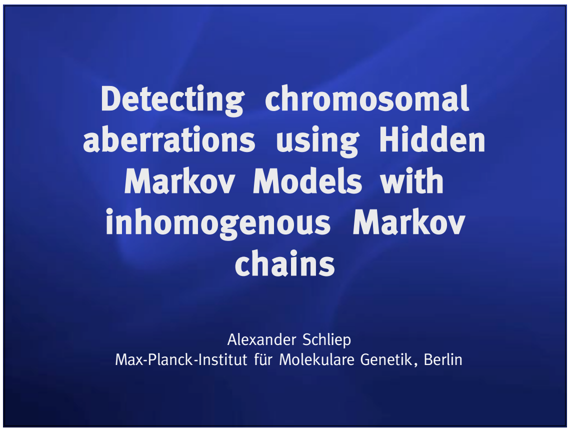
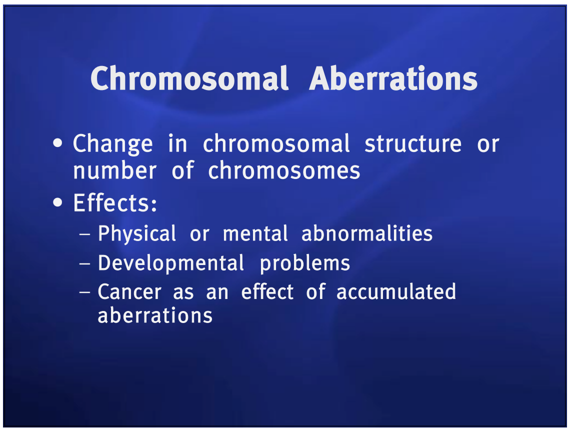


Detecting chromosomal aberrations using Hidden Markov Models with inhomogenous Markov chains

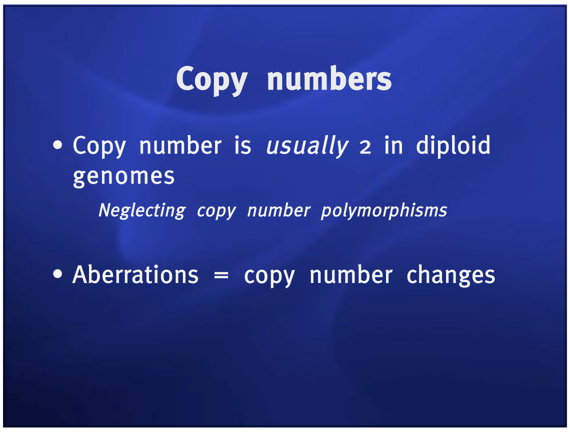
Alexander Schliep
Max-Planck-Institut für Molekulare Genetik, Berlin

Background



Chromosomal Aberrations

- Change in chromosomal structure or number of chromosomes
- Effects:
  - Physical or mental abnormalities
  - Developmental problems
  - Cancer as an effect of accumulated aberrations

