Course: Statistical Methods for Computational Biology

Course Dates/times: July 12-19, 2017 at University of Bologna; Location: TBA

Instructor: Dr. Mayetri Gupta, Reader in Statistics, University of Glasgow

Contact information: mayetri.gupta@glasgow.ac.uk

Course goals and teaching methods

Rationale:
With the dramatic increase in data generation in a variety of modern scientific fields such as molecular biology, genetics, and medical imaging due to rapid technological developments, the field of statistics has undergone a major change, as new and novel techniques of statistical modeling and data analysis are continually required. As more complex models are developed, classical statistical methods often fail in the face of huge dimensionalities and latent correlation structures in the data- Bayesian computing methods, especially Monte Carlo methods, have provided an invaluable tool to address many of these issues successfully, and this field has expanded rapidly over the last two decades.

Goals:
The goal of this course is for the student to develop a thorough understanding and set of skills in advanced statistical and computational methods used in current scientific applications, focussing on problems and issues in computational biology. The objective is also to make students aware of the possibilities (and limitations) of various methods and areas for improvement. Students will gain hands-on experience constructing, programming and implementing computational techniques in real biological applications.

Course description
This course discusses advanced statistical computing methods used in modern scientific investigation focusing on computational biology applications. We will discuss applications in genomics and computational biology, including dynamic programming, hidden Markov models, multiple sequence alignment, phylogenetic reconstruction, protein structure prediction, gene regulatory network discovery and analysis of high-throughput array and sequencing data. Methodological topics include combinatorial optimization, the EM algorithm, importance sampling, Gibbs sampling, Metropolis-Hastings and data augmentation algorithms, auxiliary variable methods, and population-based Monte Carlo.

Required texts or other materials

Given the broad nature of the topics in this course, there is no single textbook that covers all the materials. Sections from the following books may be used as references, other articles for reading will be given out in class.

[DEKM] Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. (1999). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. (Cambridge University Press)

[MW] Waterman, M. (1995). Introduction to computational biology: maps, sequences and genomes. (Chapman and Hall)

Additional readings from journal articles will be provided.

Other books that may be useful:

- Koski, T. (2001). *Hidden Markov models for bioinformatics.* Kluwer Academic.
- Baxevanis, A. D. and Ouellette, B. F. F. (1998). *Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins.* Wiley-Interscience.
- Mount, D. (2001) *Bioinformatics: Sequence and Genome Analysis.* Cold Spring Harbor Laboratory Press.
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., and Rubin, D.B. (2013). *Bayesian Data Analysis.* Chapman & Hall/CRC Press.
- (Non-technical) Ridley, M. (2000) Genome: the autobiography of a species in 23 chapters.

Required software

The statistical software which will be used in the course will mainly be based in the statistical programming interface R, which is freely downloadable, open source software. Instructions on downloading, installation, and usage of the software will be provided in class.

Class session topics and readings

Day 1. Introduction to Computational Biology.

Topics: Overview of molecular biology and database resources. Genomics and molecular biology basics; the human genome project. Data formats, methods for retrieving relevant data. Genomic sequencing and alignment. Construction of physical and genomic maps, Fragment assembly, whole genome shotgun assembly and genome annotation. Sequencing accuracy. Next generation sequencing.

Readings:

- [DEKM]: Chp 1
- [MW]: Chp 1, 7, 8
- Hunter, L. (1993). Molecular Biology for Computer Scientists. in *Artificial Intelligence and Molecular Biology*, AAAI Press, 1–46.
- Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, **26** (10), 1135–45.

Day 2. Introduction to Stochastic Modelling in Biology.

Topics: Methodology of sequence alignment, Dynamic programming methods and pairwise alignment algorithms. Global and local alignment. The BLAST algorithm. Substitution matrices. Bayesian methods for pairwise alignment.

Readings:

- [DEKM]: Chp 2
- [MW]: Chp 9, 11
- J.S. Liu and C. Lawrence. (1999). Bayesian Inference on Biopolymer Models. *Bioinformatics*, **15** (1), 38–52.

Day 3. Hidden Markov models in Biological Analysis.

Topics: Hidden Markov models for biological sequences and their estimation. Forward-backward algorithm, recursions, Viterbi and Baum-Welch algorithms. Profile HMMs, gene-finding, connections to Bayesian probability models and Monte Carlo approaches. Sequence segmentation and database scanning.

Readings:

- [DEKM]: Chp 3, 4
- Lawrence R. Rabiner (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, **77** (2), 257–286.

Day 4. Multiple Sequence Alignment and Phylogenetics.

Topics: Multiple sequence alignment and profile HMMs. Progressive alignment. Gene-finding. Modeling protein families. Introduction to analysis of molecular evolution and phylogeny from DNA sequences using likelihood-based and Bayesian methods.

Readings:

- [DEKM]: Chp 5, 6, 7, 8
- [MW]: Chp 14
- Larget, B. and Simon, D. L. (1999). Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *Molecular Biology and Evolution*, **16**(6), 750–759.

Day 5. Statistical Models in Gene Regulation Analysis and Motif Discovery

Topics: Statistical models for gene regulatory motif discovery. Combinatoric, likelihood-based and Bayesian approaches. Regulatory module detection. Gene regulatory network analysis.

Readings:

- Liu, J.S., Neuwald, A. F., and Lawrence, C. E. (1995). Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. *J. Amer. Statist. Assoc.* **90** (432), 1156–70.
- Gupta, M. and Liu, J.S. (2005). De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Nat. Acad. Sc. USA* **102** (20) 7079–84.

Day 6. Recent studies in transcription regulation and introduction to epigenetics.

Topics: Genomic tiling arrays, ChIP-chip and ChIP-sequencing data analysis. Chromatin structure and gene regulation. Introduction to epigenetics and statistical analysis of epigenomic data.

Readings:

- Gupta, M. (2007). Generalized hierarchical Markov models for discovery of length-constrained sequence features from genome tiling arrays. *Biometrics*, **63** (3), 797–805.
- Park, P. (2009). ChIP-Seq: advantages and challenges of a maturing technology. *Nat Rev Genet.* **10** (10), 669–680.
- Cuvier, O. and Fierz, B. (2017). Dynamic chromatin technologies: from individual molecules to epigenomic regulation in cells. *Nat Rev Genet.* doi:10.1038/nrg.2017.28 (Advance online publication)