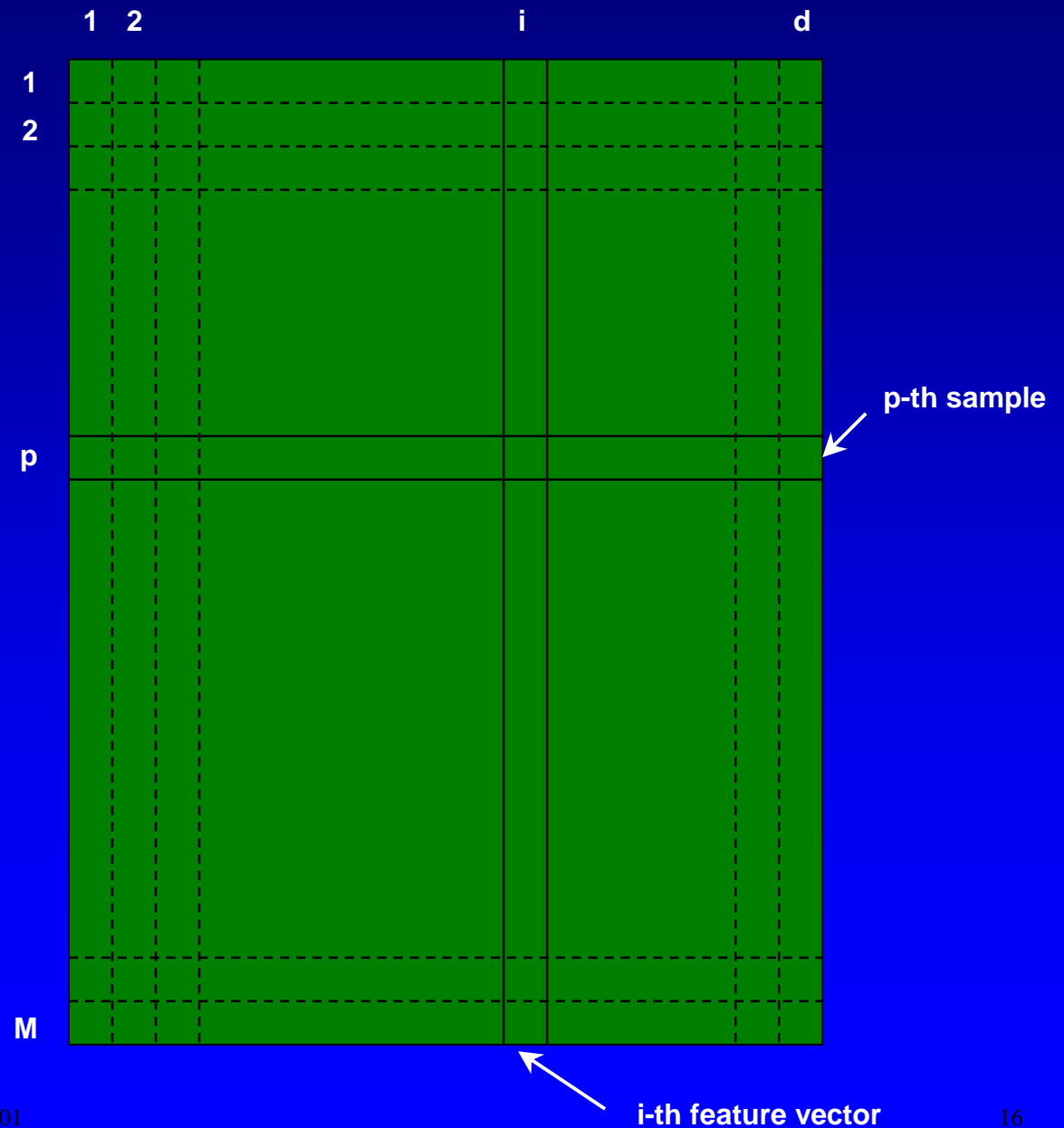
Samples: 60 cases of which 34 are unique

