Teaching bioinformatics: A student-centred and problem based approach

Yun-Can Ai
School of Life Sciences, Zhongshan (Sun Yet-Sen)University, Guangzhou 510275, People's Republic of China

Lars Jermiin and Neville Firth
School of Biological Sciences, The University of Sydney, NSW 2006, Australia

This article was first published in 'The China Papers' Issue 1, October 2002 and is reprinted here, with some modification, with permission from the authors.

Abstract

Bioinformatics is one of the fastest growing interdisciplinary sciences of the late 20th and early 21st century. For many students, bioinformatics is such a new discipline that they will not necessarily have the required background knowledge. It is therefore necessary to build bridges for students with diverse backgrounds. This paper describes teaching a bioinformatics course, explains what bridges needed to be built, and provides some examples of how the approaches of student-centred teaching and problem based learning could have been used in teaching this course using *WebCT* as the delivery tool. To understand the major concepts of bioinformatics, two conceptual frameworks are used: *similarity*, which enables analysis of predictions about structure/function; and *dissimilarity*, that allows inference of evolutionary history based on distance. These attributes are used as examples of core modules (the bridges), because all analyses need to be undertaken in appropriate biological context but involve multiple disciplines, including mathematics, statistics, computer science, and information technology, that need to be integrated into biology. The course focuses on three aspects: (i) to assist students with computing or related backgrounds understand the difficult core concepts of biology; (ii) to build critical understanding of the key mathematical and statistical principles that are crucial to bioinformatics; and (iii) to give all students hands-on training by allowing them to explore key issues using dominant bioinformatics platforms. Using these bridges, students should be able to develop the skills of creative experts rather than those of handy technicians.

Introduction

Bioinformatics is a scientific discipline that has emerged recently in response to accelerating demand for an expedient, flexible, and intelligent means of storing, managing and querying large and complex biological data sets. It is an interdisciplinary science with strong links between the life sciences and mathematics, statistics, and computer science, and it is regarded by a wide range of interest groups (governments, universities, and industry) as a crucial area of development - huge investments have been made worldwide, particularly in the USA and Europe. The ultimate goal of bioinformatics is to allow scientific exploration and exploitation of previously uncharted interdisciplinary territories. Stated differently, the aim is to enable the

discovery of new biological insights and to create a global perspective, from which unifying principles in biology can be discerned (Mount 2001).

In the beginning of the genomics era, bioinformatics was mainly concerned with the creation and maintenance of databases to store digitised biological information, such as nucleotide and amino acid sequences. Development of these types of databases involved not only design issues, but also the development of complex interfaces whereby researchers could both access existing data, and submit new or revised data (e.g. to the NCBI, http://www.ncbi.nlm.nih.gov/). More recently, emphasis has shifted towards the questions of how to analyse large data sets, particularly those stored in different formats in different databases. Ultimately, however, integration is needed (e.g. Chicurel 2002) in order to form a comprehensive picture of normal cellular and sub-cellular activities, so that researchers may study how these activities are globally regulated. The actual process of analysing and interpreting digitised biological data is often referred to as computational biology. It is commonly recognized that sub-disciplines within bioinformatics and computational biology include: (i) the development and implementation of tools that enable efficient access, management and use of various types of digitised biological information; and (ii) the development of new mathematical theorems, statistical methods and algorithms to infer relationships among members of large data sets, locate genes within nucleotide sequence, and predict protein structure and/or function.

Teaching bioinformatics is an interesting and challenging task because it requires in-depth knowledge of different disciplinary areas that are all evolving at great speed - this blend of interdisciplinary expertise is rarely available from a single individual, so courses in bioinformatics often need to be planned and taught by academics, who have widely different formal training. This, in turn, poses not only challenges, but also opportunities for the development of novel teaching strategies (Ai 2002). In the present paper, we discuss our strategies for teaching bioinformatics by outlining core themes and concepts in bioinformatics, explaining what bridges we are trying to build, and providing some examples to show how the approaches of student-centred teaching and problem based learning can be used in teaching this new discipline.

