BIOCHEMISTRY AND MOLECULAR BIOLOGY EDUCATION
Vol. 35, No. 2, pp. 119–124, 2007

Articles

Introductory Bioinformatics Exercises Utilizing Hemoglobin and Chymotrypsin to Reinforce the Protein Sequence-Structure–Function Relationship[S]

Received for publication, July 28, 2006, and in revised form, September 21, 2006

Jennifer K. Inlow‡, Paige Miller§, and Bethany Pittman§

From the Department of Chemistry, Indiana State University, Terre Haute, Indiana 47809

We describe two bioinformatics exercises intended for use in a computer laboratory setting in an upper-level undergraduate biochemistry course. To introduce students to bioinformatics, the exercises incorporate several commonly used bioinformatics tools, including BLAST, that are freely available online. The exercises build upon the students' background knowledge of hemoglobin and chymotrypsin, and foster a better understanding of how protein sequence relates to structure and function.

Keywords: Bioinformatics, BLAST, sequence alignment, bioinformatics tools.

With the rapid increase in the amount of biological sequence and structure data available [1, 2], bioinformatics methods have now been adopted and are being used routinely by scientists in all branches of the biosciences. It is, therefore, increasingly important that students of the biosciences gain a basic knowledge of bioinformatics, including a familiarity with commonly used tools, such as Basic Local Alignment Search Tool (BLAST),[1] and the ability to use databases to acquire information. The exercises described herein, which utilize bioinformatics tools, are a simple way to introduce students to these tools and help them begin to develop the skills necessary to use them. These exercises have been used, beginning in Fall 2004, in the upper-level undergraduate Biochemistry I course at Indiana State University (ISU) in which approximately 50 junior or senior science majors enroll each fall. The emphasis of the course is the structure, function, and analysis of biomolecules, and the exercises were specifically designed to build on and reinforce topics that are already covered in the course. Hemoglobin and chymotrypsin are the focus of the exercises because these two proteins are covered in detail in all the commonly used biochemistry texts [3–9]: hemoglobin in the context of ligand binding and allosterism, and chymotrypsin in the context of enzymatic reaction mechanisms. Therefore, it should be straightforward for instructors of upper-level biochemistry courses to adopt these exercises for use in their own courses.

A number of positive outcomes can be expected from the incorporation of these bioinformatics exercises into a biochemistry course. The exercises:

1. teach students to use a few web-based bioinformatics tools and biological databases;
2. reinforce facts and concepts covered in the course, particularly the relationship between protein sequence, structure, and function, but also homology, the properties of the different amino acid side chains, and enzymatic specificity;
3. foster research-appropriate thinking because students must apply facts they have learned to a new situation, using tools they have not used before;
4. engage students' interest because they are complimentary to traditional lectures in that they provide a hands-on means for students to learn actively.

The two exercises described have been used in a lecture course, and each one is designed to be finished within a 50-min class period but can easily be extended to a second class period if desired. The exercises would also be appropriate for a laboratory course, and both exercises could then be completed in a single lab period. Exercise 2 does not build upon Exercise 1, so either exercise could be done first.

PREPARATION AND MATERIALS

Resources Required—Each student must have access to a computer with internet connection; a printer should be available so that they can print and turn in their output. At most institutions this will require that the class session to be held in a computer lab. At institutions where all students are required to have their own laptop computers and wireless internet access is available in classrooms, the exercises can be carried out in the usual teaching classroom. Depending on class size and the number of computers available, students may have to work in

[S] The on-line version of this article (available at http://www.bambed.org) contains additional details concerning the supplementary materials.

‡To whom correspondence should be addressed. E-mail: inlow@carbon.indstate.edu.

§These authors contributed equally to this work.

[1] The abbreviations used are: BLAST, Basic Local Alignment Search Tool; EMBL-EBI, European Molecular Biology Laboratory-European Bioinformatics Institute; ExPASy, Expert Protein Analysis System; FASTA, FAST All; NCBI, National Center for Biotechnology Information; PDB, Protein Data Bank.