Materials and methods

Data analysis

Data analysis problem

- Data reduction
- Dimensionality reduction

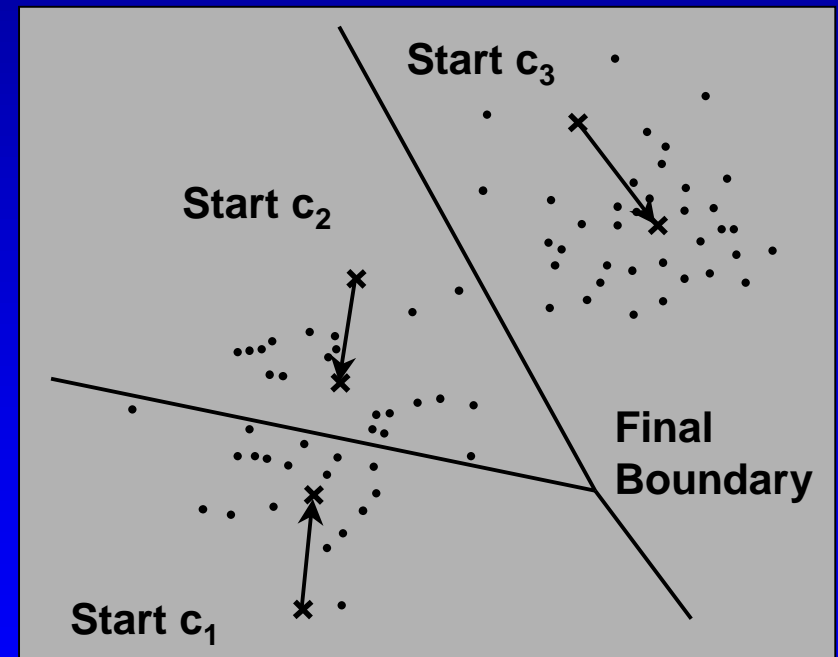


K-means clustering

- Sample feature vectors: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$.
- Means \mathbf{c}_i of vectors in cluster i .
- Vector \mathbf{x} belongs to cluster i , if $\|\mathbf{x} - \mathbf{c}_i\|$ is the minimum of all K distances to cluster centers \mathbf{c}_i .

Procedure:

- Make initial guesses for the means $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K$.
- Until there are no changes in any mean:
 - Use the estimated means to assign the samples \mathbf{x}_j to clusters
 - For i from 1 to K
 - Replace \mathbf{c}_i with the mean of samples for cluster i
 - end_for
- end_until



Fuzzy-c-means clustering

Fuzzy or soft version of the k-means procedure: A feature vector \mathbf{x} can have a degree of membership in every cluster.

Procedure:

- Make initial guesses for the means $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K$. Choose a value for the fuzziness exponent f ($f > 1$).

- Until there are no changes in any mean:

- Use the estimated means to find the degree of membership $m_{\mu,j}$ of \mathbf{x}_{μ} in cluster \mathbf{c}_j , with $d^2(\mathbf{x}_{\mu}, \mathbf{c}_j)$ being the squared distance between \mathbf{x}_{μ} and \mathbf{c}_j :

$$m_{\mu,j} = \frac{d^{\frac{2}{f-1}}(\mathbf{x}_{\mu}, \mathbf{c}_j)}{\sum_{i=1}^K d^{\frac{2}{f-1}}(\mathbf{x}_{\mu}, \mathbf{c}_i)}$$

