Nested Sampling for Motif Discovery in Biological Sequences

Motif discovery in biological sequences (DNA and protein) is a ubiquitous problem in computational and systems biology. In DNA sequences, the discovery of common motifs in a set of co-expressed genes may indicate the presence of common transcription factor binding sites, and give insights into the co-regulation of transcriptional modules of genes. In protein sequence analysis, motifs may indicate binding, catalytic or other sites of functional importance. Commonly used approaches for motif discovery utilize sequence alignment [9], or a statistical model with maximum likelihood (Expectation Maximization) [1, 5] or Gibbs sampling [4] methods for statistical inference. A novel Markov Chain Monte Carlo (MCMC) approach to this problem, using the equi-energy sample,r was suggested by Kuo et al.[3].

Nested sampling is a novel Bayesian sampling technique introduced by Skilling [7, 8], designed to explore probability distributions where the posterior mass is localised in an exponentially small area of the parameter space. It both provides an estimate of the *evidence* (also known as the *marginal likelihood*), and produces samples of the posterior distribution. Nested sampling offers distinct advantages over MCMC methods such as simulated annealing [2], parallel tempering [10] and equi-energy approaches such as Wang-Landau sampling [11], in systems characterized by first order phase transitions [6, 7]. The technique reduces multidimensional problems to one dimension and has a single key parameter in the trade-off between cost and accuracy.

This project will involve implementing nested sampling for the motif discovery problem, based on the skeleton code provided by Skilling [7]. Computational experiments will compare its performance with the commonly used expectation maximization approach (as implemented in the MEME software), and Kuo's equi-energy sampler, for which an in-house implementation exists (from Sascha Ott). A variety of synthetic and real experimental data sets will be used for these experiments. This project would suit a student with experience in C programming.

Motif discovery problems are common to many of the collaborative projects being undertaken by Warwick Systems Biology Centre. There is the possibility of developing this mini-project into a PhD project, in collaboration with our 'wet-lab' biology partners. There may also be possibilities for applying this technique to problems in next generation sequencing data analysis.

[1] T.L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, volume 2, page 28, 1994.

[2] S. Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34(5):975–986, 1984.

[3] SC Kou, Q. Zhou, and W.H. Wong. Discussion paper equi-energy sampler with applications in statistical inference and statistical mechanics. *The Annals of Statistics*, 34(4):1581–1619, 2006.

[4] CE Lawrence, SF Altschul, MS Boguski, JS Liu, and AF Neuwald. Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[5] C.E. Lawrence and A.A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Bioinformatics*, 7(1):41–51, 1990.

[6] L.B. Pártay, A.P. Bartók, and G. Csányi. Efficient sampling of atomic configurational spaces. *The Journal of Physical Chemistry B*, pages 395–463, 2010.

[7] D. S. Sivia and J. Skilling. *Data analysis: a Bayesian tutorial*. Oxford University Press, USA, 2006.

[8] J. Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–860, 2006.

[9] G.D. Stormo and G.W. Hartzell. Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America*, 86(4):1183, 1989.

[10] R.H. Swendsen and J.S. Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607–2609, 1986.

[11] F. Wang and DP Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, 86(10):2050–2053, 2001.