National Spanish Cancer Center

RJaCGH, a package for analysis of CGH arrays with Reversible Jump MCMC

Oscar Rueda, omrueda@cnio.es

Ramon Diaz-Uriarte, rdiaz@ligarto.org

1.- CGH Arrays:

National Spanish Cancer Center

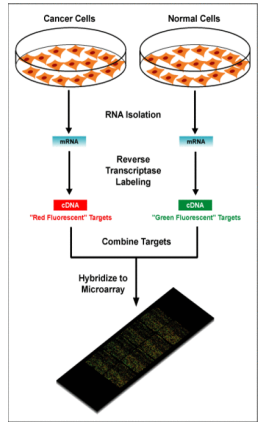
Biological problem:

Changes in number of DNA copies are associated to cancer activity.

Microarray technology:

Comparative Genomic Hybridization (CGH)

- Test DNA sample (Cancer) labeled in red
- Reference DNA sample (Control) labeled in green
- Samples are hybridized, and superimposed.
- The intensity of color is measured in log scale
- y=log (intensity of Test / intensity of reference)



http://en.wikipedia.org/wiki/Image:Microarray-schema.gif

2.- Methods for the Analysis of CGH Arrays:

National Spanish Cancer Center

Hypothesis testing based:

- Circular Binary Segmentation (Olshen et al., 2004)
- CGH-Explorer (Lingjaerde et al., 2004)
- aCGH-Smooth (Jong et al., 2004)
- SW-Array (Price et al., 2005)
- CLAC (Wang et al., 2005).
- Wavelets (Hsu et al., 2005)

2.- Methods for the Analysis of CGH Arrays (II):

National Spanish Cancer Center

Copy number estimation based:

- Hidden Markov Models (Fridlyand et al., 2003, Guha et al, 2005, Marioni et al., 2006)
- Quantile smoothing (Eilers et al., 2004)
- GLAD (Hupé et al., 2004).
- Picard et al., (2005)
- CGHMIX (Bröet and Richardson, 2006)
- Bayes Regression (Wen et al., 2006)

3.- Drawbacks of the current methods for the Analysis of CGH Arrays:

National Spanish Cancer Center

- Most of them don't have biological background.
- Some of them don't have an statistical model behind.
- Some of them have it, but make a post-processing step that invalidates the statistics.
- Most of them don't take into account distance between genes.
- Most of them have a lot of parameters to tune, with no intuitive interpretation.

4.- RJaCGH. Motivation:

National Spanish Cancer Center

There are a finite number of different copy gains / losses.

Finite Mixture Model.

We don't measure directly that number, but instead we have a gaussian noise.

Finite Mixture Model with Gaussian Distributions.

The state of every gen influences the state of its neighbours,

Hidden Markov Model with Gaussian Distributions.

This influence must be bigger the closer the genes are.

Non Homogeneous Hidden Markov Model with Gaussian Distributions.

The model uncertainty must be taken into account

NH HMM with Gaussian Distributions. with Bayesian Model Averaging: RJaCGH.

5.- RJaCGH. Main features:

National Spanish Cancer Center

- Non Homogeneous Hidden Markov Model with unknown number of states.
- Bayesian Inference through Markov Chain Monte Carlo Simulation
- Automatic selection of the number of states through Reversible Jump MCMC.
- Classification of states takes into account model uncertainty:
  - AIC or BIC are not good methods for choosing the number of hidden states.
  - Not a “purist” bayesian analysis: hidden state sequence is obtained via a point estimator of means, variances and transition matrix.
  - Bayesian Model Averaging:

$$P(S_i=r/X_i=x)=\sum P(K=i)P(S_i=r/X_i=x,K=i)$$

6.- RJaCGH. The statistical model:

