



# Statistics in Genomics and Proteomics

Wolfgang Urfer

M. Antónia Amaral Turkman, editors

Monte Estoril, Portugal,  
October 5-8, 2005

**27**

ISBN: 989-95011-0-7  
ISBN (13 dígitos): 978-989-95011-0-2  
Título: Statistics in Genomics and Proteomics  
Autor: Urfer, Wolfgang; Turkman, M. Antónia  
Data: 21-03-2006  
Editor: Centro Internacional de Matemática  
Morada: Almas de Freire, Coimbra  
Código Postal: 3040 COIMBRA  
Correio Electrónico: [cim@mat.uc.pt](mailto:cim@mat.uc.pt)  
Telefone: 239 802 370  
Fax: 239 445 380

## Contents

Foreword	1
Chris Cannings	
<i>Modelling protein-protein interaction networks from yeast-2-hybrid screens with random graphs</i>	5
Adriano V. Werhli, Marco Grzegorzczuk, Ming Tso Chiang and Dirk Husmeier	
<i>Improved Gibbs sampling for detecting mosaic structures in DNA sequence alignments</i>	25
Líbia Zé-Zé, Luzia Gonçalves and Maria Antónia Amaral Turkman	
<i>Physical and genetic mapping in whole genome sequencing era: an overview</i>	39
Margarida D. Amaral, Luka A. Clarke, Mónica Roxo-Rosa and Lisete Sousa	
<i>Genomic and proteomic approaches for studying the genetic disease cystic fibrosis</i>	51
Lígia P. Brás and José C. Menezes	
<i>Estimating gene expression missing data using PLS regression</i>	61
Joaquim F. Pinto da Costa, Hugo Alonso, Luís A.C. Roque and Manuela M. Oliveira	
<i>Supervised and unsupervised selection of genes in microarray data</i>	73
Fatma Haouari and Mohamed Limam	
<i>Gene expression measures of oligonucleotide microarray technology</i>	85
Nuno Sepúlveda, Carlos Daniel Paulino and Carlos Penha-Gonçalves	
<i>Bayesian two-gene interaction models in complex binary traits</i>	103
Elisabete Fernandes, Luísa Canto e Castro and Carlos Penha-Gonçalves	
<i>Statistical analysis in mapping quantitative genetic traits</i>	117
Adelaide V. Freitas, Miguel Pinheiro, José L. Oliveira, Gabriela Moura and Manuel Santos	
<i>A new limiting distribution for a statistical test for the homogeneity of two multinomial populations</i>	127





# Workshop on Statistics in Genomics and Proteomics

## Conference Chairman

- Wolfgang Urfer — *Dortmund University, Germany*

## Scientific Committee

- Wolfgang Urfer — *Dortmund University, Germany*
- Terry Speed — *Department of Statistics, University of California, USA*
- Maria Antónia Amaral Turkman — *University of Lisbon, Portugal*
- Luisa Loura — *University of Lisbon, Portugal*

## Local Organizing Committee

- Maria Antónia Amaral Turkman — *University of Lisbon, Portugal*
- Kamil Feridun Turkman — *University of Lisbon, Portugal*
- Lisete Sousa — *University of Lisbon, Portugal*
- Luzia Gonçalves — *Nova University of Lisbon, Portugal*

## Sponsors

- *National Science and Technology Foundation (FCT) of Portugal (FCT)*
- *International Centre of Mathematics (CIM)*
- *Centre of Statistics and Applications, University of Lisbon (CEAUL)*

- *Department of Statistics and Operations Research (DEIO), University of Lisbon*
- *Project MEDAG, POCTI/MAT/44082/2002 (FCT)*
- *Calouste Gulbenkian Foundation*
- *Fundação Luso Americana (FLAD)*
- *British Council Portugal*
- *L'Ambassade de France au Portugal*
- *Banco Espírito Santo (BES)*

# Workshop on Statistics in Genomics and Proteomics

Wolfgang Urfer \*      M. Antónia Amaral Turkman †

## Foreword

This volume contains a selection of papers (invited and contributed) presented at the *Workshop on Statistics in Genomics and Proteomics* that took place in Monte Estoril, Portugal, from 5–8 October 2005.

Genomics and proteomics aim to identify biomarkers that can answer specific clinical questions. The most obvious are markers that can be used for diagnosis and prognosis. Another important issue is to predict a patient's response to a specific drug. Diagnostic markers can themselves be candidates for drug targets. Therefore pharmaceutical companies pursue genomics and proteomics to identify markers that predict toxicity of candidate drugs. They also investigate biological interactions between all small organic molecules and construct metabolomic networks using multivariate approaches. Researchers from several areas are involved in this process, from the identification of the problems, realization of adequate experiments, collection of data, interpretation of results, etc. The analysis of such amount of data offer real challenges to statisticians. By joining their efforts with geneticists, biologists, and computer scientists, statisticians can be of great help in all this process.

There has been in Portugal a growing interest among statisticians to cooperate with researchers in these areas. The organization of this event brought fruitful discussions among statisticians and non-statisticians and we hope that will have a great impact for future collaboration and joint research.

The workshop, organized by Wolfgang Urfer, Antónia Amaral Turkman, Lisete Sousa, Luzia Gonçalves and Feridun Turkman, under the auspices of the *International Center of Mathematics* (<http://www.cim.pt>) and the *Center of Statistics and its Applications* (<http://www.ceaul.fc.ul.pt>), brought together leading researchers in the areas of statistics in genomics and proteomics, who described the state of the art and presented several challenging problems for researchers in Biostatistics and Bioinformatics.

---

\*Department of Statistics, University of Dortmund, Germany

†Department of Statistics and Operation Research, Faculty of Sciences, University of Lisbon, Portugal.  
E-mail: [antonia.turkman@fc.ul.pt](mailto:antonia.turkman@fc.ul.pt).

This workshop, had the participation of 7 keynote speakers and 5 invited speakers, covering the following topics

- Ruedi Aebersold - *Challenges in Data Analysis and Statistical Validation*
- Chris Cannings - *Random Networks in Genetics*
- Dirk Husmeier - *Detecting Mosaic Structures in DNA Sequence Alignments*
- Sophie Schbath - *Statistical Problems Arising in Physical Mapping*
- Terry Speed - *Probabilistic Modelling of Tandem Mass Spectrometry Data*
- Korbinian Strimmer - *Small Sample Statistical Modeling and Inference of Genetic Networks*
- Simon Tavaré - *Statistical Issues for Expression Analysis of Illumina Bead-Based Microarrays*
- Margarida Amaral - *Genomic and Proteomic Approaches to Study the Genetic Disease Cystic Fibrosis*
- Pedro Fernandes - *Systems Biology Approaches Based on Biological Information*
- Mário Silva - *Information Integration of Biological Data Sources*
- Rogério Tenreiro - *Phylogenies, Genome Organization and Taxonomy in Prokaryotes: Dream or Reality?*
- Libia Zé-Zé - *Physical and Genetic Mapping in Whole Genome Sequencing Era: an Overview*

Apart from the invited talks there were also 12 contributed papers and 17 posters.

Thanks are due to Terry Speed who helped to organize the invited programme. We also express our gratitude to all the speakers for their contribution to the high scientific standards of the *Workshop on Statistics in Genomics and Proteomics*.

A selection of four papers illustrating some of the statistical problems currently of interest in bioinformatics will appear in a special issue of REVSTAT - *Statistical Journal* (<http://www.ine.pt/revstat/>). In this edition we present another selection of refereed papers.

The ten papers in this volume illustrate a variety of statistical and biological problems currently of interest in genomics and proteomics.

Cannings argues that knowledge of the structure of the protein-protein interactions network and its variation within and across phyla, should provide insight into evolution, and better understanding of how such networks allow robustness, adaptability and potential for

further evolution. In this paper he presents an overview of some of the available random graph models which are available as tools in this investigation. Werhli et al. propose a modification of MCMC sampling method for detecting mosaic structures in DNA sequence alignments, reducing the computational costs and producing more reliable predictions. Zé-Zé et al. give an overview of the construction of physical and genomic maps, and discuss the advantages of a combined statistical approach during map construction in determining the overlapping probabilities of macrorestriction fragments and directing experimental procedures (saving time and money).

Microarrays are part of a new class of biotechnologies which allow the monitoring of expression levels of thousand of genes simultaneously. The analysis of data produced still constitutes one big challenge to statisticians. Hence, it is not surprising that this constituted the most popular theme for contributions to the workshop. Amaral et al. aim to use microarray data to generate a short list of genes and proteins which are differentially expressed in response to cystic fibrosis transmembrane conductance regulator, in order to propose novel hypotheses about the influence of intracellular molecular interactions on the development of cystic fibrosis pathophysiology. Brás and Menezes present a method for the estimation of missing values in DNA microarray data while Pinto da Costa et al. and Haouari and Limam discuss and propose different methods for the extraction of informative genes from microarray data.

Sepulveda et al. propose a two-gene interaction model for computing the penetrance of complex binary traits and apply their approach to cerebral Malaria data. Fernandes et al. discuss a statistical approach for mapping quantitative genetic traits and illustrate it with data originated from an intercross experiment to identify quantitative trait loci contributing to variance in the amount of immunoglobulin IgM in serum of mice. Freitas et al. propose a new statistic to test the homogeneity of codon contexts of the complete ORFeome sequences of 3 yeast species.

Finally we would like to express our thanks to Lisete Sousa and Luzia Gonçalves for all the work they put in the organization of this event.

Lisbon, February 2006

Wolfgang Urfer and M.A. Amaral Turkman <sup>1</sup>

---

<sup>1</sup>Partially supported by the project POCTI/MAT/44082/2002



# Modelling protein–protein interaction networks from yeast-2-hybrid screens with random graphs

Chris Cannings \*

## Abstract

Random graphs provide us with a tool for the study of naturally occurring networks, such as the WWW, email, metabolic, and protein-protein interaction (**PPI**). In this paper I briefly review some of the issues relating to **PPI** as produced by yeast2hybrid (**Y2H**) experiments, various classes of random graph, and suggest that the “Central Dogma of Biological Networks”, that “biological networks show a power-law distribution for their degree distribution” is not soundly based. Recent work suggests various alternate models fit better to the data.

**Keywords:** Random Graphs, Protein Nets, Power-Law, Domain Model, Sampling.

## 1 PPI from Y2H

The **Y2H** experiment takes two proteins, engineers each of these separately into yeast cells and then observes whether the two yeast types are capable of mating. The two cells can only mate if the two proteins are capable of binding, so **Y2H** screens pairs of proteins for their ability to bind together. Since proteins function in the living cell by binding to other proteins (though to other types of molecule also), this assay gives us information regarding one aspect of protein behavior.

From information on **Y2H** for many pairs of proteins we can produce a graph (or network),  $\mathbf{G} = (\mathbf{V}, \mathbf{E})$  where  $\mathbf{V}$  is the set of vertices (nodes, points) corresponding to the set of proteins considered, and  $\mathbf{E} \subset \mathbf{V} * \mathbf{V}$  is the set of edges, i.e. unordered pairs of vertices, there being an edge wherever the corresponding proteins bind together.

We should note one feature of these graphs; the presence of an edge corresponds to binding between the two proteins involved, but the absence of an edge only indicates that no binding has been observed, whether this is because the two proteins have not been used in a **Y2H** experiment, or they were assayed and failed to bind. Ideally we should be working with a graph in which there were three types of edges corresponding to binding, not binding and not

---

\*Division of Genomic Medicine, University of Sheffield. E-mail: [c.cannings@shef.ac.uk](mailto:c.cannings@shef.ac.uk).

observed. However such data is not usually reported and so we shall work with  $\mathbf{G} = (\mathbf{V}, \mathbf{E})$  as defined above.

We also note that in practice proteins will assemble in complexes, rather than in pairs, and this feature cannot be revealed by **Y2H**, though it can by mass spectroscopy.

## 2 Objectives

In studying the **PPI**'s of organisms we aim to gain insight into how the features of the network affect the robustness, adaptability and efficiency of the organism when we implement some dynamical system on the nodes. We would also like to be able to understand how **PPI**'s evolve through time under mutation, gene-duplication and other evolutionary events. Our aim here is more limited. We take the network as given and our aims are to describe, classify and differentiate between various networks.

## Motifs

A subgraph of  $\mathbf{G} = (\mathbf{V}, \mathbf{E})$  is a graph  $\mathbf{G}' = (\mathbf{V}', \mathbf{E}')$  where  $\mathbf{V}' \subset \mathbf{V}$  and  $\mathbf{E}' \subset ((\mathbf{V}' * \mathbf{V}') \cap \mathbf{E})$ . A subgraph of  $\mathbf{G} = (\mathbf{V}, \mathbf{E})$  induced by  $\mathbf{V}' \subset \mathbf{V}$  is the subgraph  $\mathbf{G}' = (\mathbf{V}', \mathbf{E}')$  where  $\mathbf{E}' = (\mathbf{V}' * \mathbf{V}') \cap \mathbf{E}$ . Within the graph certain subgraphs are of special interest, often referred to as motifs. Thus if we have  $\mathbf{G}^* = (\mathbf{V}^*, \mathbf{E}^*)$  then we might be interested in whether  $\mathbf{G} = (\mathbf{V}, \mathbf{E})$  has **any** subgraphs isomorphic to  $\mathbf{G}^* = (\mathbf{V}^*, \mathbf{E}^*)$ , or in the distribution of the number of subgraphs isomorphic to  $\mathbf{G}^* = (\mathbf{V}^*, \mathbf{E}^*)$  over the class of random graphs we are studying. We will be particularly interested in triangles and cycles since these introduce a measure of the local behaviour and correlation.

Motifs can be of importance also in assessing the similarity between networks. For example, Wuchty et al (2003) [32], examined the **PPI** networks in yeast and compared it with those in five higher organisms, by identifying which of the proteins were conserved and then comparing the subgraphs induced within those networks by that set of proteins.












They demonstrated that there was substantial conservation of many of the motifs, and that the larger motifs were more conserved than the smaller ones. This presumably reflected the fact that a complex motif corresponds to a more complex set of protein interactions which may more difficult to evolve to a fitter configuration than a simpler motif.

## 3 Random graphs

Since in practice we will only be able to obtain individual networks, or sets of distinct but related networks, these *per se* tell us little unless we can provide some model of the ways in which the network might have been generated. Ideally we would like to model the way



**Table 1 Evolutionary conservation of motif constituents**

#	Motifs	Number of yeast motifs	Natural conservation rate	Random conservation rate	Conservation ratio
1		9,266	13.67%	4.63%	2.94
2		167,304	4.99%	0.81%	6.15
3		3,846	20.51%	1.01%	20.28
4		3,649,591	0.73%	0.12%	5.87
5		1,763,891	2.64%	0.18%	14.67
6		9,646	6.71%	0.17%	40.44
7		164,075	7.67%	0.17%	45.56
8		12,423	18.68%	0.12%	157.89
9		2,339	32.53%	0.08%	422.78
10		25,749	14.77%	0.05%	279.71
11		1,433	47.24%	0.02%	2,256.67

The third column gives the number of motifs of a given kind found in the yeast protein interaction network of 3,183 proteins, which we obtained by counting all subgraphs of two-node to five-node motifs (from the set of 28 five-node motifs, we show only two, #10 and #11). We identified 678 proteins that have an ortholog in each of the five higher eukaryotes that we studied and identified all motifs for which each component belongs to this evolutionary conserved protein subset. The natural conservation rate indicates the fraction of the original yeast motifs that is evolutionarily fully conserved, meaning that each of their protein components belongs to the 678 orthologs of the list. For example, we find that 47% of the 1,433 fully connected pentagons (#11) found in yeast have each of their five proteins conserved in each of the five higher eukaryotes. If the topology of motifs does not interfere with the conservation rate of its constituting proteins, a random ortholog distribution should give the same conservation rate for specific motifs as seen in the natural sample. The random conservation rate therefore represents the fraction of motifs that is fully conserved for the random ortholog distribution. The last column gives the ratio between the natural and the random conservation ratios, indicating that all motifs are highly conserved, some (for example, #11) having a natural conservation rate 2,256 times higher than expected in the absence of correlations between protein conservation rate and the topology of a given motif.

Figure 1: Table 1 from Wuchty et al: Motifs in Yeast which are conserved in five “higher” organisms

in which biological networks are constructed and how they evolve under natural selection. There are preliminary attempts at this modeling but here we look at the more limited issue of modeling the underlying network, but not their evolution.

We shall review some of the models of random graphs below, but here we introduce what might be considered the two classical models. The Erdos-Renyi [10] random graph (which we refer to as the **ER** model)  $\mathbf{G} = (\mathbf{V}, \mathbf{E})$  assumes a set of  $n$  vertices  $\mathbf{V}$ , and that the set of edges  $\mathbf{E}$  is constructed from  $\mathbf{V} * \mathbf{V}$  by choosing each possible edge independently and with fixed probability  $p$ , see Bollobás (1985) [5]. This simple structure allows simple calculations of many quantities to be made. For the moment we only concern ourselves with one of the most basic, that of the distribution of the number of edges at a node, the degree of that node. Since for a specific node there are  $n - 1$  potential edges, each occurring independently with probability  $p$  the degree distribution is  $B(n - 1, p)$  ( $B$ =Binomial). Thus for reasonably large  $n$  and small  $p$  the degree distribution is approximately  $P_o(np)$  ( $P_o$ =Poisson). The Random Geometric Graph (referred to here as **RG**) supposes that the vertices are randomly distributed throughout some space (usually  $R^d$ ) and that an edge exists between the two points  $u$  and  $v$  if  $d(u, v) < r$ , where  $d()$  is some appropriate norm, see Penrose(2003)[21].

## 4 Power law

Zipf's law [34] states that the sizes of cities within a country have relative sizes  $1, 1/2, 1/3, \dots, 1/r, \dots$ . Thus the size of the  $r$ th largest city should be approximately  $1/r$  that of the largest. A simple generalization of this distribution is the power law distribution where the probability that some random variable  $X$  takes the value  $x$  is proportional to  $x^{-\gamma}$ , where  $\gamma$  is referred to as the exponent. Thus  $\gamma = 1$  for Zipf's law.

Albert et al (1999) [1] examined the links of the World Wide Web and observed that the degree distribution was very different from Poisson and they suggested that the observations were well fitted by a power law, the probability that a node had degree  $k$  was given by  $P(k) = \alpha k^{-\gamma}$ . Thus  $\log(P(k)) = \beta + \gamma \log(k)$  so the log-log plot is a straight line. Such networks are also often described as scale free (though precisely what this means is often not clear). I interpret it to mean that  $P(u)/P(v) = P(lu)/P(lv)$  so that the ratio of the numbers of nodes with degrees  $x$  and  $y$  does not depend on the specific values of  $x$  and  $y$  but only on their ratio. Thus if one focuses only on nodes of high degree one sees the same sort of distribution as if one focuses only on nodes of small degree.

There are various models which lead to a power distribution including self-organizing criticality (see Bak(1999) [3]), stochastic multiplicative processes with finite resources (see e.g. Wilhelm and Hanggi (2003) [30]) and preferential attachment, which we discuss in more detail below.

Since the paper of Albert et al (1999) [1] hundreds of papers have appeared reporting that other networks are power law, and scale free. These networks have included email (Newman et al (2002)[19]), scientific collaborations (Newman(2001) [18]), metabolic networks (Jeong et al (2000) [14]) and **PPI** networks, which are the focus of this paper. Certainly in the context of biological networks it has become a **Central Dogma** that **Biological Networks are Power Law**. We shall examine this assertion in some detail here, and draw attention to the mounting evidence that this is not the case, while also pointing out some of the problems in the statistical methods used.

## 5 Preferential attachment (PA)

Barabasi and Albert(1999) [2] introduced an interesting model for the growth of a random network, so called "preferential attachment", which gives rise to a power law degree distribution. The idea, which has resonance for the WWW, is that the network is initiated with a few vertices and edges, new vertices are added sequentially and each new node attaches to existing nodes independently with a probability proportional to the degree of the existing node. [2] demonstrate using differential equations that asymptotically the degree distribution is power law.

As an illustration of how the preferential attachment model works we consider the following

scenario, which is adapted from a model of Jordan(2005) [15]. We start with any small tree. At each  $t \in \{1, 2, \dots\}$  we add a new vertex and join it to one of the existing vertices with probability proportional to the degree of that vertex. Note that these assumptions mean that our graph is always a tree (i.e. has no loops).

We shall keep track of all the possible realizations for the form of the graph. Suppose for

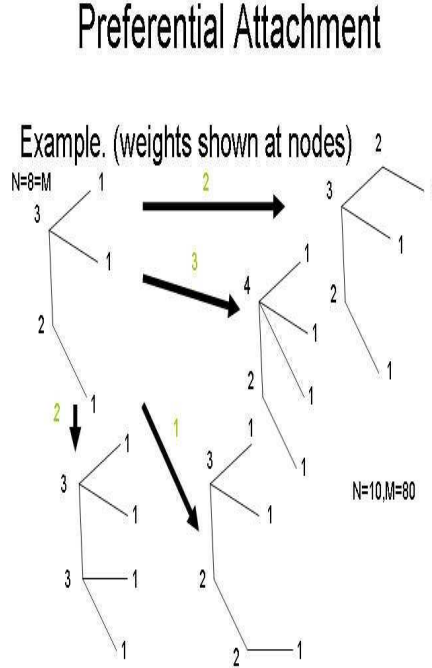


Figure 2: Growing a random graph with preferential attachment: All realisations from a single step.

example that we had the small graph shown at the top left of Figure 2, on which the degrees of the nodes are indicated. The total of the degrees is 8. Now we join a new vertex to the tree in all possible ways, giving rise to four distinct topologies, to which we attach a weight which reflect the number of ways in which that topology can arise. Each of the new trees has total degree 10 and the total of the degrees over all the weighted trees is 80. There are amongst the new trees a total weight attached to nodes of degree 1 of 29, of degree 2 of 18, of degree 3 of 21 and of degree 4 of 12. It is these various weights which we track. Thus take  $n(t)$  =total of degrees within each graph at time  $t$ ,  $m(t)$  =total of weighted degrees across all realizations at time  $t$ , and  $x_i(t)$  = total weighted degrees across all realizations of vertices of degree  $i$  at time  $t$ . Then we have  $n(t) = n(t - 1) + 2$ ,  $m(t) = m(t - 1)n(t)$ ,  $x_1(t + 1) = x_1(t)(n(t) - 1) + m(t)$  and for  $i > 1$   $x_i(t + 1) = x_i(t)(n(t) - i) + i * x_{i-1}(t)$ . It

can be proved that  $x_i(t)/x_{i-1}(t) \rightarrow i/(i+1)$  as  $t \rightarrow \infty$ , for whatever starting configuration we choose and thus the degree distribution has asymptotically  $P(i) = 4/(i * (i+1) * (i+2))$ , so that the tail is approximately of the form  $P(i) = 4/i^3$ , i.e power law with exponent equal to 3. Note here that the expected degree is equal to 2 (inevitably since we add a new node and a new edge at each stage), while the variance is  $\infty$ . See Jordan(2005) [15] for an in depth discussion and proofs.

An interesting extension of this idea is that of Dorogovtsev and Mendes (2000) [9] in which the vertices age, and the chance of a new node attaching to an existing node is proportional to the degree of the latter multiplied by some function of its age. When this function is of the form  $\tau^\pi$  where  $\tau$  is the age then [9] demonstrate that the degree distribution is still power law, with an exponent which depends on  $\pi$ .

Now there are two features of the processes described here “growth” and “preferential” attachment which are claimed as intrinsic to generating a power law for the degree distribution. Despite claims to the contrary, the **ER** graph need not be presented in the static way in which it was introduced above. Indeed many of the issues relating to the **ER** graphs address a growing graph. We illustrate some of these in the next section.

## 6 Erdős–Renyi graph

Suppose instead of the “static” model described above that we initially specify the size of  $\mathbf{V}$  and then add edges sequentially, each time picking the pair of unjoined vertices to join from amongst the candidates with equal probability. Thus the process begins with a set of isolated points and ends with a complete graph, i.e.  $\mathbf{G} = (\mathbf{V}, (\mathbf{V} * \mathbf{V}))$ . As the graph grows it goes through a number of stages, initially there are a number of small subgraphs, then around a critical threshold, where the number of edges equals  $|\mathbf{V}|/2$ , one component (a connected set of nodes isolated from the remainder of the graph) , the so-called giant component, begins to grow rapidly and dominate, there being only a few small bits separate from it, finally everything become connected into a single component. The theory of this process is elegantly developed in Bollobás(1995) [5]. We illustrate the process in Figures 3, 4 & 5 which show how tightly bounded is the size of the largest component.

It might be argued that this process is not truly growing since the number of vertices is fixed throughout. One can add this feature fairly easily, either simply adding a new vertex and then all the edges from that vertex to existing vertices with appropriate probabilities, or alternately at each stage adding a vertex with some probability  $P(v, e)$ , where  $v$  and  $e$  are the existing numbers of vertices and edges. One has the option to make the expected degree grow, shrink or remain fixed during these processes.

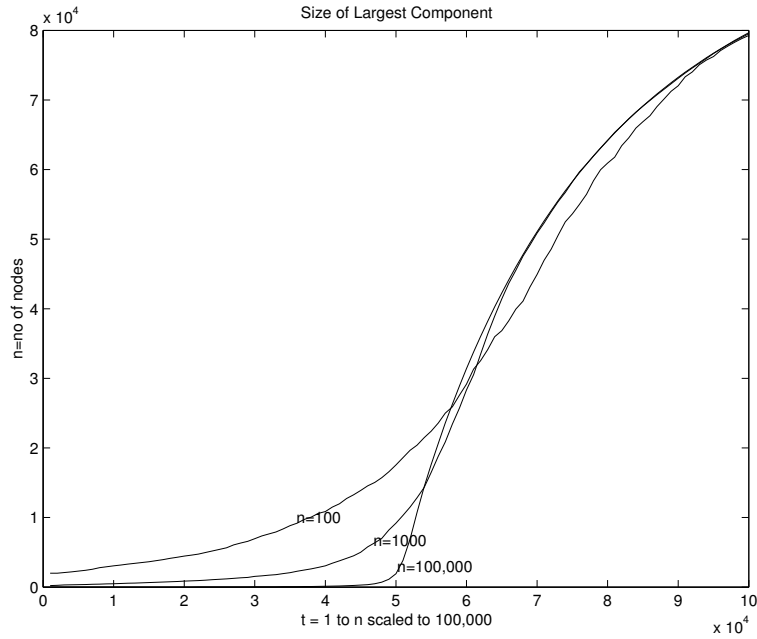


Figure 3: The growth of the largest component for graphs with  $n = 10^2, 10^3, 10^5$ ; abscissa $\cdot n/10 =$  edges added, ordinate $\cdot n/10 =$  nodes in largest component

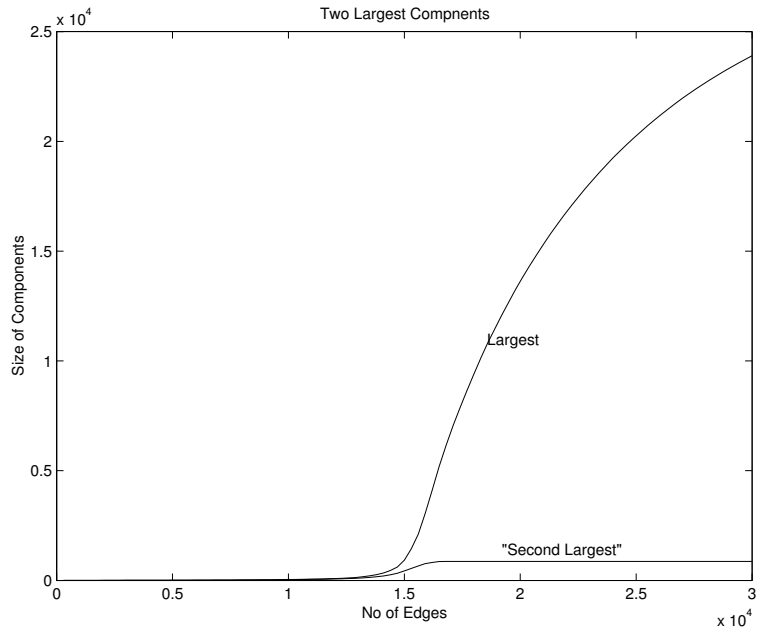


Figure 4: The growth of the largest component and the second largest ever with  $n = 3 \cdot 10^4$ ; abscissa $\cdot n/10 =$  edges added, ordinate  $\cdot n/10 =$  nodes in components.

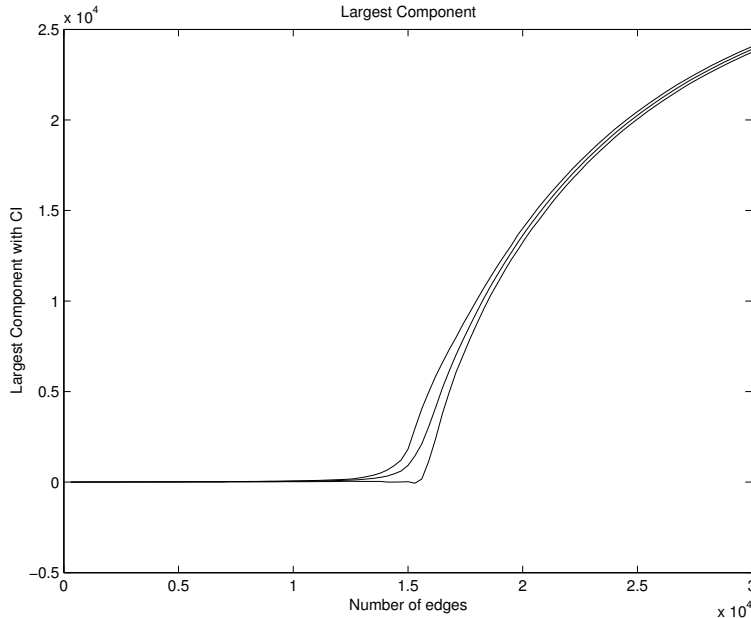


Figure 5: The growth of the largest component with 95% confidence intervals for its size

## 7 More models of random graphs

We have already introduced the **ER**, **RG** and **PA** models. We now introduce other models, the first three of which, Cunnings-Penman(**CP**), Cunnings & Penman(2003) [7], Penman(1998) [20], the Domain Model(**DM**), Thomas et al (2003) [27], and the Tag Model(**TM**) Ravasz and Barabási (2003),[23] have, in contrast to the **ER** model, an intrinsic correlation structure, the fourth, the Small World(**SW**), Watts and Strogatz (1998) [29], which has been designed to have a small diameter ( $= \max_{(i,j)} s(i,j)$  where  $s(i,j)$  is the shortest path between vertices  $i$  and  $j$ ).

### Cunnings–Penman (CP)

Suppose that the vertices are “coloured” from some set  $S = \{C_1, C_2, \dots, C_r\}$  of  $r$  colours with each vertex, independently of the others, having colour  $C_i$  with probability  $p_i$ , and that once the colours have been assigned a pair of vertices  $(i, j)$  are assigned an edge, independently of all others edges, with probability  $p_{u(i),u(j)}$  where  $u(x)$  is the colour of node  $x$ .

Now it is clear that this model has local structure. At one of the extremes if  $p_{u,v} = 1$  if  $u = v$  and  $p_{u,v} = 0$  otherwise, we have  $r$  complete subgraphs, one for each colour. In this case if vertex  $i$  has degree  $k$  and  $i$  is joined to  $j$  then vertex  $j$  also has degree  $k$ . At the other extreme where  $p_{u,v} = 0$  if  $u = v$  and  $p_{u,v} = 1$  otherwise, we have a complete  $r$ -partite graph, and two vertices not joined together (i.e. of the same colour) have the same degree.

We first comment on the degree distribution. This is not, in general Poisson. If for some specific vertex the colour is  $C_i$  then the degree distribution for that vertex, is  $B(n-1, \mu_i)$  where  $\mu_i = \sum_{j=1, r} p_j p_{C_i, C_j}$ , i.e. approximately  $Po(n\mu_i)$ . For a random node, i.e. with unspecified colour we therefore have degree distribution a mixture of Poissons i.e.  $\sum_{i=1, r} p_i Po(n\mu_i)$ , from which one can create a whole range of distributions including power-law.

As an illustration of the difference between **ER** and **CP** we consider the probability that there is a cycle on the nodes  $\{0, 1, 2, \dots, k-1\}$  with  $(i, (i+1) \bmod k) \in \mathbf{E} \forall i$ , which of course will also be the probability for any other cycle of  $k$  nodes. For **ER** this probability is simple  $p^k$ . To simplify somewhat we suppose that we have  $p_{u,v} = \theta$  if  $u = v$  and  $p_{u,v} = \chi$  otherwise, and that each colour has probability of  $1/r$ . We create a Markov chain, indexed by “time”  $\{0, 1, 2, 3, \dots\}$ . We define  $X_t = 0$  if for some pair  $(i-1, i)$  where  $i = 1, \dots, t$  there is no edge, and  $X_t = j$  if there is an edge  $(i-1, i) \forall i = 1, t$  and the vertex  $t$  is coloured  $C_j$ .  $X_t$  keeps track of the path which extends out from node 0 through nodes  $1, 2, \dots$  and the current colour at time  $t$ . We have transition probabilities matrix  $P$  given by

$$P = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ \rho & \theta/r & \chi/r & \dots & \chi/r & \chi/r \\ \rho & \chi/r & \theta/r & \dots & \chi/r & \chi/r \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho & \chi/r & \chi/r & \dots & \theta/r & \chi/r \\ \rho & \chi/r & \chi/r & \dots & \chi/r & \theta/r \end{pmatrix}$$

where  $\rho$  is such that the row sums are one. Now it is straightforward to obtain the eigenvalues  $\lambda_i$  and left eigenvectors  $\mathbf{u}_i$ . These are  $\lambda_0 = 1$  and  $\mathbf{u}_0 = (1, 0, 0, \dots, 0)$ ,  $\lambda_1 = (\theta + (r-1)\chi)/r$  and  $\mathbf{u}_1 = (\tau, 1, 1, \dots, 1)$  where  $\tau$  need not be evaluated, and  $\lambda_j = (\theta - \chi)/r$  and  $\mathbf{u}_j = \delta_j - \delta_{j+1}$  for  $j = 1, r-1$  where  $\delta_j$  is a row vector with a 1 in position  $j$  and 0 elsewhere else. In order to find the probability of the cycle in question we need to evaluate the probabilities that  $X(r-1) = i$ . Wlog we start with  $X(0) = 1$ , so the probability vector at time 0 is given by  $\mathbf{v}_0 = (0, 1, 0, \dots, 0) = \mathbf{u}_1 + \mathbf{u}_r + \dots + (r+1-j)\mathbf{u}_j + \dots + (r-1)\mathbf{u}_2$  so at time  $t$  we have  $\mathbf{v}_t = \{(\theta + (r-1)\chi)^t \mathbf{u}_1 + (\theta - \chi)^t \sum_j j \mathbf{u}_{r-j+1}\}/r^t$ . From this we can derive straightforwardly the probability that the system is in state 1 or in one of the states  $2, 3, \dots, r$  at time  $k-1$  and weighting these by the probabilities  $\theta$  and  $\chi$  that there is then a link to node 0, yield the probability that the specific  $k$ -cycle occurs

$$P(\text{k-cycle on } \{0, 1, 2, \dots, k-1\}) = \{(\theta + (r-1)\chi)^k + (r-1)(\theta - \chi)^k\}/r^k$$

For the special case where  $r = 2$ , i.e. two colours only we have

$$P(\text{2-cycle on } \{0, 1, 2, \dots, k-1\}) = \{((\theta + \chi)/2)^k + ((\theta - \chi)/2)^k\}$$

Now  $((\theta + \chi)/2)^k$  is the probability of the  $k$ -cycle in the **ER** graph with the same overall probability of an edge. In the **CP** graph the probability of a  $k$ -cycle exceeds that of the corresponding **ER** graph (i.e. with same overall edge probability) for all  $k$  when  $\theta > \chi$ , but for  $\theta < \chi$  there are more  $k$ -cycles for even  $k$  and fewer for odd  $k$ .

## 8 Domain model (DM)

This model, presented in Thomas et al (2003) [27], is specifically of relevance to the **PPI** problem. We suppose that there is a set

$$\mathbf{D} = \{A^+, A^-, B^+, B^-, \dots, U^+, U^-\}$$

of  $2d$  domains, which occur in pairs indexed by an ‘+’ or a ‘-’. These domains are entities which exist on the surface of the protein, and are such that a protein with, for example,  $L^+$  will bind to one with  $L^-$ , where  $L$  takes values  $A, B, \dots, U$ , but not to any other. Each protein is assigned a binomial sample of the  $2d$  domains with some probability  $p$  for each, and the assignment to distinct proteins is taken as independent. Two proteins will bind if for any of the “letters” one protein possesses  $L^+$  and the other  $L^-$ .

In the resulting graph the set of proteins which possess an  $L$  of one form, or the other, but not both, will form a complete bipartite component in the graph, as illustrated in Figure 6. If there are also proteins with both  $L^+$  and  $L^-$  then these will be joined to every vertex of the above bipartite component, to every other vertex with both and there will also be a loop indicating that these proteins self-bind. Figure 7 shows a piece of the PRONET database of human protein interactions, which appears to have a pattern suggestive of this model. Of course any graph is made up of complete bipartite pieces, since a single edge is complete bipartite.

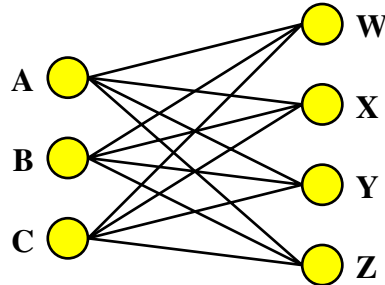


Figure 6: A bipartite subgraph resulting from a single domain; A,B,C have  $L^+$ , W,X,Y,Z have  $L^-$

The **DM** is a special case of the **CP** model but there seems utility in presenting it in this separate form. [27] derive the degree distribution (their equation 6) which is, as explained



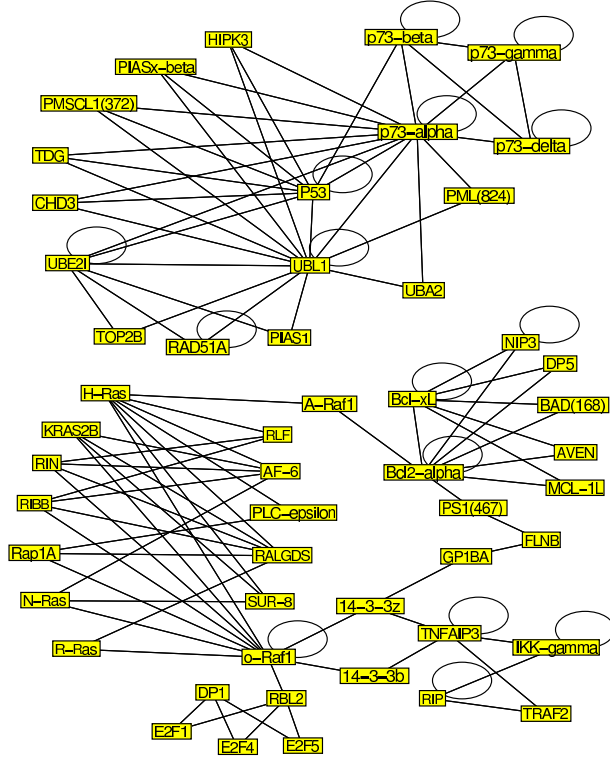


Figure 7: A small piece of PRONET

above, a mixture of binomials, and Figure 8 shows a log-log plot of the degree distribution showing the clear non-linearity.

