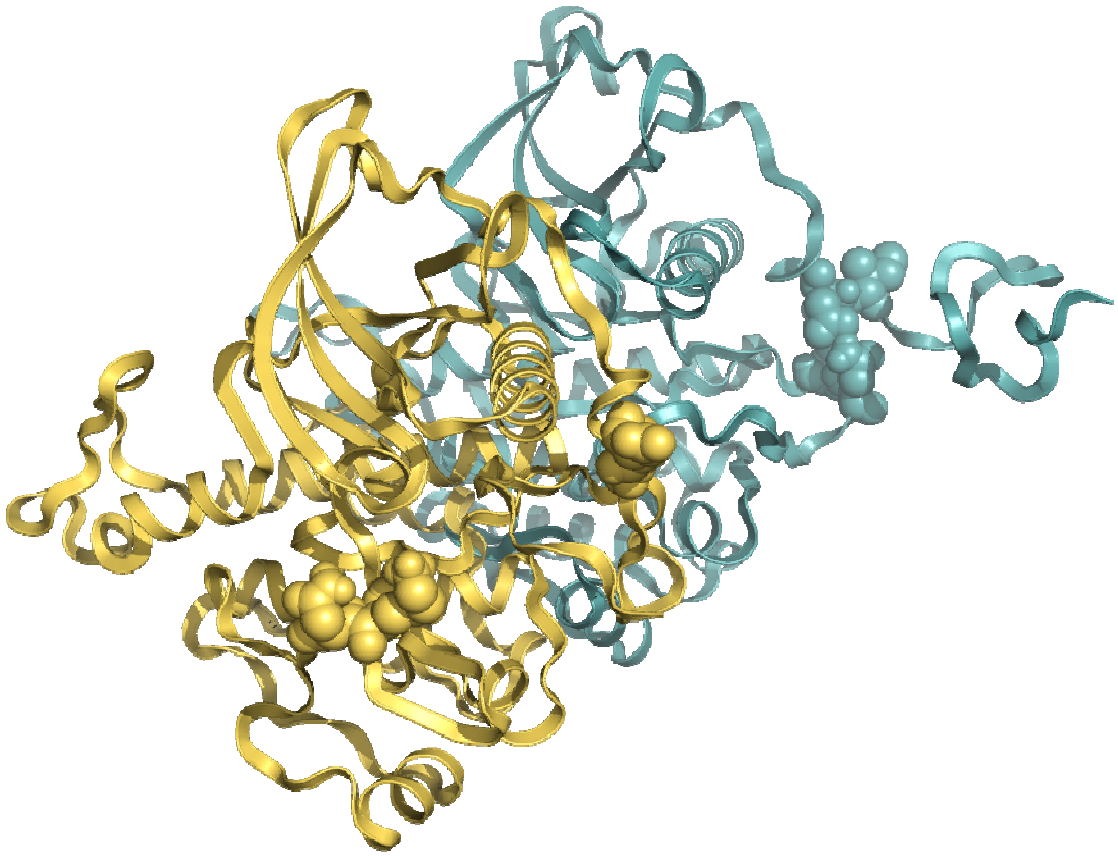
SMPGD 2014

Statistical Methods for Post-Genomic Data – 10th Workshop

January 23-24, 2014 – UPMC – Paris, France

BOOK OF ABSTRACTS



SUMMARY

Thursday, January 23, 2014

8h55 Opening – a word of introduction

9h05 Arnak Dalalyan (ENSAE / CREST, Université Paris-Est)
Big Data

10h **Submitted** (3 talks)

- **10h Julien Chiquet** (AgroParisTech, Laboratoire Statistique et génomes)
  Multi-trait genomic selection via multivariate regression with structured regularization
- **10h15 Dave duVerle** (National Institute of Advanced Industrial Science and Technology)
  A Regularisation Path-Following Approach for Discovering Interactions in High-Dimensional Survival Data
- **10h30 Franck Picard** (LBBE - CNRS)
  On the robustness of the Generalized Fused Lasso to prior specifications

10h45 Coffee break

11h30 Statistical Genomics session - Bertrand Servin (INRA)

- **11h30 Simon Boitard** (Museum National d'Histoire Naturelle)
  Inferring the past dynamics of effective population size using genome wide molecular data
- **11h55 Anne-Louise Leutenneger** (INSERM & Univ. Paris Diderot)
  Mapping genes in consanguineous and isolated populations in the era of high throughput sequencing
- **12h20 Christèle Robert-Granie** (INRA Toulouse)
  Integration of genomic information into genetic evaluation model : Is it a good statistical model?

12h45 Lunch break

14h15 Cécile Ané (Department of Statistics and Botanics, University of Wisconsin-Madison) *Probabilistic approaches for detecting and locating whole genome duplications*

15h10 Submitted (3 talks)

- **15h10 Julien Gagneur** (Gene Center of the LMU, Munich)
  Of cis, trans, and feedback regulation: Dissecting local regulation in yeast
- **15h25 Wenjia Wang** (Pharnext)
  A New Gene-Based test of Association Using Extended Rasch Models
- **15h40 Eric Frichot** (TIMC-IMAG)
  Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models

15h55 Phylogeny session - Nicolas Lartillot (LBBE UCB Lyon 1)

- **15h55 Alessandra Carbone** (CQB, Université Pierre et Marie Curie - CNRS)
  Coding of evolutionary pathways in proteins: from sequence to function
- **16h20 Gergely J. Szöllősi** (ELTE-MTA Biophysics Research Group - Eötvös University)
  Efficient Exploration of the Space of Reconciled Gene Trees

16h45 Poster Presentation & Cocktail

Friday, January 24, 2014

9h05 **David T. Jones** (Computer Science, University College London)
The future prospects for de novo protein structure prediction from evolutionary information

10h **Submitted** (3 talks)

- **10h Korbinian Strimmer** (University of Leipzig)
  Identifying Differentially Expressed Proteins by a Binary Threshold Model
- **10h15 Giulia Malaguti** (UMR 168 CNRS-UPMC)
  The Causal Mediation Analysis in Genomic Data
- **10h30 Bogdan Mirauta** (CQB, Université Pierre et Marie Curie - CNRS)
  An RNA-Seq read count emission model for transcriptional landscape reconstruction with state-space models

10h45 Coffee break

11h15 **Metabolism session - Daniel Kahn** (LBBE – INRA)

- **11h20 David Vallenet** (Laboratoire de génomique comparative, Génoscope)
  Enzyme survey and how to find new ones
- **11h50 Christoph Kaleta** (Theoretical Systems Biology, Friedrich-Schiller Universität Jena)
  Tuned for speed – Elucidation of strategies for rapid metabolic adaptations in prokaryotes
- **12h20 Frank J. Bruggeman** (Systems Bioinformatics, VU University)
  Constraints, adaptability and optimality of metabolic networks

12h50 Lunch break

14h15 **Frédéric Austerlitz** (CNRS/Museum National d'Histoire Naturelle - Université Paris Diderot)
Inference of past historical events using Approximate Bayesian Computation and Markov Chain Monte Carlo methods on population genetics data sets

15h10 **Submitted** (2 talks)

- **15h10 Fanny Pouyet** (Laboratoire de Biométrie et Biologie Evolutive)
  Evolution of Codon Usage Bias in E. coli
- **15h25 Alexander Schliep** (Department of Computer Science, Rutgers University)
  Efficient statistical computations on genome-scale data using reduced representations

15h40 Coffee break

16h15 **Submitted** (3 talks)

- **16h15 Louis Dijkstra** (Centrum Wiskunde & Informatica)
  Testing 2 x 2 Association with Uncertain Classification
- **16h30 Guillem Rigaill** (Unité de recherche en génomique végétale - INRA)
  A fast homotopy algorithm for a large class of weighted classification problems and application to phylogeny
- **16h45 Mathias Fuchs** (Medizinische Fakultät, Ludwig-Maximilians-Universität München)
  The leave-p-out estimator of the prediction error as a U-statistic and its asymptotic tests

Keynote speakers

Big Data

Arnak Dalalyan, ENSAE/Université Paris-Est, France

In this talk, we begin by reviewing some results on popular sparse estimation methods based on L1-relaxation. These methods, such as the Lasso and the Dantzig selector, require the knowledge of the variance of the noise in order to properly tune the regularization parameter. This constitutes a major obstacle in applying these methods in several frameworks-such as time series, random fields, inverse problems-for which the noise is rarely homoscedastic and its level is hard to know in advance.

In the second part of the talk, we will present a new approach to the joint estimation of the conditional mean and the conditional variance in a high-dimensional regression setting with heteroscedastic noise. An attractive feature of the proposed estimator is that it is efficiently computable even for very large scale problems by solving a second-order cone program (SOCP). We will present theoretical analysis and numerical results assessing the performance of the proposed procedure.

Probabilistic approaches for detecting and locating whole genome duplications.

Cécile Ané, University of Wisconsin – Madison, USA

Whole Genome Duplications (WGDs) can be difficult to detect when they are old and when synteny has been disrupted by genome rearrangements. To test the presence of WGDs on a species phylogeny, I will present two methods which do not require synteny information and build strength from the phylogenetic framework. They rely on a probability model for the evolution of gene families on a species tree with WGDs. Both methods use multiple gene families across multiple species. One method relies on aligned molecular sequences and the other simply uses information on gene counts. We assessed their performance with simulations and on a benchmark yeast dataset, where we recover strong evidence for a well-established WGD and a low retention rate of duplicated genes after this WGD.

The future prospects for de novo protein structure prediction from evolutionary information

David Jones, University College London, United Kingdom

Despite great strides in *de novo* prediction of protein structure from amino acid sequence over the past decade, there seem to be rapidly diminishing returns in applying these methods to real problems. At the domain level, certainly, *de novo* prediction tends to be of very limited use, particularly for globular proteins. It's now rare to find interesting protein domains without any homologues of known 3-D structure, and even in cases where no templates can be found, template-free modelling results turn out to be little better than they were 10 years ago. More worryingly, there are no reliable ways of knowing if *de novo* methods have even been successful without actually solving the structure experimentally. In the last few years, developments in contact prediction based on evolutionary information have generated a lot of excitement, but it is still unclear as to how useful these methods will ultimately prove to be.

In this talk I will be describing some of our recent work in this area, where we have made use of Sparse Inverse Covariance Estimation methods to predict inter-residue contacts in proteins, and developed algorithms to derive useful 3-D models from this information. We have recently completed a large scale test of these methods which has highlighted some of the limitations of these approaches, which I will discuss. Overcoming these limitations will be vital if they are to be of significant practical use in the future.

Inference of past historical events using Approximate Bayesian Computation methods on population genetics data sets

Frédéric Austerlitz, CNRS/Museum National d'Histoire Naturelle - Université Paris Diderot, France

New computer-intensive estimation techniques such as Approximate Bayesian Computation (ABC) and Monte Carlo Markov chains (MCMC) allows inferring unknown parts of the history of species from contemporary population genetics data. I will illustrate these possibilities with several examples. First, I will talk about a set of human populations from Western Central Africa, consisting of hunter-gatherer Pygmy populations and neighbouring non-Pygmy populations, genotyped for several kinds of genetic markers. Using ABC techniques, we could infer the history of splitting and admixture between these different groups. We could also identify sex-specific demographic processes. The second example that I will mention is the harbour porpoise (Phocoena phocoena) population from the Black Sea. Using again ABC techniques, we showed that this population underwent a strong expansion around 5000 years ago, probably as a result of the reconnection of the Black Sea with the Mediterranean Sea, but that it underwent also a drastic decline around 50 years ago, which can be linked with the intensive hunting of cetaceans performed at that time. Finally I will talk about a study on worldwide human populations, in which by applying MCMC methods on a large set of populations with different lifestyles (farmers, herder and hunter-gatherers), we were able to show that these lifestyles strongly impacted the expansion patterns of these populations. These examples illustrate well how ABC and MCMC methods allow inferring precious information on the history of populations for which archeological records are not available.

Invited session 1

Statistical Genomics

Organizer: Bertrand Servin – Laboratoire de Génétique Cellulaire
INRA, Toulouse, France

Inferring the past dynamics of effective population size using genome wide molecular data

Simon Boitard, Origine, Structure et Evolution de la Biodiversité, Museum National d'Histoire Naturelle.

Inferring the effective size of a given population, and its eventual expansions or reductions in the past, from genetic data, is a long standing question in population genetics. Due to the complexity and the high dimension of the mathematical models that are used in this context, exact inference is impossible and the most popular inference mehtods are based on numerical approaches as Markov Chain Monte Carlo, Importance Sampling or Approximate Bayesian Computaion (ABC).

