BIOCOMPUTING EDUCATION BY THE AUSTRALIAN NATIONAL GENOMIC INFORMATION SERVICE

Bruno A. Gaëta, Carolyn A. Bucholtz, Rowena Campbell, Camson Huynh, Stephanie Kim and Alex H. Reisner

Australian Genomic Information Centre, J03, The University of Sydney, NSW 2006, Australia

Introduction

The Australian National Genomic Information Service (ANGIS) was set up in 1991 to provide sequence and genomic information to the Australian molecular biology community, together with software for its analysis. The service is based at the University of Sydney, and is available across Australia through the Australian Academic and Research Network (AARNet), the Australian branch of the Internet.

ANGIS maintains over 40 comprehensive databases, including nucleotide and protein sequence databases, motif databases and genomic databases. The service also provides over 300 programs to extract information from these databases and analyse or manage sequence data. These programs include the complete GCG and Staden packages, and many others. ANGIS also provides ‘transparent’ front ends to a number of other programs running on other computers around the world and accessible by electronic mail.

The software available on ANGIS covers almost every aspect of computing applications in molecular biology, with the exception of molecular modelling. This includes a range of programs for database browsing and similarity searching, sequence editing, manipulation, analysis and comparison, restriction mapping, multiple sequence alignment, molecular phylogeny, sequencing project management, primer design, secondary structure prediction and genetic linkage analysis.

ANGIS is a UNIX-based service, but all the programs have been integrated into user interfaces which do not require a knowledge of UNIX commands. Users who have access to VT-100 compatible terminals can access the programs using a 2-dimensional menu interface. A full graphical user interface is also available for users with a fast network connection and an X windows server. It is estimated that ANGIS is used by over 2000 scientists, therapists and students from universities, research institutes and industry. Every working day, ANGIS accounts are accessed by 300-400 individuals from all over Australia.

From the inception of ANGIS, user support has been an important part of the mission of the service: two out of the six staff members are fully dedicated to the production of documentation and teaching resources, in addition to the direct user support provided by both the head and the system administrator. A survey was sent to ANGIS users in the beginning of 1994 to determine the future directions of the service. This survey confirmed that biocomputing education was seen as being part of the service’s brief: 94% of the respondents agreed with the suggestion that ANGIS should provide courses on the use of the service (1% disagreed), and 68% agreed that ANGIS should assist in the design of course work given at Australian universities to instruct in the use of bioinformatics (4% disagreed).

The ANGIS user support has therefore been gradually extended to include a number of educational activities primarily directed at researchers who use ANGIS, who range from graduate and postgraduate students to senior academics. In turn, some of these researchers who are also involved in teaching have passed on part of this knowledge and expertise to undergraduate students as components of biochemistry or genetics courses. In most cases, this undergraduate teaching is based on teaching resources produced by ANGIS, which have been made available freely in electronic form, or at cost in printed form.

Teaching Materials

The Tutorials

The January 1994 survey of ANGIS users showed that 90% of the respondents would use a self-teaching practical tutorial set if it were available. This led to the development of the *ANGIS Tutorial Book*, a set of 19 self-contained tutorials covering the major programs available on ANGIS. These tutorials have been used in the courses run by ANGIS, and are also suitable for self-teaching. They are brought up to date twice yearly, to keep up with the constant evolution in biological analytical software and the improvements in the ANGIS interface. The latest editions of the tutorial book contain over 800 pages covering the following topics:

1. **Introduction to ANGIS**
   Navigating around ANGIS, accessing on-line help.
2. **File Manipulation**
   Listing, creating and deleting files. Using the Emacs editor.
3. **Network Resources**
   File transfer, electronic mail, Usenet news, gopher, World-Wide Web.
4. **Information Retrieval**
   Retrieving sequences and other information from databases.
5. **Sequence Editing**
   Sequence entry, editing, translation and format conversion.
6. **Database Searches**
   Similarity searching of sequence databases (BLAST, FastA etc.), specialised databases.
7. **Restriction Mapping**
   Detecting restriction sites in sequences and plotting restriction maps.
8. **Sequence Comparison**
   Comparing and aligning two sequences to estimate their similarity.
9. **Multiple Sequence Alignment**
   Creating, editing and formatting an alignment of more than two sequences. Deriving a sequence profile and using it for database searches.
10. **Pattern Recognition**
    Searching sequences and sequence databases for patterns and motifs.
11. **The Ribosomal Database**
    The RDP ribosomal RNA database and the alignment editor AE2.
12. **Phylogeny Inference**
    Inferring phylogenies from sequence data and plotting phylogenetic trees.
13. **Gene Detection**
    Searching sequences for open reading frames and potential coding regions.
14. **Sequencing Project Management**
    Management of shotgun sequencing projects.
15. **PCR Planning**
    Designing oligonucleotides for polymerase chain reaction experiments.
16. **GDB and OMIM**
    The human genome database and the OMIM catalogue of human genes.
17. **Graphical Genomic Databases**
    Databases based on ACeDB (A *Caenorhabditis elegans* Database).