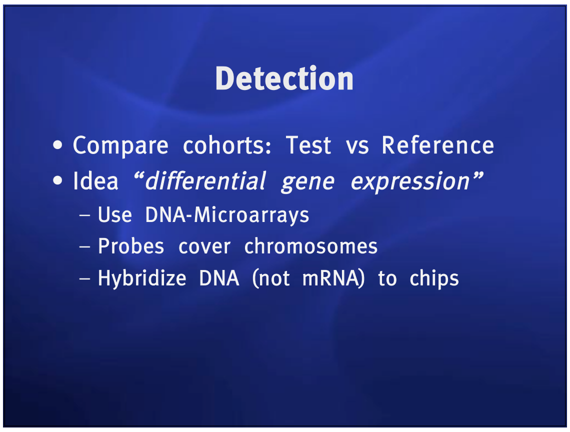


Copy numbers

- Copy number is *usually* 2 in diploid genomes

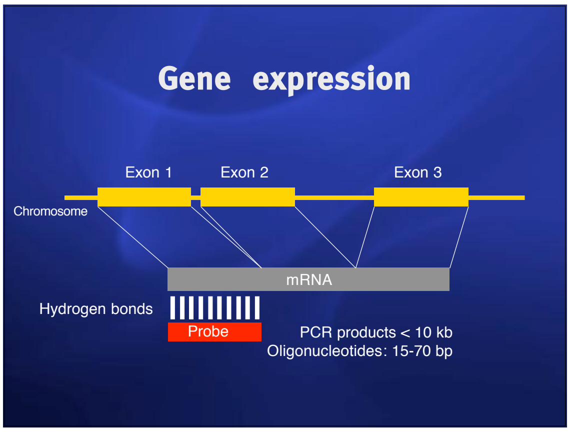
  Neglecting copy number polymorphisms

- Aberrations = copy number changes



Detection

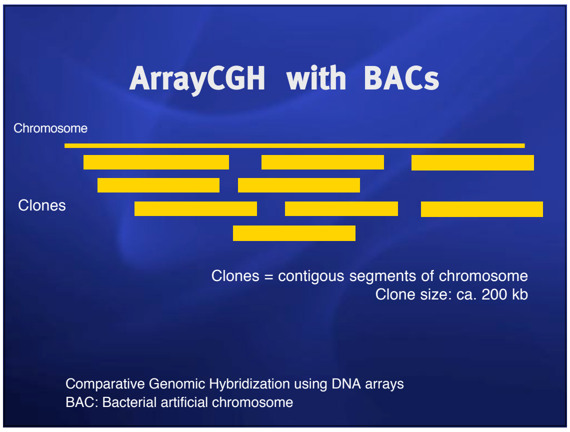
- Compare cohorts: Test vs Reference
- Idea *“differential gene expression”*
  - Use DNA-Microarrays
  - Probes cover chromosomes
  - Hybridize DNA (not mRNA) to chips



Gene expression

Exon 1
Exon 2
Exon 3
Chromosome
mRNA
Hydrogen bonds
Probe

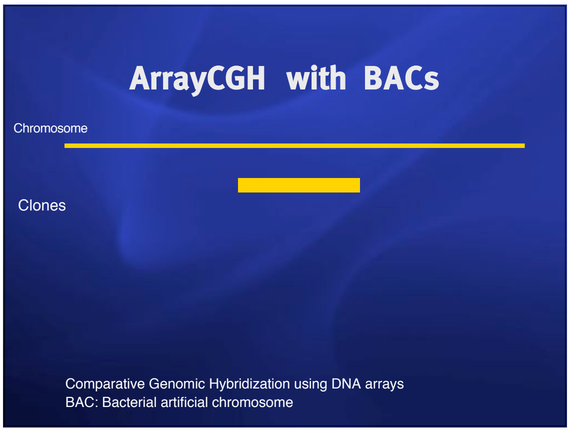
PCR products < 10 kb
Oligonucleotides: 15-70 bp



ArrayCGH with BACs

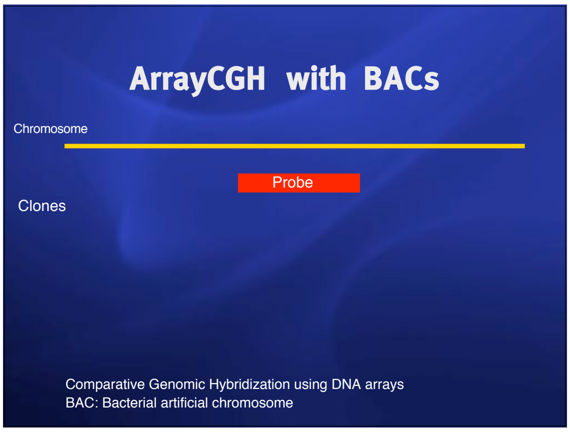
Clones = contigous segments of chromosome
Clone size: ca. 200 kb

Comparative Genomic Hybridization using DNA arrays
BAC: Bacterial artificial chromosome



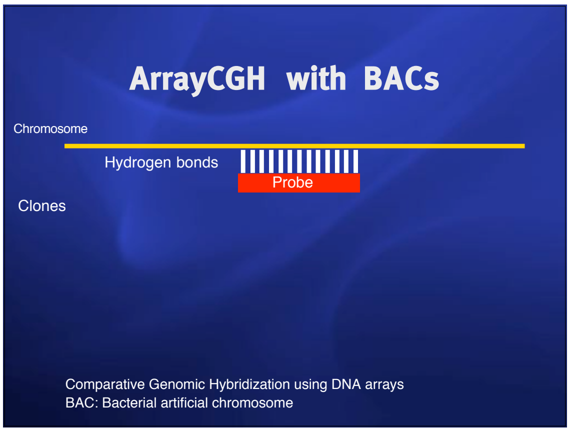
ArrayCGH with BACs

Comparative Genomic Hybridization using DNA arrays
BAC: Bacterial artificial chromosome



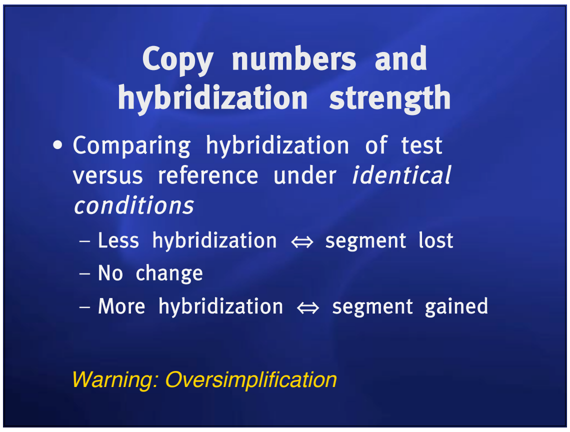
ArrayCGH with BACs

Comparative Genomic Hybridization using DNA arrays
BAC: Bacterial artificial chromosome



