Banff International Research Station

for Mathematical Innovation and Discovery

Bioinformatics, Genetics and Stochastic Computation: Bridging the Gap
July 1-6 2007

Closing report

The meeting that took place at BIRS, Banff, on July 1-6 2007, gathered 33 people from America (Canada and USA), Europe, Japan and Australia. It was quite successful, thanks to the superb organisation of the Center and the friendly help of the staff, and all invited participants showed up for the meeting. The talks were uniformly well-attended by an overwhelming number of the participants who stayed the whole week, with two exceptions only. There was no mountain accident or bear encounter to deplore, and the organisers were not aware of any complaint from the participants, during or after the meeting, but instead gathered many thanks and requests for a follow-up meeting. Several collaborations were initiated during the meeting as well.

OBJECTIVES

On the one hand, there is an explosion of complex statistical models appearing in genetics and bioinformatics. These typically highly structured systems are unfortunately very difficult to fit. On the other hand, there has been recently significant advances on MC methods but most of these methods are unknown to the applied community and/or formulated in ways that are too theoretical for direct application. By providing scientists with improved inferential methods it would allow them to consider richer models, which are more realistic than those dictated by computational constraints. The exchanges between the applied and methodological communities remain surprisingly limited.

We believe that this workshop would be an ideal place to (1) Gather people from different research communities and foster links between these communities (applied, methodological and theoretical). (2) Expose the applied community to novel statistical methodologies and advanced MC methods, and expose the MCMC community to the specifics of the complex modelling problems met in bioinformatics and

genetics. (3) Classify topologies of computational problems met by bioinformatics and genetics, and equip all participants to the workshop with benchmark problems, if possible before the workshop.

We anticipate this workshop to create an exceptional opportunity for exchanging ideas between the communities; and help to shape the future of stochastic computation within bioinformatics and genetics. Input from statisticians working in bioinformatics and genetics is absolutely crucial for development of appropriate statistical methodologies. We will encourage researchers to bring data from their own field, which could be used to implement methods and try new algorithms. We have targeted areas of stochastic computation that are of great interest to practitioners such as automatic algorithms, computational issues and parallel implementations. Bioinformatics and genetics are relatively new fields that enjoy a high representation of young talents, including many women. This workshop would be a great learning/training environment for these new talents.

SCHEDULE

Sunday

16:00-17:00	Check-in begins (Front Desk - Professional Development Centre - open 24 hours) Lecture rooms available after 16:00 (if desired)
17:00-18:00	Welcome
18:00–19:30	Buffet Dinner, Donald Cameron Hall
20:00	Informal gathering in 2nd floor lounge, Corbett Hall

Monday

7:00–8:45	Breakfast
8:45–9:00	Introduction and Welcome to BIRS by BIRS Station Manager, Max Bell 159
9:00–10:00	Keynote lecture I: David Balding, "ABC methods in population genetics"
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:00	Jean-Michel Marin "Adaptive Multiple Importance Sampling"
10:35–11:10	Ajay Jasra, "The Time Machine: A Simulation approach for the coalescent"
11:10–11:45	Anthony Brockwell, "Universal Residuals: A Multivariate Transformation"
11:45–13:00	Lunch
13:00–14:00	Guided Tour of The Banff Centre; meet in the 2nd floor lounge, Corbett Hall
14:00	Group Photo; meet on the front steps of Corbett Hall
14:30–15:30	Keynote Lecture II: Chris Holmes, "Nonparametric (distribution free) model-based hierarchical clustering: a Bayesian approach for class discovery and feature selection in high-throughput genomics"
15:30–16:00	Coffee Break, 2nd floor lounge, Corbett Hall
16:00–16:35	Adrian Dobra, "Efficient Stochastic Search Algorithms for Large p Regression with Dependent Covariates"
16:35–17:10	Sunduz Keles, "Variable selection and Dimension Reduction in Genomics with Sparse Partial Least Squares"
17:10–17:45	Marina Vannucci, "Bayesian Methods for Genomics with Variable Selection"
17:45–19:30	Dinner