Course description

Bioinformatics is the application of computers to the life sciences and in particular to genomics. This makes it possible to study biology at the genome-wide level. In teaching such a course, we aim to focus on the major applications, including data storage, retrieval, and analysis of biological information, rather than on the engineering of new bioinformatics applications. The broad aim of the course is that students will become extensively trained and develop hands-on skills in the use of bioinformatics technologies; and that they will understand and appreciate the enormous potential of bioinformatics and genomics in the contemporary life sciences. Expected outcomes include:

- an awareness of the breadth of bioinformatics resources and applications, including non-sequence-based biological information;
- skills and experience in the use of a core set of programs and databases for nucleotide and amino acid sequence analysis and phylogenetic reconstruction;

- a basic understanding of the theoretical foundation and underlying assumptions of the programs, and their relative strengths/limitations; and
- competence in the evaluation of output from the programs in appropriate biological context.

Curriculum

The details of the curriculum being taught at The University of Sydney, Australia, and at Zhongshan University, China, are shown in Table 1. There are four basic themes to be covered. Within themes 2 and 3, student-centred teaching and problem based learning are used to further student understanding of sequence alignment and phylogenetics.

Core Themes	**Main Topics**
1. Bioinformatics resources	The relationships among bioinformatics, genomics and proteomics Computational sequence analysis ('*in silico* biology') Laboratory tools Databases I - sequence databases Databases II - other biological databases
2. Sequence similarity and structure/function inference	Sequence comparison and alignment Database searching Multiple sequence alignment, motifs and profiles Signal detection Analysis of deduced products
3. Sequence dissimilarity and evolution	Structural inference Introduction to phylogenetics Simulations, multiple hits and compatibility plots Phylogenetic inference using parsimony methods Phylogenetic inference using distance-based methods
4. Integrating themes	Phylogenetic inference using maximum-likelihood methods Introduction to functional genomics

Table 1. Bioinformatics course content

Two conceptual frameworks - similarity and dissimilarity

Genomics generates unprecedented amounts of data, whereas bioinformatics focuses on converting these data into knowledge. Obviously, data generation and analysis are inter-dependent. In the context of genomics, bioinformatics is often concerned with sequence analysis,

which can usually be considered within two conceptual frameworks - *similarity* and *dissimilarity*. Analysis of similarity enables predictions about structure/function, whilst dissimilarity analysis allows inference of evolutionary history based on distance. Analyses involve disciplines including mathematics, statistics, computer science, and information technology, but need to be undertaken within the appropriate biological context. Bioinformatics represents a challenge for people coming from different disciplines - bridges are needed to help them communicate with each other.

Building bridges by using core modules

Bioinformatics requires the skills of a diverse group of specialists. Biologists want to be computational biologists, while informaticians want to be biological informaticians. They all need to be trained in theory as well as software development and applications. Core modules must cater for diverse backgrounds and offer the required bridges. We focus on three groups of students: (i) students with a computing or related background must be assisted in learning core concepts of biology; (ii) students lacking adequate mathematical/statistical expertise must be taught key principles about methods of relevance to bioinformatics; and (iii) all students must be given hands-on training in the use of predominant bioinformatics platforms. Using these educational bridges, students should become creative experts rather than simply handy technicians.

Similarity analysis module

The inference of structure and function based on biological sequence similarity is a fundamental concept in bioinformatics, and represents a good example of a bridging module. The Dotplot method is a particularly informative way to do similarity analysis between two sequences. However, although the method is conceptually simple, it can be difficult for non-biology students to understand the biological significance. Conversely biology students may be confused by the mechanisms used to reduce noise in the plot. In this case we bridge an abstract visualisation strategy directly to its biological basis to illustrate insertion/deletion/duplication, etc. This bridge provides a strong foundation so that all students can understand gene evolution. This is crucial if students are later to use or even develop sophisticated software to explore whole genome scale events.

