FLEXIBLE APPROACHES FOR TEACHING COMPUTATIONAL GENOMICS IN A HEALTH INFORMATION MANAGEMENT PROGRAM

Posted on June 24, 2013 by Administrator

Categories: Education & Careers, HIM Operations, Summer 2013
Tags: education, flexibility, genomics

by Leming Zhou, PhD; Valerie Watzlaf, PhD, RHIA, FAHIMA; and Mervat Abdelhak, PhD, RHIA, FAHIMA

Abstract

The astonishing improvement of high-throughput biotechnologies in recent years makes it possible to access a huge amount of genomic data. The association between genomic data and genetic disease has already been and will continue to be applied to personalized healthcare. Health information management (HIM) professionals are the ones who will handle personal genetic information and provide solid evidence to support physicians' diagnoses and personalized treatment strategies, and therefore they will need to have the knowledge and skills to process genomic data. In this paper, we describe flexible approaches for teaching a computational genomics course in the HIM program at the University of Pittsburgh. HIM programs at other universities may choose an appropriate approach to fit into their own curriculum.

Keywords: education, genomics, flexibility

Introduction

The improvement of high-throughput DNA sequencing technologies will make it possible for everyone to get his or her personal genome sequenced in the near future.[1, 2] These digitalized genome sequences will either become part of enhanced electronic health records (EHRs) or be closely linked with existing EHR systems.[3–6] Therefore, HIM professionals should have the knowledge and data analysis skills to extract desired information from personal genomes to support personalized healthcare practices.[7, 8] These skills and concepts are typically taught in a computational genomics course. (Computational genomics is a subject that uses computational methods to get useful information from genome sequences and related data.)

However, most current HIM programs have no computational genomics course in the curriculum. Currently, computational genomics courses are often offered by biology or computer science departments. However, the courses offered by these departments are not suitable for HIM students even though HIM students typically have taken a few college-level biology and computer science courses before they are admitted to a HIM program. Biology departments usually assume students have extensive knowledge in biology, especially genetics and biochemistry, whereas computer science departments assume students have a solid background in computer science, especially in programming and computer algorithms. HIM students tend to struggle in these types of courses, and they also may not be able to obtain what they strongly desire: the ability to identify the right tools to process genomic data efficiently and confidently. After all, biology departments mainly focus on detailed biological processes, and computer science departments often focus on data

analysis algorithms. These senior-level courses in biology or computer science departments also require a number of prerequisites, while HIM students are already fully occupied by courses in their own curriculum. These factors motivated us to develop flexible approaches for teaching computational genomics in a HIM program. Instructors in a HIM program may identify one suitable approach for their program to introduce genomics to their students. We believe this strategy would be a more efficient and effective way of teaching this subject in a HIM program.

In the next section, we describe our multilevel approaches for teaching HIM students computational genomics. In the third section, we present the evaluation methods and the results obtained for each approach. The fourth and fifth sections offer discussion and our conclusions.

Teaching Approach

The general learning objectives of our computational genomics courses and course modules are as follows:

1. Getting familiar with online tools and databases for genomic data analysis; and
2. Obtaining confidence in handling genomic data by finishing a course project.

To help students fulfill these objectives, we created four sets of course materials and taught them in different formats according to the length of the available class time in a HIM curriculum. The course materials include the following approaches:

1. a 90-minute tutorial;
2. a course module with one overview lecture and one lab session (each of which was two hours long);
3. a course module with four two-hour lectures/labs; and
4. one full-semester course.

All the course materials described in this article are posted at http://www.cpath.pitt.edu/courses.html and are freely available to everyone. The tutorial and course modules were integrated into existing and required courses in the HIM curriculum, and therefore no extra burden was added to students' workload.

The tutorial was given in a course named HIM 1480: HIM Clinical Education 3. The content of the tutorial was mainly about the importance of genomics in the HIM profession and its possible applications in personalized healthcare.

