

Analysis of aCGH data: statistical models and computational challenges

Ramón Díaz-Uriarte

2007-02-13

Outline

- 1 Analysis of aCGH data
 - Introduction
 - Alternative approaches
 - What we really want
 - RJaCGH
 - There is math (just so you believe us)
 - RJaCGH: typical output
 - RJaCGH: performance
- 2 Statistical methods need software
 - Introduction
 - ADaCGH
 - Can we make it fast?
 - MPI et al.

1 Analysis of aCGH data

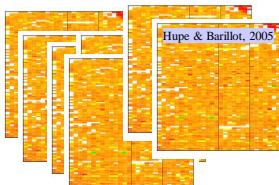
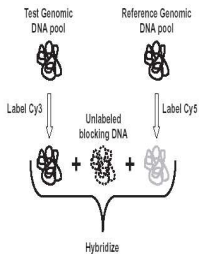
- Introduction
- Alternative approaches
- What we really want
- RJaCGH
- There is math (just so you believe us)
- RJaCGH: typical output
- RJaCGH: performance

2 Statistical methods need software

- Introduction
- ADaCGH
- Can we make it fast?
- MPI et al.

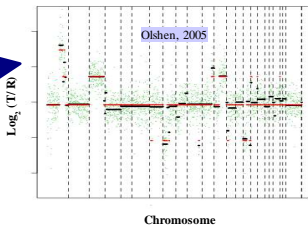
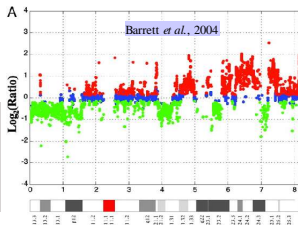
Analysis of aCGH data

I. Fridlyand et al. / Journal of Multivariate Analysis 90 (2004) 132–153



Calling gains and losses: hypothesis testing

Inferring number of copy gains/losses: estimation



Arrays: but location in chromosome matters

Available methods

Hypothesis testing based

- Circular Binary Segmentation
- CGHExplorer
- aCGHSmooth
- SWArray
- CLAC
- Wavelet-based smoothing

Copy number estimation

- Hidden markov models
- Quantile smoothing
- GLAD (adaptive nonparametric kernel smoothing)
- Gaussian mixtures
- Bayes regression
- CGHMIX

Is this status of affairs OK?

What is it we really want?

What is it we really want?

1. Probabilities of alteration.

What is it we really want?

1. Probabilities of alteration.

- **The** direct answer (to “is this gene/region gained/lost”? No, p-values and smoothed means are not a direct answer.)
- Usable in contexts from clinical to basic research: modify your thresholds as needed.
- Probabilities: of genes, of regions, of regions across subjects, etc. (No, not easy, but doable).
- Probabilities: incorporate uncertainty.

What is it we really want?

2. Account for distance between probes.

What is it we really want?

2. Account for distance between probes.

- Most platforms widely variable distance between probes.
- The larger the distance, the more likely a change.
- The larger the distance, the less information a probe provides about state of nearby probes.
- Use distance so that the information that consecutive probes provide is adequately accounted for.

What is it we really want?

3. Genome-wide and chromosome-wide analysis.

What is it we really want?

3. Genome-wide and chromosome-wide analysis.

- Chromosome-wide: alterations wrt chromosome's mean ploidy.
- Genome-wide: only way to detect whole chromosome gains/loses.

Do available methods give that to us?

No

Do available methods give that to us?

No

- Most don't provide probabilities.
- Most don't use distance between probes.
- Many ad-hoc, with parameters without intuitive meaning.
- No single methods fulfills the above three.

... thus

we will have to develop a new method

Model: characteristics

- Finite number of different copy gains / losses, number of gains/losses not measured directly, state of every gen related to the state of its neighbours: **Hidden Markov Model (HMM) with Gaussian emissions**.
- Influence larger the closer the genes are: **non-homogeneous HMM with Gaussian distributions**



Model: estimation

- Well known machinery for homogeneous HMM with known number of states

Model: estimation

- Well known machinery for homogeneous HMM with known number of states
- We have: non-homogeneous HMM with **unknown number of states** and want **probabilistic statements** about likely state and **flexible specification of model**: Bayesian inference through Markov Chain Monte Carlo

