Statistical and computational challenges for population-based segmentation of copy-number profiles

Guillem Rigaill\textsuperscript{1}, Vincent Miele\textsuperscript{2} and Franck Picard\textsuperscript{2}

\textsuperscript{1}Statistique et Génome, UMR CNRS 8071, USC INRA Université d’Evry, France
\textsuperscript{2} Biométrie et Biologie Evolutive, UMR CNRS 5558 Université Lyon 1, France

Centre Blaise Pascal, Lyon November 2013
Outline

1. Introduction
2. Statistical Analysis of single profiles
3. Statistical analysis of multiple profiles
4. Adaptation to high dimensional computing
Karyotype and chromosome copy numbers

- Gene copy number is tightly regulated
- Humans: 22 pairs (autosomal) + 1 pair sexual chromosomes
- At the chromosomal resolution, the karyotype is a visual tool to check for abnormalities
- Deviations from the reference number (2) result in massive disorders
Karyotype and chromosome copy numbers

- Gene copy number is tightly regulated
- Humans: 22 pairs (autosomal) + 1 pair sexual chromosomes
- At the chromosomal resolution, the karyotype is a visual tool to check for abnormalities
- Deviations from the reference number (2) result in massive disorders
Karyotype and chromosome copy numbers

- Gene copy number is tightly regulated
- Humans: 22 pairs (autosomal) + 1 pair sexual chromosomes
- At the chromosomal resolution, the karyotype is a visual tool to check for abnormalities
- Deviations from the reference number (2) result in massive disorders

Human karyotype (from NSF)
Karyotype and chromosome copy numbers

- Gene copy number is tightly regulated
- Humans: 22 pairs (autosomal) + 1 pair sexual chromosomes
- At the chromosomal resolution, the karyotype is a visual tool to check for abnormalities
- Deviations from the reference number (3) result in massive disorders

Human karyotype (from NSF)
Sub-Chromosomal Aberrations

- deletion
- duplication
- translocation

Mapping aberrations at low resolution has been a technical challenge in cytogenetics.
Mapping using Fluorescent In Situ Hybridization

- Consider a known sequence of \( \sim 1 \text{ Mb} \) and link it with a fluorochrome
- Mix it in presence of denatured chromosomes
- Check if the probe hybridizes somewhere
- If the probe comes from another chromosome, map the aberration

Going FISHing? (from NSF)
Multicolor Fish and Comparative Genomic Hybridization

- Consider a set of reference sequences of size $\sim x\text{Mb}$
- Link them with different fluorochromes
- Mix in presence of denatured chromosomes
- Check if the probes hybridize somewhere
- If the probe comes from another chromosome, map the aberrations
Application of the microarray technology to CGH

- The microarray technology was mainly developed for expression data. Application to CGH in 2003 (array-CGH)
- Probes are $\sim$ kbs long and fixed on a glass support
- Two genomes are compared by measuring the relative quantity of DNA at different loci

Array CGH allows a genome-wide blind search for $\sim$kbs aberrations
Tracking Genomic Aberrations in Cancer Genomes
Tracking Genomic Variation in Healthy Genomes

- In 2005 a study published the map of Copy Number Variations in healthy individuals
- Most initial studies of genetic variation concentrated on individual nucleotide sequences (SNPs)
- CNVs have become new genetic markers to study human diseases and evolution

from http://www.nature.com/
Outline

1 Introduction

2 Statistical Analysis of single profiles

3 Statistical analysis of multiple profiles

4 Adaptation to high dimensional computing
Nature of array CGH data

- The signal $Y_t$ is a log$_2$ ratio of fluorescence organized along the genome ($t$)
- When $Y_t \sim 0$ the region has no imbalance between test-reference
- When $Y_t > 0$ (resp. < 0) the test genome shows gains (resp. deletions)
- How many segments ? where ? status ?
Modeling & Computing Strategies