- For j from 1 to K

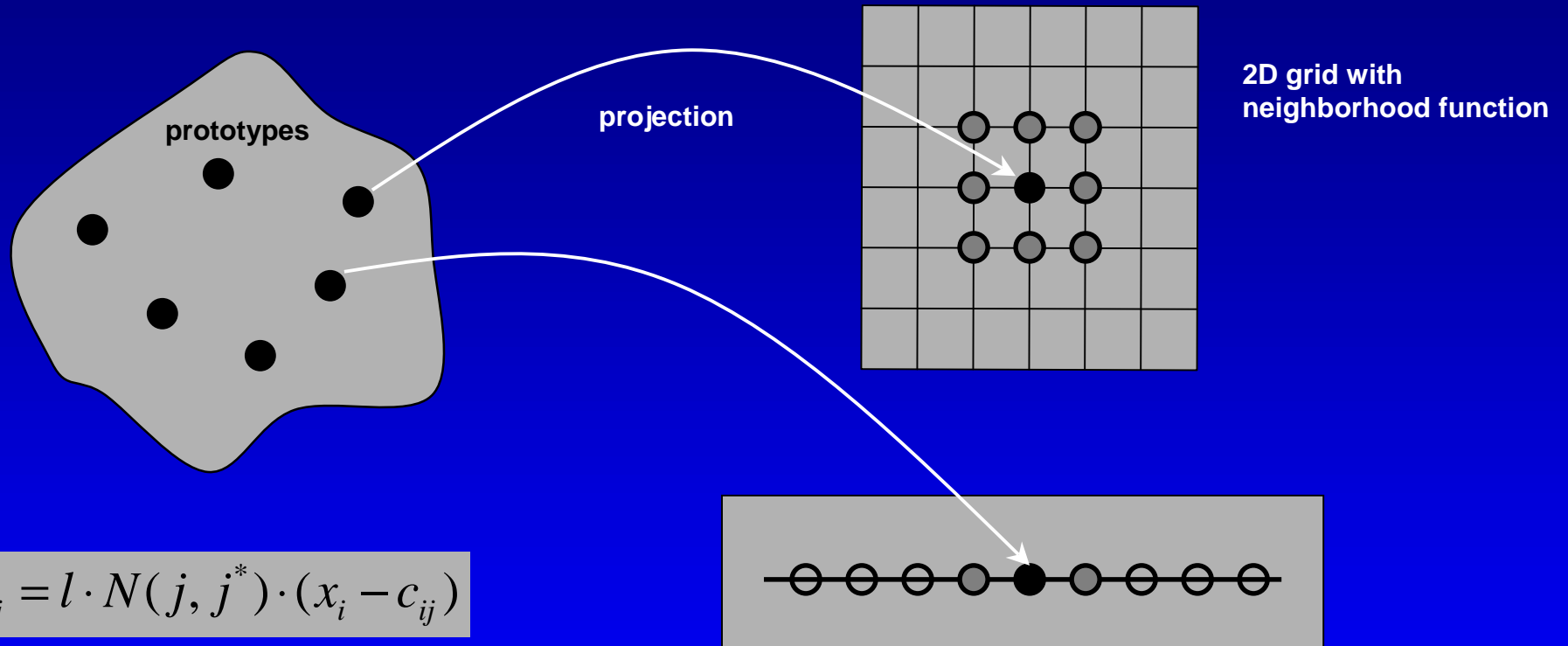
- Replace \mathbf{c}_j with the fuzzy mean of all samples for cluster j :

$$\mathbf{c}_j = \frac{\sum_{\mu=1}^N m_{\mu,j}^f \cdot \mathbf{x}_{\mu}}{\sum_{\mu=1}^N m_{\mu,j}^f}$$

- end_for

- end_until

Self-organizing feature map (SOM)



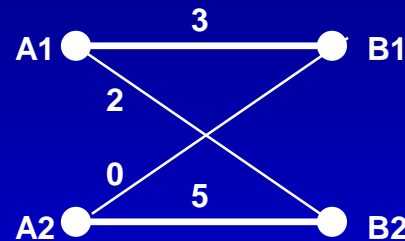
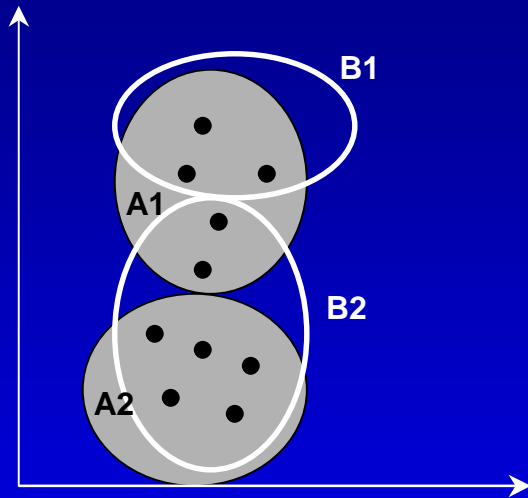
$$\Delta c_{ij} = l \cdot N(j, j^*) \cdot (x_i - c_{ij})$$

Winning neuron: j^*
Neighborhood function: $N(j, j^*)$

Example: $N(j, j^*) = \exp(- \|p(j) - p(j^*)\|^2 / \sigma^2)$, where $p(j)$ is the location of the j -th neuron in the grid or line

Measuring robustness of clusterings

Comparing different clusterings: maximum cluster assignment (MCA)