The two-session course module was incorporated into a course named HIM 1406: Data Management and Analysis for HIM Professionals. In this course module, we introduced several basic concepts such as DNA, RNA, gene, genome, and sequence alignment; a few genomics databases such as GenBank,[9] RefSeq,[10] and Protein Data Bank;[11] and frequently used genomic data analysis

tools such as BLAST[12] and the UCSC genome browser.[13]

The four-session course module was integrated into a course named HIM 1455: Quality Management. In this course module, we arranged two lectures and two labs. In the two lectures, we introduced three gene-finding methods and corresponding practical techniques. In the two lab sessions, students were guided to use several online tools and databases to analyze DNA sequences. Each student was then required to use these tools and databases to analyze an eukaryotic DNA sequence he or she selected from a list and report gene models embedded in the sequence.

The stand-alone genomics course (HRS 2425: Genomics and Personalized Care in Health Systems) does not have any prerequisite courses. We introduced basic concepts and relevant data analysis skills step-by-step without assuming students had taken any college-level biology courses. We integrated research components into the stand-alone course. Each student was required to work on one genome annotation project. The genomic sequences were taken from a few recently sequenced fruit fly genomes.[14] A team of the Genomics Education Partnership (GEP) at Washington University in St. Louis[15] performed genome finishing (a process of improving the quality of the raw sequences from DNA sequencers) and chopped the finished genome into segments of 40,000 base pairs each. Each student in this genomics course worked on one segment and reported all the genetic features of this DNA segment. Each student wrote a research report at the end of the semester to describe his or her research findings and gave a presentation to defend his or her approaches.

Evaluation Methods and Results

We used different evaluation methods for our course modules and the stand-alone course. For the 90-minute tutorial, the major purpose was to provide a brief introduction to genomics and to motivate students to pay attention to this topic. We did not assume students had any knowledge about genomics or had taken other college-level biology courses before the tutorial. We also did not assume they could solve any genomic problems after the tutorial. Therefore, we did not perform any formal evaluation on the tutorial. For the two course modules and the stand-alone course, we conducted precourse and postcourse surveys and tests and summarized students' performance.

Two-Session Course Module in a Data Analysis Class

The two-session course module was taught twice in the data analysis class in Fall 2010 and Fall 2011. There were 30 students in the Fall 2010 class and 29 students in the Fall 2011 class. Seven students chose not to submit their survey responses. Fifty-two students participated in the course evaluation. To evaluate this course module, we designed the following survey and evaluation questions:

Q1. List college level biology courses you have taken.

All 52 students had taken at least one college-level biology class such as Biology 1, Biology for Nonmajors, Biology 100, and General Biology. A few students had taken up to four college-level biology courses.

Q2. Have you heard of Human Genome Project (HGP)? If you have, what are the purposes of this project?

In this course module, HGP was explained at the very beginning of the lecture. The lecture content was also organized according to some goals of HGP (such as sequencing the human genome, creating tools to analyze DNA sequences, and developing databases to manage the raw and curated data).

Q3. Do you know the relationships among genome, chromosome, DNA, gene, and protein?

In the lecture, several students were asked to explain the relationship among these concepts. Although most of these students could not explain this relationship completely, they did have some idea of it. This relationship was then explained in detail in the lecture.

Q4. How would you store a DNA sequence on a computer disk?

During the lab session of this course module, students were guided to access the GenBank database, download a few DNA sequences, and save them onto their computers. They were then asked to open these saved files and read the content in these files.

Q5. If you are given a DNA sequence with 50 million bases, what tools may be used to view it and extract information from it?

During the lab session of this course module, students were guided to use the UCSC genome browser to view the longest chromosome in the human genome. (The length of chromosome 1 is roughly 250 million bases.) They were then taught to zoom in and focus on one particular gene in chromosome 1 and download the DNA sequence for that genomic region from the genome browser. The instructor demonstrated how to determine the specific locations of a gene and extract genetic variation records in that region.

Q6. If you are given a DNA sequence with 10,000 bases and you do not know the source of this sequence, which tool may you use to determine its source or claim this sequence has never been discovered before?

