Introduction to array CGH analysis

10-810, CMB lecture 11---Eric Xing

Tumor Cell

Chromosomes of tumor cell:

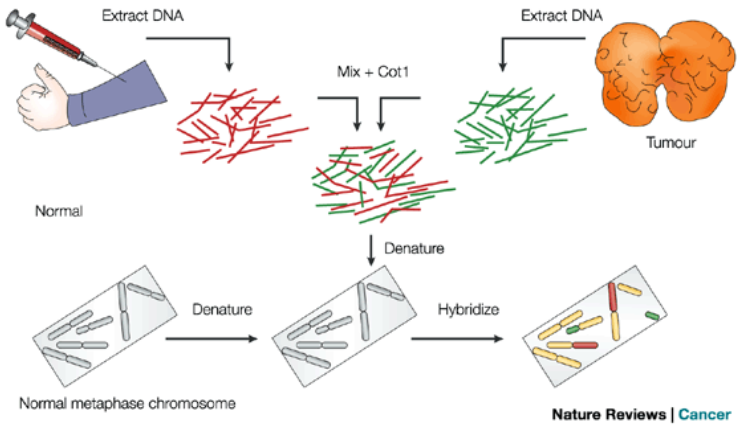
Significance of DNA copy number alternations in cancer studies

- Genetic alternations such as amplifications and deletions frequently contribute to tumorigenesis. These alternations changes the level of gene expression which modify normal growth control and survival pathways.
- Characterization of these DNA copy-number changes is important for both the basic understanding of cancer and its diagnosis.
- E.g., one of the characteristics of breast cancer is the heterogeneity in aberrations that are found.

What is CGH ?

- Comparative genomic hybridization (CGH) was developed to survey DNA copy-number variations across a whole genome.
- With CGH, differentially labeled test (i.e. tumor) and reference (i.e. normal individual) genomic DNAs are co-hybridized to normal metaphase chromosomes, and fluorescence ratios along the length of chromosomes provide a cytogenetic representation of the relative DNA copy-number variation.
- Chromosomal CGH resolution is limited to 10-20 Mb – therefore, anything smaller than that will not be detected.

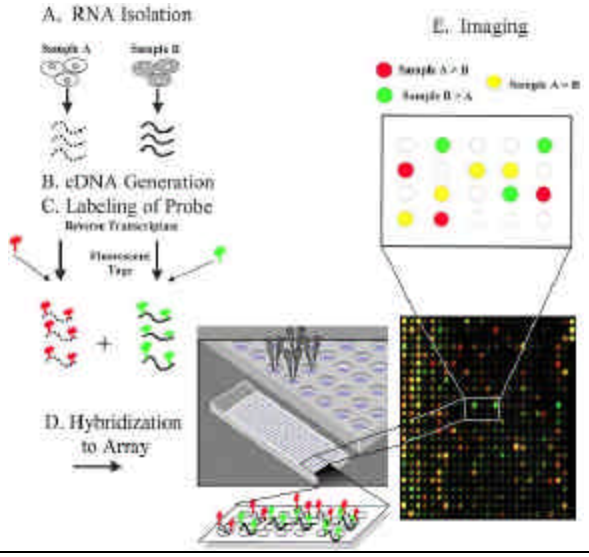
Schematic representation of CGH technique



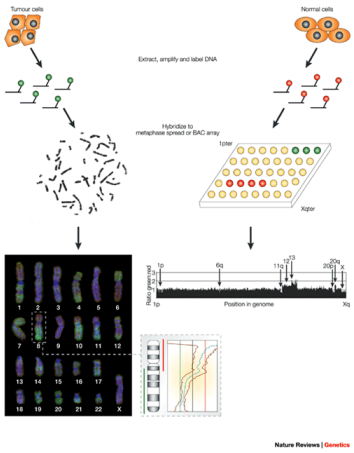
Array CGH

- In array CGH arrays of genomic BAC,P1,cosmid or cDNA clones are used for hybridization instead of metaphase chromosomes in conventional CGH technique.
- Fluorescence ratios at arrayed DNA elements provide a locus-by-locus measure of DNA copy-number variation, represents a means of achieving increased mapping resolution.

Schematic representation of Array CGH



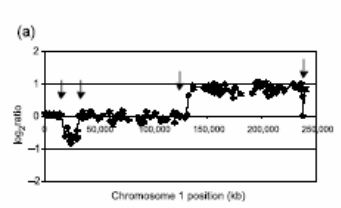
CGH compared to Array CGH



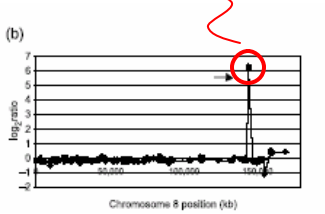
The basic assumption of a CGH experiment is that the ratio of the binding of test and control DNA is proportional to the ratio of the concentrations of sequences in the two samples.

