Introduction to Array CGH

Jane Fridlyand

CBMB
University of California, San Francisco
Genentech Hall Auditorium, Mission Bay, UCSF
October 23, 2004

Why Study DNA In Tumors?

• It is the fundamental repository of information.

• If the same DNA aberration occurs repeatedly in tumors, how can one ignore it?

• There are powerful, general methods of assessing certain types of aberrations.

• DNA is relatively robust and can be assayed specimens that have been treated in multiple ways, including archival tissue from hospital laboratories.

(One must remember that DNA is not the whole story.)
A Variety of Genetic Alterations Underlie Developmental Abnormalities and Disease

- “Point” mutation – change of one or a few bases -- leads to altered protein or change in expression level.
- Loss of gene copy reduces expression level. (tumor suppressor loss)
- Gain of gene copies increases expression level. (oncogene activation)
- (De)Methylation of gene promoters (increase) decrease expression level. ((oncogene) tumor suppressor)
- Breaking and abnormal rejoining of DNA makes novel genes.

Some examples of Copy Number Changes in Cancer

Copy number changes usually involve a DNA segment that is substantially larger than the critical gene(s)

Mutation plus terminal deletion of tumor a suppressor gene

Amplification of an oncogene and surrounding DNA; extra copies can be located anywhere in the genome
Uses of Aberration Mapping

Positional refinement → Gene identification

Aberration mapping → Genetics of Progression, Diagnosis, Staging, Prognosis, Therapy selection

Gene-specific therapy

Candidate gene / pathway (the rest of biology)

Recurrent Regions of Copy Number Change Identify Breast Cancer Genes

ERBB2

CCND1  GARP

Chromosome 17 Position, kb

Chromosome 11 Position, kb

Two Color Hybridization Allow Efficient Detection of Translocations (SKY/M-FISH)

- Cheap
- Can characterize complex rearrangements
- ~1 Mb resolution for translocation location
- Low resolution (> 10Mb) for rearrangement locations
- Unable to identify dosage alterations with acceptable precision

Array CGH

**Goals:**

- Assembly of the tiling array containing about 30,000 clones for genome-wide scanning. Current array is 2,400 clones with approximately 1 Mb resolution in human.
- Quantitatively detect single copy gains and losses and high level amplifications in an experiment
- Detect these aberrations on a single clone

*So, very high density effective resolution*
Array CGH Maps DNA Copy Number Alterations to Positions in the Genome

Test Genomic DNA

Reference Genomic DNA

Cot-1 DNA

Gain of DNA copies in tumor

Loss of DNA copies in tumor

Array CGH Analysis of a Tumor Genome

Gain

Loss

Microarrays: Case Studies and Advanced Analysis
Microarrays: Case Studies and Advanced Analysis

Examples of the array CGH profiles

Samples/Conditions:
- Gene 1
  - 1:
    - 0.46
    - 0.30
    - 0.80
    - 1.51
  - 2:
    - -0.10
    - 0.49
    - 0.24
    - 0.06
  - 3:
    - 0.15
    - 0.74
    - 0.04
    - 0.10

Microarrays: Case Studies and Advanced Analysis
Tumor copy number profiles are a reflection of two processes

- Selection for alterations in gene expression that favor tumor development. Selective advantage to maintain set of aberrations.

- Mechanisms of genetic instability promoting changes in the genome. (initiating oncogenetic event in murine models and methotrexate resistance in MMR deficient and proficient cell lines)

Can tumor classes be identified by the types and frequencies of different copy number aberrations?

Tumor Genomes are Stable
Copy Number Profiles of a Tumor & Recurrence

Initiating oncogenic event determines gene-expression patterns of human breast cancer models

Molecular expression profiling of tumors initiated by transgenic overexpression of c-myc, c-neu, c-ha-ras, polyoma middle T antigen (PyMT) or simian virus 40 T-antigen (TSa) targeted to the mouse mammary gland have identified both common and oncogene-specific events associated with tumor formation and progression. The tumors shared great similarities in their gene-expression profiles as compared with the normal mammary gland with an induction of cell-cycle regulators, metabolic regulators, zinc finger proteins, and protein tyrosine phosphatases, along with the suppression of some protein tyrosine kinases. Selection and hierarchical clustering of the most variant genes, however, resulted in separating the mouse models into three groups with distinct oncogene-specific patterns of gene expression. Such an identification of targets specified by particular oncogenes may facilitate development of lesion-specific therapeutics and preclinical testing. Moreover, similarities in gene expression between human breast cancers and the mouse models have been identified, thus providing an important component for the validation of transgenic mammary cancer models.