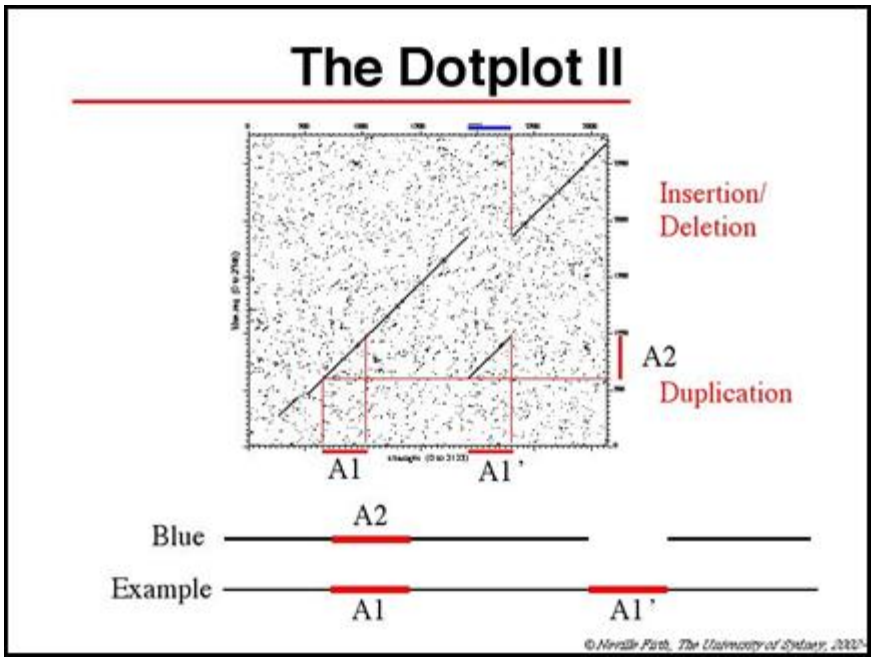

Figure 1. Similarity analysis using the Dotplot method, showing visualization of genetic insertion/deletion/duplication events on a large scale

Dissimilarity analysis module

Phylogenetics is based on the concept that taxonomic lineages diverge from one another over time and, therefore, that estimates of dissimilarity or distance between representatives of the taxonomic lineages may tell us how removed they are from one another in an evolutionary sense. Many computer programs developed to infer phylogenetic trees use a simple matrix, Q, which defines the rate of substitutions from one nucleotide to another (Figure 2).

However, the significance of this matrix is often difficult to understand for students who lack a strong mathematics background, and this can impact on their ability to use software correctly and effectively. Students with a strong mathematical background, on the other hand, will feel instantly familiar with the properties of this matrix but they may not be able to understand how it can be used to model the evolutionary process at the level of DNA. This example clearly illustrates why there is a need to build interdisciplinary bridges for students with limited interdisciplinary training and knowledge. By taking advantage of the mixed disciplinary background of a typical student cohort (e.g. biology, mathematics, statistics, computer science), it is possible to bridge the interdisciplinary gaps by asking groups of students to address problems that require knowledge from different disciplines. This has the added advantage that students are actively involved in the teaching/learning process and, in particular, that they learn to effectively communicate their knowledge to students with a different academic background. To underpin the complex knowledge that these students are acquiring, students are provided further online tutorial practice with simulations and constructions of phylogenetic trees based on their understanding of the rate matrix (Figure 2), which in turn promotes an understanding of the evolutionary history in terms of phylogenetics and the key principles and methods that are critical to bioinformatics.

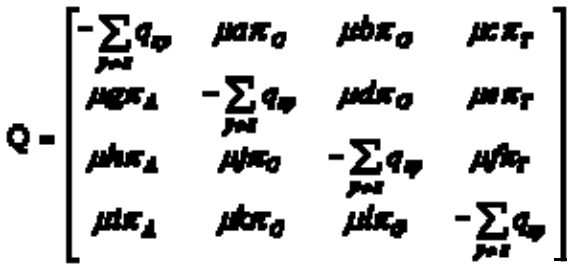

Figure 2. Each element of this matrix, qxy, represents the instantaneous rate of substitution from nucleotide x to nucleotide y during an infinitesimal amount of time (dt) (x, y = A, C, G, T) - for further details, see Swofford et al. (1996).