Until recently, these methods were designed for data sets including a small number of independent markers or non recombining DNA sequences. However, the spectacular progress of genotyping and sequencing technologies during the last decade has enabled the production of high density genome wide data in many species, so new statistical methods are needed to take benefit of this new type of data.

In this study we present an ABC approach for inferring the past effective size of a single population. This approach is based on coalescent simulations and on the use of a large number of summary statistics related to allele frequencies and linkage disequilibrium. We illustrate the performance of this approach using cross validation. We compare different ABC strategies and discuss the influence of the different summary statistics.

We finally apply this method to a set of 25 bovine sequences from the Holstein breed and compare our results with those obtained by the Pairwise Sequentially Markovian Coalescent approach of Li and Durbin (2011).

Mapping genes in consanguineous and isolated populations in the era of high throughput sequencing

Anne-Louise Leutenegger, University Paris Diderot and Inserm U946.

Great progress has been made in the identification of genetic variants for complex human traits thanks to genome-wide association studies (GWAS). However, part of the heritability remains unknown for most complex diseases, suggesting that some genetic factors remain to be discovered. The study of rare variants could help to characterize more exhaustively the genetic background of complex traits and is now facilitated by recent advances in sequencing technologies. However, those technologies remain too expensive for many academic research groups to sequence a large number of subjects from general populations, which is necessary to attain sufficient power to detect effect of such variants.

Population isolates have well-documented characteristics that can aid to identify rare variants associated with complex traits, namely reduced phenotypic, environmental and genetic heterogeneity. Besides, alleles that are rare in general populations may have become more frequent in those isolated populations, which could facilitate their identification.

We will present the strategy that has recently been proposed to study complex traits in isolated populations: selection of subset of individuals for sequencing, imputation of the sequence data in the remaining individuals to obtain sequence data on the entire population, and finally, association analysis with traits of interest.

Integration of genomic information into genetic evaluation model: is it a good statistical model?

Christèle Robert-Granié, Station d'Amélioration Génétique des Animaux (SAGA) - UR631, INRA-Toulouse

The rapid evolution in sequencing and genotyping raises new challenges in the development of methods of selection for livestock, also called genomic selection. With this genomic information, it is now possible to estimate breeding values of selection candidates at birth without waiting for phenotypic data collection. Genomic selection requires the creation of a reference population made of genotyped animals with precise phenotypes. Genomic evaluations consist in predicting phenotypes in this reference population as the sum of molecular markers. The main methodological challenge is the large number of effects to estimate, usually much larger than the number of available phenotypes. We briefly describe and compare the various families of proposed methods: genomic BLUP (Best Linear Unbiaised Prediction) based on relationship computed from marker information, Bayesian methods, variable selection methods and a single step method which combined pedigree and genomic data and raw phenotypes. The precision of genomic selection is done via cross validation among the youngest animals of the reference population. Several parameters (size of the reference population, relationship between selection candidates and the reference population, genetic parameters, etc) have a significant impact on the efficiency of genomic selection methods. Some results on real data will be presented.

christele.robert-granie@toulouse.inra.fr

Invited session 2

Phylogeny

Organizer: Nicolas Lartillot – LBBE « Biométrie et Biologie Évolutive » UCB Lyon 1

Coding of evolutionary pathways in proteins: from sequence to function

Alessandra Carbone, Laboratoire de Biologie Computationnelle et Quantitative Université Pierre et Marie Curie – CNRS

Today, networks of protein interactions do not describe protein-protein partnership at a residue level. This information is necessary to control protein behavior though. We shall present an approach based on sequence analysis to detect important information on protein binding sites and on mechanical and allosteric properties at the residue level. We shall use a fine reading of the conservation and co-evolution signals between residues in the protein sequences. This information is encoded in the tree topology of the distance tree associated to evolved sequences observable today. We shall present two methods, one applicable to protein families of variable divergence and the other to very conserved protein families.

Efficient Exploration of the Space of Reconciled Gene Trees

Gergely J. Szöllősi, ELTE-MTA Biophysics Research Group – Eötvös University

Gergely J. Szöllősi[1], Wojciech Rosikiewicz[2], Bastien Boussau[3], Eric Tannier[3] and Vincent Daubin[3]

As molecular phylogeneticists, we infer gene trees based on sequence information. Unfortunately, sequences alone contain limited signal, and as a result phylogenetic reconstruction almost always involves choosing between statistically equivalent or weakly distinguishable relationships. Although each homologous gene family has its own unique story, they are all related by a shared species history, which could be helpful for gene tree inference. We have recently published a probabilistic reconciliation model, which describes the relationships between a gene tree and a species tree as a series of events, such as duplication, transfer and loss, speciation and extinction (Szöllősi et al. Syst. Biol. 2013). We now propose an efficient way to integrate sequences and reconciliation information in the inference of gene trees.

To design a species tree aware method for reconstructing gene phylogenies, the space of reconciled gene trees must be explored using information from both a model of sequence evolution and a reconciliation model. Such an exploration can be tedious with classical approaches. To circumvent this problem, we present a general probabilistic approach to exhaustively explore all reconciled gene trees that can be amalgamated as a combination of clades observed in a sample of gene trees. For a sample derived from the posterior distribution of trees obtained from a bayesian MCMC analysis, this approach provides an accurate approximation of gene tree likelihood.

We demonstrate using both simulations and biological sequences that gene phylogenies reconstructed using the joint likelihood are dramatically more accurate than those reconstructed using sequences alone. In fact, we find that even using a simplistic model of sequence evolution, the joint reconstruction yields significantly more accurate gene trees than the sequence-based inference with the complex model used in simulations. Considering 1099 homologous gene famillies from 36 genomes of cyanobacteria we find that the majority of phylogenetic discord results from errors in sequence based reconstruction that can be corrected using information aggregated across gene families by a putative species tree. The result is a striking reduction in apparent phylogenetic discord, with resp. 24%,$59% and 46% percent reductions in the mean numbers of duplications, transfers and losses per gene family.

Our probabilistic method overcomes a fundamental limitation of recent parsimony based methods to improve gene trees given a putative species tree (David and Alm Nature 2011,Wu et al. Syst. Biol. 2013) by not having to rely on any ad hoc assumption about statistical support, while at the same time

deploying approximations that make it more efficient than methods that rely on a local search of tree space (Akerborg et al. PNAS 2009).

The open source implementation of the method is available from https://github.com/ssolo/ALE.git .

References:

Akerborg, O., B. Sennblad, L. Arvestad, and J. Lagergren (2009)
Simultaneous bayesian gene tree reconstruction and reconciliation analysis.
Proc Natl Acad Sci U S A 106:57149.

David, L. A. and E. J. Alm. (2011)
Rapid evolutionary innovation during an archaean genetic expansion.
Nature 469:93 6.

Szöllősi, G. J., E. Tannier, N. Lartillot, and V. Daubin. (2013)
Lateral gene transfer from the dead.
Systematic Biology 62:386.

Szöllősi G. J., Rosikewicz W., Boussau B., Tannier E. and Daubin V. (2013)
Efficient Exploration of the space of reconciled trees
Systematic Biology doi:10.1093/sysbio/syt054

Wu, Y.-C., M. D. Rasmussen, M. S. Bansal, and M. Kellis. (2013)
Treefix: Statistically informed gene tree error correction using species trees.
Systematic Biology 62:11020.

Invited session 3

Metabolism

Organizer: Daniel Kahn - LBBE « Biométrie et Biologie Évolutive » UCB Lyon 1

Enzyme survey and how to find new ones

David Vallenet – LABGeM, Laboratory of Bioinformatics for Genomics and Metabolism – Genoscope

Millions of protein database entries are not assigned reliable functions. This shortcoming limits the knowledge that can be extracted from genomes and metabolic models. In contrast, the «orphan enzyme activities» problem, which was reported for the first time a decade ago, corresponds to experimentally characterized activities that lack associated protein sequence. We will first present an update view of previous surveys conducted on orphan enzymes (Sorokina *et al.*, 2014). Then, two strategies will be described that may be helpful in rescuing the orphans by simultaneously combining genomic and metabolic contexts over thousands of organisms (Smith *et al.*, 2012) and in finding new activities by the exploration of the enzymatic diversity within protein families (Bastard *et al.*, 2014).

References:

Sorokina M, Stam M, Médigue C, Lespinet O, Vallenet D (2014). Profiling the orphan enzymes. *In preparation.*

Smith AA, Belda E, Viari A, Medigue C, Vallenet D (2012). The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. *PLoS Comput Biol. 2012 May;8(5):e1002540.*

Bastard K, Smith AA, Vergne-Vaxelaire C, Perret A, Zaparucha A, De Melo-Minardi R, Mariage A, Boutard M, Debard A, Lechaplais C, Pelle C, Pellouin V, Perchat N, Petit JL, Kreimeyer A, Medigue C, Weissenbach J, Artiguenave F, De Berardinis V, Vallenet D, Salanoubat M (2014). Revealing the hidden functional diversity of an enzyme family. *Nat Chem Biol. 2014 Jan;10(1):42-9.*

Tuned for speed – Elucidation of strategies for rapid metabolic adaptations in prokaryotes

Christoph Kaleta **- Research Group Theoretical Systems Biology, Friedrich Schiller University**

Martin Bartl[a], Gundían M. de Hijas-Liste[b], Martin Kötzing[a,c], Eva Balsa-Canto[b], Stefan Schuster[d], Julio R. Banga[b], Pu Li[a], Christoph Kaleta[c]

Microorganisms have to be able to quickly react to environmental challenges to survive in fluctuating environmental conditions. Dynamic optimization represents a suitable tool that allows one to identify regulatory mechanisms that provide prokaryotes with the capability to rapidly adapt after a change in environmental conditions. In this talk, I will address two works that have shed light onto such regulatory strategies. In the first work, Bartl et al. (2013), we show that protein abundance and protein synthesis capacity are key factors that determine the optimal strategy for the activation of a metabolic pathway. If protein abundance relative to protein synthesis capacity increases, the strategies shift from the simultaneous activation of all enzymes over a sequential activation of groups of enzymes to a sequential activation of individual enzymes along the pathway. In the case of pathways with large differences in protein abundance, even more complex pathway activation strategies with a delayed activation of low-abundant enzymes and an accelerated activation of high-abundant enzymes are optimal. We confirm the existence of these pathway activation strategies for a large number of metabolic pathways in several hundred prokaryotes. In the second work, de Hijas-Liste et al. (2013), we used dynamic optimization to identify optimal points of control in complex metabolic pathways. We find that in converging pathways no regulation is required around the converging reaction while in diverging pathways a regulation after the diverging reaction is necessary. Moreover, the speed at which proteins can be synthesized has a strong influence onto specific positions that are optimal for a precise control of a metabolic pathway. While organisms with a slow protein production favor the control of metabolic pathways toward the beginning of pathways, organisms with rapid protein production favor the regulation of metabolic pathways toward at the terminal step. We confirm the utilization of these regulatory strategies in a screen of several hundred prokaryotic metabolic networks.

References:

M. Bartl, M. Kötzing, S. Schuster, P. Li, C. Kaleta (2013). Dynamic optimization identifies optimal programs for pathway regulation in prokaryotes. *Nature Communications*, 4:2243.

G.-M. De Hijas-Liste, E. Balsa-Canto, J. Banga, C. Kaleta (2013). Optimal regulatory programs for the control of metabolic pathways: The case of feedback inhibition. *In preparation.*

Constraints, adaptability and optimality of metabolic networks

Frank J. Bruggeman - Systems Bioinformatics, VU University

Microorganisms show a remarkable capacity to adjust their metabolic networks to changes in conditions to restore fitness. I will present the contours of a general theory of the evolution of metabolic networks under selective pressures that can be mimicked in laboratory evolution experiments. When nutrients are in excess during batch growth, natural selection favours microorganisms with the highest specific growth rate. In the last years, we have derived theory to understand the evolutionary outcome of this selective pressure in terms of the topology of the optimal metabolic network and the gene network that steers metabolism to optimal states in dynamic environments. I will contrast these findings for the microbial evolution under batch growth conditions with the evolutionary scenario in the chemostat, which has a qualitatively different selective pressure.

