

# **Incorporation of Bioinformatics Exercises into the Undergraduate Biochemistry Curriculum**

Andrew L. Feig\* and Evelyn Jabri

Department of Chemistry, Indiana University, 800 E. Kirkwood Ave. Bloomington, IN  
47405

\* To whom correspondence should be addressed:

Andrew L. Feig  
Department of Chemistry  
Indiana University  
800 E. Kirkwood Ave.  
Bloomington, IN 47405  
U.S.A.  
Phone: 812-856-5449  
Fax: 812-855-8300  
email: afeig@indiana.edu  
web address: <http://www.chem.indiana.edu/personnel/faculty/feig/feig.htm>

**Submitted for publication to BAMBEd – November, 21, 2001**

**Keywords:** bioinformatics, data mining, molecular visualization

**Abbreviations:** BLAST – Basic Local Alignment Search Tool; EBI – European Bioinformatics Institute; ExPASy – Expert Protein Analysis System; KEGG – Kyoto Encyclopedia of Genes and Genomes; MSA – multiple sequence analysis; NCBI – National Center for Bioinformatics; OMIM – Online Mendelian Inheritance in Man; PBIL – Pole Bio-Informatique Lyonnais; PDB – Protein Data Bank; PE - Protein Explorer; URL – universal resource locator;

**Abstract**

The field of bioinformatics is developing faster than most biochemistry textbooks can adapt. Supplementing the undergraduate biochemistry curriculum with data mining exercises is an ideal way to expose the students to the common databases and tools that take advantage of this vast repository of biochemical information. An integrated collection of exercises based on “pet proteins” has been assembled. The exercises described are applicable to either a lecture or laboratory format and require only basic desktop computers, an Internet connection, a current web browser and the free Chime plug-in module. In an open-ended, inquiry-based format, the assignments ask students to explore concepts such as the relative information content of the different biopolymers, the relationship between primary sequence and tertiary structure, and how sequence conservation can be used to find an enzyme active site.

**Introduction**

The field of bioinformatics is rapidly changing the way biochemists conduct research [1]. No longer is it an arduous task to identify a gene. That which took years to do in the past often can be done in a matter of weeks or months using powerful computational algorithms and vast databases of genomic sequence information. One of the primary goals of bioinformatics as a field is to provide tools that allow scientists to find and correlate data from across disciplines – providing connectivity between different types of information (Figure 1). If one analyzes the traditional biochemistry curriculum, one finds that many of the data types that form the foundation of bioinformatics are already discussed in these courses, but in a segregated and less integrated manner. By incorporating bioinformatics into the undergraduate biochemistry curriculum, the interrelationships between these subjects become more evident to the students.

The bioinformatics databases, and the research tools necessary to access them, are used extensively by biochemists at all levels. It is, therefore, just as essential to teach students the tools of data mining, sequence analysis and molecular visualization as gel electrophoresis and enzyme kinetics. In addition to being powerful research tools, bioinformatics provides a terrific pedagogical opportunity to illustrate the enormous data content in even relatively short protein or nucleic acid sequences. For instance, students can readily discover the clustering of evolutionarily conserved sequences in the active site by performing a multiple sequence alignment (MSA). Since these tools are generally freely available via the Internet, their incorporation into the curriculum does not require a capital investment in the form of expensive software packages and simply requires the

instructor to provide some guidance and a web site with the appropriate hyperlinks. The exercises described below promote active inquiry with the curriculum and hopefully a better understanding of the underlying biochemistry. At the same time, the students become much more comfortable with the use of computers and the Internet as academic tools.

