Article

An Inquiry into Protein Structure and Genetic Disease: Introducing Undergraduates to Bioinformatics in a Large Introductory Course

April E. Bednarski, Sarah C.R. Elgin, and Himadri B. Pakrasi

Campus Box 1137, Department of Biology, Washington University, St. Louis, MO 63130

This inquiry-based lab is designed around genetic diseases with a focus on protein structure and function. To allow students to work on their own investigatory projects, 10 projects on 10 different proteins were developed. Students are grouped in sections of 20 and work in pairs on each of the projects. To begin their investigation, students are given a cDNA sequence that translates into a human protein with a single mutation. Each case results in a genetic disease that has been studied and recorded in the Online Mendelian Inheritance in Man (OMIM) database. Students use bioinformatics tools to investigate their proteins and form a hypothesis for the effect of the mutation on protein function. They are also asked to predict the impact of the mutation on human physiology and present their findings in the form of an oral report. Over five laboratory sessions, students use tools on the National Center for Biotechnology Information (NCBI) Web site (BLAST, LocusLink, OMIM, GenBank, and PubMed) as well as ExPasy, Protein Data Bank, ClustalW, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, and the structure-viewing program DeepView. Assessment results showed that students gained an understanding of the Web-based databases and tools and enjoyed the investigatory nature of the lab.

Keywords: undergraduate, bioinformatics, protein structure, genetic disease

INTRODUCTION

As computers are used more and more as tools in research laboratories, it becomes increasingly important to introduce undergraduates to bioinformatics tools and databases early in their study of biology. As stated in the National Research Council's (NRC) *Bio2010* report:

> Computer use is a fact of life for all modern life scientists. Exposure during the early years of their undergraduate careers will help life science students use current computer methods and learn how to exploit emerging computer technologies as they arise... Becoming fully conversant with databases such as the National Center for Biotechnology Information (NCBI) is important for all biology majors (NRC, 2002, 47).

Incorporating bioinformatics into undergraduate biology curriculum has been a focus of several innovative projects that have recently been described in the literature (Campbell,

2003; Centeno *et al.,* 2003; Cooper, 2001; Feig and Jabri, 2002; Honts, 2003; Jungck and Donovan, 2000). In order to reach all biology majors, the curriculum project described here introduces bioinformatics tools and databases in a broad manner in a laboratory that accompanies a large introductory biology course. Students are then offered the option to use these tools more extensively in several smaller, upper-level courses.

This curriculum was designed to incorporate elements of inquiry, while still being compatible with a large course size and limited lab time. The lab described here requires five weekly sessions of 2 h each. Our goal was to provide a laboratory experience in which students approach problems, seek information, synthesize findings, and share results as do active scientists (NRC, 1999, 2000). In particular, we wanted to incorporate evidence-based hypothesis testing, use of primary literature sources, and communication of research results in oral and written reports.

To help students develop a sense of ownership toward their work, individualized projects were developed. The projects provided an opportunity for students to work semi-independently on their own Web-based research. The projects emphasize the link between gene sequence, protein

Table 1. Bioinformatics tools used in the curriculum

Name	Web address
NCBI	http://www.ncbi.nlm.nih.gov/
PubMed (National Library of Medicine, 2004)	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
LocusLink (Pruitt and Maglott, 2001)	http://www.ncbi.nlm.nih.gov/LocusLink/
OMIM (NCBI, 2000)	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Psi-Phi BLAST (Altschul *et al.*, 1997)	http://www.ncbi.nlm.nih.gov/BLAST/
ClustalW (Thompson *et al.*, 1994)	http://www.ebi.ac.uk/clustalw/index.html
KEGG (Kanehisa, 1997; Kanehisa and Goto, 2000)	http://www.genome.ad.jp/kegg/
ExPASy	http://us.expasy.org/
DeepView (Guex and Peitsch, 1997)	http://us.expasy.org/spdbv/
SwissProt (Boeckmann *et al.*, 2003)	http://us.expasy.org/sprot/
Protein Data Bank (Berman *et al.*, 2000)	http://www.rcsb.org/pdb/
Sequence Manipulation Suite (Stothard, 2000)	http://bioinformatics.org/sms/
PSIPRED (McGuffin *et al.*, 2000), MEMSTAT (Jones, 1999)	http://bioinf.cs.ucl.ac.uk/psipred/

sequence, and gene function in genetic disease. Each project was developed around a single protein with an amino acid substitution that has been linked to a genetic disease. Students begin by reading literature related to their project, then learn more about their protein using NCBI's database, LocusLink (Pruitt and Maglott, 2001), and ExPASy's database, SwissProt (Boeckmann *et al.*, 2003). Students are then guided through a BLAST search to identify homologous proteins, a ClustalW alignment to align the homologous proteins and identify the mutation, and secondary structure prediction of their protein sequence using PSIPRED. Students perform a structure analysis of their protein and model the mutation using the program DeepView, available through the ExPASy Web site. As a result of this analysis, students are asked to predict the effect of the mutation on protein structure and function. Finally, students use the OMIM and KEGG databases to relate their protein's function to human physiology and disease states. A summary of the tools and databases used in the course is provided in Table 1. On the last day of lab, students report their results and defend their predictions in small peer groups. Students within a lab section of 20 are encouraged to serve as support for each other, facilitating group interactions within the framework of a large lecture course.

In designing the individual projects (summarized in Table 2), each protein of focus had a single amino acid substitution that had been linked to a human disease, documented in the OMIM database, and had a published crystal structure available of either the exact protein or a close homolog. Ten projects were developed that met these criteria, and we plan to add more in the future. The materials developed for this course are publicly available online (Bednarski *et al.*, 2004), and we hope they will be useful to others in designing bioinformatics labs.

Context for the Course

The introductory biology core at Washington University (WU) in St. Louis is a three-semester curriculum, Principles of Biology I, II, and III. In the first semester, biological macromolecules (in particular DNA, RNA, and proteins) are introduced in the context of cell biology and microbial genetics. In the second semester, students explore eukaryotic genetics, chromosomes, population genetics, and natural selection. In the third semester, Principles of Biology III, students study protein structure/function, metabolic pro-

cesses, and human physiology. Principles of Biology I and II are both taught with a weekly 3-h wet lab where students perform experiments in molecular biology and genetics.