Annex

Submitted talks

Multi-trait genomic selection via multivariate regression with structured regularization

Julien Chiquet[1,2], Stéphane Robin[2], and Tristan Mary-Huard[2]

[1]Laboratoire Statistique et Génome – UMR CNRS 8071/Université d'Évry, France
[2]Laboratoire MMIP – UMR INRA 518/AgroParisTech – Paris, France

Background. In genomic selection regularized methods have mostly been used for their ability to handle high dimensional data and little attention has been devoted to the development of penalty functions including prior knowledge. Moreover, while several traits are usually considered in a given experiment, most methods only perform single trait genomic selection, neglecting correlations between phenotypes and leading to poor performance for the prediction of traits with low heritability. To circumvent these limitations, we consider the general linear model to simultaneously predict q responses (output variables) using the same set of p markers (input variables) based on a training sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\dots,n}$. One has

$$\mathbf{y}_i = \mathbf{B}^T \mathbf{x}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \quad \forall i = 1, \dots, n, \tag{1}$$

where $\boldsymbol{\varepsilon}_i$ is a noise term with a q-dimensional unknown covariance matrix $\mathbf{R}$, and $\mathbf{B}$ is the $p \times q$ matrix of regression coefficients.

Structured regularization with underlying sparsity. The present work proposes a general multivariate regression framework with three purposes: *i*) to account for the dependency structure between the outputs, i.e. to integrate the estimation of $\mathbf{R}$ in the inference process; *ii*) to integrate some prior information about linkage disequilibrium to account for the dependency structure between markers and evaluate its influence on the different phenotypes; *iii*) to induce sparsity on partial covariances via a set of parameters $\boldsymbol{\Omega}$ rather than on the regression coefficients $\mathbf{B}$, since according to the Gaussian graphical models (GGM) direct effects are measured by partial covariances between predictors and responses. We present an estimator to achieve these three goals using the conditional GGM formulation proposed in [1], whose conditional likelihood L is penalized by two regularization terms: the first term accounts for sparsity of the direct effects $\boldsymbol{\Omega}$ and the second one accounts for linkage disequilibrium via a structuring matrix $\mathbf{L}$. The (convex) objective function writes

$$J(\boldsymbol{\Omega}, \mathbf{R}) = -\frac{1}{n} \log L(\boldsymbol{\Omega}, \mathbf{R}) + \frac{\lambda_2}{2} \mathrm{tr}\left({}^t\boldsymbol{\Omega}\mathbf{L}\boldsymbol{\Omega}\mathbf{R}\right) + \lambda_1 \|\boldsymbol{\Omega}\|_1.$$

This work comes with an accompanying optimization procedure to minimize J.

Genomic selection in Brassica napus. We illustrate our proposal on the study conducted by [2] where $n = 103$ lines of *Brassica napus* are considered, on which $p = 300$ genetic markers and $q = 8$ traits were recorded. Traits included are five percent winter survival for 1992, 1993, 1994, 1997 and 1999 and days to flowering after 0, 4 and 8 weeks vernalization (flower0, 4 and 8). The left panel of Figure 1 gives both the regression coefficients (top) and the direct effects (bottom). The grey zones correspond to chromosomes 2, 8 and 10, respectively. The exact location of the markers within these chromosomes are displayed in the right panel, where the size of the dots reflects the absolute value of the regression coefficients (top) and of the direct effects (bottom). The interest of

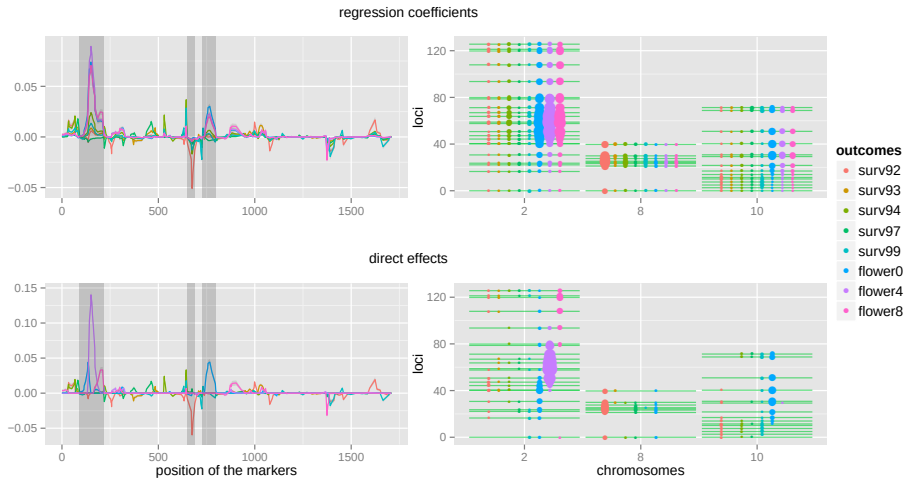


Fig. 1. Estimation of direct $\mathbf{\Omega}$ and indirect $\mathbf{B}$ genetic effects of the markers

considering direct effects rather than regression coefficients appears clearly on Figure 1, looking for example at chromosome 2. Three large overlapping regions are observed in the coefficient plot, for each flowering trait. A straightforward interpretation would suggest that the corresponding region controls the general flowering process. The direct effect plot allows to go deeper and shows that these three responses are actually controlled by separated sub-regions within this chromosome. The confusion in the coefficient plot only results from the strong correlations observed between the three flowering traits.

References

1. Sohn, K., Kim, S.: Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. JMLR **W&CP**(22) (2012) 1081–1089
2. Ferreira, M., Satagopan, J., Yandell, B., Williams, P., Osborn, T.: Mapping loci controlling vernalization requirement and flowering time in brassica napus. Theor. Appl. Genet. **90** (1995) 727–732

A Regularisation Path-Following Approach for Discovering Interactions in High-Dimensional Survival Data

David A. duVerle[1], Ichiro Takeuchi[2] and Koji Tsuda[1]

[1] Computational Biology Research Center, AIST, Tokyo,
dave.duverle@aist.go.jp
[2] Department of Computer Science, Nagoya Institute of Technology, Nagoya

Abstract

Motivation

Several efficient methods exist to relate high-dimensional gene expression data to various clinical phenotypes, however finding *combinations* of features in such input remains a challenge, particularly when fitting complex statistical models such as those used for survival studies. While such models based on linear combinations of gene expression values can be used to approximate more complex interactions, they usually falls short, when trying to identify certain combinatorial effects (e.g. synthetic lethal genes [1]). Identifying such gene combinations could lead to promising therapeutic applications, yet remains a practical impossibility with current fitting methods due to serious computational complexity issues, as they rely on the ability to enumerate potential input variables.

Method

Gene data, with its typically large dimension and small sample size, invite the use of a penalisation component in its statistical models, with ℓ_1-norm often preferred for its ability to drive sparsity of the model and select a concise set of variables (gene expression values, mutation types, etc.) [2]. Park & Hastie [3], in particular, proposed a method to compute the *regularisation path* of an ℓ_1-penalised Cox model [4] for survival data: efficiently estimating a series of Cox models with different levels of complexity and sparsity. Their approach, however, can only consider the impact of individual input variables, introduced one at a time, on the model.

In our work [5], we extended this single-variable regularisation path-following approach to handle combinatorial covariates (arbitrarily complex patterns of input variables). Using this technique, we can efficiently compute regression parameters for complex statistical models, at the optimal amount of penalisation, with the only requirement of a convex loss function. We solved the combinatorial explosion issue by taking advantage of itemset mining techniques [6] that brought the average computational complexity of the algorithm within practical reach.

With our fitting method, virtually limitless combinations of genes and phenotypes, grouped in itemsets of boolean variables, are used as single predictor variables in the model. Running multiple iterations of the algorithm on subsampled dataset, we can produce ordered lists of candidate interactions with strong predicting power.

Results

The interactions found by applying our method to cancer study survival data (breast cancer and neuroblastoma) include many genes that could not be found through linear models, yet show up in literature as strongly tied to these conditions, confirming the crucial importance of taking interaction effects into account in order to detect some of the weaker signal in gene expression data.

Beyond proportional hazards models, our itemset-based method can be applied to any regression model with convex loss, each time making use of the input's structure and sparsity to sidestep complexity issues, while at the same time guaranteeing that events along the regularisation path (values of the regularisation parameter for which a change occurs in the model structure) are exhaustively explored.

Acknowledgment

The authors would like to thank Yuko Murakami-Tonami and Kenji Kadomatsu, of the Department of Molecular Biology at Nagoya University's Graduate School of Medicine, for their suggestions and guidance regarding the medical aspects of our research into synthetic lethals for cancer cells, as well as Hiroshi Mamitsuka and Timothy Hancock, of Kyoto University, for their helpful feedback and suggestions in summarising our work.

References

1. W.G. Kaelin. The concept of synthetic lethality in the context of anticancer therapy. *Nature reviews cancer*, 5(9):689–698, 2005.
2. R. Tibshirani et al. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.
3. M.Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.
4. D. Ghosh. Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, 59(4):992–1000, 2003.
5. David duVerle, Ichiro Takeuchi, Yuko Murakami-Tonami, Kenji Kadomatsu, Koji Tsuda, et al. Discovering combinatorial interactions in survival data. *Bioinformatics*, 29(23):3053–3059, 2013.
6. T. Uno, T. Asai, Y. Uchida, and H. Arimura. An efficient algorithm for enumerating closed patterns in transaction databases. In *Discovery Science*, pages 57–59. Springer, 2004.

On the robustness of the Generalized-Fused Lasso to prior specifications

Vivian Viallon[1], Sophie Lambert-Lacroix[2], Hölger Hoefling[3], and Franck Picard[4]

[1] Université de Lyon, F-69622, Lyon, France; Université Lyon 1, UMRESTTE, F-69373 Lyon; IFSTTAR, UMRESTTE, F-69675 Bron.
[2] UMR 5525 UJF-Grenoble 1/CNRS/UPMF TIMC-IMAG, Grenoble, F-38041
[3] Novartis Pharma, Basel, Switzerland.
[4] Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558 Univ. Lyon 1, F-69622 Villeurbanne.

Abstract. Using networks as *prior* knowledge to guide model selection is a way to reach structured sparsity. In particular, the fused lasso that was originally designed to penalize differences of coefficients corresponding to successive features has been generalized to handle features whose effects are structured according to a given network. As any *prior* information, the network provided in the penalty may contain misleading edges that connect coefficients whose difference is not zero, and the extent to which the performance of the method depend on the suitability of the graph has never been clearly assessed. In this work we investigate the theoretical and empirical properties of the adaptive generalized fused lasso in the context of generalized linear models. In the fixed p setting, we show that, asymptotically, adding misleading edges in the graph does not prevent the adaptive generalized fused lasso from enjoying asymptotic oracle properties, while forgetting suitable edges can be more problematic. These theoretical results are complemented by an extensive simulation study that assesses the robustness of the adaptive generalized fused lasso against misspecification of the network as well as its applicability when theoretical coefficients are not exactly equal. Our contribution is also to evaluate the applicability of the generalized fused lasso for the joint modeling of multiple sparse regression functions. Illustrations are provided on two real data examples.

Keywords: lasso, generalized linear models, joint modeling, model selection

Of *cis*, *trans* and feedback regulation: Dissecting local regulation in yeast

Julien Gagneur, Daniel Bader

Gene Center of the Ludwig-Maximilians Universität, München, Feodor-Lynenstr. 25, 81377 Munich, Germany

Abstract. The vast majority of regulatory genetic variants are found in close vicinity of the regulated gene. The distinction between *cis*- and *trans*-acting variants is a fundamental starting point to understand the mechanisms underlying these regulatory variants. Typical *cis*-regulatory variants affect transcription factor binding sites or RNA stability. Local *trans*-regulations have been less studied and include feedbacks, an essential regulatory feature of biological systems. To understand the contribution of *cis* and *trans* regulation and their potential interplay, we devised a novel experimental design in which allele-specific expression in a hybrid cross of two yeast strains is compared to allele-specific expression in a pool of segregants of the same cross. We implemented a statistical procedure based on generalized linear models for RNA-seq count data to quantify the contribution of *cis* and *trans* effect in local regulation. Our model allows controlling for allele selection in the pool population by integrating robust estimates of allele frequency from genomic DNA sequencing. Applied to a cross of two distant yeast strains, our analysis revealed significant differences of *cis*-effects among major gene categories (essential, non-essential, and non-coding). Furthermore, our results shed light on the effects of feedback in buffering or enhancing the impact of genetic variation on gene expression.

