

ISMB 2003 Tutorial Proposal

Instructors: Professor A. Narayanan, Dr B. Olsson, Mr E. C. Keedwell.

Professor Ajit Narayanan, BSc, PhD, Professor of Artificial Intelligence, Director of Bioinformatics, School of Engineering and Computer Science, University of Exeter, Exeter EX4 4QF, United Kingdom.

Professor Narayanan was appointed Lecturer in Computer Science at the University of Exeter in 1980 and has taught various modules in computer science (artificial intelligence and machine learning techniques) cognitive science (mind/brain issues) and philosophy (philosophy of mind and language). He designed and developed the MSc/MRes programme in Bioinformatics in 1999, which now has EPSRC support as a research training programme. He teaches machine learning techniques for bioinformatics and bioethics on that programme. His CV can found at <http://www.dcs.ex.ac.uk/~anarayan/>.

Dr Björn Olsson, BSc, MSc, PhD, Director of the Bioinformatics Research Group, Department of Computer Science, University of Skövde, PO Box 408, SE-541 28 Skövde, Sweden.

Dr Olsson is Lecturer in the Department of Computer Science and coordinator for research on bioinformatics and evolutionary computation. He is a board member of the Swedish National Research School in Genomics and Bioinformatics and has taught genetic algorithms, evolutionary computation, artificial intelligence and biological sequence analysis at both the undergraduate and postgraduate levels. His recent professional activity includes being a program committee member of CBGI 2002, CBGI 2003 and ISMB 2002. Further professional details can be found at <http://www.ida.his.se/~bjorne/>.

Mr Ed Keedwell, BSc, Research Fellow, School of Engineering and Computer Science, University of Exeter, Exeter EX4 4QF, United Kingdom.

Mr Keedwell is a researcher in the School of Engineering and Computer Science and is about to submit his PhD concerning a neural-genetic model for gene expression analysis. He has several publications in the application of neural networks and genetic algorithms. Further details can be found at <http://www.dcs.ex.ac.uk/people/eckeedwe/>.

Title of tutorial: Artificial Intelligence and Machine Learning Techniques for Bioinformatics

Expected Goals, Objectives and Motivation of the Tutorial:

There is growing interest in the application of artificial intelligence (AI) techniques in bioinformatics. In particular, there is an appreciation that many of the problems in bioinformatics need a new way of being addressed given either the intractability of current approaches or the lack of an informed and intelligent way to exploit biological data. For an instance of the latter, there is an urgent need to identify new methods for extracting gene and protein networks from the rapidly proliferating gene expression and proteomic datasets. For an instance of the former, predicting the way a protein folds

from first principles may well be feasible given some algorithms for protein sequences of 20 or so amino acids, but once the sequences become biologically plausible (200 or 300 amino acids and more) current protein folding algorithms which work on first principles rapidly become intractable.

Artificial intelligence is an area of computer science that has been around since the 1950s, specialising in dealing with problems considered intractable by computer scientists through the use of heuristics and probabilistic approaches. AI approaches excel when dealing with problems where there is no requirement for 'the absolutely provably correct or best' answer (a 'strong' constraint) but where, rather, the requirement is for an answer which is better than one currently known or which is acceptable within certain defined constraints (a 'weak' constraint). Given that many problems in bioinformatics do not have strong constraints, there is plenty of scope for the application of AI techniques to a number of bioinformatics problems.

What is interesting is that, despite the apparent suitability of AI techniques for bioinformatics problems, there is actually very little published on such applications when one considers the vast and increasing amount of published papers in bioinformatics. The aim of this paper is to introduce bioinformatics researchers to three particular AI techniques, two of which may be known to bioinformaticians (neural networks and symbolic machine learning) and one of which may not be so well known (genetic algorithms). One of the intriguing aspects of the last technique is the rather philosophically appealing idea of applying techniques from AI which have been influenced by developments in our understanding of biological phenomena to biological problems themselves.

The objectives of the tutorial are to ensure that participants will have the knowledge and skills to apply machine learning techniques (both neural networks and symbolic machine learning algorithms) to biological data as well as to evaluate bioinformatics problems for their potential analysis by genetic algorithms.

Intended audience: It became clear at ISMB02 in Edmonton that there was a great deal of interest by paper presenters, poster presenters and many of the delegates in the audience concerning the application of genetic algorithms, neural networks and machine learning techniques to their problems. Current textbooks in the area do not provide an introduction to these techniques for knowledgeable biologists and bioinformaticians. The tutorial will introduce the basic algorithmic properties of the relevant techniques and then support these techniques through the presentation of examples taken from the bioinformatics literature and the research programmes of the tutorial presenters. The tutorial audience will therefore immediately see the relevance of these techniques to bioinformatics problems. Additionally, new 'discoveries' made by these techniques will be presented to demonstrate the value of applying AI and machine learning techniques to a variety of bioinformatics problems.

Detailed outline of the presentation:

The tutorial will consist of three major parts. First, classical symbolic machine learning techniques will be introduced (nearest neighbour and identification tree approaches), including examples of bioinformatics applications in secondary protein structure prediction and viral protease cleavability. Next, supervised and unsupervised neural networks will be introduced, with a number of applications described in temporal gene expression data analysis, clustering of temporal yeast data, viral protease cleavage prediction and non-temporal gene expression data analysis. Third, genetic algorithms will be covered, with applications in multiple sequence alignment and RNA folding prediction. The

presentation will have the following structure, with each of Parts A, B and C taking about 45 minutes each. The tutorial will be delivered at the pace the audience requires, with questions during presentation being encouraged. There will be two breaks of 15 minutes each during the tutorial.

- Part A:
- (i) Introduction to nearest neighbour techniques;
 - (ii) Applications of nearest neighbour techniques in bioinformatics:
 - Clustering;
 - Secondary structure protein folding prediction.
 - (iii) Introduction to symbolic machine learning;
 - (iv) Applications of symbolic machine learning in bioinformatics:
 - Viral protease cleavability;
 - Myeloma gene expression data mining.
 - (iv) Conclusion and pointers to future work.

- Part B:
- (i) Introduction to neural networks:
 - Supervised neural networks;
 - Unsupervised neural networks.

BREAK (15 minutes)

- (ii) Applications of neural networks in bioinformatics:
 - Temporal gene expression data analysis:
 - Liang networks;
 - Clustering temporal yeast gene expression data.
 - Non-temporal (classificatory) tasks:
 - Viral protease cleavage prediction;
 - Myeloma gene expression dataset;
 - Leukemia gene expression dataset.
- (iii) Conclusion and pointers to future work

- Part C:
- (i) Introduction to genetic algorithms and evolutionary computation;

BREAK (15 minutes)

(ii) Applications of genetic algorithms and evolutionary computation techniques in bioinformatics:

- Multiple sequence alignment by genetic algorithm;
- RNA folding prediction.
- Gene network reconstruction from expression data.

(iii) Conclusions and pointers to future work.

Part D: Conclusion and pointers to future applications of AI techniques in bioinformatics.

The material for the tutorial will be based on a paper ‘Artificial intelligence techniques for Bioinformatics’, written by the three presenters, recently submitted to *Applied Bioinformatics* (available from http://www.dcs.ex.ac.uk/~anarayan/publications/AI_bioinformatics_review.pdf).

Slides and examples will be taken from this paper, supplemented with further material from our research papers. The tutorial presenters will use See5 and SNNS (Stuttgart Neural Network Simulator) for demonstration purposes, as well as other purpose-built software.