18. Genetic Linkage Analysis
Genetic mapping using the Mapmaker package.

19. Secondary and Tertiary Structure
RNA and protein secondary structure prediction. Displaying tertiary structure models.

The tutorials are designed to be used without direct supervision. They detail procedures to be followed step by step, together with illustrations and explanations of the results.

The full ANGIS tutorial set has been made available by anonymous ftp at morgan.angis.su.oz.au, in /pub/ANGIS-tutorials. Microsoft Word for the Macintosh, Rich Text Format and PostScript versions are available.

Because the tutorials provide detailed step-by-step instructions, they are fairly specific to the ANGIS interface, although in many cases they can be adapted easily for more general use. To date, the tutorials on GDB and on ACeDB have been converted into stand-alone tutorials which have been made generally available by anonymous ftp. The ACeDB tutorial in particular has been very well received, and has been used as the basis for classes in the USA and in Europe. It has also been converted to HTML and made available on the World Wide Web by Sam Cartinhour of the National Agricultural Library in the US Department of Agriculture, at the URL http://probe.nalusda.gov:8000/acedocs/angis/TOC.html.

Parts of the tutorial book have also been incorporated into two smaller *Getting Started* books, *Getting Started- with the 2D Menu Interface* and *Getting Started- with the XANGIS Interface* which are sent free of charge to every ANGIS account.

The Recipe Book

Feedback obtained from a number of ANGIS users suggested that many researchers don't have the time to learn the details of how to use the computing resources available to them, but just want to be able to use these resources when they need them without spending too much time. This concern was addressed by creating a small book of protocols describing in a step-by-step fashion how to perform a number of the most commonly used sequence analysis tasks using ANGIS. The resulting *ANGIS Recipe Book* contains over 100 pages covering the following topics:

- Suggested strategies for sequence analysis
- Entering a new sequence
- Similarity searching
- Database sequence extraction
- Sequence comparison
- Searching for restriction sites
- PCR primer design
- Protein secondary structure prediction
- Motif search
- Multiple sequence alignment

Although the recipe book was not written to be a teaching tool, it has been used as such by some university lecturers wanting to teach basic bioinformatics procedures to undergraduate students, and not wanting to go to the level of detail of the tutorial book. However, because the main aim of the book is to guide ANGIS users through the ANGIS interface to reach their goal rather than to teach the in depth use of the software, it has a more limited usefulness to non-ANGIS users than the tutorial book.

In addition to the hard copy version, the *ANGIS Recipe Book* has been made available in electronic form by anonymous ftp from morgan.angis.su.oz.au, in /pub/RecipeBook.

The ANGIS Recipe Book has also been modified and converted to HTML for public access on the World Wide Web. The On-line Recipe Book can be accessed from the ANGIS home page, together with other documentation about the service, at http://morgan.angis.su.oz.au.

The Introductory Workbook

A workbook to be used in one-day introductory ANGIS workshops was also created from parts of the tutorial book and of the recipe book. This workbook covers the following topics:

- Logging on and off and navigating using the ANGIS 2D menu interface.
- A quick tour of ANGIS, using on-line help
- File manipulation
- Network resources (email, news, gopher, WWW)
- Information retrieval
- Sequence entry and editing
- Similarity searches
- Sequence alignments
- PCR primer design.
- Appendix 1: Some suggested strategies for sequence analysis
- Appendix 2: transferring files between ANGIS and a Macintosh computer

All the previously mentioned teaching materials have been made freely available to university lecturers for use in their own teaching programs. To date, seven Australian universities are known to make use of some of these materials for undergraduate and/or postgraduate teaching.

