This article reprinted from:

Neumann, M. and N. Provart. 2006. Using customized tools and databases for teaching Bioinformatics in introductory biology courses. Pages 321-328, *in* Tested Studies for Laboratory Teaching, Volume 27 (M.A. O'Donnell, Editor). Proceedings of the 27th Workshop/Conference of the Association for Biology Laboratory Education (ABLE), 383 pages.

ISBN 1-890444-09-X

The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises.

Visit ABLE on the Web at:
http://www.ableweb.org

Using customized tools and databases for teaching Bioinformatics in introductory biology courses

Melody Neumann and Nicholas Provart

Department of Zoology and Department of Botany
University of Toronto
Toronto, ON M5S 3B2
neumann@botany.utoronto.ca and provart@botany.utoronto.ca

Abstract: The developing field of Bioinformatics has revolutionized modern cell and molecular biology, in the research laboratory and classroom. It is important that students receive an introduction to Bioinformatics, and this is best accomplished by using bioinformatics tools to address biological questions, including a critical analysis of the output. Our second year Biology students (>1500) expressed frustration with the use of publicly available bioinformatics tools and databases, since these can be quite complicated, and frequently place a significant emphasis on sophisticated computing skills, complex statistical theory, or computational algorithms. This computer-based mini-workshop demonstrates exercises where students were able to develop a basic understanding of Bioinformatics by using customized tools and databases to solve real biological problems. Simplified bioinformatics tools containing the key features of common tools (eg. Blast and ClustalW) can be accessed from http://bio250y.chass.utoronto.ca/labs/lab_notes/lab4/newBioinf /index.htm . Pedagogical and logistical issues surrounding the implementation and improvement of Bioinformatics computer labs, particularly for large introductory courses are also addressed.

Introduction

The ability of the researcher to generate large quantities of data has changed modern biology from a primarily laboratory (or field) based science, to one that includes a significant information science component. The relatively recent development of Bioinformatics is a clear example of this phenomenon. Bioinformatics is a rapidly growing area that uses computational approaches to help solve biological questions. In addition to advances in basic biological research, Bioinformatics has provided tools that enable the extraction of information that is used in commercial applications such as drug discovery, clinical diagnostics, and agricultural biotechnology. For modern bioinformaticists working within the fields of cell and molecular biology and genetics, the challenges include: gene discovery and characterization; the design of better molecular modelling tools; integration of data across biological systems; and the study of phylogenetic relationships, to name a few. All of these will lead to an

increased understanding of the structure and function of genes and proteins, including how they may have evolved.

As a result of the ever-increasing role of bioinformatics tools to study cell and molecular biology (it is now difficult to find research papers in these fields that do not use some sort of bioinformatic approach), we decided to include Bioinformatics in our curriculum for our Introductory Cell and Molecular Biology course at the University of Toronto. The use of bioinformatics tools can also connect and reinforce lecture material, since students can look at real examples that show concepts explicitly. Furthermore, at an introductory level, it can be difficult to "invite students into the field". Bioinformatics labs can provide a unique way for students at all levels to join the academic discourse, since the analyses they perform are the same as those used by researchers worldwide.

The complexity of the tools and the interpretation of the output can discourage many students and we felt that having a TA-taught, real-time Bioinformatics lab was the optimal way of introducing students to Bioinformatics. Since our course contains approximately 1500 students, and we have only one three-hour lab session devoted to Bioinformatics, there are several logistical and pedagogical challenges. Experience using publicly available web resources indicated that: the majority of students were so overwhelmed that they lost sight of the biological problem that was driving the use of the tools; many students became frustrated as a result of becoming "lost in cyberspace"; and TAs lost confidence when faced with inexplicable public database annotations or student output. Due to our large class size, we have simultaneous lab sections operating between 9am and 9pm most days. This creates a further logistical concern because depending on the time of day, the use of publicly available web-based bioinformatics resources can result in extremely lengthy wait times for output. Finally, the graphical user interfaces (GUIs) for publicly available bioinformatic tools are complex and change frequently. For introductory studies, these GUIs contain many options that are rarely used and when students alter these options, it can result in an output that is non-sensical and cannot be explained by the TA. The fact that GUIs change frequently also means instructors must monitor interfaces constantly, and alter student and TA notes to reflect the changes.

Biology Student WorkBench (http://bsw-uiuc.net/) is an example of an extremely useful means of teaching Bioinformatics to undergraduate students. However, we found that these types of resources are still too complicated for introductory courses, and led us to develop our own tools and databases. The advantages of using customized tools and databases are: instructors can optimize these tools for their courses; customized databases mean wait times are short; and students and TAs are not overwhelmed by complex GUIs and outputs. In our customized Blast (Altschul et al., 1990) and ClustalW (Higgins et al., 1994) tools, we have retained the essential features of the original GUIs and output, but have removed some of the unnecessary complexity from the GUI. The output provided by these tools is similar in format to that provided by public agencies and we have retained the hyperlinks to public resources. In summary, customized tools and databases simplify the mechanics of using bioinformatics tools, thereby ensuring that instructors and students can focus on the biological problem and the complexities of real biological data.

