

THE ISBA BULLETIN

Vol. 13 No. 4

December 2006

The official bulletin of the International Society for Bayesian Analysis

A MESSAGE FROM THE PRESIDENT

by Alan Gelfand
ISBA President
alan@stat.duke.edu

I write this as my last presidential communication to the membership. I want to update you on decisions made by the Board with regard to your ISBA dues invoice for next year (and, in fact, for the next three years). With regard to dues, we have approved the following rate schedule: 2007 - 35 USD, 2008 - 35 USD, 2009 - 40 USD. As always, we can make special dispensations to those for whom these amounts are a burden. Following up on the issue of maintenance of Bayes's gravesite, we will include a tick box on the dues invoice, with space for you to indicate a donation amount. With regard to the prize funds (Savage, Lindley, DeGroot, and Mitchell), there will be one tick box on the invoice

for you to indicate a contribution, with space for a donation amount. All donations will be distributed equally across the four prize funds unless otherwise designated. Finally, we would like to maintain our choices of exciting and attractive conference venues and still be able to facilitate attendance for young researchers. Hence, the invoice will have a tick box with space for you to indicate a donation amount to help provide such support. Of course, all contribution decisions are individual.

I have enjoyed my year as ISBA President and have appreciated the help of the various board and committee members in making decisions and taking needed actions. Altogether, it is an experience I highly recommend! I believe ISBA is on a good course and I hope it will, increasingly, be the professional society that you most identify with. Finally, let me wish you all a happy, healthy holiday season.

- Alan E. Gelfand, ISDS, Duke University.

A MESSAGE FROM THE EDITOR

by J. Andrés Christen
jac@soe.ucsc.edu, jac@cimat.mx

As with every December we have elections results. The new appointments are: President-Elect, Christian Robert; Executive Secretary, Robert L. Wolpert; Directors, David Heckerman, Xiao-Li Meng, Gareth Roberts and Alexandra Schmidt. Congratulations to all and thanks to Alan Gelfand that is becoming past President this January. And good luck to Peter Green that is starting the ISBA presidency next month.

I hope you find this issue of the ISBA Bulletin in-

teresting, we have an applications section and also some News from the world with some exciting upcoming Bayesian events. Please feel free to send me, or any of the AE, suggestions for articles (or articles!) to be included in the ISBA Bulletin. Feedback and opinions are much needed to make the ISBA Bulletin of most interest to all of ISBA members.

Contents

- ▶ APPLICATIONS
 • Page 2
- ▶ NEWS FROM THE WORLD
 • Page 5

SUGGESTIONS

PLEASE, FEEL COMPLETELY FREE TO SEND US SUGGESTIONS THAT MIGHT
IMPROVE THE QUALITY OF THE BULLETIN

jac@cimat.mx

BAYESIAN MODELS FOR MOTIF DISCOVERY FROM CHIP-CHIP AND SEQUENCE DATA

Jonathan A.L. Gelfond and Mayetri Gupta

jgelfond@bios.unc.edu

gupta@bios.unc.edu

In the post genomic era, when the DNA sequence of many species and the approximate locations of the genes within are known, understanding interactions between DNA and proteins that control gene transcription is imperative. DNA interacts with proteins called transcription factors which play an important role in regulating gene expression and cellular differentiation. A particular transcription factor often binds to a specific pattern in the DNA called a *motif*, which can be thought of as a word that takes on different spellings according to a probability distribution. Mathematically, a motif of length w may be represented as a $4 \times w$ *position specific weight matrix* (PSWM), where each of the w columns gives the relative probabilities of observing any of the four letters, A, C, G or T. Motifs are on the order of 10-20 base pairs in length, and finding Transcription Factor Binding Sites (TFBSs) in the genome (consisting of 12×10^6 base pairs for yeast and 3×10^9 base pairs in humans) amounts to searching for a needle in a haystack. The Chromatin (Ch) Immunoprecipitation (IP) (ChIP) assay is a recently developed experimental technique that can concentrate the DNA which is bound to a specific transcription factor from the pool of genomic DNA (Buck and Lieb, 2004). The resulting sample of enriched DNA can be applied to a microarray (ChIP-chip) that has probes which correspond to uniformly spaced segments of the genomic DNA. A schematic of ChIP-chip data is shown in Figure 1. If a probe exhibits significantly higher binding to the IP sample than to the control genomic DNA sample, then the probe is said to be IP enriched and the corresponding segment of the DNA is likely to contain a TFBS. By pooling the IP enriched sequences, one may narrow the search for TFBSs to a much smaller set of sequences than the original genome.