Model: estimation

- Well known machinery for homogeneous HMM with known number of states
- We have: non-homogeneous HMM with **unknown number of states** and want **probabilistic statements** about likely state and **flexible specification of model**: Bayesian inference through Markov Chain Monte Carlo
- Number of states not known in advance → number of parameters different for different models → need to “jump” between models during MCMC for automatic selection and probab. asses. of number of states: reversible jump MCMC

Model: estimation

- Well known machinery for homogeneous HMM with known number of states
- We have: non-homogeneous HMM with **unknown number of states** and want **probabilistic statements** about likely state and **flexible specification of model**: [Bayesian inference through Markov Chain Monte Carlo](#)
- Number of states not known in advance → number of parameters different for different models → need to “jump” between models during MCMC for automatic selection and probab. asses. of number of states: [reversible jump MCMC](#)
- Model uncertainty must be taken into account: [models averaged using Bayesian Model Averaging](#) (improved mean square error too!).

Statistical model

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

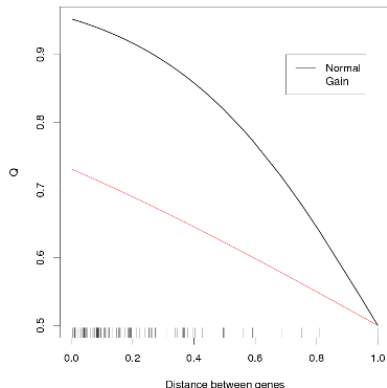
$y_t/s_t \sim N(\mu_k, \sigma_k^2)$

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

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

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

Probability of staying in the same state



Bayesian model

$p(k) \equiv$ 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, $\alpha = \text{median}(y)$, $\beta = \text{range}(y)$

$\sigma^2 \sim IG(ka, g)$

By default, $ka = 2$, $g = \text{range}^2(y)/50$

Beta $\sim \Gamma(1, 1)$

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

Reversible Jump

Birth move: A new state is sampled from the priors and accepted with probability

$p = \min(1, \text{prob.birth})$

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

Split move: A state is split into two ones and accepted with probability

$p = \min(1, \text{prob.split})$

$$\mu_{i1} = \mu_{i0} - \sigma_{i0} \epsilon_{\mu}, \quad \mu_{i2} = \mu_{i0} + \sigma_{i0} \epsilon_{\mu} \quad \text{with } \epsilon_{\mu} \sim N(0, \tau_{\mu})$$

$$\sigma_{i1}^2 = \sigma_{i0}^2 \epsilon_{\sigma}, \quad \sigma_{i2}^2 = \sigma_{i0}^2 (1 - \epsilon_{\sigma}) \quad \text{with } \epsilon_{\sigma} \sim \text{Beta}(2, 2)$$

$$\text{Split column } i_0 \quad \beta_{i1, j} = \beta_{i0, j} \epsilon_{\beta}, \quad \beta_{i2, j} = \beta_{i0, j} / \epsilon_{\beta} \quad \text{with } \epsilon_{\beta} \sim \ln(0, \tau_{\beta}) \text{ for } j \neq i_0$$

$$\text{Split row } i_0 \quad \beta_{i1, j} = \beta_{i0, j} U_j, \quad \beta_{i2, j} = \beta_{i0, j} (1 - U_j) \quad \text{with } U_j \sim \text{Beta}(2, 2) \text{ for } j \neq i_0$$

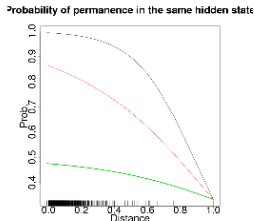
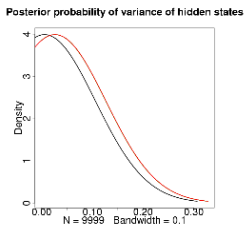
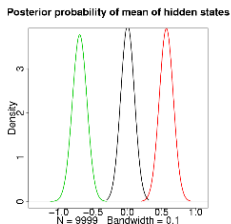
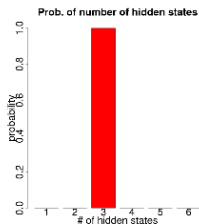
$$\beta_{i1, i2} \sim \Gamma(1, 1)$$

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

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

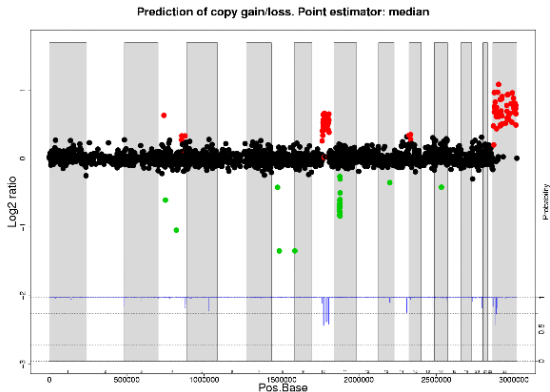
Death and combine moves are the symmetric ones, and their acceptance probabilities are the inverse of the birth and split ones..