Finding the probabilities for various cycles is fairly difficult. For a triple of vertices  $\{1, 2, 3\}$  suppose we first consider the letter  $A$ , and focuses initially only on the edges generated by  $A^+$  and  $A^-$  alone. We have, given that the probability that this signed domain is assigned to any protein is  $p$ .

$$\begin{aligned}
 g_0 &= \text{prob}(0 \text{ edges}) = 1 - 6p^2 + 6p^3 + 3p^4 - 6p^5 + 2p^6, \\
 g_1 &= \text{prob}(\text{edges} \subset \{(1, 2)\}) = 1 - 4p^2 + 2p^3 + 4p^4 - 4p^5 + p^6, \\
 g_2 &= \text{prob}(\text{edges} \subset \{(1, 2), (1, 3)\}) = 1 - 2p^2 + p^4
 \end{aligned}$$

then

$$\text{prob}(\text{triangle}) = 1 - 3g_2^d + 3g_1^d - g_0^d$$

Expanding in Maple and removing terms of higher order gives  
 Expected(No of triangles) =  $(n\lambda)^3 \lambda(3 + 3\lambda + \lambda^2)/(48d^3)$  where  $\lambda = 2dp$  equals the average number of domains per protein,  $n$  equals number of proteins. If we keep  $\lambda$  constant, and let

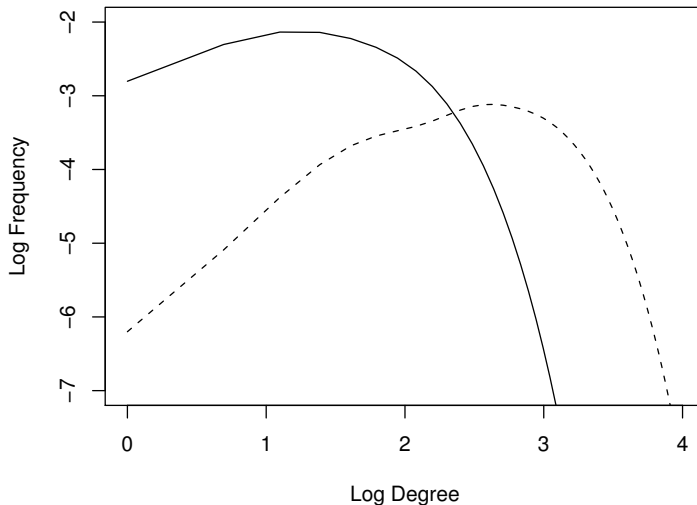


Figure 8: Log-log plot of distribution of vertex degree in an interactome with 6000 proteins, 1000 domains and an average of 1 or 2 domains per protein, shown as solid and dotted lines respectively

$n \rightarrow \infty$  while  $n = \omega d$  for some constant  $\omega$ , we have that Expected(No of Triangles) remains finite, i.e. any given triplet would almost surely not have a complete triangle.

Intuitively we can see why the above result occurs, and indeed extrapolate (informally) to other cycles lengths. Suppose for a  $k$ -cycle that completion of that cycle involves  $l$  different domains letters. Then the least number of signed-domains to achieve this occurs when the proteins with a specific letter are adjacent in the cycle, and alternate  $+$  and  $-$ . We then require  $k + l$  signed domains, except in the case where  $k$  is even and  $l = 1$ . Leaving aside the last case we have that there are  ${}_n C_k$  cycles of length  $k$ ,  ${}_d C_l$  ways of choosing  $l$  colours, we have approximately that the expected number of  $k$  cycles with  $l$  colours is of order  $n^k d^l p^{k+l}$  which remains finite under the condition that  $n = \omega d$  and  $dp$  is constant. On the other hand when  $k$  is even and  $l = 1$  there is an expected number which grows at least linearly. It is also clear that almost all cycles result from a single domain pair.

## 9 Tag models (TM)

A model of current interest is that discussed by Caldarelli et al(2002) [6], Söderberg(2002) [24] and Boguñá and Pastor-Satorras (2003) [4], but earlier described by Meester and Roy (1996)[16]. This is essentially a (potentially) continuous variation of the **CP** model. Each

vertex is assigned a “tag”, a value drawn at random from a probability distribution, and edges are then added with probabilities conditional on the tags at the vertices at the end of the potential edge. The **CP** model is this model with a distribution over a discrete set. As argued above the degree distribution is a mixture of Poisson distributions. This is also demonstrated by [4] though their argument includes an unnecessary approximation (the  $\ln$  term in their equation (19) should be outside of the  $\rho(h')$ ). They also investigate various correlations and expand somewhat on the model of [6], where the distribution of the tag  $h$  is given by density  $\rho(h) = e^{-h}$  for  $h \in [0, \infty)$ , and the probability that an edge exists between vertices with tags  $h$  and  $h'$  is  $r(h, h') = H(h + h' - \eta)$  where  $H$  is the Heavyside step function and  $\eta$  is a constant.

## 10 Small world (SW) & Erdős–Renyi–small world (RG (k,s))

Our final model is the Small World model. Following experiments on communication (see in particular Milgram (1967) [17]) the following model has been formulated. Suppose that we have a set of vertices  $V = \{0, 1, 2, 3, \dots, n-1\}$ . Initially we join nodes  $i$  and  $j$  if  $|i-j| < k$ , so if the vertices represent individuals then we expect each individual to know his near neighbours (arranged on a circle). Now, following Watts and Strogatz(1998) [29] with small probability  $p$  break each edge and replace it with a random edge. This construct has lots of local clustering, but still a smallish diameter, in contrast to the graph initially constructed.

As an illustration of some of the features of these **SW** graphs we consider the giant component again. A simple variant on the **ER** proceeds as follows; at each stage pick  $k$  vertices and join every pair of these i.e. add the clique on these vertices). Fig 9 shows how the giant component grows for several small values of  $k$  (the threshold now occurs where the number of edges added is  $n/2$  ,i.e. after  $n/(k(k-1))$  additions.

Now we introduce a model which allows us to move smoothly from the **ER** to **SW** , **RG** (k,s). For **RG** (k,s) at each stage we pick a node,  $i$  say, at random and then for  $j = 1, 2, \dots, (k-1)$  we either take vertex  $i+j$ , with probability  $s$ , or a random vertex with probability  $r = 1 - s$ . In this way we generate a set of  $k$  points and we join all pairs of these. Thus  $r = 1$  is **ER** and  $r = 0$  is equivalent to covering the circle by random arcs. Fig 10 shows the growth of the largest component as  $r$  varies. The theory of the giant component for the continuous version of **RG**(k,0) has been developed by Huillet(2003) [12].

## 11 Applying to PPI networks

There is now a substantial literature asserting that **PPI** networks are power law in their degree distribution. For example, Yook et al(2004) [33] state, with respect to the **PPI** of yeast and using four databases of such interactions, “...we show how each database supports

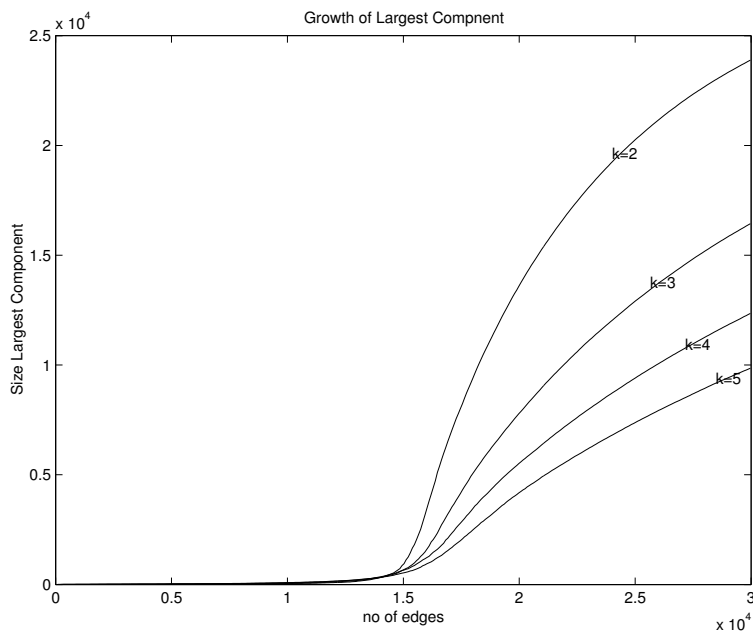


Figure 9: The growth of the largest component for graphs where at each stage a randomly selected  $r$  clique is added.

a scale-free topology ...”. However, in common with much of the literature there is no proper statistical analysis, the fit of the data (in log-log form) to a straight line is performed using least squares, but no consideration given to whether the fit was good in some sense, or whether other models might fit better, nor what, if anything, it actually tells one about the real world. They do not consider the sampling methods, nor the quality of the data. We discuss some of these issues in the context of the **Y2H** assays.

### Quality of the data

As pointed out earlier we should ideally look at a model where an absence of an edge has alternate explanations. There is also a major issue with respect to the **Y2H** positives, i.e. the edges of the network. Proteins which bind in the **Y2H** experimental system may not do so in the organism under investigation, because they never occur in the same cell (the genes switched on in a cell vary from tissue to tissue), they occur at different times or possibly because one or other of the protein binds preferentially with some other molecule and loses its capacity to bind with the other protein. For this reason the **Y2H PPI** network gives us potentially a very inaccurate picture of the real **PPI** network. Sprinzak et al (2003) [25] examined the MIPS, DIPS and BIND data bases. They checked interactions between pairs of 9347 proteins, and for each which showed a **Y2H** interaction checked to see if these proteins co-localised to a common cellular compartment, and whether they were both involved in a

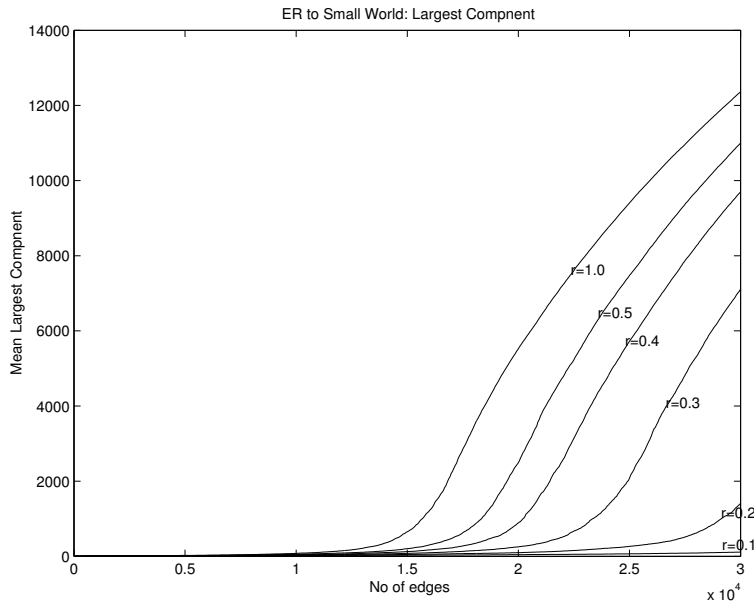


Figure 10: The growth of the largest component for  $\mathbf{RG}(1,r)$ .

common cellular role. They estimated that only 50% of the observed interactions in **Y2H** were true positives.

## Sampling

Most of the **Y2H** interactions are found by laboratories who have a particular interest in some specific protein, or group of proteins, and they compare these with various others so that when a **PPI** network is formed from many such experiments it is by no means a random sample, being more clustered, than the underlying network. Uetz et al (2000) [28] and Ito et al (2001) [13] have, in contrast, carried out substantial laboratory screens of the proteins of yeast. Figure 11 shows the **PPI** network obtained by [28]. However even such screens necessarily have sampling schemes which require careful attention.

Thomas et al(2003) [27] simulated a process imitating the laboratory protocol of [28]. They demonstrated that in both the case where the underlying model was **ER** and **DM** (with parameter values appropriate to yeast and the experimental protocol), the resulting log-log plot was approximately linear, and the same held in the case where the data used was taken from the DIPs data-base Figure 13. In fact the fit to the data was better than that of the power law, by least squares, though it is not entirely clear that this is a good method of fitting, and further work is needed on this topic.

We thus would conclude that there is no firm evidence that a power-law is an appropriate fit to the data, and more importantly it has been shown that even if the fit to the data looks

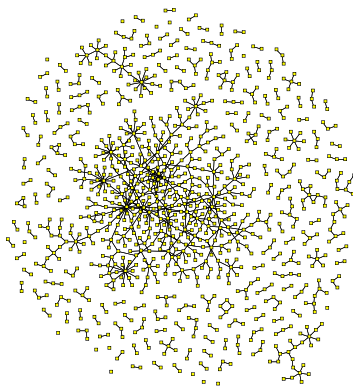


Figure 11: The **PPI** network for the Uetz et al data

approximately like a power-law it is wrong to infer that the underlying distribution is power-law. This argument has been reinforced by the recent work of Han et al (2005) [11], who sampled (in a plausible manner) from a variety of different random graph models and also concluded that this could give “..topological characteristics virtually indistinguishable from those of the currently available **Y2H**-derived partial interactome maps.”. As Stumpf et al (2005) [26] point out even randomly sampling from a power law network will not give one a power law sample. It is clear the problem of inferring the appropriate population distribution is not straightforward. The practice of fitting a distribution to the data, and ignoring the sampling is at best misleading.

## 12 Fitting by motif

Pržulj (2005) [22] fitted the **PPI** networks of both yeast and drosophila. They used the frequencies of each of the possible subgraphs on 3, 4 and 5 nodes, there being 2, 6 and 21 permutationally distinct possibilities respectively. Thus they had 29 objects whose frequencies were used. They then used as a measure of the difference between two graphs  $G$  and  $H$ ,  $D(G, H) = \sum_{i=1,29} |F_i(G) - F_i(H)|$  where  $F_i(X) = -\log(N_i(X)/T(X))$  where  $N_i(X)$  is the number of occurrences of the  $i$ th subgraph, and  $T(X)$  is the total number of the 29 possibilities. They concluded that, when they computed the measure between each of a number of real-world **PPI**'s and **ER**, **PL** and **RG**, with one exception the **RG** fitted best (i.e. had the smallest distance). Deshmukh et al [8] discuss a similar approach and demonstrate that, in the context of the **DM**, the number of unchorded four-cycles is a reliable statistic on which to base parameter estimation, and used this statistic to estimate parameters for a variety of data sets.

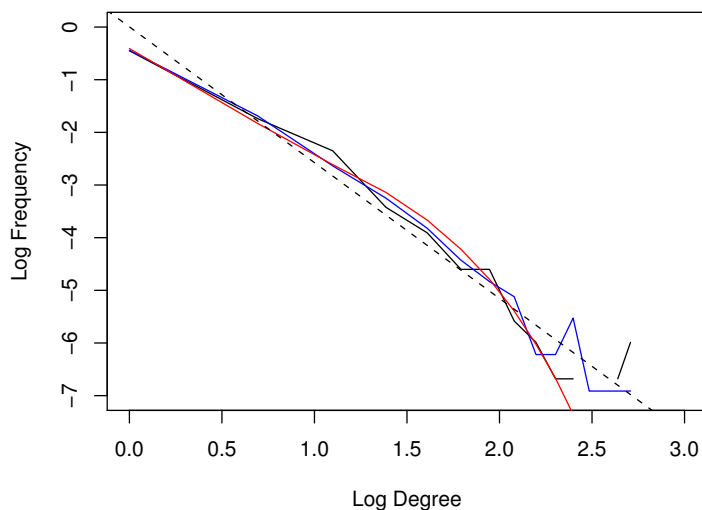


Figure 12: Log-log plots of the observed plots for Uetz and Ito data-sets, the fitted **DM** curve and the power-law (dotted line)

### 13 Discussion

Knowledge of the structure of the **PPI** network and its variation within and across phyla, should provide insight into evolution, and better understanding of how such networks allow robustness, adaptability and potential for further evolution. In this paper I have attempted to present an overview of some of the random graph models which are available as tools in this investigation.

What I have, provocatively, called the **Central Dogma of Biological Networks** says that biological networks are scale-free (power law). The objective here has been to review a variety of random graph models, including some new ones, and to demonstrate, in the context of protein-protein interaction networks, that the evidence for these being scale-free does not survive scrutiny. Inadequate statistical inference, including not carrying out any goodness-of-fit tests, having no proper model of error structure, ignoring the effect of sampling and not considering alternate models have led many authors to report yet another power law. However a series of recent papers have questioned whether **PPI** networks are scale-free (in their degree distribution), both from a theoretical and practical standpoint. It has been suggested that a variety of other models fit better to the existing data in a variety of organisms.

What is also clear is that the statistical methods for choosing between alternative hypotheses needs further development before we can be confident as to the underlying networks.

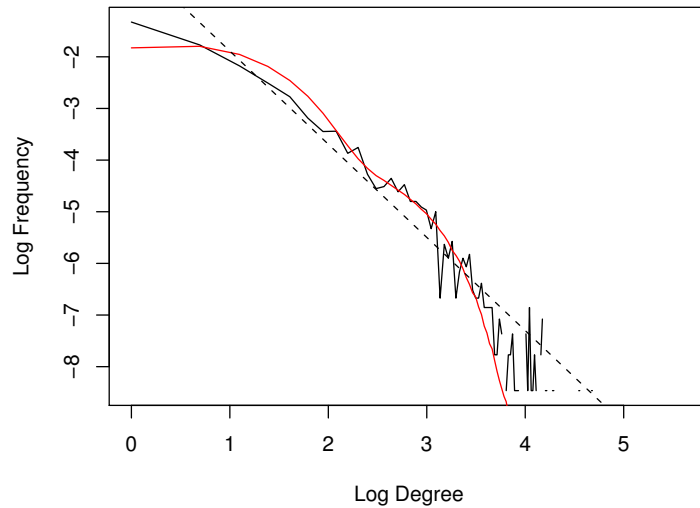


Figure 13: Comparison of the log-log plots of the DIPs data and the model fit (solid lines) and the power-law (dotted line).

Hopefully this contribution will encourage that development.

## References

- [1] Albert R, Jeong H & Barabási A-L (1999) The diameter of the world wide web. *Nature* **401** 130-131.
- [2] Barabási A-L & Albert R (1999) Emergence of scaling in random networks. *Science* **286** 409-412.
- [3] Bak P (1997) *How Nature Works*. Oxford University Press.
- [4] Boguñá M & Pastor-Satorras R (2003) Class of correlated random networks with hidden variables. *Phys Rev E* **68** 036112.
- [5] Bollobás (1985) *Random Graphs*. Acad.Press, London.
- [6] Caldarelli G, Capocci A, Rios PDL & Mñoiz MA (2002) Scale-Free networks from varying vertex intrinsic fitness. *Phys Rev Lett* **89** 258702.
- [7] Cannings C & Penman DB (2003) Models of random graphs and their applications. In *Handbook of Statistics* **21**, 51-91 Ed Shanbhag DN & Rao CR. Elsevier, Netherlands.



- [8] Deshmukh V, Cannings C & Thomas AW Estimating the parameters of a model for protein-protein interaction graphs. *In preparation*.
- [9] Dorogovtsev SN & Mendes JFF (2000) Evolution of reference networks with aging. *Phys Rev E* **62** 1842-1845.
- [10] Erdős P & Rényi A (1959) On random graphs I. *Publ. Maths Debrecen* **6** 290-297.
- [11] Han J-DJ, Dupuy D, Bertin N Cusick ME & Vidal M (2005) Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotech* **23** 839-844.
- [12] Huillet T (2003) Random covering of the circle: the size of the connected components. *Adv Appl Prob* **35** 563-582.
- [13] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M & Sakari Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS* **98** 4569-4574.
- [14] Jeong H, Tombor B, Albert R, Oltvai ZN & Barabási A-L (2000) The large scale organisation of metabolic networks. *Nature* **407** 651-654.
- [15] Jordon J (2005) The degree sequences and spectra of scale-free random graphs. *In press* *Random structures and Algorithms*.
- [16] Meester R & Roy R (1996) *Continuum Percolation*, CUP, Cambridge.
- [17] Milgram S (1967) The small world problem. *Psychol Today* **2** 60-67.
- [18] Newman MEJ (2001) The structure of scientific collaboration networks. *PNAS* **98** 404-409.
- [19] Newman MEJ, Forrest S & Balthrop J (2002) Email networks and the spread of computer viruses. *Phys Rev E* **66** 1-5.
- [20] Penman DB (1998) *Random graphs with correlation structure*. Ph.D. Thesis, University of Sheffield.
- [21] Penrose M (2003) *Random Geometric Graphs*. OUP, Oxford.
- [22] Pržulj N, Corneil DG & Jurisica I (2004) Modelling Interactome: Scale-Free or Geometric. *Bioinformatics* **20** 3508-3514.
- [23] Ravasz E & Barabási A-L (2003) Hierarchical organisation of complex networks. *Phys review E* **67** 026112.
- [24] Söderberg B (2002) General formalism for inhomogeneous random graphs. *Phys Rev E* **66** 066121.

- [25] Sprinzak E, Sattath S & Hanah M (2003) How reliable are experimental protein-protein interaction data. *J mol Biol* **327** 919-923.
- [26] Stumpf MP, Wiuf C, May RM (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. *PNAS* **102** 4221-4.
- [27] Thomas A, Cannings RC, Monk NAM and Cannings C (2003) On the structure of protein-protein interaction networks. *Biochem Soc Trans* **31** 1491-1496.
- [28] Uetz P, Giot L, Cagney G, Mansfield TA Judson RS, Knoght JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kableish T, Vijayadamodar G, Yang M, Johnston M, Fields S & Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in *Saccharimycetes cerevisiae* *Nature* **40** 623-627.
- [29] Watts DJ & Stogatz SH (1998) Collective dynamics of small-world networks. *Nature* **393** 440-442.
- [30] Wilhelm T & Hanggi P (2003) Power-law distributions resulting from finite resource. *Physica A* **329** 499-508.
- [31] Wilhelm T, Nasheuer H-P & Sühnel J (2003) Quantitative analysis of the large-scale organisation of the protein-protein interaction network in yeast. Jena Centre for Bioinformatics. Research Report.
- [32] Wuchty S, Oltvai ZN & Barabási A-L (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature* **35** 176-179.
- [33] Yook S-H, Oltvai ZN & Barabási A-L (2004) Functional and topological characterisation of protein interaction networks. *Proteomics* **4** 928-942.
- [34] Zipf GK (1949) *Human Behaviour and the Principle of Least Effort*. Addison Wesley, Cambridge MA.

# Improved Gibbs sampling for detecting mosaic structures in DNA sequence alignments

Adriano Velasque Werhli <sup>\*</sup>    Marco Grzegorzcyk <sup>†</sup>    Ming Tso Chiang <sup>‡</sup>  
Dirk Husmeier <sup>§</sup>

## Abstract

A recently proposed method for detecting mosaic structures in DNA sequence alignments is based on the combination of hidden Markov models (HMMs) with phylogenetic trees. Inference is done in a Bayesian way by sampling the model parameters and hidden state sequences from the posterior distribution with Markov chain Monte Carlo (MCMC). In an earlier method, proposed in [1], this was effected with a nested Gibbs-within-Gibbs scheme. The present article discusses a modification of the MCMC sampling method, based on a modification of the standard forward-backward algorithm and an unnested Gibbs sampling procedure. We have tested the modified algorithm on various synthetic and real-world DNA sequence alignments, where we have observed a significant improvement in the mixing and convergence of the Markov chain. As a practical consequence, the computational costs are substantially reduced, and the predictions become more reliable.

**Keywords:** Interspecific recombination, phylogenetics, hidden Markov models, Gibbs sampling.

## 1 Introduction

The underlying assumption of most phylogenetic tree reconstruction methods is that there is one set of hierarchical relationships among the taxa. While this is a reasonable approach when applied to most DNA sequence alignments, it can be violated in certain bacteria and viruses due to interspecific recombination. The resulting transfer or exchange of DNA subsequences can lead to a change of the branching order (topology) in the affected region, which results in conflicting phylogenetic information from different regions of the alignment. If undetected,

---

<sup>\*</sup>Biomathematics & Statistics Scotland (BioSS), and School of Informatics, Univ.Edinburgh, United Kingdom. E-mail: [adriano@bioess.ac.uk](mailto:adriano@bioess.ac.uk)

<sup>†</sup>Department of Statistics, Univ. Dortmund, Germany. E-mail: [Grzegorc@statistik.uni-dortmund.de](mailto:Grzegorc@statistik.uni-dortmund.de)

<sup>‡</sup>Department of Computer Science, Birbeck College London, United Kingdom. E-mail: [mingstochiang@dcs.bbk.ac.uk](mailto:mingstochiang@dcs.bbk.ac.uk)

<sup>§</sup>Biomathematics & Statistics Scotland (BioSS), United Kingdom. E-mail: [dirk@bioess.ac.uk](mailto:dirk@bioess.ac.uk)

the presence of these so-called mosaic sequences can lead to systematic errors in phylogenetic tree estimation. Their detection, therefore, is a crucial prerequisite for consistently inferring the evolutionary history of a set of DNA sequences.

A promising approach to detecting interspecific recombination, proposed in [1], is to combine phylogenetic trees and hidden Markov models (HMMs). Such phylogenetic HMMs were originally introduced in [2] to allow for autocorrelation between evolutionary rates at different sites. In [3], a phylogenetic HMM for modelling general mosaic structures in DNA sequence alignments was proposed, with wide applications in the context of functional genomics. The model in [1], which follows up on earlier work in [4] and [5], can be regarded as a special case of the more general model of [3]. The idea is to introduce a hidden state that represents the tree topology at a given site. A state transition from one topology into another corresponds to a recombination event. To introduce correlations between adjacent sites, the hidden states are given a Markovian dependence structure. Thus, the standard model of a phylogenetic tree is generalized by the combination of two probabilistic models: (1) a taxon graph (phylogenetic tree) representing the relationships among the taxa, and (2) a site graph (HMM) representing dependencies between different sites in the DNA sequence alignment. Breakpoints of mosaic segments in the alignment are indicated by state transitions in the site graph. While this method can only deal with a small number of sequences simultaneously (typically as little as 4), it was found to have the potential to predict the locations and breakpoints of recombinant regions more accurately than what can be achieved with most existing techniques [1].

An important question is how to estimate the parameters of a phylogenetic HMM from a training set. This problem was not addressed in [2], where the model parameters had to be selected by the user without reference to any training data. Siepel et al. [3] discuss the case of supervised learning, where the mosaic structure of the DNA sequence alignment is assumed to be known. This approach seems to be meaningful in the context of functional genomics: given a training set of annotated DNA sequence alignments, we want to learn a model that will make meaningful predictions on related unannotated sequence alignments. However, the application to the detection of recombination is different in that we are typically only given a single unannotated sequence alignment. Consequently, training has to be unsupervised, which means that parameter estimation and prediction have to be carried out simultaneously.

The method proposed in [1] addresses this inference problem in terms of a hierarchical Bayesian model. Parameters are first divided into groups. Parameter groups are then sampled from the posterior distribution with Gibbs sampling, that is, one group is sampled conditional on fixed settings of the other groups. While Gibbs sampling is guaranteed to converge to the true posterior distribution [6], this convergence can, in practice, be rather slow. In fact, while the results reported in [1] were very promising, the computational costs of the MCMC simulations were discouragingly high. As it turns out, the bottleneck of the sampling scheme proposed in [1] is a nested Gibbs-within-Gibbs procedure, which slows down the convergence

and mixing of the Markov chain. The objective of the present article is to discuss and evaluate a modified sampling approach. The approach itself, which is based on a modification of the forward-backward algorithm for HMMs [7], is not new: it was introduced in the context of Bayesian non-phylogenetic HMMs in [8]. We will show that its application to the detection of recombination leads to a considerable reduction in the computational costs and, thereby, significantly improves the practical viability of the Bayesian phylogenetic HMM scheme in [1].

## 2 Method

We start with a brief summary of the Bayesian phylogenetic HMM method proposed in [1]. Consider an alignment  $\mathcal{D}$  of  $m$  DNA sequences,  $N$  nucleotides long. Let a column in the alignment be represented by  $\mathbf{y}_t$ , where the subscript  $t$  represents the site,  $1 \leq t \leq N$ . Hence  $\mathbf{y}_t$  is an  $m$ -dimensional column vector that contains the nucleotides at the  $t$ th site of the alignment, and  $\mathcal{D} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ . To model topology changes caused by recombination, a hidden variable  $S_t \in \{\tau_1, \dots, \tau_K\}$  is introduced, which represents the tree topology  $\tau_i \in \{\tau_1, \dots, \tau_K\}$  at site  $t$ . To allow for correlations between nearby sites – while keeping the computational complexity limited – a Markovian dependence structure is introduced:  $P(S_1, \dots, S_N) = \prod_{t=2}^N P(S_t|S_{t-1})P(S_1)$ . The transition probabilities are defined as  $P(S_t|S_{t-1}, \nu) = \nu\delta(S_t, S_{t-1}) + \frac{1-\nu}{K-1}[1 - \delta(S_t, S_{t-1})]$ , where  $\delta(S_t, S_{t-1})$  denotes the Kronecker delta symbol, which is 1 when  $S_t = S_{t-1}$ , and 0 otherwise. Associated with each state  $S_t$  is a vector of branch lengths,  $\mathbf{w}_{S_t}$ , and a set of nucleotide substitution parameters,  $\boldsymbol{\theta}_{S_t}$ , which together define the probability of a column of nucleotides,  $P(\mathbf{y}_t|S_t, \mathbf{w}_{S_t}, \boldsymbol{\theta}_{S_t})$ . The practical computation is easily effected with the pruning algorithm [9]. To simplify the notation, we introduce the accumulated vectors  $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_K)$  and  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$  and define:  $P(\mathbf{y}_t|S_t, \mathbf{w}_{S_t}, \boldsymbol{\theta}_{S_t}) = P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta})$ . This means that  $S_t$  indicates which subvectors of  $\mathbf{w}$  and  $\boldsymbol{\theta}$  apply. The resulting model is an HMM, with emission probabilities  $P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta})$ , and transition probabilities  $P(S_t|S_{t-1}, \nu)$ . In order to proceed with inference according to the Bayesian paradigm, prior probabilities are imposed on the parameters. These priors are chosen to be vague, but proper, and conjugate where possible; see [1] for details.

Recall that the tree topology can change as a consequence of recombination. This corresponds to a state transition  $S_t = \tau_i \rightarrow S_{t+1} = \tau_{(k \neq i)}$ , at the breakpoint  $t$  of the affected region. Our main objective, thus, is the prediction of the state sequence  $\mathbf{S} = (S_1, \dots, S_N)$  or, in order to capture the intrinsic uncertainty of the prediction, the marginal posterior probability

$$P(S_t|\mathcal{D}) = \sum_{S_1} \dots \sum_{S_{t-1}} \sum_{S_{t+1}} \dots \sum_{S_N} P(\mathbf{S}|\mathcal{D}) \quad (1)$$

The distribution  $P(\mathbf{S}|\mathcal{D})$  is obtained by integrating out the model parameters:

$$P(\mathbf{S}|\mathcal{D}) = \int P(\mathbf{S}, \mathbf{w}, \boldsymbol{\theta}, \nu|\mathcal{D}) d\mathbf{w} d\boldsymbol{\theta} d\nu \quad (2)$$

This integral is analytically intractable and has to be numerically approximated with Markov chain Monte Carlo (MCMC): we sample from the joint posterior distribution  $P(\mathbf{S}, \mathbf{w}, \boldsymbol{\theta}, \nu|\mathcal{D})$  and then discard the model parameters. Sampling from the joint posterior distribution follows a Gibbs sampling procedure [6], where each parameter group is sampled separately conditional on the others. So if the superscript  $(i)$  denotes the  $i$ th sample of the Markov chain, we obtain the  $(i + 1)$ th sample as follows:

$$\begin{aligned} \mathbf{S}^{(i+1)} &\sim P(\cdot|\mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}, \mathcal{D}) \\ \mathbf{w}^{(i+1)} &\sim P(\cdot|\mathbf{S}^{(i+1)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}, \mathcal{D}) \\ \boldsymbol{\theta}^{(i+1)} &\sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{w}^{(i+1)}, \nu^{(i)}, \mathcal{D}) \\ \nu^{(i+1)} &\sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{w}^{(i+1)}, \boldsymbol{\theta}^{(i+1)}, \mathcal{D}) \end{aligned} \quad (3)$$

The order of these sampling steps, which are discussed in detail in [1], is arbitrary.

### 3 Improved sampling scheme

To sample the state sequences  $\mathbf{S}$  from  $P(\mathbf{S}|\mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}, \mathcal{D})$ , a Gibbs-within-Gibbs procedure was adopted in [1]. Here, the individual hidden states  $S_t$  of the Markov chain are sampled in separate Gibbs steps:

$$\begin{aligned} S_1^{(i+1)} &\sim P(\cdot|S_2^{(i)}, S_3^{(i)}, \dots, S_N^{(i)}, \mathcal{D}, \mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}) \\ S_2^{(i+1)} &\sim P(\cdot|S_1^{(i+1)}, S_3^{(i)}, \dots, S_N^{(i)}, \mathcal{D}, \mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}) \\ &\vdots \\ S_N^{(i+1)} &\sim P(\cdot|S_1^{(i+1)}, S_2^{(i+1)}, \dots, S_{N-1}^{(i+1)}, \mathcal{D}, \mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}) \end{aligned} \quad (4)$$

The computational complexity of this scheme is reduced considerably by application of conditional independence relations in HMMs (see also Figure 3 in [8] for a graph-theoretical explanation):

$$\begin{aligned} &P(S_t|S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \boldsymbol{\theta}, \nu) \\ &= P(S_t|S_{t-1}, S_{t+1}, \mathbf{y}_t, \mathbf{w}, \boldsymbol{\theta}, \nu) \\ &\propto P(S_{t+1}|S_t, \nu)P(S_t|S_{t-1}, \nu)P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta}) \end{aligned} \quad (5)$$

This sampling procedure was first proposed in [10], and it has also been discussed in [8]. However, Boys et al. [8] conjectured that the resulting mixing and convergence of the Markov chain might be slow due to the large number of component blocks, and they proposed the following alternative sampling scheme. Define

$$\alpha_t(S_t) = P(\mathbf{y}_1, \dots, \mathbf{y}_t, S_t) \quad (6)$$

which is the function computed in the forward pass of the forward-backward algorithm for HMMs; see, for instance, [7]. Now,

$$\begin{aligned}
& P(S_t|S_{t+1}, \dots, S_N, \mathbf{y}_1, \dots, \mathbf{y}_N) \\
& \propto P(S_t, S_{t+1}, \dots, S_N, \mathbf{y}_1, \dots, \mathbf{y}_N) \\
& = P(\mathbf{y}_{t+1}, \dots, \mathbf{y}_N, S_{t+1}, \dots, S_N | S_t, \mathbf{y}_1, \dots, \mathbf{y}_t) \\
& \quad P(S_t, \mathbf{y}_1, \dots, \mathbf{y}_t) \\
& = P(\mathbf{y}_{t+1}, \dots, \mathbf{y}_N, S_{t+1}, \dots, S_N | S_t) \alpha_t(S_t) \\
& = P(\mathbf{y}_{t+1}, \dots, \mathbf{y}_N, S_{t+2}, \dots, S_N | S_{t+1}) P(S_{t+1} | S_t) \alpha_t(S_t) \\
& \propto P(S_{t+1} | S_t) \alpha_t(S_t)
\end{aligned} \tag{7}$$

The simplifications carried out here follow directly from the independence relations in HMMs. The last step follows from the fact that the first term in the second-last line is independent of  $S_t$  and therefore cancels out in the normalization:

$$\begin{aligned}
& P(S_t = \tau_k | S_{t+1}, \dots, S_N, \mathbf{y}_1, \dots, \mathbf{y}_N) \\
& = \frac{P(S_{t+1} | S_t = \tau_k) \alpha_t(S_t = \tau_k)}{\sum_i P(S_{t+1} | S_t = \tau_i) \alpha_t(S_t = \tau_i)}
\end{aligned} \tag{8}$$

Obviously, any scaling constant also cancels out in the normalization; hence replacing  $\alpha_t(S_t)$  by some scaled version for numerical stabilization of the forward algorithm will not affect the result. The algorithm is initialized by drawing the final state,  $S_N$ , from the following distribution:

$$P(S_N = \tau_k | \mathbf{y}_1, \dots, \mathbf{y}_N) = \frac{\alpha_N(S_N = \tau_k)}{\sum_i \alpha_N(S_N = \tau_i)} \tag{9}$$

The overall algorithm can thus be summarized as follows:

- Run the (scaled) forward-backward algorithm.
- Sample  $S_N$  from (9).
- Sample the remaining states  $S_{N-1}, \dots, S_1$  recursively from (8).

Note that at the end of this recursion, henceforth referred to as the stochastic forward-backward algorithm, a whole state sequence  $\mathbf{S} = (S_1, \dots, S_N)$  has been sampled from  $P(\mathbf{S} | \mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}, \mathcal{D})$ .

## 4 Data

We have compared the two sampling schemes on various synthetic and real-world DNA sequence alignments. All alignments contain four sequences; hence the total number of possible tree topologies is  $K = 3$ .

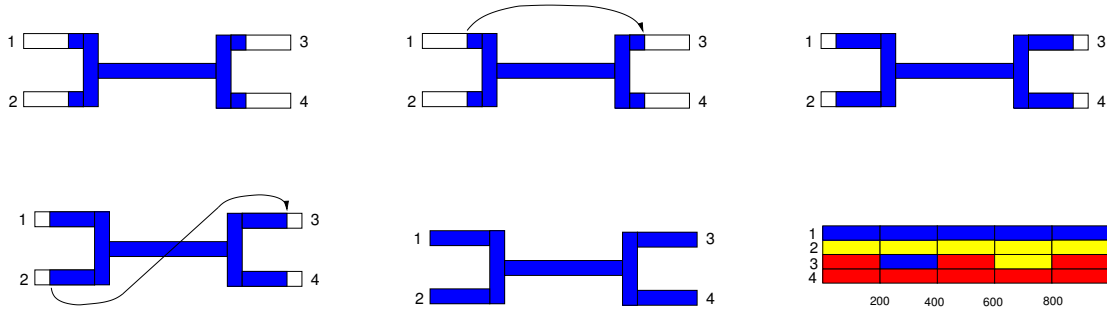


Figure 1: **Simulated recombination.** Four sequences are evolved along the interior branch and the first half of the exterior branches of a phylogenetic tree (top left). At this point, the subsequence between sites 201 and 400 in Strain 3 is replaced by the corresponding subsequence in Strain 1 (top centre). The sequences then continue to evolve along the exterior branches until the branch length is 0.75 times the final exterior branch length (top right). This is followed by a second recombination event, where the subsequence between sites 601 and 800 in Strain 2 replaces the corresponding subsequence in Strain 3 (bottom left). The sequences then continue to evolve along the exterior branches for the remaining length (bottom centre). The resulting mosaic structure is shown in the bottom right. The first, more ancient, recombination event corresponds to a transition from topology  $\tau_1$  into topology  $\tau_2$ , where  $\tau_1 = [(Strain\ 1, Strain\ 2), (Strain\ 3, Strain\ 4)]$  and  $\tau_2 = [(Strain\ 1, Strain\ 3), (Strain\ 2, Strain\ 4)]$ . The second, more recent, recombination event corresponds to a transition from topology  $\tau_1$  into topology  $\tau_3$ , where  $\tau_3 = [(Strain\ 1, Strain\ 4), (Strain\ 2, Strain\ 3)]$ . Note that this simulates a realistic scenario where an ancestor of Strain 3 incorporates genetic material from ancestors of other extant strains, which in each case is followed by subsequent evolution.

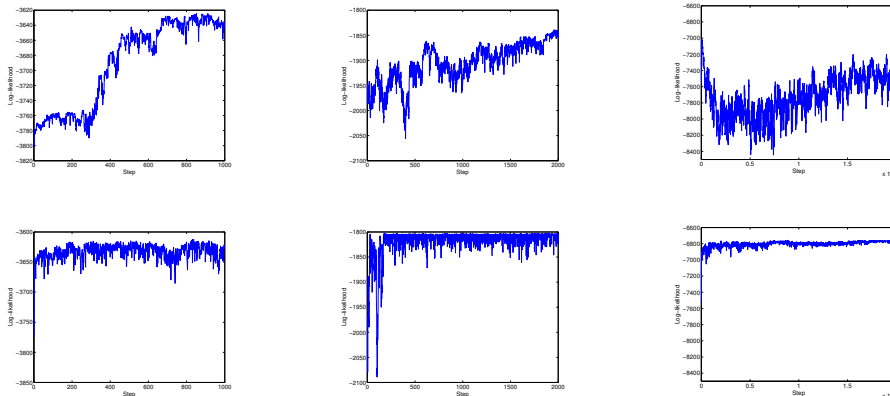


Figure 2: **MCMC trace plots** of the log likelihood. The *top row* shows results obtained with the Gibbs-within-Gibbs scheme; the *bottom row* shows results obtained with the stochastic forward-backward algorithm. The different columns refer to different DNA sequence alignments. *Left column:* Synthetic sequences, long branch lengths ( $w = 0.15$ ). *Middle column:* Synthetic sequences, short branch lengths ( $w = 0.01$ ). *Right column:* Hepatitis B virus.