A New Gene-Based test of Association Using Extended Rasch Models

Wenjia Wang, Mickael Guedj

Pharnext,Department of bioinformatics and biostatistics,
92130 Issy-Les-Moulineaux, France

Abstract. In GWAS analysis, gene-based tests of association has become an important alternative to the tradition single-marker association analysis. However, several statistical issues limited the performance of gene-based tests when assessed to real data. Here we introduce a new test to provide a p value for a gene by using extended Rasch Models. It provide a score for each individual in GWAS by aggregate genotypes of SNPs within a gene and the weight of each SNP. In a variate of simulation, this test maintained a correct false positive rate and its power exceeded other tests when the number of disease related SNP increased. This test can be generized to multivariate traits analysis.

Keywords: GWAS, gene-bases association test, Rasch Models

1 Introduction

Genome-wide association studies (GWAS) are increasingly used for identification of genetic associations for complex diseases. Traditionally, single nucleotide polymorphism (SNP) are tested individually. Recently, a gene-based approach consider association between diseases and all SNPs within a gene becomes increasingly important. As a matter of fact, in Genetics, the gene is often considered as the unit of interest as the analyses of the functional mechanisms of a disease are generally based on genes and their products such as RNA or resulting proteins[1]. To derive a gene-level measure of significance, such as a test statistic or a p-value, one needs to combine the results of all the SNPs corresponding to the gene.

Computing a single p-value per gene raises several statistical issues. First, several SNPs are usually genotyped within a gene and combining the results of each individual SNP test outcome corresponds to a multiple-testing situation. In addition, markers within a gene are usually closely located on the genome and therefore likely to be in linkage disequilibrium (LD). This LD pattern of a gene leads to a situation of multiple-testing with dependent tests. Hence, a gene-based association approach considering this two issues is often of interest.

A number of gene-based tests have been proposed such as GATES[2], margin and VEGAS[3]. However, most current gene-based tests consider combining single SNP association p value to compute a gene-based test statistic. These methods may ignores the potential joint effect of SNPs within a gene. Moreover, the permutation process to account for confounding factors required in several approaches is quite time and computation consuming.

Therefore we proposed a new gene-based association test using extended Rasch Models[4] . This test can combine the genotypes of all SNPs within a gene to produce a score for each individuals and then to derive a gene-level p value.It is also less time consuming than permutation. Rasch Model is a generalized linear model for analyzing categories data to measure variables. Rasch models are increasingly being use in many areas such as education and clinics [5], but never applied to genetics. After study, we found that Rasch models can be adapted to the analysis of GWAS data. GWAS data is consisted by a set of SNP that values are categorized in 0, 1 and 2 as items with 3 categories. The probability of linkage between a set of SNP and disease can be estimated as latent trait in the Rasch models, so a p-value can be derived for each gene. Compared to other statistical tests which calculate p-value for each SNP, Rasch models consider the pattern of every SNP in a gene. The application of Rasch on GWAS data may offer a better solution for genetic disorders research.

In a series of simulation with different scenarios (number of DSL, relative risk, LD structure of genes) we compared the extended Rasch Models to 6 other gene-based association tests: minP[6], margin test[7], goeman test[8], GATES, SKAT, Fisher's method. In the comparison, the extented Rasch models maintain correct false positive rate in different situation and it has the highest power when the number of disease related SNPs exceeds 7 in simulation.

This gene-based test can be further extended to treat GWAS data with multivariate phenotypes, which is of interest in GWAS study but few approaches have been developed for.

References

1. Eric Jorgenson and John S Witte. A gene-centric approach to genome-wide association studies. *Nature Reviews Genetics*, 7(11):885–891, 2006.
2. Miao-Xin Li, Hong-Sheng Gui, Johnny SH Kwan, and Pak C Sham. Gates: a rapid and powerful gene-based association test using extended simes procedure. *The American Journal of Human Genetics*, 88(3):283–293, 2011.
3. Jimmy Z Liu, Allan F Mcrae, Dale R Nyholt, Sarah E Medland, Naomi R Wray, Kevin M Brown, Nicholas K Hayward, Grant W Montgomery, Peter M Visscher, Nicholas G Martin, et al. A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139–145, 2010.
4. Georg Rasch. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. 1960.
5. Jeremy C Hobart, Stefan J Cano, John P Zajicek, and Alan J Thompson. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *The Lancet Neurology*, 6(12):1094–1105, 2007.
6. Ali Torkamani, Eric J Topol, and Nicholas J Schork. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92(5):265–272, 2008.
7. Wei Pan. Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic epidemiology*, 33(6):497–507, 2009.
8. Jelle J Goeman, Sara A Van De Geer, and Hans C Van Houwelingen. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493, 2006.

Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models

Eric Frichot[1], Sean D Schoville[1], Guillaume Bouchard[2], Olivier Francois[1]

[1] TIMC-IMAG UMR 5525, Universite Joseph Fourier Grenoble, Centre National de la Recherche Scientifique, Grenoble, France
[2] Xerox Research Center Europe, Meylan, France

Abstract. Local adaptation through natural selection plays a central role in shaping the genetic variation of populations. A way to investigate signatures of local adaptation, especially when beneficial alleles have weak phenotypic effects, is to identify polymorphisms that exhibit high correlation with environmental variables. However the geographical basis of both environmental and genetic variation can confound interpretation of these associations, as they can also result from genetic drift at neutral loci. Here we propose an integrated framework based on spatial statistics, population genetics and ecological modeling for scans for signatures of local adaptation from genomic data. We present a novel class of algorithms to detect correlations between environmental and genetic variation that take account background levels of population structure and spatial autocorrelation in allele frequencies generated by isolation-by-distance mechanisms. Our framework uses Latent Factor Mixed Models, a hierarchical Bayesian mixed model in which environmental variables are fixed effects and population structure is introduced as random effects. We implement fast algorithms that simultaneously estimate scores and loadings for the genotypic matrix and effects of environmental variables. Comparing these new algorithms with related methods provides evidence that LFMM can efficiently estimate random effects due to population history and isolation-by-distance patterns when computing gene-environment correlations, and decrease the number of false-positive associations in genome scans. We then apply these models to plant and human genetic data, identifying several genes with functions related to development that exhibit strong correlations with climatic gradients.

Identifying Differentially Expressed Proteins by a Binary Threshold Model

Sebastian Gibb and Korbinian Strimmer

Institute for Medical Informatics, Statistics and Epidemiology (IMISE),
University of Leipzig,
Härtelstr. 16-18, D-04107 Leipzig, Germany.

Abstract. Proteomics-based analyzes using mass spectrometry are becoming routine in clinical diagnostics, for example to monitor cancer biomarkers using blood samples. For preprocessing mass spectrometry data a wide variety of algorithms is available. Accordingly, we have recently compiled a standard analysis pipeline that includes methods for baseline correction, normalization and alignment of spectra in the freely available R package `MALDIquant` [1, 2]. However, identifying differential expression and ranking of proteomic features remains challenging due to the simultaneously discrete (absence-presence) and quantitative nature of protein expression.
Here, we present a simple yet effective approach using a binary threshold model for ranking and identifying differentially expressed proteins. This approach works by peak-wise data-adaptive thresholding of protein expression and subsequent ranking of the dichotomized features using a suitable entropy measure. Our framework may be viewed as a generalization of the 'peak probability contrast' approach of [3] and works, in contrast to [3], both in the two-group and the multi-group setting. Using data from a recent pancreas cancer study conducted at the University of Leipzig [4] we are able to identify biological relevant and statistically predictive marker peaks unrecognized in the original analysis.

Keywords: Differential expression, proteomics, mass spectrometry, peak ranking, cancer biomarker.

References

1. Gibb, S., and Strimmer, K.: MALDIquant: a versatile R package for the analysis of mass spectrometry data. Bioinformatics 28, 2270–2271 (2012)
2. http://strimmerlab.org/software/maldiquant/
3. Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q.-T.: Sample classification from protein mass spectrometry, by 'peak probability contrasts'. Bioinformatics 17, 3034–3044 (2004)
4. Fiedler, G. M., Leichtle, A. B., Kase, J., Baumann, S., Ceglarek, U., Felix, K., Conrad, T., Witzigmann, H., Weimann, A., Schtte, C., Hauss, J., Büchler, M., and Thiery, J.: Serum peptidome profiling revealed platelet factor 4 as a potential discriminating peptide associated with pancreatic cancer. Clin. Cancer Res. 3812–3819 (2009)

The Causal Mediation Analysis in Genomic Data

Giulia Malaguti, Séverine Affeldt, and Hervé Isambert

UMR 168 CNRS-UPMC, Institut Curie, 26 rue d'Ulm, 75248 Paris, France
giulia.malaguti@curie.fr

Keywords: Causal Inference Methods, Mediation Analysis, Direct *versus* Indirect Effects, Genomic Data, Whole Genome Duplication

1 Introduction

The increasing amount of high throughput data has allowed to reveal that many genomic properties are to some extent correlated, suggesting direct statistical associations. Yet, these correlations could in fact be mediated by the indirect effect of third properties and only provide partial insights on complex biological systems. To quantify direct and indirect effects in genomic data, we performed Mediation analyses following the approach of Pearl[1], a method typically used in social sciences and epidemiology. We present here the extension of the Mediation framework to the case of more than one mediating variable, and its application to the genomic properties implicated in the biased retention of whole genome duplicated genes, so-called ohnologs, in the course of vertebrate evolution.

2 The Mediation Framework

The Mediation analysis assesses the importance of a mediator M in transmitting the indirect effect of a variable X on the response variable Y. In this framework, the average direct effect $DE_{xx'}$ is defined as the change on Y due to a change of X from the initial value x to x', while keeping M fixed at the value $m(x)$. Similarly, the average indirect effect $IE_{xx'}$ is defined as the change on Y if M could be changed from the value $m(x)$ to $m(x')$ while keeping X to its original value x[1]. The Mediation formulae can be rewritten in terms of the expectation value $E(Y|x,m)$ of Y given X and M, and $P(M|x)$, that denotes a value of M drawn from a distribution P conditionally on X,

$$DE_{xx'} = Y(x', m(x)) - Y(x, m(x)) = \sum\nolimits_m [E(Y|x', m) - E(Y|x, m)]P(M|x)$$

$$IE_{xx'} = Y(x, m(x')) - Y(x, m(x)) = \sum\nolimits_m E(Y|x, m)[P(M|x') - P(M|x)]$$

$$TE_{xx'} = Y(x', m(x')) - Y(x, m(x)) = \sum\nolimits_m E(Y|x') - E(Y|x)$$

If X, M, and Y are binary variables, expectation values of Y and values of M drawn from a conditional distribution P can be estimated from the counts in the data of the triplets corresponding to each possible combination of values for (X, M, Y). This means that all the effects can be directly derived from the number n_{xmy} of triplets where $(X, M, Y) = (x, m, y)$, for each value of (x, m, y)[1].

2.1 Extension to Many Mediators

In order to study several correlated properties in genomic data, we extended the single mediator formulae to the case of more than one mediator. The simplest generalization consists in gathering all the k mediators for the ordered pair of variables (X, Y) in a single super mediator $M = \{M_1, M_2, \dots M_k\}$. This approach allows to discriminate the direct effect $DE_{xx'}$ from the effects through any other indirect path connecting X and Y.

However, it is fundamental to identify the correct set of mediators for an ordered pair of variables (X, Y) in a causal graph. The mediators are the variables that contribute to the correlation between X and Y, and belong to indirect paths between X and Y that do not include any 'v-structure' ($X... \rightarrow M \leftarrow ...Y$). The erroneous inclusion of non mediators in the super mediator M creates artificial correlations between X and Y. By iteratively collecting the mediators of each edge of a graph, we developed an efficient algorithm to recover in polynomial time the correct set of mediators for each pair of variables in a graph.

3 The Mediation Analysis Applied to Genomic Data

We considered the 20,506 protein-coding genes in human. Based on data mining analyses[2], we identified several gene properties, such as the essentiality or the implication in cancers, and we inferred the underlying causal graph linking these properties. We applied the extended Mediation analysis formulae to specifically disentangle the direct from the indirect effects of properties favoring the specific retention of ohnologs[2] during vertebrate evolution[3, 4]. Our results demonstrate that the retention of human ohnologs is surprisingly more strongly caused by their susceptibility to dominant deleterious mutations than their interactions within multi-protein complexes, unlike frequently invoked in the field.

3.1 Conclusion and Perspectives

Our study highlights the need to go beyond statistical correlations in the analysis of genomic data and to rely on more advanced inference methods to disentangle indirect from direct interaction pathways. The application of this approach to other biological properties emerging from genomic data analysis will cast new light on the function and evolution of biological properties and networks.