Materials in Development

A reference guide to the ANGIS programs is currently being written. This book is to bring together all the information available on ANGIS programs, arranged in a consistent format, together with tips for usage. Although not intended as a teaching tool *per se*, it should provide a handy reference for users and give them more confidence to experiment and learn the use of the system by experience. An on-line hypertext version of the reference book is being developed concurrently with the printed version, and can be examined on the World Wide Web from the ANGIS page at http://morgan.angis.su.oz.au.

An exercise book is also in progress to complement the tutorial book. The first part of the book is to present exercises following the tutorial book, so that after completing each tutorial in the book, students can practice what they have learned. The second part is to contain more complex exercises requiring the use of several programs, and 'real life' exercises, drawn from actual research situations, demonstrating the efficient use of the resources available on ANGIS. The exercise book is to be complemented by a book giving complete possible solutions to the various exercises, and also by a 'hint' book that would point students using the exercise book on their own in the right direction, without giving away the whole solution. It is intended that this exercise book will be made freely available in electronic form to teaching institutions who want to use it for their own courses.

Courses and Workshops

Although 94% of the respondents to the ANGIS survey agreed that courses on the use of the system should be organised, there was less agreement on the form these courses should take. Several different formats have therefore been used: one-week practical courses, two-weeks theory and practical courses, and one-day on-site workshops organised in conjunction with local staff. After each one-week and two-weeks course, feedback was obtained from the students using a questionnaire. The effectiveness of on-site workshops was assessed more informally, by talking to the local workshop organisers and to the students.

One-week practical courses

Three 5-day practical courses have been given to date: two in July 1994 and one in January 1995. The July 1994 courses were attended by a total of 77 participants. Because of equipment limitations, the January 1995 course was limited to 25 students. The courses were given in the school of Electrical Engineering of the University of Sydney, where ANGIS is located.

The major component of the course was a series of 10 practical sessions (approximately 3 hours each) making use of the DEC, Alpha and VAX stations in the Faculty's computer workstation laboratory, configured both as X and VT-100 emulator terminals accessing ANGIS. The practical sessions were based on the ANGIS Tutorial Book. Because of the wide range of research interests and prior expertise with ANGIS among the students, it was decided to keep the practical sessions self-paced: students were free to choose which tutorials they wanted to work through, and could work through them at their own pace. Students were also encouraged to bring their own data to the course so that they may apply what they were learning to problems of interest to them, in consultation with the ANGIS staff.

The practical sessions were complemented by 10 informal lectures of approximately 30 minutes each. These lectures did not discuss in detail the algorithms and theory of sequence analysis software. Instead, they gave an outline of the various programs available to carry out specific tasks and of their limitations, together with tips for their efficient use. These lectures covered the following topics:

- Introduction to ANGIS and to the course: the ANGIS interfaces.
- Information Retrieval: how to access information in the sequence databases.
- Gopher and the World Wide Web: molecular biology resources on the Internet.
- Database Search I: BLAST, FastA and other similarity search engines.
- Database Search II: motif databases (Prosite, BLOCKS etc)
- Sequence Comparison: global and local sequence alignment, dotmatrix plots.
- Multiple Sequence Alignment: creating, editing and formatting the alignment.
- Profile Analysis: protein sequence profiles and their use.
- Database Search III: specialised databases, relational databases.
- PCR primer design: programs, limitations and tips for usage.

Two-weeks theory and practical courses

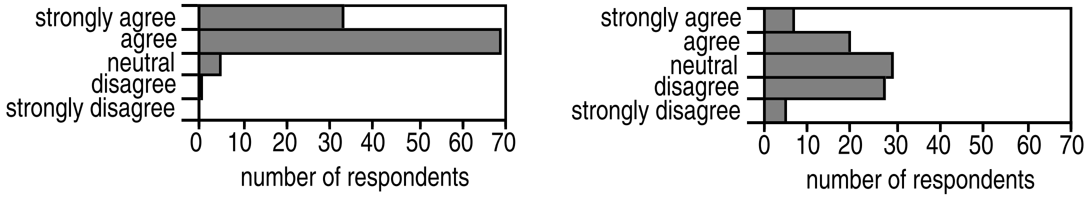
Two theory and practical courses have been given by ANGIS to date. Each of these courses ran for two weeks, and made use of the same facilities as the one-week courses, in the school of Electrical Engineering of the University of Sydney. The first course was given in July 1993, and was attended by a total of 50 participants: 25 who attended both the lectures and the practicals, and 25 who attended only the lectures, due to equipment limitations in the workstation laboratory. The second course took place in July 1995 and was attended by 45 participants, out of which 15 attended only the lectures.