Due to the rapid nature of these changes, biochemistry textbooks have not kept pace with the developments in bioinformatics. It is, therefore, necessary to provide supplemental exercises that actively engage the students and allow them to explore the realm of computational biology within the biochemistry curriculum. A single lecture or discussion on bioinformatics and the genome projects at the beginning of the course generally provides sufficient context for the students to begin the exercises described herein. Other authors have highlighted the contents of the various databases used as part of these activities and they are not described in detail here [2-4]. Instead, the focus will be on a detailed discussion of how these tools can be integrated into the undergraduate biochemistry curriculum through an integrated series of problem set assignments. We have used these exercises in both one- and two-term introductory biochemistry lecture courses. A recent report by Scott Cooper describes the implementation of related exercises into the laboratory setting [5]. The URLs for the sites used during these exercises are listed in Table 1. Direct links should be provided on a course web page every term, however, to avoid problems associated with periodic changes of these addresses. It is also essential that suitable workstations are available for student use, especially in cases where home network connections are either slow or unreliable.

When used in the lecture course, the exercises are generally parsed out over the full semester. Figure 2 provides an overview of the exercises such that one major bioinformatics task is incorporated into each problem set (5 in total usually). In this way, the computational assignments parallel the course content as much as possible. To prevent students from focusing excessively on the use of the computer rather than the analysis of the output, each exercise begins with detailed instructions that lead the students through the manipulations and mouse clicks. These instructions also describe the expected output and common mistakes, including the error messages they might generate inadvertently. As the students become accustomed to working on these web sites, the instructions become less detailed and more task-oriented. In this way, students gain independence and self-confidence in their ability to find the necessary information on their own. Toward the end of the semester, an independent assignment (usually for extra credit) is provided based on a recent report found in the popular literature. This assignment simultaneously assess the student's ability to independently navigate these sites without being told where to obtain the necessary data and demonstrates how far their data-mining skills have advanced during the term.

### **Goals**

The goal of these exercises is to introduce junior/senior level undergraduates to some of the common computational tools used by biochemists. These tools are primarily Internet-based. They require basic Internet access and a web browser such as Netscape. One site used extensively for molecular visualization, Protein Explorer [6-8], is currently not compatible with Internet Explorer, although newer versions may fix this conflict. Also required is the plug-in module Chime, available free from MSD at their website

(Table 1). Upon completion of these exercises, the students should be able to carry out simple BLAST searches, multiple sequence alignments using CLUSTALW and basic handling of structural data including accessing, viewing and on-screen manipulation of macromolecules. The students also gain familiarity with common databases that cross-index many pertinent facts about proteins, enzymes, and nucleic acid sequences including links to tabulated enzyme kinetics data and metabolic maps. Links to structural classification and conserved domains families can be explored as well as the cross-references to metabolic diseases resulting from abnormalities in specific proteins.

### **Classroom/Laboratory Exercises**

#### **Getting started**

Each student receives an unknown protein at the start of the semester in the form of two peptide sequences (Table 2). The assumption is that the students have obtained a small amount of this protein that was subjected to mass spectrometric analysis after an enzymatic digestion yielding the observed sequence fragments. The protein unknowns were selected based on enzymes the students will encounter during the course and practically all of them can be found in the index of common undergraduate textbooks. Therefore, once the protein is identified, any student who wishes can easily read some background material on it. The proteins represented by the peptides in Table 2 have been predominantly taken from *E. coli*, the exceptions being some common proteins not found in prokaryotes, in which case human proteins have been used. For various reasons, the students will not always return with the *E. coli* form of the protein, but they invariably find the correct enzyme.

#### **Project 1: Identifying an Unknown Protein**

The first objective is to identify their unknown enzyme by performing a BLAST (Basic Local Alignment Search Tool) search against a collection of non-redundant protein databases [9, 10]. There are several portals available for these types of searches, but the NCBI web page is the resource we have used. This site was chosen because of the extensive documentation available if the students care to read more about the process. Instructions lead the students through the task of inputting a single peptide into the search. The students eventually perform five BLAST searches, one using each peptide alone and then searches using both peptides but varying the order in which the peptides are entered. Finally, the students also enter one of the peptides backwards (C → N) and attempt to find the parent protein. Successful identification of all unknowns can be obtained by using the default search settings. Students are asked to find the meaning of the expectation value in the documentation before they analyze the output from these searches. Searching against both peptide fragments provides E-values of  $\approx 5 \times 10^{-4}$  or better (lower) for each protein listed in Table 2. The students are pointed to the alignment portion of the output that can be observed if they scroll down their browser window for further comparison of their query with the retrieved sequences. Finally, many students need to be informed about database accession numbers and the enzyme classification system since these identifiers are the easiest way for them to proceed through the additional assignments on later problem sets.