General Teaching Notes

In our introductory course, we place the emphasis on the biological problems that Bioinformatics can help us solve, rather than the detailed computational, statistical, and mathematical theory. Prior to the lab, students will have read an overview of the history of the development of public web-based

resources and the basic principles behind the Blast and ClustalW algorithms. Students will have also had at least one lecture where Blast and ClustalW are introduced to solve biological problems.

Students work in pairs at a single computer throughout the lab. Initially, they log on to the computers and bookmark the Bioinformatics laboratory page (http://bio250y.chass.utoronto.ca/labs/lab_notes/lab4/newBioinf/index.htm). The TA then gives a broad overview of Bioinformatics and outlines the main features of the tools the students will encounter (via a PowerPoint presentation linked to the web page). The TA subsequently leads the class (24 students) through two biological problems involving the use of the customized Blast and ClustalW tools. Biological problems contain detailed instructions and questions to guide students to an answer to the problem. Our TAs report that having the students work in pairs at a single computer, as well as going through two introductory problems as a class, greatly enhances their ability to ensure that all students understand the lab and do not become frustrated. The student pairs are then given a third, more challenging biological problem that they work on independently and hand in for marks at the end of the lab. The instructions for the third biological problem are limited and provide the opportunity for students to use what they have just learned to answer a series of questions related to a realistic biological problem. In addition to using our customized tools and databases, the third problem also makes use of Simple Modular Architecture Research Tool or SMART (http://smart.embl-heidelberg.de/) to look for protein domains. We have not modified the publicly available SMART tool since the GUI is quite straightforward and the output is clearly laid out. We feel that the blend of customized tools that are ultimately hyper-linked to publicly available databases (eg. GenBank) as well as the use of simple tools such as SMART, is a good compromise that enables us to simplify the teaching of Bioinformatics without making the experience so artificial that it does not adequately reflect publicly available databases and tools.

Sample Student Problems and Teaching Notes

Sample Biological Problem # 1: Imagine that you are working in a pathology lab and need to identify the bacterial species contained in a sample from a very sick patient. Once you know the species with which the patient is infected, the doctor will be able to recommend an appropriate antibiotic. You have purified the bacteria from the patient's sample and extracted bacterial DNA from a single colony. You then performed PCR using primers that anneal to the region containing the 16S rRNA gene. You have sequenced the PCR product and you are now ready to identify the bacterium.

Step 1: Go to the Bioinformatics Lab Page, open up the 16S rRNA gene nucleotide sequences and copy your assigned unknown bacterial sequence.

Step 2: Go back to the Bioinformatics Lab Page and click on the Blast link, which will open up the Botany Bioinformatics Blast page. Paste your sequence into the large empty window. Your sequence is in FASTA format already, but click on the link to find out what FASTA format means.

Step 3: In the pull down window on the left called "Program" choose **blastn** since you are going to perform a nucleotide Blast. In the right-hand window pull down BIO250 NUCLEOTIDE DATABASE since you want to compare your nucleotide sequence with all other nucleotide sequences in the database.

Step 4. Hit the "Search" button and scroll down to see your results.

(a) With which bacterial species is the patient most likely infected?
(b) What features of the Blast output influenced your decision?
(c) Which sequence is the query sequence and which one is the subject sequence?

Step 5: Click on the hyperlink associated with your best blast match to get to the GenBank record.
(d) What is the Accession Number of your best match?
(e) Who submitted this sequence?
(f) Why is there no CDS (sequence coding for amino acids in protein) associated with this record?
(g) If you were to perform the same blastn analysis with your bacterial sequence a year from now using a public database where sequences are constantly being added, would you expect to obtain the same Score? E-value? Explain.

Teaching Notes for Biological Problem # 1
In Sample Problem #1, students perform a Blast N of unknown bacterial 16S rRNA gene sequence. Students have already read how Blast searching works in the pre-lab component of this lab and are given the opportunity to see it in action. Our goals of this exercise are to expose students to the mechanics and theory of doing a Blast search without losing sight of the biological question. We also want to reinforce an understanding of what Score and E-values are and explain that public databases are dynamic. TAs emphasize that databases are annotated (but not always curated) and are constantly evolving. Use of the custom tools and databases allows students to concentrate on the key steps of Blast as well as the opportunity to examine realistic output. The retention of the hyperlinks to GenBank records gradually introduces students to NCBI resources and formats.