ArrayCGH with BACs

Comparative Genomic Hybridization using DNA arrays
BAC: Bacterial artificial chromosome

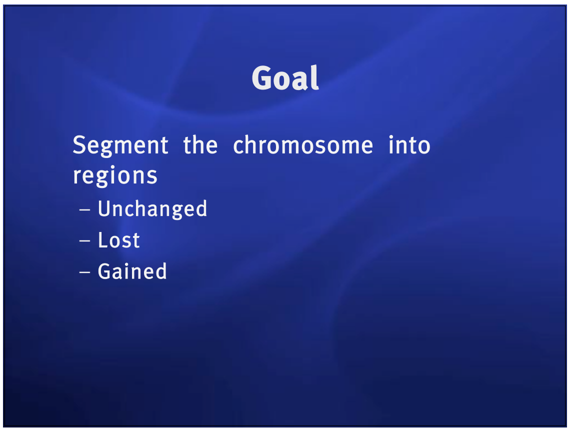


Copy numbers and hybridization strength

- Comparing hybridization of test versus reference under *identical conditions*
  - Less hybridization ⇔ segment lost
  - No change
  - More hybridization ⇔ segment gained

Warning: Oversimplification

Goal

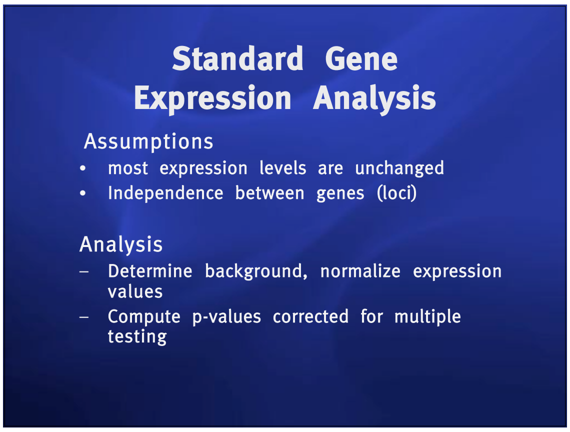


Goal

Segment the chromosome into regions

- Unchanged
- Lost
- Gained

Observations



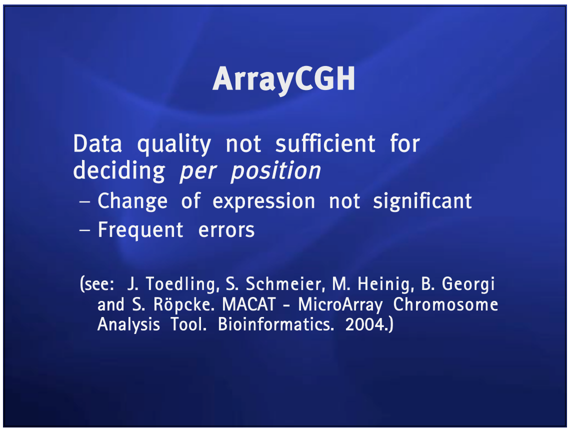
Standard Gene Expression Analysis

Assumptions

- most expression levels are unchanged
- Independence between genes (loci)

Analysis

- Determine background, normalize expression values
- Compute p-values corrected for multiple testing

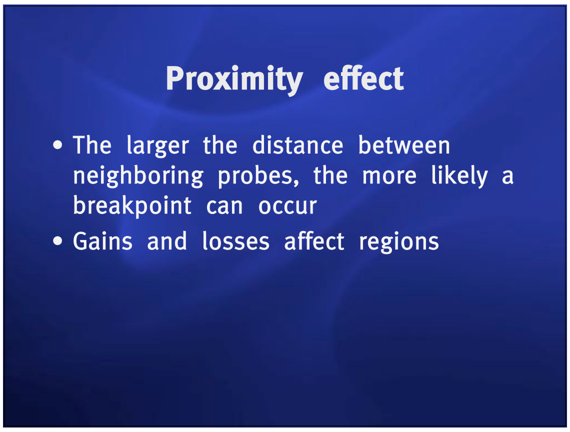


ArrayCGH

Data quality not sufficient for deciding *per position*

- Change of expression not significant
- Frequent errors

(see: J. Toedling, S. Schmeier, M. Heinig, B. Georgi and S. Röpcke. MACAT - MicroArray Chromosome Analysis Tool. Bioinformatics. 2004.)

