Ramping Up to the Biology Workbench: A Multi-Stage Approach to Bioinformatics Education

Kathleen Greene[1] and Sam Donovan[2]

[1]Departments of Education and Chemistry
Beloit College
Beloit, WI
greene@beloit.edu

[2]Department of Instruction and Learning
University of Pittsburgh
Pittsburgh, PA
sdonovan@pitt.edu

Abstract: In the process of designing and field-testing bioinformatics curriculum materials, we have adopted a three-stage, progressive model that emphasizes collaborative scientific inquiry. The elements of the model include: (1) context setting, (2) introduction to concepts, processes, and tools, and (3) development of competent use of technologically sophisticated tools. A curriculum involving the analysis of HIV sequence data is used to illustrate this framework and provide a context for discussing this student-centered, inquiry-based approach to bioinformatics education and literacy.

Keywords: bioinformatics, inquiry, HIV, phylogenetic trees, multiple sequence alignment

INTRODUCTION

The recent explosion of publicly available molecular sequence and structural data and on-line tools to analyze those data provide expanded opportunities to incorporate their use in undergraduate education. Although the technical requirement for accessing the databases and analytic tools is minimal -- browser-based Internet access -- engaging students in realistic biology problem solving is more complicated. The "ramping up" metaphor is used to describe a process in which conceptual understanding and tool usage are developed simultaneously and with progressive complexity. This approach allows us to emphasize the difference between the generally low *technical* barriers to manipulating analytic tools and the generally high *conceptual* barriers involved in the selection of appropriate data and tools, the interpretation of analytic results, and connecting the abstract molecular information to more familiar biological phenomena.

We have been involved in bioinformatics curriculum development projects aimed at bringing bioinformatics to several audiences including undergraduate biology majors, pre-service biology teachers, and teaching faculty. We have operated with the belief that meaningful learning can be promoted within carefully structured, deliberately ordered problem spaces that give students the opportunity to pursue research without becoming awash in the technical details of this dynamic and emerging field.

In this paper we present both a general approach to introducing bioinformatics problem solving and a specific instantiation of that general approach. We developed a notion of "progressive problem spaces" using a three-step approach to engage students (and faculty) in realistic research and problem solving using bioinformatics data and tools. These steps involve (1) establishing a context for the use of bioinformatics, (2) providing an introduction to the data, tools and reasoning patterns involved in bioinformatic analyses, and (3) creating open-ended opportunities for research using rich data resources and sophisticated analytic tools. This approach is exemplified with a set of activities that proceed from a basic orientation to bioinformatics to open-ended investigations of HIV evolution.

STAGE ONE

Context Setting: Seven Scenarios Activity

We developed an introductory context-setting activity we call "Seven Scenarios." Each of the seven scenarios in this collection (Parents, Police, Patents, Privacy, Patients, Profit, and Peanuts) consists of four or five short written statements, each written on a separate index card (Figure 1).

Seven Scenarios: A Context Setting Activity for Studying Bioinformatics & Biotechnology

Parents: expectant parents and a gene associated with a disabling condition.

1. Scientists have identified a gene and have developed a test for different forms of a gene.
2. One form of this gene is considered a risk factor for a disabling but not fatal condition.
3. A couple is expecting a baby.
4. Both prospective parents "carry" this form of the gene.
5. The parents are concerned about the costs of raising a child with a disability.

Police: culpability of someone accused of transmitting a virus.

1. A person is accused of sexually transmitting a virus
2. Police use blood tests to try to determine if that person is the source of the virus.
3. The virus causes a disease appearing years after the initial transmission.
4. The defendant's lawyer argued that because the viruses in the accused and the accuser were so different, her client should not be found guilty.

Patents: drug companies seeking gene patents.

1. A patent grants exclusive ownership of intellectual property so that the patent owner can profit from its use.
2. Many biotechnology firms are pursuing patents for gene sequences.
3. The companies hope that the gene sequences can be used to develop specific biological products.
4. There is currently a rush to apply for patents on any possibly useful sequences.

Privacy: a job candidate's pre-employment physical.