Once the students have run the BLAST searches and identified their protein, they are asked to compare the output from the different searches. In several of the cases, searching on just one of the two tryptic fragments provides an error that no significant hits were obtained. Many of our students find the concept of probability confusing,

especially the idea that given a database the size of the human genome, one expects to find a perfect- or near-match for short random amino acid sequences. The students see clearly how adding additional sequence information improves the statistics of the sequence match. They learn about the length of sequence required to get unambiguous hits and the importance of gap penalties. It is extremely useful for the students to talk with one another about these issues, as the outcome varies depending on the length and complexity of the short peptide fragments assigned.

Several errors appear every time this exercise is assigned. The most common occurs when the student accidentally performs a BLASTn rather than a BLASTp search – the result being an error (Warning: Blast: No valid letters to be indexed) with no additional output. Whereas the newest version of the NCBI BLAST web page separates the nucleic acid and protein search links, students still manage to perform their searches incorrectly at times. The other source of confusion derives from multiple hits to what appear to be the same protein. Even non-redundant databases contain overlap due to multiple entries and partial gene sequences. The students must look at the pairwise alignments and make a judgment regarding the correct identity of the target protein. In certain cases, one of the fragments comes from a highly conserved region of the enzyme and thus appears to be identical to the protein from several organisms; the students must rely on the searches that employ both fragments to distinguish the most likely species from which their protein derived.

Listed below are examples of some open-ended questions that we use to promote student reflection on this exercise:

- What is the name of your unknown protein and from what organism was it isolated (if you can tell)?

- Why are there multiple hits in the output for what appears to be the same enzyme?
- When you entered your peptide in backwards, did you get a significant match to anything in the database? What does this result imply about the importance of directionality and orientation in biological macromolecules?
- Compare the results from the two searches that used both peptide fragments. Did you get identical results? Why or why not? What aspect(s) of the BLAST search parameters influence(s) these outputs?
- How might you use a BLAST search for something other than identifying a protein for which you have partial sequence information?

## **Project 2: Using the Expert Protein Analysis System (ExPASy) and on-line metabolic maps**

The second phase of the project involves finding additional information about the unknown enzyme now that it has been identified. If the goal were simply to obtain the full sequence of the protein, the students could use the direct links to the Genbank entries from the BLAST output to obtain this information. The ExPASy web pages have a lot of additional information on each protein and it is therefore useful to direct the students toward this resource. The students can find the enzyme on ExPASy by one of several methods. The formatted BLAST results contain the accession number for each sequence, as well as the enzyme classification number and the name of each hit. Any one of these pieces of information can be used to trace back to the full protein sequence. In the case of the E.C. number or enzyme name, the students arrive at the NiceZyme page and must then select an organism for which they wish to proceed. In this way, the student might identify a homolog of the actual protein. The NiceZyme display has a lot of useful information about these proteins and the students are asked to identify key features about their protein in addition to finding their peptides within the intact protein sequence. By

using the accession number, the students arrive directly at the NiceProt page that has more detailed and specific information on the exact protein. Students are asked to compare the types of information available on the NiceZyme and NiceProt pages since the links to certain types of information differ. The students are also given a laundry list of biochemical characteristics that the students must find through their investigation of these pages.

After exploring ExPASy, the students are directed to the on-line metabolic pathway maps. The links through ExPASy take the students to the traditional Boehringer-Mannheim metabolic map. While quite complete, this pathway map suffers from the daunting complexity of its paper counterpart, inciting a fearful response from the students. The set of hyper-linked pathway maps available from the Kyoto Encyclopedia of Genes and Genomes (KEGG) provide a much more pedagogically friendly environment. A caveat relevant to this exercise is that a few common proteins one might consider as potential unknowns do not appear on the metabolic pathway maps. For instance, lysozyme, trypsin and carboxypeptidase are all missing from the KEGG database. While these proteins otherwise make excellent unknowns for these assignments, they have been omitted from the Table 2 for this reason.