National Spanish Cancer Center

$k=$ number of different copy numbers $s_t=$ true copy number of the gene t

$y_t=\log_2$ ratio of the gene t

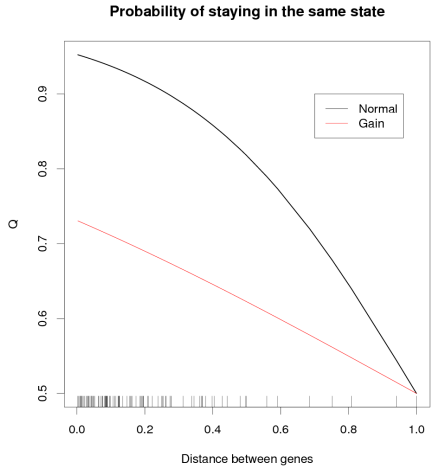
$x_t=$ distance between genes t and its predecessor

$y_t/s_t \sim N(\mu_k,\sigma_k^2)$

$p(s_t=j/s_{t-1}=i,x_t=x)=Q_{i,j,x}$

$$Q_{i,j,x}=\frac{\exp(-\beta_1+\beta_1 x)}{\sum_{p=1}^{k}\exp(-\beta_p+\beta_p x)}$$

$$\beta=\begin{pmatrix}0 & \beta_1 & \dots & \beta_{k-1}\\ \beta_k & 0 & \dots & \beta_{2k-2}\\ \dots & \dots & \dots & \dots\\ \beta_{(k-1)(k-1)-(k-1)} & \beta_{(k-1)(k-1)-k} & \dots & 0\end{pmatrix}, \quad \beta\geq 0$$



7.- RJaCGH. The bayesian model:

National Spanish Cancer Center

$p(k) \equiv$ Priori over number of hidden states
By default , is a uniform distribution

$p(\theta(k)/k) \equiv$ Priori over HMM conditioned on k
$\mu \sim N(\alpha, \beta)$
By default , $\alpha = median(y)$, $\beta = range(y)$
$\sigma^2 \sim IG(ka, g)$
By default , $ka = 2, g = range^2(y)/50$
$Beta \sim \Gamma(1,1)$

$L(y; k, \theta^k) \equiv$ Likelihood of the model

$p(k)\, p(\theta(k)/k) L(y; k, \theta^k) \equiv$ Joint distribution

8.- RJaCGH. The MCMC simulation:

National Spanish Cancer Center

Each sweep consists of three steps:

1.- Update model:

- In turn, Metropolis-Hastings step for means, variances and transition matrix.
- The hidden state sequence is not part of the of the state space of the sampler
- The dimensionality of that space is reduced.

2.-Update number of hidden states: attempt birth / death move:

3.-Update number of hidden states: attempt split / combine move:

9.- RJaCGH. The RJ moves:

National Spanish Cancer Center

Birth move:

A new state is sampled from the priors and accepted with probability prob.birth=min(1, p)

$$p = \frac{P(k=r+1)\, L(y; r+1, \theta(r+1))\, P_{death}(r+1)}{P(k=r)\, L(y; r, \theta(r))\, P_{birth}(r)}$$

Split move:

A state is split into two ones and accepted with probability prob.split=min(1, p)

$$\mu_{i1} = \mu_{i0} - \sigma_{i0}\epsilon_\mu, \quad \mu_{i2} = \mu_{i0} + \sigma_{i0}\epsilon_\mu \quad with\ \epsilon_\mu \sim N(0, \tau_\mu)$$
$$\sigma^2_{i1} = \sigma^2_{i0}\epsilon_\sigma, \quad \sigma^2_{i2} = \sigma^2_{i0}(1-\epsilon_\sigma) \quad with\ \epsilon_\sigma \sim Beta(2,2)$$
Split column i_0 $\quad \beta_{i,i1} = \beta_{i,i0}\epsilon_\beta, \quad \beta_{i,i2} = \beta_{i,i0}/\epsilon_\beta \quad with\ \epsilon_\beta \sim \ln(0, \tau_\beta) \quad for\ i \neq i_0$
Split row i_0 $\quad \beta_{i1,j} = \beta_{i0,j} U_j, \quad \beta_{i2,j} = \beta_{i0,j}(1-U_j) \quad with\ U_j \sim Beta(2,2) \quad for\ j \neq i_0$
$\beta_{i1,i2} \sim \Gamma(1,1)$