1. A candidate for a job is required to take a pre-employment physical.
2. Genetic analysis identifies a form of a gene that has been linked to high blood pressure.
3. The relationship between this gene form and high blood pressure is not well understood. Some people with the gene have normal blood pressure and many without the gene have high blood pressure.
4. The candidate is hired, but is told he will have to pay higher premiums for medical insurance.

Patients: a physician paying for genetic study of possible drugs to treat her patient.

1. A physician is treating a patient who has an aggressive and lethal cancer.
2. The physician pays a biotechnology company $37,000 to find potentially effective drugs.
3. The company identifies three drugs that are then used to treat the patient
4. The patient's cancer goes into remission.

Profit: for-profit and not-for-profit genomic enterprises.

1. The Human Genome Project is a consortium of academic research groups trying to determine the sequence of the human genome.
2. HGP (Human Genome Project) is a not-for-profit project that receives a great deal of public funding.
3. A for-profit company, Celera, uses the publicly available HGP data to check its work, fill in the gaps, and stay a step ahead.
4. Celera maintains a private database, available to corporations for subscription fees of 5 to 15 million dollars per year.

Peanuts: genetically modified organisms finding their way into the human diet.

1. A fast food restaurant recently had to dump some food because it contained an unapproved ingredient.
2. A strain of peanut has been engineered to resist a fungus known to wipe out a whole season's crop.
3. New foods do not need FDA approval if they meet three conditions: 1) the nutritional value is not lowered, 2) the food is already present in the human diet, and 3) the food is not an allergen.
4. Many people are allergic to peanuts.

Figure 1*. Scenarios used to establish a context for using bioinformatics.*

The cards can be distributed among the students, and groups are asked to assemble according to their scenario. Once the groups are assembled, each person in the small group reads his or her card to the group, and then the group discusses the scenario, considering these three questions:

- Is there any information about the scenario that you wish you had or that you felt was missing? In other words, was there enough information to consider?
- What issues (philosophical, historical, political, scientific, ethical) arise in discussion of this scenario?
- What kind of research or investigation would you consider doing based on this scenario?

After a few minutes of group discussion, the groups read their sentences to the other groups and report on the small group discussion, usually referring to the three questions. This activity usually generates discussions around people's own experiences with pre-employment physicals, their understandings of genetically modified organisms, and questions about how people pass along a serious disease. Often someone in the group has relevant specific knowledge or a personal experience to share. A variety of questions will undoubtedly arise and can be shared across the groups. It will be up to the teacher to decide if it is appropriate to attempt to answer some of these questions at this juncture or postpone them for later discussions.

There are several important outcomes of this short activity. First, we have learned much about what the students know and don't know about biotechnology and bioinformatics. Second, we have kindled their interest in these topics in part by reminding them of examples that they have read about in the news, heard about, or experienced. Third, we have involved everyone, from the beginning, in deliberately collaborative, low-stress, non-intimidating situations. By assessing background knowledge, sparking interest, and generating an experience and expectation of participation, we establish a context and a springboard for studying bioinformatics and engaging in bioinformatics activities.

STAGE TWO

Concept and Skill Establishment: Is He Guilty?

The second phase of our three-step progression provides opportunities for students to become more familiar with some of the types of data, techniques, and graphic representations that are used in sequence analysis. This activity involves working through a small problem in a series of three discrete steps to introduce learners to some of the ideas behind sequence analysis. The goal is to establish a shared conceptual understanding and introduce reasoning skills that can be applied to a variety of contexts and problems involving comparative sequence analysis.

In this example we examine the use of HIV sequence data as forensic evidence linking a Florida dentist and some of his HIV+ patients. Groups of students are provided with a series of printed materials, including raw data and the output from various bioinformatics analyses, and asked to determine whether there is evidence that the dentist is the source of HIV infection for his patients. By taking the analysis in several steps and allowing both small group and whole group discussion at each stage, it is possible to quickly bring the entire class to a relatively sophisticated understanding of how the analysis of molecular sequences can be used to support biological claims. The materials below provide an overview of our approach to establishing concepts and skills using the dentist HIV forensics example. The data files, images and additional discussion can be found on the web site <http://bioquest.org/bioinformatics/> and a similar activity is described in more detail in *Microbes Count!* (Donovan, 2003; Donovan & Weisstein, 2003).