Online hands-on use of major Bioinformatics platforms

The third kind of bridge is the hands-on use of major online bioinformatics services, including national and international resources such as ANGIS BioManager (Australia), NCBI (USA), EMBL (Europe), ExPASy (Sweden), and S-Star.org (Multinational) (see Useful Links below). There are hundreds of programs available through BioManager. NCBI was one of the first online bioinformatics resources and is still a world leader. EMBL is a major European bioinformatics resource. ExPASy is the protein sequence analysis expert system, which contains a large number of modular programs. S-Star.org is a pioneering international bioinformatics education initiative, which involves institutions from around the world (Ping et al. 2002). Students are required to use these platforms in their own time, after receiving teacher-guided help in class during tutorial sessions.

Student-centred teaching by using WebCT

Bioinformatics is well suited to the development of student-centred teaching scenarios. *WebCT* has proven particularly useful in this regard since it caters for presentation of a diverse array of teaching materials, including links to useful bioinformatics resources, *QuickTime* 'lecture movies', practical self-assessment quizzes, assignments, etc. The materials are easy to use for the students and can be updated in a cost effective manner. Recorded 'lecture movies' are instantly replayable, making them particularly useful for revision and for foreign students learning the course materials in English. Online evaluation promotes active student learning. Figure 3a shows the Bioinformatics course *WebCT* site Home page and Figure 3b shows the Bioinformatics course *WebCT* Learning Resources page.

The learning resources page includes a link to ANGIS BioManager, a service that is widely used in Australia for bioinformatics research, and increasingly also for teaching; BioManager is used throughout the teaching and learning process of this course. Online practical and tutorial activities are provided in laboratory classes, where students are provided extensive training in the programs provided through BioManager (Figure 4). With respect to phylogenetics, emphasis is placed on building an understanding of molecular evolution by using programs developed in-house, such as *hetero, zeta* and *TrExML* (Jermiin 2003a, 2003b; Wolf et al. 2000). In these learning scenarios, teachers act as supporters and facilitators rather than hands-on demonstrators,

so that students are challenged, stimulated and motivated to become experts in analysing real problems.