• Hidden Markov models [2, 7, 6]
  - Introduce a hidden Markovian sequence to model copy number
  - Recover the hidden sequence by Forward-Backward Algorithm

• Segmentation Models [10, 8]
  - Suppose that there exist abrupt changes in the signal
  - Detect jumps using a partitioning algorithm

Many comparative studies have shown the efficiency of segmentation methods on those data [12]

We focus on computational aspects of segmentations
Segmentation models: definitions and notations

- We observe a Gaussian process \( Y = \{ Y_1, \ldots, Y_n \} \) with
  \[
  Y_t \sim N(\mu_t, \sigma^2).
  \]

- We suppose that there exists \( K + 1 \) change-points \( t_0 < \ldots < t_K \) such that the mean of the signal is constant between two changes and different from a change to another.

- \( I_k = ]t_{k-1}, t_k] \): interval of stationarity, \( \mu_k \) the mean of the signal between two changes:
  \[
  \forall t \in I_k, \ Y_t = \mu_k + E_t, \ E_t \sim N(0, \sigma^2).
  \]
Calling Segments status by segmentation/clustering

- Segments can be in different states (Deleted, Normal, Amplified) which impacts the level of segments
- The idea is to introduce a hidden indicator variable $Z_{kp}$ such that

$$\forall t \in I_k, \text{if } Z_{kp} = 1, \quad Y_t = m_p + E_t, \quad E_t \sim \mathcal{N}(0, \sigma^2).$$

- Segment levels are shared across the genome ($m_{\text{deleted}}$ is the same for all deleted segments for instance).
Parameters and estimation strategy

- The parameters: $T = \{t_0, \ldots, t_K\}$, $\mu = \{\mu_1, \ldots, \mu_K\}$ and $\sigma^2$.
- The estimation is done for a given $K$ which is estimated afterwards.
- The log-likelihood of the model is:

$$
\log \mathcal{L}_K(Y; T, \mu, \sigma^2) = \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} f(y_t; \mu_k, \sigma^2).
$$

- When $K$ and $T$ are known, how to estimate $\mu$ ?
- When $K$ is known, how to estimate $T$ ?
- How to choose $K$ ?
### Parameter estimation

- When $K$ and $T$ are known the estimation of $\mu$ is straightforward:

\[
\hat{\mu}_k = \frac{1}{\hat{t}_k - \hat{t}_{k-1}} \sum_{t=\hat{t}_{k-1}+1}^{\hat{t}_k} y_t,
\]

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{K} \sum_{t=\hat{t}_{k-1}+1}^{\hat{t}_k} (y_t - \hat{\mu}_k)^2.
\]

- Find $\hat{T}$ such that:

\[
\hat{T} = \arg \max_T \{ \log \mathcal{L}_K(Y; T, \mu, \sigma^2) \}.
\]
Dynamic Programming to optimize the log-likelihood

- Partition $n$ data points into $K$ segments: complexity $\mathcal{O}(n^K)$.
- DP reduces the complexity to $\mathcal{O}(n^2)$ when $K$ is fixed.
- Shortest path problem: "subpaths of optimal paths are themselves optimal".
- $\text{RSS}_k(i, j)$ cost of the path connecting $i$ to $j$ in $k$ segments:

\[
\forall 0 \leq i < j \leq n, \quad \text{RSS}_1(i, j) = \sum_{t=i+1}^{j} (y_t - \bar{y}_{ij})^2,
\]
\[
\forall 1 \leq k \leq K - 1, \quad \text{RSS}_{k+1}(1, j) = \min_{1 \leq h \leq j} \{ \text{RSS}_k(1, h) + \text{RSS}_1(h + 1, j) \}.
\]
Model selection for segmentation

• The number of segments $K$ should be estimated:

$$\hat{K} = \arg \max_K \left\{ \log L_K(Y; \hat{T}, \hat{\mu}, \hat{\sigma}^2) - \beta \text{pen}(K) \right\}.$$ 

• Difficulty: $C_{n-1}^{K-1}$ possible partitions for a model with $K$ segments.