During the lab session, students were given a file with a DNA sequence in it. The length of the sequence was roughly 15,000 bases. The descriptive information about the sequence was removed beforehand. The students were then guided to perform a BLAST search against a nucleotide database at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) website. The instructor explained the output from this BLAST search in detail and specifically guided

students to figure out the source of the sequence or establish the uniqueness/novelty of the sequence.

Q7. If you are given two DNA sequences, and one has 3,000 bases and the other has 4,000 bases, which tool can be used to compare these two sequences and determine their similarity level?

During the lab session, students were given two DNA sequence files; one was from the human CFTR gene (6,132 base pairs) and the other was from the mouse CFTR gene (6,305 base pairs). The instructor then guided the students to perform a BLAST search between these two sequences and explained the BLAST output in detail.

The collected results for the questions were categorized into three knowledge levels:

- Level 1: The student did not know the answer at all (score = 0).
- Level 2: The student had a partial answer to the question (score = 1).
- Level 3: The student provided a complete answer to the question (score = 2).

Table 1 summarizes the survey results conducted in the two-session course module. From these numbers, we can observe that a large percentage of students grasped these basic genomic concepts and could recognize correct genomic data analysis tools for given questions after participating in the module. By matching the names of the students in pre- and posttests, we also noticed that 24 of 52 students consistently did a better job on the posttest, and 8 of 52 students consistently did not have good performance.

We roughly evaluated the overall performance of these students by calculating a total score for each question before and after the course module. The score calculation method was simple:

- If M students are assigned to knowledge level 1 (no answer or completely wrong answer) in a question, then 0 × M = 0 points are added to the score.
- If N students are assigned to knowledge level 2 (partial answer) in the question, then 1 × N = N points are added to the score.
- If P students are assigned to knowledge level 3 (complete answer) in the question, then 2 × P = 2P points are added to the score.

The total score for each question was then N + 2P points. By comparing total scores for each question on the pre- and posttests, we can see dramatic performance improvement after this course module.

In addition, we observed a weak linear relationship between the students' performance in this survey and the number of college-level biology courses they had taken (the correlation coefficient between the number of biology courses taken and the total score on the pretest was 0.14, $p = .49$; the correlation coefficient between the number of biology courses taken and the total score on the posttest was 0.22, $p = .28$; the correlation coefficient between the number of biology courses taken

and the improvement after the module was 0.12, $p = .57$). In other words, students may not need to take multiple college-level biology courses to perform well in this class.

Since this two-session class was a short course module, we believed it was sufficient for students to simply recall the concepts or tools from their memory.

Four-Session Course Module in a Quality Management Class

In this four-session course module, we applied a different approach to evaluate the performance of students because the objective of this course module is different. We still conducted a background survey; however, our major evaluation focus was whether these students could grasp the genomic data analysis skills and use online tools or databases to determine the gene structure in their selected sequences.

To reach this desired goal, the instructor delivered two lectures and offered two lab sessions, each of which was two hours, in Spring 2011 and Spring 2012. In the first lecture, the methods for extracting meaningful information from a genomic sequence were explained. In the second lecture, a practical procedure for identifying genes in a genomic sequence was explained with extensive details and specific examples. This step-by-step procedure was also written down and posted on a website with a concrete example (http://www.cpath.pitt.edu/genoAnnot.htm); in addition, lecture slides were provided to those students. In the first lab session, students were guided to perform exercises on BLAST and the UCSC genome browser. They also learned to access databases at FlyBase[16] (a major data source for fruit fly genomes) and GEP website (http://gep.wustl.edu). In the second lab session, the instructor led the students to perform a genomic sequence analysis step-by-step by using online genomic tools and databases. Each student was then assigned a project to work on. The students were required to finish the project in two weeks. They could seek help from the instructor and a teaching assistant.

A short summary from the precourse survey is provided below.