References

1. Pearl, J.: The mediation formula: A guide to the assessment of causal pathways in nonlinear models. Causality: Statistical Perspectives and Applications. John Wiley and Sons, Ltd, Chichester, UK (2012): 151-179.
2. Singh, P-P. *et. al*: On the Expansion of Dangerous Gene Repertoires by Whole-Genome Duplications in Early Vertebrates. Cell Reports, Vol. 2, Issue 5 (2012)
3. Ohno, S.: Evolution by gene duplication. Springer, New York, (1970).
4. Putnam, N.H. *et. al*: The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064-1071, (2008).

An RNA-Seq read count emission model for transcriptional landscape reconstruction with state-space models

Bogdan Mirauta[1,⋆], Pierre Nicolas[2,†] and Hugues Richard[1,†]

[1] Génomique des Microorganismes, UPMC and CNRS UMR7238, Paris, France. [2] Mathématique Informatique et Génome, INRA UR1077, Jouy-en-Josas, France.
⋆ To whom correspondence should be addressed. † Contributed equally to this work.
http://www.lgm.upmc.fr/parseq

Motivation Sequencing technologies play an increasing role in the investigation of gene expression (RNA-Seq). The most common RNA-Seq strategy is based on random shearing, amplification and high-throughput sequencing of the RNAs; yielding millions of sequence reads which serve to characterize whole-genome transcriptional profiles [1]. We developed Parseq, a statistical method to estimate the local transcription levels and to identify transcript borders. Extending previous work on tilling array [2], this transcription landscape reconstruction relies on a State-Space Model (SSM) to describe transcription level variations in terms of abrupt shifts and more progressive drifts. The transcription level at the genome position t is denoted u_t and corresponds to the expectation of y_t, the count of reads with 5'-end mapping to position t. It cannot be directly equated to the read count y_t due to randomness and biases in library preparation and sequencing. We use Particle Gibbs, a Gibbs algorithm based on a conditional Sequential Monte Carlo, to estimate the SSM parameters and reconstruct the trajectory $\mathbf{u} = (u_t)_{t\geq 1}$ from the sequence of read counts $\mathbf{y} = (y_t)_{t\geq 1}$.

A key point to obtain good results with this framework is to incorporate a realistic emission model for the distribution of y_t given u_t.

Distribution of read counts in real data sets The variability of RNA-Seq read counts has been approximated by a Poisson ($\mathcal{P}$) distribution when re-sequencing of the same library [1] and by a Negative Binomial (NB) when comparing between samples [3]. For $y_t|u_t$, a mixed Poisson distribution seems unavoidable to account simultaneously for the incompressible variance of the final sampling by sequencing (Poisson) and for the extra-variability introduced by randomness in library preparation and by position-specific biases occurring at all steps of the protocols. The Poisson-Gamma mixture is the simplest choice as it corresponds to the NB. Here, it leads to envision the relationships $y_t \sim P(u_t z_t)$ where z_t would follow a Gamma distribution with mean 1 and variance ϕ (variance $u_t + \phi u_t^2$).

We examined the distribution of read counts in regions of homogeneous expression level (*e. g.* ORFs) of two real data sets. In particular, we asked whether the NB could capture the relationships between mean and variance and simultaneously account for the fraction of positions with zero-counts (fig. 1). Marked discrepancies between the data and the NB are seen not only in the fraction

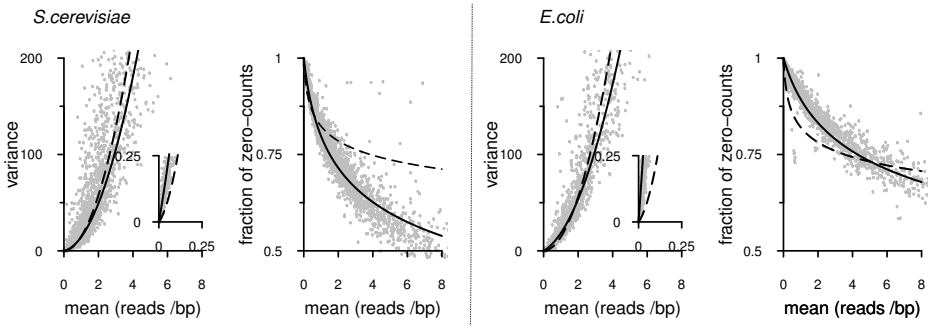


Fig. 1: Read counts distribution in regions of homogeneous expression. *S. cerevisiae* (SRR121907, SOLiD data, read coverage 1.6 reads/bp) - left; *E. coli* (SRR794838, Illumina, 2.4 reads/bp) - right. Dots: long open-reading frames (ORF, region without in-frame stop codon); Dashed lines: NB with overdispersion estimated via variance versus mean regression; Plain lines: Parseq model.

of zero-counts but also in the behaviour of the variance at low expression level where it exceeds $u_t + \phi u_t^2$. Breaking this relationship between mean and variance implies a relationships between the mixing distribution and u_t more subtle than a simple scaling.

A new read count model We adopted a mechanistic perspective to develop the more complex emission model incorporated in Parseq. Our aim is to account for the three main steps of the experimental protocol: (i) initial molecule sampling and fragmentation, (ii) amplification, and (iii) final sampling by sequencing. Namely, we write $y_t \sim \mathcal{P}(x_t a_t)$ where a_t (mean μ_a) wishes to capture the effect of randomness in amplification, and x_t (mean u_t/μ_a) is aimed at representing the number of molecules after initial sampling. The Poisson distribution accounts for the final sampling. We capture the additional variability introduced by position-specific biases in library preparation with a supplementary random term s_t that impacts on initial molecule sampling : $x_t \sim \mathcal{P}(u_t s_t/\mu_a)$ with s_t having (for simplicity) a Gamma distribution of mean 1 an variance $1/\kappa_s$.

With a Gamma distribution (shape κ and scale θ) for the amplification term a_t and after integration over all a_t the density of our emission model writes

$$\pi(y_t; u_t, s_t) = \sum_{x_t=0}^{\infty} \mathcal{P}(x_t; \frac{u_t s_t}{\kappa\theta}) \cdot \mathcal{NB}(y_t; \kappa, \frac{x_t\theta}{x_t\theta + 1}).$$

This density could also be integrated over all s_t but in Parseq we introduced a Markov dependency between s_{t+1} and s_t to account for short-range autocorrelation between counts. This model makes it possible to capture the characteristics of the read count variability (fig. 1) and increases our ability to disentangle genuine transcription breakpoints from protocol induced variations.

1. J. C. Marioni, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.
2. P. Nicolas, et al. Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics*, 25(18):2341–2347, 2009.
3. Mark D. Robinson, et al. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

Evolution of Codon Usage Bias in *E.coli*

Fanny Pouyet, Marc Bailly-Bechet, and Laurent Guéguen

Laboratoire de Biologie et Biométrie Evolutive, University Claude Bernard Lyon 1,
43 bd du 11 novembre 1918 69622 VILLEURBANNE cedex
{fanny.pouyet,marc.bailly-bechet,laurent.gueguen}@univ-lyon1.fr

Abstract. We develop an evolutionary model, inspired from Yang and Nielsen [1], that takes into account evolutionary processes at nucleotidic, codons and amino acids levels. It is implemented in Bio++ suite [2]. We apply this model in an homogeneous, non-stationary context. We study evolution of codon usage bias, preference of synonymous codons over others, in three close strains of *E. coli*. It computes equilibrium and root frequencies. We show that nucleotidic mutational bias and codon bias are counteracting: mutational bias tends to increase AT composition whereas codon bias favors GC enrichment. **Keywords :** Evolutionary Model, Codon Usage Bias, *Escherichia coli*

Introduction

The genetic code is redundant, and some codons, called synonymous, are translated in the same amino acid. Degeneracy of the code does not however lead to a uniform usage of those synonymous codons: at species or at genomic level, a particular codon is preferrentially used over other synonymous ones [3]. Codon usage bias may vary between species and between genes. These differences are easily observable and measurable, as for example with Fop statistics which are the frequency of optimal codons (preferred codons in highly expressed genes). However, to fully understand how this bias arose and is maintained in a set of genes, we need a model that studies codon usage in an evolutionary approach. We develop a model, inspired from Yang and Nielsen [1], that distinguishes evolution of coding sequences at nucleotidic, codons and amino acids level. Our model explicitely describes both the mutational bias of nucleotides and the selection of preferred codons over other synonymous ones.

Evolutionary model

Our evolutionary model is implemented in Bio++ [2] which is a set of C++ librairies for bioinformatics molecular evolution. Our model is based on a preference parameter for each codon, relative to its synonymous ones: $\Phi_{aa}(i)$ is for codon i that code for amino acid aa. This parameter corresponds to the relative codon frequency, within its amino acid, if selection for codon usage was the only selective pressure acting on sequences. We also consider Ψ_{aa_i} parameter which is the frequency of amino acid aa coded by codon i. Our sequence evolution model is such that equilibrium frequency of codon i, noted π_i^*, is proportional to:

$$\pi_i^* \alpha \quad \underbrace{\pi_{i_1}\pi_{i_2}\pi_{i_3}}_{\substack{\text{nucleotides} \\ \text{equilibrium frequency}}} \quad . \quad \underbrace{\Psi_{aa_i}}_{\substack{\text{amino acid} \\ \text{preference}}} \quad . \quad \underbrace{\Phi_{aa}(i)}_{\substack{\text{codon} \\ \text{preference}}}$$

Parameter π_{i_p} is the equilibrium frequency of nucleotide i_p with $p \in [1, 3]$ the nucleotide position within codon. With no selection in our model, every codons must have the same equilibrium frequency which is: $\pi^* = \frac{1}{61}$ (note that in this model, we do not consider stop codons). We perform non-stationary runs enabling us to compute frequencies at the root of the tree. It helps us to depict how codon bias evolves and to understand how selective pressure occurs.

Results and Perspectives

We apply this model on three strains of $E.coli$ [4]: K12, CFT073 and 0157:H7. We have 3,353 genes clustered into concatenates of 100 genes, by increasing Fop (codon usage bias intensity). We study relationships between codon usage bias and others parameters. We obtain, as awaited, a negative correlation between $\omega = \frac{dN}{dS}$ and strong usage bias (high Fop). As expected, preferred codons (codons with highest $\Phi_{aa}(i)$) and tRNA content are positively correlated. GC content at equilibrium (GC*) is influenced by both selection on codon usage and mutational bias. More precisely, codons and nucleotides levels present contrary effects on base composition: codons level tends to enrich sequences with C and G whereas nucleotides level induced enrichment by A and T. With this non-stationary model, we can infer root and equilibrium frequencies of codons and we observe a global nucleotide enrichment towards AT.

We show there are opposite forces that drive sequence evolution from which selection on codon usage and also, mutational bias. We are currently refining the model at the amino acid level by considering distance between amino acids. We plan to use deeper datasets and non-homogeneous models of codon bias evolution.

References

1. Yang, Z. and Nielsen, R: Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol. Biol. Evol 25(3):568-579, Mar (2008)
2. Gueguen L., Gaillard S., Boussau B., Gouy M., Groussin M., Rochette NC., Bigot T., Fournier D., Pouyet F., Cahais V., Bernard A., Scornavacca C., Nabholz B., Haudry A., Dachary L., Galtier N., Belkhir K., Dutheil JY: Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol. Biol. Evol 30(8):1745-50, Aug (2013)
3. Sharp, PM., Emery, LR. and Zeng K: Forces that influence the evolution of codon bias. Phil. Trans. R. Soc. B 365, 1203-1212 (2010)
4. Jordan, IK, Kondrashov, FA, Adzhubei, IA, Wolf, YI, Koonin, EV, Kondrashov, AS and Sunyaev, A: A universal trend of amino acid gain and loss in protein evolution. Nature, 433(7026):633-638, Feb (2005)

Efficient statistical computations on genome-scale data using reduced representations

Md Pavel Mahmud[1] and Alexander Schliep[1,2]

[1]Department of Computer Science, Rutgers University
[2]BioMaPS Institute for Quantitative Biology, Rutgers University
{pavelm,wiedenhoeft,schliep}@cs.rutgers.edu

1 Introduction

The amount of data generated by high-throughput technologies requires improvements to algorithms and statistics to cope with biases and noise and assess statistical significance of findings in reasonable running times. A recent insight was that many computations can be greatly accelerated by using a suitable reduced representation of the input. This strategy accelerates full Bayesian HMM to the point of competitiveness with Maximum-Likelihood HMM and also leads to large improvements in quality and speed of the analysis of high-throughput data (HTS), a technology to unravel gene sequences on a large scale, in a compressive genomics approach.