RJaCGH. The package. Examples:



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

RJaCGH. The package. Examples (II):

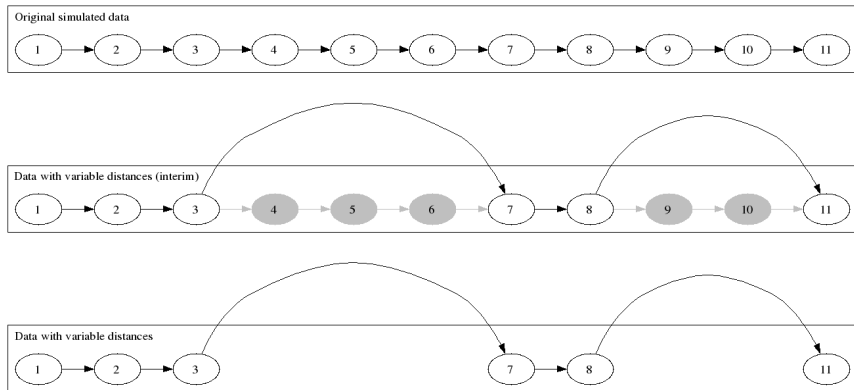


96.82% of correct classification, but only 14 transdimensional moves

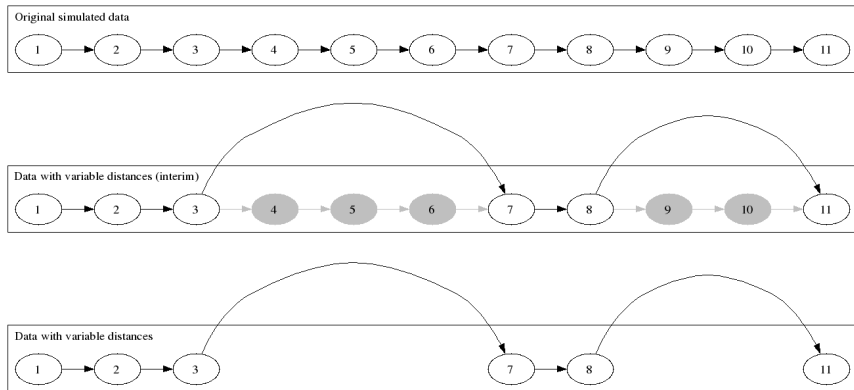
Does it really work better?

- Compare against best-performing methods (two reviews tell us which are “best”).
- Data NOT simulated under our model: Willenbrock and Fridlyand.