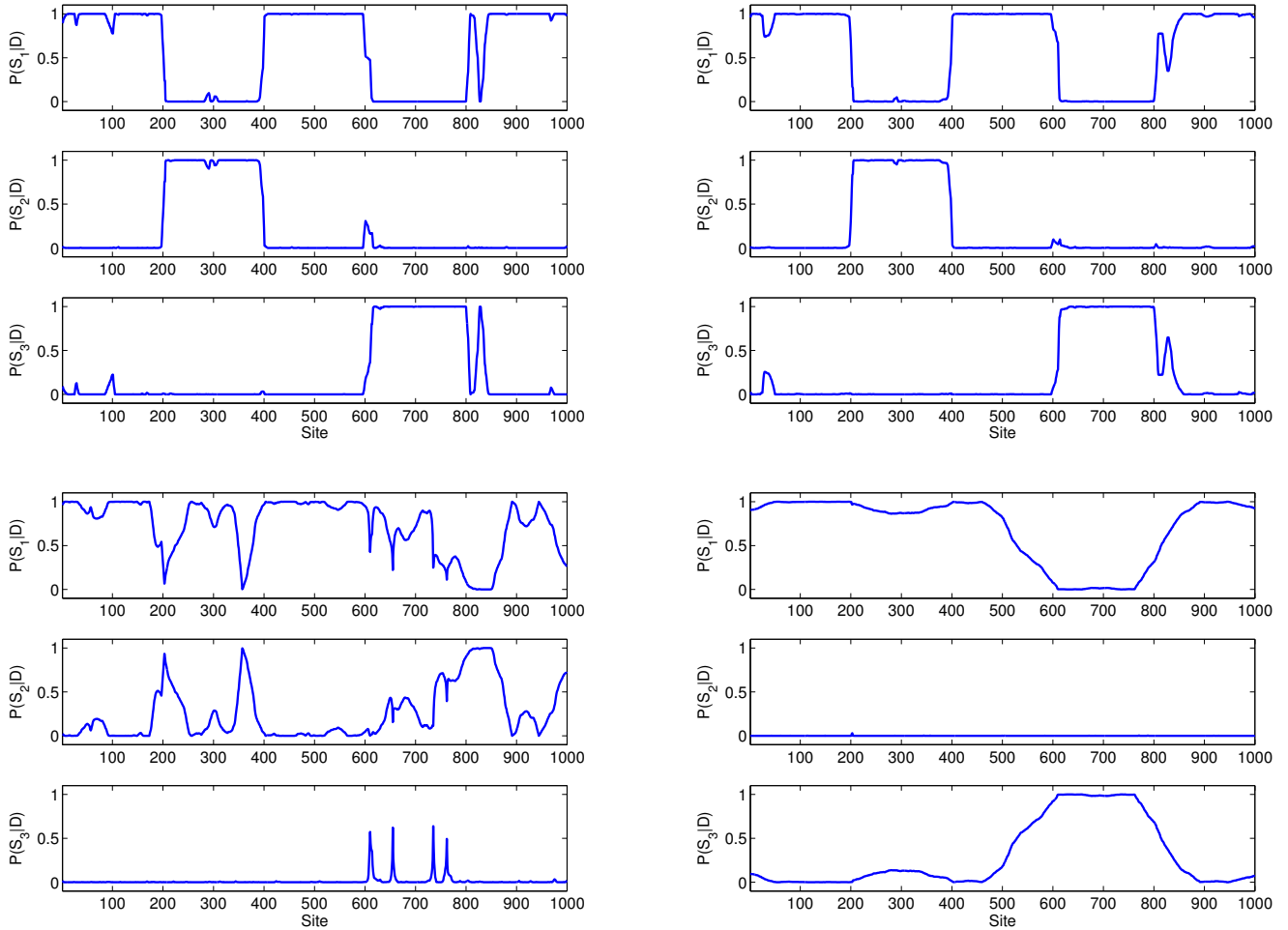


Figure 3: **Synthetic DNA sequence alignment.** The left panel shows the prediction obtained with the Gibbs-within-Gibbs scheme; the right panel shows the prediction obtained with the stochastic forward-backward algorithm. Each panel contains three subfigures, which show the predicted posterior probabilities of the three topologies,  $P(S_t = \tau_1|\mathcal{D})$  (top),  $P(S_t = \tau_2|\mathcal{D})$  (middle), and  $P(S_t = \tau_3|\mathcal{D})$  (bottom), plotted against the site  $t$  in the DNA sequence alignment. Top panel: long branch lengths. The MCMC simulations were run for 500 burn-in and 500 sampling steps. Bottom panel: short branch lengths. The MCMC simulations were run for 1000 burn-in and 1000 sampling steps.

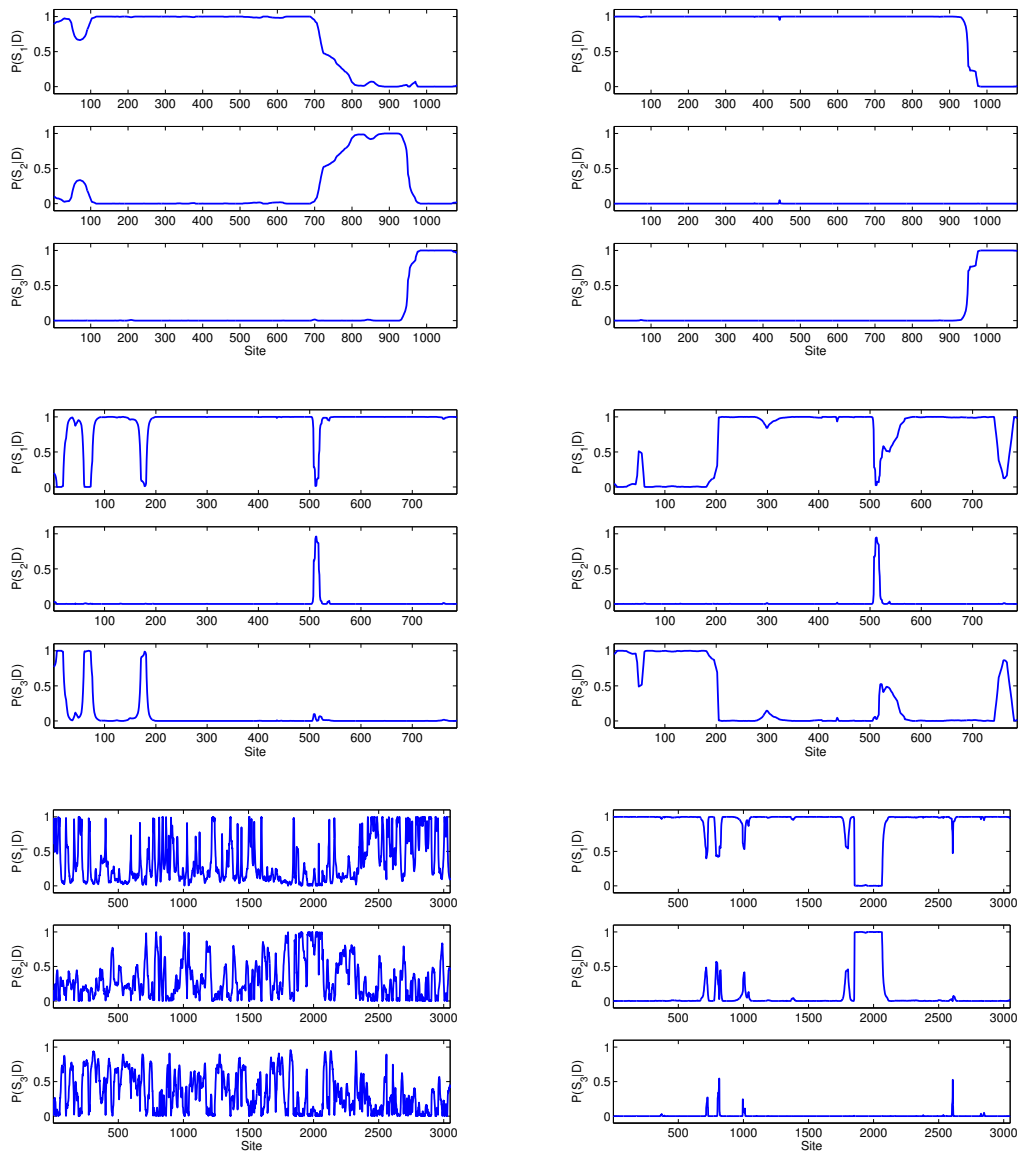


Figure 4: **Predicted mosaic structures in real-world DNA sequence alignments.** The left panel shows the prediction obtained with the Gibbs-within-Gibbs scheme; the right panel shows the prediction obtained with the stochastic forward-backward algorithm. Each panel contains three subfigures, which show the predicted posterior probabilities of the three topologies,  $P(S_t = \tau_1 | \mathcal{D})$  (top),  $P(S_t = \tau_2 | \mathcal{D})$  (middle), and  $P(S_t = \tau_3 | \mathcal{D})$  (bottom), plotted against the site  $t$  in the DNA sequence alignment. Top panel: Maize actin genes (5000 burn-in and 5000 sampling steps). Middle panel: Neisseria (5000 burn-in and 5000 sampling steps). Bottom panel: Hepatitis-B virus ( $10^4$  burn-in and  $10^4$  sampling steps).

**Simulated recombination.** DNA sequences, 1000 bases long, were evolved along a 4-species tree, using the Kimura model of nucleotide substitution [11] with a transition-transversion ratio of 2. Two recombination events were simulated, as shown in Figure 1. Topology  $\tau_1$  is the ‘true’ topology, which applies to those parts of the alignment that are not affected by recombination. The sequence alignment contains two recombinant regions: between sites 201 and 400 (topology  $\tau_2$ ), and between sites 601 and 800 (topology  $\tau_3$ ). The simulations were repeated twice, for long ( $w = 0.15$ ) and for short ( $w = 0.01$ ) branch lengths of the phylogenetic tree. Note that as the branch lengths decrease, the number of polymorphic and topology-defining sites decreases. This reduces the information content in the alignment and renders the detection of recombinant regions more difficult.

**Maize actin genes.** Gene conversion is a process equivalent to recombination, which occurs in multigene families, where a DNA subsequence of one gene can be replaced by the DNA subsequence from another. Indication of gene conversion between a pair of maize actin genes was reported in [12], who showed that the Maz56 and Maz63 genes had a gene conversion covering the last 130 nucleotides of their coding regions. We applied our algorithm to a multiple alignment of the following four maize sequences: Maz56 (GenBank/EMBL accession number U60514), Maz63 (U60513), Maz89 (U60508), and Maz95 (U60507). The sequences were aligned with Clustal-W [13], using the default parameter settings. We define the three tree topologies as follows.  $\tau_1 : [(Maz56, Maz63), (Maz89, Maz95)]$ ;  $\tau_2 : [(Maz56, Maz89), (Maz63, Maz95)]$ ;  $\tau_3 : [(Maz56, Maz95), (Maz63, Maz89)]$ . With this definition, the ‘true’ mosaic structure shows a transition from  $\tau_1$  into  $\tau_3$  at the end of the alignment.

**Neisseria.** One of the first indications for sporadic recombination was found in the bacterial genus *Neisseria* [14]. We chose a subset of the 787-nucleotide *Neisseria argF* DNA multiple alignment studied in [15], where we selected the four strains *N.gonorrhoeae* (X64860), *N.meningitidis* (X64866), *N.cinera* (X64869), and *N.mucosa* (X64873) (GenBank/EMBL accession numbers are in brackets). We define the topologies as follows.  $\tau_1 : [(N.gonorrhoeae, N.meningitidis), (N.cinera, N.mucosa)]$ ;  $\tau_2 : [(N.gonorrhoeae, N.cinera), (N.meningitidis, N.mucosa)]$ ;  $\tau_3 : [(N.gonorrhoeae, N.mucosa), (N.cinera, N.meningitidis)]$ . The mosaic structure is discussed in more detail in [1].

**Hepatitis B** is caused by a DNA virus with a short genome of 3200 bases. Evidence for recombination was found in [16]. In the present paper we investigate a subset of four strains with the following GenBank identifiers (accession numbers in square brackets): HPBADW1 [D00329], HPBADW2 [D00330], HPBADWZCG [M57663], and HPBADRC [D00630]. We define the topologies as follows.  $\tau_1 : [(HPBADW1, HPBADW2), (HPBADWZCG, HPBADRC)]$ ;  $\tau_2 : [(HPBADW1, HPBADWZCG), (HPBADW2, HPBADRC)]$ ;  $\tau_3 : [(HPBADW1, HPBADRC),$

(HPBADWZCG, HPBADW2)]. The sequences were aligned with Clustal-W, using the default parameters. Columns with gaps were discarded, giving a total alignment length of 3049 bases. Bollyky et al. [16] found a recombinant region of 189 bases in HPBADWZCG between  $t = 1865$  and  $t = 2054$  (when not removing gaps:  $t = 2014 - 2203$ ), corresponding to a transition from topology  $\tau_1$  into topology  $\tau_2$ .

## 5 Simulations

We carried out the MCMC simulations as reported in [1], but with a considerable reduction in the burn-in and sampling periods. The MCMC simulations reported in [1] were run for  $10^5 - 10^6$  Gibbs sampling steps, which required several hours of CPU time (on a Sun Ultra-60). We reduced these simulation times by two orders of magnitude to  $10^3 - 10^4$  Gibbs sampling steps. These simulations require only a few minutes of CPU time, and they are hence more realistic in terms of the computational costs a biological end-user is willing to accept. Our main objective was to compare the performance of the stochastic forward-backward algorithm with the Gibbs-within-Gibbs scheme for these shortened simulation times. Note that each iteration of the stochastic forward-backward algorithm traverses the DNA sequence alignment twice: first in the forward direction to compute  $\alpha$ , and then in the backward direction to sample new hidden states  $S_t$ . For a fair comparison between the two sampling schemes, we therefore repeated each iteration of the Gibbs-within-Gibbs scheme (4) twice also: once in the forward, and once in the backward direction, with the order chosen at random. A double traverse of the sequence was then counted as one Gibbs sampling step.

## 6 Results

Figure 2 shows MCMC trace plots of the log likelihood for three DNA sequence alignments. The top panel of the figure clearly demonstrates that the Gibbs-within-Gibbs sampling scheme consistently fails to converge. The bottom panel shows the trace plots obtained with the stochastic forward-backward algorithm, which suggest that the convergence has significantly improved. To evaluate the predictive performance of the two algorithms, we divided each MCMC trajectory into two parts of equal length: a burn-in period, which was discarded, and a sampling phase, from which the marginal posterior probabilities of the  $K = 3$  tree topologies were computed (by application of (1)). Recall that these marginal probabilities allow us to investigate the mosaic structure of a DNA sequence alignment: a plot of  $P(S_t = \tau_i | \mathcal{D})$  along the sequence alignment,  $t = 1, \dots, N$ , for all possible tree topologies,  $i = 1, \dots, K$ , provides clear indications about topology changes and, hence, recombination breakpoints.

The top panel of Figure 3 shows the results obtained on the first synthetic DNA sequence alignment. The difference between the two sampling schemes is minor, and the two recomb-

nant regions are predicted correctly; compare with the bottom right panel of Figure 1.

The bottom panel of Figure 3 shows the results obtained on the second synthetic DNA sequence alignment. Here, the branch lengths of the phylogenetic tree have been decreased from  $w = 0.15$  to  $w = 0.01$ . This considerably reduces the number of polymorphic and topology-defining sites, which renders the identification of the recombinant regions considerably more difficult. In fact, only two topology-defining sites happened to lie in the first recombinant region, which might explain why both sampling schemes failed to detect it. The second recombinant region is clearly detected with the stochastic forward-backward algorithm, while the Gibbs-within-Gibbs scheme leads to an erratic signal as a consequence of insufficient convergence of the Markov chain.

The top panel of Figure 4 shows the prediction of gene conversion in maize actin genes. Both algorithms predict a transition into topology  $\tau_3$  at the end of the alignment, in agreement with [12]. However, when applying the Gibbs-within-Gibbs scheme, this transition is preceded by a transition into topology  $\tau_2$ . This is a spurious gene conversion event, which results from insufficient convergence of the Markov chain and disappears when (considerably) increasing the number of Gibbs sampling steps.

The prediction of recombination in *Neisseria* is shown in the middle panel of Figure 4. When applying the stochastic forward-backward algorithm, we reproduce the results of [1]; these are the marginal posterior probabilities that the Gibbs-within-Gibbs scheme eventually converges to. However, for the chosen simulation length of  $10^4$  Gibbs steps, the Gibbs-within-Gibbs scheme is still far away from convergence. This is shown in the left panel of Figure 4, which strongly deviates from the findings in [1]. Also, as opposed to the stochastic forward-backward algorithm, the Gibbs-within-Gibbs scheme was found to exhibit a strong dependence on the initialization of the Markov chain (results not included), which clearly indicates insufficient convergence.

Finally, the bottom panel of Figure 4 shows the marginal posterior probabilities for the Hepatitis-B virus sequence alignment, as obtained after  $10^4$  Gibbs sampling steps (preceded by a burn-in phase of the same length). The Gibbs-within-Gibbs scheme leads to an erratic signal that does not capture any true features of the mosaic structure of the alignment. On the contrary, the stochastic forward-backward algorithm detects the mosaic structure predicted in [16]. This concurs with the marginal posterior probabilities found in [1], to which the Gibbs-within-Gibbs scheme converges after about  $10^6$  sampling steps, that is, after several hours of CPU time.

## 7 Discussion

The combination of HMMs and phylogenetic trees has proven to be a promising approach to the detection of interspecific recombination in DNA sequence alignments. However, the

Bayesian MCMC sampling scheme proposed in [1] is computationally expensive due to slow convergence and mixing of the Markov chain. At the heart of the problem is the Gibbs-within-Gibbs scheme for sampling the hidden state sequences. This scheme was proposed in [10], and it has been applied in several other applications since; see, for example, [17]. However, Boys et al. [8] conjectured that its mixing and convergence properties might be slow. The authors proposed an alternative sampling scheme, based on a modification of the forward-backward algorithm for HMMs. Their study did not include an explicit comparison between the two methods, though. In the present paper, we have compared both sampling schemes on various synthetic and real-world DNA sequence alignments. Our simulations suggest that with the stochastic forward-backward algorithm, the lengths of the MCMC simulations can be reduced by about two orders of magnitude, from  $10^5 - 10^6$  down to  $10^3 - 10^4$  Gibbs sampling steps. This corresponds to a decrease in the (Sun Ultra-60) CPU time from typically several hours (required for the simulations reported in [1]) to less than 10 minutes. We trust that this considerable reduction in the computational costs renders our phylogenetic HMM method, which has been implemented in a freely available software package [18], a more useful tool for practical applications. Since phylogenetic HMMs are increasingly being applied in functional and comparative genomics, as briefly outlined in the Introduction section, we assume that the findings of our study could be of wider interest beyond the detection of recombination.

## References

- [1] Husmeier, D., McGuire, G.: Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution* **20** (2003) 315–337
- [2] Felsenstein, J., Churchill, G.A.: A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* **13** (1996) 93–104
- [3] Siepel, A., Haussler, D.: Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology* **11** (2004) 413–428
- [4] McGuire, G., Wright, F., Prentice, M.: A Bayesian method for detecting recombination in DNA multiple alignments. *Journal of Computational Biology* **7** (2000) 159–170
- [5] Husmeier, D., Wright, F.: Detection of recombination in DNA multiple alignments with hidden Markov models. *Journal of Computational Biology* **8** (2001) 401–427
- [6] Casella, G., George, E.I.: Explaining the Gibbs sampler. *The American Statistician* **46** (1992) 167–174
- [7] Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77** (1989) 257–286

- [8] Boys, R.J., Henderson, D.A., Wilkinson, D.J.: Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics* **49** (2000) 269–285
- [9] Felsenstein, J.: Evolution trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17** (1981) 368–376
- [10] Robert, C.P., Celeux, G., Diebolt, J.: Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics & Probability Letters* **16** (1993) 77–83
- [11] Kimura, M.: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16** (1980) 111–120
- [12] Moniz de Sa, M., Drouin, G.: Phylogeny and substitution rates of angiosperm actin genes. *Molecular Biology and Evolution* **13** (1996) 1198–1212
- [13] Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22** (1994) 4673–4680
- [14] Maynard Smith, J.: Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34** (1992) 126–129
- [15] Zhou, J., Spratt, B.G.: Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Molecular Microbiology* **6** (1992) 2135–2146
- [16] Bollyky, P.L., Rambaut, A., Harvey, P.H., Holmes, E.C.: Recombination between sequences of Hepatitis B virus from different genotypes. *Journal of Molecular Evolution* **42** (1996) 97–102
- [17] Robert, C.P., Ryden, T., Titterton, D.M.: Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B* **62** (2000) 57–75
- [18] Milne, I., Wright, F., Rowe, G., Marshall, D.F., Husmeier, D., McGuire, G.: TOPALi: Software for automatic identification of recombinant sequences within DNA multiple alignments. *Bioinformatics* **20** (2004) 1806–1807





# Physical and genetic mapping in whole genome sequencing era: an overview

Líbia Zé-Zé\*    Luzia Gonçalves†    Maria Antónia Amaral Turkman‡

## Abstract

Nowadays, a huge amount of genomic sequence information becomes available everyday through genome sequencing projects. However, the physical methods for the construction of bacterial chromosome maps, by a 'top-down' approach using pulsed field gel electrophoresis (PFGE) remains a powerful tool in the study of genome structure and plasticity, namely in the framework of phylogenetics and comparative evolutionary studies.

The laboratory procedures for the construction of bacterial genomic maps, using examples of *Oenococcus oeni* chromosome mapping, will be presented. The advantages of a combined statistical approach during map construction in determining the overlapping probabilities of macrorestriction fragments and directing experimental procedures (and saving time and money) will be discussed.

**Keywords:** Physical mapping; Whole genome sequencing; Overlap probabilities; Bayes' theorem.

**Abbreviations:** STS, Sequence tagged sites; PFGE, Pulsed field gel electrophoresis; 2D-PFGE, Two dimensional-Pulsed field gel electrophoresis; bp, base pairs; kb, kilo base pairs, 1.000 bp; Mb, mega base pairs, 1.000 000 bp; LAB, Lactic acid bacteria.

## 1 Introduction

The location of landmarks along a chromosome or DNA segment constitutes a physical map. These landmarks are signature sequences that in bacterial physical maps are usually restriction sites (specific sequences of 4-8 bp or more that are recognized and cut by restriction

---

\*Universidade de Lisboa, Faculdade de Ciências, Centro de Genética e Biologia Molecular and Instituto de Ciência Aplicada e Tecnologia, Lisboa, Portugal. E-mail: [lmze-ze@fc.ul.pt](mailto:lmze-ze@fc.ul.pt)

†Unidade de Epidemiologia e Bioestatística, Instituto de Higiene e Medicina Tropical-Universidade Nova de Lisboa, and Centro de Estatística e Aplicações-UL. E-mail: [luziag@ihmt.unl.pt](mailto:luziag@ihmt.unl.pt)

‡Universidade de Lisboa, Faculdade de Ciências, Departamento de Estatística e Investigação Operacional and Centro de Estatística e Aplicações-UL. E-mail: [antonia.turkman@fc.ul.pt](mailto:antonia.turkman@fc.ul.pt)

enzymes) and in eukaryotic maps are sequence tagged sites (STS). When genes or other significant genomic targets are located in a physical map, a genetic map is generated.

Comparative analysis of genome structure at intraspecific level enables the identification of the genetic events (namely, homologous recombination, insertion/deletion, duplication, transposition) and some DNA sequences possibly involved in rearrangements (as IS elements, prophages and duplicated regions/genes). This macrodiversity is displayed through the comparative analysis of gene positioning and macrorestriction polymorphisms in the chromosome maps of different strains. Comparison of genomes of strains belonging to divergent groups can also give some insights on the genomic mechanisms driving evolution in a bacterial phylogenetic group.

## 2 The basic and technicalities of physical and genetic mapping

The first step in bacterial physical mapping typically involves the separate use of at least two rare cutting restriction enzymes to cut the chromosome in a manageable number of fragments (ideally, up to 30), called macrorestriction fragments, that are after separated by PFGE and their sizes estimated by linear interpolation with two flanking size standards. PFGE is a specific agarose gel electrophoresis system that by using alternating electric fields enables the separation of large DNA molecules, ranging from 5 kb to more than 2.0 Mb. To avoid chromosome breakage by mechanical shearing during DNA extraction (that would mess up the specific assignment of DNA fragments in the chromosome), DNA molecules are purified in agarose plugs (Sambrook *et al.*, 1989). To help the assignment of the restriction fragments, with each other and in the chromosome, several different approaches can be pursued.

In partial digestion experiments, by using lower enzyme concentration or reaction incubation times, the restriction reaction in suboptimal conditions is promoted. As a result, some of the restriction sites in the chromosome remain uncut and by PFGE electrophoresis the fragments that appear linked can be assigned as consecutive in the chromosome.

In double-digestion experiments, two restriction enzymes are used simultaneously on genomic DNA enabling the determination of fragment overlapping. These analysis of double-digestion experiments can be troublesome, namely if the fragment overlapping is very small (less than 10 kb). To help the assignment of double digestion fragments, a two dimensional PFGE (2D-PFGE) can also be used. Traditionally, in 2D-PFGE experiments (Bautsch, 1988) the restriction profiles (containing all restriction fragments) obtained with one enzyme were excised from the gel, sequentially digested with the second enzyme and submitted to a second electrophoresis. Nevertheless, by this approach the fragments that appear as spots in the gel are frequently ambiguous to assign. In a slightly different version of 2D-PFGE (Zé-Zé *et al.*, 1998), the restriction fragments obtained with one enzyme are individually excised from the gel and, after the second restriction digestion, submitted again to PFGE.

Hybridization experiments are powerful tools in the assignment of DNA probes (as genes, repetitive sequences or genomic library clones) to a specific chromosome site, enabled by the specific annealing of the probe DNA (in single strand) with complementary chromosome sequences previously denaturated (and so, also presented in single strand forms) (Sambrook *et al.*, 1989).

### 3 The case study: a wine bacterium

*Oenococcus oeni* (Dicks *et al.*, 1995) is a lactic acid bacteria (LAB) occurring naturally in wine and related habitats, which is characterized by its peculiar acidophilic nature and growth in media containing high ethanol levels. This bacterium is the main responsible for the malolactic fermentation, a fermentation that happens mostly in red wines enhancing the stability and organoleptic properties of the wine (Garvie, 1986). Regarding the benefits, *O. oeni* strains are used as starter cultures to ensure malolactic fermentation in winemaking processes. Concerning the scientific interest in this species, several phylogenetic studies have considered *O. oeni* as a fast-evolving species and highlighted their evolutionary divergence (Yang and Woese, 1989, Martinez-Murcia and Collins, 1990, Collins *et al.*, 1991, Martinez-Murcia *et al.*, 1993, Morse *et al.*, 1996).

In the bacterial chromosome mapping of *O. oeni* strains (Zé-Zé *et al.*, 1998; 2000) Southern hybridization was used to allocate DNA probes to PFGE blots with restriction profiles generated by five different restriction enzymes. The location of several housekeeping genes was achieved using homologous and heterologous probes from phylogenetically related species. Additional experiments had to be made to determine gene transcription direction of ribosomal genes. The first published physical map in *O. oeni* was of PSU-1 strain chromosome (with 1857 kb) using four restriction enzymes and enabling the mapping of 37 restriction sites using combined PFGE and hybridization experimental approaches (Zé-Zé *et al.*, 1998). Ten genetic markers were also located in PSU-1 chromosome, including IS1165 insertion sequence and the attachment sites of bacteriophages. The physical map presented an appropriate resolution as the majority of map intervals (62.5%) are smaller than 50 kb (Figure 1). In the framework of phylogenetic and comparative evolutionary studies in *Oenococcus*, an *O. oeni* divergent strain GM was also selected for physical map construction, aiming to determine the processes driving genome structure and plasticity in this species. As it can be observed in Figure 2, by the comparison of PSU-1 and GM genomic maps, we can identify several genetic events, namely insertion and deletion events and different location of some DNA sequences (IS elements and prophages attachment sites) (Zé-Zé *et al.*, 2000).

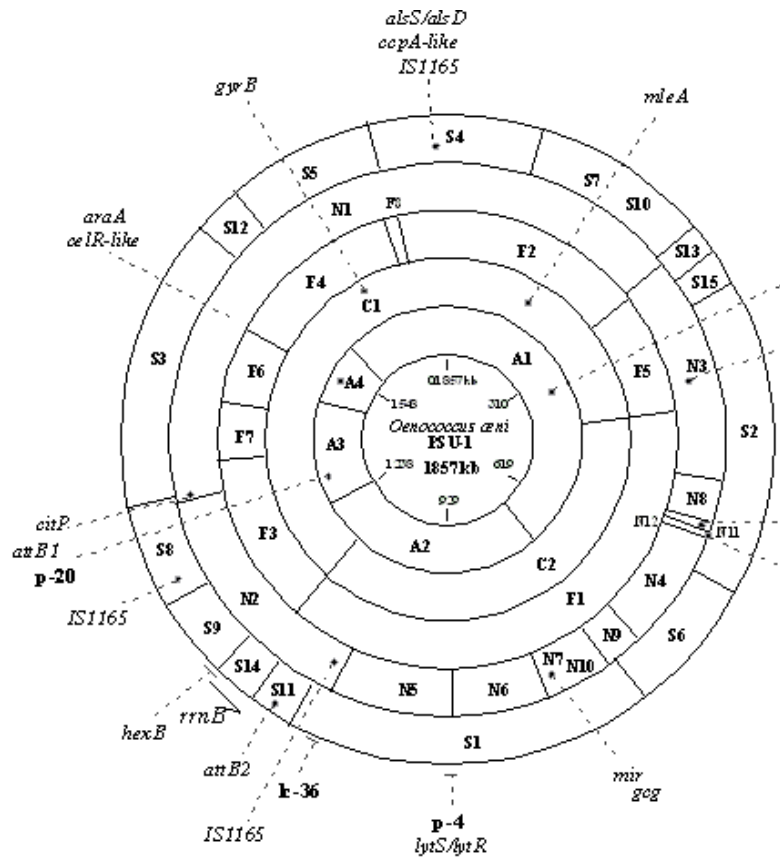


Figure 1: Physical map of the chromosome of *O. oeni* PSU-1 using enzymes *AscI*, *FseI*, *NotI* and *SfiI*. Radiating out from the centre, the four annuli show the restriction sites for the four respective enzymes. Fragments are identified by initial letter of the enzyme and numbered in order from the largest to the smallest; the scale is in kb. The location of several genetic markers is also shown, although the order of markers in a single fragment is arbitrary. Direction of transcription of *rrn* operons is indicated by an arrow (extracted from Zé-Zé *et al.*, 1998).

## 4 Statistics as a tool in genome mapping

### 4.1 Overlap probabilities based on fragment lengths

The overlap detection between fragment pairs is an important step to help in physical mapping. The statistical analysis does not provide the identification of the “true” overlaps, but it can help in simplifying the combinatorial puzzle where pieces are restriction fragments.

The lengths of fragments can be very informative to identify the overlapping relationship between two fragments. Let  $N$  be the genome length in kilobases (kb). Consider two fragments  $A_i$  and  $B_j$ , obtained applying enzymes A and B, separately, with fixed lengths  $L_i$

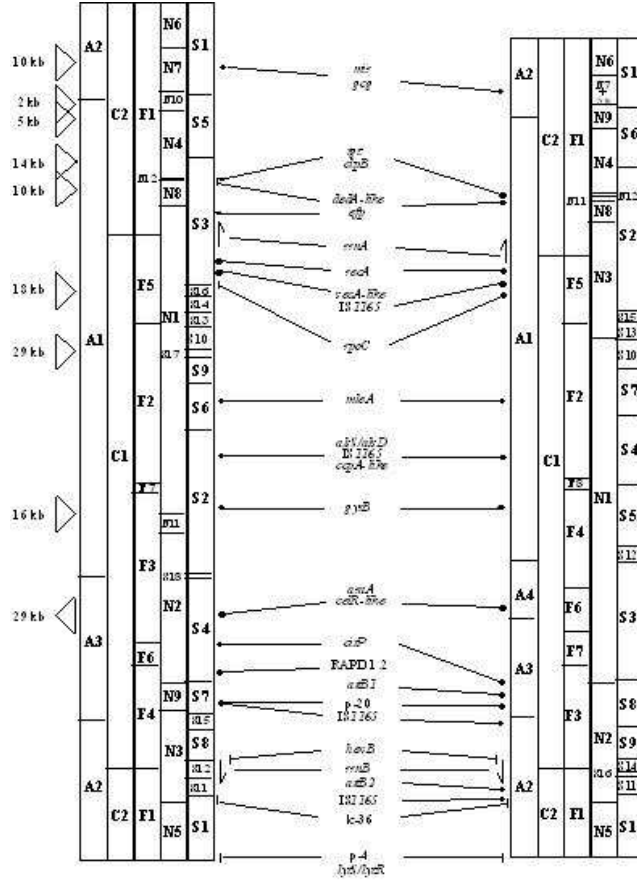


Figure 2: Comparison of the physical and genetic maps of *O. oeni* strains GM and PSU-1. Restriction sites for *AscI*, *I-CeuI*, *FseI*, *NotI* and *SfiI* are indicated. The circular genomes are shown linearized from a common *NotI* site, identified by the linking clone p-4. ▷ represents an insertion and ◁ a deletion event in GM chromosome. The location of several genetic markers to endonuclease restriction sites is also presented (extracted from Zé-Zé *et al.*, 2000).

and  $M_j$ . For example, if we have two fragments of a circular genome  $A_i$  and  $B_j$ , such that  $L_i + M_j > N$ , then they are obligatory overlapped. However, in general,  $L_i + M_j \leq N$ . The overlapping relationship between two fragments obtained by applying two different restriction enzymes, separately, is classified as *nonoverlapping*, *partial overlapping* and *total overlapping*. In mathematical terms, we can describe it using a variable  $\theta_{ij}$  which represents the fraction of overlap ( $0 \leq \theta_{ij} \leq 1$ ). If  $A_i$  and  $B_j$  are nonoverlapped, then  $\theta_{ij} = 0$ . If  $A_i$  and  $B_j$  are totally overlapped, that is, the smaller is contained in the larger, then  $\theta_{ij} = 1$  and if fragments are partially overlapped then  $0 < \theta_{ij} < 1$ .

In a Bayesian perspective, Gonçalves *et al.* (2005) proposed a mixed *prior* probability distribution for  $\theta_{ij}$ . The values 0 and 1 of  $\theta_{ij}$  are pointed out and they represent nonoverlap and total overlap, respectively. Let  $p_0^{ij}$ ,  $p_1^{ij}$  and  $p_2^{ij} = 1 - p_0^{ij} - p_1^{ij}$  be the probabilities of

nonoverlap and total overlap and partial overlap, respectively. For  $\theta_{ij} \in ]0, 1[$  we propose a Uniform distribution. Therefore, we can write,

$$f(\theta_{ij}) = p_0^{ij} I_{\{\theta_{ij}=0\}}(\theta_{ij}) + p_1^{ij} I_{\{\theta_{ij}=1\}}(\theta_{ij}) + p_2^{ij} I_{\{0 < \theta_{ij} < 1\}}(\theta_{ij}) \quad (1)$$

where,

$$I_C(\theta) = \begin{cases} 1, & \text{if } \theta_{ij} \in C \\ 0, & \text{if } \theta_{ij} \notin C. \end{cases}$$

Gonçalves *et al.* (2005) derived expressions for prior probabilities of those events, taking into account DNA fragment lengths and under the assumption that the left-hand endpoints of the two restriction fragments are independent random variables, each of which with a uniform distribution along a circular genome. Here, we only remember those expressions when  $L_i + M_j < N$ ,

$$p_0^{ij} = \frac{N - (L_i + M_j)}{N}, \quad (2)$$

$$p_1^{ij} = \frac{\max(L_i, M_j) - \min(L_i, M_j)}{N}, \quad (3)$$

$$p_2^{ij} = \frac{2 \min(L_i, M_j)}{N}. \quad (4)$$

## 4.2 Adding hybridization data

Prior information, based on fragment lengths, can be combined with hybridization data via Bayes' theorem, in order to evaluate corresponding posterior probabilities (represented by  $p_0^{*ij}$ ,  $p_1^{*ij}$  and  $p_2^{*ij}$ ). Following previous works developed in the context of clone mapping (*e.g.* Nelson and Speed, 1994), we consider a set of  $m$  probes which defines a comparison vector  $D = [d_1, d_2, \dots, d_m]$  where each  $d_s$ ,  $s$  ( $s = 1, 2, \dots, m$ ), is a random variable which can take the following values,

00, if probe  $s$  does not hybridize with neither fragment,

10, if probe  $s$  hybridizes with larger fragment only,

01, if probe  $s$  hybridizes with smaller fragment only,

11, if probe  $s$  hybridizes with both fragments.

Our probes are several genetic markers which are indicated by dashed lines in Figure 1. In the next subsection, we add more details about the assumptions to characterize hybridization data.

Now, we present an example with some pairs of fragments used in the GM map (Table 1). In this particular situation, we only used nine probes (see Gonçalves *et al.*, 2005).

Table 1: Some pairs of fragments of enzymes *AscI* and *FseI*: Prior and posterior overlap probabilities

Fragments		Prior Probabilities			Posterior Probabilities		
<i>FseI</i> vs <i>AscI</i>		$p_0^{ij}$	$p_1^{ij}$	$p_2^{ij}$	$p_0^{*ij}$	$p_1^{*ij}$	$p_2^{*ij}$
	$A_1$ (1089 kb)	0.057	0.184	0.759	0.074	0.070	0.856
$F_1$ (733 kb)	$A_2$ (520 kb)	0.352	0.110	0.538	0.138	0.233	0.629
	$A_3$ (323 kb)	0.453	0.212	0.334	0.868	0.000	0.132
$F_3$ (330 kb)	$A_3$ (323 kb)	0.662	0.004	0.334	0.516	0.000	0.484
$F_7$ (17 kb)	$A_3$ (323 kb)	0.823	0.158	0.019	0.843	0.139	0.018

Despite the reduced number of probes, it seems clear that  $F_1$  should be partial overlapped with  $A_1$  and  $A_2$ .  $F_1$  and  $A_3$  should not be overlapped. Fragments  $F_3$  and  $A_3$  have similar lengths, therefore, prior and posterior probabilities of total overlapping are insignificant values. Therefore, a partial overlapping is possible with posterior probability equal to 0.484. In general, small fragments were not hybridized by any DNA probes. Consequently, prior and posterior probabilities are very similar. The posterior probabilities corresponding to the smallest fragment of *FseI* ( $F_7$ ) tend to indicate total overlapping with the largest fragment of *AscI*.

### 4.3 Full Bayesian analysis

In this previous approach, it was assumed that fragment lengths  $L_i$  and  $M_j$  are constants such that  $\sum_{i=1}^{n_1} L_i = \sum_{j=1}^{n_2} M_j = N$ , where  $N$  is the genome size. However, in practice, the lengths of the fragments and hence the genome size will be somewhat variable due to experimental errors. Taking advantage of previous work (condensed in the last subsection) we can propose a full Bayesian analysis where fragment lengths are considered as random variables. Instead of constants  $L_i$  and  $M_j$ , we consider two random variables  $X_i$  and  $Y_j$  to describe the lengths of fragments  $A_i$  and  $B_j$ . We assume for  $X_i$  a normal distribution with mean  $L_i$  and standard deviation equal to  $rL_i$  and for  $Y_j$  a normal distribution with mean  $M_j$  and standard deviation  $rM_j$  (where  $r$  is the coefficient of variation (CV)). This way of writing the standard deviation is motivated by biological practice. In fact, small fragments can be measured very accurately but large fragments can be associated with larger measurement errors. Values of  $r$  between 1 and 5% seem to be appropriated. In physical mapping of *O. oeni*, strain GM, the maximal relative measurement error (or CV) was estimated to be less than 4.9% for fragments ranging from 6 to 1195 kb (Zé-Zé *et al.*, 2000). Comparing the PSU-1 physical map presented here (Figure 1) with sequencing data from the whole-genome (Mills *et al.*, 2005), it can suggest a value of  $r$  less than 5%.

We still assume for  $\theta_{ij}$  a mixed *prior* probability distribution, however,  $p_0^{ij}$ ,  $p_1^{ij}$  and  $p_2^{ij}$  are now random variables since they are functions of the random lengths. Combining this

information with hybridization data with a set of  $m$  probes, we can obtain posterior probabilities. Relatively to hybridization experiments, we assume that the occurrences of a probe  $s$  ( $s = 1, 2, \dots, m$ ) along the genome are according to a Poisson process with rate  $\lambda_s$  (constant). Additionally, we consider that each probe hybridizes independently of all the other probes. Our aim is to compute the posterior probabilities and for that we implemented all these specifications in WinBUGS14 program<sup>1</sup>. After 4000 burn-in iterations to ensure that the Markov-Chain Monte Carlo (MCMC) algorithm had properly moved away from its starting values, another 6 000 iterations were performed to obtain the posterior distribution of  $\theta_{ij}$ . In our example, the results obtained, with  $r = 5\%$ , are presented in Table 2.

Table 2: Posterior Means of Overlap Probabilities

Fragments		Posterior Means		
<i>FseI</i> vs <i>AscI</i>		$\bar{p}_0^{*ij}$	$\bar{p}_1^{*ij}$	$\bar{p}_2^{*ij}$
	$A_1$	0.078	0.077	0.845
$F_1$	$A_2$	0.143	0.228	0.629
	$A_3$	0.877	0.000	0.125
$F_3$	$A_3$	0.519	0.000	0.481
$F_7$	$A_3$	0.840	0.139	0.018

As can be seen, the estimated posterior probabilities using this approach are very similar to the ones obtained considering the fragment lengths fixed (Table 1 *versus* Table 2). However, this approach describes the biological problem in a more realistic manner.