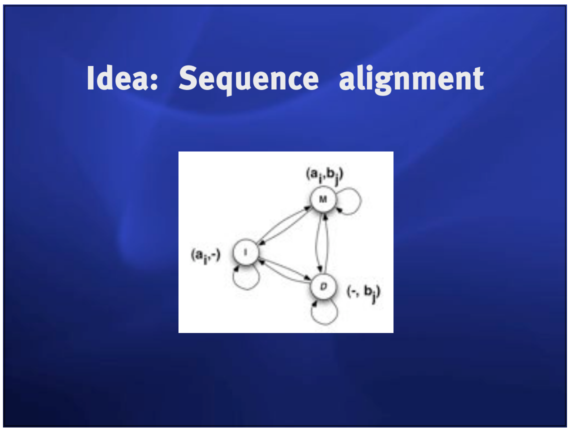


Proximity effect

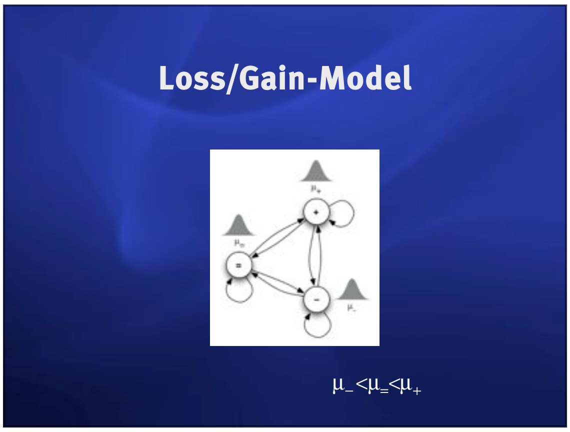
- The larger the distance between neighboring probes, the more likely a breakpoint can occur
- Gains and losses affect regions

Method

Following Jane Frydland (UCSF) et. al.

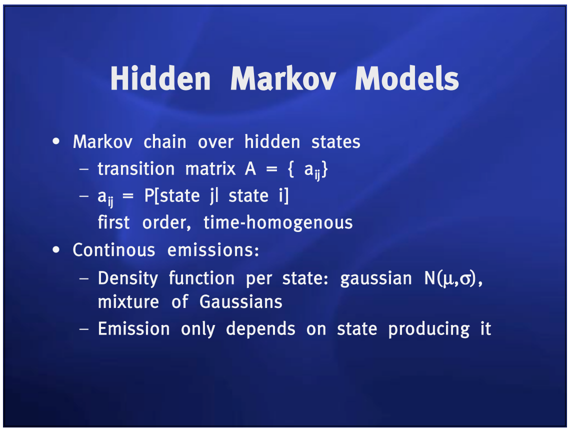


Idea: Sequence alignment



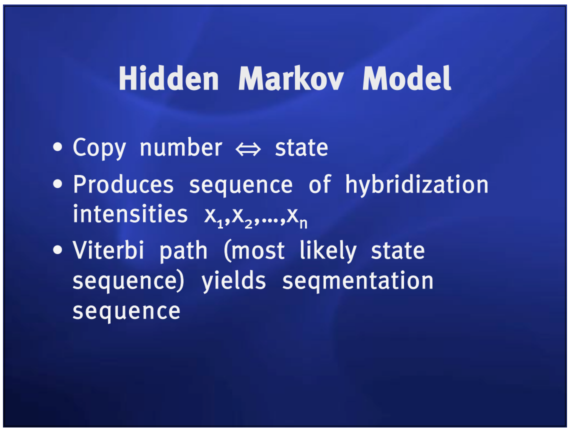
Loss/Gain-Model

$\mu_- < \mu_= < \mu_+$



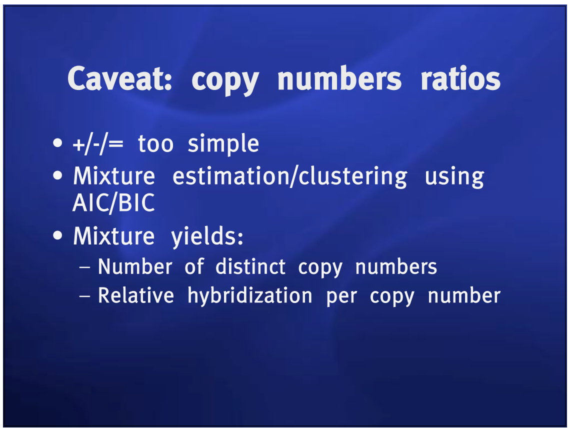
Hidden Markov Models

- Markov chain over hidden states
  - transition matrix A = { a_{ij}}
  - a_{ij} = P[state j| state i]
    first order, time-homogenous
- Continous emissions:
  - Density function per state: gaussian N(μ,σ), mixture of Gaussians
  - Emission only depends on state producing it