Tuesday

7:00–8:15	Breakfast
8:15–8:50	Dave Stephens, "Genome-wide association in the presence of high linkage disequilibrium"
8:50–9:25	Chiara Sabatti, "Sampling contingency tables and linkage disequilibrium"
9:25–10:00	Maria De Iorio, "A Bayesian Model for Phylogenetic Footprinting"
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:30	Keynote Lecture III: Elizabeth Thompson, "Relationships within and among populations: inference from genomic data"
11:30–13:30	Lunch
13:30–15:30	Break
15:30–16:00	Coffee Break, 2nd floor lounge, Corbett Hall
16:00–16:35	Jonathan Keith, "Genome Segmentation with the Generalised Gibbs Sampler"
16:35–17:10	Kerrie Mengersen, "Sensitivity of priors in Bayesian analysis of DNA sequence segmentation"
17:10–17:45	Mayetri Gupta, "Improving detection of DNA sequence motifs using chromatin structure information"
17:45–19:00	Dinner
19:00–20:00	Keynote Lecture IV: Paul Fearnhead, "Efficient Bayesian Methods for Segmenting Genetic Sequences"

Wednesday

7:00–8:15	Breakfast
8:15–8:50	Kevin Murphy, "Array CGH data analysis"
8:50–9:25	Francois Carron, "A change point model for detecting novel RNA transcripts"
9:25–10:00	Peter Green, "On clustering gene expression profiles using DP models"
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:05	Alex Lewin, "Model checks for complex hierarchical models"
11:05–11:40	Peter Mueller, "The Optimal Discovery Procedure and Bayesian Decision Rules"
11:40–13:30	Lunch
13:30–18:00	Rest or Hiking (including a possible hike up Mount Rundle led by Christian)
18:00–19:00	Dinner
19:00–20:00	Keynote Lecture V: Matthew Stephens, "Methods and models for Population Genetic Data"

Thursday

7:00–8:15	Breakfast
9:00–10:00	Keynote Lecture VI: Mike West, "Bayesian analysis and computation for stochastic models of dynamic cellular networks"
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:30	Darren Wilkinson, "The Chemical Langevin Equation: bridging many gaps"
11:05–11:40	Lurdes Inoue, "Functional Network".
11:40–12:30	Mark Beaumont, "Further advances in ABC"
12:30–13:30	Lunch
13:30–15:30	Break
15:30–16:00	Coffee Break, 2nd floor lounge, Corbett Hall
16:00–16:35	Christian Robert, "Non-informative priors for linear and generalised linear models"
16:35–17:10	Scott Schmidler, "Adaptive Markov Chain Monte Carlo for Bayesian Variable Selection"
17:30–19:00	Dinner
19:00–20:00	Keynote Lecture VII: Sylvia Richardson, "Fully Bayesian variable selection using g-priors"

Friday

7:00–8:15	Breakfast
8:15–8:50	Takashi Matsumoto, "On-line and Batch Inference for Bioinformatics Data"
8:50–9:25	Luke Bornn, "SMC for prior sensitivity analysis"
9:25–10:00	Closing remarks and informal discussion
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:30	Departs
11:30–13:30	Lunch

Checkout by 12 noon.

CONTENTS

With regards to the goal of the meeting replicated above, the meeting was clearly centred on the statistical aspects of *the complex models appearing in genetics and bioinformatics*, namely on the statistical methodology that allowed the participants to tackle the statistical analysis of those models. Most talks were therefore at the interface between statistical modelling, statistical methodology, and computational statistics. The new advances on Monte Carlo methods were indeed at the forefront of most talks, which mostly developed new tools and produced new results to handle their complex models, more realistic than those dictated by computational constraints. We believe that the presentations at the meeting enhanced the *exchanges between the applied and methodological communities*, thanks to the open schedule adopted by us. We indeed *gathered people from different research communities and foster links between these communities (applied, methodological and theoretical)*: D. Balding, M. Beaumont, R. Gottardo, M. de Iorio, S. Keles, S. Schmidler, and D. Stephens are primarily working in Genomics and, for some of them, are not statisticians, R. Craiu, A. Dobra, A. Doucet, P. Fearnhead P. Green, A. Jasra, J.-M. Marin, P. Müller, S. Richardson, C. Robert, C. Sabatti, M. Stephens, M. Vanucci, M. West and D. Wilkinson are mostly focussing on the theory of computation and of Bayesian inference, with forays into genetic and biological applications, while L. Bornn, A. Brockwell, F. Caron, M. Gupta, C. Holmes, J. Keith, A. Lewin, T. Matsumoto, K. Mengersen, K. Murphy, S. Schmidler, and E. Thompson, centre their research on the specific development of statistical methods for biological and genetic models, therefore being truly at the interface. Obviously, this classification in three classes is somehow arbitrary. The *exposure [of] the applied community to novel statistical methodologies and advanced MC methods,* was clear since five of the keynote speeches were dealing with methodological topics, mostly related to computational Statistics, and the *exposure [of] the MCMC community to the specifics of the complex modelling problems met in bioinformatics and genetics* was operated via most than half of the talks. We however failed short of *classify[ing] topologies of computational problems met by bioinformatics and genetics* in a coherent manner, and definitely did not *equip all participants to the workshop with benchmark problems,* due to the difficulty of collecting sufficiently challenging datasets that would appeal to every attendant. Although we did *encourage researchers to bring data from their own field, which could be used to implement methods and try new algorithms*, this alas did not happen. In retrospect, this aspect would have required a smaller number of participants and would have thus restricted the field processed by the meeting.