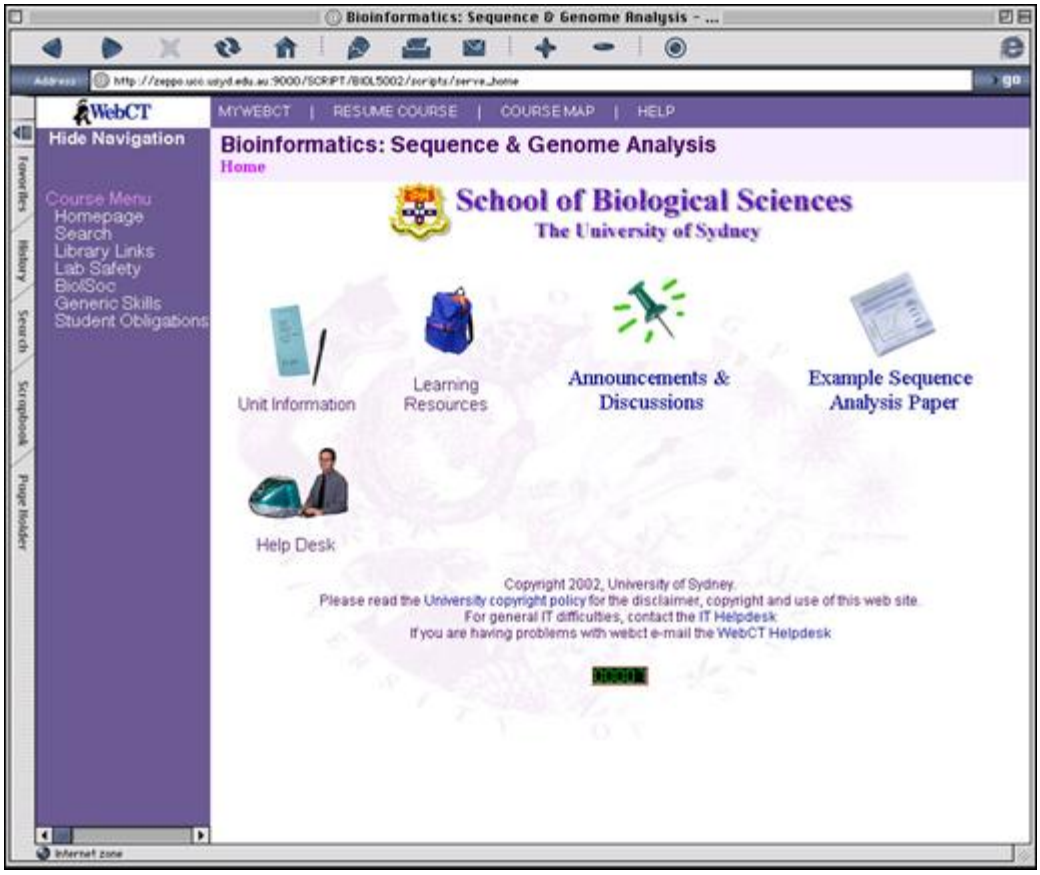

Figure 3a. *WebCT* site Home page for the Bioinformatics course at The University of Sydney

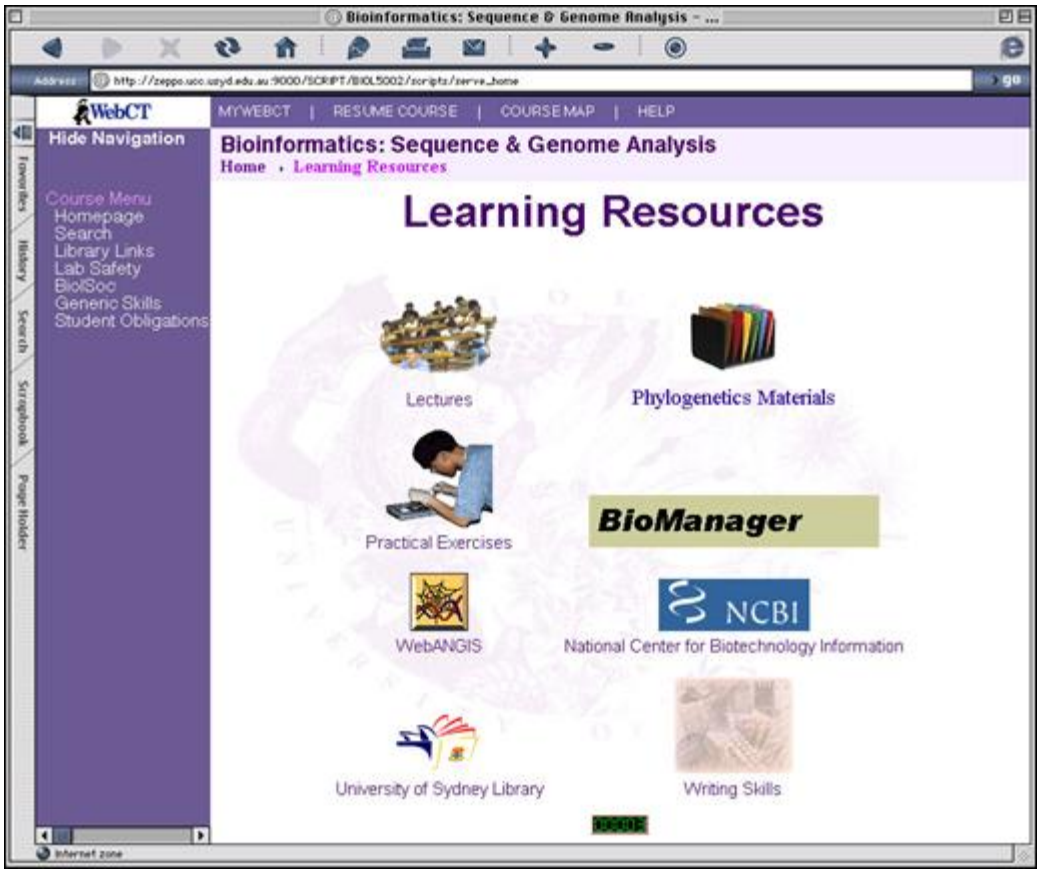

Figure 3b. ***WebCT*** **site Learning Resources page for the Bioinformatics course at The University of Sydney**