Students are required to have taken at least one semester of General Chemistry before enrolling in Principles of Biology I. By the time students are taking Principles of Biology III, they have taken Organic Chemistry I, and they are usually taking Organic Chemistry II concurrently. This allows biochemical principles to be introduced at a fairly advanced level in Principles of Biology III, and the textbook for the biochemistry topics in the course is Berg, Tymoczko, and Stryer's *Biochemistry* (2002).

The bioinformatics lab described here accompanies the third-semester course, Principles of Biology III. Care has been taken to align the laboratory projects with the proteins and metabolic processes that are discussed in the lecture portion of the course. Lectures for Principles of Biology III include topics relevant to bioinformatics such as protein evolution, multiple sequence alignments, and BLAST algorithms.

LABORATORY DESIGN

Organization

This bioinformatics lab was taught for the first time in Spring 2004 with an enrollment of 246 students. The majority of the students in the course were biology majors, but there were also a significant number of students from a wide variety of other majors, including physics and biomedical engineering (see Figure 1). The students were divided into laboratory sections of 15–20 students, and each section met weekly in a computer lab for a total of five 2-h sessions. The computer lab was equipped with 20 iMac computers (OS 10.x), a projector connected to an instructor's computer, a laser printer, and Ethernet connections for each computer. The students' desks were arranged around the wall on three sides of the classroom, so that the instructor could easily see the students' computer screens while standing in the center of the room (see Figure 2). This arrangement allowed the instructor to quickly assess the progress of the students and their focus on the individual tasks. A laboratory instructor guided each section with help from a graduate teaching assistant. In order to enhance the teaching experience for the graduate students, they had the opportunity to take the lead instructor role in 1–2 sections per week with the instructor

Table 2. Projects

Project	Gene symbol	Mutation	Associated disease (OMIM #)	Protein OMIM #	Pdb ID
Lung cancer					
K-Ras 2	KRAS2	G12D	Lung cancer (211980)	190070	1AGP (Milburn *et al.*, 1990)
Cytochrome P450 1A1	CYP1A1	I462V	Lung cancer (211980)	108330	1OG5 (Williams *et al.*, 2003)
Apoptosis					
Superoxide dismutase–1	SOD1	H46R	Lou Gehrig's disease (105400)	147450	1B4L (Hart *et al.*, 1999)
Caspase–1	CASP1	H237D	Huntington's chorea (143100)	147678	1IBC (Rano *et al.*, 1997)
Metabolic disease					
Phenylalanine hydroxylase	PAH	E280K	Phenylketonuria (261600)	261600	1KW0 (Andersen *et al.*, 2002)
Hypoxanthine-guanine phosphoribosyltransferase	HPRT1	D193N	Lesch-Nyhan syndrome (300322)	308000	1BZY (Shi *et al.*, 1999)
Cholesterol homeostasis					
HMG-CoA reductase	HMGCR	D690A	Target for statin drugs	142910	1HW9 (Istvan and Deisenhofer, 2001)
Low-density lipoprotein receptor	LDLR	W165S (144)	Familial hypercholesterolemia (143890)	606945	1N7D (Rudenko *et al.*, 2002)
Mitochondrial disease					
Cytochrome c oxidase	MTCO1	Term to KEK(513 to 516)	Lebers hereditary optic neuropathy (53500)	516030	1OCC (Yoshikawa *et al.*, 1998)
ATP synthase-6	MTATP6	L156R	Leigh syndrome (256000)	516060	1C17 (Rastogi and Girvin, 1999)

serving as an assistant. Students were also encouraged to direct questions to students at nearby computers, since many students had the same type of questions on navigating a particular Web page. This approach helped minimize frustrations over simple Web-page navigation problems.

The 10 projects that were developed are summarized in Table 2. Since there were approximately 20 students per section, generally two students worked on the same project in each section. These students were encouraged to sit near each other and help each other as they worked. Students often developed different hypotheses for the same project, and this collaboration gave them the opportunity to discuss and defend their ideas in preparation for their final report.

Projects

The 10 projects were grouped with two projects per disease category (as shown in Table 2). These disease categories were used in forming small groups on the last day of lab for students to present their oral reports to each other. These groupings worked well, because projects in the same disease category required similar background reading, and these readings provided some common ground for the presentations. As previously described, each project was formulated using a gene that was known to have a single missense mutation, resulting in a changed amino acid in the protein sequence linked to a disease state in the literature (OMIM database). This model of genetic disease is not generally applicable (since most genetic problems are multifactorial), but it lends itself well to focusing on the role of protein structure in gene function and in providing a format for

using the Web-based tools and databases. Students were introduced to more complex models of genetic disease, such as that applicable to atherosclerosis, during the lectures in population genetics given in Principles of Biology II. A review of the structural basis of inherited disease was recently published and contains several additional examples for projects that could be developed for this laboratory (Steward *et al.*, 2003).

Web Site and Materials

A Web site has been designed to provide a source of materials for the laboratory as well as to provide links to the Web addresses that the students commonly use in the lab. All the lab materials are available to download from the Web site including a general laboratory manual, which includes the syllabus, a glossary, and tutorials on the programs and databases common to all of the projects. The glossary is designed to include the content and terminology needed in the lab, regardless of the specific project.