Each course consisted of a series of practical sessions (10 for the 1993 course, 12 for the 1995 course) similar to the practical classes of the one-week courses. In the mornings, a series of 20 lectures covering the theoretical basis of biological sequence analysis software was given by Professor Douglas Brutlag of Stanford University.

The 1995 lecture series was based on the following outline:

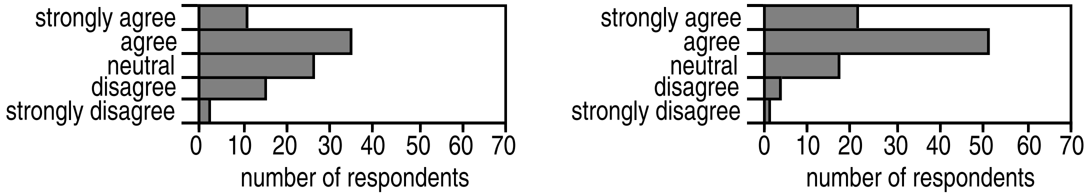
- Introduction: a survey of computer applications in molecular biology
- The Internet: structure and history, applications in molecular biology
- Sequence databases: history, structure, access
- Symbolic pattern matching in biological sequences: string searching, Prosite etc.
- Quantitative pattern matching: weight matrices, neural nets etc.
- Sequence alignment: Needleman-Wunsch, Smith-Waterman, PAM matrices etc.
- Near-optimal sequence alignments
- Rapid database search for sequence similarity: Blast, FastA etc...
- Parallel database search: Blitz and MPSearch.
- Multiple sequence alignments: Clustal, Macaw etc...
- Protein structural motifs: profiles, blocks and Hidden Markov models
- Random shotgun sequencing and sequencing by hybridisation
- Physical mapping of genomes
- Sequence phylogenies: UPGMA, neighbour joining, parsimony etc.
- The restriction mapping problem
- Positional correlations in biological sequences
- Parametric representation of sequences: protein hydrophobicity, charge etc.
- Programs for PCR primer design.

These lectures presented a theoretical background to the programs, including a description of the algorithms used. This was followed by some practical demonstrations using a computer linked to a screen projector. The lectures ranged from practical discussions of the pros and cons of various programs (for example in the PCR primer design lecture) to theoretical expositions of recent developments in bioinformatics.



(a) ANGIS should provide courses on the use of the Service

(b) ANGIS should provide a two week lecture and practical course



(c) ANGIS should provide a five day practical course

(d) ANGIS should give on-site practical workshops to its users

Figure 1: Opinion of ANGIS users on ANGIS courses: summary of the responses obtained in the relevant section of the January 1994 survey of ANGIS users.

One-day Workshops

In addition to the courses, a few on-site workshops have been run by ANGIS staff with local assistance.

The first series of workshops was run in April 1994 at the Australian National University in Canberra. It consisted of two separate workshops run on consecutive days, one for beginners and one for advanced ANGIS users. The introductory workshop was attended by 19 students and included a day-long practical session based on the *Introductory Workbook*, and two informal lectures giving a broad introduction to ANGIS. The advanced workshop was attended by 12 students. It consisted of a day-long practical session based on sections of the *ANGIS Tutorial Book* and of 2 lectures, one giving an overview of the ANGIS programs, and one covering sequence comparison and similarity search programs.

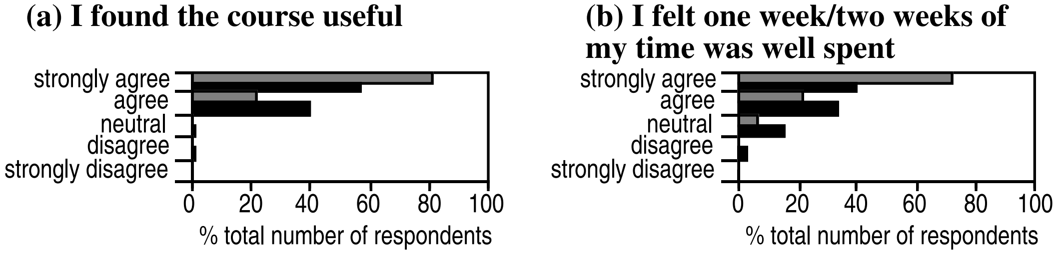
A second series of workshops using a similar format was run over a week in June 1994 at the University of Queensland. It included four one-day introductory workshops attended by a total of approximately 90 students, and one advanced workshop attended by 24 students.