ChIP-chip data and models

The ChIP-chip data can be represented as a $P \times R$ matrix Y where the probes are indexed by $p \in [1 \dots P]$, and the chip replicates are indexed by $r \in [1 \dots R]$. The elements of Y are given by Y_{pr} , and a row of this matrix which contains all measurements from a probe is denoted as Y_p . Y_{pr} is the

log-ratio of the IP sample intensity and the control sample intensity so that $Y_{pr} = \log \left(\frac{\text{IP}}{\text{Control}} \right)$.

Many currently available methods for analyzing ChIP-chip data, e.g. Keles *et al.* (2006), Ji and Wong (2005) and Li, Meyer, and Liu (2005) use a two-stage approach. In the first stage, regions of IP enrichment are identified from the microarray measurements. In the second stage, the selected sequences are searched for a motif, and binding sites are estimated by methods such as MDScan (Liu, Brutlag, and Liu, 2002).

We investigated an alternative approach, analyzing both the ChIP-chip microarray data and the sequence data simultaneously to estimate the binding motif and the locations of the binding sites under a single unified model (Gelfond, Gupta, and Ibrahim, 2006). Our method simultaneously estimates regions of IP enrichment and performs motif discovery through a hidden Markov model (HMM). The structure of a HMM fits these data particularly well because the correlated probes are in a sequential order, and the latent states of the HMM can correspond to the enriched and the non-enriched probe states. The probe intensity emission density is modeled as Gaussian where $f_0(Y_p | \mu_{\text{Control}})$ is the non-enriched density and $f_1(Y_p | \mu_{\text{IP}})$ is the enriched density for the p^{th} probe intensity vector Y_p . It is reasonable to assume that the means of the enriched log-ratios are related as $\mu_{\text{Control}} < \mu_{\text{IP}}$. Using a hierarchical prior then allows us to construct an efficient MCMC procedure to estimate the model parameters. If the probe sequence, X_p , is emitted from a non-enriched probe, its emission density is $p_0(X_p) \equiv p(X_p | \theta_0)$ where θ_0 represents the background sequence model; while if it is emitted from an enriched probe, its density is $p_1(X_p) \equiv p(X_p | \theta_0, \Theta)$ (Θ representing the PSWM for the transcription factor). The probabilities $p_0(X_p)$ and $p_1(X_p)$ are product multinomial models in which underlying motifs emit the letters of the binding site sequence independently. The calculation of the joint likelihood requires an efficient recursive algorithm which allows for all possible locations for motif occurrences. The algorithm for fitting the joint sequence and intensity model utilizes dynamic programming techniques for HMMs within a Gibbs sampler (Gupta and Liu, 2003).

Results

Simulation studies demonstrated that the joint intensity and sequence model is often more sensitive

and specific in finding TFBSs than a two-step approach. We also analyzed a yeast ChIP-chip experiment from Lieb *et al.* (2001). The analysis of the yeast study demonstrated that the joint model is capable of finding accurate estimates of the motif sites *de novo* (see Figure 2), and the joint intensity-sequence model found more TFBSs than models that used the 2 stage approach. We anticipate that in the future, similar Bayesian approaches will facilitate analyses in many such high-throughput data applications in genomics and other fields. ▲

References

- Buck, M. J. and Lieb, J. D. (2004), “ChIP-chip: considerations for the design, analysis, an application of genome-wide chromatin immunoprecipitation experiments”, *Genomics*, **83**, 349–360.
- Gelfond, J. A., Gupta, M., and Ibrahim, J. G. (2006), “A unified hidden Markov model for motif discovery from genomic sequence and ChIP-chip data”, *Working paper*, Dept. of Biostatistics, University of North Carolina at Chapel Hill.
- Gupta, M., and Liu, J. S. (2003), “Discovery of conserved sequence patterns using a stochastic dictionary model”, *J. Am. Statistical Association*, **98**, 55–66.
- Ji, H. K. and Wong, W. H. (2005), “TileMap: create chromosomal map of tiling array hybridizations”, *Bioinformatics*, **21**, 3629–3636.
- Keles, S., van der Laan, M. J., Dudoit, S., and Cawley, S. E. (2006), “Multiple testing methods for ChIP-chip high density oligonucleotide array data”, *Journal of Computational Biology*, **13**(3), 579–613.
- Li, W., Meyer, C. A., and Liu, X. S. (2005), “A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences”, *Bioinformatics*, **21**, 1274–1282.
- Lieb, J. D., Liu, X. L., and Botstein, D. and Brown, P. O. (2001), “Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association”, *Nature Genetics*, **28**(4), 327–334.
- Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002), “An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments”, *Nature Biotechnology*, **20**, 835–839.

