Web-based Protocols for Bioinformatics Education

Bruno A. Gaëta, Kirsten Balding, and Timothy Littlejohn
Australian Genomic Information Centre, The University of Sydney
kirsten@angis.org.au

The many genome projects initiated in the last few years have brought about an explosion in the amount of DNA and protein sequence and structure data available to biologists. Computers have become essential tools for the analysis of this information, and as a result, there is a growing demand among molecular biologists for education in bioinformatics, a new field of study at the forefront of computer science and molecular genetics.

Bioinformatics has a wide range of uses in molecular biology, for example the identification of the function of biomolecular sequences, the analysis and prediction of the three-dimensional structure encoded by a given sequence, and the inference of phylogenetic relationships between a group of sequences or their organism of origin. Furthermore, bioinformatics tools are useful for the management of experimental results as well as the planning and design of experiments.

The Australian National Genomic Information Service (ANGIS) is an Internet-based service providing Australian molecular biologists with access to a large collection of databases and software for genome analysis. ANGIS is based at The University of Sydney but is accessed over the Internet by over 3000 scientists from academic institutions, research institutes, hospitals and biotechnology companies throughout Australia. ANGIS uses the World Wide Web (WWW) as a front end to over 70 databases of DNA and protein sequences, genome information, molecular structures, sequence motifs and literature citations. The service also provides access to over 400 programs for biomolecular sequence analysis, database information retrieval and computational genetics. The software and databases are integrated under a simple WWW user interface that facilitates its use by biologists with limited computer literacy. This WebANGIS interface provides a useful and user friendly means of accessing bioinformatics programs and databases. It is ideally suited to biologists who have an awareness of the available bioinformatics programs and their uses. However, the user friendly nature of the service and the power of the bioinformatics databases as information resources are encouraging growing numbers of novice users to attempt bioinformatics analyses. Thus, commonly encountered questions relate to (1) the choice of suitable programs/databases and (2) the pertinent usage of programs including appropriate choice of parameters. Furthermore there is currently no guidance for biologists through what are often necessarily complex analyses involving the use of several programs.

Choosing a program or set of programs for a bioinformatics analysis is not a straightforward process, but one that requires a good understanding of the function of the programs and of the nature of the biological data being studied. Hence, it is important that users have the skills required to make informed decisions.

To this end, for the last five years, ANGIS has been providing its users with a range of educational services. These include courses, teaching manuals and online help. Disadvantages of these educational activities are that they are only available to limited audiences (courses), associate additional costs (courses, books), and require a significant investment of time which many scientists are unable to reconcile with the daily demands of practising laboratory-based molecular biology. Furthermore, detailed knowledge gained at courses tends to be most effectively retained by scientists by continual use of the service (which is often not necessary when doing molecular biology). A more effective (and time efficient) means of facilitating the learning of bioinformatics skills may be to teach the use of software and appropriate analytical decision making at the time of doing an actual analysis.

These issues are being addressed by the creation of on-line protocols describing commonly used sequence analysis procedures in step-by-step fashion. A protocol describes which programs to use, together with explanations of the program function and guidelines for the interpretation of the program outputs.

The protocols are accessed through the WWW as clickable image maps. In order to facilitate comprehension by users, these maps represent flowcharts modelled on the biochemical pathways notation familiar to most biologists. These flowcharts clearly identify the type of file required as program input, the program name and the type of output files created (in place of the chemicals and biochemical reactions of a biochemical pathway). The program name itself is a hypertext link launching the program in another WWW browser window. This allows the user to carry out the analysis with the protocol still visible. Information on the input and output files is also available by clicking on the relevant icons in the flowchart.

The flowcharts may also include ‘decision boxes’ that represent a decision made by the user when examining a program output, for example whether the output is of sufficient quality to warrant further analysis or whether it should be refined first using a different approach. These decision boxes are hyperlinked to guidelines for the interpretation of the output and other information required for making a decision. In some cases protocols may also contain steps representing manual intervention by the user, for example editing a file using a text editor, or using a non-WebANGIS drawing program to refine the final program output into a publication-quality figure.

Most protocols have been broken down into a collection of sub-protocols, or modules, as the complexity of most bioinformatics protocols results in very large flowcharts. These are sub-optimal since they are unsuitable for displaying on the WWW and are confusing for users. Collapsing a protocol into modules facilitates the display of a simple and clear overview of the entire protocol, which all fits on one page. Clicking on one of the icons representing a module displays the flowchart for the corresponding sub-protocol. This approach is suitably economical since it allows modules to be re-used in a range of different protocols.

Protocols are initially designed by ANGIS scientific staff, often in collaboration with expert biologists. The protocols are then drawn using the flowcharting program **visual thought**. One of the outputs from **visual thought** is a text file describing the flowchart in a LISP-like logic computer language, which can be processed by other computer programs. This output file is entered into an ANGIS-developed program, which creates the clickable image map automatically and inserts most of the relevant hyperlinks. This approach allows protocols to be created and modified with minimum intervention, and also maintains consistency of format between different protocols.

At the current stage of development, the initial protocols are being implemented. These include methods for molecular phylogeny, data collection, multiple sequence alignment, the design of PCR primers from a group of related sequences, and three dimensional structure prediction from protein sequences. The next step involves recruiting a number of expert users to test these protocols. Further protocols will then be developed, ideally including all of the programs available in WebANGIS in at least one protocol. These protocols may then be used for research or for *in silico* practicals teaching the basics of bioinformatics analysis to undergraduate students. It is hoped that scientists will eventually be able to create and save their own custom protocols for re-use, and ultimately, that protocols can be refined sufficiently for integration into automatic sequence analysis programs for the processing of large amounts of sequence data such as those produced by genome sequencing projects.

Web address: http://www.angis.org.au/