Seventy-six students were in this course module; 60 were undergraduate students and 16 were graduate students. Seventy-two students had taken at least one college-level biology course. Four graduate students claimed that they had taken only high-school biology courses. No students had ever taken a stand-alone genomics or genetics course before this course module. No students had done any genomic data analysis projects before. These characteristics are summarized in Table 2.

Of the 76 students who participated in the four-session genomics module, 38 students completely finished their assigned genomic sequence analysis projects, 8 students finished their projects but did not completely finish the required project reports, 4 students devoted significant efforts to their projects but did not completely finish them, 4 students worked together and finished one project, two 6-student groups worked independently and finished two projects, 8 students chose not to

work on the assigned projects because of schedule conflicts or unknown reasons, and 2 students could not attend the lab sessions and chose to work on literature surveys of related topics.

In summary, most students (66 of 76, or 86.8 percent) worked on their assigned projects and finished them partially or completely, either individually or in groups (see Table 3). This output was actually better than the instructor's expectation. After all, many of these students (32 of 76, or 47.4 percent) had no prior knowledge about genomic tools, databases, or even terminology. However, after the two lectures and two labs, they could perform genomic data analysis using online tools and databases. In this case, simple recall from memory was not sufficient. The students needed to understand the procedure and the reasoning behind the steps in order to finish the projects.

Note that the annotation projects used in this module can be done individually or in groups because each project includes multiple genes (six genes on average) and each gene has multiple exons (seven exons on average). The procedure for determining each of those genes and exons is the same or highly similar. Therefore, students can become familiar with the genomic data analysis procedure by working on several exons in a gene or multiple genes. Even when multiple students work on a project as a group, each student is required to work on individual exons and genes. The difference between individual and group projects is that students who work individually have more practice on the procedure and thereby build a more solid foundation in this field.

Computational Genomics Course

Integrating research components into undergraduate education has been shown to be an effective way of training active learners.[17, 18] As mentioned earlier, in our stand-alone computational genomics course, we integrated research components into undergraduate teaching and evaluated students according to their performance in their research projects.

In Spring 2011 and 2012, 16 students (12 undergraduates and 4 graduates) were enrolled in the computational genomics course. We performed a precourse survey to figure out these students' background. Fourteen students had taken at least one college-level biology course before taking this genomics course. None of them had ever taken a genomics or genetics course. None of them had taken the two-session or four-session genomics course module described earlier.

The instructor created research projects for each of these students. Each student was required to work on one genome segment (the length of the segment was roughly 40,000 base pairs) and figure out genetic features of the segment, including the gene structures, types and locations of genetic repeats, ortholog genes in other species, evolutionary analysis, and the genes already used in genetic disease research. Each student was also required to perform an investigation on a chosen genetic disease and to deliver a presentation to the class. All 16 students successfully finished and defended their research projects. Their research results were submitted to the GEP database and will eventually be submitted to FlyBase and the NCBI so that they can be accessed by other

researchers.

At the end of the course, we conducted another brief survey to collect students' opinions on integrating research into college education and to determine the number of hours they worked on their research projects. On average, these students worked four to six hours per week and spent four weeks on their projects. They all believed that working on research projects made them more active learners. The instructor and the teaching assistant could only provide general guidance, explanation on research methodology, and some examples to them. They themselves had to understand the course materials and figure out reasonable solutions for their research questions. At the end of the semester, a few students in the class expressed interest in working with the instructor on a larger-scale genomics research project.

Other evaluation methods, such as interviews, questionnaires, focus groups, and so forth, may be applied in the future implementations of this course. These evaluation results might be helpful for further improvement of the course.

Discussion

In this article, we describe the flexible approaches we created to deliver computational genomics knowledge and skills to HIM students. We also evaluated each approach with different methods to determine whether it was an effective way of teaching genomics.

We did not perform an evaluation of the tutorial session because it was a brief tutorial, and even a simple survey would have taken a significant portion of the allocated time. It was our hope that students who attended this tutorial could be motivated to pay more attention to topics related to genomics when they read daily news or come across other articles on genomics. One postcourse survey of these alumni might be helpful to determine if the tutorial has this intended effect.