2 Full Bayesian Hidden Markov Models

Hidden Markov Models (HMM) are widely used for analyzing Comparative Genomic Hybridization (CGH) data to identify chromosomal aberrations or copy number variations by segmenting observation sequences. For efficiency reasons often parameters of an HMM are estimated with maximum likelihood and a segmentation is obtained with the Viterbi algorithm. This introduces considerable uncertainty in the segmentation, which can be avoided with Bayesian approaches using Markov Chain Monte Carlo (MCMC) sampling. While their advantages have been clearly demonstrated, the likelihood based approaches are preferred in practice for their lower running times; datasets coming from high-density arrays and high-throughput sequencing amplify these problems.

These computational disadvantages can be alleviated by pre-processing the input to arrive at a suitable reduced representation. Leveraging spatial relations between data points in typical data sets leads to a (partial) clustering approach, inspired by ideas from using sequence compression for HMMs with discrete observations and spatial data structures such as *kd*-trees. Based on this reduced representation, we propose an approximative MCMC sampler (Mahmud . We test our approximate sampling method on simulated and biological ArrayCGH datasets and demonstrate a speed-up of 10 to 90 while achieving competitive results with the state-of-the art Bayesian approaches.

3 Compressive Genomics

The default modus operandi for analyzing high-throughput sequencing (HTS) data—which is pervasive in clinical and biological applications such as cancer research, and which is expected to gain enormous momentum in future personalized medicine applications—is individual analysis on the level of reads and on the level of genomes. This leads to non-trivial computational demands, as typical data sets consist of up to 2 billions of

sequencing reads, and large studies might provided hundreds of such data sets. Core steps of HTS analysis are read error correction, mapping to the reference genome and identifying genetic variations

We propose to compute reduced representations of HTS data for a single or multiple sequencing experiments in a way that existing and future methods can directly operate on these reduced representations of the data and enable the use of advanced statistics even on very large data sets. We arrive at a reduced representation of HTS data sets by a novel highly-efficient clustering method able to cope with many billions reads from a single or multiple sample genomes, even in repetitive genomes. The reduced representation simplifies sharing and storing of data. Additionally, existing downstream analysis methods can be used at great improvements to running times; adapted or novel downstream algorithms can operate directly on the clustered representations at even larger improvements. The most important consequence of the much increased computational efficiency is not the savings in CPU cycles, but rather the ability to use much more sensitive and accurate tools. Even for read mapping, our approach makes the routine use of Stampy feasible, negating the needs for additional advances downstream tools for calling small structural variants not identified by the speedier, but tools such as Bowtie which are less accurate in the presence of indels. Our approach constitutes a novel foundation for efficient statistical tools for analyzing HTS data.

References:

1. P. Mahmud, A. Schliep. Fast MCMC sampling for Hidden Markov Models to Determine Copy Number Variations. *BMC Bioinformatics*, 12:1, 428, 2011.
2. P. Mahmud, A. Schliep. TreQ-CG: Clustering Accelerates High-Throughput Sequencing Read Mapping. *Submitted*, 2014. `http://bioinformatics.rutgers.edu/Publications/pavel2014/`

Testing 2×2 Association with Uncertain Classification

Louis J. Dijkstra and Alexander Schönhuth

Centrum Wiskunde & Informatica (CWI), Life Sciences Group
Amsterdam, The Netherlands

{dijkstra|as}@cwi.nl

Keywords. Exact Test; 2×2 Contingency Tables; Uncertainty; Probabilistic Automata; Genome-wide Association Studies; Linkage Disequilibrium

Testing for association between two dichotomous variables plays a key role in the field of genome-wide association studies (GWASs), e.g., to determine whether two genetic variants (SNPs/indels) are in linkage disequilibrium (LD) or to assess disease risk [1,2]. A wide variety of 2×2 association tests have been developed over the years, where Pearson's asymptotic χ^2 and Fisher's exact test are among the most popular choices [3,4]. Every such approach requires the assumption that the observed items (gametes/individuals) were cross-classified with absolute certainty. In GWASs, for example, one normally assumes that genotyping proceeded free of errors. In LD studies it is common to assume that haplotypic phase in individuals has been determined perfectly. This assumption usually is rather optimistic, in particular when experiments depend on, although increasingly relevant, still notoriously noisy next-generation sequencing (NGS) data.

Although probability distributions over the 4 entries of the 2×2 table are often available for each observed item, no tests exist that can exploit this. The common approach — although often not reported explicitly — is to assign each observed item to the most likely class given the data, yielding a unique 2×2 contingency table. Data that does not allow for perfect classification, however, can give rise to a (potentially large) number of distinct tables with varying degrees of evidence in favor or against the null-hypothesis of no association.

Here, we determine the full probability distribution over all possible 2×2 tables with the same number of counts as the number of items observed (N) as a first step. We do this by way of an exact recursive polynomial-runtime, $\mathcal{O}(N^4)$, algorithm, which establishes a clear improvement over naive approaches, which require $\mathcal{O}(N!)$ runtime. The algorithm has parallels with probabilistic arithmetic automata, thereby drawing an interesting connection to pattern statistics on Markovian text models together with their highly engineered implementations [5,6].

Secondly, we determine for every possible table its degree of evidence in favor or against the null-hypothesis by applying Fisher's exact test[1]. Let $P(t)$ and $Q(t)$

[1] Fisher's test is used here for its exact character. Other tests could be applied as well.

denote, respectively, the probability of observing table t given the classification uncertainties and Fisher's (one- or two-sided) p-value. The p-value of the exact test is then defined as the expected Fisher's p-value over the set of all possible tables, T:

$$p := \mathbb{E}\big[Q(t)\big] = \sum_{t \in T} P(t) \cdot Q(t). \tag{1}$$

The test proposed here thus incorporates the evidence from all possible tables weighted by their respective likelihoods.

Since the exact test can be computationally quite demanding for large numbers of observations (say, $N > 250$), we also present a sampling approach, where we obtain a Monte Carlo estimate of the p-value in (1) through sampling 2×2 tables from T. Since we can compute the full probability distribution $P(\cdot)$, we have a handle to assess the quality of the approximation and provide some guidelines on the number of samples needed to reach a desired level of precision.

When applying the exact test presented here and the common approach in the literature to simulated data, we found the former to be more robust: slight deviations in classification uncertainties can result in large differences in p-value when the common approach was applied, while when reasoning over all possible 2×2 tables the resulting p-values are stable. In addition, we applied both methods for testing LD between SNP-SNP and SNP-deletion pairs taking from the Genome of the Netherlands[2] project [7]. We found several pairs that are likely misclassified as being in LD or not in LD, since the classification uncertainties were not accounted for.

References

1. Chapman, S.J., Hill, A.V.S.: Human genetic susceptibility to infectious disease. Nature Reviews Genetics. 13, 175-188 (2012)
2. Slatkin, M.: Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics. 9, 477-485 (2008)
3. Fisher, R.A.: Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh (Originally published 1925, 14th ed. 1970.)
4. Agresti, A.: A survey of exact inference for contingency tables. Statistical Science. 7, 131-177 (1992)
5. Marschall, T., Herms, I., Kaltenbach, H.M. and Rahmann, S.: Probabilistic arithmetic automata and their applications. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 9(6), 1737-1750 (2012)
6. Nuel, G., Prum, B.: Analyse statistique des sequences biologiques. Hermes Science Publications (2007)
7. Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., et al.: The Genome of the Netherlands: design, and project goals. European Journal of Human Genetics (2013)

[2] http://www.nlgenome.nl/

A fast homotopy algorithm for a large class of weighted classification problems and application to phylogeny

Pierre Gutierrez[1,2], Julien Chiquet[2], and Guillem Rigaill[1]

[1] Unité de Recherche en Génomique Végétale INRA-CNRS-Universitée d'Évry Val d'Essonne, Évry, France
[2] Laboratoire Statistique et Génome UMR CNRS 8071-USC INRA-Universitée d'Évry Val d'Essonne, Évry, France

Abstract. We propose a fusion penalty to aggregate a large number of groups starting from multidimensional quantitative data. In the unidimensional case it boils down to solve the one-way ANOVA problem by collapsing the coefficients of K conditions. We introduce a large class of weights for which our homotopy algorithm is in $\mathcal{O}(K\log(K))$. These weights induce a balanced tree structure and simplify the interpretation of the results. Some of these weights also enjoy asymptotic oracle properties. As an example we consider phenotypic data: given one or several traits, we reconstruct a balanced tree structure and assess its agreement with the known phylogeny.

Background

With the advent of new high-throughput technologies, it is possible to compare features across a very large number, K, of conditions. Considering for instance the case of one single feature, one typically applies one-way ANOVA to test for any significant difference between conditions. Large K leads to multiple-testing and algorithmic problems since the number of pairwise tests is in $\mathcal{O}(K^2)$. Furthermore, each test is performed independently and the resulting structure between the conditions is not necessarily simple and easily interpretable.

In this work, we propose a ℓ_1-fusion penalty achieving these goals by constructing a hierarchical structure on the conditions at a low computational cost. Our penalty collapses the coefficients within the conditions in the same manner as the fused-Lasso [1]. We prove that for a large class of weights no split can occur along the path of solutions. These weights lead to a balanced tree structure. Besides adaptive versions of these weights enjoy asymptotically an oracle property. This guarantees selection of the true underlying structure for an appropriate choice of the tuning parameter λ which control the level of aggregation.

In the unidimensional settings, an analogous strategy called "Cas-ANOVA" has been investigated in [2] for multi-factor ANOVA. They propose some weights which enjoys similar asymptotic consistency. Still, these weights to do not lead to a tree as soon as the number of individual by condition is unbalanced. Moreover,

the optimization procedure of [2] is quadratic in K and only provides the solution for a given λ. We also experienced numerical instability using their weights.

In the multidimensional setting a similar penalty was proposed in [3]. When there is just one individual per condition and for fixed weights equal to one, they showed that no split can occur along the path of solutions and proposed an efficient algorithm. However these weights typically lead to unbalanced hierarchies. We extend their results to the case of several individuals per condition and to a larger class of weights that induces a balanced tree structure.

Fast homotopy algorithm for distance decaying weights

The optimization problem that we consider can be solved by the homotopy algorithm proposed in [4]. For unspecified weights, split events may occur in this algorithm. However, the absence of splits is highly desirable because if there is no split,

1. the order of the estimated means always matches the order of the empirical means of each condition;
2. the recovered structure is a tree which simplifies the interpretation;
3. the total number of iterations is guaranteed to be small and equal to K;
4. we avoid maximum flow problems whose resolution is computationally demanding.

We prove that for a large class of weights that induced a balanced tree structure there can be no split in the path of solutions. We implemented both the general and the without split version of the algorithm in C++. For the latter, the complexity of our implementation is $\mathcal{O}(K \log K)$. We also provide a fast cross validation (CV) procedure to select λ. The main idea behind this procedure is to take advantage of the DAG structure of the path of solutions along λ to avoid unnecessary computations.

Application to phylogenetic data

We applied our penalized approach to aggregate various species of bacteria based on one ore several features. We demonstrate the good agreement between our classification and the official phylogeny.

References

1. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B **67**(1) (2005) 91–108
2. Bondell, H., Reich, B.: Simultaneous factor selection and collapsing levels in anova. Biometrics **65**(1) (2008) 169–177
3. Hocking, T., Vert, J.-P.and Bach, F., Joulin, A.: Clusterpath: an algorithm for clustering using convex fusion penalties. In: Proceedings of the 28th ICML. (2011) 745–752
4. Hoefling, H.: A path algorithm for the fused lasso signal approximator. Journal of Computational and Graphical Statistics **19**(4) (2010) 984–1006

The leave-p-out estimator of the prediction error as a U-statistic and its asymptotic tests

Mathias Fuchs

Institut für Medizinische Informationsverarbeitung Biometrie und Epidemiologie,
Ludwig-Maximilians-Universität München,
Marchioninistr. 15, 81377 München, Germany

Abstract. We outline some consequences of identifying the leave-p-out estimator of the generalization error rate of a machine learning algorithm or the prediction error of a model as a U-statistic. In particular, asymptotic normality holds true, in contrast to usual cross-validation.
Furthermore, an appropriate variance estimator leads to an asymptotically exact test of the null hypothesis that two such algorithms have equal error rate. Such a test seems to be new even for cases of few variables.
We present an application to tuning parameter choice in lasso regression on a gene expression data set and conclude with a discussion of computational aspects of the leave-p-out-statistic.

