4273π Bioinformatics for Biologists: Course Handbook

4273π, Version 1.4. http://4273pi.org

CONTRIBUTORS TO 4273π BIOINFORMATICS FOR BIOLOGISTS

Dr Daniel Barker*,
Institute of Evolutionary Biology, University of Edinburgh.

Dr Heleen Plaisier,
Institute of Evolutionary Biology, University of Edinburgh.

Dr Rosanna G. Alderson,
Department of Biological Sciences, Carnegie Mellon University.

Dr David E.K. Ferrier,
School of Biology, University of St Andrews.

Professor Peter W.H. Holland,
Department of Zoology, University of Oxford.

Dr James L. McDonagh,
Manchester Institute of Biotechnology, University of Manchester.

Dr John B.O. Mitchell,
School of Chemistry, University of St Andrews.

Professor Michael G. Ritchie,
School of Biology, University of St Andrews.

Mr Steven D. Smart,
School of Biology, University of St Andrews.

LEARNING OBJECTIVES

1. A basic understanding of bioinformatics research techniques.

2. Appreciation of how bioinformatics techniques may be applied to biological research.

* Address for correspondence: Dr Daniel Barker, Institute of Evolutionary Biology, University of Edinburgh, Charlotte Auerbach Road, The Kings Buildings, Edinburgh, EH9 3FL, United Kingdom. Email: Daniel.Barker@ed.ac.uk

COMPONENTS OF 4273π BIOINFORMATICS FOR BIOLOGISTS

4273π is split into several components, covering various subjects and levels. Each component is found in a different subdirectory of `~/4273pi/`. Table 1 gives is an indication of the level of difficulty of components and any prerequisites to be completed first. This is not intended to prescribe or limit your use of 4273π. Rather, we hope it gives some general guidance in choosing which parts of 4273π to use and when.

Component (directory)	**Level**	**Prerequisites**	**Consists of**
INTRO	Introductory	None	Lecture: Bioinformatics, sequences and genomes. Practical class: Linux, Perl and BLAST.
BLAST	Intermediate	INTRO	Practical class: Linux and protein BLAST.
FAMS	Advanced	BLAST	Practical class: Linux, Perl and delimiting gene/protein families.
PHYLO	Advanced	BLAST	Lecture: Multiple alignment and phylogeny. Practical class: Multiple alignment and phylogeny.
EVO	Intermediate	INTRO	Lecture: Gene family evolution. Practical class: Gene family evolution.
DNA	Intermediate	INTRO	Lecture: BLAST and DNA sequence analysis. Practical class: Genome annotation.
SPECIES	Advanced	PHYLO	Lecture: Looking at species differences. Practical class: PAML.
ENZYME	Introductory	None	Lecture: Enzyme function and evolution. Practical class: Enzyme function and evolution.
PROJ	Advanced	FAMS	Practical project: Comparative genomics.

Table 1. Components of 4273π. Each component is in a subdirectory, named as the component (e.g. `~/4273pi/INTRO/`). 'Introductory' components have no prerequisites. One of these would be a good place to start. 'Intermediate' components have at least one introductory-level prerequisite. 'Advanced' components have at least one intermediate-level prerequisite. Each lecture lasts slightly under one hour, and if used, should precede the practical class for the component. Each practical class lasts approximately two hours, though this will vary with experience. The practical project requires a significant effort and is intended to be carried out as self-study, spread over several weeks.

EXAMPLE TIMETABLE

A course may be created out of subsets of components. As a guide, it is useful to consider the prerequisites proposed in Table 1.

For a course spread over one semester, involving approximately 150 hours student effort (including 23 hours in class) – for example at senior undergraduate level – this is one feasible sequence of components:

Week 1: INTRO.

Week 2: BLAST.

Week 3: FAMS; PROJ assigned.

Week 4: PHYLO.

Week 5: EVO.

Week 6: DNA.

Week 7: SPECIES.

Week 8: No teaching on this course.

Week 9: Deadline to hand in PROJ.

Week 10: ENZYME.

Week 11: Student seminars.

It can be useful to extend PHYLO over two weeks, allowing more class time for the practical class (~4 hours). To make space for this, the practical class for one other component (e.g. DNA) could be removed from the course.