Similarly, the workshop did constitute *an exceptional opportunity for exchanging ideas between the communities*, as, again, shown by the involvement of all participants in every session of the meeting, despite outdoor temptations all around!. We cannot tell at this stage how much the workshop help in *shaping the future of stochastic computation within bioinformatics and genetics* but we believe major actors in this field took part in the meeting, including young talented researchers for whom this workshop truly was *a great learning/training environment.*

For instance, Ajay Jasra (Imperial College, London) presented an important advance for the processing of stochastic trees, which are so prevalent in (population) Genetics. The difficulty in handling the likelihood function was solved in this joint work with Maria de Iorio and Marc Chadeau from Imperial

using importance sampling and sequential Monte Carlo techniques. In order to handle the computational difficulty with simulating backward in time, Ajay Jasra introduced controlled approximations where the bias remained under control. Similarly, Luke Born (University of British Columbia) introduced sequential Monte Carlo methods towards computational gains in prior sensitivity, even in cases when the distribution of interest is not available analytically. Franǫis Caron (University of British Columbia) considered the problem of identifying novel RNA transcripts using tiling arrays. Standard approaches to this problem rely on the calculation of a sliding window statistic or on simple changepoint models. These methods suffer from several drawbacks including the need to determine a threshold to label transcript regions and/or specify the number of transcripts. He thus proposed a Bayesian multiple changepoint model to simultaneously identify the number of transcripts, the transcript boundaries and their associated levels. In addition, he presented a computationally efficient on-line algorithm which allows to jointly estimate both the changepoint locations and the model parameters. Using two publicly available transcription data sets, he compared his method to a common sliding window approach and a simple changepoint model. establishing that his on-line estimation procedure provides good estimates of transcript boundaries and model parameters. Alex Linwin presented a Bayesian hierarchical model for detecting differentially expressed genes using a mixture prior on the parameters representing differential effects. He formulated an easily interpretable 3-component mixture to classify genes as over-expressed, under-expressed and non-differentially expressed, and model gene variances exchangeably to allow for variability between genes. He showed how the proportion of differentially expressed genes, and the mixture parameters, can be estimated in a fully Bayesian way, extending previous approaches where this proportion was fixed and empirically estimated. Good estimates of the false discovery rates were also obtained. Different parametric families for the mixture components can lead to quite different classifications of genes for a given data set. Using Affymetrix data from a knock out and wildtype mice experiment, he showed how predictive model checks can be used to guide the choice between possible mixture priors. These checks showed that extending the mixture model to allow extra variability around zero instead of the usual point mass null fits the data better.

These talks were linked with the keynote talk of Matthew Stephens that very broadly set the challenges met in this area, as well as the directions for their resolution. Another related keynote speech was given by Peter Green (University of Bristol) on clustering gene expression profiles using Dirichlet process models. He introduced a Bayesian mixture model that allowed to express a gene expression profile across different experimental conditions as a linear combination of covariates characterising those conditions, plus error. In a standard Bayesian nonparametric formulation, the expectations and the error precisions of the expression measurements would jointly follow a Dirichlet process (DP). In this set-up the clusters generated by the process are a priori exchangeable. However in the gene expression context, it commonly occurs that some genes are not influenced by the covariates, but fall into a 'background' class. This calls for an extension to the DP model generating a background cluster that is not exchangeable with the others, and he also built regression on covariates characterising experimental conditions into the expectation structure. He defined a particular heterogeneous Dirichlet process as a mixture of a random point mass and a Dirichlet process. The location of the point mass has a partially degenerate distribution, allowing some regression coefficients to be fixed at zero for the background cluster. Standard posterior sampling methods for DP models can be extended to make use of this heterogeneous prior model. In particular, in the case of conjugacy, he thus generalised the partition Gibbs sampler/weighted Chinese restaurant process to this situation. The background or 'top-table' cluster can be identified in the posterior sample. He used a loss function approach following Lau and Green (2008) to derive a point estimate of the remaining clusters.

