Statistical Classification for Reliable High-volume Genetic Measurements

Paul Eilers
p.h.c.eilers@fss.uu.nl

Faculty of Social and Behavioural Sciences
Utrecht University
and
Faculty of Social and Behavioural Sciences
Leiden University
Modern molecular biology

- Molecular biology has gone through a revolution
- It has become possible to do massive measurements:
  - DNA composition (sequencing, SNP arrays)
  - gene expression (microarrays)
  - protein fragments (mass spectrometry)
- High-volume: $10^3$ to $10^6$ simultaneous measurements
- At a reasonable price (per measurement)
- This has all happened over one decade
The thrill has gone ...

- The new technology generated high expectations
- Medicine’s Holy Grail: reliable diagnosis and prediction
- Was it just around the corner?
- Microarrays were the big game
- With mass spectrometry in its slipstream
- Papers in high-profile journals (Nature, Science, Lancet)
- With high-profile claims
- Which did not exactly come out ...
- It is more quiet now; expectations have been lowered
Low-profile applications

- There still is a lot of room for statistics
- But at a lower level, in an earlier stage
- The technology still has many flaws
- Statistical models can be a real help
- Interesting complexities create non-standard situations
- This leads to methodological challenges
- Not so high-profile, but very rewarding
- Today’s example: SNPs (pronounced as “snips”)

What are SNPs?

- You know about DNA, the “double helix”
- It consists of long chains of nucleotides (“bases”)
- Symbols: A, C, G and T
- (adenine, cytosine, guanine, thymine)
- In normal DNA many mutations occur
- And even more in tumors
- Smallest mutation: single nucleotide polymorphism (SNP)
- Mutations in single bases
- At least a million SNPs on our genome
Why are SNPs interesting?

- They can be used as molecular markers
- Useful when searching genes that are related to diseases
- Or related to other characteristics (phenotypes)
- SNPs can also characterize changes in tumors
- This might be useful in early detection and prognosis
- Background: LUMC project on colon tumors
- (LUMC: Leiden University Medical Centre)
- Less aggressive tumors: transanal excision (no fun)
- Aggressive tumors: remove part of colon (even less fun)
- Can we detect the difference early?
Chromosomes, alleles and genotypes

- Each SNP can have two variants: alleles
- Each SNP occurs two times: chromosomes come in pairs
- Indicate the alleles by A and B
- Then we can have pairs AA, AB, BA and BB
- These are called diplotypes
- We cannot tell AB from BA
- So there are three possible genotypes: AA, AB and BB
How do we measure SNPs?

- We use Illumina technology
- A bundle of approximately 50000 glass fibers
- At the end of each fiber a tiny bead
- Each bead is covered with specific molecules
- Molecules bind to one specific SNP
- They can recognize both alleles
- Fluorescence is used for detection
- Molecules emit light when illuminated by a laser
- Green light for allele A, red for allele B
- Intensities proportional to concentrations
An Illumina array, one color
An Illumina array, detail
Raw and summarized data

- Each SNP is replicated some 30 times
- Over 1500 SNPs per array
- Arrays come in four types (panels)
- Four panels cover all chromosomes
- We will look at averages per SNP, on one panel
- Two vectors: 1500 pairs of green and red intensity
- Our goals: estimate genotypes, quantify uncertainties
Scatterplot of red and green signals

Sample 514NP

Red

Green
Logarithmic scales

Sample 514NP
Rotating the axes

Sample 514NP

log10(Green / Red)

log10(Red + Green)
Background (= minimum - 10) subtracted

Sample 514NP

log₁₀(Green / Red)

log₁₀(Red + Green)
Clustering

- Three clusters, genotypes GG = AA, GR = AB, RR = BB
- Model for each cluster: noise around straight line
- Separate slope, intercept, standard deviation per cluster
- Fit model by EM algorithm: repeated “split and fit”
  - E-step: estimate missing data (cluster memberships)
  - M-step: maximize likelihood given membership probabilities
- Assume an approximate solution to be known
  - Split: assign membership probabilities to observations
  - Fit: use posterior probabilities as weights for improved fit
- Repeat split/fit until convergence
Technical details: E-step

- Line cluster $j$: $\mu_j(x) = a_j + b_j x$, $j = 1 \ldots 3$
- Normal distribution around line: $N(\mu_j(x), \sigma_j^2)$
- Probability density for cluster $j$:

$$p_j(y) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp \left( -\frac{(y - \mu_j(x))^2}{2\sigma_j^2} \right)$$

- For each $y_i$ we get three density values: $p_{i1} = p_1(y_i)$, $p_{i2} = p_2(y_i)$ and $p_{i3} = p_3(y_i)$
- Posterior probabilities: $w_{ij} = p_{ij} / \sum_k p_{ik}$
- We have estimated cluster memberships
- (Actually more complicated: cluster probabilities $\pi_j$)
Technical details: M-step