Copy number aberration types in breast cancer cell lines

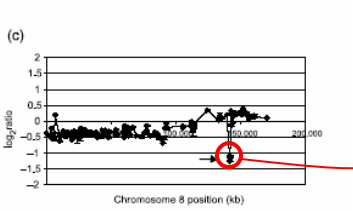
60-70 fold amplification of CMYC region



Copy number profile for chromosome 1 from 600 MPE cell line



Copy number profile for chromosome 8 from COLO320 cell line



Copy number profile for chromosome 8 in MDA-MB-231 cell line

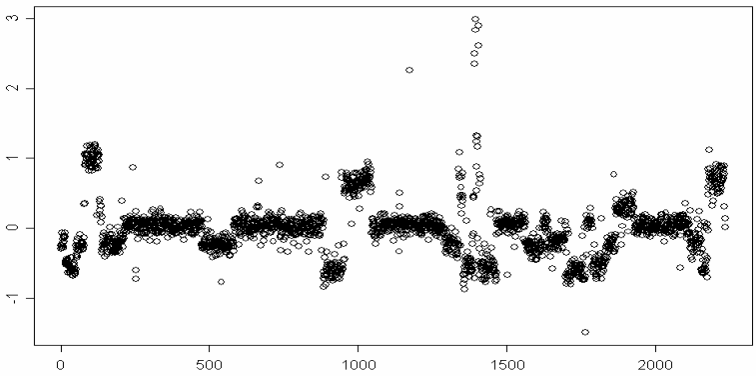
deletion

Clinical significance

- Based on the results better tests can be performed that measure the DNA copy number of a target gene.
- Monitor cancer progression and distinguish between mild and metastatic censorious lesions using FISH (Florescence in situ hybridization) probes on regions of recurrent copy number aberrations in several tumor types.
- It can be used to reveal more regional copy number markers that can be used for cancer prediction.
- Identifying and understanding the genes that are involved in cancer will help to design therapeutic drugs that target the dysfunction genes and/or avoid therapies that cause tumor resistance.

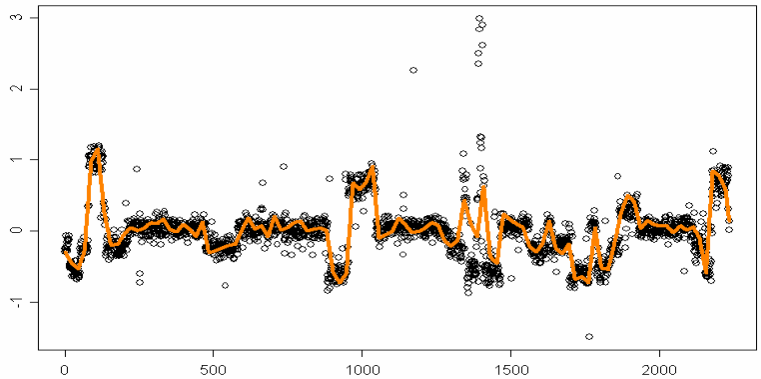
Analyzing CGH Data

⇑ Copy #

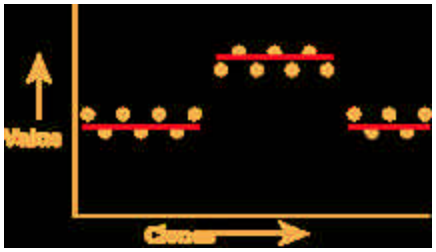


Clones/Chromosomes ⇒

Naïve Smoothing



"Discrete" Smoothing



Copy numbers are integers

Why Smoothing ?

- Noise reduction
- Detection of Loss, Normal, Gain, Amplification
- Breakpoint analysis

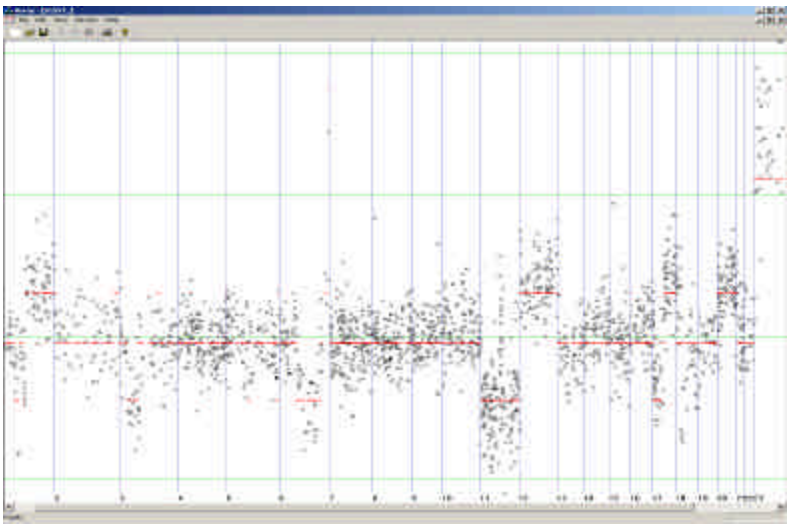
Recurrent (over tumors) aberrations may indicate:

- an oncogene or
- a tumor suppressor gene

Is Smoothing Easy?

