COMPUTERS MEET BIOLOGY – AN INTERDISCIPLINARY AND COLLABORATIVE APPROACH

Joanna Urban and Mila Kwiatkowska
Thompson Rivers University
Kamloops, BC, Canada

Abstract

Interdisciplinary methods create connections between traditionally distinct disciplines such as sciences, social studies, arts, or languages. This paper presents the experience of team-teaching Bioinformatics by faculty from Computing Science (MK) and Biological Sciences (JU) in a Computing Science undergraduate course, "Introduction to Biomedical Informatics." The main purpose, to involve students, with little or no biological background, in actual research projects related to Biology and Medicine, was accomplished by hands-on exercises and assignments using real data from a medical clinic, and real DNA and RNA data from the sequencing of genes related to seed dispersal of a parasitic plant, Dwarf Mistletoe.

Introduction

Biology depends on chemistry and physics to explain certain biological phenomena, leading to the development of the fields called *biochemistry* and *biophysics*. Correspondingly, the substantial amount of data collected by biologists needs analysis requiring tools developed within computing science. As the result, modern biologists are required to learn new skills that go beyond standard biology and have to use methods from bioinformatics, a computational branch of molecular genetics (also called *in silico* biology). Likewise, modern computer scientists have to learn interdisciplinary skills that allow them to apply their knowledge to other disciplines (Cohen, 2004).

This paper shows how to introduce computer undergraduate students to bioinformatics, and how to create practical exercises to utilize and interpret the real data gathered in genomic research. The underlying motivation for our collaborative work is to demonstrate that both disciplines can benefit from this symbiotic relationship. Therefore, the course included topics ranging from specific programming tasks (e.g., Perl regular expressions for DNA sequence processing), applications of specialized software developed for biologists (e.g., BLAST, GenePattern, Protein Database), through computational modeling techniques (e.g., sequence analysis using Hidden Markov Model) to specialized areas of computing science (e.g., machine learning techniques used in the analysis of gene expression).

Our general pedagogical approach is based on constructivism (Piaget, 1970) in which the teacher's role is to facilitate learning through creation of a learning environment and the students have an active participant role (discovering, constructing, experimenting, and validating new knowledge through analysis and interactions with other group members). As the result, the students (in consultations with the researchers) are required to determine the skills they need to solve a real world problem (Rodgers, 2002). In particular, our pedagogical approach uses methods from the following interrelated styles of learning: (a) inquiry-based learning (IBL) (Bruner, 1961) including problem-based

learning (PBL) (Barrows, 1996), (b) experiential learning (Kolb, 1984), and (c) a creative research process (Hmelo-Silver & Barrows, 2006). Figure 1 shows the interactions between teaching/research environment and the students involved in studying and learning.

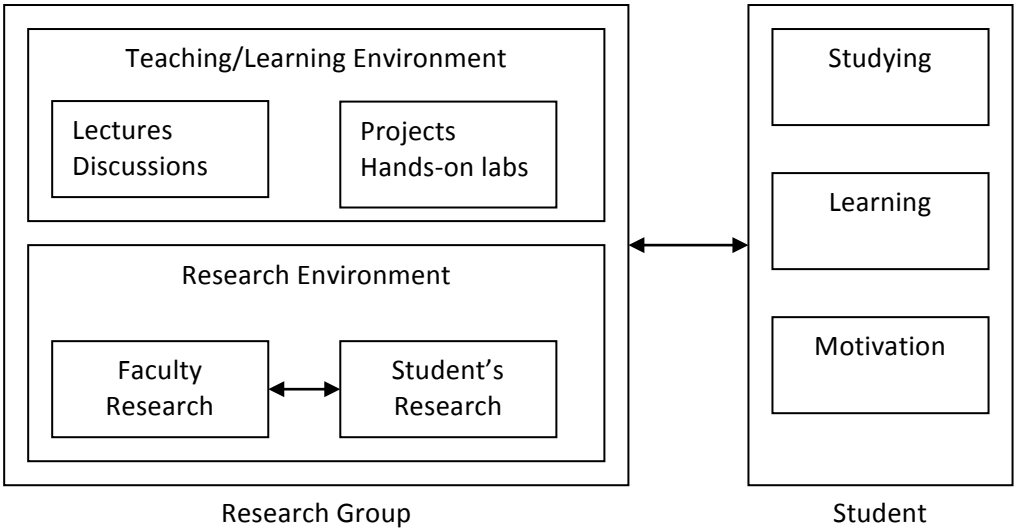

Figure 1. Interactions between educational components. (Adapted with permission from Kwiatkowska & Matthews, 2007.)