- We have estimated cluster memberships $w_{ij}$
- We use them as weights to estimate cluster lines
- Weighted fitting: minimize

$$SS_j = \sum_i w_{ij} (y_i - a_j - b_j x_i)^2$$

- Regression for each cluster separately
- Residual variance gives $\hat{\sigma}_j^2 = \sum_i w_{ij} (y_i - \mu_j(x_i))^2 / \sum_i w_{ij}$
- Result: improved cluster parameters
- We repeat process with new E-step
- And so on, until changes are (very) small
Starting values

- Good starting values are needed
- That is simple here
- Horizontal cluster lines
- At heights -1.5, 0, 1.5
- Standard deviations all equal to 0.3
- Convergence takes 10 to 30 iterations
- For decent (i.e. most) arrays
Result of fit (dots colored by posterior probabilities)

Sample 514NP

log10(R + G)

log10(R/G)
Some more examples, good samples
Some more examples, bad samples

- **511TV**
- **106NB**
- **106TV**
- **106TP**
Quality control

- Dull nickname: RS plot ($R = \text{ratio}, \ S = \text{sum}$)
- We also call it the shaslik plot
- Whatever its name, it is informative
- Cluster separation easily assessed visually
- Feel for numbers (cluster positions, widths) should grow
- For each SNP we get three posterior probabilities
- Highest probability determines genotype
- Distribution of highest probabilities informative
- Show cumulative distribution
Looking at the probabilities

- Membership probabilities add up to one
- They are used to color the dots (RGB system)
- GG cluster: green, GR cluster: blue, RR cluster: red
- Color purity gives good impression
- Genotype decision: select cluster with largest probability
- Good classification when largest probability close to 1
- Distribution of largest probabilities shows performance
Cumulative distribution of membership probabilities

Sample 514NP
Why a custom EM algorithm?

- Clustering software is available for R
- So-called model based clustering
- Data distribution as sum of 2-D normal distributions
- Raftery’s library mclust; why not use that?
- Experience shows that it is quite slow
- And that it tries to do too much
- Standard deviations along clusters (bivariate normal)
- We don’t need them
- Custom algorithm is just 15 lines of R code and fast
Working along chromosomes

- Our data cover a number of chromosomes
- Individual chromosomes are interesting
- Vertical axis: $\log_{10}(\text{Red}/\text{Green})$ again
- Horizontal axis: position on chromosome
- Position is measured in megabases (Mb)
- We now fit horizontal lines (zero slopes)
- We compare tumor and matching normal sample
Red/Green ratio (log) along chromosome

Normal sample, chromosome 8

Tumor sample, chromosome 8
Red/Green ratio (log) and model

Normal sample, chromosome 8

Tumor sample, chromosome 8
Loss of heterozygosity

- In tumors, parts of chromosomes may get lost
- The cell may produce a copy of the remaining chromosome
- Then we can only encounter genotypes AA and BB
- Loss of heterozygosity (LOH): no AB genotype
- Chromosome fragments with LOH are interesting
- Especially those that show frequent LOH
- They may indicate regions with special genes
- Tumor-promoter or tumor-suppressor genes
- We hope they will be useful for early warning
A more dramatic LOH example

**Normal sample, chromosome 7**

**Tumor sample, chromosome 7**
Detecting LOH without normal samples

- LOH detection is easy with tumor plus normal sample
- You simply look where AB genotype disappears
- But normal sample is not always available
- Certainly not for many tumor “archives”
- Removed tumors, deep frozen, or in paraffin
- Long homozygote regions make LOH more probable
- Fit a smooth curve for the probability of LOH
The smooth LOH model

- We observe 1 or 0 (homozygote or not)
- At all SNPs, with positions $x$
- Smooth logistic regression:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = f(x_i)$$

- Probability $\pi$ of homozygote
- Smooth curve $f(x)$, modeled by splines
- My favorite: P-splines
- A sum of 100 scaled B-splines to build $f(x)$
- Difference penalty on coefficients to tune smoothness
Logistic smoothing I

Tumor sample, chromosome 8

Position on chromosome (Mb)

log10(Red/Grn)

Homozygote probability (L2 penalty)

Position on chromosome (Mb)
Logistic smoothing II

Tumor sample, chromosome 7

LOH probability

Position on chromosome (Mb)

Homozygote

Position on chromosome (Mb)

log10(Grn/Red)
Optimal smoothing

- Model for log-odds:
  \[ \log\left(\frac{\pi_i}{1 - \pi_i}\right) = f(x_i) = \sum_j B_j(x_i) \alpha_j \]