 DOI 10.1002/bambed.30

pairs. In the Biochemistry class at ISU, students worked in pairs or groups of three or four. Pairing the students worked well because it encouraged discussions in which the students learned from each other. Groups of three or four students did not work well because the students could not gather around a single computer easily, and at least one person tended to be left out of the discussions.

The instructor must have access to a website where the protein sequence needed for Exercise 1 can be posted. Students can then go to the site and simply copy the sequence. Alternatively, the instructor could email the sequence to all students before class.

Background on Bioinformatics and BLAST—One class period prior to doing the exercises, students should be given a very brief introduction to bioinformatics, explaining what bioinformatics is, the types of scientists who might use bioinformatics methods or tools, and how bioinformatics is relevant to biochemistry. Students should then be introduced to BLAST [10, 11], explaining that it allows the comparison of a protein or DNA sequence to a database of other protein or DNA sequences, respectively, and the output from a BLAST search should be shown and explained as an example. The class should be encouraged to ask questions and informed that they will be required to do their own BLAST searches and interpret the output in the next class period. The Powerpoint slides used to introduce bioinformatics, sequence alignment, and BLAST to the Biochemistry class at ISU are available for download from JKI's website (carbon.indstate.edu/inlow/bioinformatics/exercises.htm). Alternatively, one could make these slides available to students and require them to review the material on their own, rather than using class time.

Materials for Exercises—A detailed experimental procedure for each exercise, along with questions to be answered and turned in, should be prepared and a copy provided to each student. The exercises are outlined below with potential discussion questions included, but the detailed procedures with discussion questions that have been used at ISU are available from JKI's website (carbon.indstate.edu/inlow/bioinformatics/exercises.htm).

CLASSROOM/COMPUTER LABORATORY EXERCISES

Exercise 1: Protein of Unknown Function from Mus musculus

Purpose—Exercise 1 is based on hemoglobin, and it is appropriate after the structure and function of hemoglobin have been covered in the course. It is important that students are familiar with a few of the residues essential to the function of hemoglobin, such as the proximal and distal histidines, or residues that stabilize the T state by ion pairing. The exercise is intended to make students think about the protein sequence-structure–function relationship in the context of a single protein (hemoglobin) from a variety of species, and enables them to see that certain residues are essential to a protein's structure and function, while others can vary without affecting function. The exercise utilizes BLAST, so students become familiar with this tool and the types of tasks it can be used for, and they learn to interpret pairwise amino acid sequence alignments as they examine the BLAST output.

Procedure—

1. Students obtain the "sequence of a protein of unknown function from *Mus musculus*" in FASTA format by copying it from the course website. The sequence is actually the amino acid sequence of the β-chain of mouse hemoglobin (gi 31982300), but the accession number and protein name have been removed so that students do not know the identity of the protein. (This sequence can be obtained from JKI's website at carbon.indstate.edu/inlow/bioinformatics/exercises.htm or from the NCBI protein database at www.ncbi.nlm.nih.gov.)
2. Students are given detailed instructions for using the mouse protein to carry out a protein-protein BLAST search (blastp) against the *Homo sapiens* Refseq database at the NCBI website (www.ncbi.nlm.nih.gov/BLAST/), including a search against the Conserved Domain Database.
3. Students examine the BLAST results to determine which human protein is most similar in sequence to the mouse protein, and work together to answer the discussion questions:

   Is there a conserved domain in the mouse protein? If so, click on the domain and explore the links, then describe what you learned about this domain.
   Which human proteins are similar in sequence to the mouse protein?
   Speculate about the identity and function of the mouse protein.