Adding variable distances to the data

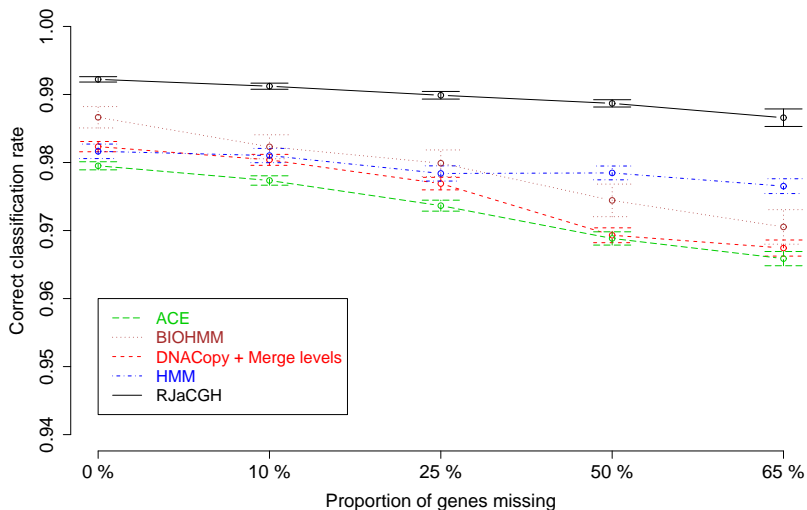


Adding variable distances to the data

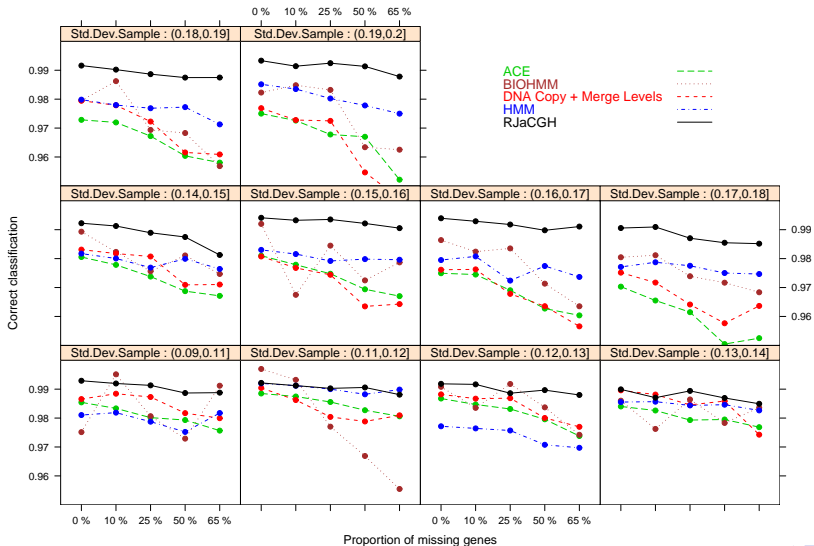


- Location of gaps: Uniform (100)
- Length of gaps: Poisson.
- Increasing lambda of Poisson (recall: mean = variance = λ) increases variance of inter-probe distance and % of gaps.

RJaCGH: and how does it do?



RJaCGH: and how does it do? (II)



aCGH analyses: what next?

- Minimal common regions (joint \neq product of marginals !!!)
- Computation: many MCMCs ...

1 Analysis of aCGH data

- Introduction
- Alternative approaches
- What we really want
- RJaCGH
- There is math (just so you believe us)
- RJaCGH: typical output
- RJaCGH: performance

2 Statistical methods need software

- Introduction
- ADaCGH
- Can we make it fast?
- MPI et al.

Statistical methods need software

“(...) a *reference implementation*, some code which is warranted to give the authors’ intended answers in a moderately-sized problem. It need not be efficient, but it should be available to anyone and everyone.”

Brian D. Ripley, RSS 2002 Plenary Lecture.

“(...) publishing figures or results without the complete software environment could be compared to a mathematician publishing an announcement of a mathematical theorem without giving the proof.”

Jonatahn B. Buckeit and David L. Donoho. Wavelab and Reproducible Research.

Need not be efficient?

Or who is our audience?

Need not be efficient?

Or who is our audience?

- Statisticians, bioinformaticians
- Wet-lab researchers

For wet-lab researchers

Implementing a user friendly tool for a
Bayesian-based approach not easy

For wet-lab researchers

Implementing a user friendly tool for a Bayesian-based approach not easy

- Will the user get tired of waiting (faster, faster)
- All the extra stuff (convergence, restarting chains, etc)

For wet-lab researchers

Implementing a user friendly tool for a Bayesian-based approach not easy

- Will the user get tired of waiting (faster, faster)
- All the extra stuff (convergence, restarting chains, etc)
- Working on it. In the meantime . . .

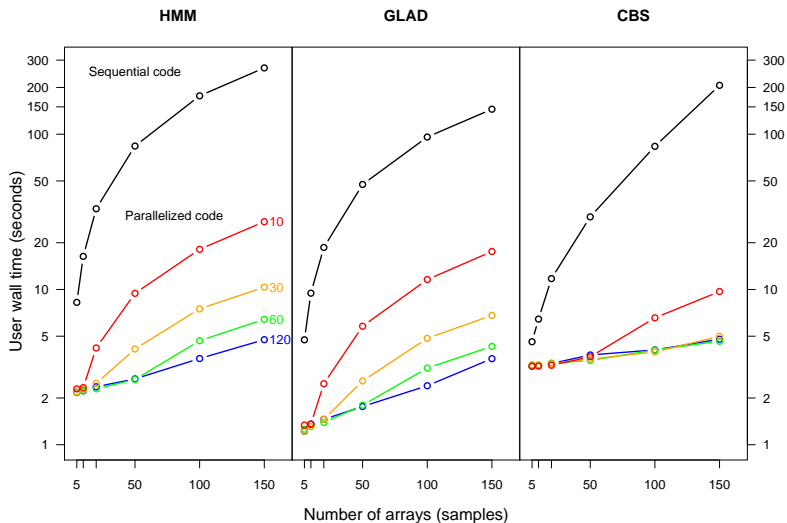
A simpler problem

ADaCGH: A web-based tool for the analysis of aCGH data.