Another of the keynote talks was given by Mike West (Duke University) where he presented a wide ranging survey of the recent results he and his team obtained on statistical inference for dynamic cellular networks in systems biology. Advances in bioengineering technologies are generating the ability to measure increasingly high-resolution, dynamic data on complex cellular networks at multiple biological and temporal scales. Single-cell molecular studies, in which data is generated on the levels of expression of a small number of proteins within individual cells over time using time-lapse fluorescent microscopy, is one critical emerging area. Single cell experiments have potential to develop centrally in both mechanistic studies

of natural biological systems as well as via synthetic biology – the latter involving engineering of small cellular networks with well-defined function, so providing opportunity for controlled experimentation and bionetwork design. There is a substantial lag, however, in the ability to integrate, understand and utilise data generated from single-cell fluorescent microscopy studies. The talk highlighted aspects of this area from the perspective of Mike West's forays in single cell studies in synthetic bacterial systems that emulate key aspects of mammalian gene networks central to all human cancers. The most relevant aspects were about

1. The data in those studies come as movies of colonies of cells developing through time, with a need for imaging methods to estimate cell-specific levels of fluorescence measuring mRNA levels of one or several tagged genes within each cell. This is complicated by the progression of cells through multiple cell divisions that raises questions of tracking the lineages of individual cells over time.

2. In the context of their synthetic gene networks engineered into bacterial cells, they have developed discrete-time statistical dynamic models inspired by basic biochemical network modelling of the stochastic regulatory gene network. These models allow the incorporation of multiple components of noise that is "intrinsic" to biological networks as well as approximation and measurement errors, and provide the opportunity to formally evaluate the capacity of single cell data to inform on biochemical parameters and "recover" network structure in contexts of contaminating noise.

3. Last and not least, in their approaches to model fitting, they have developed Bayesian methods for inference in non-linear time series. This involves MCMC methods that impute parameter values coupled with novel, effective Metropolis methods for what can be very high-dimensional latent states representing the unobserved levels of mRNA or proteins on nodes in the network as well as contributions from "missing" nodes.

In connection with the keynote talk of Sylvia Richardson (Imperial College London) on the Bayesian and computational tools required to run models selection in "large p small n" linear models, namely models with many more covariates than observations, which requires the use of default priors across all models like Zellner's g-priors, Jean-Michel Marin (Université Paris Sud) presented a talk on an hierarchical extension of the g-prior that allowed for less informative inputs as well as computational gains, avoiding the recourse to complex techniques like reversible jump. In the same spirit, Marina Vanucci (Rice University) addressed in her talk methods for Bayesian variable selection for high-dimensional data. While initially dealing with the simple linear regression model, she extended the setup to probit models for classification and to clustering settings, as well as survival data. Her talk included many examples from genomics, in particular DNA microarray studies. In addition, the analysis of the high-dimensional data generated by her studies challenges standard statistical methods and she discussed the performances of those methods both on simulated and real data. Additional talks in this quite important area included Anthony Brockwell's, Adrian Dobra's, Sundunz Keles', Alex Lewin's, Peter Mueller's and Chiara Sabatti's.

In his keynote talk, Paul Fearnhead considered Bayesian analysis of a class of multiple changepoint models. While there are a variety of efficient ways to analyse these models if the parameters associated with each segment are independent, there are few general approaches for models where the parameters are dependent. Under the assumption that the dependence is Markov, he proposed an efficient online algorithm for sampling from the an approximation to the posterior distribution of the number and position of the changepoints. In a simulation study, Paul Fearnhead showed that the approximation introduced is negligible. He illustrated the power of his approach through fitting piecewise polynomial models to data, under a model which allows for either continuity or discontinuity of the underlying curve at each changepoint. This method is competitive with, or out-performs, other methods for inferring curves from noisy data; and uniquely it allows for inference of the locations of discontinuities in the underlying curve.

CONCLUSION

While quantifying the impact of a workshop always is a delicate task, we are convinced that this meeting has had an influence on the community of computational statisticians working in Bioinformatics and Genomics. The recent rise of ABC (standing for Approximate Bayesian Methods) methods can for instance be partly connected to the debate about this method initiated during the meeting after the talk of David Balding. Similarly, the current revival of Bayesian model choice evaluation has links with the talks of Jean-Michel Marin, Sylvia Richardson, and Scott Schmidler. The fact that many of us keep referring to this meeting as a highlight, even two years later, is also a significant indicator that some alchemy took on during the workshop, even though putting a finger on exactly what happened is not possible. Once again, we are immensely grateful to PIMS for its support and trust, as well as to the BIRS centre and its friendly staff for a superb organisation that let us concentrate 110% on scientific issues.

Arnaud Doucet
Raphael Gottardo
Christian P. Robert