Note the three main clusters consisting of (a) MMTV-myc; (b) T-antigen group and (c) of the MMTV-neu, MMTV-ras and PyMT groups.

Need for automation

The manual process of characterizing individual genomic profiles is time-consuming, prone to human error and non-reproducible. We present an automated method for identifying and characterizing DNA copy number changes in a given sample.

Previously, it was possible to evaluate results of FISH and chromosomal CGH for overall number of genomic aberrations or to search for recurrent changes. The more detailed taxonomy has emerged only recently since the use of genome-wide array CGH allows one to accurately determine the number, types and breakpoints of DNA copy number aberrations throughout the genome.
Types of genomic changes

- Copy number transitions (DNA double stranded breaks)
- Whole chromosomal changes (failures of karyokinesis (mitosis) or cytokinesis)
- Focal aberrations (localized double stranded breaks)
- High-level focal amplifications (localized event replicated multiple times)
**Goals and underlying model**

**Goal:** To partition the clones into sets with the same copy number and to characterize the genomic segments.

**Biological model:** genomic rearrangements lead to gains or losses of sizable contiguous parts of the genome, possibly spanning entire chromosomes, or, alternatively, to focal high-level amplifications.

Want to make use of the physical dependence of the nearby clones, which translates into copy number dependence.

---

**An approach ignoring spatial dependence**

For k between 1 and K, repeat the following:

1. For a pre-specified number of components, k, fit a Gaussian mixture model to the observed log2ratios.

2. Compute log-likelihood of the fit.

3. Choose optimal number of components using a model selection criterion, e.g. Schwartz BIC.

*This approach will not work most of the time.*
Observed clone value and spatial coherence

Hidden Markov models (LE Baum et al 1970)

Markov model:

Hidden Markov model:

**Hidden State**: H/L pressure
Markov process

**Observation**: Sunny/Rainy
State dependent
(Hidden) Markov Model

State = underlying copy number; Observation = observed log2ratio

Outline of an algorithm

1. Partition clones on individual chromosomes into sets with the same underlying copy number using unsupervised Hidden Markov Model partitioning. Often the copy number gain or loss can be inferred by thresholding log2 ratios in a given sample.

2. Characterize individual chromosomes according to whether there have been any copy number transitions present or whole chromosome has been gained or lost. We also look for focal aberrations corresponding to the individual clones such as low level gains and losses or high level amplifications.

Refs: Fridlyand et al., 2004, JMVA
Why NOT estimate copy number directly?

Confounding by many experimental and biological quantities such as purity and ploidy of sample.

• Normal cell contamination (<50% of the normal cells)

• Aneuploidy (tumor is not diploid. We approximate the ploidy by the median copy number of the loci represented on the array)

• Heterogeneity (not all cells acquired a given aberration)

Fitting k-state HMM

• Any state is reachable from any state
• Emissions are distributed normally with state-specific mean and overall variance
• Re-estimation is done with forward-backwards algorithm
• Initialize state means with k-means partitioning
• Transition probabilities are set so staying in the same state is most likely and the rest is split equally among possible transitions
• Initial probabilities are initialized so the largest weight falls on “no change” state (with the mean closest to 0)
Algorithm, Part I

Algorithm 1: Segment clones into sets with the same underlying copy number

1. For \( k = 1 \ldots K \) states:
   1. Fit \( k \)-state HMM
   2. Calculate penalized negative log-likelihood \( \psi(k) \).
2. Choose the model corresponding to the number of states with the smallest \( \psi(k) \), \( k^* = \text{argmin}_k \psi(k) \).
3. If \( k^* = 1 \), then STOP
4. Calculate the median for each state and identify the two states whose medians are the closest. Compute \( d = \min_{k_1 \neq k_2} |\text{med}_{k_1} - \text{med}_{k_2}| \).
5. While \( d < \text{threshold} \):
   1. Merge the two closest states
   2. Recompute \( d \).