The Web site also contains a separate page for each of the 10 projects where the students can download their project manual, starting DNA sequences, and any articles they are assigned to read. The project manuals contain a short introduction, reading assignments, and guide sheets specific to each project. The reading assignments include short sections from the textbook (Berg *et al.*, 2002), a review article about the disease their protein is related to, and an excerpt from the article that described the crystal structure they will be analyzing. The guide sheets contain instructions and questions specific to each project and the database or
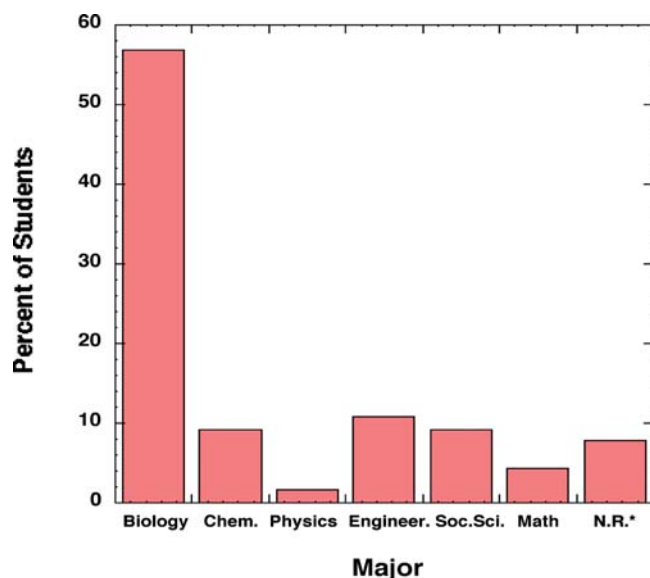
Figure 1. Distribution of student majors in the Bioinformatics lab Spring 2004. *No response.

program in use. These guide sheets are designed to be very detailed, so that students can easily use the databases and tools with minimum frustration. Although this approach does not encourage a lot of trial and error, it allows students to work at their own pace and to approach the work without anxiety.

In summary, the Web site gives students access to all the lab materials from their home computers as well as other campus computers. While lab time is usually sufficient, students can finish their work outside of class if needed.

IMPLEMENTATION

Lecture

The lecture component of the course presents the key topics needed for understanding the explorations in the laboratory. For example, the topics of protein evolution, multiple sequence alignments, and the BLOSUM-62 substitution matrix (Henikoff and Henikoff, 1992) are presented in lecture before the students work with BLAST (Altschul *et al.*, 1997) and ClustalW in the laboratory. The students have lectures on protein structure, including an introduction to the methods of solving protein structures, before working with the crystal structure data in the laboratory. The lecture also includes discussion of protein function, basic metabolism, enzyme regulation, and cellular energetics that are important in helping the students understand the effects of loss of function of the protein they study in the laboratory, and how that could play a role in a disease. The students are also asked to memorize the amino acid side chains (chemical structure and three- and one-letter codes) during the first week of class. This greatly facilitates working with the multiple sequence alignments and the crystal structure data in the laboratory. Requiring students to be able to use this information immediately in the lab helps reinforce the need to be able to recall the chemical nature of the amino acid side chains based on the one- and three-letter codes.

In the laboratory component of the course, each session begins with a brief review of the topics from lecture that are important for the focus of that day's activities and a demonstration lasting approximately 30 min to show the basics about the Web sites and programs the students are using that day. During the rest of the laboratory session, the students work at their own speed using the guide sheets, with the instructor, the teaching assistant, and other students available for assistance.

Laboratory

In the laboratory, students use a variety of bioinformatics tools and databases. This curriculum emphasizes availability and familiarity of resources, which is consistent with our first goal. To begin each project, students download a cDNA sequence from the lab Web site for a human gene containing a single nonsynonymous mutation. Through their guide sheets, students are instructed to first translate their sequence using a tool on the Sequence Manipulation Suite. Next, students perform a BLAST search (Altschul *et al.*, 1997) with the wild-type human protein sequence to obtain a group of diverse, yet homologous, sequences. Students are instructed to select five sequences, each from a different organism, from the BLAST results. They are asked to include sequences with a range of E-values, so that the selected sequences will be similar, but not almost identical, to their search sequence. To accomplish this, students select sequences with E-values ranging from 10^{-25} to 10^{-75}. Students then create a document including the homologous sequences, the wild-type sequence, and their mutant sequence in FASTA format, which they load into ClustalW (Thompson *et al.*, 1994) to obtain a multiple sequence alignment. Students use this alignment to identify the mutation and to observe regions of high and low conservation. Next the students use the secondary structure prediction program PSIPRED (McGuffin *et al.*, 2000) to identify and map secondary structure predictions onto their multiple sequence alignment. The aim of this activity is to encourage students to think about the three-dimensional protein structure and provide an opportunity for students to check the prediction method. Students compare the results of the secondary structure predictions with the crystal structure data and generally find some disagreement, although the predictions are very close. This activity helps emphasize the difference between predictions using bioinformatics programs and experimental data obtained from a database.

Before students begin working with the crystal structure data, they download the pdb file for either the human protein or a homologous protein from the Protein Data Bank (Berman *et al.*, 2000). Students then view the crystal structure using DeepView (Guex and Peitsch, 1997). In viewing the structure, students are asked to develop a hypothesis for the role of the wild-type amino acid residue and the effect the mutated residue might have on protein structure and/or function. To accomplish this, students study changes in the noncovalent interactions of the amino acid side chain when the residue is mutated. The main focus of this portion of the lab is to examine the ball-and-stick view of the side chain where the missense mutation occurs and predict noncovalent interactions of the side chain based on its chemical nature and the distance of the neighboring atoms. Examination of noncovalent interactions gives the students a chance to investigate a portion of the crystal structure in depth. The students then use DeepView

Figure 2. Students working in the Bioinformatics computer lab.

to "mutate" the selected side chain to the missense mutation they are studying, examine the possible effects of the change on the local noncovalent interactions, and predict how such interactions might be maintained or changed. This activity helps reinforce the concept of non-covalent interactions in a protein structure as well as the different chemical properties of the amino acid side chains. In combination with the structure analysis using DeepView, the students are asked to draw a noncovalent interaction with the residue using Fisher projections to represent the amino acid side chains. This type of activity has been shown to help students interpret three-dimensional structures (Richardson and Richardson, 2002). Additional examples of curriculum developed to aid students in analyzing crystal structures have recently been described in the literature (Centeno *et al.*, 2003; Feig and Jabri, 2002; Honey and Cox, 2003; Richardson and Richardson, 2002).

After studying the structure, students analyze the metabolic pathway(s) containing their protein using the KEGG database. At this point, students can examine their hypothesis regarding the change in protein function and the effect on downstream events in the pathway. Finally, students develop the link between the gene they are studying and

human disease using the OMIM database (NCBI, 2000). With the "Allelic Variants" portion of the OMIM entries, the students can read about the specific mutation they are examining and compare their hypothesis with the clinical (and sometimes biochemical) data available. Finally, students organize their results into a report that contains a one-page written summary, their multiple sequence alignment, and figures created using DeepView. By the completion of each project, students have traveled from genotype to phenotype, beginning with a DNA sequence and ending with clinical data.