This paper presents the real-life experience of team-teaching by faculty from Computing Science (MK) and Biology (JU). It concentrates on teaching and learning bioinformatics in an undergraduate course, "Introduction to Biomedical Informatics." Biomedical informatics is a rapidly growing field that examines biological and medical data, information, and knowledge and their storage, retrieval, and optimal use for problem solving. The main purpose of this course is to involve the students in authentic research projects in order to create interest in biomedical informatics as important discipline and as possible future career. The course is inquiry-based, providing students with anonymized data from a clinic and RNA and DNA data from the sequencing of an aquaporin gene, taken directly from JU's research. The modelling and analysis of medical data and Perl programming is based on the research of MK. The course uses experiential learning by involving a faculty from biology (JU), who presents her research in a *wet* biological lab and the processing of the data in a computer lab. Furthermore, the course utilizes group research projects to engage the students in the research process.

The practical components include programming laboratory, data analysis assignments, and field trips to a biology wet laboratory. The group projects, carried throughout the course, encompass project proposals, formal presentations, and project reports. These hands-on students' projects develop rudimentary skills necessary for future undergraduate and graduate research (Airasian, Cruikshank, Mayer, & Pintrich, 2001; Bloom, 1984).

The paper is organized as follows. The first section describes the biomedical course context and presents the main course units. The second section discusses an example of students' research projects. The third section provides examples of the course assessment. In the last section, we present discussion and future work.

Course Description

The biomedical course has been designed to provide the undergraduate Computing Science (CS) students with basic knowledge and data analysis skills in bioinformatics. The course is offered as an upper-level elective course in a four-year Bachelor of Science program at Thompson Rivers University (TRU). Although most undergraduate CS students should have a basic level of knowledge in Biology from high school or introductory Biology courses, the international students and mature students require brief introduction to basic molecular biology before they can challenge more difficult problems related to JU's research on genes. The introduction entails basic information on cellular and molecular biology (DNA, RNA, Central Dogma, proteins, and genes) and basic techniques used in molecular genetics: gel electrophoresis and Polymerase Chain Reaction (PCR). Whenever possible, we used real data to help students understand complex problems, such as gene expression, applications of sequence alignments (BLAST), cloning, real-time quantitative reverse transcription PCR (qRT-PCR), sequencing, and microarrays.

Sample Course Units

All units are designed using the following four assumptions: (a) content must be based on a real-life problem, must use real-life data, and provide (or contribute to) solutions to an important real-life problem, (b) learning environment for the students should be hands-on including practical laboratories using several software packages, (c) the learning process must combine individual and collaborative group work, and (d) the students should be given opportunities to work gradually with several smaller problems and to integrate their skills and knowledge in a larger meaningful research project.

Unit 1. In the first unit, students were introduced to biological concepts, for example, they studied a parasitic plant, *Arceuthobium americanum* (Dwarf Mistletoe). In addition, the real DNA & RNA data from the sequencing of genes related to seed dispersal of Dwarf Mistletoe (JU research) was used. The mode of dispersion of seeds was explained (see Figure 2) and then students were asked specific questions about the mechanism of dispersion: "Why does water move to a fruit?" and "How is that such a large pressure enables seeds to be dispersed a distance of 20 meters with a speed of 100 km^{-}per hour?" (Hawksworth & Wiens, 1996; deBruyn, Paetkau, Ross, Godfrey & Ross Friedman, 2015).