Step 1: Sequence data

In the Spring of 1990 Kimberly, a 22 year old living in Fort Pierce, Florida, tested positive for HIV -- she had no identifiable risk factors for contracting the virus. Epidemiological research focused on an invasive dental procedure performed by an HIV+ dentist several years earlier. From the dentist's records a number of other HIV+ patients were identified, several of whom had no known risk factors for contracting the virus. The Centers for Disease Control and Prevention (CDC) became involved and the case received a great deal of media attention based on the public's concern that HIV+ health care workers may be a threat to their patients (Gentile, 1991; MMWR, 1990, 1991a, 1991b). Multiple lawsuits were filed by patients claiming that they were infected by the dentist and seeking damages. In court, attempts to link the patients' HIV to the dentist's HIV rested in part on comparative analyses of the virus sequences (Ou, et al., 1992).

In order to explore the role of sequence analysis in resolving this type of question, groups of students are provided with a small collection of raw amino acid sequence data from HIV viruses collected from three patients, the dentist, a local control and an outgroup (see Fig 2). We also distribute a printout of the abstract from a paper reporting on the analysis of these sequences, as well as one of the GenBank® sequence records (Ou, et al., 1992). These resources make the scenario more nearly "real" for students by allowing them to see the mechanisms used by scientists to share their data and report their results.

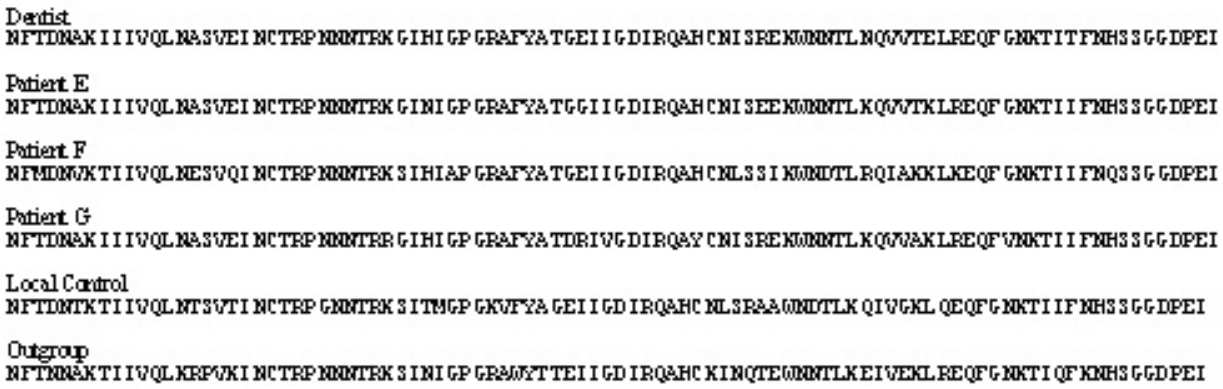

Figure 2. *Amino acid sequence data from HIV found in the dentist, three patients, a local control and an outgroup.*

As we progress through each of the sections of this activity, the driving question for groups to consider is whether they feel they have evidence to link the dentist to the HIV in any of these patients. For this section they are also asked to consider:

- What sorts of patterns do you see within/between these sequences?
- How are these sequences similar (different)? Are they all similar (different) in the same ways?
- How do you think this information could be used to determine if the dentist were the source of the HIV in the patients?

Groups are encouraged to keep a list of questions that arise during their discussion. Students often use highlighters or other visual methods to begin comparing the order of the letters in each sequence. They generally notice that the sequences have different lengths. Many also recognize that because the sequences are so similar, it is efficient to look for differences rather than similarities.

In the subsequent discussion of what the groups have learned, there should be an abundance of promising ideas and questions to address. There are often specific questions about how to interpret the data (e.g., What do the letters mean? What is a local control?). Other questions may focus on the nature of the virus (e.g., Are all HIV viruses identical within a person? How fast does HIV change? Do any two people have identical HIV viruses?). Still other questions might relate to making comparisons of the sequences (e.g., Is it significant that some letter combinations don't seem to change much and others change quite a bit? Are there likely to be more changes in one area of these sequences compared with other areas?). It is interesting to watch a shared language develop as students work to describe what the groups have seen. Some of these questions can be addressed directly, some are postponed, and others are reflected back to students to help them integrate their existing biological knowledge.