Once the student finds their protein within the metabolic pathways, they can also obtain extensive information about the enzyme on KEGG. The hyperlinks within the maps allow them to easily overlay the contents of other databases upon the metabolic pathway. In particular, we use the superposition of PDB and OMIM (On-line Mendelian Inheritance in Man) databases. The latter database catalogs inheritable diseases due to

inborn genetic errors [11]. In this way, students with interest in medicine can determine whether their unknown protein has been linked to a genetic disorder.

As part of this assignment, the students are asked to perform a treasure hunt and obtain various pieces of information about their enzyme. This project familiarizes them with the types of data collected in these entries so that the students can return to these sites later in the semester to find related information for future assignments. They are encouraged to explore beyond the realm of the exercises and experiment with the bioinformatics tools available through these sites.

### **Project 3: Finding, Manipulating and Understanding 3-D structure at PDB web site**

The unknown enzymes selected for inclusion as part of these projects have all been crystallographically characterized. This criterion was imposed in order to facilitate using the same set of enzymes for every step of the process. In certain instances, the structure was solved from a different organism, so the students must be made aware that the sequence may not match exactly with the one they identified as part of exercise 2.

The first phase of this project requires the student to use a tutorial to learn how to access the PDB, find a structure, and view it by using Protein Explorer (PE) and Chime [8]. For various reasons, we found that our students were reluctant to use the detailed PE tutorial available on the PE web site to learn the full gamut of structural manipulations available within the program. A shorter tutorial was therefore developed to familiarize them with the content of the PE windows and the embedded Chime menu commands. The first page of the tutorial includes mouse actions that zoom, rotate, and translate the molecule. The remainder of the tutorial walks the student through the structure of a small protein such as Protein G. Students prefer to start with a short polypeptide with simple

topology because they can “see” the secondary structure as well as specific amino acids without extensive manipulation of the structure. Initially, the students are guided through the structure with detailed instructions on how to select residues, color them, and display them in a various representations. Students then learn to obtain distance measurement and other data through manipulations of the structure with the mouse. Lastly, students are made aware of command syntax required to manipulate structures containing multiple polypeptide chains and non-proteinaceous ligands.

Once the students have become familiar with the set of simple commands, they proceed to an analysis of their unknown protein. They are instructed to find the structure of their protein in the PDB using the Search-lite option and the name of the protein. Since the PDB search sometimes yields a large number of structures, the student should be advised to choose a structure with a bound substrate or inhibitor, if available, as such structures facilitates the localization of the active site. They are also advised to pick the structure with the most complete polypeptide sequence rather than a small fragment. Once the structure has been obtained, the student must view the model in space-filling and cartoon modes, and provide a print out each model after some minor manipulations. The latter exercise serves two purposes; it shows the grader that the student has a) found the correct model, and b) learned to use PE to manipulate the on-screen image of the structure. Most students are able to complete these tasks with little difficulty.

Although these exercises allow assessment of the student’s mastery of PE, they do not provide information on the student’s ability to use their acquired structure manipulation skills to understand macromolecular models. These skills are assessed with a writing assignment in which the student describes the juxtaposition of secondary

structure elements in their assigned protein as well as any other structural features they find interesting. An example description is provided on the class web page to help students formulate their own paragraph. In addition to the prose description, they are instructed to include a sketch of the topology and a detailed figure generated in PE that illustrates one of their points, such as a close-up view of a ligand binding interaction.

Most of our students find this task extremely challenging. Introductory biochemistry texts often contain fabulous graphics of protein structures these days, but the students often see these images simply as pretty pictures. They make little attempt to understand these structures at a deeper level. Hence, when confronted with their protein, they have a hard time selecting a place to start their investigation. Furthermore, they cannot deconstruct the architecture. Getting students to let go of their bias toward organizing their discussion based on the primary structure is particularly difficult. The process of transferring their visual cues into a coherent description, however, teaches them how to examine and understand the complexities of biomolecular structures.