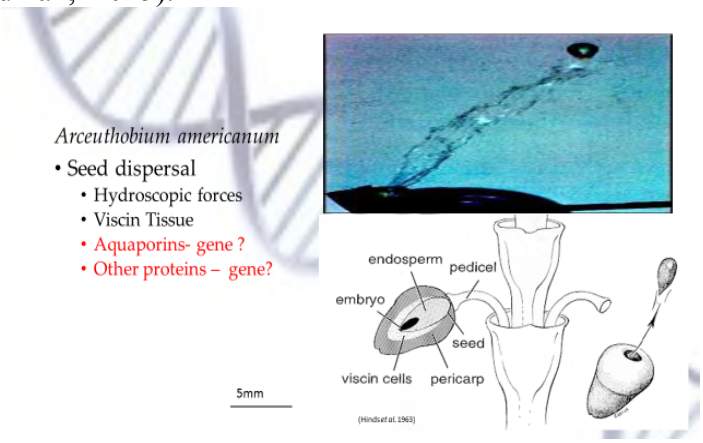

Figure 2. Arceuthobium americanum seed dispersal mechanism. (Adapted from Hawksworth & Wiens, 1996; Kelly, Ross Friedman, & Smith, 2009.)

In order to stimulate students' interest, the next diagram was shown (Figure 3), and the students were asked to explain, in their own words, which techniques can be utilized (based on their biological and CS knowledge) to obtain results seen below. The process was discussed in relation to the flow of information in which DNA holds a vast amount of data.

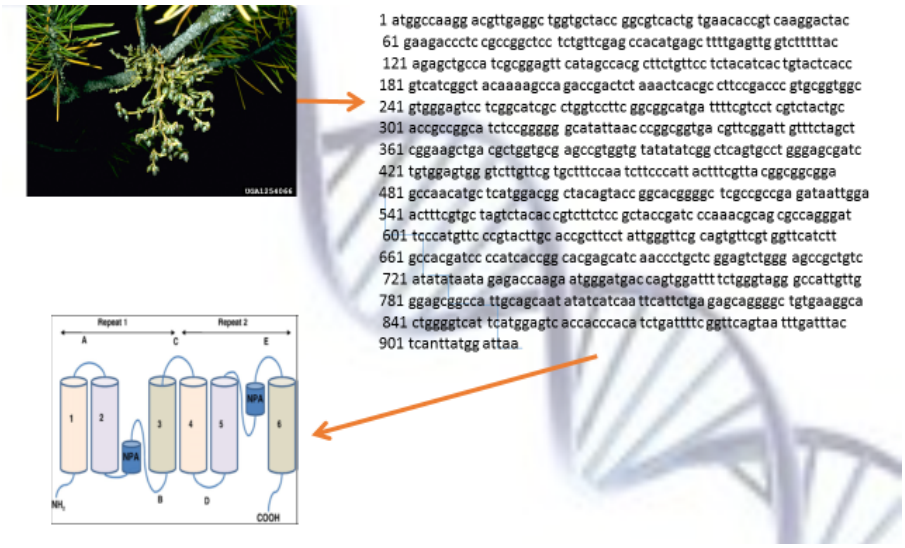

Figure 3. From plant tissue to sequence of bases (agtc) in DNA to protein (Aquaporin). (Aquaporin structure adapted from Da Ines, 2008.)

Students felt very engaged at this point, as they could ask discipline-specific questions and they were guided through the creation of a work-flowchart. At that time the students were given the detailed information about JU's research.