• Non-asymptotic theory provides a general form for $\beta \text{pen}(K)$ [5]:

$$\beta \text{pen}(K) = \frac{K}{n} \sigma^2 \times \left( c_1 + c_2 \log \frac{n}{K} \right).$$

• Other methods are based on an adaptive estimation of $K$ [4, 10] or on a modification of the BIC [13].
Outline

1. Introduction

2. Statistical Analysis of single profiles

3. Statistical analysis of multiple profiles

4. Adaptation to high dimensional computing
Using Multiple Arrays to assess CNA/CNV

- Population-based analysis for cancer and human genetics
- Multiple Arrays Analysis
  - Find breaks using all samples
  - Find *reccurrent* breaks
- What is specific/common?
  - Shared biases
  - Specific CN

Use multiple samples to increase the power of detection
Modelling individual-specific breakpoints \[^{[10]}\]

- \(Y_i(t)\): the signal for individual \(i = 1, \ldots, I\) with segments \(\{I^i_k\}\)
  \[
  \forall t \in I^i_k, \ Y_i(t) = \mu_{ik} + \varepsilon_i(t), \ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2).
  \]

- \(\mu_i\): specific levels of segments
- \(T_i\): specific incidence matrix of the breaks

\[
Y_i = T_i \mu_i + E_i
\]

- Signal levels associated to CN status are shared across arrays:
  \[
  \{Z^i_{kp} = 1\}, \ \forall t \in I^i_k, \ Y_i(t) = m_p + \varepsilon_i(t), \ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2).
  \]

\[
Y_i = T_i Z_i m + E_i
\]
Segmentation of Multiple Arrays [9]

- The RSS is additive wrt the series and to the number of segments.

\[ \text{RSS}_K(\mu, T) = \|Y - T\mu\|^2 = \sum_{i=1}^{l} \sum_{k=1}^{k_i} \text{RSS}_k^i(\mu_i, T_i) \]

- Global DP would lead to a \( O(n^2l^2) \) complexity.
- But there is a constraint: \( \sum_i k_i = K \), (K unknown) thus:

\[ \min_{\{T,\mu\}} \text{RSS}_K(T, \mu) = \min_{k_1 + \ldots + k_l = K} \left\{ \sum_{i=1}^{l} \min_{T_i, \mu_i} \text{RSS}_k^i(T_i, \mu_i) \right\} . \]
A two-stage Dynamic Programming procedure - 1

- Find all optimal breaks for each profile using a “classical DP”
- $\hat{T}_i(k_i)$ the set of optimal breaks with $k_i$ segments for profile $i$.
- Find $\hat{T}_i(k_i), \forall k_i = 1, \ldots, k_{\text{max}}$ segments by minimizing $RSS^i_{K_i}(T_i, \mu_i)$ for each series.

$$\forall i \in [1, l] \{\hat{T}_i, \hat{\mu}_i\} = \arg\min_{T_i, \mu_i} \{RSS^i_{K_i}(T_i, \mu_i)\}$$
A two-stage Dynamic Programming procedure - 2

- Optimal allocation of segments to series

\[
\forall i \in [1 : l],
\{\hat{k}_1, \ldots, \hat{k}_i\} = \arg\min_{k_1 + \ldots + k_i = K} \text{RSS}_K \left( \hat{T}_1^1(k_1), \ldots, \hat{T}_i^i(k_i) \right)
\]

\[
\hat{T}(K) = \left\{ \hat{T}_1^1(\hat{k}_1), \ldots, \hat{T}_l^l(\hat{k}_l) \right\}.
\]

- This procedure is optimal with a complexity \(\mathcal{O}(ln^2 k_{max} + k_{max}^2 l^3)\).
Enrich the model to account for common genomic biases?