4. Students are then given a handout showing a multiple sequence alignment of human α-globin, β-globin, and myoglobin with the locations of several residues important to the structure and function of hemoglobin highlighted. The multiple sequence alignment is available at JKI's website (carbon.indstate.edu/inlow/bioinformatics/exercises.htm); the highlighted residues are the proximal and distal histidines, a histidine that contributes to the Bohr effect, three residues involved in BPG binding, and three residues that stabilize the T state by forming ion pairs. Rather than making this multiple sequence alignment available, students could be instructed to use their textbook to determine the locations of important residues. Many of the biochemistry texts mentioned in the introduction provide a figure or sequence alignment in which at least some of these residues are labeled. Students are instructed to compare the multiple sequence alignment (or their textbooks) with their BLAST output and work together to answer the discussion questions:

   Does the mouse protein have all of the highlighted residues that you know to be important to the structure and/or function of human hemoglobin?
   What can you conclude about the importance of these residues to the structure and/or function of hemoglobin in both organisms?
   What might happen if one of these residues was replaced by some other residue as the result of a genetic mutation?
   Identify a few residues in the mouse protein that are different from those at the corresponding

position in the human globins. Why do you think these residues differ, and what can you conclude about the importance of these residues to the structure and function of this protein in both organisms?

5. Students are given detailed instructions for using the mouse protein to carry out a second protein-protein BLAST search, this time against the *Takifugu rubripes* (pufferfish) nr database at NCBI. They are asked to compare the BLAST results with those from the first BLAST search and answer the discussion questions:

   Is the mouse protein more similar to human β-globin or to pufferfish β-globin? Is this what you might have expected before doing the BLAST search? Why?
   Does pufferfish β-globin have all of the residues that are highlighted on the multiple sequence alignment handout, which you know to be important to the structure and/or function of human β-globin? If any of these residues differ between pufferfish β-globin and human β-globin, are the substitutions conservative?

6. Students are given detailed instructions for using the mouse protein to carry out a third protein-protein BLAST search, this time against the *Arabidopsis thaliana* nr database at NCBI. They are asked to examine the BLAST results, paying particular attention to the total lengths of the aligned proteins, and answer the discussion questions:

   Is it surprising that an Arabidopsis β-globin is not among the BLAST hits from this BLAST search? Why or why not?
   Name the Arabidopsis protein that is the top hit from the BLAST search. On the basis of its degree of similarity to the mouse protein, in terms of total protein sequence length and percent amino acid identity, do you think it is likely to have the same structure and function as the mouse protein? Explain your reasoning.

Extension—Exercise 1 could be extended if time permits. For instance, students could choose from a list of extensions provided by the instructor. Some possibilities are as follows: (1) Students could conduct BLAST searches against a variety of other species that are of interest to them and construct a phylogenetic tree using the β-globin sequences from these species. Phylogenetic trees can be created when using the ClustalW multiple sequence alignment tool from the EMBL-EBI Toolbox at www.ebi.ac.uk/clustalw/. (2) Instruct students to obtain the sequences of alpha, beta, gamma, delta, and epsilon globins from a single species (e.g. gi numbers 4504347, 4504349, 28302131, 4504351, and 4885393, respectively, for Homo sapiens), construct a multiple sequence alignment using ClustalW, and compare and contrast the sequences of these different globins. For instance, they could determine whether gamma, delta, and epsilon globins contain helix D like β-globin, or lack helix D like α-globin. (3) Students could obtain β-globin sequences from a variety of species, along with several sequences of mutant human β-globins (e.g. gi numbers 18418633 and 229965). They could then construct a multiple sequence alignment in order to think about the consequences of the mutations. They could be asked to determine whether or not a particular mutation results in a conservative amino acid substitution relative to normal human β-globin, and how the mutant residue compares to the residues at this position in β-globins from other species. (4) The exercise could also be expanded in a more open-ended manner by asking students to pose their own questions or hypotheses based on what they have learned so far. They could be given the URLs to sites such as the ExPASy Proteomics Server (www. expasy.ch/) or the EMBL-EBI Toolbox (www.ebi.ac.uk/Tools/), where they could find tools to pursue answers to their questions.