- Measurements are relative to a reference sample
- Printing, labeling and hybridization may be uneven
- Tumor sample is inhomogeneous
  - vertical scale is relative
  - do expect only few levels

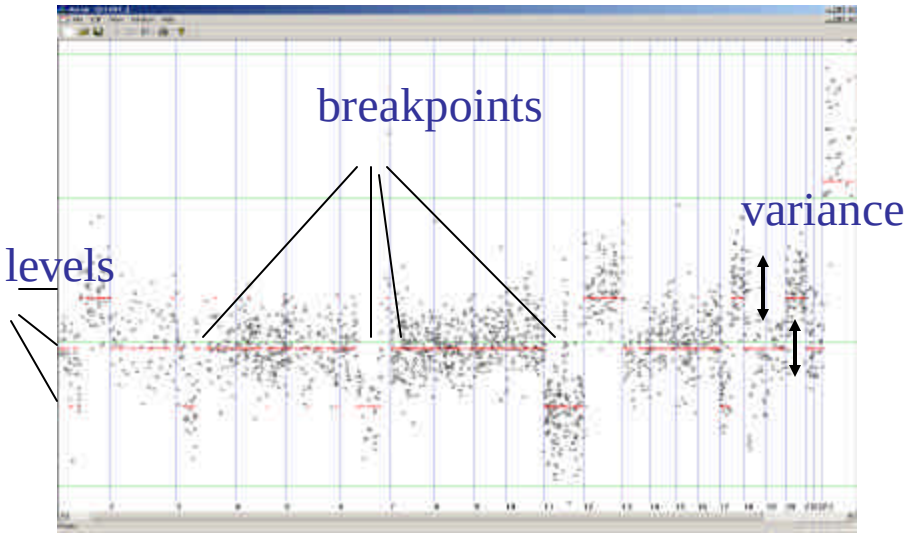
Smoothing: example



Problem Formalization

- A smoothing can be described by
  - a number of breakpoints
  - corresponding levels
- A fitness function scores each smoothing according to fitness to the data
- An algorithm finds the smoothing with the highest fitness score.

Smoothing



Fitness Function

We assume that data are a realization of a Gaussian noise process and use the maximum likelihood criterion adjusted with a penalization term for taking into account model complexity

We could use better models given insight in tumor pathogenesis

Fitness Function (2)

CGH values: $x_1, \ldots, x_n$

breakpoints: $0 < y_1 < \cdots < y_N < x_n$

levels: $m_1, \ldots, m_N$

error variances: $s_1^2, \ldots, s_N^2$

$$(x_1, \ldots, x_{y_1}), (x_{y_1+1}, \ldots, x_{y_2}), \ldots, (x_{y_N+1}, \ldots, x_n)$$

likelihood:

$$\prod_{i=1}^{y_1} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2} \cdots \prod_{i=y_N+1}^{n} \frac{1}{\sigma_{N+1}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_{N+1}}{\sigma_{N+1}}\right)^2}$$

Fitness Function (3)

Maximum likelihood estimators of m and s^2 can be found explicitly

Need to add a penalty to log likelihood to control number N of breakpoints

penalty

$$f(y_1, \ldots, y_N) = \sum_{i=1}^{N+1} (y_{i+1} - y_i) \log \hat{\sigma}_i + \lambda N$$

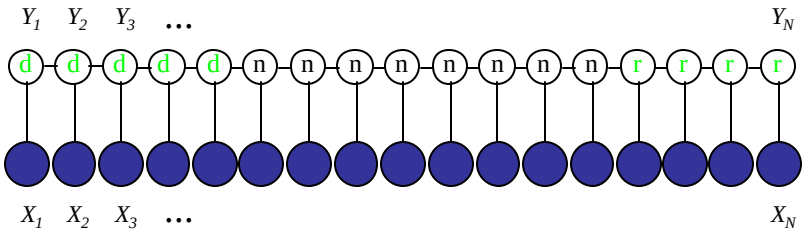
Breakpoint Detection

- Identify possibly damaged genes:
  - These genes will not be expressed anymore
- Identify recurrent breakpoint locations:
  - Indicates fragile pieces of the chromosome
- Accuracy is important:
  - Important genes may be located in a region with (recurrent) breakpoints

Algorithms

- Maximizing Fitness is computationally hard
- Genetic algorithm + local search was used to find approximation to the optimum (Kees Jong, et. al.)
- Can we do it more elegantly?