- Common interface
- Implements all decently performing methods: CBS, HMM, BioHMM, GLAD, cghSeg, ACE, PSW, Wavelet-based smoothing.
- Dynamic graphics and links to additional functional information.
- And it is FAST.

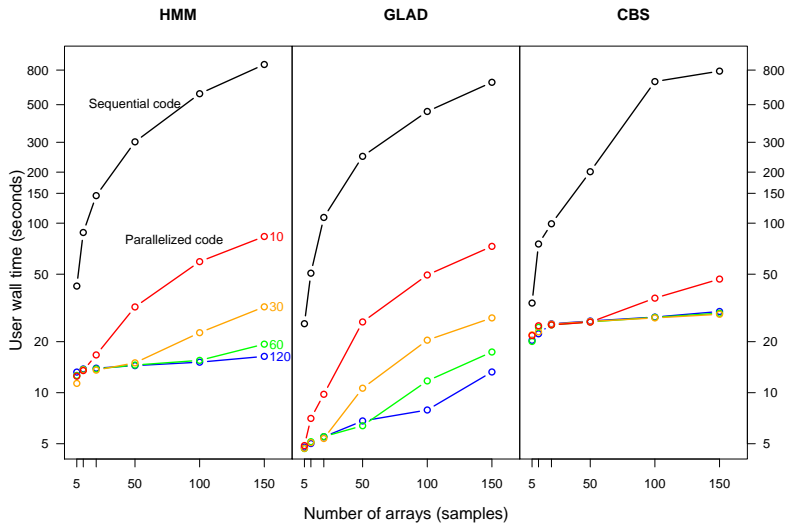
How fast is fast? (I)

Small data set (2271 genes)

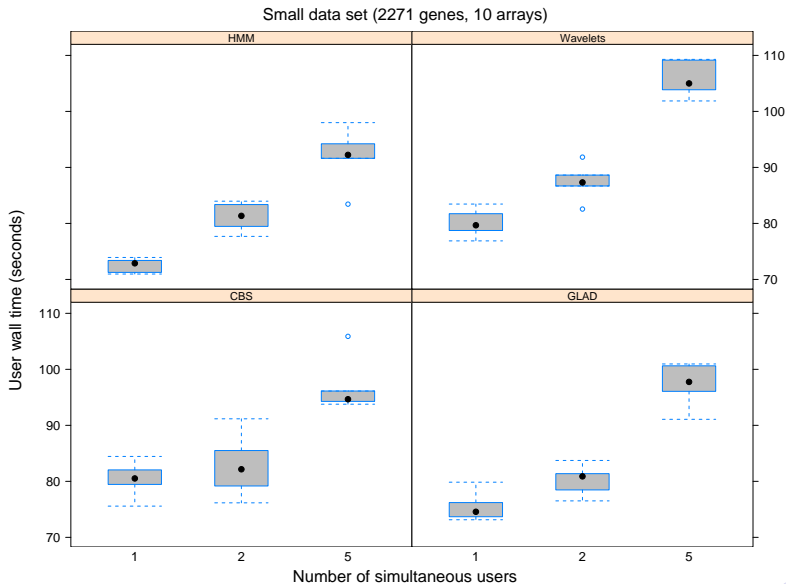


How fast is fast? (II)

Medium data set (10,000 genes)