To help the students improve their scientific writing skills, the short descriptions are evaluated and returned with extensive comments. The students are then encouraged to resubmit their paragraphs after making the appropriate revisions. For many students, the initial exercise of writing about their unknown protein is insufficient to raise them to a deeper level of understanding regarding protein structure. Our experiences indicate that the revision process actually leads to the most significant conceptual development.

#### **Project 4: Multiple Sequence Alignments and Sequence Conservation**

A number of current biochemistry textbooks incorporate multiple sequence alignments (MSA) and phylogenetic trees to illustrate certain simple evolutionary

concepts. It is not always clear to the students, however, that these types of sequence comparisons can be performed on practically any protein and used to discern a lot more than the evolutionary date at which two species diverged from one another. The MSA can be used for a variety of biochemical pursuits. For instance, one can find relationships between seemingly disparate enzymes or show how a seemingly random pattern of primary sequence conservation clusters in three-dimensional space in the active site or key regulatory sites of enzymes. This assignment lets the students discover this clustering of conserved residues in three dimensions.

The students are asked to align the sequences of their unknown protein from five or six different organisms. They obtain fasta-formatted sequences from the ExPASy site via the NiceZyme page they have visited previously. Included among the sequences they collect should be the sequence from the structure they found in the PDB. As the sequences are gathered, they can copy them into a Microsoft Word document or other text editor so that the file can be transferred electronically and not retyped for submission into ClustalW. The student can then access the ClustalW program [12] at any number of different web sites, including the EBI or PBIL. For the purposes of this exercise, default alignment parameters have been quite successful. The qualities of the alignments depend on the organisms chosen by the student. With the alignments in hand, the students are then asked to go back to the PDB and create a figure that superimposes the alignment on the structure. For the 3-D alignment, a 10-15% sequence identity seems to be optimal and the students are instructed to add sequences to their alignment if the collected proteins are too closely related. The structural alignment can be done manually, but is substantially easier to accomplish with the new MSA3D feature in PE. An example of the output from

this exercise performed on the enzyme phenylalanine hydroxylase is shown in Figure 3 [13]. In most cases, the exercise quite dramatically shows a 3-dimensional clustering of conservation around the active site of the enzyme. The students then must describe the relationship between the two forms of MSA output – the primary sequence alignment file and the structural overlay.

Some of the questions that can be asked to make the students think about the implications of their findings include:

- When you look at the primary sequence and the amino acid conservation across organisms, do you see any patterns or organization?
- Is there a relationship between the sequence conservation and the overall 3-D structure?
- How would you use MSA on a nucleic acid such as a tRNA or rRNA and what information might you obtain from such an exercise? Will you always look simply for conservation or are there other factors that might come into play?

### **Project 5: Enzyme activity comparisons and the BRENDA database**

The final project involves analysis of kinetic data available for their unknown enzyme. These data are compiled in the BRENDA database, allowing rapid access to information on the various homologs. Here, the students must explore the reasons why kinetic data for an enzyme varies from organism to organism. Students use these data as a springboard through which they can discover the patterns evident in sequence conservation and the effect that mutations have on catalytic activity. Furthermore, they can explore the impact that experimental conditions such as pH and temperature have on a specific enzyme's activity as they learn about how enzyme kinetics is used to decipher catalytic mechanisms.

In addition to data on the natural substrate, the database also contains entries for non-natural substrates that have been tested and various inhibitors. Students can compare the structures the inhibitors to the natural substrates and learn about the use of transition state analogs as reversible inhibitors and suicide substrates as their irreversible cousins. The project culminates in the use of this information, together with their textbook and other literature sources, to propose a mechanism for the enzyme catalyzed transformation in as much detail as possible.

### **Learning Assessment**

Several levels of assessment must be performed to determine the efficacy of these exercises within the context of any biochemistry course. The instructions are sufficiently detailed that simply obtaining the printed output from the on-line exercise does not necessarily show understanding. Therefore, problem set questions ask the students to interpret their results as well. These questions require the students to make judgments regarding the significance of the output and to relate the findings to other topics in the curriculum.