During the last lab session, the students meet in small groups and present their projects to each other in the form of an oral report. For these presentations, the students who are working on the same project present together to the other pair who are working on the other project in the same disease category. For example, in the Lung Cancer group, the two students working on the K-Ras project present to the two students working on the Cytochrome P450 project (see Table 2). To help ensure that students understand the presentations and the relationship between the two projects, they work together to provide answers to a group quiz about the two projects. This quiz involves four essay questions (see Table 3) that help generate a discussion about

Table 3. Quiz questions

Disease category	Joint quiz questions
Apoptosis	1. Briefly describe what apoptosis is, how reactive oxygen species (ROS) are involved, and the roles for SOD1 and caspase–1 (in either preventing or carrying out apoptosis). 2. Briefly explain how apoptosis can lead to diseases like Huntington's chorea and familial amyotrophic lateral sclerosis (FALS). 3. Explain how caspase–1 activity can lead to Huntington's disease as well as how the H237D mutation in caspase–1 is likely to affect caspase–1 activity and why. 4. Explain, at the molecular level, why SOD1 is not functioning well in the patient with the FALS-related H46R mutation in SOD1.
Lung cancer	1. What do polycyclic aromatic hydrocarbons (PAHs) do, and how can their presence lead to an increased risk for lung cancer? 2. If a patient had a mutation in either K-Ras 2 or CYP1A1, how could knowing that lead to more effective treatment (either of the cancer or for prevention of cancer)? 3. Explain, at the molecular level, why a patient with a G12C K-Ras 2 mutation may have a higher risk for lung cancer than a person with wild-type K-Ras 2. 4. Explain, at the molecular level, why the patient with the I462V CYP1A1 mutation may have an increased risk for lung cancer.
Cholesterol homeostasis	1. What is the primary gene involved in autosomal dominant familial hypercholesterolemia (FH)? Is the gene mutated in the same position in most people with FH? 2. Statins inhibit HMG-CoA reductase, which blocks a key step of cholesterol biosynthesis in liver cells. How does the decrease in cholesterol in the liver lead to decreased cholesterol in the blood? 3. Explain, at the molecular level, why the patient with a mutation in HMG-CoA reductase (D690A) is not responding to statin therapy. 4. Explain, at the molecular level, why the LDLR is not functioning correctly in the patient with a mutation in LDLR (W144S).
Mitochondrial diseases	1. Briefly describe the mitochondrial genome and how it is inherited. 2. Briefly explain why mitochondrial diseases can vary greatly between individuals, even if the individuals inherit exactly the same mutation in a mitochondrial gene. 3. Explain, at the molecular level, why the patient with a mutation in MTATP6 (L156R) is defective in oxidative phosphorylation. 4. Explain, at the molecular level, why the patient with the KEK mutation in cytochrome c oxidase is defective in oxidative phosphorylation.
Metabolic diseases	1. Briefly explain how the lack of function of an enzyme in a metabolic pathway can result in human disease. 2. How can the effects of a metabolic disease be prevented? 3. Explain as best you can, from the molecular level to the physiology level, why the patient with a mutation in HPRT1 (D193[4]N) results in Lesch-Nyhan disease. 4. Explain, at the molecular level, how the phenylalanine hydroxylase with the E280K mutation functions.

the presentations and the hypotheses that the students develop for their projects.

An Emphasis on Providing an Inquiry-Based Experience

The parts of the curriculum that emphasize inquiry are in analyzing the protein structure, developing a hypothesis about the structure, and writing and discussing the report summaries. In analyzing the protein structure, students are asked to assess the impact of the missense mutation on the protein structure, then relate this impact to effects on protein activity. With this activity, we are asking students to perform the first step in laboratory mutagenesis studies, predicting the experimental outcome of the mutation before the experiment has been performed. For most of the student projects, the mutagenesis experiment has not yet been performed, so the students are not able to look at experimental results of the mutations. In the cases where the mutagenesis results are available, such as with the phenylalanine hydroxylase project, the students read about the results and often discover that the results are not as straightforward as

their predictions. For example, in the case of phenylalanine hydroxylase, the students discover that many mutations lead to folding and stability problems. The hypotheses that the students develop concerning the consequences of the mutation are graded on the basis of how well the students explain their reasoning and how well they incorporate the chemical concepts. Students often develop different hypotheses than their partners who are working on the same project, yet both can receive full credit.

After students read about the consequences of the mutation they are studying under the "Allelic Variants" section of the OMIM database and the KEGG database, they are asked to develop a conclusion for their final report that ties together all that they have learned about the mutation they are studying. As part of this conclusion, they are asked to hypothesize how the change in amino acid could lead to a symptom of a disease. Since students often develop different hypotheses to explain the same data, these differences lead to lively discussions and debates during their oral presentations. This activity helps to emphasize the importance of sharing ideas in the process of scientific research.

ASSESSMENT TOOLS

The assessment tools are designed to:

1) determine how well the students understand bioinformatics terms and the purpose of each database and tool used in the lab;
2) assess how students feel about the inquiry-based approach. Did they enjoy working on the projects, and did they feel they had gained important skills?

To address the first aspect of the assessment, a multiple-choice test was given before and after the lab. The questions were based on information that was available in the glossary designed for the lab, although students were expected to have learned these topics through the process of their Web-based research projects. The second aspect of the assessment was addressed by a questionnaire designed to obtain students' opinions by ranking their level of agreement to a series of statements. This questionnaire also solicited comments from students on any subject related to the lab. As both of these tools were available online and class time was provided to complete the forms, both tools had a high response rate, and many comments were obtained.