#### 4.4 After ordering the fragments of an enzyme

It is also possible to combine, for example, the fragment length information with partial digest data to explore other aspects relative to the fragment ordering of one enzyme. By partial digest, in mapping of *O. oeni* PSU-1, the chromosomal order of the four fragments of *AscI* was deduced as  $A_1 - A_2 - A_3 - A_4 - (A_1)$  (Zé-Zé *et al.* (1998)).

In general, given a specific ordering of all fragments ( $n_1$ ) of enzyme A, consider a fragment  $B_j$  of another enzyme with  $n_2$  fragments. The above expressions of partial overlapping (4) and total overlapping probabilities (3) can be used to obtain some probabilities for the location of  $B_j$  fragment in relation to enzyme A fragments. For example, given an order of the fragments of enzyme *AscI* (Figure 1), if we choose a fragment of enzyme *NotI* then  $N_j$  can be located as described in the first column of Table 3. Columns 2 and 3 show this type of probabilities for two fragments of *NotI* ( $N_1$  and  $N_2$ ). In the laboratory, the biologist can fix some fragments of enzyme *NotI* on *Asc* fragments step by step. For example, note  $N_9$  as described in Figure 1. If more than one fragment is located on *Asc* fragments, the expressions of conditional probabilities became more complicated. In order to solve some situations, we have performed

<sup>1</sup>is freely distributed over the internet at <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>



simulations of maps and we have obtained frequencies of events with practical interest. For example, 30 000 000 maps were simulated with all fragments of *AscI* and *NotI* enzymes. We obtained the conditional frequencies referring to  $N_1$  and  $N_2$ , given that  $N_9$  was located on union  $A_1A_2$  (as shown in Figure 1). These conditional frequencies are presented in last two columns of Table 3.

Table 3: Probabilities<sup>§</sup>/Frequencies<sup>†</sup> of events, relating fragments  $N_1$  and  $N_2$  of enzyme *NotI* with *AscI* fragments.

Where is $N_j$ ?	§No fragments of <i>NotI</i> were located		† $N_9$ was located on union $A_1A_2$	
	$N_1$	$N_2$	$N_1$	$N_2$
1. Inside $A_1$	0.102	0.388	0.228	0.478
2. Inside $A_2$	0.000	0.137	0.000	0.265
3. Inside $A_3$	0.000	0.000	0.000	0.000
4. Inside $A_4$	0.000	0.000	0.000	0.000
5. Only on union $A_1A_2$	0.280	0.142	0.000*	0.000*
6. Only on union $A_2A_3$	0.000	0.118	0.000	0.130
7. Only on union $A_3A_4$	0.000	0.046	0.000	0.031
8. Only on union $A_4A_1$	0.070	0.071	0.101	0.039
9. Contains only $A_1$	0.000	0.000	0.000	0.000
10. Contains only $A_2$	0.118	0.000	0.000	0.000
11. Contains only $A_3$	0.039	0.024	0.137	0.020
12. Contains only $A_4$	0.118	0.073	0.181	0.037
13. Overlaps with all <i>Asc</i> fragments	0.272	0.000	0.354	0.000

\*  $N_9$  is here.

Another type of probabilities can also be deduced to help the biologist. Let  $X_j$  be a random variable representing the number of enzyme A fragments which  $B_j$  can overlap. Using the above expressions of partial overlapping and total overlapping probabilities, we can obtain the probability mass function of  $X_j$ . General theoretical expressions can be more or less complicated, but using schematic representations, at least in some practical examples, we can obtain them more easily. For example,  $N_1$  can be overlapped with 1, 2, 3, and 4 *AscI* fragments with probabilities: 0.102, 0.350, 0.276 and 0.273, respectively.

Computation of probabilities of events of interest and simulation studies can help the biologists in this process of inferring the arrangement of restriction fragments along a genome.

#### 4.5 The promoting role of physical and genetic mapping in whole genome sequencing era

In the last decade, the automation of sequencers with the use of dye molecules and the development of capillary electrophoresis has enabled the development of a huge sequencing

capacity. A 96 capillary sequencer can read about 70.000 high quality DNA bases in one hour. So, most of the Institutes involved in whole genome sequencing projects report about 3 billion bases read in one month (<http://www.jgi.doe.gov/sequencing>). As most whole-genome sequencing projects use shotgun strategies, involving the construction of several genomic libraries with different size of cloned fragments, clone redundancy is one of the problems in genome sequencing, namely in the assembling of clones with repetitive sequences. To overcome these problems, a consensus map is constructed. So, the finishing work in whole-genome sequencing projects is usually the most time consuming, involving the gap closing, the improvement of sequences quality and the verification of the assembly process (Strachan and Read, 1999).

Probabilistic characterization and statistical methods can be powerful aids in clone mapping and sequencing. Standard Lander and Waterman (1998) theory has been extended to attempt new biological development in these fields (see, for example, Wendl (2005), Piau (2005) Wendl and Yang (2004) and Wendl and Waterston (2002)).

## 5 Concluding remarks

Albeit we are in the whole-genome sequencing era, the construction of physical and genomic maps still remains a powerful tool in molecular genetics playing a main role, by giving the chromosome landmarks that can help to assemble sequences in whole-genome sequencing projects. The usefulness of physical and genetic mapping is clearly stated by the selection of *O. oeni* strain PSU-1 as the first strain of this species for whole genome sequencing by the Joint Genome Institute (Mills *et al.*, 2005). In fact, PSU-1 was the only *O. oeni* strain with a known physical and genetic map until 2000, the starting data of *O. oeni* whole genome sequencing project. The statistical analysis of data from *O. oeni* strains PSU-1 and GM presented promising results, and showed that there is coincidence between statistical and experimental results (Gonçalves *et al.*, 2005). Laboratory experiments in physical map construction and sequencing can be less time consuming if the experiments are directed with the help of statistical analysis.

## Acknowledgements

This work has been partially supported by the project POCTI/MAT/44082/2002 (FCT - Portugal). We also acknowledge the valuable suggestions from Prof. Rogério Tenreiro.

## References

- [1] BAUTSCH, W. (1988). Rapid mapping of the *Mycoplasma mobile* genome by two-dimensional field inversion gel electrophoresis techniques. *Nucleic Acids Res* **16**, 11461–11467.
- [2] COLLINS, M. D., RODRIGUES, U., AGUIRRE, M., FARROW, J., MARTINEZ-MURCIA, A., PHILLIPS, B., WILLIAMS, A. AND WALLBANKS, S. (1991). Phylogenetic analysis of the genus *Lactobacillus* and related lactic acid bacteria as determined by reverse transcriptase sequencing of 16S rRNA. *FEMS Microbiol Lett* **77**, 5–12.
- [3] DICKS, L. M. T., DELLAGLIO, F. AND COLLINS, M. D. (1995). Proposal to reclassify *Leuconostoc oenos* as *Oenococcus oeni* [corrig.] gen. nov., comb. nov. *Int J Syst Bacteriol* **45**, 395–397.
- [4] GARVIE, E. I. (1984). Separation of species of the genus *Leuconostoc* and differentiation of the leuconostocs from other lactic acid bacteria. In *Methods in Microbiology* **16**, 147–178. London Academic Press.
- [5] GONÇALVES, L., ZÉ-ZÉ L., PINHEIRO, H.P. AND AMARAL-TURKMAN, M.A. (2005) Statistical Aspects in Physical Mapping - Application to the genome of *O. oeni* strain GM. *Biometrics*, **61**, 481–487.
- [6] MARTINEZ-MURCIA, A. J. AND COLLINS, M. D. (1990). A phylogenetic analysis of the genus *Leuconostoc* based on reverse transcriptase sequencing of 16S rRNA. *FEMS Microbiol Lett* **70**, 73–84.
- [7] MARTINEZ-MURCIA, A. J., HARLAND, N. M. AND COLLINS, M. D. (1993). Phylogenetic analysis of some leuconostocs and related organisms as determined from large-subunit rRNA gene sequences: assessment of congruence of small- and large-subunit rRNA derived trees. *J Appl Bacteriol* **74**, 532–541.
- [8] MILLS, D. A., RAWSTHORNE, H., PARKER, C., TAMIR, D. AND MAKAROVA, K. (2005). Genomic anlysis of *Oenococcus oeni* PSU-1 and its relevance to winemaking. *FEMS Microbiol Rev* **29** 465–475.
- [9] MORSE, R., COLLINS, M. D., O'HANLON, K., WALLBANKS, S. AND RICHARDSON, P. T. (1996). Analysis of the beta subunit of DNA-dependent RNA polymerase does not support the hypotesis inferred from 16S rRNA analysis that *Oenococcus oeni* (formerly *Leuconostoc oenus*) is a tachytelic (fast-evolving) bacterium. *Int J Syst Bacteriol* **46**, 1004–1009.

- [10] NELSON, D.O. AND SPEED, T.P.(1994). Statistical Issues in Constructing High Resolution Physical Maps. *Statistical Science* **9**, 334–354.
- [11] PIAU, D. (2005) Invariance Principle for the coverage rate of genomic physical mappings. *The Annals of Applied Probability* **15**, 2553–2574.
- [12] SAMBROOK, J., FRITSCH E.F. AND MANIATIS T. (1989). *Molecular Cloning: a Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- [13] STRACHAN, T. AND READ, A. P. (1999). *Human Molecular Genetics 2*, BIOS Scientific Publishers Ltd, Oxford, UK.
- [14] WENDL, M.C. AND WATERSTON, R.H. (2002) Generalized Gap Model for Bacterial Artificial Chromosome Clone Fingerprint Mapping and Shotgun Sequencing. *Genome research* **12**, 1943 –1949.
- [15] WENDL, M.C.(2005) Probabilistic assessment of clone overlaps in DNA fingerprint mapping via a priori models. *Journal of Computational Biology* **12**, 283–297.
- [16] WENDL, M. C. AND YANG, S. P. (2004) Gap Statistics for Whole Genome Shotgun DNA Sequencing Projects. *Bioinformatics* **20**, 1527–1534.
- [17] YANG, D. AND WOESE, C. R. (1989). Phylogenetic structure of the "Leuconostocs": An interesting case of a rapidly evolving organism. *System Appl Microbiol* **12** 145–149.
- [18] ZÉ-ZÉ L., TENREIRO, R., BRITO, L., SANTOS, M. A. AND PAVEIA, H. (1998). Physical map of the genome of *Oenococcus oeni* PSU-1 and localization of genetic markers. *Microbiology*, **144**, 1145–1156.
- [19] ZÉ-ZÉ L., TENREIRO, R. AND PAVEIA, H. (2000). The *Oenococcus oeni* genome: physical and genetic mapping of strain GM and comparison with the genome of a 'divergent' strain, PSU-1. *Microbiology*, **146**, 3195-3204.

# Genomic and proteomic approaches for studying the genetic disease cystic fibrosis

Margarida D. Amaral \*    Luka A. Clarke †    Mónica Roxo-Rosa ‡  
Lisete Sousa §

## Abstract

Cystic Fibrosis (CF), which is the most common lethal genetic disease with a recessive and autosomal pattern among caucasians, is caused by mutations in the CF transmembrane conductance regulator (CFTR) gene. CFTR protein is an ATP-binding cassette (ABC) transporter expressed at the apical membrane of epithelial cells where it mainly functions as a chloride channel. Our aim is to identify marker genes and proteins showing robust differential expression in human CF-vs-normal epithelial cell lines and native tissues.

To identify genes that are differentially expressed at the level of the messenger ribonucleic acid (mRNA) we used human 40K microarrays covering essentially every gene in the human genome to measure gene expression in two pairs of CF-vs-non CF cell lines and human nasal epithelia. To identify differential abundance of cellular proteins, we analysed samples from nasal cells from CF patients and non-CF controls by two dimensional (2D) gel electrophoresis.

We thereby aim to generate a short list of genes and proteins which are differentially expressed in response to CFTR mutations. We will then be able to propose novel hypotheses about the influence of intracellular molecular interactions on the development of CF pathophysiology.

**Keywords:** Genomics; proteomics; microarrays; Bayesian; rank products; Cystic Fibrosis.

---

\*Department of Chemistry and Biochemistry, Faculty of Sciences, University of Lisbon, and Centre of Human Genetics, National Institute of Health, Portugal. E-mail: [mdamaral@fc.ul.pt](mailto:mdamaral@fc.ul.pt).

†Department of Chemistry and Biochemistry, Faculty of Sciences, University of Lisbon, Portugal. E-mail: [laclarke@fc.ul.pt](mailto:laclarke@fc.ul.pt).

‡Department of Chemistry and Biochemistry, Faculty of Sciences, University of Lisbon, Portugal. E-mail: [maiglesias@fc.ul.pt](mailto:maiglesias@fc.ul.pt).

§Department of Statistics and Operation Research, Faculty of Sciences, University of Lisbon, Portugal. E-mail: [lmsousa@fc.ul.pt](mailto:lmsousa@fc.ul.pt).

## 1 Introduction

Cystic Fibrosis (CF), the most common lethal autosomal recessive disease among caucasians, is caused by mutations in the CF transmembrane conductance regulator (CFTR) gene (Collins, [2]). CFTR protein is an ATP binding cassette (ABC)-transporter that mainly functions as a chloride ( $\text{Cl}^-$ ) channel at the apical membrane of epithelial cells lining the airways, pancreatic ducts, intestine, biliary tree, sweat duct and vas deferens (Collins, [2]). Many studies have revealed that CFTR plays other roles, being implicated in the regulation of several cellular mechanisms (Schwiebert, [7]) (*e.g.*, regulation of other channels and transporters, regulation of vesicle trafficking, modulation of glycosylation) which may explain why CFTR protein dysfunction causes such a pleiotropic and phenotypically variable disease. Indeed, CF affects several organ systems, being clinically characterized by several distinct manifestations, including progressive lung dysfunction, increased saline concentration in sweat and exocrine pancreatic insufficiency (Welsh, [8]).

Despite the many advances in our understanding of CF, the knowledge of the basic mechanisms governing its pathophysiology remains limited. We are therefore very interested in the identification of genes/proteins, which are consistently affected in CF tissues in comparison to normal controls, and which therefore may act as CF biomarkers. For this purpose we have made use of the recent technical advances in the fields of genomic and proteomic research.

## 2 Genomics

In the post-genomic era, following on from the sequencing of the entire human genome, it is no longer sufficient to study individual genes, since most cellular processes, and by extension most diseases, involve highly complex networks of interacting genes. Such interactions can be revealed using the various new technologies and informatics tools which have emerged alongside the international genome sequencing efforts. These include DNA microarrays, which allow the simultaneous measurement of the activities of tens of thousands of genes, and potentially the whole genome. The simultaneous measurement of gene activity might more accurately be termed transcriptomics, since abundance of gene transcripts, *i.e.*, messenger ribonucleic acid (mRNA) molecules, is measured.

As a paradigmatic monogenetic disorder, classical CF is caused solely by mutations in the CFTR gene, but in most cases of CF the expression of many other genes in cascades "downstream" of the CFTR defect will also be altered. If such alterations are consistently induced by CFTR mutations in various different cell or tissue types, then the altered gene sets might be used as "CF markers" in a prognostic or diagnostic capacity similar to the "genetic signatures" recently proposed for some forms of cancer (Ebert, [3]). In our studies, we compared gene expression across the whole human genome in two pairs of human cell lines (with mutated vs. wild type CFTR) and in nasal epithelial cells freshly collected from CF

patients and sex- and age-matched non-CF healthy controls. The human nasal epithelium is an easily collectable tissue yielding numerous and well-preserved dissociated cells. These are representative of the human superficial respiratory mucosa which reflects the physical and biological properties of the CF target tissues of the lower airways (Beck, [1]; Ratjen, [6]). We used such cells as an "in vivo" model tissue for validation of results obtained in cell lines.

For each of three CF-vs-non CF comparisons (two pairs of CF-vs-non CF cell lines and CF-vs-non CF nasal cells) we prepared two replicates and a colour swop, using a total of twelve 40K genomewide arrays (MWG Biotech, Ebersburg, Germany). All experimental steps were performed according to the protocols defined by the supplier. Total RNAs were extracted from samples and in vitro transcription of cRNA incorporating Cy-3 (green) or Cy-5 (red) -labelled nucleotides was performed to achieve labelling of the resulting target molecules. Green or red labelled CF cRNAs were then mixed in a 1:1 molar ratio with their oppositely labelled non-CF counterparts and hybridized overnight to the 40K arrays. Non-specific cRNA was washed off the following morning, and the arrays were scanned at the emission wavelengths for Cy-3 and Cy-5. The raw data consisted of fluorescent emission intensity values for Cy-3 and Cy-5 for each of the approximately 40,000 spots on each of the twelve microarrays, corresponding to the mRNA abundance for each of the human genes represented by the spots in the CF and non-CF total RNA samples. Intensity values from CF samples were divided by non-CF values to give a relative expression ratio for each spot/gene independent of the colour direction (Cy-3/Cy-5 or Cy-5/Cy-3). The gene expression ratios were then "normalized" (Quackenbush, [5]), essentially to bring the average fluorescence intensity/gene expression ratio across all spots and replicate arrays to 1:1 for maximum comparability.

Statistical analyses were undertaken using a Bayesian methodology. This approach was found to improve on the *t*-statistic-based methods because of the large number of genes that are examined simultaneously, the noisiness of the data and the small number of replicates. The method applied was the empirical Bayes method for analysing replicated two-channel microarray data (paired data) proposed by Lönnstedt and Speed [4].

Here the number of arrays is 4, for each of the three pairs of cells studied. Each array consists of approximately 40,000 spots. Therefore,  $M_{ij}$  is the summarised measure, log-ratio of gene expression, for gene  $i$  on array  $j$ ,

$$M_{ij} = \log_2 \frac{X_{ij}^{CF}}{X_{ij}^{NCF}} = Y_{ij}^{CF} - Y_{ij}^{NCF},$$

where  $Y_{ij}^s$  is the logarithm of the expression level,  $X_{ij}^s$ , of gene  $i$  in array  $j$  for sample  $s$  ( $i = 1, \dots, 40,000$ ;  $j = 1, \dots, 4$ ;  $s \in \{CF, NCF\}$ ).

An extreme value (positive or negative) of  $M_{ij}$  suggests that the corresponding gene is differentially expressed between both cell lines. Let  $I_i$  be the indicator for whether a gene is

differentially expressed:

$$I_i = \begin{cases} 0 & \text{if gene } i \text{ is not differentially expressed} \\ 1 & \text{if gene } i \text{ is differentially expressed} \end{cases}.$$

For each gene  $i$  one is interested in whether the gene is differentially expressed, i.e., in  $P(I_i = 1|\mathbf{M}_i)$ . Therefore, Lönnstedt and Speed [4] propose a target measure - a posterior logarithm of odds, for being differentially expressed for each gene  $i$ ,

$$B_i = \ln \frac{P(I_i = 1|\mathbf{M}_i)}{P(I_i = 0|\mathbf{M}_i)},$$

so that  $P(I_i = 1|\mathbf{M}_i) > P(I_i = 0|\mathbf{M}_i)$  if and only if  $B_i > 0$ . Note that  $\mathbf{M}_i = (M_{i1}, M_{i2}, M_{i3}, M_{i4})$ . They rank all genes according to this  $B$ -statistic and name the first few as the potentially differentially expressed genes.

The components of  $\mathbf{M}_i$  are random variables from Normal distribution with mean  $\mu_i$  and variance  $\sigma_i^2$ , so that independently

$$M_{ij}|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i) \quad i = 1, \dots, 40,000, j = 1, \dots, 4.$$

$\tau_i$  is set as  $na/2\sigma_i^2$  to simplify calculations, where  $n$  is the number of arrays ( $n = 4$ ). Consequently, the indicator  $I_i$  is equivalent to

$$I_i = \begin{cases} 0 & \text{if } \mu_i = 0 \\ 1 & \text{if } \mu_i \neq 0 \end{cases},$$

indicating whether gene  $i$  is unchanged ( $\mu_i = 0$ ) or differentially expressed ( $\mu_i \neq 0$ ). The  $M_{ij}$ 's are independent of each other.

To use all their knowledge about means,  $\mu_i$ , and variances,  $\sigma_i^2$ , Lönnstedt and Speed collect the information gained from the complete set of genes in estimating the joint prior distributions for  $\mu_i$  and  $\sigma_i^2$ . They let the prior distribution of the precision  $1/\sigma_i^2$  be Gamma and that of  $\mu_i$  given  $\sigma_i^2$  be Normal. This is a conjugate prior distribution allowing the calculation of  $B_i$  explicitly. For  $\nu$  degrees of freedom and scale parameters  $a > 0$  and  $c > 0$ , suppose

$$\tau_i \sim \Gamma(\nu, 1)$$

$$\mu_i|\tau_i \begin{cases} = 0 & \text{if } I_i = 0 \\ \sim N(0, cna/2\tau_i) & \text{if } I_i = 1 \end{cases},$$

for all  $i = 1, \dots, n$ . The parameter  $c$  expressing dependence between the priors for  $\mu_i$  and  $\tau_i$  is necessary for the calculations.

By Bayes' theorem and independence across genes

$$B_i = \ln \frac{P(I_i = 1|\mathbf{M}_i)}{P(I_i = 0|\mathbf{M}_i)} = \ln \left( \frac{p}{1-p} \times \frac{f_{I_i=1}(\mathbf{M}_i)}{f_{I_i=0}(\mathbf{M}_i)} \right), \quad (1)$$



where  $p$  is the proportion of differentially expressed genes in the experiment,  $p = P(I_i = 1)$ , for any  $i$  in  $1, \dots, 40,000$ . Usually  $p$  refers to a small portion of genes, say 1%.

The final form of the  $B$ -statistics depends on the densities  $f_{I_i=1}(\mathbf{M}_i)$  and  $f_{I_i=0}(\mathbf{M}_i)$ , which correspond to integrating the joint densities  $f_{I_i=k}(\mathbf{M}_i, \mu_i, \tau_i)$  (for  $k=1$  and  $0$ , respectively) in order to  $\mu_i$  and  $\tau_i$ . Hence for gene  $i$ , from (1),

$$B_i = \ln \frac{p}{(1-p)\sqrt{1+nc}} + \left( \nu + \frac{n}{2} \right) \ln \left( \frac{a + s_i^2 + \overline{M}_i^2}{a + s_i^2 + \frac{\overline{M}_i^2}{1+nc}} \right).$$

The only gene specific part of  $B_i$  is the last ratio, which is always  $> 1$  since  $1/(1+nc) < 1$ . It is deducible that an increasing differential expression (and hence an increasing  $\overline{M}_i^2$ ) increases  $B_i$ , all the more if the variance is small. If  $\overline{M}_i^2$  is small too,  $a$  ensures that the ratio cannot be expanded by a very small variance.

There are four global parameters in the model for  $B_i$ :  $p$ ,  $\nu$  and  $a$  (parameters in the prior distribution of the variance) and  $c$  (in the prior distribution of the mean). Unfortunately, there is no consistent estimate of  $(p, \nu, a, c)$ . Therefore, Lönnstedt and Speed fix  $p$  and estimate  $\nu, a|p$  and  $c|p, \nu, a$ . This approach imposes a light modelling structure on the observations, and is described primarily as a way of ranking genes, which suits well this exploratory type of study. Furthermore, it has been shown that this approach has lower false negative and false positive rates than  $t$ -statistic-based methods.

As this method is available in  $R$ , through the function `stat.bay.est` (library `sma`), it was very simple to determine the differentially expressed genes. Setting the proportion of differentially expressed genes to be 0.0015, 55 and 65 genes were selected in the cell lines pairs, which are designated here as CL1 and CL2, respectively, but the thresholds had to be set as -6.1 for CL1 and -1.1 for CL2. The estimates of the parameters were  $\nu = 3.03$ ,  $a = 10.60$  and  $c = 4.15$  for cell lines pair CL1, and  $\nu = 2.18$ ,  $a = 0.55$  and  $c = 5.84$  for cell lines pair CL2. The volcano plots shown in figure 1 are not very symmetrical, leading to selecting a different number of up- and down-regulated genes for each cell lines pair: 36 down- against 19 up-regulated for cell lines pair CL1; 51 down- against 14 up-regulated for cell lines pair CL2. Note that if the threshold was 0, no genes would be selected for cell lines pair CL1 and only 27 would be selected for cell lines pair CL2. To consider gene  $i$  to be differentially expressed only if  $B_i > 0$ ,  $p$  has to be increased. If  $p = 0.337$  the same 55 genes are selected for cell lines pair CL1 and if  $p = 0.0043$  the same 65 genes are selected for cell lines pair CL2, all corresponding to  $B_i > 0$ . These results may be due to lack of Normality of the data and to the small number of replicates.

Lists of the most significantly up- or down-regulated genes in CF-vs.-non CF cell lines and nasal cells were compared, but it was found that very few individual genes were shared between data sets. This may be due to high experimental variability and/or to phenotypic

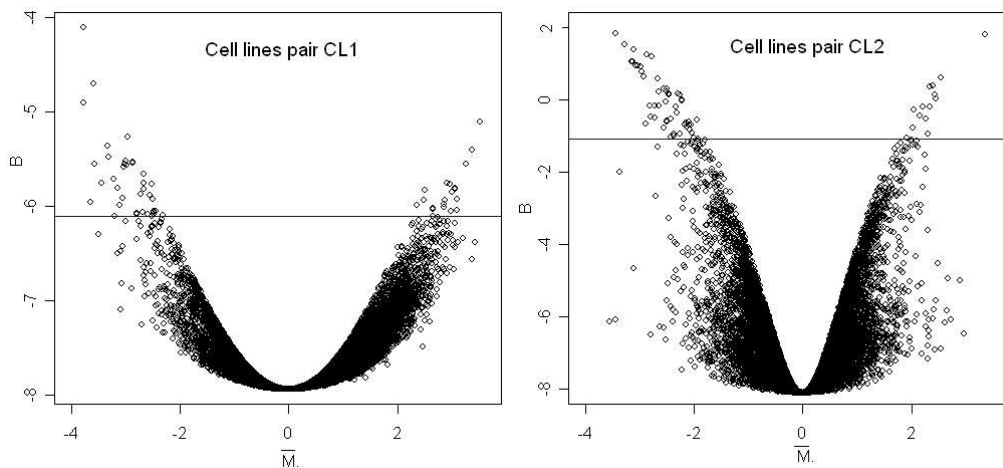


Figure 1:  $B$  vs  $\bar{M}$  plots for cell lines pairs CL1 and CL2. The dots above the horizontal line correspond to genes selected as differentially expressed for  $p = 0.0015$  and thresholds of  $-6.1$  (CL1) and  $-1.1$  (CL2).

differences between the cultured human cell lines and the fresh native tissues from patients. However, when genes are assigned to their functional groups it becomes clear that many closely related, although not identical, genes are common among the data sets. Gene families involved in cellular processes highly relevant to the pathophysiology of CF, such as inflammation, proteolysis, extracellular matrix remodelling, ion transport, cell adhesion and bone development, are thus identified in all data sets. This finding highlights the importance of more global approaches to the analysis of such vast amounts of data using specific software that identifies gene clusters and pathways.

### 3 Proteomics

To identify potential CF biological markers at the protein level, we compared the protein profiles of nasal epithelial cells from CF versus non-CF individuals using a classical proteomic methodology. Proteins from those cell samples were separated by two-dimensional electrophoresis (2-DE), i.e., first according to their isoelectric point in an isoelectric focusing step and then according to their relative mass, in a range of 14-150 kDa, using gradient sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE). In this way, thousands of proteins can be analysed at the same time, under the same experimental conditions.

After protein visualization by silver-staining, the digitalized images of the 2-DE gels were analysed using a specific software package (ImageMaster<sup>TM</sup> 2D Platinum, Geneva Bioinformatics SA/ Amersham Biosciences/ GeneBio, Geneva, Switzerland). This allows detection of all protein spots present in each gel, spot-to-spot matching among all the 2-DE gels under

study and spot quantification. Statistical analysis can be performed on groups of protein spots for which up- or down-regulation is observed. Moreover, with this bioinformatic tool which has powerful algorithms specifically developed for the analysis of 2DE-gels, we were able to overcome some reproducibility problems inherent in this type of experiment (namely distortions, staining problems, etc). Indeed, in this study the presence of mucus, especially in CF nasal cell samples, made the separation and visualization of the proteins much more difficult. Therefore, in order to standardize the intensities of silver-staining among spots present in the several 2-DE gels (from CF and non-CF nasal cell samples), analysis was carried out by taking into account the standardized relative intensity volume of spots (or %Vol, i.e., the volume of each spot divided by the volume of all spots on the gel). Furthermore, through heuristic clustering analysis, i.e., an artificial intelligence-based analysis to automatically classify sets of gels into different classes according to their characteristic spots, the 2-DE gels were automatically (and correctly) separated into two distinct classes (CF and non-CF). For a given protein, the difference in expression levels between these two classes of gels was statistically assessed by using the two-sample  $t$  test, for  $n$  observations, where  $n$  is the total number of individuals analysed (CF + non-CF). Differences were considered statistically significant for  $p < 0.05$ .

Finally, all spots of interest were excised from the 2-DE gels and hydrolysed with trypsin enzyme. The determination of the mass of all peptides resulting from this trypsin digestion, i.e. peptide-mass fingerprinting (PMF) analysis, was performed using matrix-laser desorption/ionization - time of flight (MALDI-TOF) spectrometry. These peptide mass spectra were used to search for homologies and protein identification in the mass spectrometry protein sequence DataBase (MSDB) <http://csc-fserve.hh.med.ic.ac.uk/msdb.html>.

From the non-CF 2DE-gels, 65 spots were identified in this way and a reference 2DE-map was thus established. Comparison of this protein profile with the one similarly obtained for CF nasal cells, revealed a set of differentially expressed proteins ( $p < 0.05$ ) that could be related to CF symptomatology. These included proteins related to chronic inflammation, and some others involved in oxidative stress injury. Alterations were also observed in the levels of cytoskeletal proteins, which may be implicated in changes to cytoskeletal organization which have been described in CF airways. Lower levels were found for some mitochondrial proteins suggesting an altered mitochondrial metabolism in CF. Differential expression was also found for two more enzymes that have not been previously associated with CF.

Further studies will clarify the involvement of such proteins in CF pathophysiology and whether they are good targets for CF therapy.

## 4 Conclusion

Transcriptomic and proteomic approaches such as those described above, aimed at identifying gene products differentially expressed in association with a given pathology are just the first step towards understanding the pathways that are putatively associated with the respective disease. However, no functional information nor direct relationship with the pathology is established. Elucidation of protein function is thus the next post-genomic challenge towards the understanding of biological processes in health and disease. Strategies and tools are thus critically needed for distinguishing genes and proteins with mere pathologic association from those primarily responsible for the basic cellular defect(s) in such pathologies in order to establish causal relationships in a global manner.

## Acknowledgements

Work was supported by the EU (Project # QLK3 / CT / 2001 / 01982 "CF-CHIP"), the FCT (Grants # POCTI / MGI / 47382 / 2002 and # POCTI / MAT / 44082 / 2002), CIGMH (Centro de Investigação em Genética Molecular Humana) and FCT postdoctoral fellowships to LAC (SFRH / BPD / 20730 / 2004) and MRR (SFRH / BPD / 19935 / 2004).

## References

- [1] BECK, S.; PENQUE, D.; GARCIA, S.; GOMES, A.; FARINHA, C. and MATA, L. *et al.* (1999). Cystic fibrosis patients with the 3272-26A<sub>Δ</sub>G mutation have mild disease, leaky alternative mRNA splicing, and CFTR protein at the cell membrane, *Hum Mutat*, **14**, 2, 133–144.
- [2] COLLINS, F.S. (1992). Cystic fibrosis: molecular biology and therapeutic implications, *Science*, **256**, 774–779.
- [3] EBERT, B.L. and GOLUB, T.R. (2004). Genomic Approaches to Haematologic Malignancies, *Blood*, **104**, 923–932.
- [4] LÖNNSTEDT, I. and SPEED, T. (2002). Replicated microarray data. *Statistica Sinica*, **12**, 31–46.
- [5] QUACKENBUSH, J. (2002). Microarray data normalization and transformation, *Nat Genet*, **32**, Suppl: 496–501.
- [6] RATJEN, F. and DORING, G. (2003). Cystic fibrosis, *Lancet*, **361**, 9358, 681–689.

- [7] SCHWIEBERT, E.M.; BENOS, D.J.; EGAN, M.E.; STUTTS, M.J. and GUGGINO, W.B. (1999). CFTR is a conductance regulator as well as a chloride channel, *Physiol Rev*, **79**, 145–166.
- [8] WELSH, M.; TSUI, L.-C.; BOAT, T.F. and BEAUDET, A.L. (1995). *Cystic fibrosis*. In “The Metabolic and Molecular Basis of Inherited Disease” (C.R. Scriver, A.L. Beaudet, W.S. Sly and D. Valle, Eds.), New York: McGraw-Hill, Inc., 3799–3876.



# Estimating gene expression missing data using PLS regression

Lígia P. Brás \*      José C. Menezes †

## Abstract

We present a method for the estimation of missing values (MVs) in DNA microarray data that is based on partial least squares (PLS) regression and involves the reprocess of imputed data and the use of correlations between genes and between arrays in an iterative manner. The method was called alternating PLS imputation (APLSimpute).

The imputation efficiency of APLSimpute was assessed under different conditions (type of data, fraction of data missing and missing structure) by the normalised root mean squared error and the squared Pearson correlation coefficients between actual and estimated values, and compared with that of other imputation methods. Namely, we considered the weighted K-nearest neighbours imputation (KNNimpute), local least squares imputation (LLSimpute), partial least squares imputation (PLSimpute), and Bayesian principal component analysis (BPCA).

For the different proportions of missing data, LLSimpute and BPCA showed the best performance in time-series and mixed data. However, PLS-based methods are preferable when imputing non-time series data. Combining gene-based and array-based correlation in the estimation process by APLSimpute enhances the prediction ability in the presence of non-time series data with high missing rates. Results also suggest that, when dealing with time course data, the PLS-based methods should be further improved and optimised in terms of variable selection, because they are not as capable as LLS imputation or BPCA to take advantage of local similarity structures present in the data.

**Keywords:** Gene expression data, DNA microarray data, missing value estimation.

**AMS Classification:** 92C99.

## 1 Introduction

DNA microarrays allow simultaneously monitoring the mRNA levels of thousands of genes in particular cells or tissues under a variety of conditions [1, 2, 3]. The primary aim of the

---

\*Department of Chemical and Biological Engineering, Technical University of Lisbon, IST, Portugal. E-mail: [ligia.bras@ist.utl.pt](mailto:ligia.bras@ist.utl.pt).

†Department of Chemical and Biological Engineering, Technical University of Lisbon, IST, Portugal. E-mail: [cardoso.menezes@ist.utl.pt](mailto:cardoso.menezes@ist.utl.pt).

various analysis techniques that have been developed for microarray data is at identifying regulatory patterns or similarities in expression under different experimental conditions. In this context, many clustering or class discovery (unsupervised) techniques (e.g. hierarchical clustering, self-organizing maps, principal components analysis), and various supervised learning methods (e.g. support vector machines, classification trees) have been proposed (for a review see [4, 5, 6]).

Microarray data is usually represented by large matrices of gene expression (rows) under different experimental conditions (columns). It is common to encounter missing values (MVs) in such matrices, since imperfections in any of the various steps of the microarray experiment create suspicious values, which are usually discarded and regarded as missing [7]. Due to economic reasons or biological sample availability, repeating the microarray experiment is usually not a feasible option. However, many downstream analysis methods used for microarray data (e.g. classification and model-based clustering techniques) require complete matrices, being affected by the estimates used to replace the MVs. Thus, MV estimation can be regarded as a preprocess step essential to minimally bias the performance of microarray analysis methods. Recently, de Brevern et al. [8] studied the stability of gene clusters of microarray data including or not MVs, defined by various hierarchical clustering algorithms, showing that the MVs (even at a low rate) have important effects on the stability of the gene clusters.

Weighted K-nearest neighbours (KNNimpute) [9], Bayesian principal component analysis (BPCA) method [10], local least squares imputation (LLSimpute) method [11], and partial least squares (PLS) imputation [12] are examples of procedures for gene expression MV estimation that try to use available information to preserve relationships in the entire dataset. In this paper, we propose a new PLS-based method called alternating PLS imputation (APLSimpute). The performance of APLSimpute is compared with that of other MV imputation methods using four publicly available gene expression datasets for various rates of MVs and type of missing structure.

## 2 Methods

The imputation methods considered herein can be classified into three main groups: the cluster-based method (KNNimpute), the Bayesian approach (BPCA), and methods based on regression (LLSimpute, PLSimpute and APLSimpute). In this section we give a brief overview of these imputation methods and propose a new PLS-based method called alternating PLS imputation (APLSimpute).

Using the notation of [12], a particular gene with MVs to be estimated is called *target gene*, whereas the genes with available information for estimating the MVs of the target gene are called *candidate genes*. Each target gene  $i$ ,  $\mathbf{x}_i$ , can be split into an available and missing part, which will be denoted, respectively, by  $\mathbf{x}_i^{\text{obs}}$  and  $\mathbf{x}_i^{\text{miss}}$ .



## 2.1 Cluster-based method

In the weighted  $K$ -nearest neighbour imputation, KNNimpute [9], MVs are imputed by combining the expression levels of  $K$ -nearest genes chosen based on a weighted Euclidean distance. The value for  $K$  must be empirically determined since there are no theoretical standards for selecting its value. For a given target gene  $\mathbf{x}_i$ , the weighted Euclidean distance  $d_{ik}$  between the target gene  $i$  and each candidate gene  $k$  is calculated. Then, the missing entry  $j$  of  $\mathbf{x}_i$  is estimated by the weighted average of the expression values of the  $K$  selected candidate genes in experiment  $j$ . The weight  $w_{ik}$  for the  $k$ th neighbour gene of target gene  $i$  is normalized by the sum of the inverse distances for all  $K$  neighbours (1). The code for KNN imputation method was downloaded from the Helix group at Stanford University (<http://smi-web.stanford.edu/projects/helix/pubs/impute>).

$$w_{ik} = \frac{\frac{1}{d_{ik}}}{\sum_{k=1}^K \frac{1}{d_{ik}}} \quad (1)$$

## 2.2 Bayesian method

Oba et al. [10] proposed an estimation method for MVs that is based on Bayesian principal component analysis (BPCA). BPCA method consists of three steps: principal component (PC) regression, Bayesian estimation, and an expectation-maximization (EM)-like repetitive algorithm. Because of the PC regression step, the method depends on the number of principal axes (eigenvectors),  $K$ . However, Oba et al. [10] have found that BPCA exhibits its best results with  $K = n - 1$  (where  $n$  is the number of samples or arrays), obviating the need to tune the  $K$  value in advance. Further details about BPCA method are given in [10] and the code was available at <http://hawaii.aist-nara.ac.jp/~shige-o/tools/>.

## 2.3 Regression-based methods

The imputation methods based on regression that will be considered herein are the local least squares method (LLSimpute) of Kim et al. [11], and the partial least squared imputation method (PLSimpute) proposed by Nguyen et al. [12]. The main difference between regression-based and cluster-based methods lies in the fact that the latter uses a weighted average to obtain the coefficients (weights) for the linear combination of the nearest neighbours. On the other hand, rather than employing a heuristic rule, regression-based methods use an objective function (least squares) for optimising the weights for combining the candidate genes. Therefore, we expect that regression-based estimates will surpass those from cluster-based methods.

### 2.3.1 LLS imputation

LLSimpute exploits local similarity structures in the data together, representing a target gene as a linear combination of  $K$ -nearest neighbour genes selected using the Euclidean distance as a similarity measure [11]. The method comprises two steps. Given a target gene  $\mathbf{x}_i$ , in the first step, the  $K$ -nearest neighbours of  $\mathbf{x}_i$  are found based on the Euclidean distance by ignoring in each gene the missing positions of the target gene. In the second step, least squares regression is used to optimise the coefficients of the linear combination of the  $K$  similar genes in the non-missing positions of the target gene. Then,  $\mathbf{x}_i^{\text{miss}}$  is estimated by a linear combination of the expression values of the neighbour genes in the target gene missing positions. To take advantage of non-missing entries of neighbouring genes that have MVs, each MV is initially estimated by gene-wise averages. Kim et al. [11] implemented a heuristic method for estimating the parameter  $K$  inside LLSimpute, where different values of  $K$  are used to estimate missing positions artificially generated in the non-missing data matrix. Further details about LLSimpute algorithm can be found in [11], and the method was available in Matlab code at <http://www.cs.umn.edu/~hskim/tools.html>.

### 2.3.2 PLS imputation

PLSimpute applies PLS regression to estimate the missing expression values [12]. PLS belongs to a class of regression models that attempts to find the relationship between explanatory and response variables by assuming that they are generated by a common set of underlying factors [13, 14, 15].