Other one-day workshops without hands-on practical component have been run in all the major Australian centres by one member of the ANGIS staff. In most cases they comprise a lecture (about 60-90 minutes) covering what ANGIS is about and what software is available, followed by an extensive discussion time, and/or practical demonstrations.

Feedback from the courses

Course formats

The respondents to the questionnaire sent to ANGIS users in the beginning of 1994 were strongly in favour of ANGIS giving courses in the use of the service (figure 1), with the on-site workshop being the favourite format (77% in favour, 5% against). This format is certainly the most convenient for researchers in that they need not travel or be away from the bench for a long time. In addition, the on-site workshop usually allows the programs to be demonstrated and practiced on the very



(a) I found the course useful

(b) I felt one week/two weeks of my time was well spent

(c) I would recommend this course to colleagues

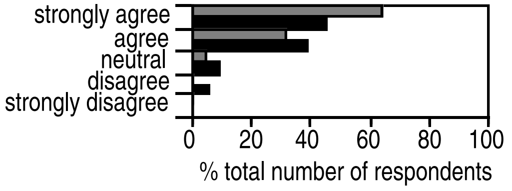
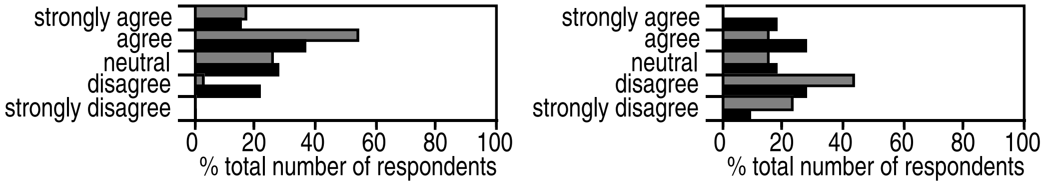
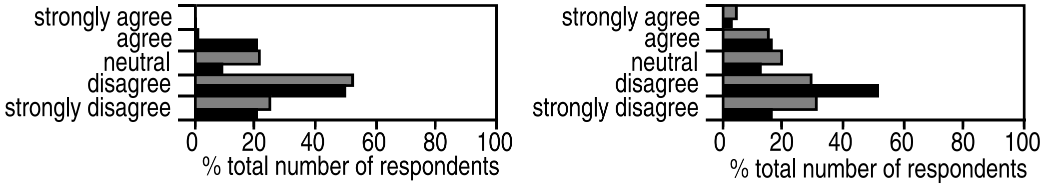


Figure 2: Appreciation of the ANGIS courses by participants: summary of the responses of the participants of the one-week practical courses (grey bars) and of the two-weeks theory/practical courses (black bars) to the relevant statements in the course questionaire.


(a) The course was what I expected

(b) The course emphasized computer technology at the expense of molecular biology


(c) The course relied too much on the use of jargon

(d) The course assumed previous knowledge that I did not possess

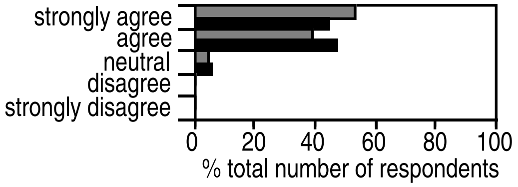

(e) The content of the course was relevant to my work

Figure 3: Perceived relevance of the ANGIS courses to their participants: summary of the responses of the participants of the one-week practical courses (grey bars) and of the two-weeks theory/practical courses (black bars) to the relevant statements in the course questionnaire.

computers the participants are going to use for their research at a later date. However, this format has severe limitations. The time available (usually one day) is only sufficient to practice the use of a few programs. In addition, a suitably equipped computer laboratory is required at the site of the workshop if the participants are to have any hands-on training. For this reason this type of workshop is only suited to universities and not to research institutes which do not have computer laboratories. Finally, the on-site workshop is more labour-intensive for the instructors than the course format, in that the same material has to be presented many times in different places. However, this format is the only one that allows users with only a limited interest in the topic to be reached. These users may then continue their study, either on their own using the tutorial book, or by attending a course at a later date.