RESULTS AND DISCUSSION

Pre/Posttest

The pre/posttest consists of 19 multiple-choice questions concerning Web-based bioinformatics terms, tools, and databases. Although most students had not previously used the bioinformatics tools in the way required for this lab, several students had used these tools in an undergraduate research experience, and all of the students were introduced to the BLAST program (Altschul *et al.,* 1997) in a previous biology laboratory that was a prerequisite for this course (Principles of Biology I, see *Introduction*). Students were asked to complete the test online from the lab Web site (Bednarski *et al.,* 2004) on the first day of lab and again on the last day of lab. The average score on the pretest was 42%; this increased to an average score of 77% on the posttest, indicating that the students gained in knowledge of the tools used during this course. There were no significant differences on the scores of the pre/posttest when the scores were separated by lab section. A summary of these results is described in Table 4, and the questions used in the pre/posttest are provided in Appendix A. Although improvement was generally very high, three of the questions were significantly lower than average in the improvement between pre- and posttest, indicating a need to emphasize those points in the future. Overall, we found some of the

questions more informative than others, and we are in the process of making revisions to this test.

Postquestionnaire and Comments

Statements on the postquestionnaire were designed to obtain students' perceptions of their own learning and their views on the curriculum materials, instruction, project assignment, and overall usefulness of the lab. The students were asked to respond to the statements using the Likert scale (Anderson *et al.,* 1983; Likert, 1932). This scale and the results of the survey are summarized in Table 5.

The postquestionnaire is available online from the laboratory Web site (Bednarski *et al.,* 2004), and the students completed the questionnaire during class on the day of the final presentations after completing the posttest. The questionnaires were anonymous and included an option to provide additional comments about the laboratory. Two hundred twenty-nine students out of 246 enrolled students completed the questionnaire, and 137 students chose to write additional comments. In analyzing the comments, we found that many students chose to make similar statements, and that the comments could be summarized by grouping them into categories. This grouping provides a quantitative view of the comments (see Table 6) (Wolcott, 2001).

Assessment of Goals

The major goals we had for the lab were to introduce all biology students to bioinformatics tools and databases, so that they had a basic knowledge of the types of databases and tools commonly used in biomedical research. Additionally, we wanted to create an inquiry-based lab experience for a large course by providing independent projects, requiring evidence-based predictions, encouraging a collaborative atmosphere, and requiring students to communicate their results.

In regards to our first goal, the pre/posttest results show that students gained a general understanding of bioinformatics terms, as well as the tools and databases used in the lab (Table 4 and Appendix A). On the questionnaire, students also agreed that they had learned how to access and use these tools. They responded with an average of 4.0 ± 0.9 (indicating agreement on a scale of 1–5) to six different questions on the postquestionnaire assessing their comfort in using the tools and databases. In addition, 31 students chose to comment that they were now confident that they could use the bioinformatics tools on their own (Table 6). The students were slightly less confident that they would use these tools in the future; they responded with an average of 3.6 ± 1.2 that they would use the NCBI Web site (including OMIM, LocusLink, and PubMed) in the future and an average of 3.8 ± 1.1 that they would use any of the skills learned in this laboratory in the future. However, 19 students commented directly that they could see themselves using these tools in the

---

Table 4. Pre/posttest results

Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Average[a]
Pretest (% correct)	23	48	14	20	28	35	59	43	54	41	33	29	61	65	51	26	63	69	32	42 ± 17
Post-test (% correct)	44	80	34	80	87	79	78	96	86	73	49	74	91	90	78	88	87	89	77	77 ± 17

[a]n = 218 for the pretest, and n = 180 for the posttest.

Table 5. Laboratory assessment questionnaire responses

Statement	Response average $\pm$ SD[a]
Confidence with bioinformatics skills	
Before course, I could effectively use the tools on the NCBI Web site.	1.6 ± 1
After the course, I understand how to effectively use the tools on the NCBI Web site.	4.0 ± 0.8
I understand how to interpret entries on the LocusLink site.	4.0 ± 0.8
I feel comfortable using DeepView.	4.0 ± 0.9
I have gained skills in performing Web-based research projects.	4.1 ± 0.8
This course helped me increase my skills in viewing and interpreting three-dimensional structures of proteins.	4.0 ± 0.9
This course helped me improve my computer skills.	3.1 ± 1.2
I understand how Web-based bioinformatics tools are used in biomedical research.	4.2 ± 0.8
Review of materials	
The guide sheets were too detailed.	2.0 ± 0.9
I would have liked more detailed instructions for the projects.	2.6 ± 1.2
The textbook readings were helpful.	3.3 ± 1.2
I was able to understand my assigned literature readings.	4.0 ± 0.9
Lab instruction	
Too much time was spent on the tutorials.	2.3 ± 1.0
The tutorials given by instructors and teaching assistants provided adequate introduction to the Web-based tools.	4.1 ± 0.9
Projects and group work	
Everyone in a lab section should work on the same project in the future.	2.0 ± 0.9
I would have liked to work more independently throughout the labs.	2.1 ± 1.1
Future usefulness of learned skills	
The skills I gained in this lab will be useful to me in the future.	3.8 ± 1.1
I can see myself using the NCBI Web site (OMIM, LocusLink, PubMed) in the future.	3.6 ± 1.2

In the original questionnaire, the statements in the table were randomly distributed, and students were given the following set of directions: "Please select the choice that best describes how you feel about each statement: 1 = disagree completely, 2 = disagree somewhat, 3 = neither disagree nor agree, 4 = agree somewhat, 5 = agree completely." Two hundred twenty-nine out of 246 students completed the questionnaire; 112 respondents were female, and 110 respondents were male; seven respondents did not indicate male or female. The questionnaire did not ask for any identifying information beyond "male or female" and their choice of major. Students submitted their responses anonymously online.

future (Table 6). As students encounter these tools in upper-level courses and undergraduate research experiences, we anticipate that student expectations will shift.

Our second goal was most directly addressed by the comments submitted by students. In Table 6, the first three categories of comments, totaling 44 comments, focused on gaining a better understanding of what scientists do and enjoyment of the investigative nature of the projects. Since the inquiry-based approach encourages students to mimic scientists, these comments suggest that this curriculum was successful in providing that experience for many students in the course. By providing independent projects, we hoped students would develop a sense of ownership toward their research; responses to the postquestionnaire and submitted comments suggest that the majority of students enjoyed working on their own projects. In the postquestionnaire, students disagreed with statements that everyone should work on the same project (2.0 ± 0.9), indicating that most students liked the current structure of the laboratory with respect to having their own project to work on. In order to encourage students to read journal articles, we included reading assignments from primary literature sources (the crystal structure paper and several abstracts) and a review

article on the disease under study. The assessment results showed that the students slightly preferred the journal articles to the textbook readings. On the questionnaire, students agreed with statements about the helpfulness of the textbook reading assignments with an average of 3.3 ± 1.2 while reporting an understanding of the journal articles with an average of 4.0 ± 0.9. These results suggest that students enjoyed the challenge of reading journal articles. Instructors also noted students commented that they particularly appreciated the introduction to the PubMed database for finding journal articles.