In PLSimpute, the set of candidate genes for target gene  $i$ ,  $\mathbf{X}^{C_i}$ , is constructed by selecting those genes with available values in all the missing entries of the target gene  $i$ . Then,  $\mathbf{X}^{C_i}$  is partitioned according to the available values and missing positions of the target gene into  $\mathbf{X}^{C_i, \text{obs}}$  and  $\mathbf{X}^{C_i, \text{miss}}$ , respectively. Shortly, in the PLS regression context, one seeks to capture the most important mode of covariation between  $\mathbf{X}^{C_i, \text{obs}}$  and  $\mathbf{x}_i^{\text{obs}}$  (training data) subject to orthogonality constraints for the linear combination of the candidate genes. In order to improve the estimates from the PLS algorithm, instead of using all candidate genes, Nguyen et al. [12] uses only those genes with highest sum of squared PLS weights from the PLS fit. Further details about the algorithm can be encountered in [12]. PLS computations require  $\mathbf{X}^{C_i, \text{obs}}$  to be complete, so it is necessary to use a method to perform initial estimates and fill in the missing entries. Following Nguyen et al. [12], PLSimpute was implemented using initial estimates from KNN imputation to form a complete matrix  $\mathbf{X}^{C_i, \text{obs}}$ . The code for PLSimpute was implemented in Matlab using the algorithms available in the Internet page <http://dnguyen.ucdavis.edu/.html/supplemental.html> [12]. For PLS regression, data was mean-centred and scaled to unit variance.

### 2.3.3 APLS imputation

Regarding the data space used for clustering and MV imputation, most of the methods proposed in the literature for handling MVs in microarray gene expression data use only relationships between genes (gene or row space). The use of relationships in the gene space as a basis for estimation is driven by the cellular gene co-regulation and co-expression in functional processes. However, for time course experiments, we may also have relationships in the array space (column space) between adjacent time points. This may also occur when biologically similar samples are used in the array hybridisations.

Considering the idea of combining the estimates obtained based on relationships in the gene space or in the array space, we developed a new imputation method based on PLS regression called alternating partial least squares imputation (APLSimpute). In this method, the estimates obtained from gene-based relationships and array-based relationships are used in an iterative fashion. Given a gene expression matrix  $\mathbf{X}$  with  $N$  missing elements, the outline for APLSimpute is given below:

**Step 0** Initialisation: replace the MVs in  $\mathbf{X}$  by the estimates given by gene averages, obtaining  $\mathbf{X}^{\text{complete}(0)}$ ;

**Step 1** Using  $\mathbf{X}^{\text{complete}(0)}$ , apply the PLS algorithm in the gene space (i.e. using the genes as variables) in order to obtain a vector of estimates ( $\hat{y}^{(1)} = \hat{y}^{\text{gene}}$ ) for the  $N$  missing entries, and update the complete gene expression matrix into  $\mathbf{X}^{\text{complete}(1)}$ ;

**Step 2** For each  $h$ th cycle ( $h = 2, \dots$ ):

- a) Using  $\mathbf{X}^{\text{complete}(h-1)}$ , re-estimate the MVs by PLS regression in the array space ( $\hat{y}^{\text{array}}$ ) and update the complete expression matrix;
- b) Re-estimate the MVs using the PLS algorithm in the gene space, obtaining  $\hat{y}^{(h)} = \hat{y}^{\text{gene}}$  and  $\mathbf{X}^{\text{complete}(h)}$ ;
- c) Determine  $\delta^{(h)} = \sum_{j=1}^N [\hat{y}_j^{(h-1)} - \hat{y}_j^{(h)}]^2$ , the sum of squared differences between the MV estimates.

**Step 3** If  $\delta^{(h)} < \tau$ , stop. Otherwise return to step 2 and iterate until convergence.

The convergence criterion was set to  $\tau = 10^{-3}$ , and in general, it was reached for  $h = 3$ . When compared to PLSimpute, APLSimpute handles the need to obtain a complete matrix,  $\mathbf{X}^{C_i, \text{obs}}$ , for the PLS modelling in a different way. Whereas in PLSimpute the MVs are initially imputed by KNN imputation whenever a target gene is considered, APLSimpute performs a single step of initial imputations by substituting all MVs in  $\mathbf{X}$  by gene averages before the PLS-based imputation. By doing this, the matrix of candidate genes for a given target gene comprises all genes except the latter and it has no missing elements, so  $\mathbf{X}^{C_i, \text{obs}}$  is ready to be

employed in the PLS modelling.

While various schemes have been developed to perform variable selection in PLS regression, we opted to simply apply in APLSimpute the same strategy as that of PLSimpute to restrict the number of candidate genes (i.e. variables) to use in the MVs estimation. So, although variable selection is an important issue in regression analysis, it will not be regarded herein.

### 3 Experimental

In this study, we utilised four publicly available datasets. The first dataset comes from a study of the cell cycle regulated genes in the yeast *Saccharomyces cerevisiae* [16]. It consists of time-series microarray data from a *cdc15*- and *cdc28*-based synchronisation, and will be denoted by TS1. We considered a further dataset (TS2), which only comprises the *cdc28*-based synchronisation data from [16]. The third dataset belongs to a study of gene expression regulated by the calcineurin/Crz1p signalling pathway in *S. cerevisiae* [17]. This dataset can be classified as a mixed experiment, since it comprises both time course and non-time series data, and will be referred to as MIX dataset. The fourth dataset is from a study of human cancer cell lines [18]. This dataset, which can be classified as a non-time series, will be called NTS. All datasets consist of cDNA microarray experiments, and the dimensions of the data matrices before (original dataset) and after removing all genes with MVs (complete dataset) are presented in Table 1, together with the structure of missing entries in the original datasets. Prior to the analysis, data were logarithmically (base 2) transformed (except for the cases where data sets were already downloaded in log2 scale).

	Datasets			
	TS1	TS2	MIX	NTS
$p \times n$ (original dataset)	$6178 \times 41$	$6178 \times 17$	$6166 \times 24$	$9712 \times 64$
$p \times n$ (complete dataset)	$869 \times 41$	$1383 \times 17$	$4380 \times 24$	$6115 \times 64$
Total missing rate (%)	8.3	6.1	3.8	3.9
Missing rate per gene (%)				
< 5%	82.5	22.4	84.8	82.2
5–10%	9.3	66.8	4.4	6.2
10–20%	5.0	9.1	4.3	5.9
20–50%	2.5	0.9	5.2	4.6
$\geq 50\%$	0.7	0.8	1.3	1.1

Table 1: Dimension of the data matrices before and after removing the missing elements, total missing rate and missing rate per gene in the original datasets.

In order to evaluate the accuracy of the imputation methods, we introduced artificial missing entries to a complete (i.e. without MVs) expression matrix. First, the complete gene expression matrix was constructed from the real dataset ( $\mathbf{G}$ ) by discarding the missing elements. Next, the test set  $\mathbf{X}$  was constructed by randomly removing a specific proportion of the entries (1%, 5% and 10%) from the complete expression matrix. Moreover, with the aim of mimicking realistic missing patterns, we constructed test datasets in a different way. We randomly assigned MVs to the elements in the  $p$  rows of the complete matrix by sampling  $p$  rows (genes) of  $\mathbf{G}$ , and using their missing locations. This led to a similar missing structure for the test dataset  $\mathbf{X}$  as that of the original set  $\mathbf{G}$ .

To obtain results unbiased with regard to the portion of the data that is missing, we run five independent rounds of both procedures. The following notation will be used for identifying the type and rate of MVs in the test matrices. E.g., TS1<sup>1%</sup> will denote the TS1 dataset with 1% of missing entries introduced randomly, while TS1<sup>uneq</sup> represents the TS1 dataset with unequally distributed MVs.

The imputation methods were run in Matlab 6.5.0 (The MathWorks Inc., 2002), except KNNimpute, which was run in the free software R (Version 2.0.1, 2004). Different values for the model parameters were tested for KNNimpute, PLSimpute and APLSimpute, and only the results obtained using the values yielding the best prediction performance will be presented herein.

For every dataset, each imputation method was applied to recover the introduced MVs, and the accuracy of the method was evaluated by calculating the error between actual ( $y_j$ ) and imputed values ( $\hat{y}_j$ ) using the normalised root mean squared error (NRMSE):

$$\text{NRMSE} = \frac{\sqrt{\frac{\sum_{j=1}^N (y_j - \hat{y}_j)^2}{N}}}{\sigma_y} . \quad (2)$$

$\sigma_y$  is the standard deviation for the  $N$  real values corresponding to the total missing entries over the whole matrix. The squared Pearson correlation coefficients between actual and imputed values were also calculated.

The bias on the methods, i.e. a consistent under- or overestimation of the true values was also tested using a robust estimator based on a rank test, the Wilcoxon signed rank test [19]. Considering the residuals  $\varepsilon_j = y_j - \hat{y}_j$  associated with a given method, we are interested in testing if the residuals are evenly distributed around zero. The null hypothesis ( $H_0$ ) tested is that negative and positive residuals coming from a given imputation method are equally likely.

The performance of two different imputation methods was compared by comparing their mean squared errors of prediction (MSEP), which is equivalent to a test comparing the variances of two groups of samples. Herein, we used the Levene's test [20], which was proposed as an alternative to using the variance ratio as a test statistic to compare two samples in terms

of the amount of variation that they possess [21]. So, the data values are transformed and subject to an analysis of variance to produce the usual  $F$ -statistic for a test of whether the means vary significantly between the samples. We opted to use the form of the test where the absolute deviation from the sample median (instead of the sample mean) is applied as a transformation, since it provides good robustness against many types of non-normal data and retains good power [21]. In this test,  $H_0$  states that the variance (or MSEP, in our case) is equal across both methods, while the alternative hypothesis ( $H_1$ ) states that the variances are different between the two estimates. For both Wilcoxon signed rank test and Levene’s test, a significance level of 5% was considered, so  $H_0$  was rejected if the obtained p-value was inferior or equal to 0.05.

## 4 Results and discussion

When dealing with MVs from TS experiments (especially TS1), the prediction ability is very much influenced by the rate of missing entries (Figure 1). Concerning the type of missing entries (i.e. uniformly or unequally distributed), results indicate that the structure of missing entries in the test dataset is a relevant aspect. Worse estimates are in general obtained when dealing with data with a pattern of MVs resembling that of the original experiment (even if these ones have similar or higher global missing rates; Figure 1 and Table 1).

For every data type and missing rate and pattern, the prediction performance of the distinct imputation methods was compared using the Levene’s test at a 5% significance level. BPCA and LLSimpute give statistically similar predictions for MIX and NTS data ( $0.09 < \text{p-values} < 0.75$ ). For the imputation of TS2 data, BPCA outperforms LLSimpute (p-values = 0), while the opposite is seen on TS1 data (p-values  $< 6.4 \times 10^{-6}$ ). Both BPCA and LLSimpute give particularly good results when applied to TS data, outperforming KNNimpute (p-value = 0) and PLS-based methods (p-values = 0). This supports the fact that BPCA and LLSimpute were designed to efficiently take advantage of the similarity structures on the data. BPCA method is able to capture useful information by a Bayesian optimisation process, while LLSimpute optimises the neighbouring (local) similarity structures encountered in TS data by least squares in order to obtain the estimates. However, although generating better estimates for TS missing entries, BPCA and LLSimpute are less robust than cluster-based or PLS-based methods to increments of the proportion of MVs in this type of experiments (Figure 1).

In spite of not having an optimisation criterion, KNN-based estimates outperform those from PLS-based methods in TS2 data with more than 1% missing rate (Figure 1; p-values  $< 7 \times 10^{-3}$ ). This suggests that a more stringent procedure should be employed for selecting significant variables (genes or arrays) in PLS regression in order to optimally explore the local similarity structures in the data. However, for the other datasets, PLS-based methods outperform KNNimpute (p-values  $< 1.9 \times 10^{-7}$ ).

For the estimation of NTS missing data, PLSimpute and APLSimpute outperform BPCA (p-values  $< 1.15 \times 10^{-6}$  and  $< 5.07 \times 10^{-9}$ , respectively), LLSimpute (p-values  $< 3.6 \times 10^{-3}$  and  $< 1.5 \times 10^{-4}$ , respectively) and KNNimpute (p-values = 0).

Although the prediction errors obtained with APLSimpute are smaller than those from PLSimpute (Figure 1), the estimation ability of both methods is statistically equivalent ( $0.13 < \text{p-values} < 1.0$ ), except for TS1<sup>10%</sup> data (p-value =  $2.8 \times 10^{-7}$ ) and NTS<sup>10%</sup> data (p-value  $< 1.3 \times 10^{-5}$ ).

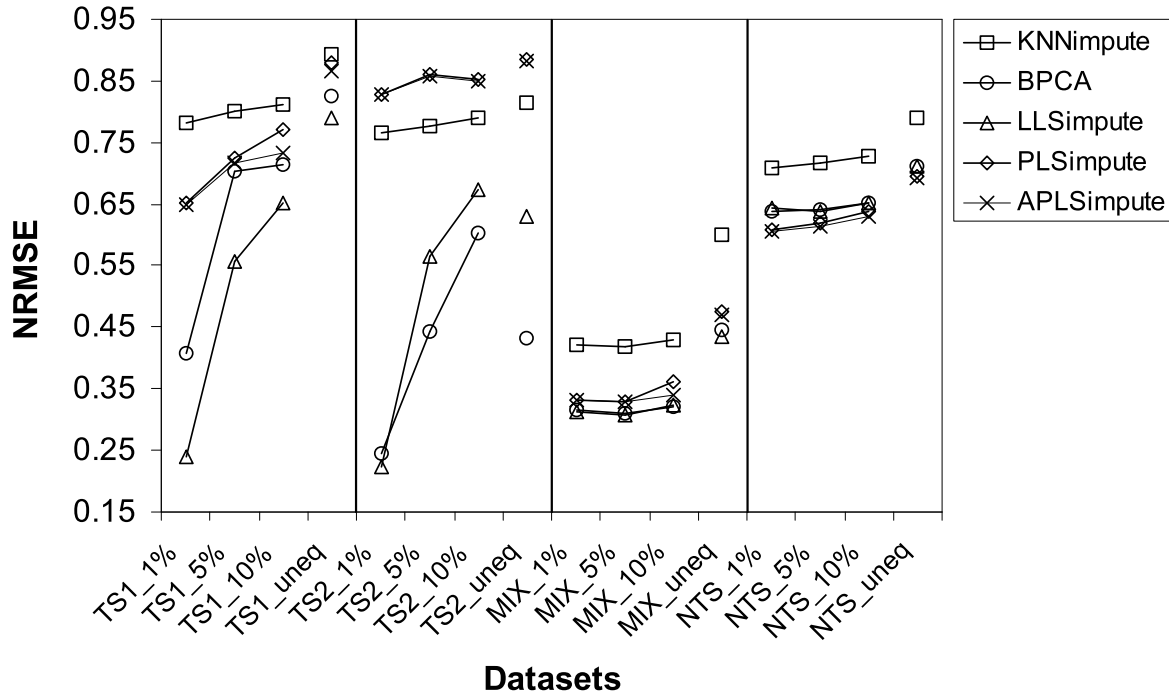


Figure 1: NRMSE for the imputation methods in the different types of datasets.

The methods were further evaluated in terms of bias using the Wilcoxon signed ranks test (data not shown). Results indicate that methods based on PLS regression or in Bayesian analysis are less biased than KNNimpute or LLSimpute. PLS-based imputation methods are particularly preferable in terms of generating unbiased estimates when dealing with NTS data.

## 5 Conclusion

Overall, for the different proportions of missing data, LLSimpute shows the best performance on TS1, while BPCA is preferable for TS2 data. For MIX data, both BPCA and LLSimpute give statistically equivalent estimates. For the imputation of NTS data, PLS-based methods are preferable. In NTS data with 10% total missing rate, the combined use of both gene-based and array-based correlations by the APLSimpute method provides superior prediction performance than PLSimpute, suggesting that the combined utilization of correlation between both genes and arrays can be of value.

Results also suggest that, when dealing with time course data, the PLS-based methods should be further improved and optimised in terms of variable selection, because they are not as capable as LLS imputation or BPCA imputation to take advantage of local similarity structures present in such data.

## Acknowledgements

Lígia P. Brás would like to thank Foundation for Science and Technology in Portugal for financial support (POSI BD/10302/2002).

## References

- [1] LOCKHART, D.J. and WINZELER, E.A. (2000). Genomics, gene expression and DNA arrays, *Nature*, **405**, 827–836.
- [2] SCHULZE, A. and DOWNWARD, J. (2001). Navigating gene expression using microarrays - a technology review, *Nature Cell Biol.*, **3**, E190–E195.
- [3] BARRETT, J.C. and KAWASAKI, E.S. (2003). Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression, *Drug Discov. Today*, **8**, 8, 134–141.
- [4] BRAZMA, A. and VILO, J. (2000). Gene expression data analysis, *FEBS Letters*, **480**, 17–24.
- [5] QUACKENBUSH, J. (2001). Computational analysis of microarray data, *Nat. Rev. Genet.*, **2**, 418–427.
- [6] BUTTE, A. (2002). The use and analysis of microarray data, *Nat. Rev. Drug Discov.*, **1**, 12, 951–960.
- [7] ALIZADEH, A.A.; EISEN, M.B.; DAVIS, R.E.; MA, C.; LOSSOS, I.S.; ROSENWALD, A.; BOLDRICK, J.; SABET, H.; TRAN, T.; YU, X. ET AL. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, **403**, 503–511.



- [8] DE BREVERN, A.G.; HAZOUT, S. and MALPERTUY, A. (2004). Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering, *BMC Bioinformatics*, **5**: 114.
- [9] TROYANSKAYA, O.; CANTOR, M.; SHERLOCK, G.; BROWN, P.; HASTIE, T.; TIBSHIRANI, R.; BOTSTEIN, D. and ALTMAN, R.B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 6, 520–525.
- [10] OBA, S.; SATO, M.A.; TAKEMASA, I.; MONDEN, M.; MATSUBARA, K. and ISHII, S. (2003). A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics*, **19**, 16, 2088–2096.
- [11] KIM, H.; GOLUB, G.H. and PARK, H. (2005). Missing value estimation for DNA microarray expression data: local least squares imputation, *Bioinformatics*, **21**, 187–198.
- [12] NGUYEN, D.V.; WANG, N. and CARROLL, R.J. (2004). Evaluation of missing value estimation for microarray data, *J. Data Sci.*, **2**, 347–370.
- [13] GERLACH, R.W.; KOWALSKY, B.R. and WOLD, H.O.A. (1979). Partial least-squares path modelling with latent variables, *Anal. Chim. Acta*, **112**, 417–421.
- [14] GELADI, P. (1988). Notes on the history and nature of partial least squares (PLS) modelling, *J. Chemometr.*, **2**, 231–246.
- [15] HÖSKULDSSON, A. (1988). PLS regression methods, *J. Chemometr.*, **2**, 211–228.
- [16] SPELLMAN, P.T.; SHERLOCK, G.; ZHANG, M.Q.; IYER, V.R.; ANDERS, K.; EISEN, M.B.; BROWN, P.O.; BOTSTEIN, D. and FUTCHER, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell.*, **9**, 12, 3273–3297.
- [17] YOSHIMOTO, H.; SALTSMAN, K.; GASCH, A.P.; LI, H.X.; OGAWA, N.; BOTSTEIN, D.; BROWN, P.O. and CYERT, M.S. (2002). Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in *Saccharomyces cerevisiae*, *J Biol Chem.*, **277**, 34, 31079–31088.
- [18] ROSS, D.T.; SCHERF, U.; EISEN, M.B.; PEROU, C.M.; REES, C.; SPELLMAN, P.; IYER, V.; JEFFREY, S.S.; VAN DE RIJN, M.; WALTHAM, M. ET AL. (2000). Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genet.*, **24**, 227–235.
- [19] SIEGEL, S. and CASTELLAN, N.J. (1988). *Nonparametric Statistics for Behavioral Sciences*, McGraw-Hill, Inc., 2nd ed., New York, USA.

- [20] LEVENE, H. (1960). *Robust tests for the equality of variance*. In "Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling" (I. Olkin et al., Eds.), Palo Alto, CA, 278–292.
- [21] MANLY, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Chapman and Hall, London, 2nd ed., UK.

# Supervised and unsupervised selection of genes in microarray data

Joaquim F. Pinto da Costa \*    Hugo Alonso †    Luís A.C. Roque ‡  
Manuela M. Oliveira §

## Abstract

In this work we consider the problem of selecting informative genes from the thousands of genes whose expression is usually measured in microarray experiments. Firstly, the selection is done by taking into account the information about the class membership (disease) of each sample; we try to find which of the measured genes have relevant information to discriminate between the different diseases (classes) by using Decision Trees [1]. Surprisingly, in the five datasets analysed, only a few of the thousands of genes were selected; it seems that most of the genes are not good to discriminate between the diseases. Secondly, we approach the problem by finding the Principal Components of the most expressed genes. Two variants are used: the usual Principal Component Analysis (PCA) using the Pearson's correlation matrix and a "weighted" version which is introduced in this work. This "weighted" PCA consists in using an adaptation of a new rank correlation coefficient that gives more importance to higher ranks and which was introduced by Pinto da Costa & Soares in [7].

**Keywords:** Microarrays, Decision Trees, PCA, Weighted Rank Correlation.

## 1 Introduction

This work is concerned with the extraction of informative genes and principal components of genes (known as "eigengenes" [10]) from microarray data. Part of this problem has already been considered by Pinto da Costa & Silva in [6], where the authors used other methods of Data Analysis, mainly Partial Least Squares and Clustering. Here the aim is first of all to see if the genes whose expression is usually measured in microarray experiments have discriminatory information to distinguish between the various sample conditions. Then, a "weighted"

---

\*Dep. de Matemática Aplicada, FCUP, Universidade do Porto, Portugal. E-mail: [jpcosta@fc.up.pt](mailto:jpcosta@fc.up.pt).

†Dep. de Matemática Aplicada, FCUP, Universidade do Porto, Portugal. E-mail: [hugo.alonso@fc.up.pt](mailto:hugo.alonso@fc.up.pt).

‡Dep. de Matemática, ISEP, Instituto Politécnico do Porto, Portugal. E-mail: [lar@isep.ipp.pt](mailto:lar@isep.ipp.pt).

§Dep. de Matemática, Universidade de Évora, Portugal. E-mail: [mmo@uevora.pt](mailto:mmo@uevora.pt).

version of Principal Component Analysis (PCA) is introduced to find some linear combinations of gene expression containing the main information present in the data. Contrarily to the usual PCA, the one introduced by us gives more importance to larger expression values, as we think these are the most important in microarray studies.

The outline of the remaining of the paper is as follows. In the next section we briefly describe the technique of building decision (classification) trees, and focus mainly in the popular method CART [1]. Then, five microarray datasets are analysed with this technique. In section 3 we consider the problem of finding a few linear combinations of gene expression amongst the most expressed genes and that account for most of the variation in the microarray datasets. This is done by the popular technique of PCA which finds new variables, represented by linear combinations of gene expression, which have maximum variance and are uncorrelated. In the first part, the usual PCA using the Pearson's correlation matrix is used; then, a "weighted" PCA version is introduced. This "weighted" PCA gives higher weights to larger expression values. These two versions of PCA are then compared in the five datasets considered. Finally, section 4 presents our work conclusions.

## 2 Selecting genes with decision trees

In this section we use Decision Trees, particularly CART [1], to see which of the thousands of genes whose expression is usually measured in microarray experiments are good predictors of the different diseases (classes). We start by a summary description of this technique.

CART [1] builds regression and classification trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The classic algorithm was popularized by Breiman et al. [1] (see also the work by Ripley in [8]). Besides this algorithm, there are numerous more for predicting continuous variables or categorical variables from a set of continuous predictors and/or categorical factor effects. In many cases, these algorithms specify a linear combination (design) of continuous predictors and categorical factor effects (e.g., with two-way and three-way interaction effects) to predict the response variable. This is the case, for instance, of GLM (General Linear Models), GRM (General Regression Models) and GDA (General Discriminant Function Analysis).

Here, we are interested in the case of predicting a categorical response variable, corresponding to the different classes (diseases) in microarray data, and we do not seek a linear combination of the predictors to that end. For instance, in figure 2 we can see that the (decision) classification tree will determine a set of logical if-then conditions (instead of linear equations) for predicting or classifying cases instead. The interpretation is straightforward: for instance, in the NCI dataset case, given a sample, if the gene 2024 has an expression value greater than 0.18501 and the gene 6277 has an expression value smaller than -0.24 and the gene 44 has an expression value greater than 0.03, then the tree predicts COLON for such a

sample.

Tree classification techniques, producing accurate predictions or predicted classifications based on few logical if-then conditions, have a number of advantages over many of those alternative techniques pointed out before. In most cases, the interpretation of results summarized in a tree is very simple. This simplicity is useful not only for purposes of rapid classification of new observations, but can also often yield a much simpler “model” for explaining why observations are classified or predicted in a particular manner. Additionally, tree techniques are nonlinear and nonparametric because there is no implicit assumption that the underlying relationships between the predictor variables and the dependent variable are linear, follow some specific non-linear link function or that they are even monotonic in nature.

The computational details involved in determining the best split conditions to construct a simple yet useful and informative tree are quite complex and we suggest the reader the book by Breiman et al. [1], which gives an excellent mathematical description of this powerful method. A good general discussion of tree classification and regression methods, and comparisons with other approaches to pattern recognition and neural networks, is also provided by Ripley in [8].

A major issue that arises when applying regression and classification trees to “real” data contaminated by random error noise of high level concerns the decision when to stop splitting tree branches. If not stopped, the tree algorithm will ultimately “extract” all information from the data, and will eventually predict with perfect accuracy the classes of our training set. Of course, it is far from clear whether such (complex) results (with many splits) will replicate in a sample of new observations; most likely they will not. This is because the last splits were obtained with only very few observations and have therefore no generalisation power. In the beginning, this problem was solved by defining a stopping criterion. Many criteria were tested, but the results were never satisfactory because sometimes it seemed that nothing was gained by continuing splitting but, further down the tree, the decrease in error was again significant. This problem was solved in CART [1], by building a very large tree, eventually with only one observation in each leaf, and then prune it backwards until finding the best subtree. Nevertheless, this brought serious computational problems, because the number of subtrees increases exponentially with the size of the tree; again, this problem was elegantly solved by the authors of CART, who have demonstrated how to find the optimal subtree in linear time.

## 2.1 Results

We considered five different datasets containing gene expression from samples (instances) having or not one of possibly various forms of cancer (classes). Obtained from the urls <http://www.lsi.us.es/~aguilar/datasets.html> and <http://www-stat.stanford.edu/~tibs/ElemStatLearn/data.html>, the datasets are identified as follows: Colon cancer (2

classes: 1 - Tumor, 2 - Normal); Embryonal tumours (of the central nervous system) (2 classes: 1 - Tumor, 2 - Normal); Global cancer map (14 classes: 1 - Breast, 2 - Prostate, 3 - Lung, 4 - Colorectal, 5 - Lymphoma, 6 - Bladder, 7 - Melanoma, 8 - Uterus Adeno, 9 - Leukemia, 10 - Renal, 11 - Pancreas, 12 - Ovary, 13 - Mesothelioma, 14 - CNS); Leukemia (2 classes: 1 - ALL, 2 - AML); NCI (14 classes: 1 - CNS, 2 - Renal, 3 - Breast, 4 - NSCLC, 5 - Unknown, 6 - Ovarian, 7 - Melanoma, 8 - Prostate, 9 - Leukemia, 10 - K562B-repro, 11 - K562A-repro, 12 - Colon, 13 - MCF7A-repro, 14 - MCF7D-repro). Figure 1 illustrates the distribution in the training and (whenever available) test sets of the instances among the classes; it can be seen that the most balanced distribution is that related to the Global cancer map dataset. All experiments, whose results are shown below, were carried out using bash-scripting and Matlab 7.0.0.19901 (R14) under Linux.

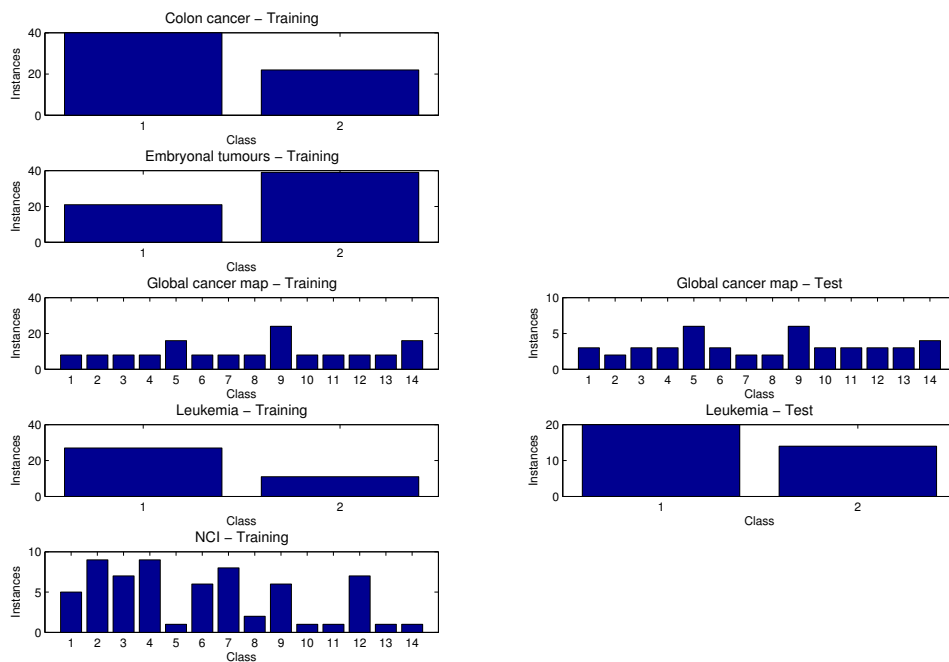
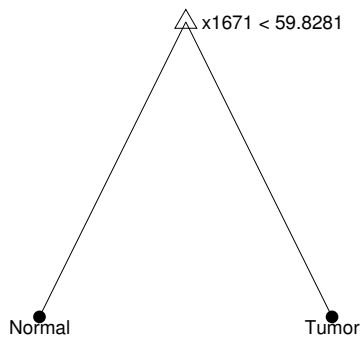
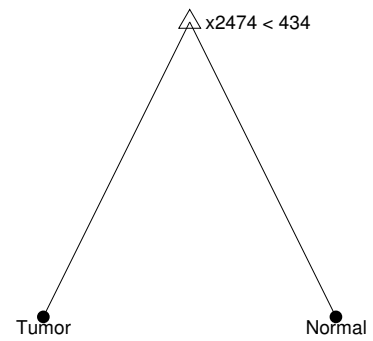


Figure 1: Distribution in the training and (whenever available) test sets of the instances among the classes for the five datasets.

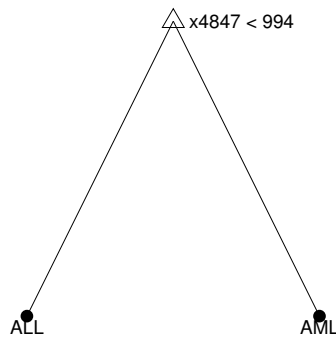
For each of the datasets, decision tree (performance) assessment and selection were done using, respectively, 10-fold cross-validation and the one standard deviation rule.



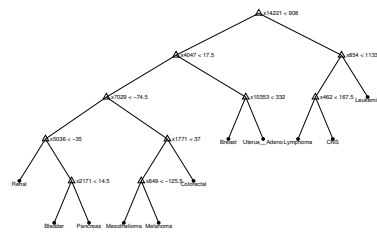
(a) Colon cancer



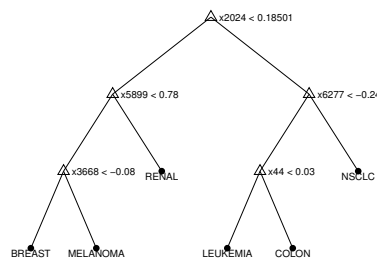
(b) Embryonal tumours



(c) Leukemia



(d) Global cancer map



(e) NCI

Figure 2: Decision trees for the five datasets, where  $x_n$  denotes the expression of the  $n$ -th gene.

Dataset	Available genes	Decision tree					
		Selected genes	Terminal nodes	Classification error			
				Training	10-fold cross-validation	Test	<i>A priori</i>
Colon cancer	2000	1	2	13%	24%		35%
Embryonal tumours	7129	1	2	17%	20%		35%
Global cancer map	16063	10	11	23%	44%	65%	83%
Leukemia	7129	1	2	0%	13%	9%	29%
NCI	6830	5	6	34%	81%		86%

Table 1: Statistics relative to the decision trees for the five datasets.

Figure 2 depicts the chosen trees, and table 1 shows some statistics concerning the number of selected genes for class discrimination, the number of terminal nodes which in this case correspond to the number of identified classes, and the classification error on the training and (whenever available) test sets, where the columns named “10-fold cross-validation” and “*A priori*” contain, respectively, an estimate of the true test error and the error one would get if all instances were labelled with the most representative class in the training set. The results reveal that only a few of the thousands of genes have a relevant expression to discriminate between classes; moreover, in two of the five datasets, some of the classes were not identified by the decision trees. Hence, we conclude that not only unusable measure of gene expression is being made, but also that important genes are not being considered at all.

### 3 Finding eigengenes using weighted and unweighted PCA

In this section we seek a few linear combinations of the genes that account for most of the variation present in microarray data. This is done by using Principal Component Analysis (PCA), introduced by Karl Pearson (1901) and Hotelling (1933) (see [2, 4, 3]). Let us designate by  $\mathbf{X} = (X_1, X_2, \dots, X_g)^T$  a vector containing all expression measurements for the  $g$  genes. Thus, our data consists in  $n$  vectors  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  in a space of  $g$  dimensions, where  $n$  is the number of samples. Mathematically, the PCA problem consists in finding a subspace of the original space which maximizes the dispersion of the points projected onto that subspace. Let  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$  denote the projected points onto the subspace corresponding to the first  $M$  principal components of the sample  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  in the original space. This subspace is the one that maximizes the trace of the total dispersion matrix of the projected points,  $\tilde{S}_T = \sum_{i=1}^n (\mathbf{Y}_i - \hat{m})(\mathbf{Y}_i - \hat{m})^T$ , where  $\hat{m}$  is the mean vector of the projected sample. The solution to this optimisation problem (see [4]) is given by the eigenvectors corresponding to the  $M$  largest eigenvalues of the covariance matrix of the initial sample:



$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \hat{\mu})(\mathbf{X}_i - \hat{\mu})^T$ , where  $\hat{\mu}$  is the mean vector of the original sample. For various reasons, it is common to start by standardising the data, so that all the variables have the same importance in the analysis. In our case this consists in subtracting to each observation the average gene expression and divide by the corresponding standard deviation. With this initial transformation, the principal components obtained are a linear combination of the original gene expression and the coefficients of these linear combinations are given by the components of the eigenvectors of the usual correlation matrix based on Pearson's correlation coefficient  $r$ .

In this work, we introduce a “weighted” version of PCA. This “weighted” version consists in giving more importance to observations whose expression values are higher, which we think makes sense for microarray data, given that the higher the expression value the more probable is that the gene in question is related to the particular disease. To that end, this “weighted” PCA uses a correlation coefficient that gives higher weights to observations that take larger values and that corresponds to a new rank correlation coefficient, introduced by Pinto da Costa & Soares in [7], that gives more importance to higher ranks. We now describe this coefficient based on Spearman's one [9].

Usually, to calculate Spearman's rank correlation coefficient we must first rank the observations in each sample  $x$  and  $z$ , corresponding to the expression of genes  $g$  and  $g'$ , from 1 (highest rank) to  $n$  (lowest rank). We thus obtain  $r(x_i)$  and  $r(z_i)$ , where  $x_i$  and  $z_i$  are the pair of values corresponding to observation  $i$  in each sample and  $r(x_i)$  returns the rank of value  $i$  in the first series. For sake of simplicity, let us use the ranks directly rather than the values in the series, that is,  $R_i = r(x_i)$  and  $Q_i = r(z_i)$ . The Spearman's rank correlation coefficient,  $r_S$ , is given by the expression:

$$r_S = \frac{\sum_{i=1}^n (R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (Q_i - \bar{Q})^2}},$$

where  $\bar{R}$  and  $\bar{Q}$  are the average ranks. However, for computational purposes, a more convenient expression which assumes there are no ties is the following:

$$r_S = 1 - \frac{6 \sum_{i=1}^n (R_i - Q_i)^2}{n^3 - n}.$$

It is clear from this rewritten form of  $r_S$  that the calculation of the distance between two ranks in Spearman's coefficient is given by

$$D_i^2 = (R_i - Q_i)^2,$$

which does not take rank importance into account. In [7], Pinto da Costa & Soares propose the following alternative distance measure:

$$\begin{aligned} W_i^2 &= (R_i - Q_i)^2 \left( (n - R_i + 1) + (n - Q_i + 1) \right) \\ &= D_i^2 \left( (n - R_i + 1) + (n - Q_i + 1) \right). \end{aligned}$$

The first term of this product is  $D_i^2$ , exactly as in Spearman’s coefficient, and represents the distance between  $R_i$  and  $Q_i$ ; the second term represents both the importance of  $R_i$  and  $Q_i$ . Taking this expression as the distance, the authors obtain the weighted rank measure of correlation

$$r_W = 1 - \frac{6 \sum_{i=1}^n (R_i - Q_i)^2 ((n - R_i + 1) + (n - Q_i + 1))}{n^4 + n^3 - n^2 - n},$$

which yields values between +1 and -1. Some properties of the distribution of the statistic  $r_W$ , including its sample distribution, are analysed in [7, 5]: in particular, the expected value of this statistic is zero when the two variables are independent and its sampling distribution converges to the gaussian when the sample size increases. A table of the most significant percentiles is given in [7].

Our aim now is to use the two correlations (Pearson’s  $r$  and “weighted”  $r_W$ ) as inputs for the PCA and compare the results obtained. First of all, because we have many more variables (genes) than observations (samples) in the considered datasets, we will start by filtering the genes that we think are most important. This is done by considering only the most expressed genes. Secondly, we apply the “weighted” and unweighted PCA to the chosen genes and find the new variables, corresponding to the principal components, which are a linear combination of the chosen genes. These principal components have been called “eigengenes” [10]. Then, as suggested in [2], suppose that for instance the first principal component is  $\sum a_i x_i$ , where the  $a_i$  are the coefficients in that component and  $x_i$  is the expression level for gene  $i$ . Restricting attention to those genes for which  $|a_i| > c$ , for some chosen cut-off value  $c$ , allows us to focus on a small set of genes that might be used in a future microarray experiment, for instance.

### 3.1 Results

For each of the datasets, the 15 most expressed genes were used in the usual (Pearson’s  $r$ ) and in the “weighted” ( $r_W$ ) PCA.

Figure 3 shows the cumulative explained variance for a number of principal components, *i.e.*, eigengenes, ranging from 1 to 15, and two major conclusions are drawn: although the eigengenes are different, the cumulative explained variance is almost the same in both approaches to PCA; apart from the Global cancer map dataset, the minimum number of eigengenes needed to explain at least 80% of the total variance is about 5. Table 2 presents the selected genes in the first two eigengenes, ordered from top to bottom by decreasing order of importance, which is measured by the absolute value of the correspondent coefficient in the linear combination representing the eigengene.

## 4 Conclusions

Under the context of supervised selection of genes in microarray data by means of decisions trees, the present work showed that only a few of the thousands of genes whose expression

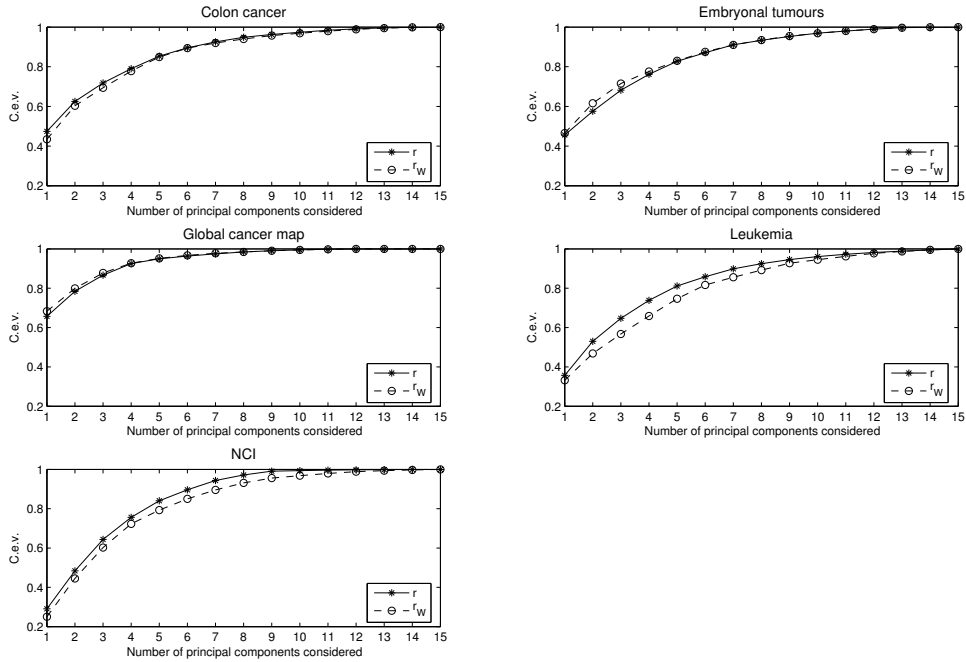


Figure 3: Cumulative explained variance (C.e.v.) for each of the approaches to PCA - usual (Pearson's  $r$ ) and "weighted" ( $r_W$ ) - in the five datasets.

is measured are relevant to discriminate either between the existence or not of a cancer or between different forms of cancer; moreover, in some cases, there was not enough information to carry out such discrimination. Hence, we conclude that not only unusable measure of gene expression is being made, at the expense of high monetary costs let us say, but also that important genes are not being considered at all.