Step 2: Interpreting a multiple sequence alignment (MSA)

For the next round of group work, we introduce one of the standard techniques for comparing sequences, a multiple sequence alignment. This is, in many respects, what some students will have already begun to work on in their groups when presented with the raw sequence data. They are readily convinced that the longer the sequences and the more of them there are, the more cumbersome it becomes to align them by hand. They see that this is the perfect type of work for a computer. What is more difficult to have them understand, though, is that the parameters used in the algorithm for a sequence alignment reflect a set of assumptions about the relationships between those molecules. We distribute to the groups an alignment of the sequences with which they have been working, and provide them with information about the pairwise comparisons between sequences (Figures 3 & 4). Once again they work in their groups, and we prime their discussions with the following questions:

- Does the information presented in these outputs support the patterns you saw when you looked at the raw sequence data?
- Do you think some of the amino acid changes are more important than others?
- Why do you think that two of the sequences needed to have a "gap" (-) inserted to make them align with the others?
- How do you think this information could be used to determine if the dentist were the source of the HIV in the patients?

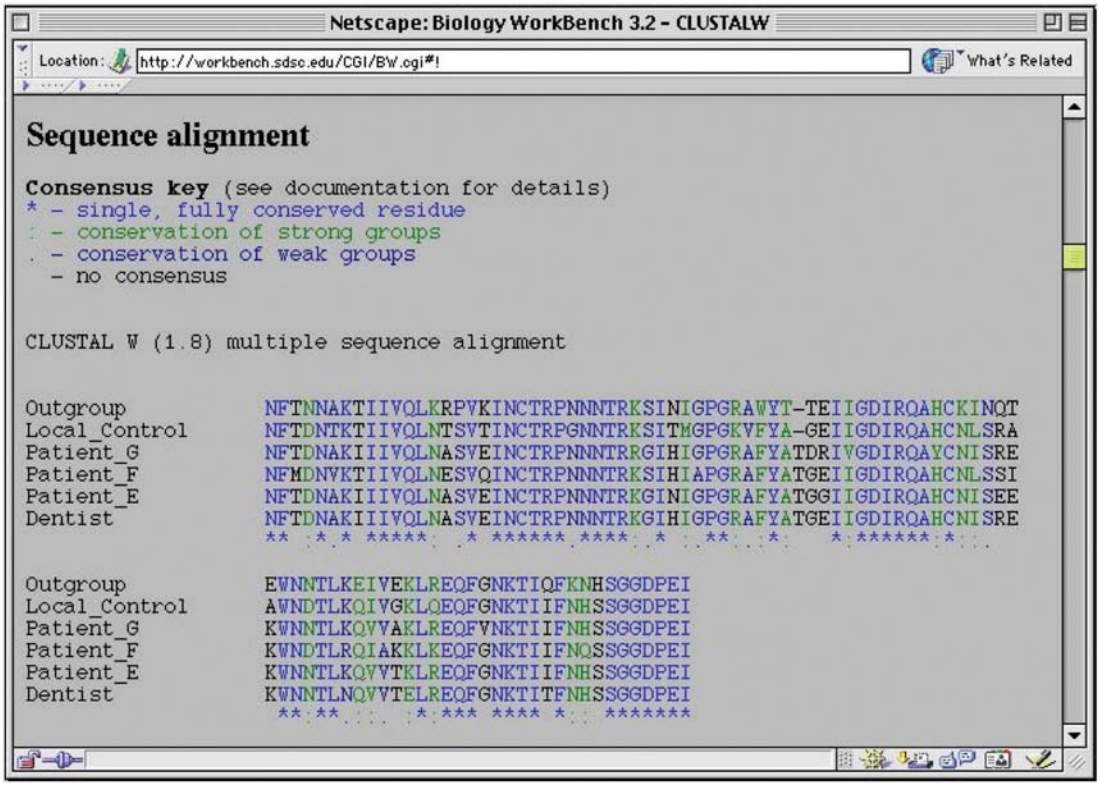

Figure 3. *A multiple sequence alignment of the HIV amino acid sequences listed in Figure 1.*