ISBA/SBSS ARCHIVE FOR ABSTRACTS

All authors of statistics papers and speakers giving conference presentations with substantial Bayesian content should consider submitting an abstract of the paper or talk to the ISBA/SBSS Bayesian Abstract Archive. Links to e-prints are encouraged. To submit an abstract, or to search existing abstracts by author, title, or keywords, follow the instructions at the abstract’s web site,

<http://www.isds.duke.edu/isba-sbss/>

Figure 1: Representation of ChIP-chip data. The black bars over the sequence represent the probes. The blue letters represent the underlying DNA sequence, and the red letters represent a putative binding site.

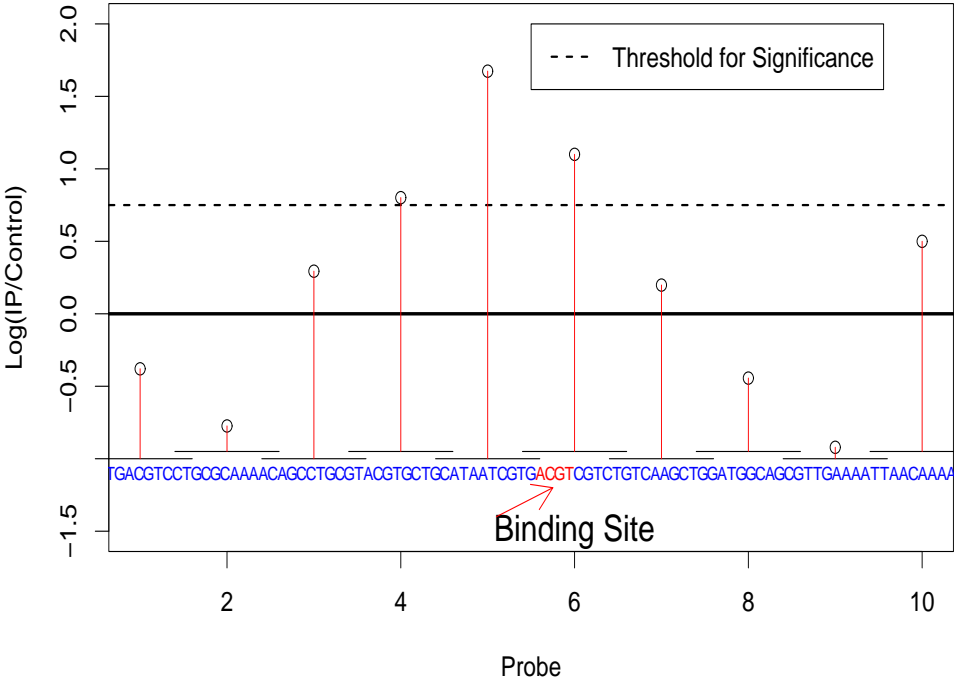
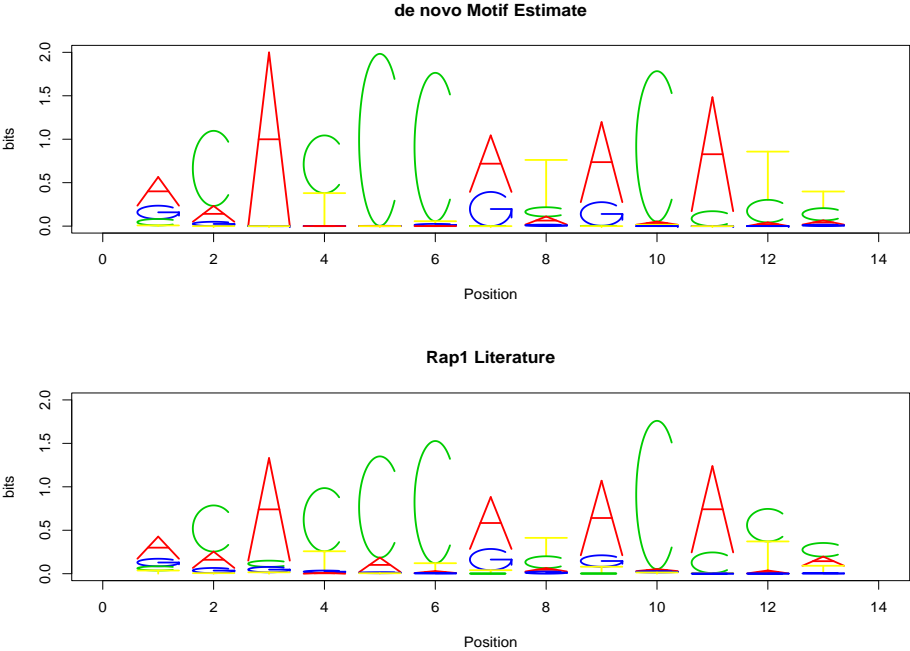