In the unsupervised approach, the usual (Pearson's  $r$ ) and a "weighted" ( $r_W$ ) Principal Component Analysis (PCA) were used to select both the so-called eigengenes, a compressed representation of gene expression through principal components, and the most important genes in some of the most important eigengenes. Most of the times, the two forms of PCA selected different genes or the same genes but in a different order of importance; nevertheless, we observed that the associated eigengenes explain almost the same total variance of gene expression.

## 5 Acknowledgements

The second author would like to thank the Fundação para a Ciência e a Tecnologia under the Third Community Support Framework for the financial support during the course of this

project.

Dataset	Colon cancer				Embryonal tumours				Global cancer map			
Correlation coefficient	$r$		$r_W$		$r$		$r_W$		$r$		$r_W$	
Eigengene	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
	1	2	10	2	7096	1222	5199	5938	44	19	11503	9869
	23	3	23	3	46	5997	6226	47	7184	7158	7184	1765
	9	24	1	878	6773	19	19	45	47	4199	44	1222
	7	10	6	24	5199	5199	5997	7096	7187	1765	7187	5058
	6	878	24	7	19	7096	46	5997	11503	9869	47	4199
Selected genes	10	23	2	26	6226	46	7096	5199	10712		10712	7158
	26	22	3	11	45	45	6773	19	11414		4199	19
	4		9	22	4199		5507	1222	11438		11414	
	2		7		1765		1765		1222		1765	
	3						4199		1765		1222	
	16								4199		11438	
									5058		5058	

Dataset	Leukemia				NCI			
Correlation coefficient	$r$		$r_W$		$r$		$r_W$	
Eigengene	1st	2nd	1st	2nd	1st	2nd	1st	2nd
	1765	4199	4454	4044	6150	6391	6393	6391
	879	1222	1765	879	6393	6392	6152	6392
	896	5998	896	912	6151	6393	6151	256
	1222	6026	4044	5648	6152	6415	6150	6415
Selected genes	4454	912	4482	930	6392	4701	6392	257
	4482	879	4299	4482	6391	6150	6391	6393
	4299		1222	896	5732	6152	3518	6150
	6026		6026	4178	256	6151		6151
				5998	257			4701

Table 2: Selected genes in the first two eigengenes for each of the approaches to PCA - usual (Pearson’s  $r$ ) and “weighted” ( $r_W$ ) - in the five datasets.

## References

- [1] Breiman, L., Friedman, J.H., Olshen, A. and Stone, C.J.. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.
- [2] Ewens, W.J. and Grant, G.R.. *Statistical Methods in Bioinformatics - An Introduction*, 2nd edition. Springer, New York, 2005.
- [3] Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A. and Dudoit, S. (eds). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, 2005.
- [4] Hastie, T., Tibshirani, R. and Friedman, J.. *The Elements of Statistical Learning*. Springer, 2002.
- [5] Pinto da Costa, J.F. and Roque, L.. Limit Distribution for the Weighted Rank Correlation Coefficient,  $r_W$ . (submitted to *REVSTAT - Statistical Journal*) 2005.
- [6] Pinto da Costa, J.F. and Silva, L.. Feature Selection in DNA microarrays. In *ACTES Du X<sup>eme</sup> Congrès de la Société Francophone de Classification (SFC 2003)*, 10-12 September 2003, Neuchâtel, Suisse, pp. 103-108.
- [7] Pinto da Costa, J.F. and Soares, C., 2005. A weighted rank measure of correlation. *Australian & New Zealand Journal of Statistics*, Vol. 47(4), pp. 515-529.
- [8] Ripley, B. D.. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
- [9] Spearman, C.. The proof and measurement of association between two things. *American Journal of Psychology*, Vol. 15, pp. 72-101, 1904.
- [10] Speed, T.P. (ed.). *Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall/CRC Press, London, 2003.



# Gene expression measures of oligonucleotide microarray technology

Fatma Haouari \*

Mohamed Limam †

## Abstract

To search for more optimal methods with better detection of differential gene expression, different standard gene expression measures are explored. Gene expression measure is a process consisting of four main steps: background correction, normalization, Perfect Match correction and summarization. By combining different steps from several performing methods of gene expression measurement, four new measures are proposed. Their performances are assessed based on model fit, differential expression between probe sets and differential expression between RNA samples, using oligonucleotide microarray data from spike-in study and part of a dilution study. MAS method has the best fitting to the model expressing the relation between gene expression values and concentrations since it shows slopes very near to one. The RMA with LiWong summarization (RLW) method shows a good fitting to the model better than Robust Multi-array Average (RMA) method. We notice also that ignoring the background step, in the RMA method, improves the detection of differential expression between RNA samples and between probe sets. Also, we noticed that, in the RMA method, by replacing the quantile normalization method by the contrast one, this gives good results in detecting differential expression between probe sets.

**Keywords:** oligonucleotide microarray, gene expression, differential expression, model fit.

## 1 Introduction

High density oligonucleotide microarray technology, developed by Lockhart et al. ([12]), is a promising biotechnology which measures the expression level of thousands of genes simultaneously in parallel and in a single hybridization assay. It is widely used in many research applications such as: identification of genes that express differently under various experimental conditions, improvement of the process of disease diagnosis, in pharmacogenomics and in toxicogenomics. After the scanning step, a huge amount of gene expression intensities is

---

\*LARODEC, Institut Supérieur de Gestion, Tunis. E-mail: [Fatma.haouari@laposte.net](mailto:Fatma.haouari@laposte.net).

†LARODEC, Institut Supérieur de Gestion, Tunis. E-mail: [Mohamed.limam@isg.rnu.tn](mailto:Mohamed.limam@isg.rnu.tn).

obtained at the probe level. Then, to generate at the gene level, expression measures suitable for analysis, it is necessary to summarize this probe level data. Li et al. ([11]), Naef et al. ([13]), Affymetrix ([1]), Chu et al. ([4]), Irizarry et al. ([9]), Irizarry et al. ([10]), Cope et al. ([5]), Gautier et al. ([7]), Wu et al. ([16]), and others, proposed gene expression methods and evaluated them according to different criteria. Gautier et al. ([7]) showed that gene expression measure is a process with four main steps. The first is background correction consisting in the removal of the optical noise and non-specific hybridization affecting intensity data. The second is normalization which is the process of eliminating, or at least reducing, variation of non-biological origin mentioned by Hartemink et al. ([8]). The third is Perfect Match correction which describes the way of correcting for non-specific binding. The last step is summarization of probe level data into summary expression values. These four modules can work independently of the process, which provides the possibility to remove a step or to replace it by another one. In this paper, we used functions provided by the Affy package implemented by Gautier et al. ([7]) using the R statistical software, introduced by Venables et al. ([15]). We used in addition other R packages to generate four new gene expression measures. Those packages are implemented in the bioconductor project and can be downloaded (<http://www.bioconductor.org>). These new methods consist in combining different steps of the process of gene expression from different methods such as Mas 5.0 of Affymetrix ([1]), Multiplicative Model-based expression index method of Li et al. ([11]) and robust multiarray analysis method (RMA) of Irizarry et al. ([9]) and other normalization methods such as the contrast method of Astrand ([2]) and the quantile method of Bolstad et al. ([3]). These new methods are evaluated based on criteria of special interest in biomedical research such as, model fit and differential expression between RNA samples and differential expression between probe sets using spike-in studies and dilution study. Different gene expression methods are discussed in Section 2. Next, different steps of gene expression methods are combined in Section 3. Then, the new methods are evaluated based on model fit, differential expression between probe sets and differential expression between RNA samples in section 4.

## 2 Methods of gene expression measurement

In this section, we discuss three standard gene expression measures namely: RMA, MBEI and Mas 5.0. Irizarry et al. ([10]) conducted an evaluation study and showed that RMA and MBEI provide more consistent detection of differential expression between RNA samples and probe sets than those presented by Affymetrix ([1]), Mas 5.0 and Mas 4.0. The three methods perform also similarly in terms of accuracy (bias) but with better precision with respect to the RMA measure.



## 2.1 Multiplicative Model-based expression index

MBEI method proposed by Li et al. ([11]), and implemented in their software DNA-Chip Analyzer (dChip), uses a non-linear normalization method to account for differences between intensities across arrays. Then, it fits the following statistical model for one probe set in multiple arrays:

$$y_{ij} = PM_{ij} - MM_{ij} = \theta_i \Phi_j + \epsilon_{ij}, \quad (1)$$

where  $PM_{ij}$  and  $MM_{ij}$  are the intensities of the  $j^{th}$  probe pair in the  $i^{th}$  sample,  $\theta_i$  represents the expression level for a gene in the  $i^{th}$  sample,  $\Phi_j$  is the additional rate of increase for the  $j^{th}$  PM probe pair and a random error  $\epsilon_{ij}$  assumed to be  $N(0, \sigma^2)$ . This model is identifiable only under the constraint  $\sum_j \Phi_j^2 = J$ , where  $J$  is the number of probe pairs in the probe set. The expression measure is computed as follows:

- a) Normalization of probe intensities using the non-linear normalization method of Li et al. ([11]).
- b) Fitting of the model defined in (1) to the normalized data to obtain gene expression measures.

This method is the first gene expression measure based model. And one of his disadvantages is that her model parameters must be estimated from a training data set, and can therefore be less accurate in the case of a small or heterogeneous data set.

## 2.2 Mas 5.0 signal method

Mas 5.0 signal method is the latest version of Affymetrix ([1]) algorithms for gene expression measurement. It consists in computing a signal value, a gene expression measure, from the combined background-adjustment, both PM and mismatch (MM) values of the probe set. And by using Tukey biweight algorithm we compute a robust average of the values  $\log(PM_{ij} - IM_{ij})$ , where  $PM_{ij}$  represents PM intensities of the  $j^{th}$  probe pair in the corresponding  $i^{th}$  probe set, and  $IM_{ij}$  is the ideal mismatch of the  $j^{th}$  probe pair in the corresponding  $i^{th}$  probe set. The signal value is calculated through the following steps:

- a) Cell intensities are preprocessed for global background.
- b) An ideal mismatch value is calculated and subtracted to adjust the PM intensity.
- c) The adjusted PM intensities are log-transformed to stabilize the variance.
- d) The biweight estimator is used to provide a robust mean of the resulting values.
- e) Finally, signal is scaled using a trimmed mean.

In contrast to previous methods based on PM - MM differences, Affymetrix uses the IM quantity, never bigger than PM, to correct for non-specific binding. Due to this alternative, this Affymetrix method has become more accurate than all other gene expression measures.

### 2.3 Robust Multi-array Average

RMA method proposed by Irizarry et al. ([9]) uses the average of log transformed, background subtracted PM values for gene expression calculation and fits a robust linear model, by applying a median polish technique, described by Venables et al. ([15]), on this model,

$$y_{ij} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}, \quad (2)$$

where  $y_{ijn}$  is the background adjusted, normalized, and log-transformed PM intensities,  $\mu_{in}$  is the log expression level for array  $i$  on the probe set  $n$ ,  $\alpha_{jn}$  is the probe-effect on probe set  $n$  and  $j$  represents probes in one probe set,  $j = 1, \dots, J$ ,  $\varepsilon_{ijn}$  is an independent identically distributed error term with mean 0. For the identifiability of the parameters it is assumed that  $\sum_j \alpha_j = 0$  for each probe set. The expression measure is computed through the following steps:

- a) Background correction of the PM probes using a model based on observed intensity, assumed to be the sum of signal and noise.
- b) Normalization of corrected PM probes using quantile normalization.
- c) Making a log transformation to obtain the additive model defined in (2).
- d) Fitting of the additive model using median polish technique to calculate the gene expression measure.

As shown by Irizarry et al. ([10]), RMA gene expression method is considered a standard since it gives the most precise gene expression measures and it is the best in detecting differential expression between probe sets and samples.

## 3 Combination and generation of gene expression measures

In this section, different methods used for the generation of the new gene expression measures are discussed and different combined methods and the way of producing them are presented.

### 3.1 Background correction

This step of the process is not implemented in many gene expression measures. The widely used background correction methods are those given by Affymetrix ([1]), which is the procedure mentioned in this paper in section 3.3.1, and the one given by Irizarry et al. ([9]) in their RMA method.

### 3.1.1 The RMA background correction.

Irizarry et al. ([9]) ignore the MM probes and assume that PM intensities are composed of signal and background signal caused by noise and non-specific binding. The background is denoted by  $bg_{ijn}$ , where  $i$  represents arrays,  $i = 1, \dots, I$ ,  $j$  represents probes in one probe set,  $j = 1, \dots, J$ , and  $n$  represents probe sets,  $n = 1, \dots, N$  and it follows a normal distribution with parameters  $\mu$  and  $\sigma$ . Background corrected values are estimated as follows,

$$B(PM_{ijn}) \equiv E(s_{ijn}|PM_{ijn}), \quad (3)$$

where

$$E(s_{ijn}|PM_{ijn} = s) = a + b \frac{\phi(\frac{a}{b}) - \phi(\frac{s-a}{b})}{\Phi(\frac{a}{b}) + \Phi(\frac{s-a}{b}) - 1}, \quad (4)$$

with  $PM_{ijn}$  denotes the PM intensities,  $s_{ijn}$  stands for signal,  $a = s - \mu - \sigma^2\alpha$  and  $b = \sigma$ .

## 3.2 Normalization

Many normalization methods are presented in the literature. Schadt et al. ([14]) and Li et al. ([11]), proposed a non-linear baseline method based on non-linear relations. Astrand ([2]) proposed a contrast based method consisting in fitting a smooth curve to the feature intensities at the log scale. Bolstad et al. ([3]) introduced cyclic loess method based on M vs A method of Dudoit et al. ([6]). Bolstad et al. ([3]) proposed also the quantile method consisting in transforming the distribution of probe intensities in a set of arrays the same. Bolstad et al. ([3]) compared these several methods and showed that the quantile method reduces significantly bias and variance as compared to other methods, and that it is the most time consuming.

### 3.2.1 Contrast based method

Astrand ([2]) introduces the contrast based method as a non-linear procedure for normalizing the raw feature intensities, i.e. the PM and MM intensities. As the cyclic loess method, it is another extension of the M versus A method which fits a smooth curve in scatter plots, with the log feature intensity differences on the y-axis, and the log intensity means on the x-axis. This method is faster than the cyclic loess method, however, the computation of the loess smoothers is time consuming.

### 3.2.2 Quantile method.

The principle of this method, as presented by Bolstad et al. ([3]), is to make the distribution of probe intensities in a set of arrays the same. The method is motivated by the idea that we could give to two separate data sets the same distribution by transforming the quantiles of each to have the same value. This could be done by projecting probe intensities of the set of

arrays onto the unit vector  $(\frac{1}{\sqrt{N}}, \dots, \frac{1}{\sqrt{N}})$  to get normalized data. The projection is equivalent to taking the average of the quantile in a particular row of a matrix, having arrays as columns and intensities as rows. By substituting this value, for each of the individual elements in that row, we have

$$proj_d q_i = (\frac{1}{N} \sum_{j=1}^N q_{ij}, \dots, \frac{1}{N} \sum_{j=1}^N q_{ij}), \quad (5)$$

where  $d$  is the unit vector mentioned above,  $q_{ij}$  is the row of probe  $i$  in column  $j$  and  $N$  denotes the number of arrays. The quantile based method provides a fast procedure to normalize multiple chips based on the assumption of a common distribution.

### 3.3 PM correction

This step cannot be independent of the process of gene expression measurement, since it is a way of correcting the PM values which are implemented in each gene expression method. Except the Mas 5.0 method, which adopts a PM correction method based on alternatives and parameters, the majority of gene expression measures correct PM intensities by simply ignoring the MM intensities or subtract the MM quantity from the PM quantity.

#### 3.3.1 Mas 5.0 PM correction method

As reported in Affymetrix ([1]), and for many reasons, the MM value can be larger than PM value. In this case, it can't be an estimator for the amount of signal in the PM intensity. Instead, an idealized value (IM) can be estimated based either on the average ratio between PM and MM, or a value slightly smaller than PM. Three cases for determining the ideal mismatch, for probe pair  $j$  in probe set  $i$ , can be distinguished depending on parameters:

$$IM_{i,j} = \begin{cases} MM_{i,j}, & \text{if } MM_{i,j} < PM_{i,j} \\ \frac{PM_{i,j}}{2^{(SB_i)}}, & \text{if } MM_{i,j} \geq PM_{i,j} \text{ and } SB_i > contrast\tau \\ PM_{i,j}, & \text{if } MM_{i,j} \geq PM_{i,j} \text{ and } SB_i \leq contrast\tau \end{cases} \quad (6)$$

where contrast  $\tau = 0.03$ , scale  $\tau = 10$  and  $SB_i$  represents the value of the specific background ratio for the probe set which is calculated as follows:

$$SB_i = T_{bi}(\log_2(PM_{i,j}) - \log_2(MM_{i,j})), \quad (7)$$

where  $j$  represents the probe pairs,  $j = 1, \dots, n_i$  and  $i$  the respective probe set and  $T_{bi}$  is the one-step Tukey's Biweight algorithm. The first case, where the mismatch value provides a probe-specific estimate of stray signal is the best situation. In the second case, the estimate is not probe-specific, but at least it provides information specific to the probe set. The third case involves the least informative estimate, and it is weakly based only on probe-set specific data.

### 3.4 Summarization

#### 3.4.1 Median Polish Technique

Venables et al. ([15]) present the median polish technique as a robust data analysis technique used to test the significance of various factors in a multifactor model. It models the response variable as:  $\text{data} = \text{common value} + \text{factor-1 effects} + \dots + \text{factor-n effects} + \text{residual}$ . For each factor, we do the following:

- a) Calculate the medians for each level of a factor (these are the factor effects).
- b) For each value of the response variable, subtract the corresponding level median.

The above steps are repeated until the ratio of the sum of the residuals from the current step and the residuals from the previous step are less than some cut-off value (normally only 2 or 3 passes are required).

### 3.5 Combination

As shown by Irizarry et al. ([10]), RMA gene expression method is considered a standard since it is the best in detecting differential expression between probe sets and samples, and it gives the most precise gene expression measures. Different steps of the RMA method are interchanged with the different steps of the other mentioned methods. By this way, we generate four other gene expression measures mentioned in table 1. Wu et al. ([16]) adopted this procedure where they replaced only the background correction step of the RMA method by a GC-background method that they implemented in their PM-only GC-RMA method. The modification of the background step gave more accurate gene expression measure.

Method	Background correction	Normalization	PM correction	Summarization
RMA	rma	quantile	pmonly	median polish
RWB	-	quantile	pmonly	median polish
RIM	rma	quantile	mas	median polish
Rcontras	rma	contrast	pmonly	median polish
RLW	rma	quantile	pmonly	LiWong

Table 1: Different combinations for proposed gene expression methods.

### 3.5.1 RMA without background correction (RWB)

Expression measures are computed using RMA method but by omitting the background correction step. This method will allow us to see the effect of not removing noise and non-specific binding from data.

### 3.5.2 RMA using the ideal IM method as PM correction (RIM)

Expression measures are computed by replacing the PM correction method mentioned in the RMA method by the one implemented in the MAS 5.0 expression measure.

### 3.5.3 RMA with contrast normalization method (Rcontras)

In this method, we ignore the same distribution feature of the quantile method to replace it with the smooth curve of the contrast normalization method. So, we take the PM probes in pairs to compute the differences and means of the MVA plots on which is based the contrast based method. We chose this normalization method because it proved her performance in reducing variance and bias as reported by Bolstad et al. ([3]).

### 3.5.4 RMA with LiWong summarization (RLW)

In this method, we applied on the PM corrected intensities the multiplicative model of the summarization step followed by MBEI method of Li et al. ([11]) to summarize the probe level data.

## 4 Results

In this section, we evaluate the four proposed methods based on model fit and differential expression between RNA samples and probe sets using the dilution study and the spike-in study. For the second criterion, we use the observed fold change in expression which is usually used to assess differential expression between replicate arrays.

### 4.1 Data description

Evaluation is based on data used by Irizarry et al. ([9]), Bolstad et al. ([3]) and others. It is a benchmark, composed of two sets of experiments, a dilution/mixture experiment for liver tissue and central nervous system (CNS) samples and a spike-in experiment. The two studies are available from the web at <http://qolotus02.genelogic.com/datasets.nsf/>.

- 1) The dilution/mixture data

The dilution/mixture data series consist of 75 HG-U95A arrays, where two sources of RNA, liver and central nervous system cell line are investigated. There are 30 arrays for each source, broken into 6 groups at 5 dilution levels. The remaining 15 arrays, split into 3 groups of 5 chips, involve mixtures of the two tissue lines in the following proportions: 75:25, 50:50, and 25:75.

2) Spike-in study:

The spike-in data sets consist of experiments where 11 different cRNA fragments were added to the hybridization mixture of 98 HG-U95A GeneChip arrays at different picomolar (pM) concentrations. The 11 control cRNAs were BioB-5, BioB-M, BioB-3, BioC-5, BioC-3, BioDn-5 (all *E. coli*), CreX-5, CreX-3 (phage P1), and DapX-5, DapX-M, DapX-3 (*B. subtilis*).

3) The varying concentration series data set, B1

For an individual array, all of the 11 control cRNAs were spiked-in at the same concentration and this concentration is varied across arrays, taking the values 0.0, 0.5, 0.75, 1, 1.5, 2, 3, 5, 12.5, 25, 50, 75, 100 and 150 pM.

4) The spike-in data sets B2

This data set consists of a 12 x 12 cyclic Latin square, with each concentration appearing once in each row and once in each column. The 12 concentrations used are 0.5, 1, 1.5, 2, 3, 12.5, 25, 37.5, 50, 75, and 100 pm. The 12 combinations of concentrations used on the arrays are taken from the first 11 entries of the 12 rows of this Latin square. Of the 12 combinations used, 11 are conducted on three arrays and one on just two arrays.

## 4.2 Model fit Comparison

In order to evaluate the gene expression measures in terms of model fit, as in Bolstad et al. ([3]), we look to this criterion in the context of the spike-in series. In particular we use the data set B1 for this criterion just described. We compute the gene expression measures for each of the 26 arrays as a group using each of the gene expression methods mentioned in this paper. To the spike-in probe sets, we fit the following linear model in the log scale,

$$\text{Log}_2 E = \beta_0 + \beta_1 \text{Log}_2 c + \epsilon \tag{8}$$

where E are the values of the expression measures, c are the concentrations,  $\beta_0$  is a constant,  $\beta_1$  is the slope estimate, and  $\epsilon$  is a random error. To guarantee a good fit to the model, we eliminate the array with spike-in concentration 0 because of the log use. The ideal result would be to have slopes that are near one. Table 2 shows the slope estimates for each of the 11 spike-in probe sets, relatively to the 11 control cRNA chosen in this data set.

Probe set	MAS	MBEI	RMA	RIM	back	Rcontras	RLW
BioB-5_at	0.976	0.842	0.339	0.868	0.752	0,081	0,917
DapX-M_at	0.834	0.792	0.829	0.795	0.725	0,052	0,749
DapX-5_at	0.889	0.695	0.813	0.737	0.608	0,127	0,815
CreX-5_at	0.982	0.870	0.962	0.901	0.776	0,050	0,998
BioB-3_at	1.095	0.829	0.900	0.842	0.737	0,092	0,941
BioB-M_at	1.157	0.780	0.799	0.892	0.696	0,142	0,873
BioDn-3_at	0.781	0.594	0.591	0.583	0.559	0,084	0,652
BioC-5_at	0.970	0.826	0.856	0.828	0.751	0,073	0,858
BioC-3_at	1.304	0.773	0.860	0.803	0.677	0,103	0,884
DapX-3_at	0.891	0.820	0.863	0.873	0.748	0,063	0,864
CreX-3_at	0.291	0.029	0.046	0.043	0.021	-0,042	0,086

Table 2: Regression slope estimates for each of the 11 spike-in probe sets.

If we exclude the CreX-3\_at probe set from the rest of the cited spike-in probe sets, the RMA, MAS and RLW methods give slopes closer to one and the MAS method shows the best results. It indicates a good relationship between gene expression values and concentrations. For the rest of the methods, we observe a downward bias in the slope estimates. The average  $R^2$  for the spike-in probe sets, excluding CreX-3, are 0.882 for the MAS method, 0.954 for the MBEI method, 0.871 for the RMA method, 0.953 for the RIM method, 0.953 for the RWB method, 0.013 for the Rcontras and 0.885 for the RLW method. And the median standard errors of each gene expression method are 0.055 for the MAS method, 0.033 for the MBEI method, 0.040 for the RMA method, 0.035 for the RIM method, 0.029 for the RWB method 0.212 for the Rcontras and 0.043 for the RLW method. We notice that between all the gene expression methods the RMA, MAS and RLW method have high slopes, a better fitting model and more precise slope estimates.

### 4.3 Comparison of differential expression between RNA samples

To achieve this comparison, we refer to the analysis mentioned in Cope et al. ([5]). In optimal conditions, the Affymetrix protocol calls for 15  $\mu\text{g}$  of RNA, but in practice the amount of target mRNA available for the hybridization reactions can be amplified or decreased. It depends on the cells, the tissue type under study and from experiment to experiment. Expression ratio estimates (fold changes) are relative and should not co-vary with RNA quantity. To simulate extreme variation in total quantity of RNA, we use a data set from the dilution study, where the lowest concentration at 1.25  $\mu\text{g}$  and the highest at 20  $\mu\text{g}$  are used. For each



gene, observed expressions are first averaged across replicates to obtain a single mean value for each sample and dilution. Then, for each gene, we computed log fold change estimates between liver and CNS samples using the 10 arrays in the 1.25  $\mu\text{g}$  concentration group with each of the five expression measures. After that, we computed estimates using the arrays in the 20  $\mu\text{g}$  concentration group. Log base 2 fold change estimates of gene expression between liver and CNS samples, computed from arrays hybridized to 1.25  $\mu\text{g}$  of cRNA, were plotted against the same estimates obtained from arrays hybridized to 20  $\mu\text{g}$  for all five measures. The ideal result would be to have a curve fit with one as coefficient correlation. The plot resulting from applying RMA method is presented in figure 1 in order to make a clear comparison to this method, taken as reference.

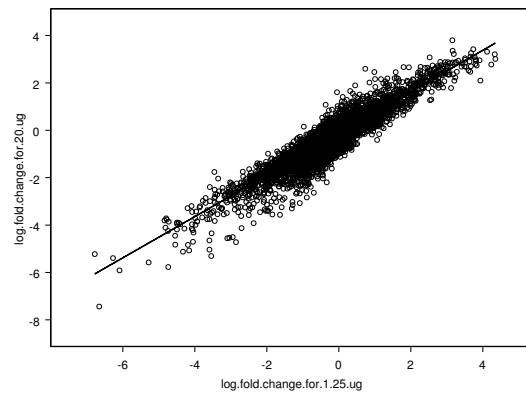


Figure 1: Plot of RMA method.

Figure 1 shows a good linear fit relation between the distributions in the two samples, which is justified by the correlation coefficient given in table 2.

Method	RMA	RWB	RIM	Rcontras	RLW
<b>Correlation coef</b>	0,912	0,932	0,753	-0,235	0,422

Table 3: The correlation coefficients of the five methods.

The correlation coefficient of fold change estimates from the different concentrations, given in Table 2, shows that RMA and RWB provide more consistent estimates than the other methods. Also, we notice that RWB performs better than RMA as it is shown from figure 2 and Table 2. The figures and tables results that the noise corrected by the RMA background may be informative. In effect, these results are expected since it is shown in Cope et al. ([5]),

Naef et al. ([13]), and others, that MM probes may also detect signal. The situation is also similar here and the modification performed on data by applying RMA background correction may lead to the deformation of results. The worst result is noticed in figure 4 relative to the RMA with contrast normalization method. The methods using the two feature probes PM and MM detect better differential expression between RNA samples. The RLW, figure 5, shows many genes with n-fold discrepancy. The plot of figure 3 shows a result better than the Rcontras and RLW methods.

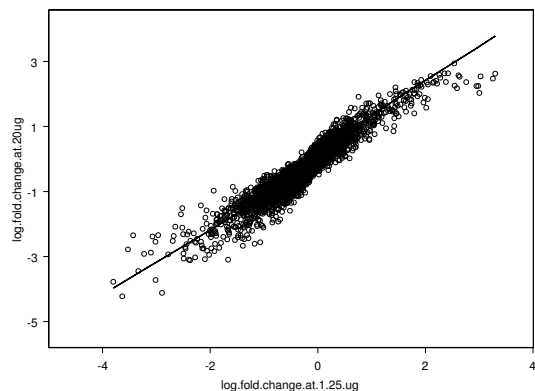


Figure 2: Plot of RWB method.

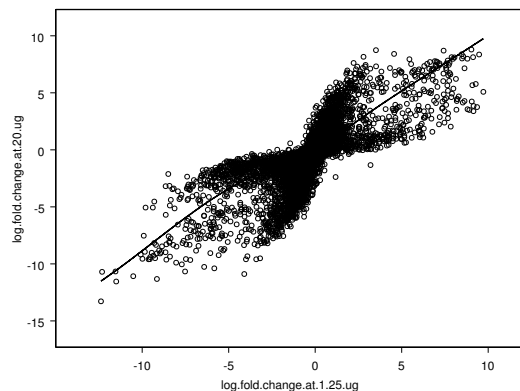


Figure 3: Plot of RIM method.

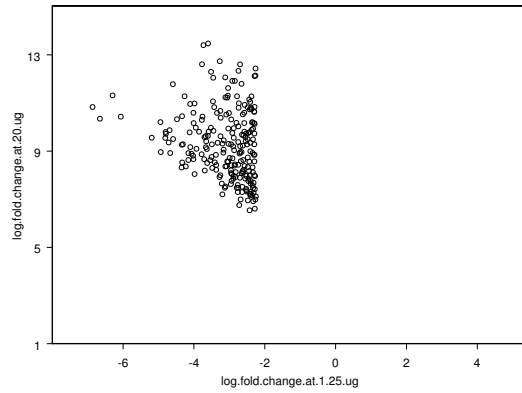


Figure 4: Plot of Rcontras method.

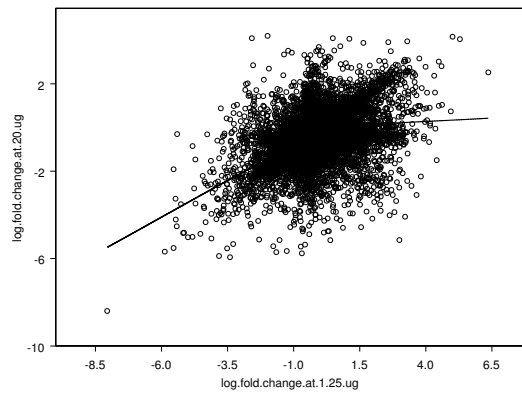


Figure 5: Plot of RLW method.

#### 4.4 Comparison of differential expression between probe sets

We now compare the performance of the five methods in detecting differential expression between probe sets using the data set from the spike in study. Each probe set is expressed at a different concentration between the two triplicates, this allow us to compare differential expressions at the probe set level. We compute expression measures for each of the six arrays using each of our methods. For each of the six arrays studied, expression measures  $E_{11n}$ ,

$E_{12n}, E_{13n}, E_{21n}, E_{22n}, E_{23n}$  were obtained in their respective scale, log for the four first methods and natural for the RLW method, for each probe set  $n = 1, \dots, N$ . We then compute the averages  $E_{i-n} = (1/3)\sum_{k=1}^3 E_{ikn}$ ,  $i = 1, 2, n = 1, \dots, N$ . After that, for the probe sets representing spike-in RNAs, we compute fold changes as the ratio of the average expressions. The following figures show MVA plots of the average expressions values versus fold change values. A great deal of information about the distribution of observed fold changes can be read from such plot. The ideal is to have high absolute differential expressions and the method can distinguish them explicitly. That is to say we can see points drawn separately from the agglomeration of points in the plot. From the first plot, figure 6, relative to the RMA method and the figure 7 and figure 9, we see probe sets drawn distinctly out of the cloud of points, and they are differentiated by their large absolute fold changes. Those differentially expressed probe sets are the 10 out of 11 genes differentially expressed relative to the probe sets mentioned in Table 4. Figure 8 and figure 10, relatively to the RIM and RLW methods, show a large cloud. It didn't distinguish the appropriate probe sets differentially expressed. In the plots 8 and 10 the probe sets with high differential expressions are not actually spike-in ones. We notice a mixture of points, which is difficult to differentiate. Therefore, the method with smaller cloud of points spread, RMA and RWB methods, are the best in terms of detection of differentially expressed probe sets. The RWB method, by reference to RMA method, turns out to be the best because it has less spread cloud points than the other. The plot related to the RMA with contrast normalization method shows good results for this comparison criterion and demonstrates a good detection capability of differentially expressed probe sets. This is an expected result since the contrast normalization method is based on MV A plots as is mentioned earlier.

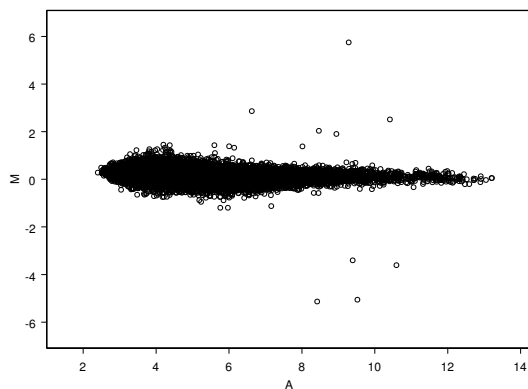


Figure 6: Plot of the RMA method.

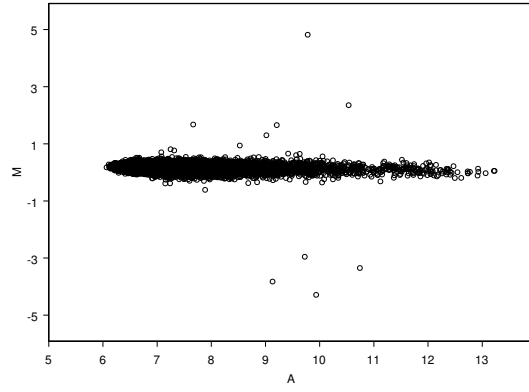


Figure 7: Plot of the RWB method.

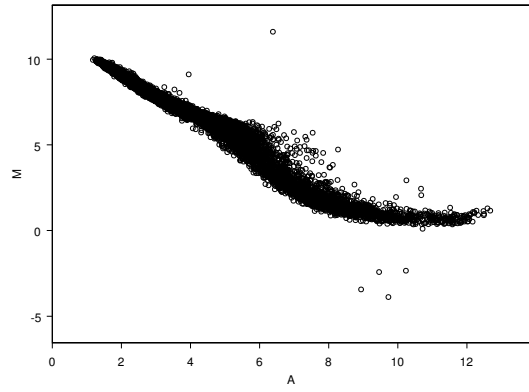


Figure 8: Plot of the MAS method.

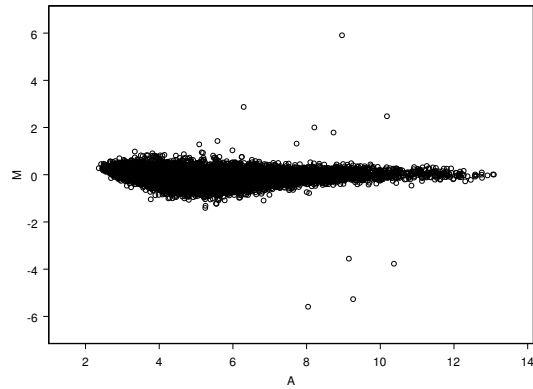


Figure 9: Plot of the Rcontras method.

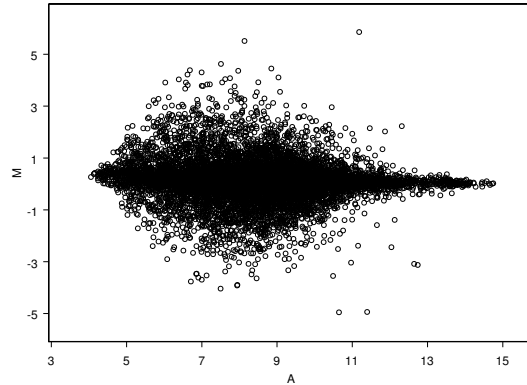


Figure 10: Plot of RLW method.

## Acknowledgments

Authors would thank GeneLogic Corporation for providing them datasets to assess results.

## References

- [1] AFFYMETRIX. (2002). Statistical Algorithms Description Document, *Technical report*, Affymetrix Inc, Santa Clara, CA.
- [2] ASTRAND, M. (2003). Contrast normalization of oligonucleotide arrays, *Journal of Computational Biology*, **10**, 95–102.

Probe set	Concentration	
	Triplicates1	Triplicates2
BioB-5	100.0	0.5
BioB-3	0.5	25.0
BioC-5	2.0	75.0
BioB-M	1.0	37.5
BioDn-3	1.5	50.0
DapX-3	35.7	3.0
CreX-5	12.5	2.0
BioC-3	25.0	100.0
DapX-5	5.0	1.5
DapX-M	3.0	1.0

Table 4: Concentrations of each spike-in gene in two sets of triplicates.

- [3] BOLSTAD, B. M., IRIZARRY, R. A., ASTRAND, M. AND SPEED, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, **19**, 185–193.
- [4] CHU, T., WEIR, B. AND WOLFINGER, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments, *Math.Biosci*, **176**, 35–51.
- [5] COPE, L. M., IRIZARRY, R. A., JAFFEE, H. A., WU, Z. AND SPEED, T. P. (2003). A Benchmark for Affymetrix GeneChip Expression Measures, *Bioinformatics*, **19**, 1–10.
- [6] DUDOIT, S., YANG, Y. H., CALLOW, M. J. AND SPEED, T. P. (2002). Statistical methods for identifying differential expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, **12**, 111–139.
- [7] GAUTIER, L., COPE, L., BOLSTAD, B. M. AND IRIZARRY, R. A. (2004). Affy-analysis of Affymetrix GeneChip data at the probe level, *Bioinformatics*, **20**, 307–315.
- [8] HARTEMINK, A. J., GIFFORD, D. K., JAAKKOLA, T. S. AND YOUNG, R. A. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization, *Bioinformatics*, **20**, 307–315.
- [9] IRIZARRY, R. A., HOBBS, B., COLLIN, F. AND SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Bio-statistics*, **4**, 249–264.
- [10] IRIZARRY, R. A., BOLSTAD, B. M., COLLIN, F., COPE, L. M., HOBBS, B. AND SPEED, T. P. (2003). Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research*, **31**, 4–15.

- [11] LI, C. AND WONG, W. H. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, *Genome Biology*, **2**, 1–11.
- [12] LOCKHART, D. L., DONG, H., BYRNE, M. C., FOLLETTIE, M. T., GALLO, M. V., CHEE, M. S., MITTMAN, M., WANG, C., KOBAYASHI, M., HORTON, H. AND BROWN, E. L. (1996). Expression monitoring by hybridisation to high-density oligonucleotide arrays, *Nature Biotechnology*, **14**, 1675–1680.
- [13] NAEF, F., HACKER, C. R., PATIL, N. AND MAGNASCO, M. (2002). Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays, *Genome Biology*, **3**, 1–10.
- [14] SCHADT, E. E., LI, C., ELLIS, B. AND WONG, H. W. (2001). Feature Extraction and Normalization Algorithms for High-Density Oligonucleotide Gene Expression Array Data, *Journal of Cellular Biochemistry Supplement*, **37**, 120–125.
- [15] VENABLES, W. N. and RIPLEY, B. D. (1980). *Modern Applied Statistics with S-Plus*, Springer Verlag.
- [16] WU, Z., IRIZARRY, R. A., GENTLEMAN, R., MARTINEZ-MURILLO, F. AND SPENCER, F. (2004). A Model Based Background Adjustment for Oligonucleotide Expression Arrays, *Journal of the American Statistical Association*, **99**, 909–917.