The one-week practical course was the next favourite format with the ANGIS survey respondents: 51% were in favour of this format, and 20% were against. From the additional comments made on the survey, it seems that the major problem the respondents have with this format is the need for interstate users to travel to Sydney in order to take the course, and the need to be away from the bench for a week. However, the feedback from the 3 one-week courses which have been given so far suggests that one week is a good duration when taking into account the material to be taught, and that many participants appreciate the opportunity to work at the computer and learn without interruptions and distractions (13 out of the 66

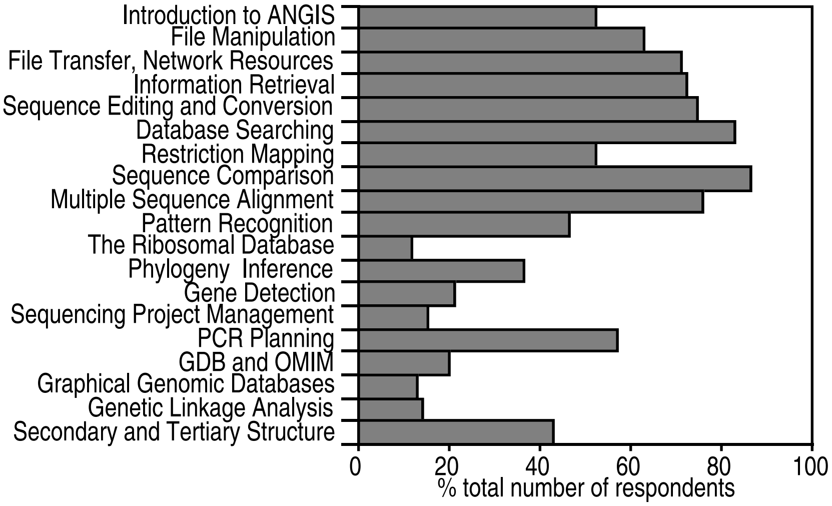


Figure 4: Level of interest in the various practical tutorials given in the 1994 and 1995 ANGIS courses. The participants were given a list of the topics and asked to check the topics they were most interested in.

course participants who returned their questionnaire considered this one of the best features of the course). Of the course participants who returned their questionnaires, 70% strongly agreed and 23% agreed with the statement ‘I felt one week of my time was well spent’. No one disagreed or strongly disagreed with the statement (figure 2, panel b).

The general opinion of the respondents to the 1994 ANGIS user survey was less in favour of a two-weeks theory and practical course: 36% of ANGIS users were against the idea, 31% were in favour (figure 1(b)). However, most of the users who did attend such a course found it useful (figure 2). Again, from discussion with users and from the feedback obtained from course participants, it appears that the major reason for the opposition to two-weeks courses has to do with the time commitment required. In addition, a lot of biologists feel that they do not need to know about the algorithms behind the software they use, but just require enough practical knowledge of the software to use it efficiently and to understand the program output. A two-week course with a strong emphasis on algorithms and recent developments in bioinformatics is perceived as a waste of time and resources by many users. This does not imply that there is no need for such a course. The feedback obtained from biologists with an interest in the more theoretical aspects of bioinformatics and who attended a two-weeks course confirms that such a course is definitely needed in Australia, even if its target audience is limited, and that ANGIS is in an excellent position to offer such a course because of its resources and expertise. However, the emphasis of the lectures on the theory of biocomputing needs to be clearly spelled out when advertising the course, otherwise it may attract students who are mostly interested in the practical use of the programs and would find a shorter practical course more suitable. This accounts in part for the slightly lower level of satisfaction with the two-week courses when compared to the one-week courses (figure 2) and for the fact that the two-week courses met the expectations of their participants less than the one-week courses (figure 3).

Practicals

The practical sessions were run in the same way in the one-week and the two-week courses: the students were left to choose which areas of the tutorial book to work on, and were free to proceed at their own pace, with the staff of ANGIS available to answer questions. The tutor to student ratio was on average one tutor per 10-12 students. Guidelines as to which tutorials to do first in order to avoid confusion were given, but only as a recommendation. This allowed students with varying levels of expertise to all work at a level that was suitable for them. The self-paced nature of the practical sessions was in general appreciated by the students, and 10 students nominated it in their questionnaire as one of the best features of the course. Only one student commented that he much preferred a more structured format, with set tutorials to do each session, and a formal demonstration of how to go through these tutorials at the beginning of each session. The idea of a formal demonstration was supported by a few other students. The lack of suitable equipment made such demonstrations impractical, but they are being considered for future courses.