A broader form of assessment is used toward the end of the semester. By following the popular literature every term, a recent example from a local or national newspaper story is identified and the students are asked to find information about the underlying gene/protein by using the tools they have learned during these exercises. Here, with only minimal direction, they are asked to find specific pieces of biochemical information. This assessment tool mimics the application of these tasks to a personal or research problem such as might be encountered beyond the realm of the classroom. A recent example derives from a report on the link between the ACC2 gene and obesity in

mice, taken from a local newspaper [14], but based on a paper published in Science [15]. The students were required to find the full sequence of the gene, the crystal structure and perform a sequence alignment between ACC2 and ACC1, a related gene disruption of which is fatal. The students were also informally polled regarding how long it took to locate the necessary information in the databases.

Skill retention was also assessed in subsequent courses, such as the physical biochemistry course (C481), taken by the biochemistry majors. Approximately 50 percent students in C481 during spring, 2001 had participated in these assignments during the fall term of 2000 where as the other half had taken a more traditional biochemistry course at Indiana University or elsewhere. During the first week of physical biochemistry course, when the instructor reviewed macromolecular structure, students with previous bioinformatics and molecular visualization experience immediately accessed the appropriate web pages and used these resources as a study aide. Students without this background were significantly more reluctant to do so, even though they were provided with the identical PE and Chime tutorial used in the previous semester's biochemistry course. While anecdotal in nature, we feel quite strongly that exposure to these resources will change the perception of our students toward computer-assisted-learning of biochemistry.

## **Conclusions**

On-line data mining tools provide students ready access to powerful tools for computational biology and biochemistry. The exercises described above can be incorporated into the biochemistry curriculum to provide students with practice performing complex computational tasks. The hands-on experience with these problems

allows the students to explore a rapidly changing area of biochemistry. The conceptual tools involved in these exercises teach data analysis and critical thinking skills. These exercises also bring to life the impact that genomics and proteomics is having on the field of biochemistry. The ability to find, parse and evaluate information from these databases will be essential for their continuing education in the field and are particularly valuable tools for students who plan to pursue careers in the laboratory sciences. The long term ramifications of these curricular changes will, of course, require more systematic analysis of student learning and the perception students hold regarding the role technology should play in this process.

**Acknowledgements.** We would like to thank Dr. Steven Wietstock in the Chemistry Department Instructional Support Office for assistance maintaining the web sites for the courses and Tim O’Dea, Tara Lorenz, Peter Mikulecky, Cheri Stowell, and Jaime Vaeth who have assisted in the classes as associate instructors during the development of these exercises.

**References**

- [1] C. Gibas, P. Jambeck, *Developing Bioinformatics Computer Skills*, O'Reilly, Beijing, 2001.
- [2] A. Bairoch, R. Apweiler, *Nucleic Acids Research* 28(1) (2000) 45-48.
- [3] C.E. Sansom, C.A. Smith, *Biochemical Education* 28(3) (2000) 142-149.
- [4] D.L. Wheeler, D.M. Church, A.E. Lash, D.D. Leipe, T.L. Madden, J.U. Pontius, G.D. Schuler, L.M. Schriml, T.A. Tatusova, L. Wagner, B.A. Rapp, *Nucleic Acids Res* 29(1) (2001) 11-16.
- [5] S. Cooper, *Biochemistry and Molecular Biology Education* 29(4) (2001) 167-168.
- [6] E. Martz, *Trends in Biochemical Sciences* (2002) in press.
- [7] E. Martz in S.A. Krawetz, D.D. Womble (Eds.) *Introduction to Bioinformatics*, Humana Press, Totowa NJ, 2002, pp. in press.
- [8] Protein Explorer Software, E. Martz (2001) <http://proteinexplorer.org>.
- [9] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, *J Mol Biol* 215(3) (1990) 403-410.
- [10] S.F. Altschul, W. Gish, *Methods Enzymology* 266((1996) 460-480.
- [11] Online Mendelian Inheritance in Man, OMIM™. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000. World Wide Web URL: <http://www.ncbi.nlm.nih.gov/omim/>.
- [12] J.D. Thompson, D.G. Higgins, T.J. Gibson, *Nucleic Acids Research* 22(22) (1994) 4673-4680.