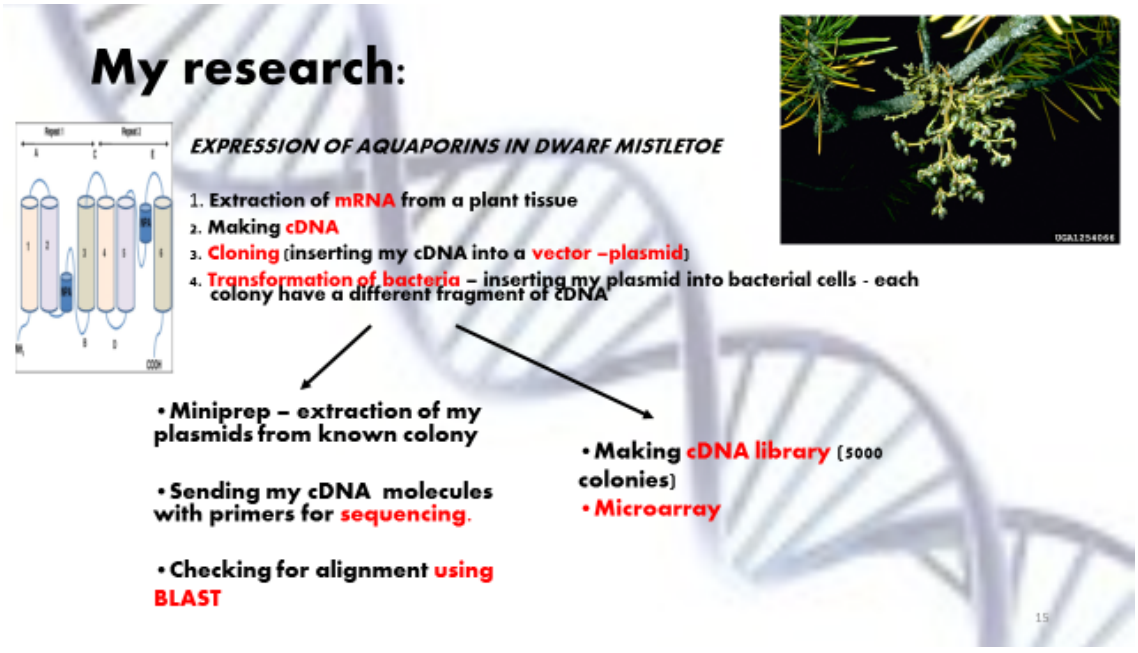

Figure 4. A flowchart with details of JU research lab work.

Finally, a conference poster (Figure 5) was shown to demonstrate how biology connects with CS.