A consequence of the self-paced nature of the course was the possibility for students to bring their own data to analyse, so they could apply the experience they had gained to problems of direct relevance to them, with the ANGIS staff at hand to assist them. This feature was appreciated by the students, and in a few cases nominated as the best feature of the course.

One of the main complaints about the practical sessions was that they made use of computers unfamiliar to the students. A number of students commented that they would have much preferred using a computer of the type they normally use to access ANGIS (Macintosh or IBM-compatible PC), rather than UNIX workstations. This complaint has been partly addressed in the more recent courses by organising demonstrations on how to access ANGIS using Macintoshes or PC's, and on how to transfer files between these microcomputers and ANGIS using some readily available software. Little can be done at this stage to address the more specific questions on how to set up individual microcomputers to access ANGIS. Due to the wide variety of possible computers, software and network configurations, these queries are best answered by local computer systems administrators.

The participants of the 1994 and of the 1995 courses were asked to list the tutorial they were most interested in. The results are summarised in figure 4. Overall, the most popular tutorials were those demonstrating the types of programs

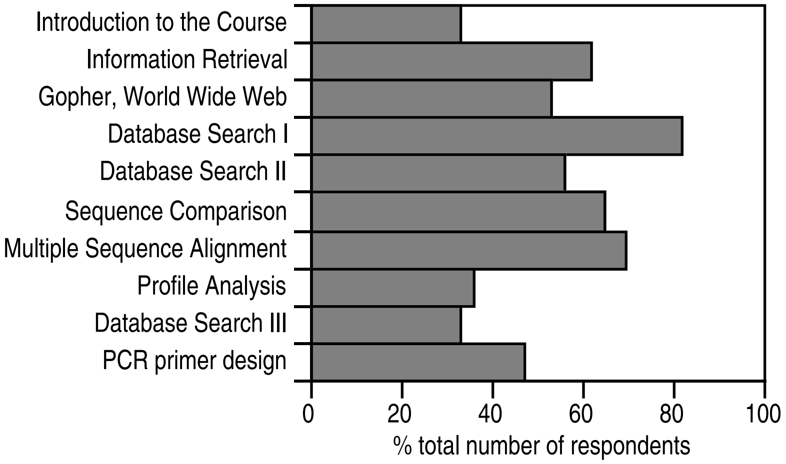


Figure 5: Level of interest in the various practical talks given in the one-week ANGIS practical-only courses. The participants were given a list of the talks and asked to check the talks they were most interested in.

which are most used on ANGIS (sequence comparison, database similarity searching, database information retrieval, sequence editing and conversion and multiple sequence alignment, together with the more general programs such as those involved in file manipulation and network resources). The least popular tutorials were those dealing with specialised databases such as GDB, ACeDB and RDP, which are accessed only by a small subset of ANGIS users.

Short Practical Talks

The talks given in the one-week courses deliberately took a very practical approach. Instead of going into details of the mechanisms used by the programs, they exposed the pros and cons of the programs available on ANGIS to carry out a specific task, and gave a few tips for their use. Overall, the practical orientation of the talks was appreciated (over 83% of the respondents agreed or strongly agreed with the statement 'the talks were useful', and less than 2% disagreed). However, 5 out of the 66 respondents commented in their questionnaire that they would have liked more details about the programs, especially in relation to customising program parameters for specific uses, and interpreting the results.

The level of interest in each of the talks is summarised in figure 5. As seen with the practicals, the talks which generated the most interest were those which covered the programs most used on ANGIS (Database Search I, Sequence Comparison, Multiple Sequence Alignment and Information Retrieval).

Theory Lectures

The theory lectures in the two-weeks courses were given by Professor Douglas Brutlag from Stanford University and covered program algorithms, tips for efficient use, as well as recent developments and advances in the field of bioinformatics.

