Phylogenetic Footprinting of Co-expressed Genes by Tree-Gibbs Sampling

Stefan Van Yper [(1)], Olivier Thas [(1)], Jean-Pierre Ottoy [(1)] and Wim Van Criekinge [(2)]

[(1)] Department of Applied Mathematics, Biometrics and Process Control, Ghent University,
Coupure Links 653, B-9000 Ghent, Belgium
stefan.vanyper@biomath.UGent.be
olivier.thas@biomath.UGent.be
jeanpierre.ottoy@biomath.UGent.be
[(2)] Department of Molecular Biotechnology, Ghent University,
Coupure Links 653, B-9000 Ghent, Belgium
wim.vancriekinge@UGent.be

Keywords. Gibbs sampling; Motif; Co-expressed genes; Phylogenetic footprinting; Tree-Gibbs sampling.

Introduction

Promotors and transcription factors regulate gene expression in eukaryotes. Eukaryotic transcription factors are highly modular proteins that activate or inhibit the transcription of genes by binding simultaneously to a recognition sequence element on the DNA and to the transcription machine. Gibbs sampling is a stochastic process that looks for overrepresentation of those so-called motifs against a background model ([1].,[2],[3],[4]).This approach is used for finding the genomic binding sites for the transcription factors, mainly on sets of co-expressed genes and their corresponding promoter sequences

Motivation

Given the amount of different genomes available, one could reason that orthologous genes, defined as vertical descendants from a common ancestor and therefore likely to have a similar function, have similar regulation. This is reflected in conserved motif(s) in what can be called an orthologous promoter.
By combining both data sources and taking their hierarchical structure into account, additional information is available for the algorithm, resulting in improved, both in speed and accuracy, motif finding. It enables to seed genomic data with experimental data which will become a major theme in bioinformatics..

Results

The proposed Tree-Gibbs sampling algorithm makes this approach possible. It is an extension of the existing site/motif Gibbs sampler, programmed in C. On simulated data, the Tree-Gibbs algorithm works in situations where the classic site/motif sampler fails, but scores worse in others. Based on simulation studies a combination of the site Gibbs sampler with the Tree-Gibbs sampler gives the best results. Biological data will probably resemble a mixture of the generated datasets which makes the Tree-Gibbs sampler a valuable technique for phylogenetic footprinting of co-expressed genes.

Availability

The executable of the Tree-Gibbs sampler is available: http://biomath.rug.ac.be/ ~stefan/treegibbs,http://www.vancriekinge.com/treegibbs. The source code will be available at a later date. Both the 20/5 75 homology datasets and the perl datagenerating script are available.

Acknowledgements

The decision of Neuwald *et al,* to make the source code of the site and motif Gibbs sampler public is greatly appreciated. This research was financed with a specialisation grant from the Flemish institute for the promotion of scientific-technologic research in the industry (IWT).

References

[1] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C., Detecting Subtle Sequence Signals: A Gibbs Sampling strategy for Multiple Alignment. *Science*, 262: 208-214, 1993.

[2] Liu, J.S., Neuwald, A.F. and Lawrence, C.E., Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *JASA*, 90, no.432, 1156-1170, 1995.

[3] Neuwald, A.F., Liu, J.S. and Lawrence,C.E., Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science*, 4, 1618-1632, 1995.

[4] Rouchka,E.C. (1997) A Brief Overview of Gibbs Sampling. IBC Statistics Study Group, Washington University, Institute for Biomedical Computing, 1997.