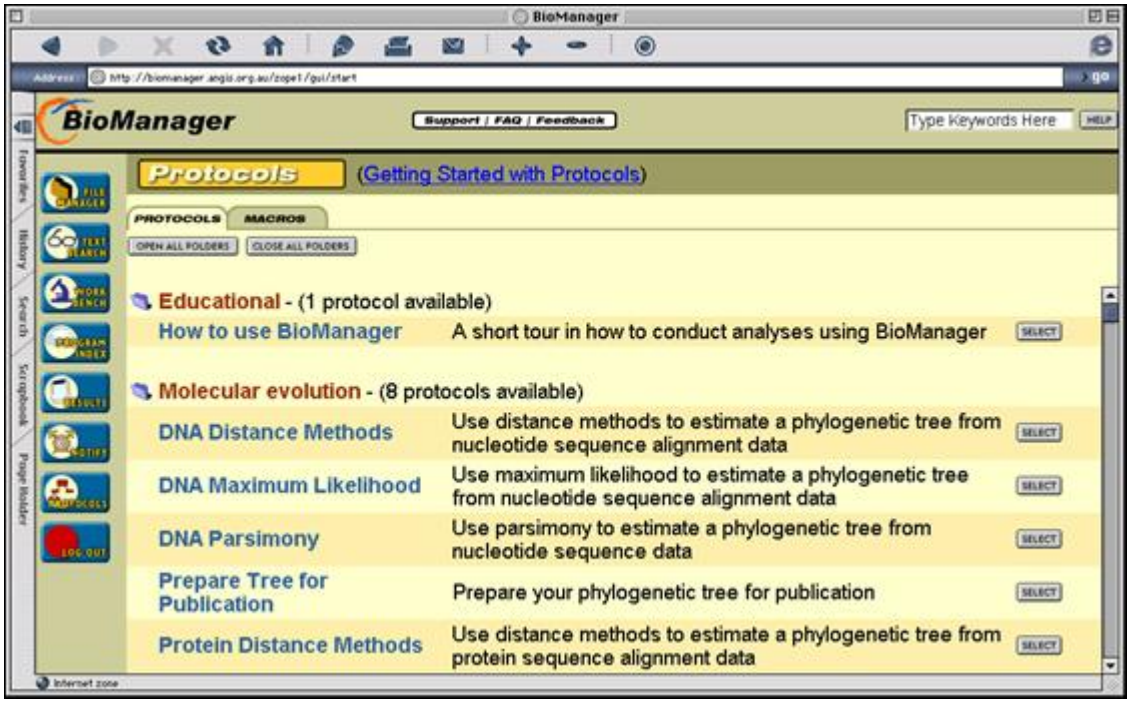

Figure 4. BioManager web site for online use in lecture and tutorial practical sessions

Online problem based learning

During the course, students are required to work in groups to explore and evaluate bioinformatics resources on the Internet, such as ExPASy and NCBI. Each student group is then required to give a seminar to the class, outlining the function, features and methods of the resource, and the group's evaluation of its utility. This type of teamwork encourages students to work cooperatively, and it establishes a foundation for a data mining project that is undertaken individually. This task is intended to challenge the students to actively learn core skills and reinforce key bioinformatics concepts and generic skills, by requiring them to undertake and report an analysis of a real DNA sequence. Support for sequence analysis, academic writing and reference citation is provided in the form of an example journal manuscript (Apisiridej et al. 1997) that is available on the course *WebCT* site. These activities form the basis for group-wide and individual assessment of performance, and provide a rational basis for constructive feedback during the course. Several criteria are considered to promote, and allow assessment of, both academic and generic skills (see Table 2 for example).

Criterion	Available Marks	Aspects Considered (marks allocated)
Analytical approach, interpretation and content	17	Appropriate selection and use of programs (3) Evaluation of results in appropriate biological context (3) Demonstrated understanding of use of bioinformatics tools to gain biological insight (3) Adequacy of supporting information (e.g. parameters used) (3) Citation of relevant references and bibliography (3) Independence (2)
Presentation	8	Clarity and logic (4) Standard (e.g. spelling, grammar, clarity of figures and tables) (4)

Table 2. Assessment of the data mining project

Talented students can be further challenged by asking them to write a proposal for a project grant application and/or engage in scholarly debate in class. This provides extremely useful generic skills for students who may become academics in future.