	B1	B2
A1	3	2
A2	0	5

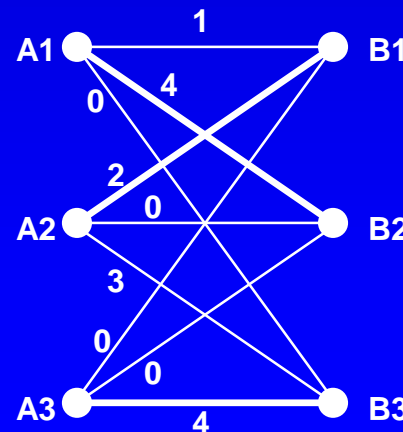
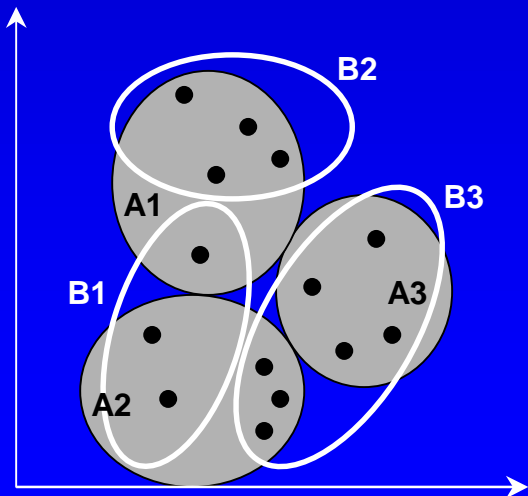
A1↔B1
A2↔B2

$$Score = \frac{3+5}{10}$$

A1↔B2
A2↔B1

$$Score = \frac{0+2}{10}$$

- Count number of samples in the intersection set
- Find correspondence with the max. score