# Bayesian two-gene interaction models in complex binary traits

Nuno Sepúlveda <sup>\*</sup>   Carlos Daniel Paulino <sup>†</sup>   Carlos Penha-Gonçalves <sup>‡</sup>

## Abstract

Current two-gene interaction models for complex binary traits assume linear effects of either gene in a given mathematical scale of penetrance. However, this assumption is much closer to a statistical than to a genetic description of penetrance. Therefore, few genetic interpretations can be taken from these models. To address this problem, we propose a novel statistical framework, the allelic penetrance approach, that models dominance and recessiveness in a single diallelic gene with reduced penetrance. Using this approach, we develop new two-gene interaction models with more genetic information, including the specification of the phenotype-conferring alleles at each gene as well as their genetic behavior. As an example, we fit the models to data on the susceptibility to experimental cerebral malaria in mice using Bayesian analysis via Gibbs sampling.

**Keywords:** binary traits, genetic interaction, allelic and external penetrance, Gibbs sampling, Bayesian model fitting.

## 1 Introduction

Binary traits are biological characters classified into two categories (*e.g.*, susceptibility or resistance to a certain disease). Their inheritance can be simple or complex. Simple binary traits are only affected by a single gene and exhibit often Mendelian dominant or recessive inheritance. Therefore, one can predict the phenotype through the genotype of an individual. In contrast, the so-called complex binary traits are the outcome of an intricated network of genetic and environmental factors, that makes extremely difficult to predict the phenotype from the genotype. At an extreme case, genetically identical individuals can manifest either one or the other phenotype.

Penetrance is the most important concept to analyze complex binary traits (Griffiths *et al.*, 2000) and embodies the conditional probability of an individual inheriting the phenotype

---

<sup>\*</sup>Instituto Gulbenkian de Ciência, Portugal. E-mail: [nunosep@igc.gulbenkian.pt](mailto:nunosep@igc.gulbenkian.pt).

<sup>†</sup>Departamento de Matemática, Instituto Superior Técnico, Portugal. E-mail: [dpaulino@math.ist.utl.pt](mailto:dpaulino@math.ist.utl.pt).

<sup>‡</sup>Instituto Gulbenkian de Ciência, Portugal. E-mail: [cpenha@igc.gulbenkian.pt](mailto:cpenha@igc.gulbenkian.pt).

of interest given its genotype. It is often stated that simple binary traits have complete penetrance, which take the values 1 or 0 depending whether the genotype is conferring or not the phenotype, respectively, while complex binary traits have reduced penetrance (values between 0 and 1).

Several models for penetrance have been proposed to infer the genetic interaction of two genes in complex binary traits. Those models are generalized linear models (GLM) (McCullagh and Nelder, 1989) with random component following a Binomial-product distribution and systematic component

$$g(\pi_{ij}) = \lambda + \alpha_i + \beta_j, \quad (1)$$

where  $\pi_{ij}$  is the penetrance of genotype  $i$  of gene 1 and of genotype  $j$  of gene 2,  $g(\cdot)$  is a link function and  $\sum_i \alpha_i = \sum_j \beta_j = 0$ . The most popular link functions are the identity (additive model), logarithm (multiplicative model; Hodge, 1981), complementary logarithm (heterogeneity model; Risch, 1990), logit (logistic model; Stewart, 2002) and probit (liability model; Pearson, 1900; Dempster and Lerner, 1950). However, it is recognized that these models are much closer to a statistical rather than genetic description of the penetrance, and so few genetic interpretations can be taken from these models (Cordell *et al.*, 2001).

The recent study on susceptibility to cerebral malaria in mice illustrates the difficulties of interpreting current two-gene interaction models (Bagot *et al.*, 2002). The authors reported the identification of two disease-associated loci through a genetic mapping on a  $F_2$  progeny of a backcross between a resistant and a susceptible strain, where  $F_1$  progeny was crossed with susceptible parental strain. GLM were fitted to data to infer about genetic interaction of the two loci. The usual goodness-of-fit tests did not rule out some of these models, but provided evidence against others. Even though some GLM could fit the data, Bagot *et al.* (2002) could not from that infer what kind of genetic interaction of the two loci is actually controlling the trait.

This paper proposes new two-gene interaction models with improved genetic interpretation, namely, the phenotype-conferring alleles are clearly specified as well as their genetic behavior (dominant or recessive). These models are based on a new approach to reduced penetrance, the allelic penetrance approach (Sepúlveda, 2004; Sepúlveda *et al.*, 2004, 2005). The basic idea of the approach is to decompose penetrance in an internal component pertaining to the penetrance of the alleles in the genotype and in an external component attributable to factors other than the genotype under study.

The structure of the paper is the following. First, we introduce the allelic penetrance approach for a single diallelic gene (Section 2). The framework is then extended in order to embody different genetic interaction mechanisms for two-gene diallelic genes (Section 3). A Bayesian analysis via Gibbs sampling is suggested to fit the models to experimental data (Section 4) and is applied to cerebral malaria data (Section 5). Finally, concluding remarks are presented (Section 6).

## 2 The allelic penetrance approach

The allelic penetrance approach aims to decompose penetrance in its main components. To do this, consider the most simple case: a single diallelic locus. In such a system, the main factors affecting penetrance would be: (i) the two alleles at the controlling locus, (ii) the contribution of the rest of the genome and (iii) the variability of environmental factors. The usual interpretation for reduced penetrance postulates the existence of environmental and other genetic factors, which can modify the action of the phenotype-conferring allele (Griffiths *et al.*, 2000; Nadeau, 2001). However, experimental genetics is rich in examples of reduced penetrance in pure lines maintained under strict environmental conditions. This suggests that somehow the genotype has an intrinsic stochastic property of being expressed at the level of the phenotype. Moreover, there are some evidences that reduced penetrance can actually be attributed to the alleles themselves, such as the cases of loss-of-function mutations (Lalucque and Silar, 2004) and the "metastable epialleles" (Rakyan *et al.*, 2002).

Based on these observations, penetrance can be decomposed in a sum of two components: an *internal component* attributable to the probability of the alleles of the genotype being expressed at the level of the phenotype and an *external component* pertaining to the probability of the phenotype being affected by other (genetic and/or environmental) factors than the gene under study. The probability of an allele being expressed at the level of the phenotype is from now on referred to as the allelic penetrance.

Consider a gene with alleles  $a$  and  $b$  that control partially the expression of a binary trait. Let  $a$  be a dominant allele over allele  $b$  with respect to the phenotype of interest. In absence of external factors, the phenotype is ascribed to the expression of at least one allele  $a$ .

Let  $\pi_g$  be the penetrance of genotype  $g = aa, ab, bb$ , respectively. Denote the penetrances of alleles  $a$  and  $b$  by  $\pi_a$  and  $\pi_b$ , respectively. On the basis of the above reasoning, we define the internal component of penetrance as the allelic expression probabilities of each genotype towards the phenotype of interest. Considering independent allelic expressions, the internal component of penetrance is

$$\pi_g^{int} = \begin{cases} \pi_a^2 + 2\pi_a(1 - \pi_a), & \text{if } g = aa \\ \pi_a, & \text{if } g = ab \\ 0, & \text{if } g = bb \end{cases} \quad (2)$$

In this framework the external component of penetrance refers to the action of external factors to the gene under study in promoting the phenotype. This action is assumed to be only relevant in the absence of expression of allele  $a$ . That is, external factors acts as a "backup mechanism" for the expression of the phenotype of interest. In this line of thought external factors can force the phenotype to be the one of interest when none of the alleles of the genotype is being expressed or even when the non-conferring alleles  $b$  are being expressed. Considering  $\pi_{ext}$  as the probability of the external factors favoring the phenotype of interest,

the external component of penetrance is then

$$\pi_g^{ext} = (1 - \pi_g^{int}) \pi_{ext} = \begin{cases} (1 - \pi_a)^2 \pi_{ext}, & \text{if } g = aa \\ (1 - \pi_a) \pi_{ext}, & \text{if } g = ab \\ \pi_{ext}, & \text{if } g = bb \end{cases} . \quad (3)$$

The final formula of penetrance is just the sum of internal and external components, i.e.,

$$\pi_g = \pi_g^{int} + \pi_g^{ext} = \begin{cases} \pi_a^2 + 2\pi_a(1 - \pi_a) + (1 - \pi_a)^2 \pi_{ext}, & \text{if } g = aa \\ \pi_a + (1 - \pi_a) \pi_{ext}, & \text{if } g = ab \\ \pi_{ext}, & \text{if } g = bb \end{cases} . \quad (4)$$

Consider now that allele  $a$  is recessive with respect to phenotype. Here it is assumed that the phenotype is acquired when there is no expression of the dominant allele  $b$  and there is expression of at least one allele  $a$ . Using the above reasoning, the internal component of penetrance is

$$\pi_g^{int} = \begin{cases} \pi_a^2 + 2\pi_a(1 - \pi_a), & \text{if } g = aa \\ \pi_a(1 - \pi_b), & \text{if } g = ab \\ 0, & \text{if } g = bb \end{cases} . \quad (5)$$

The external component of penetrance follows previous reasoning and therefore external factors act solely when the alleles  $a$  are not expressing the phenotype. In this line of thought we have the following external component of penetrance

$$\pi_g^{ext} = (1 - \pi_g^{int}) \pi_{ext} = \begin{cases} (1 - \pi_a)^2 \pi_{ext}, & \text{if } g = aa \\ (1 - \pi_a)(1 - \pi_b) \pi_{ext}, & \text{if } g = ab \\ \pi_{ext}, & \text{if } g = bb \end{cases} . \quad (6)$$

Finally, by summing (5) and (6) the penetrance of the phenotype follows

$$\pi_g = \begin{cases} \pi_a^2 + 2\pi_a(1 - \pi_a) + (1 - \pi_a)^2 \pi_{ext}, & \text{if } g = aa \\ \pi_a(1 - \pi_b) + (1 - \pi_a)(1 - \pi_b) \pi_{ext}, & \text{if } g = ab \\ \pi_{ext}, & \text{if } g = bb \end{cases} . \quad (7)$$

### 3 Statistical models for two-locus joint action

Here we extend the decomposition of penetrance for the two diallelic gene case, i.e.,

$$\pi_{g_1g_2} = \pi_{g_1g_2}^{int} + \pi_{g_1g_2}^{ext}, \quad (8)$$

where  $\pi_{g_1g_2}^{int}$  and  $\pi_{g_1g_2}^{ext}$  are the internal and the external components of penetrance, respectively, for the combined genotype  $g_1$  in gene 1 and  $g_2$  in gene 2. As stated in the single gene case, external factors are only relevant when there is no allelic expression of the two interacting

genes towards the phenotype of interest. Therefore, the external penetrance can be factorized as

$$\pi_{g_1g_2}^{ext} = (1 - \pi_{g_1g_2}^{int}) \pi_{ext}, \quad (9)$$

which substituted in equation (8) leads to the following general formula of penetrance

$$\pi_{g_1g_2} = \pi_{g_1g_2}^{int} + (1 - \pi_{g_1g_2}^{int}) \pi_{ext}. \quad (10)$$

Different genetic interaction mechanisms can be considered by specializing the internal component of penetrance. Some of them are described as follows.

### 3.1 Independent Action Models

The independent action models (IAM) rely on the so-called genetic heterogeneity (Risch, 1990; Vieland and Huang, 2003), which establish that the phenotype is acquired by the independent action of each gene. Here, each gene has a phenotype-conferring allele, which can be either dominant or recessive. Thus, there are four types of IAM according to the genetic behavior of the phenotype-conferring alleles at each gene: dominant-dominant, dominant-recessive, recessive-dominant and recessive-recessive.

Derivation of the penetrances according to IAM follows almost straightforwardly from what has been written so far. The internal component of penetrance is simply defined by the probabilities of expression of the phenotype-conferring alleles at either gene towards the phenotype of interest. Since heterogeneity means independent action of the two genes at the level of the phenotype, the internal component of penetrance satisfies the probabilistic relationship for the union of two independent events, each one referring to the allelic expressions of each gene, i.e.,

$$\pi_{g_1g_2}^{int} = \phi_{g_1} + \phi_{g_2} - \phi_{g_1}\phi_{g_2}, \quad (11)$$

where  $\phi_{g_1}$  and  $\phi_{g_2}$  are the probabilities of expression of the phenotype-conferring alleles at genotypes  $g_1$  and  $g_2$ , respectively. If the phenotype-conferring allele at one gene is dominant, then the corresponding  $\phi_{g_i}$  follows equation (2). Analogously, if the phenotype-conferring allele is recessive, then the respective  $\phi_{g_i}$  is determined by equation (5). Finally, external factors are included in the model through (10) with  $\pi_{g_1g_2}^{int}$  determined by (11).

### 3.2 Inhibition Models

In the early days of Genetics, Bateson (1909) described a phenomenon termed epistasis whereby an allele of a given gene prevents an allele of another gene from manifesting its effect. Nowadays, epistasis has been used in many contexts with different and often conflicting meanings that has led to a lack-of-consensus about its formal definition (Phillips, 1998;

Cordell, 2002). As a consequence, many authors have proposed different models to deal with epistatic effects: additive models adapted from quantitative genetics considering penetrance as a quantitative trait (Fisher, 1918; Cockerham, 1954; Kao and Zeng, 2002), multiplicative models (Hodge, 1981), and heterogeneity models (Risch, 1990; Vieland and Huang, 2003). Every author claims that epistasis is present when his model does not fit the data under study.

Here we recall Bateson’s definition of epistasis to develop inhibition models (IM). These models describe a type of interaction where one gene confers the phenotype by the expression of its respective phenotype-conferring allele, whereas the other gene simply inhibits the phenotypic expression of the former by its inhibiting alleles. Phenotype-conferring or -inhibiting alleles can be considered either dominant or recessive.

Consider that gene 1 confers the phenotype of interest and gene 2 inhibits the expression of the former. In this case, the internal penetrance relates to the probability of the phenotype-conferring alleles at gene 1 being expressed with no expression of phenotype-inhibiting alleles at gene 2. Thus, the internal component of penetrance satisfies

$$\pi_{g_1g_2}^{int} = \phi_{g_1} (1 - \phi_{g_2}^*), \quad (12)$$

where  $\phi_{g_1}$  is the probability of genotype  $g_1$  expressing the phenotype of interest and  $\phi_{g_2}^*$  is the probability of genotype  $g_2$  having an inhibitory behavior. Dominance and recessiveness are included in the model by replacing  $\phi_{g_1}$  and  $\phi_{g_2}^*$  by the single-gene internal penetrances (2) and (5), respectively. Finally we include external factors in penetrance through equation (10) with internal penetrance given by the above equation.

### 3.3 Minimum Alleles Models

Falconer (1965) coined the term liability to refer to an unobserved quantitative continuous trait influenced by genetic and environmental factors that controls the inheritance of a multifactorial disease. Under this concept, the phenotype is inherited when liability of an individual exceeds a certain threshold. Usually liability is assumed to follow a (standard) Gaussian distribution with genes contributing additively to its mean (Pearson, 1900; Dempster and Lerner, 1950; Risch *et al.*, 1993). One can also model liability by the (standard) Logistic distribution, which leads to the popular logistic model.

Recently Stewart (2002) proposed the concept of allelic liability to model the inheritance of multifactorial diseases. In his work, disease occurs when the number of disease-conferring alleles in an individual exceeds a critical value. Based on this idea the minimum allele models (MAM) establish that the phenotype of interest is inherited when the joint expression of the phenotype-conferring alleles at both genes exceeds a certain level. Note that dominance and

recessiveness are not included in the model, because what matters here is the cumulative expression of phenotype-conferring alleles.

It is worth noting that MAM requiring at least one phenotype-conferring allele being expressed is identical to IAM with dominant phenotype-conferring alleles at both genes, because both models rely on the same condition for the expression of the phenotype. This shows that different genetic interaction mechanisms can be described mathematically using the same requisite for the inheritance of phenotype.

Let  $x_i$  represent the number of phenotype-conferring alleles in the genotype of gene  $i = 1, 2$ . Let also  $Y_i$  be the random variable that indicates the number of those alleles expressing the phenotype at gene  $i = 1, 2$ . According to the allelic penetrance approach,  $Y_i|x_i$  has a Binomial distribution with  $x_i$  trials and probability of success given by the allelic penetrance  $\pi_i$  of the phenotype-conferring allele at gene  $i = 1, 2$ . Assuming independence between  $Y_1$  and  $Y_2$ , the probability mass function of the total number  $Y$  of phenotype-conferring alleles expressing the phenotype given combined genotype  $(x_1, x_2)$  is determined by

$$P[Y = y|x_1, x_2] = \sum_{l=0}^{x_1} P[Y_1 = l|x_1, x_2] P[Y_2 = y - l|x_1, x_2], \quad (13)$$

where

$$P[Y_i = y_i|x_1, x_2] = \binom{x_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{x_i - y_i}. \quad (14)$$

Thus, MAM entails the following internal penetrance

$$\pi_{x_1 x_2}^{int} = P[Y \geq k|x_1, x_2] = \sum_{y=k}^{x_1+x_2} P[Y = y|x_1, x_2], \quad k = 1, \dots, 4, \quad (15)$$

where  $P[Y = y|x_1, x_2]$  is given by (13) with  $k \leq x_1 + x_2$ . As in IAM, the effect of the external factors are included in the model by (10) with  $\pi_{g_1 g_2}^{int}$  as calculated by (15).

## 4 Bayesian analysis

Penetrance data from experimental crosses are usually represented in the form of  $I \times J \times 2$  contingency tables, where  $I$  and  $J$  are the number of genotypes of each gene and the dimension of the third variable refers to binary trait under study. For instance,  $I = J = 3$  and  $I = J = 2$  when analyzing data from a  $F_2$  progeny of an intercross or a backcross, respectively. Like in GLM, we consider the Binomial-product distribution as sampling model.

We follow a Bayesian analysis for the purpose of model fitting. Here it is assumed prior independence among the parameters of the models. Since allelic and external penetrances are new concepts in genetic interaction modeling, it is reasonable to specify non-informative (Uniform) prior distributions for them.

Table 1: Cerebral malaria data where  $s_i$  and  $r_i$  denote the alleles inherited from susceptible and resistant strains in locus  $i = 1, 2$ , respectively. Penetrance refers to susceptibility.

Genotypes		Phenotype		
Locus 1	Locus 2	Susceptible	Resistant	Penetrance
$s_1s_1$	$s_2s_2$	35	10	0.78
	$r_2s_2$	25	23	0.52
$r_1s_1$	$s_2s_2$	27	21	0.56
	$r_2s_2$	9	40	0.18

The great complexity of the proposed models shows *per se* that posterior distributions for parameters are impossible to be handled analytically. To overcome this problem, we use Markov Chains Monte Carlo methods via Gibbs sampling with the help of WinBUGS (Spiegelhalter *et al.*, 2003). Sepúlveda (2004) shows that full conditional distributions are log-concaves, which allows the software to use the adaptive rejection method (Gilks, 1992). The diagnostics of convergence is performed in the Bayesian Output Analysis software (Smith, 2003) and can be found elsewhere (Sepúlveda, 2004).

Since many models can fit the data, we design a strategy of model selection and comparison which can be performed in real time. First, we evaluate empirically the effects of the alleles of each gene with respect to the phenotype of interest. In this way we can establish which are the most plausible phenotype-conferring alleles as well as the potential phenotype-inhibiting alleles. Second, we calculate the deviance information criterion (DIC) (Spiegelhalter, 2002) and posterior mean of Pearson’s parametric function (PMP) for each model, selecting those with the lowest values for both measures. Finally, we compare the selected models through prior predictive probability (PPP) and sum of logarithm of conditional predictive ordinates (SLNCPO). The best-fitted models are the ones with the highest values for both measures.

Bayesian estimation in the best-fitted models requires essentially the calculation of posterior mean, median and standard deviation for allelic, external and genotypic penetrances. Highest posterior density (HPD) credible intervals are determined through a method proposed by Chen and Shao (1999). We also compute a credible region for  $c$  genotypic penetrances by the cartesian product of the  $\gamma^{1/c} \times 100\%$  individual HPD credible intervals for these parameters.

## 5 Application to cerebral malaria data

Here we fit our models to the cerebral malaria data referred to in Section 1, considering susceptibility as the phenotype of interest. With this purpose we first present the notation for our models: IAM( $s_1/S_2$ ) is an IAM with phenotype-conferring alleles  $s_1$  and  $S_2$  at loci 1 and 2, respectively, where capital and small letters denote dominant and recessive alleles, respec-



tively. For inhibition models we use the same kind of notation, but we add the superscripts  $c$  and  $i$  in each allele to denote phenotype-conferring and -inhibiting alleles, respectively (for instance,  $\text{IM}(s_1^c/R_2^i)$ ). Finally,  $\text{MAM}_k(s_1/s_2)$  is a MAM with phenotype-conferring  $s_1$  and  $s_2$  at loci 1 and 2, respectively, which requires at least the expression of  $k$  alleles to acquire the phenotype. Alleles in these models are always represented by small letters.

Table 1 shows the genotypes and phenotypes for the  $F_2$  generation of the experimental cross. The observed data suggests that the alleles derived from the susceptible strain at either gene tend to increase the penetrance of susceptibility. Taking into account this observation, we first select IAM and MAM with phenotype-conferring alleles derived from the susceptible strain, whereas IM are chosen with one phenotype-conferring allele in one locus derived from the susceptible strain and one phenotype-inhibiting allele from the resistant strain in the other locus. However, there are some IM that parameterize the data solely by external penetrance (for instance,  $\text{IM}(s_1^c/r_2^i)$ ). We exclude these ones, because they provide very little information about the underlying genetic interaction. Therefore, we select 12 models: the four types of IAM with phenotype-conferring alleles  $s_1$  and  $s_2$ ,  $\text{MAM}_k(s_1/s_2)$ ,  $k = 1, 2, 3, 4$ ,  $\text{IM}(S_1^c/R_2^i)$ ,  $\text{IM}(R_1^i/S_2^c)$ ,  $\text{IM}(R_1^i/s_2^c)$  and  $\text{IM}(s_1^c/R_2^i)$ .

Note that the models with some recessive allele are overparameterized. This is a consequence of modelling internal penetrance of recessive alleles using two allelic penetrances, one for the dominant allele and another for the recessive allele (see equation (5)). Since the resistant parental strain was observed to be 100% resistant to the disease while the other strain was not fully susceptible, we could avoid overparametrized models by assuming that phenotype-conferring and -inhibiting alleles have reduced allelic penetrance while the remaining alleles have complete allelic penetrance.

Table 2 presents the observed values of PMP and DIC for the models described above. There are 5 models that show low values for these measures, namely,  $\text{IAM}(S_1/s_2)$ ,  $\text{IAM}(s_1/s_2)$ ,  $\text{IM}(S_1^c/R_2^i)$ ,  $\text{MAM}_2(s_1/s_2)$  and  $\text{MAM}_3(s_1/s_2)$ . Following our model selection procedure, we calculate PPP and SLNCPO estimates for these 5 models (see again Table 2). The results distinguish clearly  $\text{IAM}(s_1/s_2)$  and  $\text{MAM}_3(s_1/s_2)$  from the remaining models. Taking account of the previous results, we conclude that these two models are the ones that best fit the experimental data.

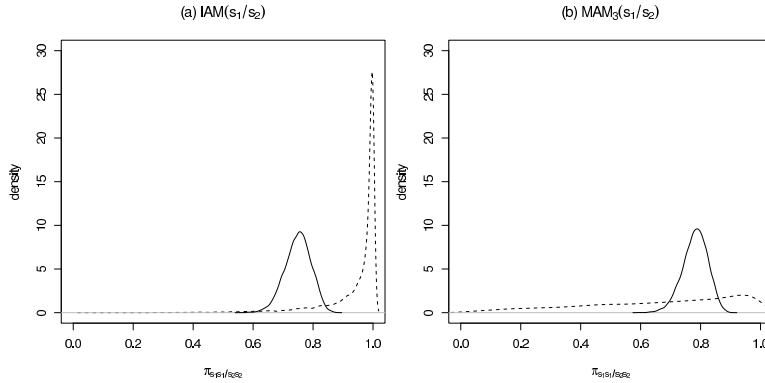
At this stage of the analysis, we can conclude that the two following genetic interaction mechanisms can explain the observed data: (1) the alleles derived from the susceptible strain at either locus are recessive and sufficient to cause the disease, or (2) both *loci* act together so that susceptibility is inherited when there are at the two loci jointly at least three alleles derived from the susceptible strain expressing the phenotype.

To conclude this analysis, Table 3 presents estimates for different parameters of the models, where the credibility degree of each individual HPD interval for genotypic penetrances is setup at 98.7% in order to obtain a 95% overall credibility degree for the respective credible regions.

Table 2: Comparison and selection of genetic interaction models.

Model	PMP	DIC	SLNCPO	PPP
IAM( $S_1/S_2$ )	16.660	36.262	—	—
IAM( $S_1/s_2$ )	7.449	26.335	-118.587	$6.06 \times 10^{-8}$
IAM( $s_1/S_2$ )	9.215	28.249	—	—
IAM( $s_1/s_2$ )	3.210	22.643	-116.424	$1.28 \times 10^{-6}$
IM( $S_1^c/R_2^i$ )	7.859	22.878	-118.797	$1.45 \times 10^{-7}$
IM( $R_1^i/S_2^c$ )	9.219	26.248	—	—
IM( $R_1^i/s_2^c$ )	15.140	34.600	—	—
IM( $s_1^c/R_2^i$ )	17.840	37.359	—	—
MAM <sub>1</sub> ( $s_1/s_2$ )	16.660	36.262	—	—
MAM <sub>2</sub> ( $s_1/s_2$ )	6.090	25.266	-117.806	$2.21 \times 10^{-7}$
MAM <sub>3</sub> ( $s_1/s_2$ )	3.038	22.035	-116.243	$1.41 \times 10^{-6}$
MAM <sub>4</sub> ( $s_1/s_2$ )	19.660	37.734	—	—

Figure 1: Prior (dashed line) and posterior (solid line) densities of  $\pi_{s_1 s_1 / s_2 s_2}$  for the best-fitted models.



The estimates show that both models are almost indistinguishable in terms of genotypic penetrance estimates. Figure 1 displays the plots of prior (dashed line) and posterior (solid line) densities of  $\pi_{s_1 s_1 / s_2 s_2}$  for each model. This figure shows that there is a strong update of the respective prior distributions by experimental data.

## 6 Concluding remarks

The present work proposes an allelic penetrance approach to model genetic interactions in complex binary traits. This framework is based on two types of basic parameters: allelic and external penetrances. The allelic penetrance is an intrinsic stochastic property of phenotypic expression of the alleles and agrees with current observations of stochastic gene expression.

Table 3: Posterior estimates of IAM( $s_1/s_2$ ) and MAM<sub>3</sub>( $s_1/s_2$ ). <sup>a</sup>95% and <sup>b</sup>98.7%.

IAM( $s_1/s_2$ )					
Parameters	Mean	Median	SE	HPD IC	
$\pi_{s_1}$	0.243	0.278	0.057	0.125	0.348 <sup>a</sup>
$\pi_{s_2}$	0.278	0.278	0.056	0.167	0.388 <sup>a</sup>
$\pi_{ext}$	0.195	0.191	0.053	0.102	0.306 <sup>a</sup>
$\pi_{s_1s_1/s_2s_2}$	0.759	0.761	0.042	0.649	0.855 <sup>b</sup>
$\pi_{s_1s_1/s_2r_2}$	0.537	0.539	0.065	0.379	0.691 <sup>b</sup>
$\pi_{s_1r_1/s_2s_2}$	0.579	0.581	0.062	0.418	0.727 <sup>b</sup>
$\pi_{s_1r_1/s_2r_2}$	0.195	0.191	0.055	0.077	0.335 <sup>b</sup>
MAM <sub>3</sub> ( $s_1/s_2$ )					
Parameters	Mean	Median	SE	HPD IC	
$\pi_{s_1}$	0.702	0.704	0.111	0.485	0.915 <sup>a</sup>
$\pi_{s_2}$	0.776	0.781	0.111	0.584	1.000 <sup>a</sup>
$\pi_{ext}$	0.211	0.208	0.057	0.105	0.322 <sup>a</sup>
$\pi_{s_1s_1/s_2s_2}$	0.782	0.785	0.042	0.663	0.871 <sup>b</sup>
$\pi_{s_1s_1/s_2r_2}$	0.512	0.512	0.058	0.374	0.651 <sup>b</sup>
$\pi_{s_1r_1/s_2r_2}$	0.542	0.543	0.056	0.397	0.670 <sup>b</sup>
$\pi_{s_1r_1/s_2r_2}$	0.211	0.208	0.057	0.087	0.358 <sup>b</sup>

The external penetrance is the probability of external factors, other than the genes under study, favoring the phenotype and it is assumed to be constant over the genotypes of genes under analysis. In future work, we intend to model external penetrance in order to include covariates embodying different environmental factors or genetic backgrounds. In this regard it would be particularly important to include pertinent information on parental strains and  $F_1$  generation of experimental genetic crosses.

The proposed genetic interaction models have the advantage of providing clear genetic interpretation. However, some models present the disadvantage of being saturated or over-parametrized models in backcross data. In the cerebral malaria example, this problem was overcome by attributing reduced allelic penetrance for phenotype-conferring and -inhibiting alleles and complete allelic penetrance for the non-conferring alleles, because susceptible parental strains exhibit reduced penetrance and its resistant counterpart show complete penetrance. Therefore, the usage of the models in backcross data is only recommended when the phenotype of interest shows reduced penetrance while the alternative phenotype has complete penetrance in their respective parental strains.

Finally, the two best-fitted models for cerebral malaria data suggest future experiments in the genetic dissection of the trait. Thus, we expect that the susceptible single-locus congenic

strains would be either susceptible or resistant to cerebral malaria infection if IAM( $s_1/s_2$ ) or MAM<sub>3</sub>( $s_1/s_2$ ) hold, respectively.

## Acknowledgments

We are grateful to Dan Holmberg laboratory at the Instituto Gulbenkian de Ciência (IGC), specially, to Dra. Susana Campino for kindly providing the cerebral malaria data. We would like also to thank IGC for financial support and to Jorge Carneiro, Rui Gardner, Tiago Paixão and Ramiro Magno for valuable discussions.

## References

- [1] BAGOT, S., CAMPINO, S., PENHA-GONÇALVES, C., PIED, S., CAZENAVE, P. and HOLMBERG, D. (2002). Identification of two cerebral malaria resistance loci using an inbred wild-derived mouse strain. *Proceedings of The National Academy of Sciences* **99**, 9919–9923.
- [2] BATESON, W. (1909). *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge.
- [3] CHEN, M.-H. and SHAO, Q.-M. (1999). Monte carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* **8**, 69–92.
- [4] COCKERHAM, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39**, 859–882.
- [5] CORDELL, H. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* **11**, 2463–2468.
- [6] CORDELL, H., TODD, J., HILL, N., LORD, C., LYONS, P., PETERSON, L., WICKER, L. and CLAYTON, D. (2001). Statistical Modeling of Interlocus Interactions in a Complex Disease: Rejection of the Multiplicative Model of Epistasis in Type I Diabetes. *Genetics* **158**, 357–367.
- [7] DEMPSTER, E. R. and LERNER, I. M. (1950). Heritability of threshold characters. *Genetics* **35**, 212–236.
- [8] FALCONER, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics* **29**, 51–76.
- [9] FISHER, R. A. (1918). The correlations between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* **52**, 399–433.

- [10] GILKS, W. (1992). *Derivative-free adaptive rejection sampling for Gibbs sampling*. In "Bayesian Statistics 4" (Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., Eds.) Oxford University Press, Oxford, 641–665.
- [11] GRIFFITHS, A., MILLER, J., SUZUKI, D., LEWONTIN, R. and GELBART, W. (2000). *An Introduction to Genetic Analysis*. 7th Edition, W. H. Freeman, New York.
- [12] HODGE, S. (1981). Some epistatic two-locus models of disease. I. Relative risks and identity-by-descent distributions in affected sib pairs. *American Journal of Human Genetics* **33**, 381–395.
- [13] KAO, C.-H. and ZENG, Z.-B. (2002). Modelling epistasis of quantitative trait locus using Cockerham's model. *Genetics* **160**, 1243–1261.
- [14] LALUCQUE, H. and SILAR, P. (2004). Incomplete penetrance and variable expressivity of a growth defect as a consequence of knocking out two K(+) transporters in the eukaryotic fungus *Podospora anserina*. *Genetics* **166**, 125–133.
- [15] MCCULLAGH, P. and NELDER, J. (1989). *Generalized Linear Models*. 2nd Edition, Chapman and Hall, London.
- [16] NADEAU, J. H. (2001). Modifier genes in mice and humans. *Nature Reviews in Genetics* **2**, 165–174.
- [17] PEARSON, K. (1900). Mathematical contributions to the theory of evolution. VIII. On the inheritance of characters not capable of exact quantitative measurement. *Philosophical Transactions of the Royal Society of London Series A* **195**, 79–121.
- [18] PHILIPS, P. C. (1998). The Language of Gene Interaction. *Genetics* **149**, 1167–1171.
- [19] RAKYAN, V., BLEWITT, M., DRUKER, R., PREIS, J. and WHITELAW, E. (2002). Metastable epialleles in mammals. *Trends in Genetics* **18**, 348–351.
- [20] RISCH, N., GOSH, S. and TODD, J. A. (1993). Statistical evaluation of multiple locus linkage data in experimental species and relevance to human studies: application to murine and human IDDM. *American Journal of Human Genetics* **53**, 702–714.
- [21] RISCH, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. *American Journal of Human Genetics* **46**, 222–228.
- [22] SEPÚLVEDA, N. (2004). *Modelos estatísticos para a acção conjunta de dois loci em fenótipos binários complexos*. Master's Thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

- [23] SEPÚLVEDA, N., PAULINO, C. D. and PENHA-GONÇALVES, C. (2004). *Modelos de Interação Genética: Uma abordagem por Penetrâncias Alélicas*. In "Estatística com Necessidade e Acaso" (Rodrigues, P., Rebelo, E. and Rosado, F., eds.), SPE Editions, Lisbon, 735–746.
- [24] SEPÚLVEDA, N., PAULINO, C. D. and PENHA-GONÇALVES, C. (2005). *Modelos de interação genética de dois genes em fenótipos binários complexos: nova modelação e aplicação de metodologia bayesiana*. In "Estatística Jubilar" (Braumann, C. A. , Infante, P. , Oliveira, M.M.,Alpizar-Jara, R. and Rosado, F., Eds.), SPE Editions, Lisbon, 703–716.
- [25] SMITH, B. (2003). Bayesian Output Analysis Program (BOA) Version 1.0 User's Manual. Department of Biostatistics, College of Public Health, University of Iowa.
- [26] SPIEGELHALTER, D., BEST, N., CARLIN, B. and VAN DER LINDEN, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of Royal Statistical Society Series B* **64**, 583–640.
- [27] SPIEGELHALTER, D., THOMAS, A., BEST, N. and LUNN, D. (2003). WinBUGS User Manual, Version 1.4. MRC Biostatistics Unit, Institute of Public Health & Department of Epidemiology and Public Health, Imperial College School of Medicine.
- [28] STEWART, J. (2002). Towards the genetic analysis of multifactorial diseases: the estimation of allele frequency and epistasis". *Human Heredity* **54**, 118-131.
- [29] VIELAND, V. and Huang, J. (2003). Two-locus heterogeneity cannot be distinguished from two-locus epistasis on the basis of affected-sib-pairs. *American Journal of Human Genetics* **73**, 223–232.

# Statistical analysis in mapping quantitative genetic traits

Elisabete Fernandes <sup>\*</sup>    Luísa Canto e Castro <sup>†</sup>    Carlos Penha-Gonçalves <sup>‡</sup>

## Abstract

The majority of the measurable inherited traits have quantitative nature and present wide variation in the population. The genetic factors controlling such traits are named quantitative trait loci (QTLs).

The statistical problem of localization of QTLs has come to deserve great attention in the last decades. Interval mapping using the likelihood approach has been the most commonly used method for QTLs mapping in experimental crosses. This method assumes that quantitative phenotypes follow a normal distribution in the population. The interval mapping model is illustrated with data originated from an intercross experiment to identify QTL contributing to variance in the amount of immunoglobulin IgM in serum of mice.

The application of interval mapping method revealed the presence of a controlling locus located between the markers D13Mit115 and D13Mit266 on chromosome 13.

**Keywords:** quantitative trait loci; interval mapping; immunoglobulin IgM.

**AMS classifications:** 49A05, 78B26.

## 1 Introduction

There are many biologically and economically important traits in higher organisms which are quantitative, not qualitative. Traits such as body weight in humans, beef cattle, or mice; blood pressure or glucose level in humans, and oil content in corn, are all examples of *quantitative traits*. The underlying distribution of the quantitative traits is continuous. The values of the traits are measured, not observed or counted, and these traits may be influenced by many genes and environmental factors.

*Quantitative trait loci* (QTLs) are genes that influence quantitative traits. QTL refers to a locus or a specific region in the genome (not necessarily a single gene) associated with a quantitative trait.

---

<sup>\*</sup>DEIO and CEAUL, Faculdade de Ciências, Universidade de Lisboa, Instituto Gulbenkian de Ciência. E-mail: [ebfernandes@fc.ul.pt](mailto:ebfernandes@fc.ul.pt).

<sup>†</sup>DEIO and CEAUL, Faculdade de Ciências, Universidade de Lisboa. E-mail: [lloura@fc.ul.pt](mailto:lloura@fc.ul.pt).

<sup>‡</sup>Instituto Gulbenkian de Ciência. E-mail: [cpenha@igc.gulbenkian.pt](mailto:cpenha@igc.gulbenkian.pt).

For a long time, geneticists and breeders have tried to obtain QTL information in order to manipulate them and through that to improve traits. Recently, with the rapid advancement in molecular biology, it has become possible for scientists to gain fine scale genetic maps for various organisms by determining the genomic positions of a number of *genetic markers* (restriction fragment length polymorphism, isozymes, random amplified polymorphic DNA), and to obtain a complete classification of marker genotypes by using codominant markers (Liu [5]). These advances greatly facilitate the investigation of individual QTL.

Various statistical approaches have been developed to identify QTLs by using markers (DNA portion, with position and genotype known). Lander and Botstein [4] proposed *interval mapping* to detect and localize QTL by using a maximum likelihood method. Their method combined Today's [7] idea, which used two markers flanking a region where two QTL might fall, and Wellers idea [8], which used the maximum likelihood methods to estimate the location and effects of QTL. The *expectation maximization* (EM) algorithm (Dempster *et al.* [3]) was used to calculate the maximum likelihood estimates.

Calculating *thresholds* is a very important practical issue in the design and analysis of QTL. Due to the multiple tests on the whole genome, the usual pointwise significance level based on the chi-square approximation is inappropriate, since the tests are not independent among marker loci. A shuffling approach can be used to determine the empirical distributions of statistics (Churchill and Doerge [2]).

In this article, we describe the typical intercross experimental and data structure. We illustrate the use of the interval mapping with one data set on the amount of immunoglobulin IgM in serum of mice.

## 2 Interval mapping

### 2.1 Intercross Experimental and Data Structure

To generate a population for QTL mapping, two phenotypically distinct inbred lines differing in the phenotype of interest,  $P_1$  and  $P_2$ , are chosen as parents. It is assumed that these lines are homozygous for all loci. At each locus, the genotypes of parents  $P_1$  and  $P_2$  are  $MM$  and  $mm$ , respectively. The crossbred  $F_1$  generation is then heterozygous with genotype  $Mm$  at all loci, receiving one allele from each parent. The  $F_1$  individuals are mated among themselves to produce a  $F_2$  generation of intercrossed mice, in which each mouse has probability  $1/4$ ,  $1/2$  and  $1/4$  of being  $MM$ ,  $Mm$  and  $mm$ , respectively.

In an *intercross* experiment, each individual of the progeny is scored for one or more traits. Here, the quantitative phenotype is the amount of immunoglobulin IgM in serum of mice. The mice are typed at a number of genetic markers spread out over the all genome. At each of these marker loci, it is determined for each individual of the progeny whether the genotype is  $MM$ ,  $Mm$  and  $mm$ .



A *genetic map*, specifying the relative locations of the markers, may be known or estimated using the data of the current experiment. Such map gives the linear order of markers on the various chromosomes. The distance between markers in a genetic map is given by *map distance*, in the units centiMorgans (cM). Two markers are separated by  $d$  cM, if  $d$  is the expected number of crossovers between the markers in 100 meiotic products.