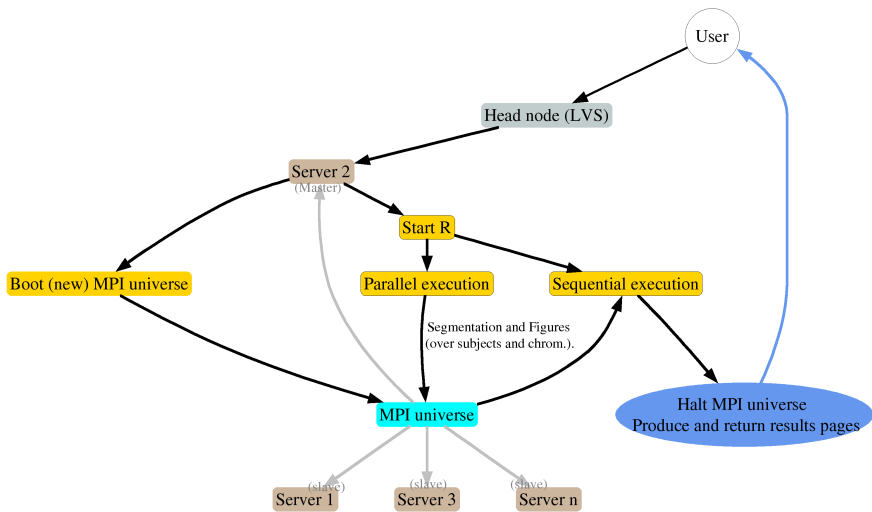


How many users?



Underneath

Combine **MPI** with **web-service load-balancing** plus systematic rotation of MPI master and slave nodes.



What can we parallelize?

What can we parallelize?

Segmentation (heavy number crunching) and figures . . .

What can we parallelize?

Segmentation (heavy number crunching) and figures . . .

over arrays, or chromosomes, or chromosomes by arrays.

What can we parallelize?

Segmentation (heavy number crunching) and figures . . .

over arrays, or chromosomes, or chromosomes by arrays.

. . . for everything else, there is load-balancing.

Is this the way to go? (I)

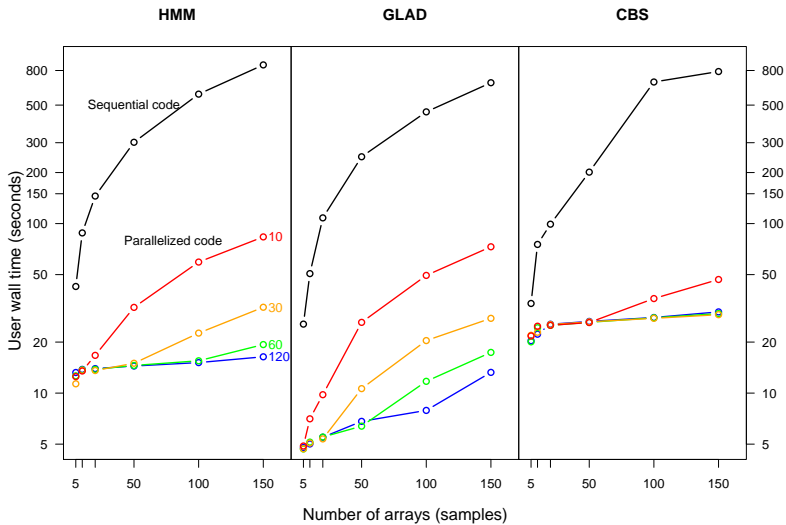
Gains are spectacular.

Is this the way to go? (I)

Gains are spectacular.

Strange (Striking?): we seem to be the only ones to use, systematically, parallelization for bioinfo/biostats web-based applications.

Medium data set (10,000 genes)



Is this the way to go? (II)

Hard to tell.

Is this the way to go? (II)

Hard to tell.

(Automatically) managing, monitoring, sanitizing MPI and R + MPI a pain

Is this the way to go? (II)

Hard to tell.

(Automatically) managing, monitoring, sanitizing MPI and R + MPI a pain

MPI is fragile (use a fault-tolerant PVM instead?)

Is this the way to go? (II)

Hard to tell.

(Automatically) managing, monitoring, sanitizing MPI and R + MPI a pain

MPI is fragile (use a fault-tolerant PVM instead?)

I wish R were Erlang . . .

Is this the way to go? (II)

Hard to tell.

(Automatically) managing, monitoring, sanitizing MPI and R + MPI a pain

MPI is fragile (use a fault-tolerant PVM instead?)

I wish R were Erlang . . .

However it is every day more evident that “The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software”: H. Sutter.

Acknowledgements

- Oscar Rueda: for the work on RJaCGH.
- Funding from Fundación de Investigación Médica Mutua Madrileña and Project TIC2003-09331-C02-02 of the Spanish Ministry of Education and Science
- Ramón y Cajal Programme of the Spanish Ministry of Education and Science
- CNIO users for interesting problems and tool testing
- useRs and developereRs for a vibrant statistical computing community and amazing platform