Exercise 2: Chymotrypsin—Active Site and Specificity

Purpose—This exercise allows students to compare chymotrypsin and trypsin, focusing on the active site residues as well as a residue of the specificity pocket, which is in part responsible for the differing specificities of these two serine proteases. The exercise is appropriate after students have studied the reaction mechanism of chymotrypsin in the course and know the identities of the three residues comprising the catalytic triad. They should also know the specificities of chymotrypsin (cleaves on C-terminal side of large aromatic residues) and trypsin (cleaves on C-terminal side of Lys and Arg residues). The exercise is intended to reinforce what students have learned about these differing specificities by showing them, at the amino acid sequence level, part of the explanation for the difference. It also serves to emphasize the fact that the three residues of the catalytic triad are essential to the function of the enzyme and are, therefore, conserved in various species. During the exercise students learn to search protein sequence databases and use pairwise and multiple sequence alignment software. By changing the alignment parameters they must think about what constitutes a "good" alignment.

Procedure—

1. Students are given detailed instructions for obtaining the amino acid sequences of bovine chymotrypsin (gi 576117) and trypsin (gi 60593450) in FASTA format from the NCBI protein database (www.ncbi.nlm.nih.gov/).
2. Students are given instructions for constructing a pairwise global alignment of the two sequences using the pairwise alignment tool available from the EMBL-EBI Toolbox (www.ebi.ac.uk/emboss/align/index.html). They are instructed to examine the alignment carefully to determine if it is a good alignment. (Using the default parameters does produce a good alignment.) At this step they should be focusing on the catalytic triad, making sure that the active site Ser, His, and Asp of chymotrypsin are aligned with those of trypsin. They are encour-

FIG. 1. **Multiple sequence alignment from Step 5 of Exercise 2, generated by students.** B_Try is bovine trypsin (gi 60593450), B_Chy is bovine chymotrypsin (gi 576117), and C_Chy is chymotrypsin of the biting midge, *Culicoides sonorensis* (gi 56199440). Residues of the catalytic triad are boxed in black; three residues lining the specificity pocket are boxed in gray.

aged to change the default parameters of the alignment tool and try again, comparing the results. Then they select the alignment they believe is best and work with other students to answer the discussion questions:

> What constitutes a "good" alignment?
> Why did you choose this particular alignment as the best relative to other alignments you generated?

3. Students are then informed that chymotrypsin residues 189, 190, and 228 are three of the residues lining the specificity pocket of the enzyme [12]. They are asked to name the residues at these positions for chymotrypsin and determine the corresponding residues for trypsin, based on their sequence alignment. (The residue numbers will be different for trypsin.) If the students have generated a good alignment, Ser189 of chymotrypsin will be aligned with an Asp of trypsin, Ser190 of chymotrypsin will be aligned with a Ser of trypsin, and Tyr228 of chymotrypsin will be aligned with a Tyr of trypsin. If the exercise is to be completed in a 50-min class period, students are then given a handout showing a molecular rendering of chymotrypsin, which includes a close-up view looking into the specificity pocket with residue 189 highlighted. The handout is available on JKI's website (carbon.indstate.edu/inlow/bioinformatics/exercises.htm). If time permits, the handout can be omitted and students can instead use Cn3D to view the three-dimensional structure of chymotrypsin as described below in the "Extensions" for Exercise 2. Discussion question:

   > Although other features of these two enzymes are essential for their differing specificities, the residues lining the specificity pocket make an important contribution to their specificities [12]. Residue 189 of chymotrypsin and the corresponding residue of trypsin are each at the "base" of the specificity pocket, as you can see on the handout. Reconcile the identity of this residue in chymotrypsin and trypsin with the differing specificities of the two enzymes.