Although we gave students individual projects, we encouraged a collaborative atmosphere in the lab, and students often worked closely with another student nearby. The students also worked in small groups in order to present their research results and work on a joint quiz. In response to these activities, students disagreed with the statement that they would rather work more independently (2.1 ± 1.1). In Table 6, seven comments fit into the category of "Enjoyed partner and group work." Some comments further stated that the discussions with the group were important to understanding the projects. We interpreted these results to mean that most students enjoyed the collaborative experiences in

Table 6. Student comment data

Category	Number of comments	Examples
The lab helped me understand what scientists do	15	"I feel that this is real-life biology." "It made me realize how people can dedicate their careers to studying something like a protein or genetic disease."
Raised awareness of how biological information is shared	12	"(I) learned more about how genetic sequences are organized in a world-wide network." "I enjoyed seeing the large amount of biological information that is shared around the world."
Enjoyed focus of the projects	17	"I enjoyed the fact that we were given a real-life disease and that each lab section is designed for us to learn a little more each time about the disorder." "I enjoyed working on a specific protein and learning what diseases are associated with it and how it works on a cellular/molecular level as well as at a phenotypic level."
Will use Web sites/tools in the future	19	"I have found the Web sites helpful, and I am very likely to use them for future projects." "Most of the programs I used I can see myself using again in the future."
Enjoyed partner and group work	7	"I enjoyed working in a group...since I could discuss my results with my group and see if they agreed or had a different interpretation of the data." "I enjoyed being able to work with a partner. It made things more fun."
Felt confident in skills gained in using Web-based tools	31	"I've actually been able to use some of my skills during research that I'm doing on a protein." "I learned a lot about bioinformatics. ...The project made it clear how to use the programs."
Better understanding of links between DNA-protein-disease	6	"The best aspect of the lab was seeing in the best way possible how a gene goes from DNA to protein to phenotype." "It was extremely interesting and rewarding when we finally figured out how our protein and our disease and our mutation fit together."
Overall satisfaction with lab	34	"I thoroughly enjoyed this lab." "Overall, this was a very meaningful experience."
Overall dissatisfaction with lab	5	"I cannot truthfully say that I learned much from the lab." "It was pretty boring just poking around on the Internet."

the laboratory, which would not be possible if students worked completely independently on their projects.

Additional Observations

We observed that there were two areas where students commonly had difficulties in working through their projects. First, students often had a hard time understanding the importance of including sequences of proteins from distantly related species in their sequence alignment, and then interpreting their alignment once they had obtained it. Because of this, we plan to add an activity where students will need to predict which residues are important to protein function based on two different alignments. The first alignment will contain protein sequences from six closely related species, and the second will contain protein sequences from six distantly related species.

The second common difficulty was with predicting hydrogen bonds between amino acid side chains while studying the protein structure. Since many crystal structures are not solved to a resolution that can determine the location of hydrogen atoms, students were required to predict which acidic and basic residues would be protonated in order to

predict hydrogen bonding. To help them prepare, we will be including a new written activity with amino acid representations created in DeepView to give students some experience with drawing hydrogen bonds between two residues before they work with the protein structures.

By observing the final presentations and talking with the students, it was clear that the preparation for the report and oral presentation was important for helping students assimilate the information that they had been collecting from the various bioinformatics databases and programs. Students became more motivated to understand the big picture in the Web-based research they were conducting when they needed to explain it to others. In the future, it may be beneficial to encourage students to begin preparing earlier in the course for their presentations to help them make important connections as they work through their projects. The main feature of understanding the big picture is understanding how genotype relates to phenotype. In fact, several students commented on the postquestionnaire that they enjoyed making the connections between DNA sequence, protein sequence, and phenotype as they worked through their projects (Table 6).

Figure 3. Students presenting their oral reports.

The group quizzes given after the oral presentations generated significant discussion within the groups and were important in helping the students understand each other's work. These discussions were often lively, while students fought to defend their own ideas. The presentations were kept informal to encourage discussion, but they gave students an opportunity to use scientific language that was new to them. Figure 3 shows several groups while they were presenting their final reports to each other.

Although the curriculum is written so that most students could complete the projects working on their own time, our observations and student comments suggest that the collaboration with partners, aid from instructors, and the group work were important in helping students to avoid simple but frustrating problems, to understand the exploratory nature of their projects, and to develop their hypotheses.

CONCLUSIONS

Overall, we are satisfied that both goals for the lab were met with this curriculum. The goal of providing an inquiry-based experience for students is difficult to assess. We used the students' comments and our observations to determine that students began to think and act more like scientists while working on the lab projects (Hassard, 2005). Many students commented that the lab gave them a better idea about the ways in which scientists approach problems and share information. The independent projects helped students focus on a single research problem and develop an interest toward that problem. Students used evidence to form hypotheses related to their projects, and, through our informal discussions with the students, we noticed changes in the way they discussed their projects, using more and more scientific language. The oral and written reports gave students a chance to practice their skills in communicating scientific information, and the collaborative nature and group work in the lab gave students opportunities to both defend their ideas and learn from their peers.

In regards to our first goal, the pre/posttest and questionnaire results showed that students learned about accessing and using a variety of Web-based tools and databases. This curriculum allows students to develop a foundation of bioinformatics skills that they will be able to build on in upper-level courses. There are several small, upper-level biology courses at WU that provide additional opportunities for students to work on bioinformatics-based projects.

It is too soon to determine the impact of this lab on the upper-level courses, since this lab was taught for the first time in Spring 2004 to sophomores, and students have 2 yr more to complete upper-level biology courses. It will be interesting to monitor both enrollment and student performance in these courses. Faculty of these upper-level courses have indicated that they found the topics covered in this curriculum to be relevant to the bioinformatics projects in their courses, and that they expect students to perform better on these projects. It appears likely that this will enable us to add additional projects to the upper-level curriculum in the future.