Figure 2: Logo plot for the estimate resulting from the motif finding algorithm. The upper plot represents the motif giving the highest likelihood; the lower one is the logo for the Rap1 motif given in the literature.



NEWS FROM THE WORLD

by Alexandra M. Schmidt
alex@im.ufrj.br

As this is the last issue of this year, I would like to take this opportunity and wish you all a happy new year. As usual, if you are organizing any event around the World, please, get in touch with me to announce it here.

Events

Spatial and Spatio-Temporal Statistics, Fayetteville, Arkansas, USA, April 12th-14th, 2007.

The Department of Mathematical Sciences at the University of Arkansas is organizing the 32nd Annual Spring Lectures in the Mathematical Sciences: "Spatial and Spatio-Temporal Statistics", to take place April 12-14, 2007 at the Center for Continuing Education, 2 E Center Street, Fayetteville, AR 72701. The main goal of the 32nd Spring Lecture Series is to bring together leading experts, young researchers and graduate students in the area of Spatial and Spatio-Temporal Statistics. The conference will review recent developments, present the state of the art in the field and point to important challenges and open problems. The conference will have a series of five one-hour lectures, nine one-hour talks delivered by invited speakers on topics that complement and expand the content of the lectures. The conference will also include contributed talks and a poster session, which provide the main avenue for active participation of graduate students and junior researchers. For further details, please visit the conference web page, <http://comp.uark.edu/~jjsong/SLS2007/>.

Sixth International Workshop on Objective Bayesian Analysis, Università "La Sapienza", Piazzale Aldo Moro, 5 Roma - ITALY, June 9-12th, 2007.

Following earlier meetings on objective Bayes methodology (held in Purdue, USA, 1996; Valencia, Spain, 1998; Ixtapa, Mexico, 2000; Granada, Spain, 2002; Aussois, France, 2003; Branson, MO, USA, 2005), the principal objectives of OBayes6 are to facilitate the exchange of recent research developments in objective Bayes methodology, to provide opportunities for new researchers to shine, and to establish new collaborations and partnerships that will channel efforts into pending problems and open new directions for further study. OBayes6

will also serve to further crystallize objective Bayes methodology as an established area for statistical research. The workshop will consist of a series of invited talks followed by a discussion and one or more sessions dedicated to contributed posters. On June 8 there will be a short course on Objective Bayes methodology for graduate students and other interested participants. The admission to the short course is free. For registration please visit the website <http://3w.eco.uniroma1.it/OB07>. For information please send e-mail to Brunero Liseo brunero.liseo@uniroma1.it.

Bayesian Inference in Stochastic Processes (BISP5), Valencia, Spain, June 14th-16th, 2007.

CALL FOR PAPERS

People interested in presenting a paper are kindly invited to send an abstract by December 15, 2006 to Fabrizio Ruggeri fabrizio@mi.imati.cnr.it.

The workshop follows the ones held in Madrid (Spain) in 1998, in Varenna (Italy) in 2001, in La Manga (Spain) in 2003 and in Varenna (Italy) in 2005 and is aimed to encourage discussion and promote further research in the field of Bayesian inference in stochastic processes and on the use of stochastic processes for Bayesian inference.

The workshop, organised by the Universitat de València, is endorsed by ISBA (International Society for Bayesian Analysis). The number of participants is limited to 80 people. For more details please visit <http://www.uv.es/bisp5/>.

International Workshop on New Direction in Monte Carlo Methods, Fleurance, France, June 25th - 29th, 2007.