Summary

We have outlined some of the strategies that we have used in teaching bioinformatics, an example of an emerging interdisciplinary science, so as to take advantage of a combination of contemporary teaching strategies and online information technology. The approaches of building bridges, student-centred teaching, and problem based learning have been used to successfully promote student's active learning.

Acknowledgements

This paper is based on a presentation given in 2002 while attending a professional development program at The University of Sydney, and was critically reviewed as well as extensively edited for language usage by Associate Professor Mary Peat. The program *Teaching Sciences in English: a professional development course for Chinese university science teachers* is a collaborative project between the China Scholarship Council and The University of Sydney, coordinated by Associate Professor Mike King (Faculty of Education) and Associate Professor Mary Peat (Faculty of Science). Yun-Can Ai is a recipient of The National Universities Distinguished Teachers by the Chinese Ministry of Education and The Guangdong Provincial Universities Outstanding Talents by the Guangdong Provincial Education Department. Our warmest thanks go to all of the students who provided valuable insight into the course.

Useful links

ANGIS (Australian National Genomic Information Service)
http://www.angis.org.au/new/
EMBL (European Molecular Biology Laboratory)
http://www.embl-heidelberg.de/
ExPASy (Expert Protein Analysis System)
http://www.expasy.org/
NCBI (National Center for Biotechnology Information)
http://www.ncbi.nlm.nih.gov/
S-Star.org (BioInformatics Education Alliance)
http://www.s-star.org/

References

1. Ai, Y-C. (2002) The development of a Microbiology course for a large class of students of diverse backgrounds: A review of seven years of change in Zhongshan (Sun Yet-Sen) University, China. *The China Papers: Tertiary Science and Mathematics Teaching for the 21st Century*, **1**, 36-41.
2. Apisiridej, S., Leelaporn, A., Scaramuzzi, C. D., Skurray, R. A. and Firth, N. (1997) Molecular analysis of a mobilizable theta-mode trimethoprim resistance plasmid from coagulase-negative *Staphyloccoci. Plasmid*, **38**, 13-24.
3. Chicurel, M. (2002) Bioinformatics: bringing it all together. *Nature*, **419**, 751-757.
Jermiin, L. S. (2003a) Hetero: program to simulate the evolution of nucleotide sequences on a tree, with the inclusion of some more realistic substitution models (ANSI C code available soon).

4. Jermiin, L. S. (2003b) Zeta: a program to assess stationarity in aligned nucleotide and amino acid sequences (ANSI C code available soon).
5. Mount, D. W. (2001) *Bioinformatics: Sequence and genome analysis*. New York: Cold Spring Harbor Laboratory Press.
6. Ping, L. Y., Höög, J. O., Gardner, P., Ranganathan, S., Andersson, S., Subbiah, S., Wee, T. T., Hide, W. and Weiss, A. S. (2002) International online education: the S-Star trial bioinformatics course. *CAL-laborate*, **8**, 18-19.
7. Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. (1996) Phylogenetic inference. In D. M. Hillis, C. Moritz, C. and B. K. Mable (Eds) *Molecular Systematics*, 2nd Ed. Sunderland, Massachusetts: Sinauer Associates, 407-514.
8. Wolf, M. J., Easteal, S., Kahn, M., McKay, B. D. and Jermiin, L. J. (2000) TrExML: a maximum likelihood approach for extensive tree-space exploration. *Bioinformatics*, **16**, 383-394.

Yun-Can Ai
School of Life Sciences
Zhongshan (Sun Yet-Sen)University
Guangzhou 510275
People's Republic of China
lssayc@zsu.edu.cn

Lars Jermiin
School of Biological Sciences
The University of Sydney
NSW 2006
Australia

Neville Firth
School of Biological Sciences
The University of Sydney
NSW 2006
Australia