- Maximize penalized likelihood
  \[ L^* = L - \lambda \sum_j (\alpha_j - \alpha_{j-1})^2 / 2 \]

- Binomial likelihood:
  \[ L = \sum [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)] \]

- Use AIC to optimize λ: \[ AIC = -2L + 2 \times ED \]

- \( ED \) is effective dimension

- It is the trace of the (linearized) smoother matrix

- Vary λ on grid, search for minimum of AIC
A pleasing AIC profile
An other penalty

- Penalty $\lambda \sum_j (\Delta \alpha_j)^2$ gives smooth results
- Biologists like to see sharper jumps
- Between more or less flat segments
- This represents the actual process better
- We can improve results with penalty $\lambda \sum_j |\Delta \alpha_j|$
- This can be implemented with adaptive weights
- Penalty $\lambda \sum_j v_j (\Delta \alpha_j)^2$ with $v_j = 1/|\Delta \alpha_j|$ 
- Little stability trick: $v_j = 1/\sqrt{a^2 + (\Delta \alpha_j)^2}$
- With small number (0.001) for $a$
Logistic smoothing with L1 penalty

Tumor sample, chromosome 8

Homozygote probability (L1 penalty)
AIC for L1 penalty

Optimal logistic smoothing (L1 penalty)
The model is too simple

- We get a curve of apparent LOH probabilities
- Useful, but not the full story
- Even intact DNA has chance $\pi_i$ of homozygous SNP $i$
- Let latent probability of change be $q_i$
- With latent change, probability of LOH is 100%
- Then $p_i = q_i + (1 - q_i)\pi_i$
- Model logit of $q$ with P-splines
- Work in progress, no results yet
Future challenges

- Tumor samples are seldom “pure”
- They are mixed with normal tissue
- You might get mixture of LOH and not-LOH
- This will blur the R-S plot
- “Copy number” changes also occur
- Locally only one chromosome part is found
- Or more than two chromosome parts
- Genotypes, A, B, AAA, AAB, ABB, BBB, etc.
More future challenges

- Present data sets are relatively small: 1500 SNPs
- But larger arrays and other technologies exist
- Really high volume: $10^5$ SNPs and more
- Example on somewhat larger scale: Affymetrix ($10^4$ SNPs)
- Point clouds look curved
Affymetrix data \(10^4\) SNPs
Scatterplot visualized as a density

Array 137_AA_T0N0

log10(R + G)

log10(R/G)
Modelling opportunities

- Clusters with quadratic axes?
- Or semi-parametric models (P-splines again)?
- Or approach it as density estimation problem?
- First reduce data to counts on 2-D grid
- Fit sum of three semi-parametric densities
- Relative probabilities follow from ratios of densities
- Original data points get probabilities
- By referring to their 2-D bins
The EM algorithm with densities

- Assume that we have a good 2-D density smoother
- One that favors log-concave shapes
- E-step: split counts per bin, proportionally to densities
- M-step: smooth three sets of split counts separately
- Repeat until convergence
- Starting values not difficult to find
- Smooth upper, middle and lower third of 2-D histogram
Density smoothing in one dimension

- Observed counts \( y_i, \ i = 1 \ldots m \)
- Expected values \( \mu_i = \exp(\eta_i), \ i = 1 \ldots m \)
- Poisson likelihood \( \sum_i (y_i \eta_i - \mu_i) \)
- B-splines to model \( \eta: \eta = B \alpha \)
- Third order difference penalty: \( \sum_i (\Delta^3 \alpha_i)^2 \)
- The penalty gently pushes \( \eta \) towards quadratic shape
- Result: \( \mu \) gently pushed to log-concave shape
Density smoothing in two dimensions

- Observed counts $y_{ij}$, $i = 1 \ldots m, j = 1 \ldots n$
- Expected values $\mu_{ij} = \exp(\eta_{ij})$
- Poisson likelihood $\sum_i (y_{ij}\eta_{ij} - \mu_{ij})$
- Tensor products B-splines to model $\eta$
- Third order difference penalty in two directions
- This fits in a large research project
- Collaboration with Currie, Durbàn, Lambert and Marx
- Penalized likelihood, mixed models, Bayes
- Fast array algorithms make it efficient
- I use the penalized likelihood approach
Three semi-parametric density estimates

Array 137_AA_T0N0

GG density

RG density

RR density
A bright view of the mixture

Array 137_AA_T0N0

log10(R/G)

log10(R + G)
Acknowledgements

- Work started at Leiden University Medical Centre
- In cooperation with:
  - Ronald van Eijk
  - Esther Lips
  - Anneke Middeldorp
  - Hans Morreau
  - Jan Oosting
  - Marjo van Puijenbroek
  - Tom van Wezel