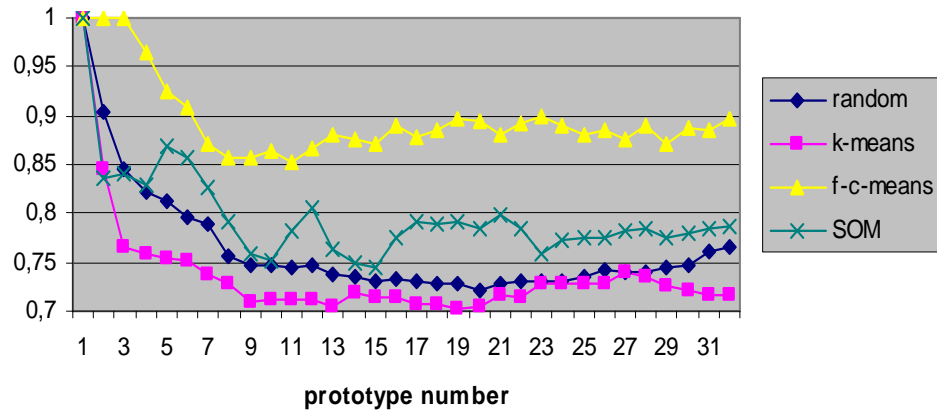


1	4	0
2	0	3
0	0	4

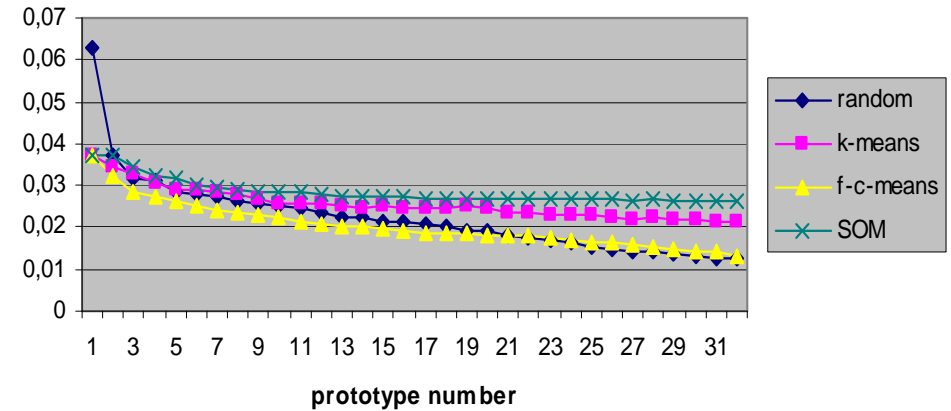
$$Score = \frac{4+2+4}{14}$$

Results I

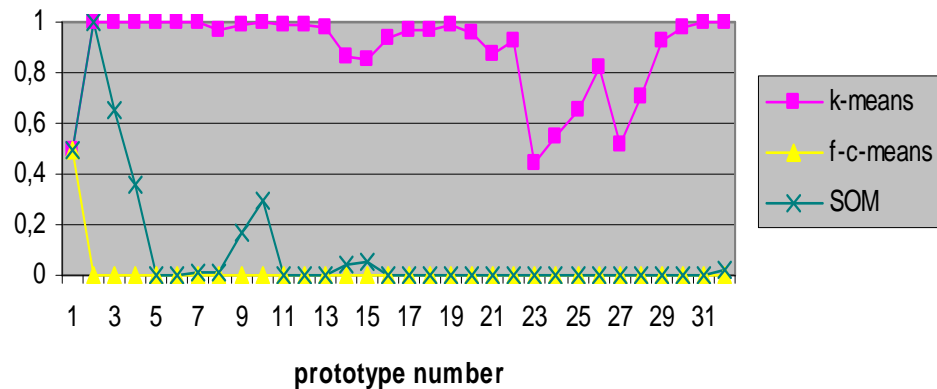
Robustness (MCA)



Quantization Error



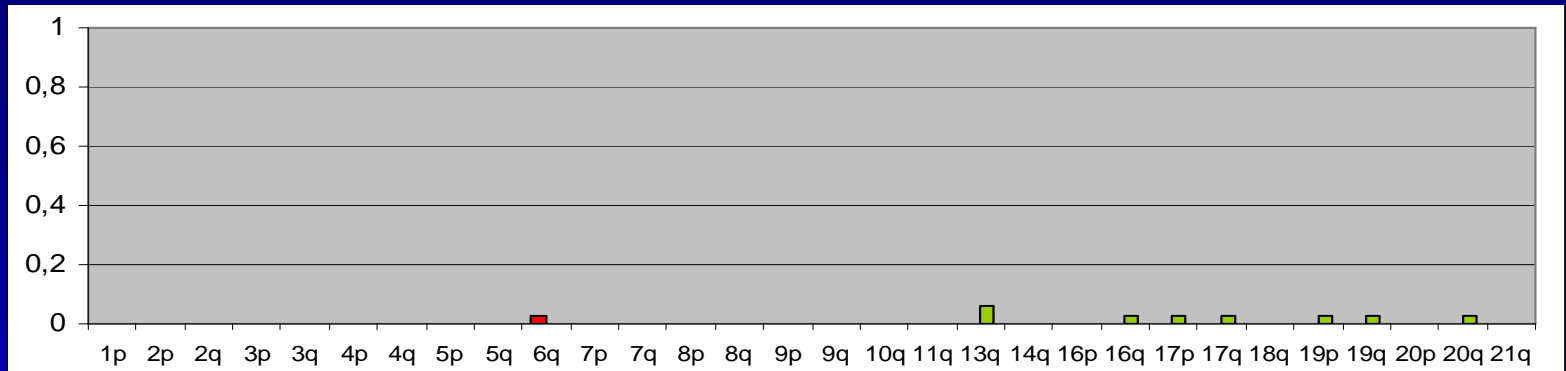
Mann-Whitney Test (p-values)