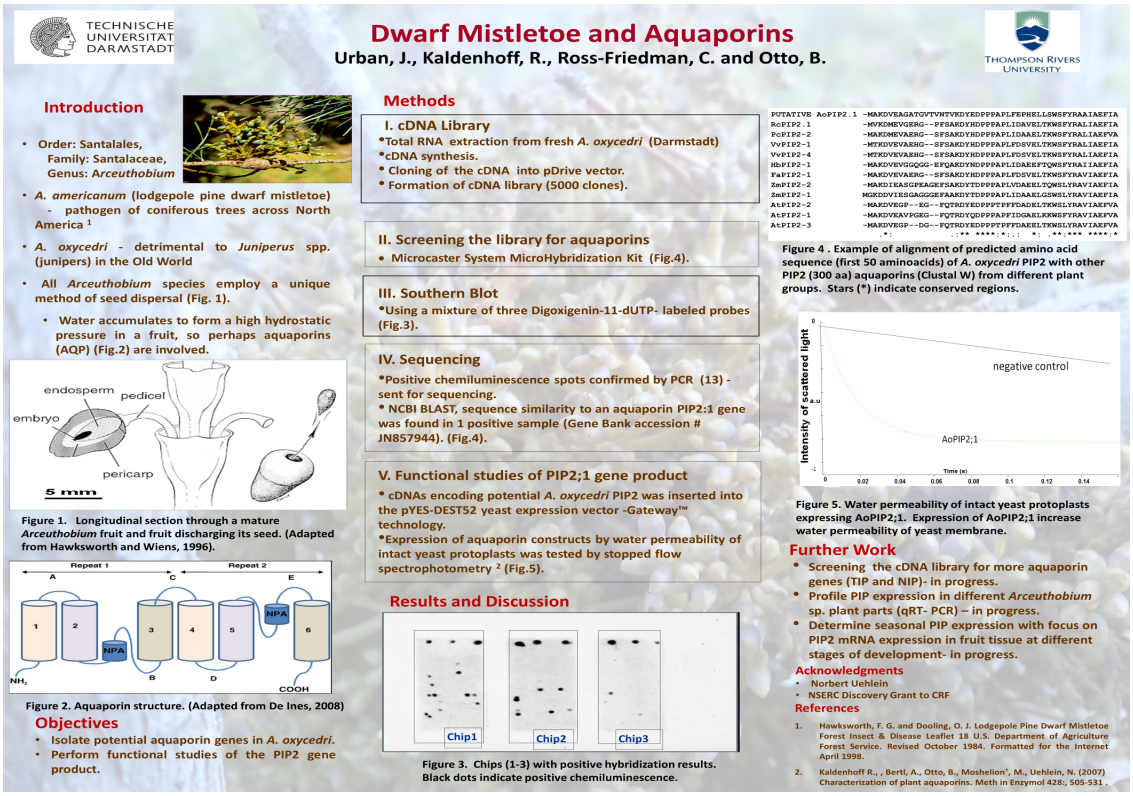

Figure 5. JU conference poster presented at Botany Conference, Canada (2013).

Unit 2. In the second unit, MK introduced a programming language, Perl. This language has been widely used in bioinformatics, and it does not have a steep learning curve for the CS students. The students were introduced to processing DNA sequences (strings in DNA alphabet), Open Reading Frames (ORF), and FASTA file format. The following is an example of an exercise that uses the data from JU research on aquaporin:

> Write a Perl program to calculate the number of four nucleotides in a sequence, the percentage of each nucleotide, the number of errors, and the percentage of G and C in the DNA. Create a FASTA file containing the gene JN857944 *Arceuthobium oxycedri* aquaporin. The gene can be found at http://www.ncbi.nlm.nih.gov/nuccore/JN857944.

Unit 3. In the third unit, MK introduced the students to the Basic Local Alignment Search Tool (BLAST) for comparing sequences (similarity measures) using existing sequences (Aquaporin *Areuthobium oxycedri* gene). The students were introduced to NCBI, BLAST and PubMed database. The following is an example of an exercise using the real data and BLAST algorithm from the NCBI server: "Use JN857944 *Arceuthobium oxycedri* aquaporin gene sequence http://www.ncbi.nlm.nih.gov/nuccore/JN857944 and use NCBI BLAST server to perform BLAST analysis."

Unit 4. In the fourth unit, we introduced an exercise on gene expression in cancer research. The unit was organized into four steps: (a) introduction to the necessary biological knowledge in order to understand gene expression and interpret the

microarray data (combination of lecture, video, and small exercises), (b) familiarization with the biology laboratory and the sequence scanner machine (visit to the biology lab and demonstration of the sequence scanning machine), (c) introduction to the machine learning software, Weka (in-lab hands-on exercise to view and understand the microarray array data), and (d) working on a larger assignment to perform gene expression classification using the leukemia data set (studying relevant research papers using several software tools, presenting the solutions in class, and preparing a written documentation).

The content of the unit was organized around the key problem: molecular classification of cancer data. In the unit, we utilized a well-known paper and we used the publicly available leukemia data sets. The students were asked to read the paper to learn about the problem of the classification of leukemias and to understand the related data. The original data file has 72 examples with about 7,000 genes. The data set contains bone marrow samples obtained from adult acute leukemia patients at the time of diagnosis, before chemotherapy. The data are for two populations of patients: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The simplified file has 72 examples with 40 genes (identified as the best classifiers). We provided the students with both files in a csv format. The leukemia data are available from the Broad Institute server with the GenePattern software (http://www.broadinstitute.org). The students had to use data mining techniques (decision tree induction J4.8 algorithm from the Weka repository) for the analysis of gene expression.

Task 1. The students were asked to read a paper written by Golub, Slonim and Tamayo (1999) and answer the following questions: (a) Why is the classification of leukemias important? and (b) Describe the distinction between a class discovery and a class prediction.

Task 2. The students were asked to (a) use the provided leukemia data sets to run the decision tree induction algorithm, (b) run the induction algorithm, and (c) visualize the decision tree and document the results. Furthermore, the students were asked to compare the genes distinguishing ALL from AML (genes listed in the Golub et al. [1999] paper with the genes included in the decision tree). The induced decision tree is shown in Figure 6.