$$p = \frac{P(k=r+1)\, P(\theta(r+1))\, L(y; \theta(r+1))\, (r+1)}{P(k=r)\, P(\theta(r))\, L(y; \theta(r))\, 2\mathrm{p}(\epsilon_\mu)\, p(\epsilon_\sigma) \prod P(\epsilon_\beta) \prod P(U_j)} J_{split}$$

$$J_{split} = 2^r \sigma^3_{i0} \prod_{r-1} \beta_{i0,j} \prod_{r-1} \frac{\beta_{i,i0}}{\epsilon_{beta}}$$

Death and combine moves are the symmetric ones, and their acceptance probabilities are the inverse of the birth and split ones..

10.- RJaCGH. The package:

National Spanish Cancer Center

Main function:

```
RJaCGH(y, Chrom = NULL, Pos = NULL, model = "genome", burnin = 0,
 TOT =1000, k.max = 6, stat = NULL, mu.alfa = NULL, mu.beta = NULL, ka = NULL,
g = NULL, prob.k = NULL, jump.parameters=list(), start.k = NULL, RJ=TRUE)
```

The object returned can be of several classes:

- RJaCGH.array: if y was a matrix or data frame of arrays.
- RJaCGH.genome: if we fit the same model to the whole genome.
- RJaCGH.Chrom: if we fit a different model to each chromosome.
- RJaCGH: a fit to a sequence without chromosome index.

```
jp <- list(sigma.tau.mu=rep(0.05, 6), sigma.tau.sigma.2=rep(0.01, 6),
    sigma.tau.beta=rep(0.5, 6),tau.split.mu=0.5, tau.split.beta=0.5)

fit <- RjaCGH(y=gm01523$LogRatio, Pos=gm01523$PosBase,
    Chrom=gm01523$Chromosome,model="genome", burnin=50000,
    TOT=100000, jump.parameters=jp)
```

11.- RJaCGH. The package:

National Spanish Cancer Center

The objects returned are lists:
if fit has class 'RjaCGH' we can access its elements:
- fit$k : models visited
- fit[[1]] : model with 1 hidden state
  - fit[[1]]$mu : means visited by the sampler.
  - fit[[1]]$sigma.2 : variances visited by the sampler.
  - fit[[1]]$beta : betas visited by the sampler.
- fit[[r]] : model with r hidden states.

If fit has class 'RjaCGH.genome', it's the same as before.
If fit has class 'RjaCGH.Chrom' it's a list with sublists as before:
- fit[[1]] : model for the first chromosome
  - fit[[1]]$k, fit[[1]][[1]]$mu, etc.

if fit has class 'RjaCGH.array' it's again a list with sublists:
- fit[[1]] : first array
  - fit[[1]][[1]] : first array, first chromosome (if model=Chrom)

12.- RJaCGH. The package. Methods:

National Spanish Cancer Center

Summary:

summary(fit) -> summary of the fit. Point estimator (mean, median or mode) of means, variances and transition matrix.

States:

states(fit) -> sequence of hidden states. Not a part of the model, computed via a point estimator of the means, variances and transition matrix and the backward filtering probabilities. Not computed by viterbi.