Sample Biological Problem # 2: You and your colleagues in the pathology lab have sequenced the 16S rRNA gene sequence for a total of five bacterial species. You are curious to see how similar your 16S rRNA gene sequence is to the other four sequences. You have learned that multiple sequence alignments may provide some information about this and would like to try it.

Step 1: Go back to the Bioinformatics Lab Page and copy all five 16S rRNA gene sequences.

Step 2: Click on the Multiple Sequence Alignment link on the Bioinformatics Lab Page. Paste your bacterial sequences into the window. You do not need to give your output a name in the box on the right.

Step 3: Under "Desired Output Style" choose the black and white setting. Under the heading "Type" you should choose DNA since that is what you are aligning.

Step 4: Click the **Align** button and wait for your results to appear.

Step 5: Scroll down to see your alignment.
(a) What do you think it indicates when the nucleotides are shaded in black? What do you think it means when they are on a white background?
(b) Would you estimate that all five sequences are quite similar or not?

(c) Which bacterial sequence appears to differ the most from the others? Click on "View Tree" at the top of the output page for a different visual representation. (Note that trees generated by ClustalW analyses represent sequence similarity and are not intended to be interpreted as a tree indicating evolutionary descent).

(d) Does your bacterial sequence cluster with any other sequence? Briefly describe the similarity between your sequence and those of your colleagues.

Teaching Notes for Sample Biological Problem #2

In Sample Problem #2, students perform a Multiple Sequence Alignment (ClustalW) of bacterial 16S rRNA gene sequences from five bacterial species. Students have also read a little background about ClustalW prior to coming to the computer lab. The goals of this exercise are to reinforce an understanding of sequence similarity and provide a further exploration of sequence alignment, especially multiple sequence alignment. Using the simplified custom tools allows students to do this within the context of a realistic biological question and practice what they have learned in lecture. This exercise also gives TAs a chance to explain that trees drawn using ClustalW analyses are not phylogenetic trees but are similarity trees. If desired, this can open up a discussion regarding "homology" versus "similarity".

Sample Biological Problem #3. You are about to begin a summer job in a research lab that is working on human eye diseases. You want to learn a bit more about your project before you start experiments in the lab, so your supervisor has given you a file with the amino acid sequence of the human protein you will be studying and has suggested that you find out as much as you can using Bioinformatics tools. You have decided you would like to find out the identity of your sequence, whether it is similar to sequences in other organisms, and whether there are any interesting protein domains.

Step 1: This time you have an amino acid sequence so you need to do a BlastP. Make sure you choose the correct database with which to compare your sequence.

(a) How long was the amino acid sequence that you submitted?

(b) What protein does your sequence appear to code for? What is the score and E-value for this Blast match?

(c) Look at the alignment of your query sequence and the subject sequence. Why have some regions been filtered when your sequence was blasted?

(d) What is the Accession number of your best blast match? Is there a CDS? How long is the mRNA for this sequence? Is this what you would expect based on (a)? Explain.

Step 2: Click on the hyperlink called "Domains" on the far right-hand side of the top of the GenBank record. Mouse over the conserved domain region. You should see a pale yellow box with information regarding the domain. You are now viewing a summarized entry in the NCBI Conserved Domain Database (CDD). If you click on the domain, you will get the full entry that contains a lot of extra information.

(e) Concentrate on the information in the yellow box. With which gene family does your protein share homology?

(f) Is the sequence you are analyzing likely to code for the same size of protein as that encoded by the VMD2 gene described? Explain.

(g) What type of protein are you likely to be studying and where have these proteins been found to be localized?

Step 3: Based on the information obtained in Step 2, you would like to find out more about the protein structure and would like to do a domain search. Go back to our Bioinformatics web page, copy your amino acid sequence again and click on the link to the SMART domain search. You should be in a window with a blue background corresponding to SMART in "Normal" default mode. Paste your sequence into the "Sequence" window and click **Sequence SMART**.

(h) List the domain(s) found and give the amino acid location(s) of this/these domain(s).

(i) Click on a domain (blue bar) and examine the amino acid composition. What characteristic is common to the majority of amino acids in each domain? How does this correlate with (g) above?

Step 4: You noticed when you did the blastp that there were several sequences from other organisms that resembled the sequence you are working with. You decide to align some of these sequences. Assume that you already copied some sequences of interest and created a file called "Protein Sequences for Multiple Sequence Alignment" to use for this analysis (it is on the Bioinformatics Lab Page).

(j) Which tool did you choose? Why? Were there any additional features of this tool that give you more information about your sequence?

(k) Based on the output, which organism's sequence is most similar to your human sequence? Which is least similar?

(l) Suggest an explanation for the greater number of differences between the amino acid composition of all three organisms in the region of the alignment after about position 360 (towards the carboxy end of the protein).