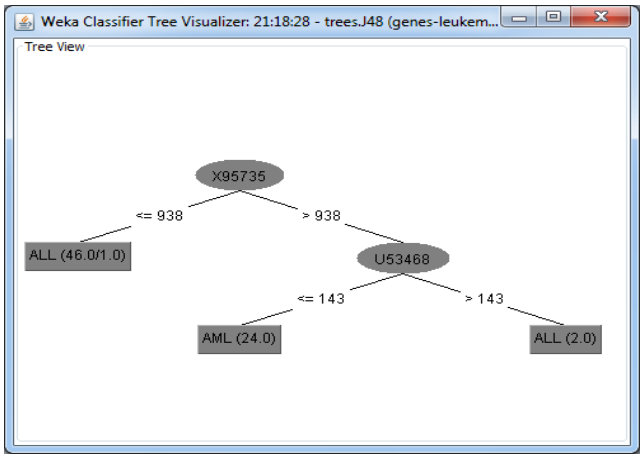

Figure 6. Decision tree for the leukemia data set.

Task 3. The students were asked to use the GenePattern (Kuehn, Liberzon, Reich, & Mesirov, 2008; Reich, Liefeld, Gould, Lerner, Tamayo, & Mesirov, 2006) server to

produce visualization of the most predictive genes. They had to use a heatmap and create gene profiles. Figure 7 shows an example of a profile for X95735 (Zyxin). Furthermore, the students were asked to use the National Center for Biotechnology Information (NCBI) web site to find the *cytogenetic location* (on the chromosomes). For example, the zyxin (X95735 *Homo sapiens*) gene has the location 7q32 (Gene ID: 7791).

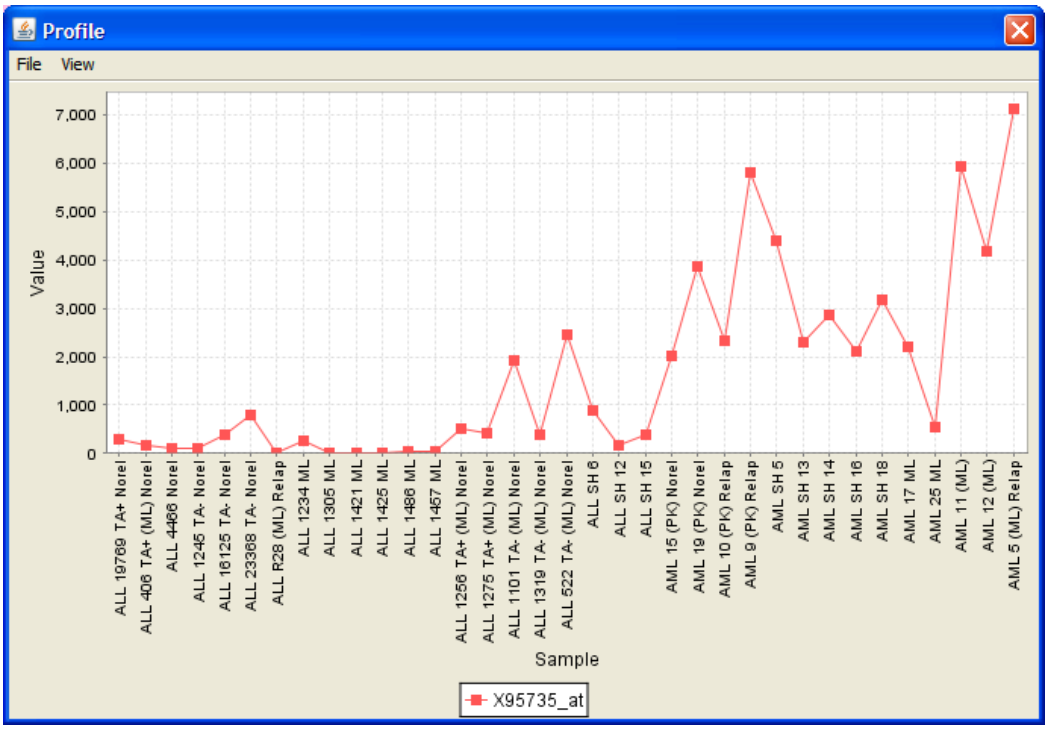

Figure 7. Profile for X95735 (Zyxin) based on the leukemia data set.