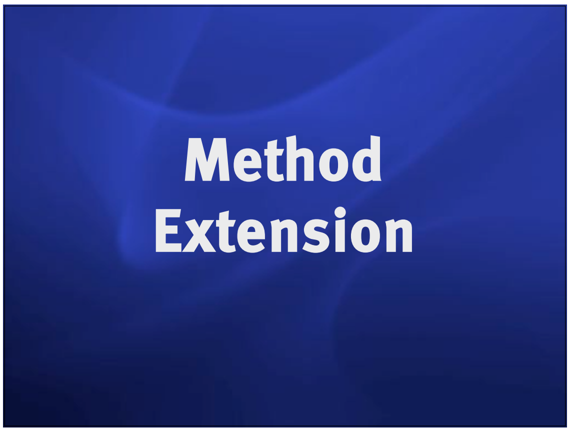
Hidden Markov Model

- Copy number ⇔ state
- Produces sequence of hybridization intensities $x_1,x_2,\ldots,x_n$
- Viterbi path (most likely state sequence) yields seqmentation sequence

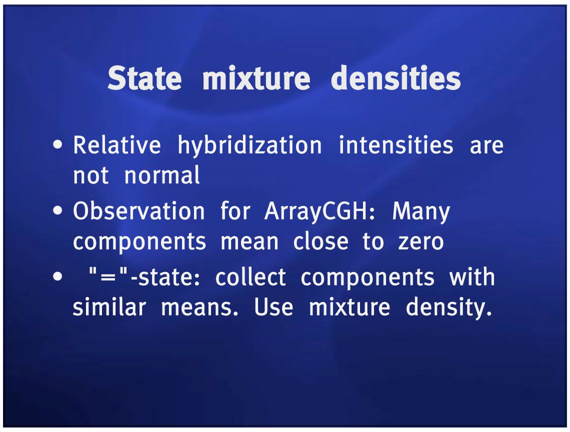


Caveat: copy numbers ratios

- +/-/= too simple
- Mixture estimation/clustering using AIC/BIC
- Mixture yields:
  - Number of distinct copy numbers
  - Relative hybridization per copy number

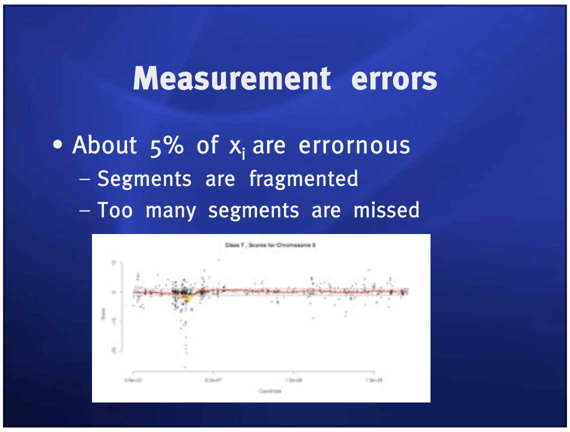


Method Extension



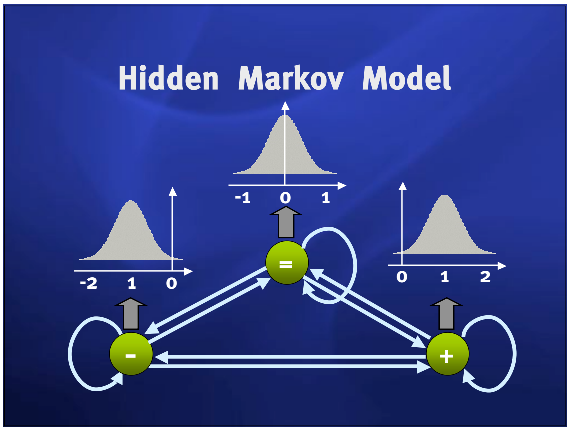
State mixture densities

- Relative hybridization intensities are not normal
- Observation for ArrayCGH: Many components mean close to zero
- "="-state: collect components with similar means. Use mixture density.

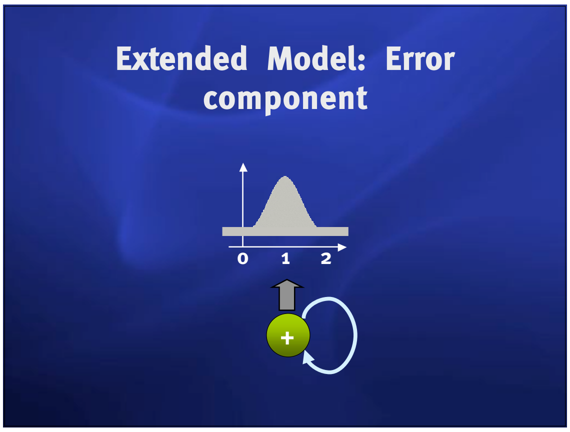


Measurement errors

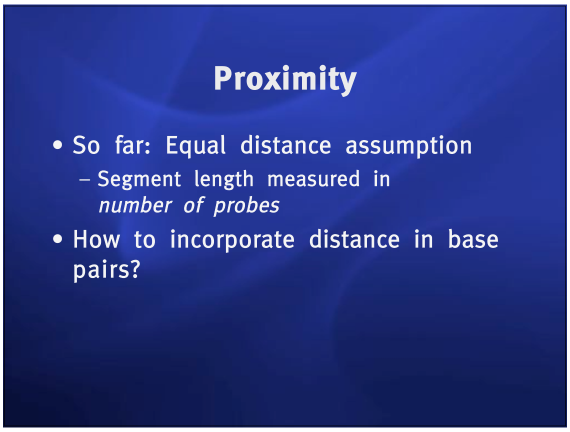
- About 5% of x_i are errornous
  - Segments are fragmented
  - Too many segments are missed