The evaluation of the two-session course module was not very sophisticated. After all, in one two-hour lecture and one two-hour lab in a single week, we could not include many concepts and data analysis tools. Only the very basic and critically important ones were selected. The pre- and postcourse surveys were conducted within a short period of time, and therefore students still had fresh memory of the materials and could easily recall them even if they did not fully understand the concepts and the online tools. Even so, the evaluation results might still be helpful as a basis for decisions about future course offerings. For instance, college-level biology courses might not be required before students take this genomics course module. After all, courses such as General Biology or Biology 101 typically would not spend much time on genetics concepts and would be unlikely to mention the genomics data analysis tools that were the focus of this course module. Students could make significant progress on this topic if we took the time to explain the concepts to them and guided students to work on real genomic data with online tools.

The evaluations of the four-session course module and stand-alone course were more detailed.

Students in these two classes were evaluated on their performance in research projects. They could not find existing solutions to their projects or obtain quick answers from their teaching assistant or instructor. After all, they worked on newly sequenced and finished genomes, and no one had the correct answer before the students figured it out. Therefore, these students had to fully understand the concepts introduced in the classes and completely grasp the required genomic data analysis skills to finish their projects. Students who could independently finish their projects certainly had developed the desired knowledge and skills. One can be fairly certain that these students obtained strong confidence in dealing with genomic data from this research project. In fact, three students (two undergraduate students and one graduate student) approached the instructor to work on more genomic projects after the classes were over. The graduate student annotated several more genome segments; one undergraduate student performed extensive evolutionary analysis on several annotated genes and their ortholog genes obtained in these two classes; and the other undergraduate student worked with the instructor for two years (2010–2012) on the mechanisms of gene intron-exon structure evolution and published two first-author research papers.[19, 20]

There were limitations in the implementation and evaluation of the teaching approaches. First, we did not have one single cohort that participated in all these levels of genomic course materials. Some students had the tutorial and the two-session course module. Some students had the two-session module and the four-session module. Some students only took the stand-alone course. It is reasonable to expect that students who go through all these levels of classes have better performance than the ones who only take some of them. We also plan to make some adjustments to the course materials so that there is not much repetitive material in the tutorial, course modules, and stand-alone course. Second, we did not have a control group available for comparison because we only had one session of each class in each year. A controlled comparison should be done for at least two groups, with one having a research component and one having only lectures, labs, and traditional homework assignments. Therefore, the results in this study cannot be used to suggest that integrating a research component in class is better than the traditional teaching approach, even though the students' self-evaluations reported that they became more active learners after working on the assigned research projects.

One possible improvement is to introduce students who have taken these classes to some nontraditional internship sites, such as DNA sequencing centers or molecular pathology labs. Students who have this type of internship experience are likely to obtain more solid training in genomics and become familiar with various application fields, which would greatly help them to be more competitive in the job market.

Conclusions

By distributing computational genomics course materials into multiple levels and integrating them into a few required courses, we successfully delivered this content to HIM students without

requiring multiple prerequisites or adding any extra burden to these students' workload.

For students who wanted to learn computational genomics systematically, we offered a stand-alone course and incorporated research components into the course. With proper arrangement of course contents according to the available class time (see Table 4), HIM students can obtain a solid background in computational genomics and gain confidence in handling genomic data. Both of these are beneficial for their future careers, in which they may be called upon to handle personal genomic data sets and provide evidence to support personalized healthcare practices.

Leming Zhou, PhD, is an assistant professor in the Department of Health Information Management at the University of Pittsburgh in Pittsburgh, PA.

Valerie Watzlaf, PhD, RHIA, FAHIMA, is an associate professor in the Department of Health Information Management at the University of Pittsburgh in Pittsburgh, PA.

Mervat Abdelhak, PhD, RHIA, FAHIMA, is the department chair and associate professor in the Department of Health Information Management at the University of Pittsburgh in Pittsburgh, PA.