Biology Laboratory Experience and Students' Research

The CS students had an opportunity to visit the biology wet lab and to work on their research projects. The research projects provide a good foundation for life-long learning and problem solving. The students had to learn six fundamental research skills: (a) managing a research project (planning, scheduling, time management, and communication), (b) conducting a literature review, (c) building computational models, (d) generating and testing hypotheses, (e) conducting data analysis using statistical and data mining methods, and (5) report writing (Roach et al., 2001). For example, two students (LB, MM) worked on a research project "Genetic tree analysis of aquaporin genes of *Arceuthobium oxycedri* and similar genes using Molecular Evolutionary Genetics Analysis 5 (MEGA 5)." The following is an excerpt from students' research project's introduction:

> The goal for this research project was to explore how similar or different the proposed aquaporin protein of *Arceuthobium oxycedri* is from the aquaporin proteins of other plant species. We are also interested in plants that have aquaporins that are structurally similar (therefore functionally similar) to those found in *Arceuthobium oxycedri*. To explore these topics we queried protein databases to find similar aquaporin proteins to that of *Arceuthobium oxycedri* and built a phylogenetic tree of these proteins, thus looking into their history and similarity. As part of this process we performed sequence alignments between

Arceuthobium oxycedri's aquaporin and the aquaporins found in the database and were able to visualize the proteins' similarities and differences.

To study this relationship, the students built a phylogenetic tree using MEGA 5 program and used Multiple Sequence Comparison by Log Expectation (MUSCLE), a sequence alignment tool.

Assessment Process

This course used the CS standard assessment, which includes: labs, assignments, a research project, a midterm exam, and a final exam. We discussed the labs, assignments and the research project in the previous sections. The midterm and final exam had several interdisciplinary questions. The following are the examples showing the level of the required knowledge.

1. Given the following DNA sequence, write a **reverse-complement** DNA (mark the **5'& 3')** **5' T A T G C A 3'**

2. Complete the following short (6-base) DNA palindrome GTG _? _? _?

3. What is the difference between mRNA and tRNA?

4. Using the following scoring schema: **match = +2 mismatch = -1 gap = -1** , calculate the score for the alignment between two sequences:

   ```
   A G G A T A C C
   _ G G _ A A T C
   ```

5. Explain the following terms: **in silico**, **in vitro**, **in vivo.**

6. Translate the sequence (**use the second and third frame**) into amino acids: **UUGACGGAGUAG.**

7. Geneticists use maps to describe the location of a particular gene on a chromosome. One type of map uses the cytogenetic location to describe a gene's position. Specify the cytogenetic location for a gene which is located on a short arm of chromosome 6 in region 2, band 1, sub-band 3.

Discussion and Future Work

In this paper we discussed the importance of team-teaching in interdisciplinary courses. We described the real-life based learning units, assignments, research project, and the assessment process. We stressed the role of using real-life problems and connecting the students with the researcher who is actually conducting the studies. We described how we have engaged the CS students in each step of the JU's research process, and how the students were using the real data.

In each of the offerings of this biomedical course, we used an approach that is learner-centered, interdisciplinary, problem-based, and inquiry-based. The students were able to work with minimal supervision on real world problems. This interdisciplinary research was both challenging and highly motivating for the computing science undergraduate students. Additionally, the students were able to learn how to communicate

professionally with faculty and students from other disciplines (biology and environmental studies). We have conducted a standard course evaluation at the end of each course offering, and the students commented positively about their learning experience. For example, they asked for a second course in bioinformatics: "There should be a Bioinformatics II." Moreover, they expressed their enthusiasm about learning the applications of computers in biology: "It was good to learn something about biology and its relating to computing." Additionally, one student, MM, presented his learning experience during educational seminar at TRU and two students, LB and MM, continued their research in a final capstone project course in CS, and they have published a paper in bioinformatics.