Hidden Markov Model

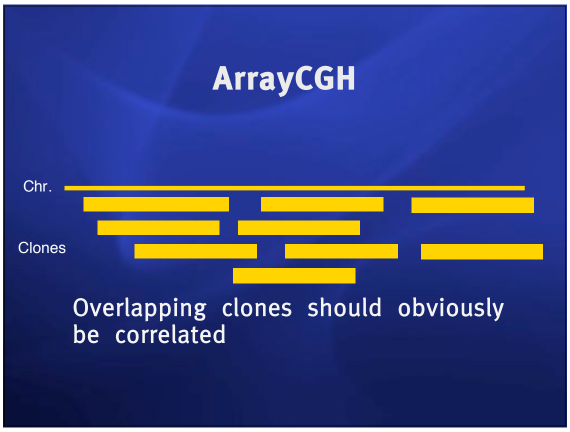


Extended Model: Error component



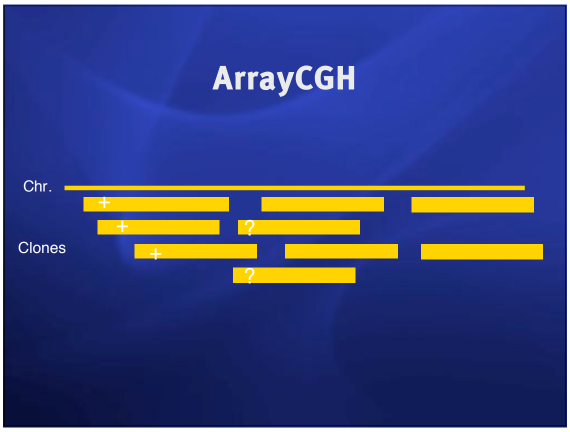
Proximity

- So far: Equal distance assumption
  - Segment length measured in *number of probes*
- How to incorporate distance in base pairs?

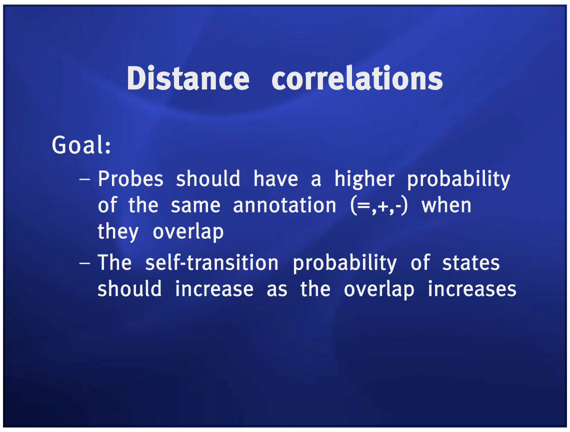


ArrayCGH

Overlapping clones should obviously be correlated



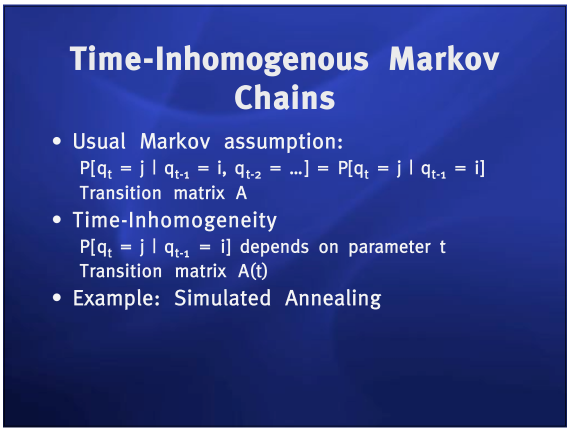
ArrayCGH



Distance correlations

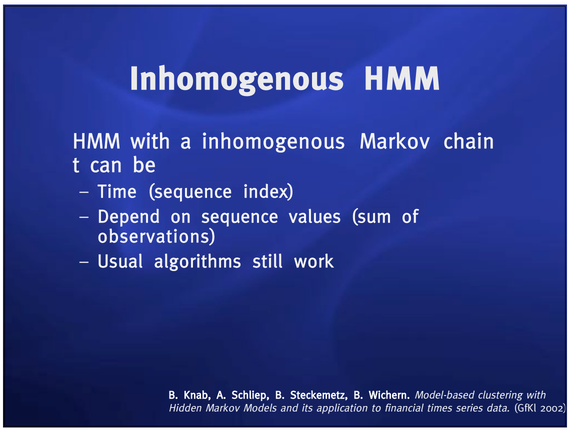
Goal:

- Probes should have a higher probability of the same annotation (=,+,-) when they overlap
- The self-transition probability of states should increase as the overlap increases



Time-Inhomogenous Markov Chains

- Usual Markov assumption:
  $P[q_t = j \mid q_{t-1} = i, q_{t-2} = \dots] = P[q_t = j \mid q_{t-1} = i]$
  Transition matrix A
- Time-Inhomogeneity
  $P[q_t = j \mid q_{t-1} = i]$ depends on parameter t
  Transition matrix A(t)
- Example: Simulated Annealing

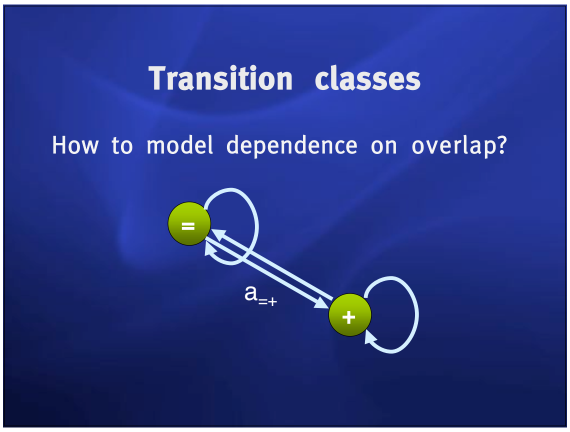


Inhomogenous HMM

HMM with a inhomogenous Markov chain
t can be

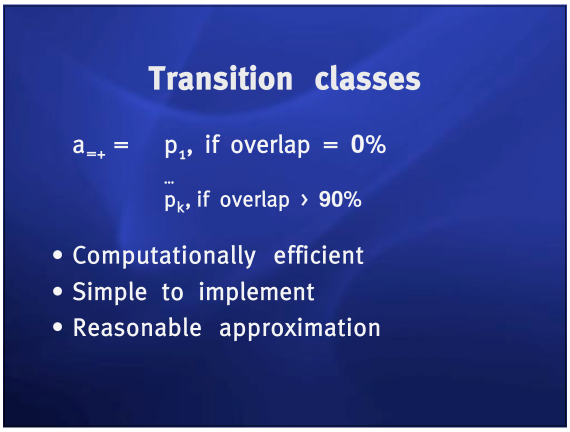
- Time (sequence index)
- Depend on sequence values (sum of observations)
- Usual algorithms still work

B. Knab, A. Schliep, B. Steckemetz, B. Wichern. *Model-based clustering with Hidden Markov Models and its application to financial times series data.* (GfKl 2002)



Transition classes

How to model dependence on overlap?



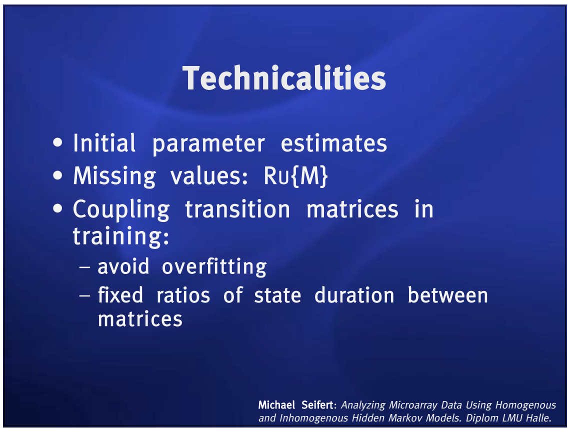
Transition classes

$a_{=+}$ = p_1, if overlap = **0%**

...

p_k, if overlap > **90%**

- Computationally efficient
- Simple to implement
- Reasonable approximation

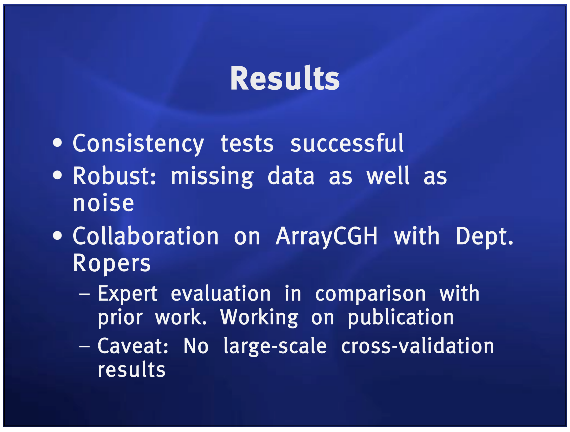


Technicalities

- Initial parameter estimates
- Missing values: R∪{M}
- Coupling transition matrices in training:
  - avoid overfitting
  - fixed ratios of state duration between matrices