Acknowledgments

This work was supported by grant IIS-0938393 from the National Science Foundation CPATH program. We thank Paul Yenerall (the teaching assistant) at the University of Pittsburgh and the Genomics Education Partnership at Washington University in St. Louis for their valuable support of this project.

Notes

1. Mardis, Elaine R. "Anticipating the $1,000 Genome." *Genome Biology* 7 (2006): 112.
2. Metzker, Michael L. "Sequencing Technologies—the Next Generation." *Nature Reviews Genetics* 11 (2010): 31–46.
3. National Human Genome Research Institute. "DNA Sequencing Costs." Available at http://www.genome.gov/sequencingcosts/ (accessed October 2, 2012).
4. Al-Ubaydli, Mohammad, and Rob Navarro. "Genomic Electronic Health Records: Opportunities and Challenges." *Genome Medicine* 1 (2009): 73.
5. Kohane, Isaac S. "Using Electronic Health Records to Drive Discovery in Disease Genomics." *Nature Reviews Genetics* 12 (2011): 417–28.
6. Mardis, Elaine R. "A Decade's Perspective on DNA Sequencing Technology." *Nature* 470 (2011): 198–203.
7. McGuire, Amy L., et al. "The Future of Personal Genomics." *Science* 317 (2007): 1687.
8. Guttmacher, Alan E., et al. "Personalized Genomic Information: Preparing for the Future of

Genetic Medicine." *Nature Reviews Genetics* 11 (2010): 161–65.
9. Benson, Dennis A., et al. "GenBank." *Nucleic Acids Research* 39 (2011): D32–D37.
10. Pruitt, Kim D., et al. "NCBI Reference Sequences (RefSeq): Current Status, New Features and Genome Annotation Policy." *Nucleic Acids Research* 40 (2012): D130–D135.
11. Rose, Peter W., et al. "The RCSB Protein Data Bank: Redesigned Web Site and Web Services." *Nucleic Acids Research* 39 (2011): D392–D401.
12. Altschul, Stephen F., et al. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (1990): 403–10.
13. Kuhn, Robert M., David Haussler, and W. James Kent. "The UCSC Genome Browser and Associated Tools." *Briefings in Bioinformatics* 14 (2013): 144–61.
14. Heger, Andreas, and Chris P. Ponting. "Evolutionary Rate Analyses of Orthologs and Paralogs from 12 Drosophila Genomes." *Genome Research* 17 (2007): 1837–49.
15. Lopatto, David, et al. "Genomics Education Partnership." *Science* 322 (2008): 684–85.
16. McQuilton, Peter, et al. "FlyBase 101: The Basics of Navigating FlyBase." *Nucleic Acids Research* 40 (2012): D706–D714.
17. Seymour, Elaine, et al. "Establishing the Benefits of Research Experiences for Undergraduates in the Sciences: First Findings from a Three-Year Study." *Science Education* 88 (2004): 493–534.
18. Shaffer, Christopher D., et al. "The Genomics Education Partnership: Successful Integration of Research into Laboratory Classes at a Diverse Group of Undergraduate Institutions." *CBE Life Sciences Education* 9 (2010): 55–69.
19. Yenerall, Paul, Bradlee Krupa, and Leming Zhou. "Mechanisms of Intron Gain and Loss in Drosophila." *BMC Evolutionary Biology* 11 (2011): 364.
20. Yenerall, Paul, and Leming Zhou. "Identifying the Mechanisms of Intron Gain: Progress and Trends." *Biology Direct* 7 (2012): 29.

Printer friendly version of this article.

Leming Zhou, PhD; Valerie Watzlaf, PhD, RHIA, FAHIMA; and Mervat Abdelhak, PhD, RHIA, FAHIMA. "Flexible Approaches for Teaching Computational Genomics in a Health Information Management Program." *Perspectives in Health Information Management* (Summer 2013): 1-13.

There are no comments yet.