ACCESSING MATERIALS

The curriculum described in this article is available online at http://www.nslc.wustl.edu/courses/Bio3055/bio3055.html

ACKNOWLEDGMENTS

Hafer, Frances Thuet, David Heyse, Marcia Mannen, and Rob Compton at Washington University in St. Louis for their teaching and computer support during this course.

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Andersen, O.A., Flatmark, T., and Hough, E. (2002). Crystal structure of the ternary complex of the catalytic domain of human phenylalanine hydroxylase with tetrahydrobiopterin and 3-(2-thienyl)-L-alanine, and its implications for the mechanism of catalysis and substrate activation. J. Mol. Biol. 320, 1095–1108.

Anderson, A., Basilevsky, A., and Hum, D. (1983). Measurement: theory and techniques. In: Handbook of Survey Research, ed. P. Rossi, J. Wright, and A. Anderson. New York: Academic Press, 231–287.

Bednarski, A., Heyse, D., and Pakrasi, H. (2004). The course webpage for Biology 3055 at Washington University in St. Louis. http://www.nslc.wustl.edu/courses/Bio3050/bio3055.html (accessed 2 May 2005).

Berg, J., Tymoczko, J., and Stryer, L. (2002). Biochemistry. New York: WH Freeman and Company.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. 28, 235–242.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370. http://us.expasy.org/sprot/ (accessed 2 May 2005).

Campbell, M.A. (2003). Public access for teaching genomics, proteomics, and bioinformatics. Cell Biol. Educ. 2, 98–111.

Centeno, N.B., Villa-Freixa, J., and Oliva, B. (2003). Teaching structural bioinformatics at the undergraduate level. Biochem. Mol. Biol. Educ. 31, 386–391.

Cooper, S. (2001). Integrating bioinformatics into undergraduate courses. Biochem. Mol. Biol. Educ. 29, 167–168.

Feig, A.L., and Jabri, E. (2002). Incorporation of bioinformatics exercises into the undergraduate biochemistry curriculum. Biochem. Mol. Biol. Educ. 30, 224–231.

Guex, N., and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18, 2714–2723. http://www.expasy.org/spdbv (accessed 2 May 2005).

Hart, P.J., Balbirnie, M.M., Ogihara, N.L., Nersissian, A.M., Weiss, M.S., Valentine, J.S., and Eisenberg, D. (1999). A structure-based mechanism for copper-zinc superoxide dismutase. Biochemistry 38, 2167–2178.

Hassard, J. (2005). Assessing Active Science Learning: The Art of Teaching Science. New York: Oxford University Press.

Henikoff, S., and Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. 89, 10915–10919.

Honey, D.W., and Cox, J.R. (2003). Lesson plan for protein exploration in a large biochemistry class. Biochem. Mol. Biol. Educ. 31, 356–362.

Honts, J.E. (2003). Evolving strategies for the incorporation of bioinformatics within the undergraduate cell biology curriculum. Cell Biol. Educ. 2, 233–247.

Istvan, E.S., and Deisenhofer, J. (2001). Structural mechanism for statin inhibition of HMG-CoA reductase. Science 292, 1160–1164.

Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.

Jungck, J.R., and Donovan, S. (2000). Evolution as a basis for bioinformatics education. Mol. Biol. Cell 11, 26a.

Kanehisa, M. (1997). A database for post-genome analysis. Trends Genet. 13, 375–376.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30. http://www.genome.ad.jp/kegg/ (accessed 2 May 2005).

Likert, R. (1932). Technique for the measurement of attitudes. Arch. Psych. 21, 140.

McGuffin, L., Bryson, K., and Jones, D. (2000). The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405.

Milburn, M.V., Tong, L., deVos, A.M., Brunger, A., Yamaizumi, Z., Nishimura, S., and Kim, S.H. (1990). Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. Science 247, 939–945.

National Center for Biotechnology Information (2000). OMIM—Online Mendelian Inheritance in Man. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM (accessed 2 May 2005).

National Library of Medicine (2004). PubMed. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi (accessed 2 May 2005).

National Research Council (1999). How People Learn: Brain, Mind, Experience, and School. Washington, DC: National Academy Press.

National Research Council (2000). Inquiry and the National Science Education Standards: A Guide for Teaching and Learning. Washington, DC: National Academy Press.

National Research Council (2002). Bio2010: Undergraduate Education to Prepare Biomedical Research Scientists. Washington, DC: National Academy Press.

Pruitt, K., and Maglott, D. (2001). RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 29, 137–140.

Rano, T.A., Timkey, T., Peterson, E.P., Rotonda, J., Nicholson, D.W., Becker, J.W., Chapman, K.T., and Thornberry, N.A. (1997). A combinatorial approach for determining protease specificities: application to interleukin-1beta converting enzyme (ICE). Chem. Biol. 4, 149–155.

Rastogi, V.K., and Girvin, M.E. (1999). Structural changes linked to proton translocation by subunit c of the ATP synthase. Nature 402, 263–268.

Richardson, D.C., and Richardson, J.S. (2002). Teaching molecular 3-D literacy. Biochem. Mol. Biol. Educ. 30, 21–26.

Rudenko, G., Henry, L., Henderson, K., Ichtchenko, K., Brown, M.S., Goldstein, J.L., and Deisenhofer, J. (2002). Structure of the LDL receptor extracellular domain at endosomal pH. Science 298, 2353–2358.

Shi, W., Li, C.M., Tyler, P.C., Furneaux, R.H., Grubmeyer, C., Schramm, V.L., and Almo, S.C. (1999). The 2.0 A structure of human hypoxanthine-guanine phosphoribosyltransferase in complex with a transition-state analog inhibitor. Nat. Struct. Biol. 6, 588–593.

Steward, R.E., MacArthur, M.W., Laskowski, R.A., and Thornton, J.M. (2003). Molecular basis of inherited diseases: a structural perspective. Trends Genet. 19, 505–513.

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. BioTechniques 28, 1102–1104. http://bioinformatics.org/sms/ (accessed 2 May 2005).

Thompson, J.D., Higgins, D.G., and 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 Res. *22,* 4673–4680.