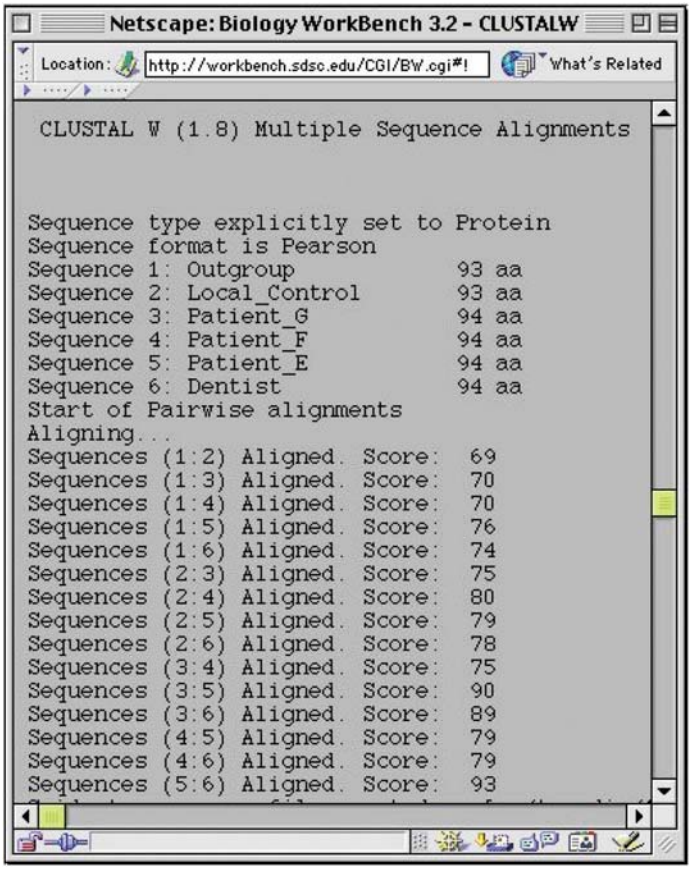

Figure 4. *Pairwise sequence comparisons listing the percent identity between each pair of sequences from the sequences in Figure 1.*

The ensuing discussions provide great teachable moments. Groups that we have worked with have brought up the common origins of sequences as a source for their similarity, the notion of mutation "hot-spots," conservation of sequence for conservation of structure and function, and the similarities and differences between groups of amino acids. Still, even with MSA and pairwise comparisons, it is difficult for students to argue effectively for role of the dentist in transmitting HIV to certain patients.

Step 3: Reading trees

In the final step of this second-stage activity we provide each group with an unrooted tree built from their aligned sequence data (Figure 5). They then have another opportunity to work with their group to address the following questions:

- Does the information presented in this tree representation support the patterns you saw when you looked at the raw sequence data and the multiple sequence alignment?
- Why do you think some of the lines are longer than others? Do you think the places where the lines connect with one another is important? What does it mean?
- How do you think this information could be used to determine if the dentist were the source of the HIV in the patients?