- [13] H. Erlandsen, F. Fusetti, A. Martinez, E. Hough, T. Flatmark, R.C. Stevens,  
Nature Structural Biology 4(12) (1997) 995-1000.
- [14] Scripps Howard News Service, Herald Times, Bloomington, IN, March 30, 2001.
- [15] L. Abu-Elheiga, M.M. Matzuk, K.A. Abo-Hashema, S.J. Wakil, Science  
291(5513) (2001) 2613-2616.

**Table 1. Internet addresses (URLs) for the primary web sites used in these exercises.**

<b>Course web sites that use these exercises</b>	
C483 (IU)	<a href="http://chemlearn.chem.indiana.edu/c483/">http://chemlearn.chem.indiana.edu/c483/</a>
C484 (IU)	<a href="http://chemlearn.chem.indiana.edu/c484/">http://chemlearn.chem.indiana.edu/c484/</a>
<b>Bioinformatics sites important for these exercises</b>	
Blast:	<a href="http://www.ncbi.nlm.nih.gov/blast/">http://www.ncbi.nlm.nih.gov/blast/</a>
BRENDA:	<a href="http://www.brenda.uni-koeln.de/">http://www.brenda.uni-koeln.de/</a>
ExPASy:	<a href="http://www.expasy.ch/">http://www.expasy.ch/</a>
IUBMB	
Nomenclature:	<a href="http://www.chem.qmw.ac.uk/iubmb/enzyme/">http://www.chem.qmw.ac.uk/iubmb/enzyme/</a>
KEGG:	<a href="http://www.genome.ad.jp/kegg/kegg2.html">http://www.genome.ad.jp/kegg/kegg2.html</a>
MSA:	<a href="http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html">http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html</a>
PDB:	<a href="http://www.rcsb.org/pdb/">http://www.rcsb.org/pdb/</a>
Protein	
Explorer:	<a href="http://proteinexplorer.org/">http://proteinexplorer.org/</a>
PubMed:	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed</a>
<b>Download free software for use in these exercises</b>	
CHIME:	<a href="http://www.mdli.com/cgi/dynamic/product.html?uid=\$uid&amp;key=\$key&amp;id=6">http://www.mdli.com/cgi/dynamic/product.html?uid=\$uid&amp;key=\$key&amp;id=6</a>
CN3D:	<a href="http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dinstall.shtml">http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dinstall.shtml</a>
Rasmol:	<a href="http://www.bernstein-plus-sons.com/software/rasmol/">http://www.bernstein-plus-sons.com/software/rasmol/</a>
Shockwave:	<a href="http://sdc.shockwave.com/shockwave/download/frameset.fhtml?">http://sdc.shockwave.com/shockwave/download/frameset.fhtml?</a>
VRML:	<a href="http://www.karmanaut.com/cosmo/player/">http://www.karmanaut.com/cosmo/player/</a>