The Workshop will cover both Monte Carlo methodologies and applications. It will focus on adaptive Markov Chain Monte Carlo methods, population Monte Carlo algorithms, and on machine learning strategies on which the on-line optimisation of Monte Carlo algorithms heavily rely. Applications cover Mathematical Finance, Bayesian Statistics and Quantum Physics. The Workshop will gather together graduate students and researchers.

There will be 5 short courses (3h30), contributed sessions, and plenty of time for informal discussion. The number of participants is limited to 50 persons. For more details please visit <http://www.adapmc07.enst.fr>.

5th International Symposium on Imprecise Probability: Theories and Applications, Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic, July, 16th-19th, 2007.

The ISIPTA meetings are one of the primary international forums to present and discuss new results on the theories and applications of imprecise probability. Imprecise probability is a generic term for the many mathematical or statistical models which allow us to measure chance or uncertainty without using sharp numerical probabilities. These models include belief functions, Choquet capacities, comparative probability orderings, convex sets of probability measures, fuzzy measures, interval-valued probabilities, possibility measures, plausibility measures, upper and lower expectations or previsions, and sets of desirable gambles. Imprecise probability models are needed in both inference and decision problems where the relevant information is scarce, vague or conflicting, and where preferences may therefore also be incomplete. For further details about (pre)registration, paper submission, scientific and cultural programme, programme committee, please consult the ISIPTA '07 web site at <http://www.sipta.org/isipta07/>.

Tenth IMS Meeting of New Researchers in Statistics and Probability University of Utah, Salt Lake City, UT, USA, July 24 - 28, 2007.

The IMS Committee on New Researchers is organizing a meeting of recent Ph.D. recipients in Statistics and Probability. The purpose of the conference is to promote interaction among new researchers, primarily by introducing them to each other's research in an informal setting. Participants

will present a short, expository talk or a poster on their research and discuss interests and professional experiences over meals and social activities organized through the conference and the participants themselves. The meeting is to be held immediately prior to the 2007 Joint Statistical Meetings in Salt Lake City, UT. Application deadline: February 1, 2007. Co-chairs: Mayetri Gupta and Xiaoming Sheng, nrc@bios.unc.edu. For more details please visit <http://www.bios.unc.edu/~gupta/NRC>.

Ninth Workshop on Case Studies of Bayesian Statistics, Carnegie Mellon University, Pittsburgh, PA, USA, October 19th and 20th, 2007.

The Workshop will feature in-depth presentations and discussions of substantial applications of Bayesian statistics to problems in science and technology, poster presentations of contributed papers on applied Bayesian work and, new this year, contributed presentations by young researchers. In conjunction with the workshop, the Department of Statistics' Tenth Morris H DeGroot memorial lecture will be delivered by Professor Larry Brown, University of Pennsylvania. Abstracts are due February 1st. Please submit abstracts via <http://workshop.stat.cmu.edu/bayes9> which contains additional information, including abstracts of previous, successful case studies. If you have questions, please contact Jay Kadane kadane@stat.cmu.edu or any of the other organizers.



INTERNATIONAL SOCIETY FOR BAYESIAN ANALYSIS

Executive Committee

President: Alan Gelfand
Past President: Sylvia Richardson
President Elect: Peter Green
Treasurer: Bruno Sansó
Executive Secretary: Deborah Ashby

Program Council

Chair: Kerrie Mengersen
Vice Chair: Peter Müller
Past Chair: José Miguel Bernardo

Web page:
<http://www.bayesian.org>

Board Members 2006–2008:

Marilena Barbieri, Wes Johnson, Steve MacEachern, Jim Zidek,

Board Members 2006–2007:

Carmen Fernandez, Valen Johnson, Peter Müller, Fernando Quintana.

Board Members 2006:

Brad Carlin, Merlise Clyde, David Higdon, David Madigan.

EDITORIAL BOARD

Editor

J. Andrés Christen <jac@soe.ucsc.edu, jac@cimat.mx>

Associate Editors

Annotated Bibliography
Marina Vannucci <mvannucci@stat.tamu.edu>
Applications
Catherine Calder
<calder@stat.ohio-state.edu>
Interviews
Brunero Liseo <brunero.liseo@uniroma1.it>
News from the World

Alexandra M. Schmidt <alex@im.ufrj.br>
Software Review
Ramses Mena <ramses@sigma.iimas.unam.mx>
Student's Corner
Vacant
Bayesian History
Vacant