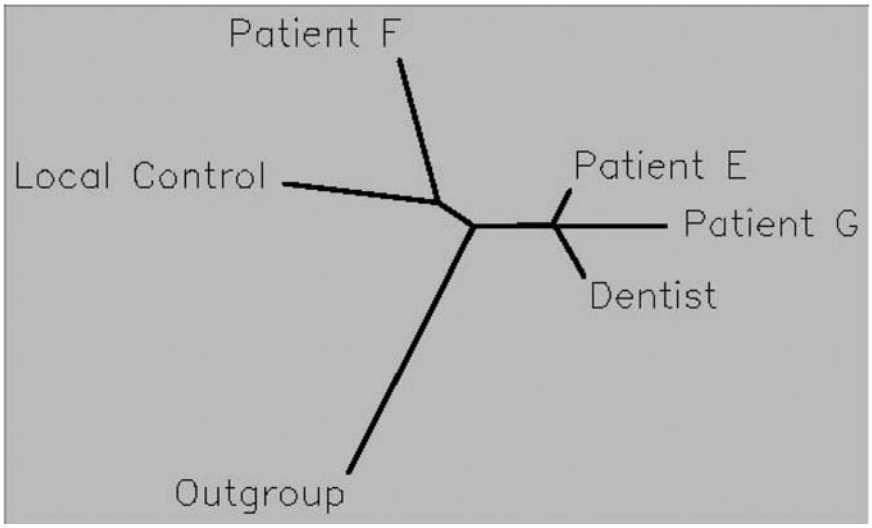

Figure 5*. An unrooted tree representing the genetic distance between the sequences in Figure 1.*

By the conclusion of this final discussion, we find that students have become conversant about a variety of important aspects of bioinformatics research including but not limited to: (1) the types of information that are associated with sequence data submitted to public research databases (2) ways to read similarities and differences between sequences from multiple representations of molecular data (3) how a multiple sequence alignment summarizes the comparisons of sequences (4) how a phylogenetic tree graphically represents the differences between sequences and can be used to develop hypotheses about evolutionary relationships (5) how evolutionary relationships between sequences can be used as forensic evidence.

STAGE THREE

Exploring HIV Evolution: An Opportunity to Do Your Own Research

The third stage of developing bioinformatics inquiry skills builds from the previous activity to engage students in investigations of their own questions using molecular data. This exercise is open, in that it provides students opportunities to make decisions and develop their own research strategies, but it is not unstructured. The use of a published data set both simplifies the problem, so students don't need to search for sequences or decide if particular sequences are appropriate to compare, and limits the range of questions that can be addressed, because the data set lends itself to certain types of analyses and is not appropriate for others.

This stage brings all of the pieces together. It involves students in, and connects them to, the biology, the analytical tools and a specific data set. We discuss the biology of HIV, emphasizing information that is pertinent to the data set they have been provided. We use a collection of several hundred sequences from 15 HIV+ patients taken over a period of time (Markham, et al., 1998). We hand out a summary table of the data that are available and discuss briefly how the data were collected (Figure 6). The groups are then asked to look over the summary data table for interesting patterns, and to think about possible research questions. As a

whole class, we brainstorm possible research ideas, which accomplishs several teaching objectives. We get some concrete ideas in the air, further orient students to the data available, link the data to what we know about HIV biology, and illustrate the range of potentially fruitful investigations that one could undertake. At this time, we also call attention to the idea that the same data set can be used to generate multiple hypotheses.

Summary of the data set

Subjects: 15
Number of visits: 3-9
Number of clones per visit: 2-18
CD4 cell counts for each visit

Subject	Total Number of Visits	Total Number of Clones	Visit Number	Number of Clones	CD4 Count[1]
1	3	42	1	13	464
			2	16	305
			5	13	15
2	3[2]	24	1	6	715
			3	9	825
			4	9	830
3	5	39	1	4	819
			3	10	375
			4	9	265
			5	10	100
			6	6	45

Figure 6*. Part of the data summary table for the Markham, et al. 1998 dataset.*

As a next step, we select one or two questions and model what the students will be asked to do in their own investigations. This allows us to work through the process of focusing a general question like, "Is there a particular change in the HIV sequence that causes the T-cell count to drop?" Together, we generate some specific analytic ideas that could be used to address this broad question. For example, one could compare sequences from individuals who did not have T-cell count drops to those who did, or maybe compare sequences from one time to those from a later time. These discussions help students recognize that a variety of decisions need to be made in order to make progress on any research question and that these decisions will be central to the process of relating their results to their scientific claims.

Next, the groups work together at their tables to begin defining their research questions and methods. Depending on the setting, we might introduce students to the Biology WorkBench website <http://workbench.sdsc.edu> to show them the mechanics of choosing sequences and running analyses. We build in multiple opportunities for feedback and peer review by getting groups to share their preliminary results with one another. The outputs the groups see are the same as the printed materials they worked with in the second stage. We encourage groups to print their findings and bring them back to the conference room, where they have table space to lay them out and consider the results in light of their research questions. This also promotes the important idea that the generation of computer outputs is only a preliminary step to answering their research question, which requires careful analysis and interpretation of those outputs, as well as a coherent synthesis and presentation of the investigation and results.