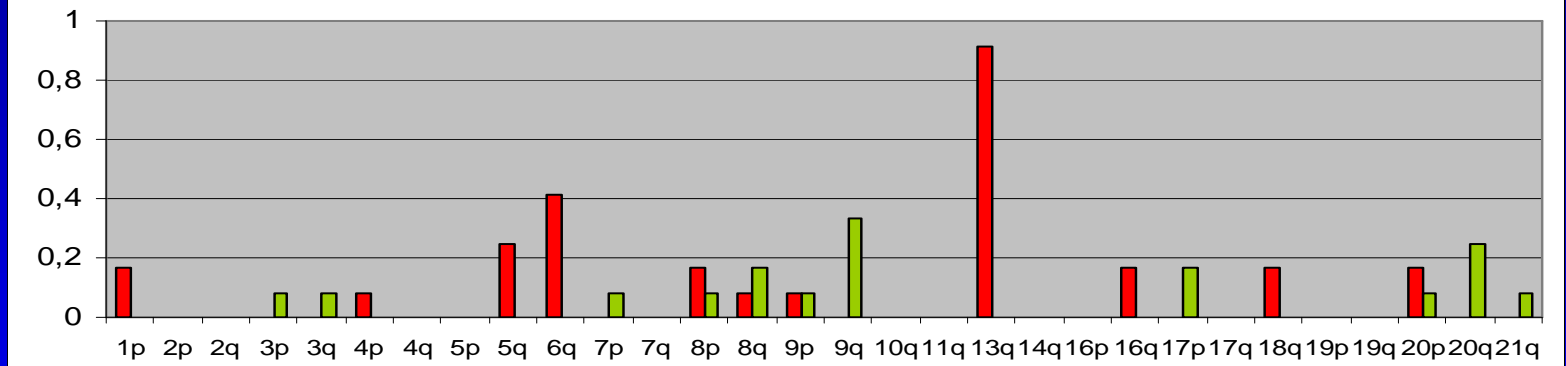
- Quantization Error: Sum of distances (l^2) to nearest prototype normalized by dimension and number of samples.