It is suggested that assessment be based on coursework, in the form of PROJ (50%); and a separate two-hour written exam, as outlined below ('Exam'), perhaps held about a fortnight after the student seminars.

The weeks with no classes allow students time to focus on the coursework (PROJ).

The suggested seminar (proposed for the final week) is not assessed and has no teaching material in 4273π. Students give a brief presentation on their coursework (PROJ), to an audience of each other and the instructor, with or without slides as the student sees fit. This gives students chance to share results from PROJ, and ask any questions before the exam.

RECOMMENDED READING

The main recommended books are listed during the first lecture in the INTRO component. Other appropriate reading is recommended within specific components.

EXAM

Examinations are not provided with 4273π. The below instructions are just as an example, which may prove useful to anyone setting an exam based on this material.

The exam lasts two hours. Please bring a calculator.

The exam consists of four questions, all of which are compulsory. Questions are broadly of a 'problem-solving' kind. There are 20 marks for each question and questions are weighted equally. Marks allocated to each part of each question are indicated.

Any subject from the lectures and practical classes may appear. A thorough practical and theoretical understanding of the material covered will help. However, an in-depth knowledge of the details of Perl programming will not be required in the exam. If you have used Perl successfully as required for the practical project (PROJ component), you should have sufficient knowledge of Perl for the exam, and can focus your revision on other aspects of the course.

To assist with revision, one 'mock' question is provided, below. Although this is intended to be helpful, it has not been subject to the same level of checking that would be applied to an actual exam question.

Mock Exam Question: Sequence alignment and phylogeny

A multiple alignment of orthologous proteins for 20 species of eukaryote and one of archaea is to be used for phylogeny reconstruction. The multiple alignment is submitted to the phylogenetic model selection program, Modelgenerator. Part of Modelgenerator's output is as follows:

```
****Bayesian Information Criterion (BIC)****

Model Selected: LG+I+G+F
-lnL = 226052.80535126594
k = 60 (Branch lengths included as model parameters)
BIC = 452680.14719995187

Substitution Model (with Rate Distribution):
Model of substitution: LG (Le and Gascuel, MBE 2008 25:1307-20)

Amino acid frequencies:
pi(A) = 0.06499
pi(R) = 0.05272
pi(N) = 0.04510
pi(D) = 0.06103
pi(C) = 0.01571
pi(Q) = 0.03788
pi(E) = 0.08217
pi(G) = 0.05936
pi(H) = 0.02181
pi(I) = 0.06304
pi(L) = 0.09289
pi(K) = 0.07879
pi(M) = 0.02303
pi(F) = 0.04061
pi(P) = 0.04163
pi(S) = 0.06028
pi(T) = 0.04958
pi(W) = 0.00801
pi(Y) = 0.03123
pi(V) = 0.07017


Model of rate heterogeneity: Discrete Gamma + Invariable sites
Number of rate categories: 1 + 4
Gamma distribution parameter alpha: 0.89
Proportion of invariable sites: 0.04


Relative rates and their probabilities:

    Rate      Probability
1   0.00000   0.04453
2   0.12499   0.23887
3   0.49881   0.23887
4   1.10482   0.23887
5   2.45778   0.23887
```

(a) Which phylogenetic model has been selected by the Bayesian Information Criterion (BIC)? Which substitution matrix and other features appear in this model? **[6 marks]**

(b) If we wish to reconstruct the phylogeny not from the protein sequences used in (a), but from the DNA sequences which code for them, outline the necessary steps to obtain a multiple alignment and select a phylogenetic model. **[4 marks]**

(c) The globin protein superfamily is present in all domains of life. Three-dimensional protein structures of globins are reasonably well-conserved, but their protein sequences are more variable. A researcher wishes to gather a representative sample of globins from the public sequence database. These will be used for phylogeny reconstruction, in a study of gene duplication, loss and divergence over evolutionary time. To find globin sequences, the researcher uses several known globin sequences as queries in searches of the protein database at the National

Center for Biotechnology Information (NCBI). Why might PSI-BLAST be helpful here? **[4 marks]**

(d) Using sketches to illustrate your answer, compare global and local pairwise sequence alignment. **[6 marks]**