Students prepare posters and hold a research meeting at the end of this third stage. We generally see a high level of student engagement with each other's research. Having worked with the same data set and struggled with the same conceptual issues, the class has a chance to become a real research community.

DISCUSSION

At least five core aspects characterize our bioinformatics curricular model. First, the entire curriculum is set within an inquiry context, it is a question-based curriculum. Second, it is a collaborative model, in which students and teacher think and talk with one another in groups both large and small. Third, students are placed in a decision-making role; it is they who ask the questions, which they investigate. Fourth, the science and tools are taught in context. Finally, and this is where the "ramping up" occurs, there is an intentional progression of concepts and procedures from simple to complex,

and from the more conceptual to the more technological. Complex concepts build on simple ones, and sophisticated technological tools are used to carry out tasks that are based on and emerge from the spectrum of concepts from the most basic to the most complex.

In our curriculum model, it is the students whose work is "up front" and the instructor whose work is "behind the scenes." That is not to say that we do not play an active or directorial role. We carefully construct biological scenarios that are interesting and that promote certain types of questions appropriate to the setting. We select molecular data sets that are rich with possibility. Our choices in designing these problem spaces are guided by four goals: (1) to establish real-world and science context, (2) to review and provide necessary biology content and relevant concepts, (3) to guide progressive exposure to and experience with bioinformatics data, techniques and representations, and (4) to develop an awareness of and facility with bioinformatics tools, such as Biology WorkBench, all in a context of inquiry. Over the course of the activities, we review our goals, and adjust our facilitation as necessary to accomplish them.

Understanding the uses of bioinformatics, what the various data represent, which tools to use, and what inferences are reasonable are essential to a successful bioinformatics educational experience. Conceptual, procedural and technological understandings are dynamic and fluid, and all must be present for meaningful learning and understanding in bioinformatics. Let us use our example to describe these three overlapping categories and some of the understandings within them.

Conceptual understanding means that students have robust knowledge that allows them to work with ideas in appropriate and meaningful ways. To be successful in bioinformatics, students need to be familiar with and understand large biological ideas such as inheritance, evolution, genetics, mutation and the somewhat more specific biological notions of DNA, transcription, translation, replication, amino acids, and protein synthesis, etc. More specialized bioinformatics concepts include knowledge of molecular databases and sequencing and other analytic heuristics and tools.

Procedural knowledge includes general scientific procedures, such as those associated with collaborative inquiry (Bruce & Levin, 1997) and problem solving (Peterson & Jungck 1988), but also specific procedures, such as multiple sequence alignment and analysis and gel electrophoresis. Interpretation of analytic outputs such as phylogenetic trees requires both conceptual and procedural knowledge.

Technological understandings likewise extend from the general to the specific, including the use of computers and basic applications for a variety of tasks, such as word processing and internet searching. Specific technological understandings include knowledge and skills associated with bioinformatics technology, both those using computers as a central tool, such as molecular database searching, sequence selection and retrieval, and subsequent analysis, as well as other tools such as wet lab apparatus. Use of bioinformatics technology can aid students in both the generation and interpretation of analytic outputs such as the phylogenetic trees mentioned above. Our bioinformatics curriculum provides experiences that highlight all of the components described above. The first stage, "Seven Scenarios," does not so much develop as elicit, assess, and lay groundwork for developing conceptual, procedural and technological understanding. It provides an opportunity for students to retrieve and demonstrate their existing conceptual knowledge. This also provides a form of global problem-solving procedural practice, as students work in groups to think and talk through possible problems to pose and pursue. Also through discussion, students set the stage for developing an awareness of technological possibilities.

In the second stage, students confront various representations of data, practice asking biological and procedural questions, and experience and develop fundamental technical knowledge of the basic processes on which bioinformatics is based, such as sequence comparison and analysis, tree building, and the interpretation of all of these. Their development of conceptual and procedural meaning allows them to develop an appreciation and see a need for using technology to investigate interesting biological questions.