Concurrent Approaches

**Olshen and Venkatraman**: novel modification of binary segmentation, *circular binary segmentation*. Looks for change points along each chromosome which represents the copy number transitions.

**Jong et al**: *genetic local search algorithm* to best segment clones into clusters.

Some others.

All existing methods amount to unsupervised, location-specific partitioning and operate on individual chromosomes.
Defining Genomic Alterations

Recall that we characterize tumor profiles using 4 types of genomic alterations:

• Copy number transitions (DNA double stranded breaks)
• Whole chromosomal changes (failures of karyokinesis or cytokinesis)
• Focal low-level aberrations (localized double stranded break)
• Focal high-level amplifications (localized event replicated multiple times)

After partitioning clones into the states, we need to define parameters allowing us to call alteration types.

Algorithm: PART II

• Estimate sample standard deviation
  (median of MADs of clones belonging to the same state. The selection of clones is robust.)

• Identify outliers
  (individual clones whose values differs from the median value of the state containing them. May indicate mismapped or problematic clones)

• Find focal aberrations
  (individual clones whose states differ from the state of the neighboring clones. May indicate mismapped clones or copy number polymorphisms.)
Algorithm: PART II (Cont)

• Determine *copy number transitions*  
(exclude focal aberrations and place copy number transitions between two regions whose states differ. The location between the last clone of the first region and the first clone of the second region is of interest.)

• Find *whole chromosomal changes*  
(chromosomes with no copy number transitions whose median significantly differs from 0 and is above a threshold)

• Identify *focal amplifications*  
(narrow focal regions generated by an event happening more than once)

Examples of the alterations

Refs: Snijders et al., 2003, Oncogene.
When true alterations are known …

Coriel cell line data

The data consists of single array CGH experiments on 15 fibroblast cell lines containing cytogenetically mapped partial or whole-chromosome aneuploidy.

We use this dataset as a proof of principle to demonstrate the feasibility of the HMM approach to structure discovery. We also perform limited simulations with several chromosomes from these data.

Application of the HMM to Coriel Cell lines

Relationship between underlying genetic instability and copy number is poorly understood for most solid tumors

Mismatch repair (MMR) cell line data

- 10 MMR deficient and 10 MMR proficient cell lines. The deficiency is for MLH1 and MSH2/MSH6 complex.

- Tumors with defective Mismatch Repair (MMR)
  - Defects in MMR genes (MLH1, MSH2, MSH6)
  - High level of microsatellite instability

Application to MMR cell lines

Expectation:
MMR deficient cell lines promote growth by mutation which is not corrected by MMR. Thus, MMR def cell lines will have fewer alterations than MMR prof cell lines.

Conformation:
Cytogenetically, it has been shown MMR def cell lines have fewer alterations than MMR proficient cell lines. We show that this holds for all types of alterations but whole chromosomal changes. Thus, we indirectly demonstrate utility of our algorithm.

Data-generated hypothesis:
Cell lines deficient for MLH1 show more copy number alterations than cell lines deficient for MSH2/MSH6.
Copy Number Aberrations in Mismatch Repair Deficient and Proficient Cell Lines

Summary

- Automatic computational approaches are necessary for the analysis of the aCGH profiles.
- Different copy number phenotypes associated with different tumors are likely to reflect both selection and genetic instability.
Software

- aCGH (R/BioConductor) (Fridlyand and Dimitrov)
- DNAcopy (R/Bioconductor) (Olshen et al)
- Other software is available:
  (Stanford, Princeton, Netherlands, France)

Acknowledgements

Dan Pinkel  
Donna Albertson  
Boris Bastian  
Fred Waldman  
Antoine Snijders  
John Curtin  

Ajay Jain  
Taku Tokuyasu  
Chris Kingsley  
Peter Dimitrov  
Adam Olshen  
Terry Speed

Many other UCSF Cancer Center researchers