

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 17 January 2011.**

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 nine main lecture topics listed below, or may be on a sub-topic of one of these nine. 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 are encouraged to find further relevant material.

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 (2 lectures). Yang, Z. (2006), Computational Molecular Evolution, chapters 1 + 2. Oxford University Press.
- Phylogenies (2 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 (2 lectures). Jiang, T. (2002), Current Topics in Computational Molecular Biology, chapter 4. 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, RNA and Sequence Analysis (2 lectures). Notes by R. Lyngsø available from teaching web-page.
- Annotation and Signals in Sequences (2 lectures). Wasserman, W.W. & Sandelin, A. (2004), Applied bioinformatics for the identification of regulatory signals. Nature Reviews Genetics 5, 276-287. Brent, M.

(2008), Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature Review Genetics* 9, 62-73.

- Comparative Biology (1 lecture).
- Networks (2 lectures). Sharan, R. & Ideker, T. (2006), Modeling cellular machinery through biological network comparison. *Nat. Biot.* 24, 427-433. Bumgarner, R.E. & Yeung, K.Y. (2009), Methods for the Inference of Biological Pathways and Networks, chapter 11 (p225-245) in McDermott, J. et al. (eds.), *Computational Systems Biology*, vol. 541 Humana Press.
- Integrative Genomics (2 lectures). Davies, J., Rafnar, T., Hellenenthal, G. & Hein, J. (2008), Integrative Genomics and Functional Explanation (notes - <http://www.stats.ox.ac.uk/research/genome/publications>).

The material for each lecture is available at

http://www.stats.ox.ac.uk/research/genome/teaching2/ms2a_2010/compbiol10

and was also sent to all students in advance of the class. The slides from lectures go through a series of key problems, and on each slide are some relevant references that were behind that problem.

Example outlines of possible reports

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

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

Example 1. “Genome assembly algorithms” - What are the basic principles of genome assembly? Which algorithms can do the assembly? Why are predictions useful and how reliable are these prediction? What are the major challenges?

1. Li Y, Hu Y, Bolund L, Wang J. *State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics.* 2010 Apr;4(4):271-7.
2. Flicek P, Birney E. (2009) *Sense from sequence reads: methods for alignment and assembly. Nat Methods.* 2009 Nov;6(11 Suppl):S6-S12

This could lead to a report with the following content:

- Introduction. What is the basic genome assembly problem? (3 pages)
- Explain the algorithms. (3 pages)
- How reliable is the assembly? (1 page)
- How has the assembly problem changed over time? (2 pages)
- How can present methods be improved? (1 page)

Example 2. “Protein gene prediction methods” - Why is it a challenge to predict protein genes in genomes? What are the major approaches? Are there still challenges in this field?

1. Burge, C and S. Karlin. (1997) *Prediction of Complete Gene Structures in Human Genomic DNA. J. Mol. Biol.* 268, 78-94
2. Brent, M. (2008) *Steady progress and recent breakthroughs in the accuracy of automated genome annotation Nature Review Genetics vol. 9 January 62-73*

This could lead to a report with the following content:

- Introduction. What is the structure of a protein gene and its characteristics? (2 pages)
- Technical background and short sketch of central algorithms. (3 pages)
- What are the main advantages of different methods. (4 pages)
- What are the main challenges now? (1 page)

Example 3. “Combinatorics of phylogenies” - What is a tree topology? How do you count them? Why is it useful to count tree topologies? Are there still interesting problems to be explored?

1. *J. Felsenstein (1978) The Number of Evolutionary Trees Syst Biol 27 (1): 27-33*
2. *J. Felsenstein (2004) Inferring Phylogenies, chapter 3 (How many Trees), Sinauer Assoc.*

This could lead to a report with the following content:

- Introduction. What is a tree topology? (1 page)
- What kinds of tree topologies are there? How do you count them? Are there alternatives to exact counting? (4 pages)
- Can enumerative algorithms be used to devise tree reconstruction algorithms? (3 pages)
- Are there still interesting problems to be explored? (2 pages)

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

[End of Mini-Project Assignment]