Annex

Submitted posters

Statistical Integration of Genotypic and Phenotypic Data Towards Starter Prediction

Sandrine Laguerre [1,2,3], Amandine Dhaisne [1,2,3], Muriel Cocaign-Bousquet [1,2,3]

[1] Université de Toulouse; INSA, UPS, INP; LISBP, Toulouse, France

[2] INRA, UMR792 Ingénierie des systèmes biologiques et des procédés, Toulouse, France

[3] CNRS, UMR5504, Toulouse, France

Keywords: Principal Component Analysis, Hierarchical Clustering, Genotype, Phenotype, Lactococcus lactis

The mesophilic lactic acid bacterium *Lactococcus lactis* is one of the most extensively exploited microorganisms. It is used in particular in the manufacture of dairy products. In this work, the diversity of nine dairy strains of *Lactococcus lactis* subsp. *lactis* in fermented milk was investigated by both genotypic and phenotypic analyses. The objective was to classify well known strains in order to position new strains and evaluate their originality for further developments in dairy products.

Pulsed-field gel electrophoresis and multilocus sequence typing were used to establish an integrated genotypic classification.

To assess phenotypic diversity, 82 variables were selected as important dairy features. They included physiological descriptors and the production of metabolites and volatile organic compounds (VOCs). Principal component analysis (PCA) and variables selection were used to compare phenotypes and reduce the amount of variables to be measured when another strain will have to be tested.

Genotypic and phenotypic data were included in the same classification to provide an integrated picture of the diversity of the studied strains.

This work proposes an original method for the differentiation of closely related strains in milk and may be the first step toward a predictive classification for the manufacture of starters.

REFERENCES

1. Dhaisne A, Guellerin M, Laroute V, Laguerre S, Cocaign-Bousquet M, Le Bourgeois P, Loubiere P.: Genotypic and phenotypic analysis of dairy Lactococcus lactis biodiversity in milk: volatile organic compounds as discriminating markers. Appl Environ Microbiol. 2013 Aug;79(15):4643-52.

Cytopathic Effects Influence the Phenotype

Bettina Knapp[1,2,*], Kathleen Börner[3,*], Christoph Sommer[4], Petr Matula[5], Chenchen Zhu[3], Bärbel Glass[3], Diana Schwendener Forkel[3], Christine E. Engeland[3], Julian Kunkel[6], Nigel P. Brown[6], Johannes Hermle[3], Nina Beil[7], Jürgen Beneke[7], Karl Rohr[6], Christian Lawerenz[6], Fred Hamprecht[4], Dirk Grimm[3], Holger Erfle[7], Lars Kaderali[2], Maik J. Lehmann[3], Hans-Georg Kräusslich[3]

[1] Institute of Computational Biology (ICB), German Research Center for Environmental Health (GmbH), Helmholtz Zentrum München, München, Germany
[2] Institute for Medical Informatics and Biometry, Medical Faculty Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
[3] Department of Infectious Diseases, Virology, Heidelberg University Hospital, Heidelberg, Germany
[4] Insitute of Biochemistry, Swiss Federal Institute of Technology Zurich, Zurich, Switzerland
[5] Department of Computer Graphics and Design, Faculty of Informatics, Masaryk University, Brno, Czech Republic
[6] BioQuant, University of Heidelberg, Heidelberg, Germany
[7] CellNetworks RNAiScreening Facility, Bioquant, University of Heidelberg, Heidelberg, Germany*These authors contributed equally

Abstract. Previously published RNA interference (RNAi) screens aimed at identifying HIV-1 host factors show only little overlap, despite studying the same viral target. One possible explanation for this oddity is that for cellular screening data, the population context of a single cell largely influences the phenotypic outcome. Here, we analyze cellular phenotypes in an in-house HIV-1 RNAi screen and distinguish desired effects due to target gene silencing from false-positives caused by the viral infection itself (cytopathic effects). Our data shows that for HIV-1, specific RNAi phenotypes can indeed be enhanced by correcting for virus-induced cytopathic effects such as fusion of infected cells into multinucleated super cells (syncytia). In the future, this knowledge will facilitate the identification and segregation of host factors which control either the viral life cycle or secondary cellular phenotypes such as syncytia formation.

1 Correcting for Virus Induced Effects Can Mask the Phenotype

In HIV-1 research the dynamics of individual steps of virus replication, as well as the involvement of host factors remains largely unknown. Recent advances in screening technologies and high-resolution live microscopy have enabled the development of new approaches towards a better understanding of the cellular machinery in HIV-1 infections. Surprisingly, three different published large-scale RNAi screens have only a few identified factors in common [1]. One of the reasons for this small overlap might be cell population context effects which have been shown to greatly influence the phenotype (e.g. [2, 3]). We are studying the HIV mediated effect of RNAi screening data using four different settings: 1) GFP-control cell line which is a virus-free control to evaluate the virus independent effects of a siRNA knockdown on the GFP expression. 2) VSV-G-pseudotyped HIV-1-AGFP cells which show single round infections. Virus particles

bypass the normal HIV-1 receptor-ligand-mediated entry steps and do not form syncytia and consequently can be used for controlling the effect of the fusion of infected cells. 3) HIV HelaP4 cells which show the full infectious cycle and thus, are used to determine the HIV replication based GFP expression. 4) HIV-Ago-2 cells which allow an improved knockdown capacity of the siRNAs.
For the data analysis we use an approach published by Knapp *et al.* [2] which takes the population context features (such as cell size, nucleus size, cell shape) and technical features of each cell into account. The data is normalized against the features to correct for the cytopathic effects such as the formation of syncytia which influence the phenotype. Then an enrichment score (ES) is computed for each siRNA to determine the effect size of a knockdown. Doing this for each of the four experimental settings explained above, results in considerably different observed effects of individual perturbations in each screen. These phenotypic differences are induced by the different types of HIV infections in the four screens. This means that the effect on viral infection of a knockdown can be masked by correcting against the local population context which itself, can be a virus induced effect rather than a pure population context effect. To correct for this, the ES of each siRNA of the GFP-control screen is used as a baseline effect to normalize the ES of the corresponding siRNA in the HIV-1-AGFP and Ago-2 screens. Thereby, a normalization against cell population context and technical features is combined with a normalization against effects induced by the virus.
The corrected ES are furthermore summarized for each gene taking the median of the replicates. A one-sample, two-sided Welchs t-test is used to compute significance values of genes having an influence on HIV-1 infection. Hit genes are defined to have a FDR corrected p-value smaller or equal than a significance level of 0.2 and an absolute median ES smaller than a certain threshold. This threshold is determined by the standard deviation of the ES of the GFP screen.
Using this approach, we are able to identify host factors such as FASN and PSMD14 which influence HIV-1 infection. Another example is BRD4 for which we unraveled a function in the post-entry steps of infection. As a whole, these observations, which are supported by additional experimental validation, help to understand the HIV-1 life cycle in more detail.

References

[1] Bushman, F.D., Malani, N., Fernandes, J., D'Orso, I., Cagney, G., Diamond, T.L., Zhou, H., Hazuda, D.J., Espeseth, A.S., Konig, R., Bandyopadhyay, S., Ideker, T., Goff, S.P., Krogan, N.J., Frankel, A.D., Young, J.A., Chanda, S.K.: Host cell factors in HIV replication: meta-analysis of genome-wide studies. PLoS Pathog. **5**(5) (May 2009) e1000437

[2] Knapp, B., Rebhan, I., Kumar, A., Matula, P., Kiani, N.A., Binder, M., Erfle, H., Rohr, K., Eils, R., Bartenschlager, R., Kaderali, L.: Normalizing for individual cell population context in the analysis of high-content cellular screens. BMC Bioinformatics **12** (2011) 485

[3] Snijder, B., Sacher, R., Ramo, P., Liberali, P., Mench, K., Wolfrum, N., Burleigh, L., Scott, C.C., Verheije, M.H., Mercer, J., Moese, S., Heger, T., Theusner, K., Jurgeit, A., Lamparter, D., Balistreri, G., Schelhaas, M., De Haan, C.A., Marjomaki, V., Hyypia, T., Rottier, P.J., Sodeik, B., Marsh, M., Gruenberg, J., Amara, A., Greber, U., Helenius, A., Pelkmans, L.: Single-cell analysis of population context advances RNAi screening at multiple levels. Mol. Syst. Biol. **8** (2012) 579

An exact method to calculate the expected allelic diversity

Luis Alberto Garcia Cortes[1] and M. Angeles R. de Cara[2]
[1] Dept. Mejora Genética Animal
Instituto Nacional de Investigación Agraria y Alimentaria (INIA)
Ctra. de La Coruña km. 7.5, Madrid 28040 Spain.
[2] Laboratoire d'Eco-anthropologie et Ethnobiologie
UMR 7206 CNRS/MNHN/Universite Paris 7
Museum National d'Histoire Naturelle
CP 139 57 rue Cuvier F-75231 Paris Cedex 05 France

Abstract

Allelic diversity or allelic richness is the number of different alleles averaged over a range of genes. It is commonly used as a measure of genetic diversity in populations of different sizes via the rarefaction, as it provides a better comparison between them than other measures of genetic diversity like expected or observed heterozygosity. Allelic diversity is crucial in long term conservation programmes, as it is more sensitive to historical changes in population size. Furthermore, it determines the limit of the selection response.

Predicting the loss of alleles for a given pedigree is a difficult task. It is usually carried out by computer simulations or gene dropping, that is, simulating gene flow between founder alleles down the pedigree and averaging over replicates. Usually, the loss of variability is obtained by assuming that each founder carries two different alleles at each locus. This idealised case assumes an infinite number of alleles in the population where the founders were sampled.

Here we will describe an exact method to calculate the expected number of surviving alleles in the current cohort with a known pedigree. For this purpose, we calculate all the probabilities of each possible state of identity for multiple relationships, which has partially been resolved before by Thompson (1974) and Karigl (1982). As expected, we will show that calculations grow exponentially with the number of individuals in the current cohort. Therefore, the use of such exact method is limited due to computational demmands. However, its implications are of great importance, as it can potentially handle including inbred individuals in the affected pedigree methods.

Keywords: identity by descent, identity coefficients, allelic diversity

Tracts of identity: length distribution, evolutionary history and applications

M. Angeles R. de Cara[1], Frederic Austerlitz[1] and Luis Alberto Garcia Cortes[2]
[1] Laboratoire d'Eco-anthropologie et Ethnobiologie
UMR 7206 CNRS/MNHN/Universite Paris 7
Museum National d'Histoire Naturelle
CP 139 57 rue Cuvier F-75231 Paris Cedex 05 France
[2] Dept. Mejora Genética Animal
Instituto Nacional de Investigación Agraria y Alimentaria (INIA)
Ctra. de La Coruña km. 7.5, Madrid 28040 Spain.

Abstract

The study of haplotypes, their linkage disequilibrium structure and diversity patterns has only become feasible with the advent of high density genotype and now whole genome sequence data. This has reopened the theoretical questions opened by Fisher in 1954 about the distribution of junctions and the distribution of lengths of segments of identity, which so far, had been mostly explored theoretically. The power of such dense data sets to infer evolutionary processes remains to be explored, as it is unclear how to distinguish demographic processes from selective processes, let alone the two processes occurring together. Here we present simulation results for the distribution of lengths of segments of identity under neutrality, with bottlenecks and with selection on polygenic traits. These distributions not only provide insight into the history of the populations, but, furthermore, they appear as a useful tool to measure inbreeding and coancestry in the populations for which genealogies are not available. Therefore, our results are relevant not only in the context of disentangling the processes that the populations have undergone, but can also provide helpful insights on how the populations may respond to future changes in their environment.