Following students' comments and our discussions with other faculty members in CS and biology, we are planning to expand our work in three directions: (a) development of new course modules in wet laboratory to have hands-on experience with extraction of DNA from plants, running PCR, and studying the obtained DNA sequence using computer-based techniques such as BLAST, (b) application of other programming languages, for example, Python, and other ML techniques, for example, associative rules induction, and (c) creation of a course-specific students' evaluation (currently, the students use generic evaluation forms).

Acknowledgements

We would like to thank all students who took the "Introduction to Biomedical Informatics course" and whose enthusiasm kept us motivated to teach complex interdisciplinary skills. In particular, we would like to thank Lee Bergstrand and Matthew McInnes for allowing us to use their abstract from their research project.

References

Airasian, P., Cruikshank, K., Mayer, R., & Pintrich, P. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. L.W. Anderson & D. R. Krathwhol (Eds.), *JSTOR* (1st ed.). Pearson. Retrieved from http://www.jstor.org/stable/pdf/42926529.pdf

Barrows, H. S. (1996). Problem-based learning in medicine and beyond: A brief overview. *New Directions for Teaching and Learning, 68*, 3–12. doi:10.1002/tl.37219966804

Bloom, B. S. (1984). *Taxonomy of educational objectives*. Boston, MA: Allyn and Bacon.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31*(1), 21–32.

Cohen, J. (2004). Bioinformatics— An introduction for computer scientists. *ACM Computing Surveys, 36*(2), 122-158. doi:10.1145/1031120.1031122

Da Ines, O. (2008). Functional analysis of PIP2 aquaporins in Arabidopsis thaliana (Unpublished doctoral dissertation). Ludwig-Maximilians-Universität, München, Germany.

deBruyn, R. A. J., Paetkau, M., Ross, K. A., Godfrey, D. V., & Ross Friedman, C. (2015). Thermogenesis-triggered seed dispersal in dwarf mistletoe. *Nature Communications*, *6*, 6262.

Golub, T. R., Slonim, D. K., & Tamayo, P. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression. *Science*, *286*(5439), 531-537.

Hawksworth, F. G., & Wiens, D. (1996). *Dwarf mistletoes: Biology, pathology and*

systematics. Agricultural Handbook 709. Washington DC: United States Department of Agriculture, Forest Service.
Hmelo-Silver, C. E., & Barrows, H. S. (2006). Goals and strategies of a problem-based learning facilitator. *Interdisciplinary Journal of Problem-based Learning*, *1*, 21-39.
Kelly, S. K., Ross Friedman, C. M., & Smith, R. G. (2009). Vesicular cells of the lodgepole pine dwarf mistletoe (*Arceuthobium americanum*) fruit: Development, cytochemistry, and lipid analysis. *Botany, 87*, 1177–1185.
Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
Kuehn, H., Liberzon, A., Reich, M., & Mesirov, J. P. (2008). Using GenePattern for gene expression analysis. *Current Protocols in Bioinformatics*, *7*(12), 1-7.
Kwiatkowska, M., & Matthews, L. (2007, May). *A creative framework for research courses*. Paper presented at 12th Western Canadian Conference on Computing Education (WCCCE-2007). Thompson Rivers University, Kamloops, BC, Canada.
Piaget, J. (1970). *Structuralism*. New York, NY: Basic Books.
Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., & Mesirov, J. P. (2006). GenePattern 2.0. *Nature Genetics, 38(5),* 500-501.
Roach, M., Blackmore, P., & Dempster, J. A. (2001). Supporting high-level learning through research-based methods: a framework for course development. *Innovations in Education and Teaching International, 38*(4)*,* 369-382.
Rodgers, C. (2002). Defining reflection: Another look at John Dewey and reflective thinking. *Teachers College Record*, *104*(4), 842-866.

Author Details

Joanna Urban
jurban@tru.ca
Mila Kwiatkowska
mkwiatkowska@tru.ca