- There exist common genomic biases that are shared by all profiles. How to correct them?
- The simplest way to model this trend is to introduce a common background function $b(t)$ such that:

$$
\forall t \in [t_{k-1}^i, t_k^i], \quad Y_i(t) = \mu_{ik} + b(t) + E_i(t).
$$

- This produces a new model that mixes piece-wise constant functions and other undetermined functions.
Regularization of the trend using splines

- Control the second derivative of $b$ using a penalty:

$$
\min_{T,\mu,\theta} \left\{ \frac{1}{2} \| Y - T\mu - Xb \|_2^2 + \lambda \int [b''(t)]^2 dt \right\}. 
$$

- $\{W\}_{jk} = W_j(t_k)$ a $n$-dim. set of natural spline functions: $b = W\theta$

$$
\min_{T,\mu,\theta} \left\{ \frac{1}{2} \| Y - T\mu - XW\theta \|_2^2 + \lambda \theta^T \Omega \theta \right\},
$$

- The solution is given by:

$$
\hat{\theta} = \left( W^T W + \lambda \Omega \right)^{-1} W^T \left( X^T \left[ Y - T\mu^{[h]} \right] / I \right).
$$
Outline

1. Introduction

2. Statistical Analysis of single profiles

3. Statistical analysis of multiple profiles

4. Adaptation to high dimensional computing
Pruning Strategy

• Pruning Strategies reduce the computational burden of Dynamic Programming ([11, 3])
• The idea is to prune the set of candidates while computing potential segmentations
• The complexity is reduced from $O(Kn^2)$ to $O(n)$ or $O(n \log(n))$
• This linearization allows segmentation to be used on very long signals (microarrays, sequencing)
Parallelization in cghseg

- cghseg is a R-package dedicated to segmentation
- Most computers have multi-core architectures (from laptops to many-core servers)
- It has become essential to adapt software to computer architectures

Use straightforward parallelization to perform segmentation on large-numerous signals
Computing scheme

**Individual Segmentations**
- For a given level of segments
- Use DP on each profile
- Propose potential segmentations

**Global Segmentation**
- Estimate the level of segments
- Use DP to find the best global segmentation
- Perform model selection

Pruning for each segmentation

High number of profiles

Can be parallelized !!!
Parallelization in cghseg

- Compare the observed speedup to theoretical speedup (Amdahl’s law [1])
- The speedup of cghseg follows the Amdahl’s law when the number of profiles is high
- The gain decreases with the number of profiles due to overheads associated with the used of the parallel R-package

\[
\text{total time} \approx \text{time(sequential)} + \frac{\text{time(parallel)}}{\text{nb of cores}}
\]
Next Gen. Computing/Next Gen. experiments

- Considering multiple Arrays allows the joint assessment of Copy Number Aberrations/Variations at the cohort level
- We solve the computational issue of joint segmentation using a X2 Stage DP with linearization
- The method is implemented in the cghseg package

<table>
<thead>
<tr>
<th>$n$ (observations/profile)</th>
<th>$l$ (number of profiles)</th>
<th>20,000</th>
<th>100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>256</td>
<td>512</td>
<td>1024</td>
</tr>
<tr>
<td>Average CPU time (min)</td>
<td>6</td>
<td>15</td>
<td>54</td>
</tr>
<tr>
<td>Memory usage (Gb)</td>
<td>0.4</td>
<td>0.8</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>512</td>
<td>1024</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>70</td>
<td>253</td>
</tr>
<tr>
<td></td>
<td>1.7</td>
<td>3.7</td>
<td>7.9</td>
</tr>
</tbody>
</table>
Conclusions

• The analysis of copy number profiles has been very challenging from a statistical and computational point of view
• When providing methods, check the scalability (500K probes)
• Developments had to be done using mathematical+computing skills
• Many question arise from these data, in particular the impact of the inter-individual variability
• Project: shift towards functional mixed models for genomics
References