Model averaging:

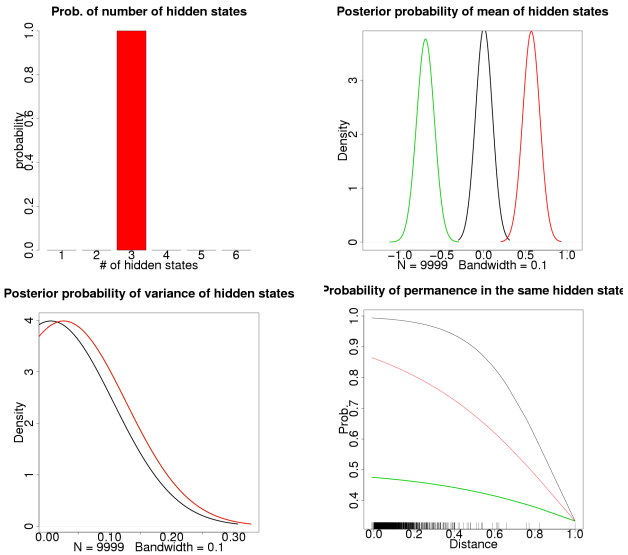
model.averaging(fit) -> sequence of hidden states computed via a call to states for every model fit, weighted by the posterior probability of that number of states.

Plot:

plot(fit) -> plot fitted model. Plot a single chromosome, the whole genome, bayesian model averaging of several arrays, region of common gains / losses of several arrays.

13.- RJaCGH. The package. Examples:

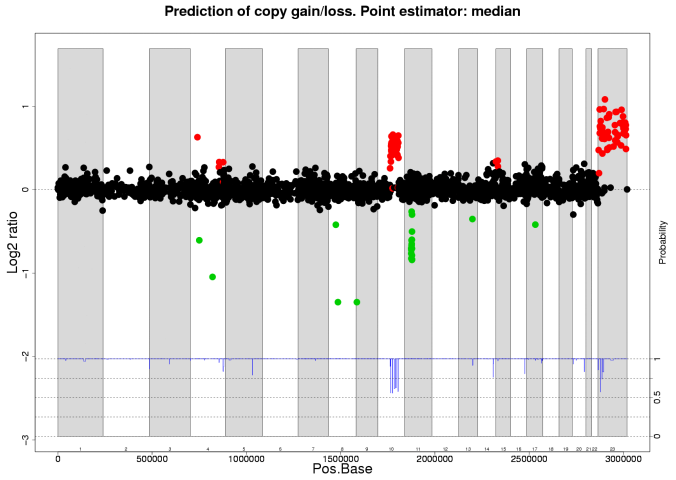
National Spanish Cancer Center



Data from cell line GM05296 from Snijders et al. (2001).

13.- RJaCGH. The package. Examples (II):

National Spanish Cancer Center



96.82% of correct classification, but only 14 transdimensional moves

14.- RJaCGH. The package. Checking convergence:

National Spanish Cancer Center

We recommend running several parallel chains and join them in a list.

Gelman & Brooks plot:

gelman.brooks.plot(list) -> plot of R values (Gelman and Brooks, 1998.
General Methods for Monitoring Convergence of iterative simulations.
Journal of Computational and Graphical Statistics.
The sequence of R values must converge to 1.

Collapse chain:

collapse.chain(list) -> if the chains have converged, they can be joined and make inference with all of them.

15.- RJaCGH. Implementation issues:

National Spanish Cancer Center

First attempt:

Developing the whole package in R language: Too slow for typical array sizes

Second try:

Writing parts of the algorithm in C language: gaining speed

Final version:

Whole sweep algorithm in C language.

Things to improve:

Speed up with the inclusion in Asterias suite: http://www.asterias.info
Improve split / combine moves to achieve better mixing rates...

16.- RJaCGH. Things to improve:

National Spanish Cancer Center

- Speed up with the inclusion in Asterias suite: http://www.asterias.info
  Parallelize algorithm for R: (Rmpi, papply).
  Parallelize algorithm for C: (MPI, UPC).
- Improve split / combine moves to achieve better mixing rates...

ACKNOWLEDGMENTS:
Funding provided by Fundación de Investigación Médica Mutua Madrileña and Project TIC2003 – 09331-C02 – 02 of the Spanish Ministry of Education and Science.

Oscar Rueda, omrueda@cnio.es Ramon Diaz-Uriarte, rdiaz@ligarto.org