Keyowords: identity by descent, runs of homozygosity, statistical inference

The impact of agricultural emergence on the genomes of African rainforest hunter-gatherers and agriculturalists

Etienne Patin[1,2], Katherine Siddle[1,2], Guillaume Laval[1,2], Hélène Quach[1,2], Christine Harmant[1,2], Noémie Becker[3,†], Alain Froment[4], Béatrice Régnault[5], Laure Lemée[5], Simon Gravel[6], Jean-Marie Hombert[7], Lolke Van der Veen[7], Nathaniel J. Dominy[8], George H. Perry[9], Luis B. Barreiro[10], Paul Verdu[3], Evelyne Heyer[3], Lluís Quintana-Murci[1,2]

[1]Institut Pasteur, Unit of Human Evolutionary Genetics, 75015 Paris, France; [2]Centre National de la Recherche Scientifique, URA3012, 75015 Paris, France; [3]CNRS, MNHN, Université Paris Diderot, Sorbonne Paris Cité, UMR7206, 75005 Paris, France [4]IRD, MNHN, CNRS UMR 208, 75005 Paris, France; [5]Genotyping Platform, Institut Pasteur, 75015 Paris, France; [6] McGill University and Genome Quebec Innovation Centre, H3A 1A4 Montréal, Canada; [7]Dynamique du Langage, CNRS UMR 5596, Université Lumière-Lyon 2, 69007 Lyon, France; [8]Department of Anthropology, Dartmouth College, Hanover, NH 03755, USA; [9]Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA 16802 USA; [10]Université de Montréal, Centre de Recherche CHU Sainte-Justine, H3T 1C5 Montréal, Canada

1. Abstract

The emergence of agriculture in West-Central Africa, ~5,000 years ago, profoundly modified the cultural landscape and mode of subsistence of most human sub-Saharan populations. How this major innovation has impacted the genetic history of rainforest hunter-gatherers — historically referred to as “pygmies” — and agriculturalists, however, remains poorly understood. Here, we report genome-wide SNP data from these populations located west-to-east of the equatorial rainforest. We find that hunter-gathering populations present up to 50% of farmer genomic ancestry, and that substantial admixture began only within the last 1,000 years, estimated from the observed decay of admixture linkage disequilibrium. Furthermore, we show that the historical population sizes characterising these communities already differed before the introduction of agriculture, by fitting observed and simulated genome-wide levels of linkage disequilibrium. Our results suggest that the first socio-economic interactions between rainforest hunter-gatherers and farmers introduced by the spread of farming were not accompanied by immediate, extensive genetic exchanges and occurred on a backdrop of two groups already differentiated by their specialisation in two ecotopes with differing carrying capacities.

Modeling the evolution of gene relationships

Magali Semeria, Laurent Guéguen, Eric Tannier

Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Lyon 1, 69622 Villeurbanne, France

Abstract. A genome is not only a set of independent genes. It can be seen as an organized and functioning set of interactions between genes that are embedded in their environment. To study the evolution of genomes, some integrative methods have been developed that reconstruct species and gene history simultaneously, and take into account sequence evolution, gene birth, duplication, transfer and loss. These methods give a first hint at the gene content of ancestral species, but since they assume that every gene evolves independently from each other, they do not give information about the relationships between these genes.

We propose a method to model the evolution of these relationships. Our method aims to be general but, at first, we apply it to modeling the evolution of gene adjacencies on chromosomes. Given a species tree, a set of gene trees, a set of present gene adjacencies, and a model of adjacencies evolution, we calculate the probability of adjacencies along the branches of the gene trees. As genes undergo evolutionary events such as duplication and rearrangement, adjacencies can be kept, gained or lost. The pseudo-likelihood of observed adjacencies can be computed using the usual dynamic algorithm proposed by Felsenstein.

We thus establish the methodological basis that will enable us to integrate information about gene relationships into models of genome evolution. This information will allow us to improve the reconstruction of gene and species histories. It will also be helpful to reconstruct the genomes of ancestral species.

Meta-Analysis of incomplete Microarray Studies

Alix Leboucq, Anthony C. Davison, Darlene R. Goldstein

EPFL SB MATHAA STAT, station 8, 1015 Lausanne

Abstract. Meta-analysis of microarray studies to produce an overall gene list is relatively straightforward when complete data are available. When some studies lack information, for example, have only a ranked list of genes instead of complete primary data, it is common to reduce all studies to ranked lists prior to combining them. Since this entails a loss of information, we consider a hierarchical Bayes modeling approach to combine studies using the type of information available in each study: the full data matrix, summary statistics, or ranks for each gene. The model uses an informative prior for the parameter of interest, which eases the detection of differentially expressed genes. Simulations show that the new approach can give substantial power gains compared to classical meta analysis and list aggregation. A large meta-analysis based on 11 published studies providing data of the types cited above, and comparing serous ovarian cancer with normal tissue, is also performed.

References

1. C. M. Carvalho, N. G. Polson and J. G. Scott, the horseshoe estimator for sparse signals, Biometrika 97 (2), 465-480, 2010.
2. J. E. Griffin and P. J. Brown, Inference with normal-gamma prior distributions in regression problems, Bayesian Analysis 5(1), 171-188, 2010.
3. H. Ishwaran and J. S. Rao, Spike and Slab Gene selection for multigroup microarray data, Journal of the American Statistical Association 100(471), 764-780, 2005.

Causal inference of gene expression in a neoadjuvant phase II trial of breast cancer

Nils Ternès[1,2], Stefan Michiels[1,2], Serge Koscielny[1], Emilie Lanoy[1]

[1] Gustave Roussy, Service de biostatistique et d'épidémiologie, Villejuif, F-94805, France

nils.ternes@gustaveroussy.fr

[2] Univ. Paris-Sud, Le Kremlin-Bicêtre, F-94276, France

Abstract. Omics technologies have become an essential part of clinical trials in oncology to provide a better comprehension of molecular mechanisms, which implies adequate modeling of the omics data. High-dimensional omics data have some statistical characteristics: high number of measured variables, multi-colinearity between variables and confounding bias inherent to the observational setting. To overcome the limitations of standard methods in this context, non-causal and causal methods have been proposed and are reviewed here.

We described non-causal methods including penalized regressions (Ridge, Lasso and Elastic-net) which constraint model coefficients of standard regression in order to ease variable selection and to improve prediction accuracy. We also considered causal methods (IDA and CStaR) which aim to estimate causal effects in presence of confounding.

In a non-randomized neoadjuvant phase II study of letrozole that included 56 women with invasive breast cancer, the gene expression measurements were available from 22,283 probes sets before treatment, after 10 to 14 days of treatment (early change) and after 3 months of treatment (late change). The objective was to identify early markers of the treatment response, defined as late change in the expression of a proliferation gene (AURKA). Studied markers were the early change in expression of all candidate genes. Except for IDA, the lists of genes with top effects found with causal and non-causal methods were very similar.

In this particular example, causal inference methods did not outperform the non-causal counterparts.

Keywords: causal inference, penalized regression, gene expression, breast cancer, clinical trial

Allele Scoring to investigate shared risk loci in comorbid disorders

Jack Euesden[1], Jelmar Quist[2], Cathryn Lewis[1,2], and RADIANT Depression Consortium[1]

[1] Social, Genetic and Developmental Psychiatry Centre, King's College London
[2] Division of Genetics and Molecular Medicine, King's College London

Abstract. Major Depressive Disorder (MDD) is frequently reported as comorbid with other disorders - that is to say, is found alongside another disorder within individual patients more often than would be predicted by chance. These range from physical disorders, such as Rheumatoid Arthritis, to psychiatric disorders, such as Schizophrenia (SCZ). In this paper, we use a novel form of risk profile scoring in order to explore the genetic basis of the comorbidity between MDD and SCZ.

1 Introduction

In order to investigate the evidence for shared risk genes between SCZ and MDD, we adapted a widely used statistical genetics technique, Polygenic Risk Scoring (PRS), [1]. This requires SNPs from an independent discovery dataset, clumped for LD to achieve linkage equilibrium. When investigating a disorder such as SCZ, this becomes concerning, as it is associated with genotype at loci in the Major Histocompatability Complex (MHC) [2]. The MHC has such high short-range linkage disequilibrium that it is often omitted from PRS in order to obtain linkage equilibrium. We propose a novel method to circumvent this limitation, using allelic information from imputation at the MHC in order to study risk conferred by individual loci there.

2 Methods

We used classical PRS to calculate risk profiles for SCZ in MDD patients, using publicly available SCZ GWAS results as a discovery dataset [2] and the RADIANT MDD samples as a test dataset [3]. We combined these predictions with HLA alleles weighted by their association with SCZ, as reported by the ISC [4]. We can use penalised regression to fit models of SCZ risk at each HLA locus to minimise confounding due to correlated predictors, whilst investigating association between SCZ risk alleles and MDD.

3 Results

The PRS for SCZ showed significant association with MDD status, explaining a modest but significant proportion of variance in MDD status. We found evidence that SCZ risk in HLA loci also predicted MDD, despite the predictive signal captured by PRS.

4 Discussion

The concept of generalised psychiatric liability, p, has been proposed to explain studies finding genetic overlap between psychiatric phenotypes. It would be possible to explore this by expanding our approach to investigate other psychiatric phenotypes, such as bipolar disorder, which have not yet been associated with immune dysregulation.

5 Conclusion

Here we have developed a novel method for incorporating HLA allelic data into risk prediction models. We have found evidence for a shared genetic substrate between SCZ and MDD, not just genome-wide, but also in part localised to HLA alleles.

Acknowledgments

The authors would like to thank all participants whose data was used in this analysis. This research was supported by an MRC PhD Studentship to JE.

References

1. Dudbridge, Frank. (2013). PLoS Genetics, 9(3).

2. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kahler, A. K., Akterin, S., . . . Sullivan, P. F. (2013). Nat Genet, 45(10), 1150-1159.

3. Lewis, C. M., Ng, M. Y., Butler, A. W., Cohen-Woods, S., Uher, R., Pirlo, K., . . . McGuffin, P. (2010). American Journal of Psychiatry, 167(8), 949-957.

4. Irish Schizophrenia Genomics, Consortium, & the Wellcome Trust Case Control, Consortium. (2012). Biological psychiatry, 72(8), 620-628.

Genotype calling in polyploid (and pooled) genomes.

Emanuele Raineri, Luca Ferretti, Sebastian Ramos-Onsins

CNAG (Barcelona), Collège de France(Paris),CRAG(Barcelona)

Abstract. We would like to outline, in a brief talk, some thoughts on how to infer genotypes (and hence, SNPs/SNVs) from sequenced polyploid genomes and from pools of genomes belonging to different individuals that have been sequenced together. In both cases there are two interesting technical questions to ponder upon : the first is how to model the variability inherent in the sequencing/mapping pipeline; the second relates to the combinatorial aspects of dealing with reads which may not be evenly distributed across all the homologous chromosomes, or across the different individuals in a pool. These are topics which are interesting per se but also have practical applications for example in calling variants when the ploidy is partially unknown, as is the case for certain human cancer samples. Our description will be very concrete, down to presenting the algorithms we have devised so far; we will refer to already published material and to work in progress as well.

Analysis of genomic markers: make it easy with the R package MPAgenomics

Quentin Grimonprez[1], Alain Celisse[1,2], Serge Iovleff[1,2], Guillemette Marot[1,3]

[1] Modal team, Inria Lille-Nord Europe, France
[2] Laboratoire Paul Painlevé, Université Lille 1, France
[3] EA 2694, Université Lille 2, France

Abstract. `MPAgenomics`, standing for multi-patients analysis of genomic markers, is an R-package which enables to study several copy number and SNP data profiles at the same time. It offers wrappers for commonly used packages (`aroma` [Bengtsson, 2004], `changepoint` [Killick *et al.*, 2013], `cghcall` [van de Wiel *et al.*, 2007] and `glmnet` [Friedman *et al.*, 2010]), providing a pipeline for beginners in R. Using `MPAgenomics`, normalization, segmentation and calling of copy-number or SNP data profiles can be easily performed at the same time. The special architecture of [Bengtsson, 2004] packages is automatically created and outputs to right formats for further analyses are suggested. Therefore, more advanced users can use their own method at one point if they prefer to separate the global analysis in several steps. In addition to these useful wrappers, we propose a strategy to improve the choice of the penalty parameter used in [Killick *et al.*, 2013] for segmentation. As far as multi-patients analysis is concerned, we suggest to select relevant markers associated with a given response thanks to wrappers with the R-packages `glmnet` [Friedman *et al.*, 2010] and `HDPenReg` (still under development on the R-forge). Functionalities already implemented in our package `HDPenReg`, standing for penalized regression in high dimension, will be briefly presented.

References

[Bengtsson, 2004] Bengtsson,H. (2004) aroma - An R Object-oriented Microarray Analysis environment, *Preprint in Mathematical Sciences*.

[Killick *et al.*, 2013] Killick,R., Eckley,E.(2013) changepoint: An R package for changepoint analysis.

[van de Wiel *et al.*, 2007] van de Wiel,M. *et al.* (2007) CGHcall: Calling aberrations for array CGH tumor profiles, *Bioinformatics*, **23**, 892-894.

[Tibshirani , 1994] Robert Tibshirani (1994) Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

[Friedman *et al.*, 2010] Friedman J. *et al* (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, **33**, 1-22.