**Table 2. Tryptic fragments of common metabolic enzymes for use as unknowns.**

#	peptide 1	peptide 2
1	DVQFGLATMCIGLGQGIATVFER	INLNGGAIALGHPLGCSGAR
2	ATIANMSPEYGATCGFFPID AVTLDYMR	WQDGNVTEEDIHALAGWLK
3	DRPMAVNGQVEILPMMYLALSVDHR	QQASLEEQNNDALSPAIR
4	NIFSVIYDGTQPNPTTENVAAGLK	AVTMIAENLPLAVEDGSSNAK
5	AGSVFVNNYNDGDMTVPFGGYK	VYGEVATTSSHELAMIVR
6	HNLPHNSLNFVVFHGGSGSTAQEIK	ENNFALPAVNCVGTDSINAVLETAAK
7	MDPPCTNTTAASTYLNNPYVR	FYATNDTEVAQSNFEALQDFFR
8	TVGWIAHWSEMHSDBGMK	AMGIPSSMFTVIFAMAR
9	FQTAFAQLADNLQSALEPILADK	ADAPLIQWDATSATLK
10	AYFTSATMIIAIPITGVK	DIGTLYLLFGAWAGVLGTALSLLIR
11	SGETEDATIADLAVGTAAGQIK	GMPLYEHIAELNGTPGK
12	QHEFSHATGELTALLSAIK	LLYECNPMAFLAEQAGGK
13	MDSYVDQLQAQGGSMIMLAK	YYDELPTEGNEHGQAFR
14	QGLQLGFSFSFPCHQTGLDR	QGLSGQSLPLGFTFSFPCR
15	NHLNMHFVSNVDGTHIAEVLK	AVGEFGIDTANMFEFWDWVGR
16	EIPQVAGSLEEALNELDLDR	EQHVTIPAHQVNAEFFEFGK
17	YAGQDIVSNASCTTNCLAPLAK	WDEVGVDVVAEATGLFLTDETR
18	VDGGAVANNFLMQFQSDILGTR	IPISGIAGDQQAALFGQLCVK
19	STQVYQDQVWLPAETLDLIR	FPEHCGIGIKPCSEEGTK
20	GYTSWAIGLSVADLAESIMK	ITVVGAVGMACAISILMK
21	TPGAVNACHLSCSALLQDNI ADAVACAK	GISLANWMCLAK
22	VAVLGAAGGIGQALALLLK	ACIGIITNPVNTTVAIAAEVLK
23	ATLLIETLPAVFQMDEILHALR	EQDAPITADQLLAPCDGER
24	AFSQFLNLANAEQYHSISPK	AVESLSLELVLTAHPTEITR
25	FANQILSYGAELDADHPGFK	ILADSINSEIGILCSALQK
26	HSFNPHILAIQAIAEER	AGLTLGVDPLGGSGIEYWK
27	IAAVAEDGEPCVITYIGADGA GHYVK	AGAGTDAIDSLKPYLK
28	ISYISTGGGAFLEFVEGK	YAALCDVFVMDAFGTAHR
29	EEGYSFDFAYTSVLK	AIHTLWNVLDELQAWLPVEK

30	EQGLNSENFVAFNLTER	GVLTNLGAVAVDTGIFTGR
31	DLPEDIYVVIEIPANADPIK	FMSTAMFYPCNYGYINHTLS LDGDPVDVLVPTPYPLQPGSVIR
32	VILFIGDGSLQLTVQEISTMIR	MIEVMLPVFDAPQNLVEQAK
33	DVNVPDIGSDEVEVTEILVK	EVNVPDIGGDEVEVTEVMVK
34	QDLIFGCEQGVDFVAASFIR	GDLGVEIPVEEVIFAQK
35	TGFHMLHTLFTSLQFPQIQR	AATAGNGNEAAIEAQAAGVEQR

---

**Figure Legends**

**Figure 1.** Overlap between bioinformatics and the topics commonly taught in introductory biochemistry. Arrows indicate the cross-referencing of information common in the bioinformatics databases. Topics on dark backgrounds are those already covered in the majority of introductory biochemistry courses and textbooks. (Adapted from reference 1.)

**Figure 2.** Schematic overview of the bioinformatics exercises and their division into 5 general projects performed over the course of the semester.

**Figure 3. A.** A portion of the multiple sequence alignment used for the analysis of phenylalanine hydroxylase generated by the program ClustalW. Swiss-Prot accession numbers for the sequences used in this exercise were: P00439 (*Homo sapiens* PAH), P04176 (*Rattus norvegicus* PAH), P16331 (*Mus musculus* PAH), P30967 (*Chromobacterium violaceum* PAH), P43334 (*Pseudomonas aeruginosa* PAH), P17276 (*Drosophila melanogaster* PAH), P90925 (*Caenorhabditis elegans* PAH) and 1PAH, the sequence from the crystallographic solved PAH fragment from *Rattus Norvegicus* [13].

**B.** Superposition of the sequence conservation on the structure of phenylalanine hydroxylase based on the ClustalW alignment shown. **C.** Overall statistics of the ClustalW alignment.