Results II

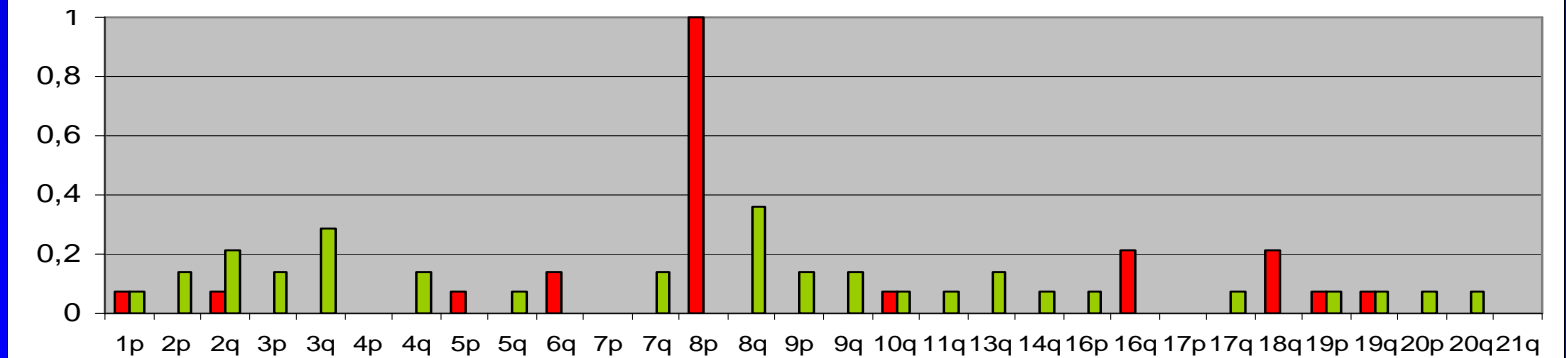
Cluster 0



Cluster 1



Cluster 2

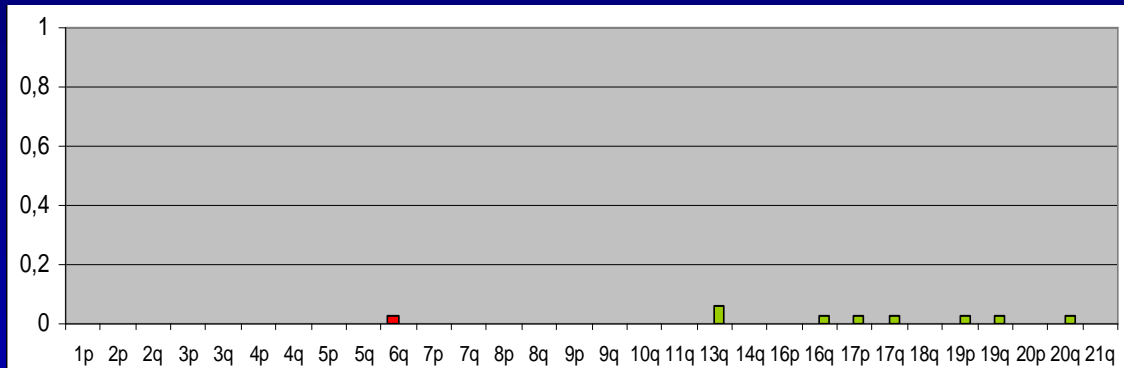


Mean values

■ losses ■ gains

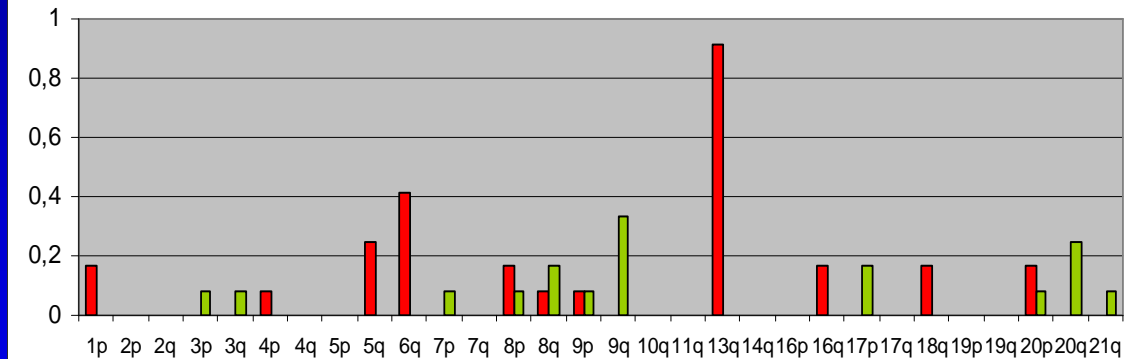
Results III

Cluster 0



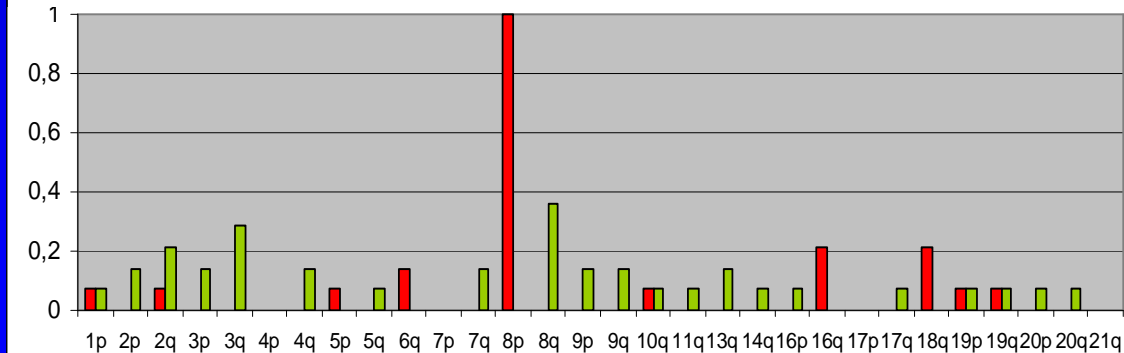
Samples: 34
 Tumor progression: 2
 No progression: 7
 Mean Gleason score: 5.7
 Mean WHO grade: 1.9

Cluster 1



Samples: 12
 Tumor progression: 1
 No progression: 2
 Mean Gleason score: 5.8
 Mean WHO grade: 1.9

Cluster 2



Samples: 14
 Tumor progression: 7
 No progression: 1
 Mean Gleason score: 7.3
 Mean WHO grade: 2.5

■ losses ■ gains

F-c-means

Summary & Conclusion

- Different clustering algorithms (k-means, SOM, F-c-means, random) were used to analyze three valued CGH data from prostate carcinomas.
- A simple measure of clustering robustness based on counting was used to find the configuration which is most robust over different initializations, i.e. find the algorithm which is most robust in comparison to a random clustering, has the lowest quantization error and the lowest number of prototypes.
- Clustering revealed a grouping into different types of malignancy:
 - Cluster 0 had the lowest number of gains and losses.
 - Cluster 1 with approx. the same malignancy as cluster 0 had losses at 13q, 6q and 5q.
 - Cluster 2 consisted of cases with a high degree of malignancy and had most prominently a loss at 8p.
- A loss at 8p seems to have the highest prognostic importance.