4. Students are asked to choose a species of organism that is distantly related to *Bos taurus* and to obtain the sequence of chymotrypsin or trypsin from this organism in FASTA format by searching the NCBI protein database (www.ncbi.nlm.nih.gov). They are cautioned to read the sequence entry carefully and be certain that they have obtained a full-length sequence. Before continuing to step 5, they are asked to propose some hypotheses:

   > Predict whether or not the enzyme you chose will contain the residues forming the catalytic triad. Explain your reasoning.
   > Predict the identities of the three residues lining the specificity pocket which you examined in step 3 for bovine chymotrypsin and trypsin. Explain your reasoning.

5. Students are given detailed instructions for constructing a multiple sequence alignment for bovine chymotrypsin, bovine trypsin, and the serine protease they selected in step 4, using ClustalW from the EMBL-EBI Toolbox (www.ebi.ac.uk/clustalw/index.html). An example generated by a student group is shown in Fig. 1. Using the default parameters will usually generate a good alignment, but students are encouraged to examine their alignment and try changing the parameters if they have reason to believe this is necessary. Next they locate the catalytic triad and residues 189, 190, and 228 for bovine chymotrypsin and determine

the corresponding residues for the serine protease they chose. Finally, they are asked to return to the hypotheses they proposed in step 4:

> Were your hypotheses from step 4 correct? Name any residues that turned out to be different from your predictions and discuss the implications in terms of enzymatic activity and specificity.
>
> On the basis of amino acid sequence alone, do you expect the serine protease from the organism you chose to have chymotrypsin-like activity or trypsin-like activity? Justify your answer.
>
> Compare your multiple sequence alignment to that of two other groups who have chosen a serine protease from a different organism; focus on the catalytic triad and the three residues lining the specificity pocket. Were the identities of the three residues lining the specificity pocket for the serine proteases chosen by the other groups different from those of the serine protease you chose? Discuss any differences.
>
> Looking at your multiple sequence alignment, are there any regions that are highly conserved (many identical amino acids for all three sequences)? Are there any regions that are not well conserved (few identical amino acids for all three sequences)? Do any of these regions surround the residues of the catalytic triad or the three residues lining the specificity pocket? How might the locations of any highly conserved regions relate to the structure and function of the three enzymes?

Extension—If time permits, Exercise 2 could be extended. Potential extensions might include the following: (1) Students could search for the sequences of bovine elastase and thrombin, two other serine proteases. They could then generate a multiple sequence alignment and compare the active site and specificity pocket residues to those of chymotrypsin and trypsin and think about the implications of any differences. (2) Rather than giving students the handout in step 3, they could view the three-dimensional structures of chymotrypsin and trypsin using molecular visualization software such as Cn3D. Cn3D is a free helper application for viewing molecular structures, which can be quickly and easily downloaded from NCBI (www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml). An online tutorial is available, but the software is easy to use and many of its features are intuitive. Students should easily be able to open the structure files for bovine chymotrypsin (PDB code: 1GCD) and trypsin (PDB code: 1S0Q) and then highlight residues of interest to see their locations. (3) Students could obtain both a chymotrypsin and a trypsin sequence from a variety of species and use ClustalW to generate a chymotrypsin phylogenetic tree and a trypsin phylogenetic tree. (4) To learn more about what constitutes a "good" alignment, students can try different combinations of alignment parameters when using ClustalW to align bovine trypsin and chymotrypsin with the partial sequence of a distantly related serine protease. The trypsin-like protease from *Metarhizium anisopliae*, gi number 40804912, is one sequence that can be used. (5) It may also be desirable to take a less structured approach to this exercise and allow students to explore questions and hypotheses of their own using Cn3D and other tools available from the ExPASy Proteomics Server (www.expasy.ch/) or the EMBL-EBI Toolbox (www.ebi.ac.uk/Tools/).