Williams, P.A., Cosme, J., Ward, A., Angove, H.C., Matak Vinkovic, D., and Jhoti, H. (2003). Crystal structure of human cytochrome P450 2C9 with bound warfarin. Nature *424,* 464–468.

Wolcott, H.F. (2001). Writing up Qualitative Research. Thousand Oaks, CA: Sage Publications.

Yoshikawa, S., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., Yamashita, E., Inoue, N., Yao, M., Fei, M.J., Libeu, C.P., Mizushima, T., Yamaguchi, H., Tomizaki, T., and Tsukihara, T. (1998). Redox-coupled crystal structural changes in bovine heart cytochrome c oxidase. Science *280,* 1723–1729.

Appendix A

PRE/POSTTEST

Directions: Click in the circle next to the option that BEST answers the question.

1. Which of the following search programs/databases is NOT found at the NCBI Web site?

    A. LocusLink
    B. OMIM
    C. dbSNP
    D. PSIPRED
    E. PubMed

2. Which of the following programs DO NOT use pdb files to create a picture of a three-dimensional molecule?

    A. RasMol
    B. Chime
    C. Swiss-PdbViewer/DeepView
    D. Protein Explorer
    E. BLAST

3. All the coordinate files (.pdb) of three-dimensional protein structures that have been published are stored in the same online database. This database is:

    A. ExPASy
    B. RCSB
    C. PubMed
    D. PSIPRED
    E. OMIM

4. The online database for genetic disease, located through the NCBI Web site, is called what?

    A. ExPASy
    B. RCSB
    C. PubMed
    D. PSIPRED
    E. OMIM

5. To make a multiple sequence alignment, the sequences are saved in FASTA format in a text document, then copied and pasted into what program?

    A. BLAST
    B. ClustalW
    C. PSIPRED
    D. Swiss-PdbViewer
    E. Sequence Manipulation Suite

6. Which of the following best describes the scoring pattern by homology search programs such as BLAST?

    A. identical residue = 10 pts, gap = 0 pts
    B. identical residue = 10 pts, conservative substitution = 5 pts, gap = 0 pts
    C. identical residue = 10 pts, conservative substitution = 1–9 pts, gap = –3 pts

    D. identical residue = 10 pts, conservative substitution = 1–9 pts, gap = 0 pts
    E. identical residue = 10 pts, conservative substitution = 10 pts, gap = –3 pts

7. In the CPK coloring mode used in Swiss-PdbViewer, what atom type is blue?

    A. hydrogen
    B. carbon
    C. nitrogen
    D. oxygen
    E. phosphorus

8. The correct FASTA format for a multiple sequence alignment is best explained by which of the following statements:

    A. Use one-letter amino acid code, all capital letters, spaces separating each letter, and a return separating each line. Headings are put in quotation marks.
    B. Use one-letter amino acid code, no spaces between letters. Returns are used to separate different sequences and headings. Heading lines are marked with a ">."
    C. Use one-letter amino acid code, all capital letters, no spaces between letters, with returns separating each line of the sequence. Headings are put in all capital letters, followed by a return.
    D. Use three-letter amino acid code; capitalize the first letter of each amino acid, with a space between each amino acid.
    E. Use one-letter amino acid code, put spaces every 10 amino acids, and put the numbering system above each line of amino acids.

9. Which font is used for alignments because each letter is the same width?

    A. Times
    B. Geneva
    C. Helvetica
    D. Courier
    E. Arial

10. Which of the following sets of information is NOT usually found in a SwissProt entry?

    A. the three-dimensional coordinates for the crystal structure
    B. tissue and subcellular localization
    C. amino acids in the active site
    D. alternate names for the protein
    E. the E.C. number for an enzyme

11. Which of the following databases would you select for a BLAST search if you would like to obtain the most homologous protein sequence to your search sequence that has its three-dimensional structure solved?

    A. nr
    B. swissprot
    C. pdb
    D. month
    E. pat

12. Which of these sites combines the following information for a particular gene: RefSeq files, KEGG links, PubMed articles, OMIM entries, and dbSNP entries?

    A. GenBank
    B. ExPASy
    C. LocusLink
    D. RCSB
    E. PSIPRED

13. Why is GenBank called a redundant database?

    A. because it contains two versions of every entry
    B. because it is saved in multiple locations for safety
    C. because it contains similar sequences for different organisms
    D. because it contains a historical record of sequence discoveries which sometimes includes multiple versions of the same sequence
    E. because it is reread many times to check for mistakes

14. RefSeq files are useful to researchers for what reason?

    A. RefSeq files have been edited and contain the most up-to-date sequence information for a particular site.
    B. RefSeq files are edited to contain just the most biologically interesting sequence information.
    C. RefSeq files are generated solely by a computer so contain no mistakes.
    D. RefSeq files contain sequence information generated only from the Human Genome Project.
    E. RefSeq files are the only sequence files that have been obtained experimentally.

15. Approximately how close together should two atoms be in a crystal structure to assume hydrogen bonding?

    A. 0.25 angstroms
    B. 0.5 angstroms
    C. 3 angstroms
    D. 6 angstroms
    E. 20 angstroms

16. Which of the following BEST describes the KEGG database?

    A. a database of all macromolecular structure files
    B. a database of metabolic pathways
    C. a database of genetic disease genes
    D. a database of phylogenetic trees
    E. a database of enzyme substrates

17. Which of the following terms describes SNP's that result in an amino acid change in the protein?

    A. synonymous change in the noncoding region
    B. synonymous change in the coding region
    C. nonsynonymous change in the noncoding region
    D. nonsynonymous change in the coding region
    E. mutation in the promoter region

18. A conservative mutation means:

    A. an amino acid change in a conserved region of the protein
    B. an amino acid change in a nonconserved region of the protein
    C. an amino acid change from acidic to basic
    D. an amino acid change from hydrophobic to hydrophilic
    E. an amino acid change to an amino acid with similar size and chemical properties

19. If you were interested in reading a referenced summary about the research done on sickle cell anemia, which Web site would be best to go to?

    A. KEGG
    B. LocusLink
    C. Sequence Manipulation Site
    D. OMIM
    E. SwissProt