In the third stage of our model, the students use, and develop skill with powerful and specific technological bioinformatics tools. They use these tools within an authentic research and content context; they use and build their conceptual and biology content knowledge, and they engage in real biological inquiry. Students are getting training in technology, as researchers, who can direct the technology to meet needs that *they* identify rather than as technicians who perform tasks set out by others.

Having constructed the scenarios and selected the data sets with the previously stated principles and goals in mind, we share them with students, and literally invite them to engage in discussion and inquiry. We provide a flexible structure in which to discuss and inquire, but the students are decision makers at each stage. In the first stage, students evaluate the information they have been given and they decide what else they need to know. In the second and third stages, students pose and pursue questions of and with their data, progressively applying and enhancing conceptual, procedural and technological knowledge as they pursue their investigations, making and reflecting on their research decisions as they proceed.

The curriculum model we have designed is inquiry-based, collaborative, student-centered, and intentional in both sequence and context. In their use of technologically sophisticated tools that rapidly carry out familiar procedures based on understood concepts, collaboratively investigating answers to interesting questions they themselves have set, students develop rich and contextual knowledge and understanding of biology and bioinformatics concepts, procedures, and technologies.

REFERENCES

Bruce, B., & Levin, J. (1997). Educational technology: Media for inquiry, communication, construction, and expression. *Journal of Educational Computing Research*. 17(1) 79-102.

Donovan, S. (2003). Molecular forensics. *In* J. Jungck, M. Fass & E. Stanley (Eds.) *Microbes count: Problem posing, problem solving and peer persuasion in microbiology*. American Society for Microbiology Press. pp. 129-136.

Donovan, S., & Weisstein, A. (2003). Exploring HIV evolution: An opportunity for research. *In* J. Jungck, M. Fass & E. Stanley (Eds.) *Microbes count: Problem posing, problem solving and peer persuasion in microbiology*. American Society for Microbiology Press. pp. 137-148.

Gentile, B. (1991). Doctors with AIDS. *Newsweek*. 48-56.

Markham, R., Wang, W., Weisstein A., Wang, Z., Munoz, A., Templeton, A., Margolick, J., Vlahov, D., Quinn, T., Farzadegan, H., Yu, X. (1998). Patterns of HIV-1 evolution in individuals with differing rates of CD4 T cell decline. *Proceedings of the National Academy of Sciences 95*(21): 12568-73.

MMWR (1990). Possible Transmission of Human Immunodeficiency Virus to a Patient during an Invasive Dental Procedure. 39(29) 489-493. Available at <http://www.cdc.gov/mmwr/preview/mmwrhtml/00001679.htm>

MMWR (1991a). Epidemiologic Notes and Reports Update: Transmission of HIV Infection during an Invasive Dental Procedure -- Florida. 40(2) 21-27,33. Available at <http://www.cdc.gov/mmwr/preview/mmwrhtml/00001877.htm>

MMWR (1991b). Epidemiologic Notes and Reports Update: Transmission of HIV Infection During Invasive Dental Procedures --Florida. 40(23) 377-381. Available at <http://www.cdc.gov/mmwr/preview/mmwrhtml/00014428.htm>

Ou, C.Y.; Ciesislski, C.A.; Myers, G. et al. (1992). Molecular Epidemiology of HIV Transmission in a Dental Practice. *Science*. 256,1165-1171.

Peterson, N. & Jungck, J. (1988). Problem-posing, problem-solving, and persuasion in biology. *Academic Computing.* 2(6), 14-17; 48-50.

Call for Reviewers

We are looking for persons who are willing to review manuscripts for *Bioscene*. We need reviewers for a wide variety of subject areas. Reviewers should be willing to provide in depth reviews and detailed suggestions for authors concerning revisions necessary to improve their manuscript for possible publication. Reviewers should be willing to provide a rapid turn-around time for the manuscripts they review. If you are interested in reviewing for *Bioscene*, please send an email that includes your phone number, FAX number, and a list of the areas for which you are willing to review to: William Brett, Chair of the Editorial Board, at lsbrett@isugw.indstate.edu.