ASSESSMENT AND CONCLUSIONS

The two bioinformatics exercises were designed for an upper-level undergraduate biochemistry course focused on structure, function, and analysis of biomolecules. They are intended to introduce students to some of the tools of bioinformatics, and also to use bioinformatics to reinforce the protein sequence-structure–function relationship in students' minds. The potential questions included with the procedures are a way to help guide and focus the students' efforts, and also serve as a concrete means for the instructor to assess what students have learned. The majority of students in the ISU Biochemistry course were able to answer all of the questions correctly, although many students did need some hints from the instructor in answering some of the questions.

In addition to exposing students to bioinformatics, these exercises increase their general problem-solving ability because they cannot simply look up the answers in a textbook. Rather, they must apply facts learned earlier in the course to a new problem in an unfamiliar context. We believe this is an approach many students enjoy, because each semester a number of students informed the instructor after class that they had enjoyed doing the exercises and learning about bioinformatics, and no students made negative comments about the exercises on their final course evaluations.

No problems that we consider significant were encountered with the exercises in any of the semesters they were used. Students did not have trouble understanding and following the instructions. However, despite the fact that the output from a BLAST search was shown and explained prior to Exercise 1, weaker sudents needed some initial help from the instructor in understanding the output from their first BLAST search in this exercise. Students who had some background in evolutionary biology or phylogenetics from life science courses seemed somewhat better able to grasp the significance of the sequence differences between mouse, human, and pufferfish hemoglobins than their classmates who had completed little life science coursework. Some students could not remember facts essential for the exercises, such as the three residues of the catalytic triad or the significance of the proximal and distal histidines to the structure of hemoglobin, but they were allowed to use their textbooks or notes to look up this information. It has been our experience that occasionally the NCBI web server is unavailable for brief periods of time. Although this did not happen while our students were working on the exercises, this problem is always a possibility. On the other hand, we have never experienced problems with the EMBL-EBI web server.

Since almost all of the students who typically enroll in ISU's Biochemistry course are unfamiliar with bioinformatics tools, and since the work must be completed in a 50-min class period, the exercises are purposely very structured, with detailed instructions and specific questions to answer. However, both exercises could be conducted in a more open-ended and unstructured way to promote creative thinking and resourcefulness. For instance, students could be assigned to learn how to use the bioinformatics tools on their own and allowed to pose and explore their own questions or hypotheses. This might be a more desirable format for courses with a small enrollment, or courses in which most students plan to attend graduate school in the biosciences.

REFERENCES

[1] http://www.ncbi.nlm.nih.gov/Genbank/index.html.
[2] http://www.rcsb.org/pdb/contentGrowthChart.do?content=total&seqid=100.
[3] J. M. Berg, J. L. Tymoczko, L. Stryer (2007) Biochemistry, 6th ed., Freeman, New York.
[4] R. H. Garret, C. M. Grisham (2005) Biochemistry, 3rd ed., Thompson Brooks/Cole, Belmont, CA.
[5] H. R. Horton, L. A. Moran, K. G. Scrimgeour, M. Perry, J. D. Rawn (2006) Principles of Biochemistry, 4th ed., Prentice Hall, Upper Saddle River, NJ.
[6] C. K. Mathews, K. E. van Holde, K. G. Ahern (2000) Biochemistry, 3rd ed., Addison Wesley Longman, San Francisco, CA.
[7] D. L. Nelson, M. M. Cox (2005) Lehninger Principles of Biochemistry, 4th ed., Freeman, New York.
[8] D. Voet, J. G. Voet (2004) Biochemistry, 3rd ed., Wiley, Hoboken, NJ.
[9] D. Voet, J. G. Voet, C. W. Pratt (2006) Fundamentals of Biochemistry Life at the Molecular Level, 2nd ed., Wiley, Hoboken, NJ.
[10] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman (1990) Basic local alignment search tool, *J. Mol. Biol.* **215**, 403–410.
[11] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res*. **25**, 3389–3402.
[12] J. J. Perona, C. S. Craik (1997) Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold, *J. Biol. Chem.* **272**, 29987–29990.