Mini-Project Assignment:
MS2a Bioinformatics and Computational Biology

Deadline

TWO COPIES of your completed mini-project for MS2a Bioinformatics and Computational Biology should be handed in at the Examination Schools by the deadline of **12 noon on Monday 18 January 2010.**

The details of this mini-project assignment are on pages 2-4 below.

Mini-Project Assignment:
MS2a Bioinformatics and Computational Biology

The assignment is to write a report devoted to a topic that has been covered in the course.

Your report may be on one of the six main lecture topics listed below, or may be on a sub-topic of one of these six. The starting point for your report is for you to choose a topic, or sub-topic, on which to write your report. Then you should consult the appropriate references given in the lecture slides and below. You may of course consult further references if you wish.

The report should contain:

- Introduction. This includes motivation for the problem chosen, a short sketch of its history, and a preview of what comes in the rest of the report.
- Technical background.
- Main results in the literature on the problem.
- Discussion, including critical comments and suggestions about current unsolved problems.
- References.

Assessment and marking

As there is clearly not a precise model solution for this assignment, each mini-project will be double-marked.

The evaluation criteria used when marking will be:

- How well has the topic been motivated? (15%)
- How well has the technical background to the topic been described? (30%)
- Is there good explanation and presentation of the topic, its main problems, and their suggested solutions? (40%)
- Is there a good assessment of the current state of work on this topic, and are there possible suggestions for what to do next? (15%)

Lecture topics

The lectures have presented a series of key topics – listed here with a few key references:

- Stochastic Models of Sequence Evolution (4 lectures). Yang, Z. (2006), Computational Molecular Evolution, chapters 1 + 2. Oxford University Press.
- Phylogenies (3 lectures). Yang, Z. (2006), Computational Molecular Evolution, chapter 3. Oxford University Press.
- Detection of Recombinations in Sequences (1 lecture). Hein, J., Schierup, M.H. & Wiuf, C. (2005), Gene Genealogies, Variation and Evolution, chapter 5. Oxford University Press.
- Alignment (3 lectures). Jiang, T. (2002), Current Topics in Computational Molecular Biology, chapter 3. MIT Press. Miklos, I., Novak, A., Satija, R., Lyngsø R. & Hein, J. (2009), Stochastic Models of Sequence Evolution including Insertion-Deletion events. Stat. Methods Med. Res. 18(5):453-85.
- Grammars, Annotation and Sequence Analysis (4 lectures). Notes by R. Lyngsø available at web page http://www.stats.ox.ac.uk/research/genome/teaching2/ms2a_2008/teaching_material_and_exercises.

- Signals in Sequences (2 lectures). Wasserman, W.W. & Sandelin, A. (2004), Applied bioinformatics for the identification of regulatory events. Nature Reviews Genetics 5, 276-287.

Each lecture is available at
http://www.stats.ox.ac.uk/research/genome/teaching2/ms2a_2008/teaching_material_and_exercises.
The slides from lectures go through a series of key problems, and on each slide are the relevant references that were behind that problem.

Example outlines of possible reports

Below are 3 example outlines of possible reports on 3 different topics. Please note that these 3 examples are only suggestions.

In addition to the references mentioned below, you may consult other references, such as references from the lectures.

Example 1. "RNA folding algorithms" - What are the basic principles of structure prediction? Which algorithms can do the prediction? Why are predictions useful and how reliable are predictions? What are the major applications?

1. *Eddy S.R. (2004) How do RNA folding algorithms work? Nat Biotechnol. 2004 Nov;22(11):1457-8.*
2. *Robin D. Dowell and Sean R. Eddy (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics 2004, 5:71.*

This could lead to a report with following content:

- Introduction. What are the principles of reconstruction such as maximizing basepairs, minimizing the free energy, or maximizing the likelihood in the model? 3 pages
- Explain the algorithms. 3 pages
- How reliable are the predictions? 1 page
- Interesting applications. 2 pages
- How can present methods be improved? 1 page

Example 2. "Multiple Alignment Algorithms" - Explain the basic principles of multiple alignment algorithms. How reliable are the alignments they propose? Since the original papers (discussed in the lectures), which major ideas have been proposed? Are there still open challenges in this field?

1. *Notredame C. (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol.3(8):e123.*
2. *Miklos I, Novak A, Satija R, Lyngsø R, Hein J. (2009) Stochastic models of sequence evolution including insertion-deletion events. Stat Methods Med Res.*

This could lead to a report with following content:

- Introduction. The basic principles. 2 pages
- Technical background and algorithms. Short sketch of the dynamic programming algorithms. 3 pages
- What are the main advantages of different methods? 4 pages
- What are the main challenges now? 1 page

Example 3. "Computational Promoter Prediction" - What is a promoter? Which principles are used to predict them? What are some conclusions that have come out of computational studies? What are the main challenges now?

1. *Stark et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures Nature 450, 219-232.*
2. *Abeel et al. (2009) "Toward a gold standard for promoter prediction evaluation" Bioinformatics 25(12):i313-20.*

This could lead to a report with following content:

- Introduction. What is a promoter, and what are its main features? 1 page
- What are the key principles used for prediction? Which algorithms are used? 4 pages
- How reliable are predictions? What are the interesting conclusions from large scale investigations? 3 pages
- Is there still room for improvement? Can other data sources than sequences be used? 2 pages

Again, please remember that these 3 examples were only suggestions.

[End of Mini-Project Assignment]