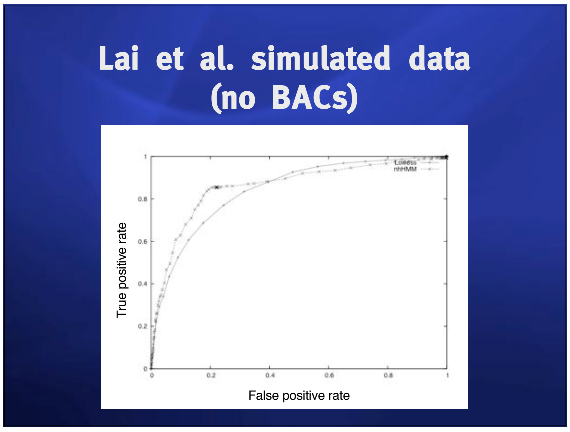
Michael Seifert: *Analyzing Microarray Data Using Homogenous and Inhomogenous Hidden Markov Models. Diplom LMU Halle.*

Discussion

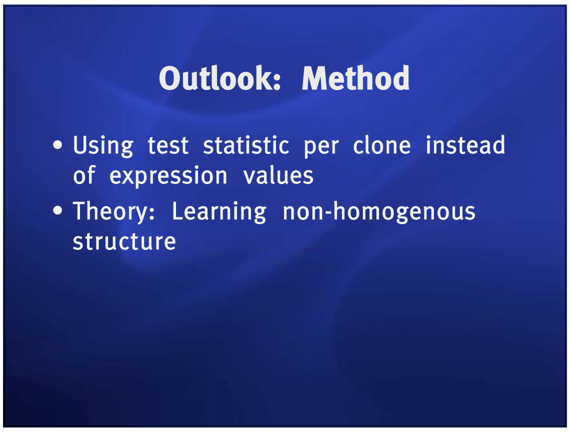


Results

- Consistency tests successful
- Robust: missing data as well as noise
- Collaboration on ArrayCGH with Dept. Ropers
  - Expert evaluation in comparison with prior work. Working on publication
  - Caveat: No large-scale cross-validation results

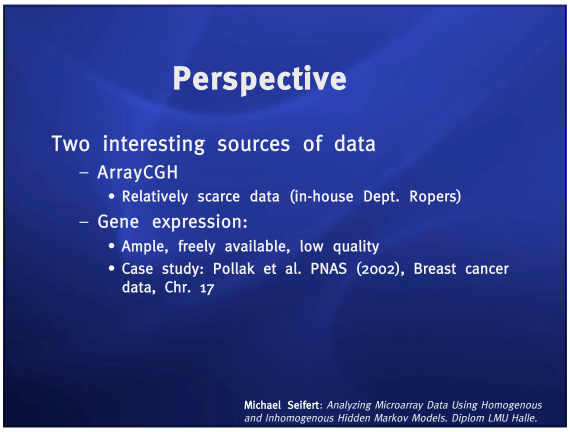


Lai et al. simulated data (no BACs)



Outlook: Method

- Using test statistic per clone instead of expression values
- Theory: Learning non-homogenous structure



Perspective

Two interesting sources of data

- ArrayCGH
  - Relatively scarce data (in-house Dept. Ropers)
- Gene expression:
  - Ample, freely available, low quality
  - Case study: Pollak et al. PNAS (2002), Breast cancer data, Chr. 17

Michael Seifert: *Analyzing Microarray Data Using Homogenous and Inhomogenous Hidden Markov Models. Diplom LMU Halle.*

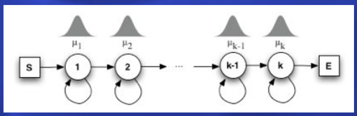
Perspective

- Combination of ArrayCGH and gene expression:
  - Regulation in presence of aberrations

k-Segmentation (Picard)

Find k Normals and k-1 segment boundaries maximizing the obvious likelihood function for $x_1,x_2,\ldots,x_n$

k-Segmentation (Picard)



- All transition probabilities are 1/2
- Assuming homogenous variance σ
- Can show: Maximum likelihood HMM and Viterbi-path converge to solution of k-Segmentation problem when σ goes to zero

Acknowledgements

- Wei Chen, MPI Molgen: ArrayCGH
- Stefan Röpcke, MPI Molgen
- Michael Seifert, LMU Halle
- GHMM: Wasinee Rungsarityotin, Benjamin Georgi, Alexander Schönhuth, Ben Rich, Matthias Heinig, Alexander Riemer, Janne Grunau
- Jane Frydland for helpful discussions

Thanks

http://ghmm.org

http://algorithmics.molgen.mpg.de