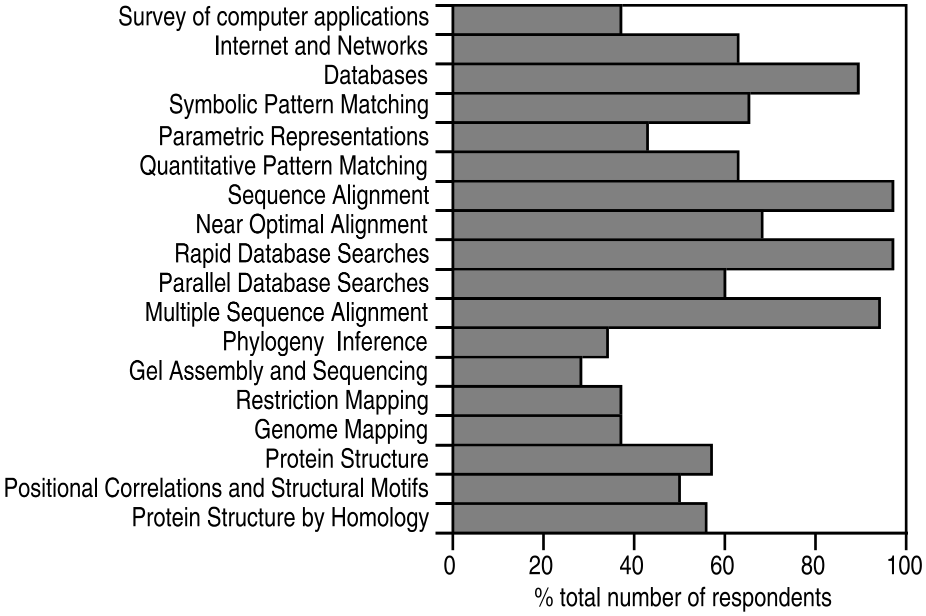


Figure 6: Level of interest in the theory lectures given in the two-weeks ANGIS courses. The participants were given the list of lectures and asked to select the lectures they were most interested in. The lecture on PCR primer design is not listed here because it was a late addition and therefore was not listed as an option on the course assessment questionnaires.

Similar lectures have been given by Douglas Brutlag in other courses in the USA and in Europe. Most of the lectures were illustrated by practical demonstrations using a computer linked to a screen projector. Because Professor Brutlag was not familiar with ANGIS, these demonstrations made use of software not always available to the course participants. As a result, a number of students commented about the lack of correspondence between lectures and practical sessions. This concern should be addressed in future courses by including demonstrations of ANGIS software in the lectures.

The coverage in the lectures of algorithms and of other material with a strong mathematics/computer science background was met with mixed reactions from the participants. Some listed it as the best feature of the course, while others considered it the worst feature, and commented that the lectures were aimed more at mathematicians and computer scientists than molecular biologists (this mixed reaction is apparent in the response to the statement 'the course emphasised computer technology at the expense of molecular biology', shown in figure 3(b), black bars). Since an understanding of its algorithm is often necessary for a fully informed, efficient use of a program, the solution to this problem probably lies in a careful advertising of this type of course, so that it attracts mostly advanced users with an interest in bioinformatics and adequate background knowledge. Other students should be directed towards more practical courses, which can introduce them to many of the statistics and computer science concepts, as well as give them a good familiarity with ANGIS and its interfaces.

The interests of the various participants in the lectures are summarised in figure 6. Again, the topics which engendered the most interest corresponded to the programs that are used most on ANGIS.

Conclusion

The fact that every ANGIS course so far has been filled to capacity indicates that a lot of molecular biologists in Australia feel the need to learn how to use computer software appropriately and efficiently for their research. This interest is mainly focussed on the computer applications that they use most in their field, and for most molecular biologists this includes software for sequence alignment, database searching, database information retrieval and multiple sequence alignment. Overall, ANGIS users preferred a practical approach to this teaching, preferably one that they could apply directly to their own data. In general, the presentation of the algorithms and statistics behind the programs should be made directly relevant to problems in molecular biology, and should be directly applicable to practical tasks such as understanding a program's limitations, modifying program parameters to suit specific circumstances, and evaluating a program's results.

There is however a need for a more in-depth instruction in bioinformatics, one that is required if Australia is to make a significant contribution to the field instead of just reaping the rewards of software developed overseas. A number of molecular biologists have expressed interest in learning in detail about program algorithms and recent developments and research directions in this area. In addition to fostering this interest, the lectures in the two-weeks ANGIS courses have provided an excellent introduction to these topics.