Rocall the HMM



- The shaded nodes represent the observed nucleotides at particular sites of an organism's genome
- For discrete Y_i, widely used in computational biology to represent segments of sequences
  - Then, gene finders and motif finders
  - Now, hybridization signal

Unsupervised HMM partition

(Jane Fridlyand et al)

- K , the number of states in the model. The states are hidden and, generally, physically meaningful. Typically, the states are interconnected in a way that any state can be reached from any other state. We denote the individual states as $S=S_1,...,S_K$, and the state at the location l as s_l, $1 \le l \le L$.

- The initial state distribution $\pi=\{\pi_k\}$ where

$$\pi_k = P\{s_1 = S_k\}, 1 \le k \le K.$$

- The state transition probability distribution $A=\{a_{mp}\}$ where

$$a_{mp} = P\{s_{l+1} = S_p | s_l = S_m\}, 1 \le m, p \le K$$

Unsupervised HMM partition, cont.

- The emission distribution or probability density function $B=\{b_k(O)\}$ where

$$\{b_k(\mathbf{O})\} = \mathcal{G}(\mathbf{O}, \mu_k, \mathbf{U}_k), 1 \le k \le K.$$

Where O is the vector being modeled, *G* is Gaussian density with mean vector μ_k and covariance matrix U_k. More generally *G* is any log-concave or elliptically symmetric density and the probability density function, {bk(O)} is a finite mixture (see Rabiner (1989)).

The Forward Algorithm

We can compute $f_k(t)$ for all k, t, using dynamic programming!

Initialization:

$f_0(0) = 1$
$f_k(0) = 0$, for all $k > 0$

Iteration:

$f_l(t) = e_l(\mathbf{X}_t) \Sigma_k f_k(t-1) a_{kl}$ (a_{0k} is a vector of initial probability)

Termination:

$P(x) = \Sigma_k f_k(T)$

The Backward Algorithm

We can compute $b_k(t)$ for all k, t, using dynamic programming

Initialization:

$b_k(T) = 1$, for all k

Iteration:

$b_k(t) = \Sigma_l e_l(\mathbf{X}_{t+1}) a_{kl} b_l(t+1)$

Termination:

$P(x) = \Sigma_l a_{0l} e_l(x_1) b_l(1)$

A-CGH vs. Expression

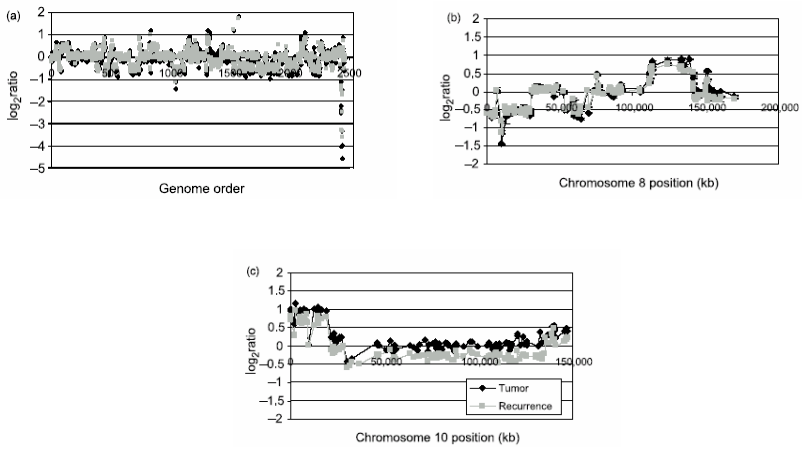
a-CGH

- DNA
  - In Nucleus
  - Same for every cell
- DNA on slide
- Measure Copy Number Variation

Expression

- RNA
  - In Cytoplasm
  - Different per cell
- cDNA on slide
- Measure Gene Expression

Comparison of a tumor and recurrence by CGH



Clinical significance

- Based on the results better tests can be performed that measure the DNA copy number of oncogenes and TSGs.
- Monitor cancer progression and distinguish between mild and metastatic censorious lesions using FISH (Florescence in situ hybridization) probes on regions of recurrent copy number aberrations in several tumor types.
- It can be used to reveal more regional copy number markers that can be used for cancer prediction.
- Identifying and understanding the genes that are involved in cancer will help to design therapeutic drugs that target the dysfunction genes and/or avoid therapies that cause tumor resistance.