*Crossover* is the exchange of genetic material between a homologous pair of chromosomes in meiosis. The gametes resulting from the odd number of crossovers between two loci are recombinants. Recombination frequency, or recombination fraction  $r$  is defined as the proportion of recombinants individuals in the population. The *recombination fraction* between two loci depends on the distance between these two loci on the chromosome. An appropriate mapping function, such as Haldanes function converts the recombination fraction in map distance. In this article, we use the map distances of the markers included in the database of the Whitehead Center for Genome Research (USA) (*Mouse Genome Informatics*, MGI).

Typical experiments involve between 100 and 1000 progenies and use between 90 and 300 genetic markers. Here, we use 136 mice and 99 markers.

## 2.2 Model

The interval mapping approach is based on a search for a putative QTL, with alleles  $a$  and  $b$ , between two adjacent markers  $M_1$  and  $M_2$ . Flanking markers  $M_1$  and  $M_2$  are linked with recombination fraction  $r$  and the QTL is located between the two markers with recombination fraction  $r_1$  from marker  $M_1$  and  $r_2$  from marker  $M_2$ .

When  $r$  is small (two markers are tightly linked) no double crossover is assumed and the relationship of these recombination fractions is

$$r = r_1 + r_2 \tag{1}$$

We consider a sample of  $n$  individuals from a  $F_2$  population. Let  $y_j$  and  $\mathbf{m}_j$  denote the quantitative phenotype and the multipoint marker data, respectively, for the individual  $j$ . In the  $F_2$  model, given the QTL genotype  $aa$  or  $ab$  or  $bb$ , the trait is assumed normally distributed with mean  $\mu_1 = \mu_{aa}$  or  $\mu_2 = \mu_{ab}$  or  $\mu_3 = \mu_{bb}$  and common standard deviation  $\sigma$ . The likelihood function of the parameter vector  $\boldsymbol{\theta}^* = (\mu_{aa}, \mu_{ab}, \mu_{bb}, \sigma, p_{ij})$  is given by

$$L(\boldsymbol{\theta}^* | \mathbf{y}, \mathbf{m}) = \prod_{j=1}^n \sum_{i=1}^3 p_{ij} \frac{1}{\sqrt{2\pi}\sigma} e^{\left[ \frac{-(y_j - \mu_i)^2}{2\sigma^2} \right]} \tag{2}$$

where  $p_{ij}$  is the conditional probability that the individual  $j$  has genotype  $i$ , given the marker data and the QTL position. For example  $p_{1j} = \Pr(\text{QTL genotype of individual } j \text{ is } aa \mid \text{genotype of markers } M_1 \text{ and } M_2 \text{ for individual } j \text{ is } \mathbf{M}_j)$ . These probabilities are in table 1 (Liu [5]).

Number	Genotypes of markers	$P(aa M_j)$	$P(ab M_j)$	$P(bb M_j)$
1	$M_1M_1M_2M_2$	1	0	0
2	$M_1M_1M_2m_2$	$r_2/r$	$r_1/r$	0
3	$M_1M_1m_2m_2$	$(r_2/r)^2$	$2r_1r_2/r^2$	$(r_1/r)^2$
4	$M_1m_1M_2M_2$	$r_1/r$	$r_2/r$	0
5	$M_1m_1M_2m_2$	$\frac{r_1r_2}{(1-r)^2+r^2}$	$\frac{(1-r)^2+r_1^2+r_2^2}{(1-r)^2+r^2}$	$\frac{r_1r_2}{(1-r)^2+r^2}$
6	$M_1m_1m_2m_2$	0	$r_2/r$	$r_1/r$
7	$m_1m_1M_2M_2$	$(r_1/r)^2$	$2r_1r_2/r^2$	$(r_2/r)^2$
8	$m_1m_1M_2m_2$	0	$r_1/r$	$r_2/r$
9	$m_1m_1m_2m_2$	0	0	1

Table 1: Expected QTL genotype frequency conditional on genotypes of the flanking markers  $M_1$  and  $M_2$  in  $F_2$  populations with no double crossover.

### 2.3 Parameter Estimation

The likelihood function in 2 is complicated and it is difficult to obtain analytical estimates of all unknown parameters,  $\theta^* = (\mu_{aa}, \mu_{ab}, \mu_{bb}, \sigma, p_{ij})$ . In practice, however, the following can be done:

- "chromosome walking" technique: since  $M_1$  and  $M_2$  are known markers on the map with known recombination fraction  $r$ , start maximizing likelihood at marker  $M_1$ , then move along the chromosomal segment toward marker  $M_2$  in small steps (e.g. 1 cM), by iterative change of  $r_1$  and  $r_2$ , therefore assuming  $r_1$  and  $r_2$  known. Thus, the parameters  $p_{ij}$  ( $j = 1, 2, 3$ ) are known, because they are function of  $r_1$  and  $r_2$ .

For each position, maximize the likelihood and obtain *maximum likelihood* estimates (MLEs) for  $\theta = (\mu_{aa}, \mu_{ab}, \mu_{bb}, \sigma)$  by *expectation maximization* (EM) algorithm (Dempster *et al.* [3]). Next, a *likelihood ratio* (LR) statistic is constructed to test the hypotheses,

$$\begin{aligned}
 H_0 & : \mu_{aa} = \mu_{ab} = \mu_{bb} \text{ (no QTL in the interval)} \\
 H_a & : \mu_{aa} \neq \mu_{ab} \vee \mu_{aa} \neq \mu_{bb} \vee \mu_{ab} \neq \mu_{bb} \text{ (a QTL in the interval)}
 \end{aligned}$$

This statistic is distributed as a chi-square variable with two degrees of freedom.

- If there are more markers on the chromosome, repeat the procedure for each pair of markers. Thus, the hypotheses are tested at each position in an interval, for all intervals of the genome.
- The position with significantly largest LR statistic or LOD score (base 10 logarithm of

the likelihood ratio) is inferred to be the location of the gene, and the MLEs at the position are the estimates of the parameters.

The EM algorithm is a popular method for maximum likelihood analysis in *incomplete data* problems. The major reasons for its popularity are that its M-step involves only complete data maximum likelihood estimation and its convergence is stable (Meng *et al.* [6]). For QTL mapping problem, EM has been a powerful tool to obtain MLEs and applied by several researchers (Broman [1], Lander and Botstein [4], Zeng [9]). In this section, the EM algorithm is applied to derive the MLEs of equation 2 by treating the putative QTL as missing information.

Under the full model (equation 2), assume at iteration  $k + 1$  we have estimates of the parameters  $\hat{\boldsymbol{\theta}}^{(k)}$  by EM algorithm. Then, let  $z_{1j}$ ,  $z_{2j}$  and  $z_{3j}$  be unobserved variables,

$$z_{1j} = \begin{cases} 1, & \text{if the QTL genotype for individual } j \text{ is } aa; \\ 0, & \text{otherwise.} \end{cases}$$

$$z_{2j} = \begin{cases} 1, & \text{if the QTL genotype for individual } j \text{ is } ab; \\ 0, & \text{otherwise.} \end{cases}$$

$$z_{3j} = \begin{cases} 1, & \text{if the QTL genotype for individual } j \text{ is } bb; \\ 0, & \text{otherwise.} \end{cases}$$

The complete data likelihood function (equation 2) may be written as

$$L_c(\boldsymbol{\theta}|\mathbf{x}) = \prod_{j=1}^n \prod_{i=1}^3 \left[ p_{ij} \frac{1}{\sqrt{2\pi}\sigma} e^{\left[ \frac{-(y_j - \mu_i)^2}{2\sigma^2} \right]} \right]^{z_{ij}} \quad (3)$$

where  $\mathbf{x} = (\mathbf{y}, \mathbf{m}, \mathbf{z})$  is vector of complete data. The complete data log likelihood function is

$$\text{Log}L_c(\boldsymbol{\theta}|\mathbf{x}) = \sum_{j=1}^n \sum_{i=1}^3 \left[ z_{ij} \log(p_{ij}) - z_{ij} \log(\sqrt{2\pi}\sigma) - z_{ij} \frac{(y_j - \mu_i)^2}{2\sigma^2} \right]$$

In the E-step we compute the conditional expected complete data log likelihood function given the observed phenotypes, that is

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E \left[ \text{Log}L_c(\boldsymbol{\theta}|\mathbf{x}) | \mathbf{y}, \mathbf{m}, \boldsymbol{\theta}^{(k)} \right] \quad (4)$$

for such, it is enough to calculate

$$z_{ij}^{(k)} = \hat{E} \left[ z_{ij} | \mathbf{y}, \mathbf{m}, \boldsymbol{\theta}^{(k)} \right] = \frac{p_{ij} \phi(y_j; \mu_i^{(k)}, \sigma^{(k)})}{\sum_{i=1}^3 p_{ij} \phi(y_j; \mu_i^{(k)}, \sigma^{(k)})}$$

where  $\phi(y_j; \mu_i, \sigma)$  is the density function for normal distribution with mean  $\mu_i$  and standard deviation  $\sigma$ . Thus,

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right) = \sum_{j=1}^n \sum_{i=1}^3 \left[ z_{ij}^{(k)} \log(p_{ij}) - z_{ij}^{(k)} \log\left(\sqrt{2\pi}\sigma^{(k)}\right) - z_{ij}^{(k)} \frac{(y_j - \mu_i^{(k)})^2}{2(\sigma^{(k)})^2} \right]$$

In the M-step we find  $\hat{\boldsymbol{\theta}}^{(k)}$  to maximize the conditional expected log likelihood function  $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)$ , simply, by taking the derivatives of  $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)$ , such as

$$\begin{aligned} \frac{\delta Q}{\delta \mu_i} &= \sum_{j=1}^n z_{ij}^{(k)} \frac{(y_j - \mu_i)^2}{\sigma^2} \\ \frac{\delta Q}{\delta \sigma} &= \sum_{j=1}^n \sum_{i=1}^3 \left[ z_{ij}^{(k)} \frac{-1}{\sigma} + z_{ij}^{(k)} \frac{(y_j - \mu_i)^2}{\sigma^3} \right] \end{aligned}$$

The MLEs of  $\mu_{aa}$ ,  $\mu_{ab}$ ,  $\mu_{bb}$  and  $\sigma$  are given by

$$\hat{\mu}_{aa}^{(k+1)} = \frac{\sum_{j=1}^n z_{1j}^{(k)} y_j}{\sum_{j=1}^n z_{1j}^{(k)}} \quad (5)$$

$$\hat{\mu}_{ab}^{(k+1)} = \frac{\sum_{j=1}^n z_{2j}^{(k)} y_j}{\sum_{j=1}^n z_{2j}^{(k)}} \quad (6)$$

$$\hat{\mu}_{bb}^{(k+1)} = \frac{\sum_{j=1}^n z_{3j}^{(k)} y_j}{\sum_{j=1}^n z_{3j}^{(k)}} \quad (7)$$

$$\hat{\sigma}^{(k+1)} = \sqrt{\frac{1}{n} \sum_{j=1}^n z_{1j}^{(k)} \left(y_j - \hat{\mu}_{aa}^{(k+1)}\right)^2 + z_{2j}^{(k)} \left(y_j - \hat{\mu}_{ab}^{(k+1)}\right)^2 + z_{3j}^{(k)} \left(y_j - \hat{\mu}_{bb}^{(k+1)}\right)^2}$$

We iterate until the estimates converge.

## 2.4 Overall significance level

Determining the threshold of the test statistic is complicated. Many factors, such as the genome size, genetic map density and the proportion of missing data, may affect the distribution of the test statistic under the null hypothesis. The usual pointwise significance level

based on the chi-square approximation is inadequate because the entire genome is tested for the presence of a QTL (multiple testing).

Churchill and Doerge [2] proposed a *permutation procedure* to obtain an empirical threshold. Trait and marker data are separated and trait data are "shuffled". In doing so, the linkage relationship between QTL and the markers is broken. Thus, we are actually analyzing data under null hypothesis. This "shuffling" is repeated 1000-10000 times and each time the obtained value of the test statistics is stored. After obtaining many replicates, the values of these test statistics are sorted from the smallest to the largest. The 90th or 95th or 99th percentile of the empirical test statistics distribution is taken as a "threshold". If the value of the test statistic obtained from the original data exceeds this threshold, linkage is detected.

### 3 Application

Using genetic marker information, a number of statistical methods have been developed (Lander and Botstein [4], Liu [5]) to identify and estimate the positions of QTLs. The interval mapping is the method most used to identify QTLs. In this section, real data are used to detect and to locate QTLs by interval mapping.

To search genes controlling the levels of serum immunoglobulin in mice, a genetic intercross experimental was performed with the C57BL/6 and BALB/c mouse strains. We considered in this data set (our unpublished data) the phenotype of immunoglobulin IgM levels that was analyzed in serum of 136  $F_2$  female mice. The mice were also typed for 99 markers on the 19 chromosomes. We did not include markers on  $X$  chromosome in this analysis because in this study design only two genotypic classes can be found for chromosome  $X$ . We used the map distances (centiMorgan, cM) of the markers included in the database of the Whitehead Center for Genome Research (USA) (*Mouse Genome Informatics*, MGI).

We applied the interval mapping to these data. Genome-wide LOD thresholds were obtained by permutation tests (Churchill and Doerge [2]), using 1000 permutation replicates. The estimated 90%, 95% and 99% genome-wide LOD thresholds for the interval mapping were 2.88, 3.07 and 3.39, respectively. For each chromosome, the position with the highest LOD score is in table 2.

Figure 1 shows the maximum LOD score as a function of map position (cM) for 19 chromosomes. The position with significantly largest LOD score is accepted as the most likely position of the QTL. Thus, the results indicated evidence for one QTL within the region defined by the markers D13Mit115 and D13Mit266 (Lod score = 4.38), more exactly in the position 16 cM from the centromere of chromosome 13. This QTL was unknown. Figure 2 shows the detailed chromosome 13 LOD score plot. The estimates of parameters are  $\hat{\mu}_{aa} = 568$ ,  $\hat{\mu}_{ab} = 421$ ,  $\hat{\mu}_{bb} = 335$  and  $\hat{\sigma} = 192$ . The interval mapping was implemented by

software R/qtl (Broman [1]).

Chr	Pos. (cM)	LOD	$\hat{\mu}_{aa}$	$\hat{\mu}_{ab}$	$\hat{\mu}_{bb}$	$\hat{\sigma}$
1	43.1	1.53	521	397	432	201
2	52.5	1.31	537	425	384	202
3	2.4	1.11	488	415	378	203
4	66.0	0.64	389	420	471	204
5	0.0	0.86	480	397	455	203
6	34.5	1.42	453	387	493	202
7	37.0	1.45	370	467	380	201
8	43.0	0.88	483	418	387	203
9	57.0	1.81	444	456	327	200
10	59.0	0.94	470	442	369	203
11	10.0	0.40	402	417	469	205
12	16.0	1.21	400	407	505	202
13	16.0	4.38	568	421	335	192
14	40.0	2.10	536	405	391	199
15	29.2	0.86	469	397	473	203
16	33.9	0.33	420	442	389	205
17	23.2	0.81	384	460	400	204
18	47.0	1.05	366	454	442	203
19	53.0	0.81	394	457	386	204

Table 2: Peak of LOD scores and estimates of the respective parameters, obtained through the interval mapping.

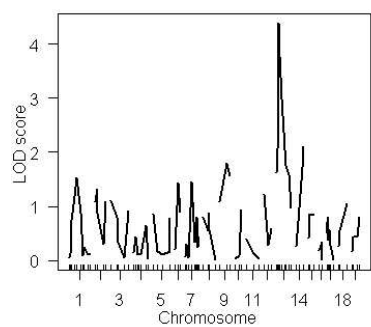


Figure 1: Plot for 19 chromosomes.

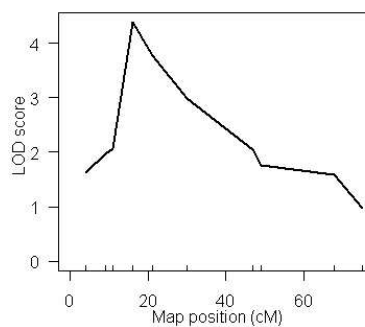


Figure 2: Plot for chromosome 13.

## 4 Discussion

The QTL mapping methods can be classified into three categories based on the number of markers used in the analysis. They are single-marker, two-marker and multiple-marker methods. With only one marker being used in QTL mapping, the effects are underestimated and the position cannot be determined by the t-test. In addition, the power is not high with the single marker likelihood ratio test (LRT).

The interval mapping is a two-marker method. With a fine scale genetic marker map throughout the genome, interval mapping can be performed at any position covered by markers, and the method can create a systematic strategy for detecting QTLs. Compared with the t-test, interval mapping has several advantages. These include:

- The probable position of the QTL can be inferred by the support interval;
- It is more powerful and needs fewer progeny to detect QTLs than the one-marker approach (Lander and Botstein [4], Zeng, [9]).

However, there are still some problems associated with the interval mapping:

- We cannot find out how many QTLs we have in the segment;
- If there is no QTL in the tested interval, the likelihood profile can still exceed the threshold if there is a QTL in another region on the chromosome ("ghost QTL").
- It is not efficient to use only two markers at a time to do the test, because the information from other markers is not used.

The application of the interval mapping detected only one QTL between the markers D13Mit115-D13Mit266 of chromosome 13. The maximum LOD score is 4.38 and the estimates of parameters are  $\hat{\mu}_{aa} = 568$ ,  $\hat{\mu}_{ab} = 421$ ,  $\hat{\mu}_{bb} = 335$  and  $\hat{\sigma} = 192$ . In the future, it will be useful to analyze more markers including those that can produce a more dense mapping of QTLs and better LOD score peak definitions. A biological role of this locus in controlling the level of serum IgMs is likely to have relevance in the homeostasis of the immune system.

## Acknowledgments

This work was supported by a fellowship from Fundação para a Ciência e Tecnologia (FCT), code FFRH/BD/5987/2001, reference 438.01.

We want to thank the Fundação para a Ciência e Tecnologia (FCT), Departamento de Estatística e Investigação Operacional (DEIO) da Faculdade de Ciências da Universidade de Lisboa, Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL) and Instituto Gulbenkian de Ciência (IGC).

## References

- [1] BROMAN, K. W. (2003). Mapping Quantitative Trait Loci in the Case of a Spike in the Phenotype Distribution, *Genetics*, **163**, 1169–1175.
- [2] CHURCHILL, G. A. and DOERGE, R. W. (1994). Empirical Threshold Values for Quantitative Trait Mapping, *Genetics*, **138**, 963–971.
- [3] DEMPSTER, A. P.; LAIRD, N. M. and RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm, *Journal of Royal Statistical Society, B*, **39**, 1–38.
- [4] LANDER, E. S. and BOTSTEIN, D. (1989). Mapping Mendelian Factors Underlying Quantitative Traits using RFLP Linkage Maps, *Genetics*, **121**, 185–199.
- [5] LIU, B. H. (1998). *Statistical Genomics, Linkage, Mapping and QTL Analysis*, CRC Press LLC, New York.
- [6] MENG, X. L. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika*, **2**, **80**, 267–278.
- [7] TODAY, J. M. (1961). Location of polygenes. *Nature*, **191**, 368–370.
- [8] WELLER, J. I. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers, *Biometrics*, **42**, 627–640.
- [9] ZENG, Z. B. (1994). Precision Mapping of Quantitative Trait Loci, *Genetics*, **136**, 1457–1466.



# A new limiting distribution for a statistical test for the homogeneity of two multinomial populations

Adelaide Valente Freitas \*      Miguel Pinheiro, José Luís Oliveira †  
Gabriela Moura, Manuel Santos ‡

## Abstract

In this paper, we propose a new statistic test  $T_m$  defined as the maximum term of the  $m - 1$  components of the partitioning of the Pearson chi-squared statistic  $X_P^2$  formulated by Kimball [6]. The limiting distribution of  $T_m$  is also derived.

Furthermore, we compare the results obtained from the statistics  $X_P^2$  and  $T_m$  to test the homogeneity of codon contexts of the complete ORFeome sequences of 3 yeast species, namely *Saccharomyces cerevisiae*, *Saccharomyces mikatae* and *Schizosaccharomyces pombe*. The statistic  $T_m$  has the advantage that their components identify the codon contexts responsible for eventual rejection of the homogeneity leading to results that are easier to interpret than those obtained using adjusted residuals as suggested by Haberman [4].

**Keywords:** contingency table, Pearson chi-squared statistic, partitioning, ORFeome, codon context, phylogenetic tree, residual analysis, extreme value distribution.

**AMS classifications:** 62P10, 60G70, 92D15.

## 1 Introduction

Consider two populations  $A_1$  and  $A_2$  each one described by a multinomial probability distribution with  $m$  mutually exclusive categories. Let  $(p_{11}, p_{21}, \dots, p_{m1})$  and  $(p_{12}, p_{22}, \dots, p_{m2})$  denote  $m$  unknown parameters in the populations  $A_1$  and  $A_2$ , respectively, where  $\sum_{i=1}^m p_{ij} = 1$ , for  $j = 1, 2$ . Note that, for each  $j$ , known  $m - 1$  parameters, the formulae  $\sum_{i=1}^m p_{ij} = 1$  allows to calculate the remaining one.

Suppose that a sampled data set cross-classified from the populations  $A_1$  and  $A_2$  are in a  $m \times 2$  contingency table.

We deal with the general goodness-of-fit problem for testing the null hypothesis  $H_0 : p_{i1} = p_{i2}$ ,  $i = 1, 2, \dots, m$ , against the alternative hypothesis  $H_1 : p_{i1} \neq p_{i2}$ , for at least one  $i$ .

---

\*Department of Mathematics, University of Aveiro, Portugal. E-mail: [adelaide@mat.ua.pt](mailto:adelaide@mat.ua.pt).

†IEETA, University of Aveiro, Portugal. E-mail: [monsanto@ieeta.pt](mailto:monsanto@ieeta.pt), [jlo@det.ua.pt](mailto:jlo@det.ua.pt).

‡Department of Biology, University of Aveiro, Portugal. E-mail: [gmoura@bio.ua.pt](mailto:gmoura@bio.ua.pt), [msantos@bio.ua.pt](mailto:msantos@bio.ua.pt).

One method used to test  $H_0$  is based on the Pearson chi-squared statistic  $X_P^2$ . As Lancaster [7] and Irwin [5] demonstrated (see [3] and references therein), the statistic  $X_P^2$  can be decomposed into  $m - 1$  mutually independent components. Each component is asymptotically distributed as a chi-square random variable with one degree of freedom (df) and each one is an appropriate statistic for testing a particular contrast. Kimball [6] derived one general formulae for these components.

We propose to use the  $T_m$  statistic test defined by the maximum term of these  $m - 1$  components instead of the sum of these  $m - 1$  components.

It is worth mentioning that Choulakian and Mahdi [2] had also proposed a statistic test defined by the maximum term of  $m - 1$  asymptotically chi-squared distributed random variables with one df, although assuming an alternative one-side hypothesis. However, Choulakian and Mahdi's representation does not define a partition of  $X_P^2$ .

Applying classical Extreme Value results<sup>1</sup>, we derive the limiting distribution of  $T_m$ , under linear normalization, as both  $m$  and the total sample size tend to  $+\infty$ .

The rest of the paper is organized as follows. The formulae of computation of the statistic  $T_m$  and its limiting distribution are given in the next section. In Section 3 we conduct a simulation study to see how the statistics  $X_P^2$  and  $T_m$  work. In the last section we apply Pearson's statistic  $X_P^2$  and the statistic  $T_m$  to analyze the homogeneity associated to gene primary structure on three yeast species: *Saccharomyces cerevisiae*, *Saccharomyces mikatae* and *Schizosaccharomyces pombe*. The real data set on the complete coding sequences (OR-Feome) of the genome of these three yeasts involves a large number of observations and can be analyzed under several categorized random variables each having several mutually exclusive categories. Attending to some codon context rules in the gene primary structure unveiled by Moura et al.[9], we test the homogeneity of the distribution of the codons for which the nucleotide 3'-neighbor is Adenine for *S. cerevisiae* versus *S. mikatae* and for *S. cerevisiae* versus *S. pombe*. We compare the results obtained when we use the statistics  $X_P^2$  and  $T_m$ . Considering the statistic  $T_m$  we identify the categories responsible for the rejection of the null hypothesis of homogeneity leading to results that are easier to interpret than when residual analysis is applied to identify the cells responsible for a significant chi-square value.

## 2 Theoretical results

Let  $N_{ij}$  be a random variable representing the number of observations falling into category  $i$  that are sampled from the population  $A_j$ , for  $i = 1, 2, \dots, m$  and  $j = 1, 2$ , and let  $N_{i\cdot} = N_{i1} + N_{i2}$ . Let  $n_{ij}$  and  $n_{i\cdot}$  denote one realization of  $N_{ij}$  and  $N_{i\cdot}$ , respectively,  $n_{\cdot 1}$  and  $n_{\cdot 2}$  the sample size obtained from the populations  $A_1$  and  $A_2$ , respectively, and  $n = n_{\cdot 1} + n_{\cdot 2}$  the total sample size.

---

<sup>1</sup>In, for instance, Leadbetter et al. [8].

The Pearson chi-squared statistic

$$X_P^2 = \sum_{j=1}^2 \sum_{i=1}^m \frac{(N_{ij} - \frac{n_{.i}N_{i.}}{n})^2}{\frac{n_{.i}N_{i.}}{n}},$$

can be partitioning into  $m - 1$  asymptotically chi-squared components each having one df. Kimball [6] has derived convenient formulae to apply the partitioning method of  $X_P^2$  showed by Lancaster [7] and Irwin [5]. Kimball partitioned  $X_P^2$  into  $m - 1$  asymptotically independent chi-squared statistics  $X_1^2, X_2^2, \dots, X_{m-1}^2$ , deriving the following general formulae for  $X_i^2$ ,  $i = 1, 2, \dots, m - 1$ ,

$$X_i^2 = \frac{n^2 \left( N_{i+1,2}(N_{11} + N_{21} + \dots + N_{i1}) - N_{i+1,1}(N_{12} + N_{22} + \dots + N_{i2}) \right)}{n_{1.}n_{2.}N_{i+1.}(N_{1.} + N_{2.} + \dots + N_{i.})(N_{1.} + N_{2.} + \dots + N_{i+1.})}.$$

The statistic  $X_i^2$  is the Pearson chi-squared statistic used to test the null hypothesis  $H_{0,i} : p_{i+1,1} = p_{i+1,2}$  and it is calculated from the  $2 \times 2$  contingency table constructed from the original  $m \times 2$  contingency table in the following way: its first row is obtained collapsing the first  $i$  categories of the  $m \times 2$  contingency table, and its second row is the  $(i + 1)th$  row of the  $m \times 2$  contingency table.

Note that the null hypothesis  $H_0$  of homogeneity between  $A_1$  and  $A_2$  can be written, equivalently, as a intersection of  $m - 1$  null hypotheses, each one corresponding to the homogeneity of two sub-populations of  $A_1$  and  $A_2$  with two categories: one is common with the original populations and the other is obtained by collapsing the remaining in some appropriate way. In particular, we take

$$H_0 = \bigcap_{i=1}^{m-1} H_{0i}.$$

Thus, the statistic test

$$T_m = \max(X_1^2, X_2^2, \dots, X_{m-1}^2),$$

may also be used to test the null hypothesis  $H_0$  of homogeneity between  $A_1$  and  $A_2$ .

The statistic  $T_m$  represents the maximum term of  $m - 1$  independent random variables each asymptotically chi-squared distributed with one df, as  $n \rightarrow +\infty$ . Hence,

$$\lim_{n \rightarrow +\infty} P(T_m \leq t) = F^{m-1}(t), \quad t \in \mathbb{R},$$

where  $F$  is the chi-squared distribution function with one df.

Results well known from Extreme Value Theory allows for obtaining the limiting distribution of  $T_m$ , under convenient normalization, as  $m, n \rightarrow +\infty$ . Namely, since the chi-squared distribution function belongs to the domain of attraction of a Gumbel distribution<sup>2</sup>, there exist

---

<sup>2</sup>See, for instance, Leadbetter et al. [8]

sequences of normalization constants  $\{a_m > 0 : m \in \mathbb{N}\}$  and  $\{b_m > 0 : m \in \mathbb{N}\}$  such that

$$\begin{aligned} \lim_{m \rightarrow +\infty} \lim_{n \rightarrow +\infty} P(T_m \leq a_{m-1}t + b_{m-1}) &= \lim_{m \rightarrow +\infty} F^{m-1}(a_{m-1}t + b_{m-1}) \\ &= \exp(-\exp(-t)) , \forall t \in \mathbb{R}. \end{aligned} \quad (1)$$

One can take  $b_m$  such that  $1 - F(b_m) = \frac{1}{m}$  and  $a_m = \frac{1}{mF'(b_m)}^3$ .

In practical situations, when  $n$  becomes large we reject  $H_0$ , at level of significance  $\alpha$ , if the observed value of  $T_m$  is greater than  $t_m$ , where  $F^{m-1}(t_m) = 1 - \alpha$ . If both  $n$  and  $m$  become large, then we can take the limiting behavior in (1) and, we reject  $H_0$ , at level of significance  $\alpha$ , if the observed value of  $T_m$  is greater than  $t_m^* = a_{m-1}(-\ln(-\ln(1 - \alpha))) + b_{m-1}$ . In this case we can identify each of the categories  $i$  which have a  $X_i^2$  greater than  $t_m$ , and we can plausibly say that each of these shows a deviation from the null hypothesis.

### 3 Simulation study

For  $m = 5, 10, 20, 50, 100$  we simulated 10000  $m \times 2$  contingency tables with equally likely categories and  $n_{.1} = n_{.2} = 10000$ .

For each simulated contingency table we tested the null hypothesis of homogeneity using both Pearson's statistic  $X_P^2$  and the statistic  $T_m$ . The observed p-values are available at <http://www.bio.ua.pt/genomica/lab/>. In Table 1 are summarized some results for  $m = 5$  and  $m = 100$ .

For levels of significance  $\alpha = 0,05; 0,01; 0,001$  both the statistics  $X_P^2$  and  $T_m$  reject the null hypothesis of homogeneity for a small percentage of simulated contingency tables. For example (see Table 1), that percentage is 5,26% ( $= (318 + 208)100\%$ ) for  $m = 5$ ,  $\alpha = 0,05$ , when one used the statistic  $T_m$  to test  $H_0$ , and it is 5,03% when one consider the statistic  $X_P^2$ .

Our simulation results showed that the percentage of contingency tables for which the statistic  $T_m$  leads to the rejection of  $H_0$ , but not the statistic  $X_P^2$ , is similar to the percentage of contingency tables for which the statistic  $X_P^2$  leads to the rejection of  $H_0$  and no rejection when the statistic  $T_m$  is used.

The advantage of the statistic  $T_m$  is to allow identification of the categories responsible for the rejection of  $H_0$  with easy interpretation (see Section 4).

### 4 Application to real data

Using the software ANACONDA (available at <http://www.bio.ua.pt/genomica/lab/>) we constructed maps of codon pairs context for *S. cerevisiae*, *S. mikatae* and *S. pombe*. These

---

<sup>3</sup>See Reiss and Thomas [10], p. 52.

	$m = 5$		$m = 100$		$\alpha$
	$p_{T_m} \leq \alpha$	$p_{T_m} > \alpha$	$p_{T_m} \leq \alpha$	$p_{T_m} > \alpha$	
$p_{X_p^2} \leq \alpha$	318	185	76	370	0,05
$p_{X_p^2} > \alpha$	208	9289	386	9168	
$p_{X_p^2} \leq \alpha$	41	48	3	85	0,01
$p_{X_p^2} > \alpha$	46	9865	78	9827	
$p_{X_p^2} \leq \alpha$	1	3	0	9	0,001
$p_{X_p^2} > \alpha$	6	9989	8	9983	

Table 1: For each  $m$  and  $\alpha$  considered, is indicated the number of  $m \times 2$  contingency tables, between 10000 simulated with  $m$  equally likely categories, with p-value greater than  $\alpha$  and lower than  $\alpha$ . The p-values were calculated using Pearson’s statistic  $X_p^2$  and the statistic  $T_m$ .  $p_{T_m} = P(T_m > T_{mObs}|H_0)$  where  $T_{mObs}$  represents the observed value of the statistic  $T_m$  for each simulated contingency table. A similar notation was used for  $p_{X_p^2}$ .

maps clearly show that each codon has a set of preferred 3<sup>1</sup>-codon neighbors and rejects a set of other codons, indicating that codon context is highly biased in these three yeast species. For *S. cerevisiae* an important feature of gene primary structure that modulates mRNA decoding accuracy is codon pairs that have the nucleotide A (Adenine) in the first position of the second codon <sup>4</sup>.

Since *S. cerevisiae* and *S. mikatae* have diverged from each other 5 million years ago only but *S. pombe* has diverged 420 million years from the other two (Figure 1), one was expecting that the codon context rules for the latter species would be rather different from those of the former two species and very similar for *S. cerevisiae* and *S. mikatae*.

We tested this working hypothesis using the statistics  $X_p^2$  and  $T_m$  and their limiting distributions herein established. We used the complete ORFeome sequences of *S. cerevisiae*,

*S. mikatae* and *S. pombe* genomes and we tested the homogeneity of the codon contexts  $XYZ - A1$ , where each of  $X, Y, Z$  is the nucleotide A, C (Cytosine), G (Guanine), and T (Thymine), respectively, on the corresponding position in the first codon of a codon-pair and  $A1$  indicates that the first nucleotide of the second codon of the codon-pair is  $A$ . Note that all  $4^3$  codons except the three stop codons belong to the set of codon contexts  $XYZ - A1$ . We summarized the data in two  $61 \times 2$  contingency tables and we took the categories alphabetically ordered. The contingency tables and all the intermediate and final computations are available at the URL <http://www.bio.ua.pt/genomica/lab/>.

In Table 2 are summarized the observed values for the 16 first components  $X_i^2$  and the

<sup>4</sup>For example, XXU-AYY. For details see Moura et al [9].

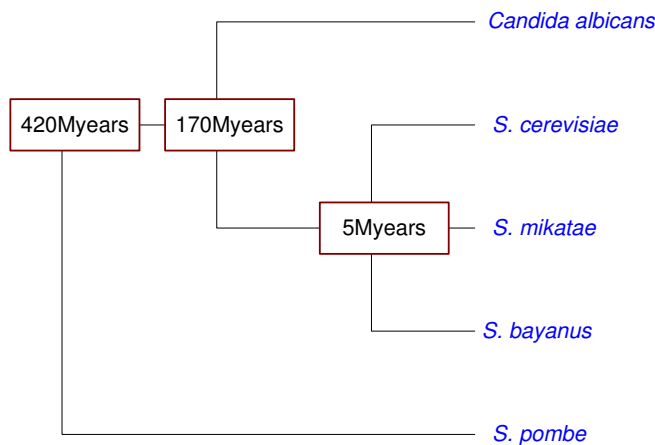


Figure 1: The phylogenetic tree showing the divergency times of several yeast species used in this study.

observed adjusted residual values<sup>5</sup>.

The observed values for the statistics test are: for *S. cerevisiae* vs *S. mikatae*  $X_p^2 = 326,49$  and  $T_m = 37,70$  and for *S. cerevisiae* vs *S. pombe*  $X_p^2 = 28196,3$  and  $T_m = 3646,4$ . For  $m = 61$ , we have  $a_{m-1} = 1,756$  and  $b_{m-1} = 5,731$  and,  $t_m = 18,54$  and  $t_m^* = 17,86$ , for  $\alpha = 0,001$ . Then, both the statistics  $X_p^2$  and  $T_m$  lead to reject  $H_0$  at level of significance 0,001. For *S. cerevisiae* vs *S. pombe* we obtained values of the statistics test clearly greater. Considering the  $m - 1 = 60$  components of the statistic  $T_m$  and the adjusted Pearson residual values we can identify codon contexts responsible for a significant value for the statistics  $T_m$  and  $X_p^2$ , respectively.

When we tested the homogeneity between *S. cerevisiae* and *S. mikatae* we observed that the components  $X_i^2$ ,  $i = 1, 2, \dots, 60$ , are not so large. This suggests that the context rule  $XYZ - A1$  is *almost* identical for these two yeasts. Indeed, only the contexts AAG-A, GAA-A, GAG-A and GAU-A contribute for the rejection of  $H_0$ , at level of significance 0,001. However, this conclusion is not so clear when the analysis of residuals is applied. Some statistically significant values of the adjusted Pearson residual are not associated to one row but only to one cell by row in the  $61 \times 2$  contingency table. Our approach leads to results that are easier to interpret than those obtained from the analysis of residuals. The adjusted residuals values are more appropriate to complement testing of independence.

For the comparison of *S. cerevisiae* vs *S. pombe*, we obtain large values for the statistics  $X_p^2$  and  $T_m$ , and its components, clearly leading to the rejection of the homogeneity of the codon context rule  $XYZ - A1$  between these two yeasts species, thus confirming the divergence of these yeasts observed in the phylogenetic tree (Figure 1).

<sup>5</sup>For its formulae, see Everitt [3], p. 47, or Agresti [1], p. 81.

codon contexts $XYZ - A1$	$X_i^2$		adjusted residual			
	<i>Cer</i> vs <i>Mik</i>	<i>Cer</i> vs <i>Pom</i>	<i>Cer</i>	<i>Mik</i>	<i>Cer</i>	<i>Pom</i>
<i>AAA - A</i>			4,24	-5,09	-5,15	5,15
<i>AAC - A</i>	9,256	494,5	-0,63	0,76	24,0	-24,0
<i>AAG - A</i>	37,707	730,4	-5,57	6,69	38,8	-38,8
<i>AAU - A</i>	2,568	23,04	-2,32	2,79	22,4	-22,4
<i>ACA - A</i>	9,039	3,101	2,78	-3,10	15,2	-15,2
<i>ACC - A</i>	1,709	5,723	1,13	-1,36	14,8	-14,8
<i>ACG - A</i>	0,432	39,13	0,59	-0,71	2,58	-2,58
<i>ACU - A</i>	0,831	627,1	-1,05	1,26	-1,31	1,31
<i>AGA - A</i>	13,287	857,2	-4,11	4,95	42,5	-42,5
<i>AGC - A</i>	2,832	100,3	1,31	-1,57	0,79	-0,79
<i>AGG - A</i>	3,191	190,3	1,52	-1,82	22,9	-22,9
<i>AGU - A</i>	17,838	85,86	-4,67	5,62	2,57	-2,57
<i>AUA - A</i>	0,0005	81,19	-0,64	0,76	28,5	-28,5
<i>AUC - A</i>	5,209	190,4	-3,06	3,68	5,11	-5,11
<i>AUG - A</i>	0,943	107,4	-1,85	2,22	-15,5	15,5
<i>AUU - A</i>	13,627	935,9	2,94	-3,53	14,6	-14,6
⋮						
<i>UUU - A</i>	5,979	569,0	2,45	-2,93	-2,37	2,37

Table 2: The observed values for the 16 first components  $X_i^2$  of the statistic  $T_m$  and the observed adjusted residual values associated to the statistic  $X_P^2$ , when is tested the homogeneity of the codon contexts  $XYZ - A1$  between *S. cerevisiae* and *S. mikatae* and between *S. cerevisiae* and *S. pombe*.

This type of analysis can be useful for comparing mRNA decoding rules between different genomes.

## Acknowledgments

We are thankful to the referee for his useful suggestions. We thank FCT (Plurianual Funding - R&D Unit 'Mathematics and Applications', University of Aveiro and POCTI/BME/39030/2001) and IEETA.

## References

- [1] AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd Edition, Wiley.
- [2] CHOULAKIAN, V. and MAHDI, S. (2001). A new statistic for the analysis of association between trait and polymorphic marker loci, *Math. Biosciences*, **164**, 139–145.
- [3] EVERITT, B.S. (2000). *The Analysis Of Contingency Tables*, 2nd Edition, Wiley.
- [4] HABERMAN, S.J. (1973). The analysis of residual in cross-classified tables, *Biometrics*, **29**, 205–220.
- [5] IRWIN, J.O. (1949). A note on the subdivision of  $\chi^2$  into components, *Biometrika*, **36**, 130–134.
- [6] KIMBALL, A. W. (1954). Short-cut formulae for the exact partition of chi-square in contingency tables. *Biometrics*, **10**, 452–458.
- [7] LANCASTER, H.O. (1949). The derivation and partition of  $\chi^2$  in certain discrete distributions, *Biometrika*, **36**, 117–129.
- [8] LEADBETTER, M.R.; LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes And Related Properties Of Random Sequences And Processes*. Springer-Verlag.
- [9] MOURA, G.; PINHEIRO, M.; SILVA, R.; MIRANDA, I.; AFREIXO, V.; DIAS, G.; FREITAS, A.; OLIVEIRA, J.L. and SANTOS, M. (2005). Comparative context analysis of codon pairs on an ORFeome scale. *Genome Biology*, **6**, R28.
- [10] REISS, R.-D. and THOMAS, M. (1997). *Statistical Analysis Of Values With Applications To Insurance, Finance, Hydrology And Other Fields*. Birkhäuser.