Teaching Notes for Sample Biological Problem # 3

Sample Problem #3 is more difficult than the other problems. This problem is very realistic and allows students to put together what they have learned (Blast and ClustalW) as well as explore the concept of protein domains using the NCBI's Conserved Domain Database (CDD) and SMART. We have asked additional conceptual questions, but have provided less guidance in the exploration of the problem. Students may explore many concepts including: "DNA makes RNA makes Protein"; the relationship between mRNA and protein sequences (eg. to look at start and stop codons); protein domains and conserved domains, the physico-chemical characteristics of amino acids and how these characteristics affect arrangements of amino acids within protein domains; and the concept of conserved versus variable protein domains and how this might relate to evolution.

Technical Notes

All of the Bioinformatics resources in this workshop may be used directly by other ABLE members. This 3-hour lab is launched from a web page located at http://bio250y.chass.utoronto.ca/labs/lab_notes/lab4/newBioinf/index.htm and all analyses are run on the Botany Beowulf Cluster webserver (bbc.botany.utoronto.ca). Just e-mail us (neumann@botany.utoronto.ca) a few weeks in advance, if you would like to use our exercises directly. If you choose to use our exercises, databases and web server, the only other thing you need to provide your students with are computers with internet access and a web browser. No other tools are required.

If you would like to customize your own databases and host these, along with our customized tools at your own institution, a webserver (eg. Apache) able to run PERL is required. We used OpenSource or Shareware to customize the Blast and ClustalW tools, and created our customized protein and nucleotide databases by downloading databases from The Arabidopsis Information Resource (TAIR). The nucleotide sequences of *Arabidopsis thaliana* coding regions used in the nucleotide database and amino acid sequences for *A. thaliana* proteins can be found at: ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/blast_datasets/ATH1_cds_cm_20040228 and ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/blast_datasets/ATH1_pep_cm_20040228. We used these databases as a starting point, and then added the sequences we needed (bacterial and animal sequences) for the lab exercises. Any small database (or portion thereof) could be used instead. All of the required software needed for installation on the web server is OpenSource or Shareware and is listed as follows:

- NCBI Blast, blastable databases (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/) GD for graphics generation (http://www.boutell.com/gd/) ClustalX (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/)MView for viewing and formatting Multiple Sequence Alignments (http://bioweb.pasteur.fr/docs/mview/) TreePlot for drawing trees (http://www.pge.cnrs-gif.fr/bioinfo/treeplot/index.php)Interface code (available upon request from provart@botany.utoronto.ca)

Acknowledgements

We would like to thank Anne Cordon for searching the NCBI databases to provide the bacterial nucleotide sequences, and Jonathan Taylor for his assistance with the construction of the customized protein and nucleotide sequence databases. The webserver on which this lab runs is part of a computer cluster funded by Genome Canada.

References and Web Sites

Altschul SF, Gish W, Miller W, Myers EW, Lipman, DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403-410.

Baxevanis AD and Ouellette BFF. 2001. Bioinformatics: A practical guide to the analysis of genes and proteins. 2nd ed. Toronto: Wiley Interscience; 470 pp.

Higgins D., Thompson J., Gibson T., Thompson J.D., Higgins D.G., Gibson T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673-4680.

TAIR website and URL's for databases
ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/blast_datasets/ATH1_cds_cm_20040228
(nucleotide sequence of *A.thaliana* coding sequences)
ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/blast_datasets/ATH1_pep_cm_20040228
(amino acid sequence of *A.thaliana* proteins)

For general and specific information regarding BLAST:
NCBI site: http://www.ncbi.nlm.nih.gov/About/index.html

For general and specific information regarding ClustalW:
EBI Site: http://www.ebi.ac.uk/clustalw/

Web sites of interest:
Biology Student WorkBench: http://bsw-uiuc.net/
Howard Hughes Medical Institute: www.hhmi.org/grants/lectures/biointeractive/vlabs/index/htm

About the Authors

Melody Neumann received her BSc in Biology from Simon Fraser University, her MSc from the University of British Columbia, and her PhD from the University of Sydney, Australia. She is a Lecturer in the Department of Zoology and develops laboratories and tutorials for introductory and advanced cell and molecular biology courses. Her professional interests include the development of investigative laboratories, technology in the classroom, and the teaching of scientific writing at the undergraduate level, especially for large classes.

Nicholas Provart received both his BSc in Molecular Genetics and Molecular Biology, and his MSc in Plant and Microbial Biology from the University of Toronto. He received his PhD in Plant Molecular Biology from the Free University in Berlin. He is an Assistant Professor in the Botany Department and his primary research focus is using large-scale biological data sets for hypothesis generation in plant stress biology. Hypotheses are tested in his "wet lab".