International Journal of Bioengineering & Biotechnology
2017; 2(3): 13-23
http://www.openscienceonline.com/journal/ijbb

Open Science

Applied Bioinformatics for Exploring College Freshmen and High School Students

Zach Lozier, Sridhar Ramachandran*

Department of Informatics, Indiana University SE, New Albany, Indiana, USA

Email address

zlozier@ius.edu (Z. Lozier), sriramac@ius.edu (S. Ramachandran)

*Corresponding author

To cite this article

Zach Lozier, Sridhar Ramachandran. Applied Bioinformatics for Exploring College Freshmen and High School Students. *International Journal of Bioengineering & Biotechnology*. Vol. 2, No. 3, 2017, pp. 13-23.

Received: July 28, 2017; **Accepted**: November 23, 2017; **Published**: December 20, 2017

Abstract

Bioinformatics is a growing professional field characterized by the combination of skills required in the fields of biology, computer science, and information technology. Generally, bioinformatics programs offered at the undergraduate level and beyond focus on preparing their students to develop programs that will be used to analyze and organize data generated by biologists. There is an apparent disconnect between the programs being developed and biologists willing to use them. This paper describes the field of bioinformatics and attempts to establish a need for a course that teaches students how to use various tools and programs currently available for use in analyzing biological data, primarily in the form of genes and proteins. Additionally, a course design and a framework is presented that can be easily adopted by interested instructors.

Keywords

Informatics, Bioinformatics, Undergraduate, Biotechnology, Bioengineering

1. Introduction

Bioinformatics is a term used to describe the field of study that is the intersection of biology, computer science, and information technology with goals to manage and analyze data in an effort to understand and model living systems [1]. This intrinsically means that bioinformatics is a field that requires competencies in biology, computer science, and information technology. Many institutions offer programs of study in each of these areas separately. However, few universities are starting to offer specific bioinformatics programs. A search on the College Board's website for "must have" majors in bioinformatics for universities returned only 48 results[1]. Sczyrba et al. have also suggested that education in the field of bioinformatics at the undergraduate level is limited [2]. This observation is alarming considering the growing trend for fields rooted in biology to become increasingly data intensive and reliant on computational methods [3]. Bioinformatics study and training needs to expand to match the growth seen in the fields that utilize bioinformatics skills. Ideally, training should begin early to maximize the exposure of relevant materials and techniques that students will receive during their course of study. A problem often encountered is that when students choose fields to study in, most gravitate towards well known and established fields such as biology. Introducing the concepts of bioinformatics to students early on in their coursework could open them up to pursuing a bioinformatics career path sooner. A course that introduces bioinformatics to undergraduate freshmen or high school students preparing for college could serve as a tool for cultivating a better prepared professional workforce.

1.1. Defining Bioinformatics

As mentioned previously, bioinformatics is a field of study that incorporates the practices of biology, computer science, and information technology [1]. A more formal definition adopted by the National Institute of Health is the "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store,

1 https://bigfuture.collegeboard.org/college-search?major=897_Bioinformatics

organize, archive, analyze, or visualize such data" [4]. This definition highlights the broad nature of bioinformatics. It includes the development of programs to store, access, and manipulate data as well as the functional use of such programs to research and perform tasks.

Initially, bioinformatics can appear to be another name for computational biology. While the two fields do share many qualities, there is a distinction. Computational biology focuses on computational methods to address theoretical and experimental questions in biology, whereas bioinformatics focuses on making biological data useful and understandable [4]. This distinction will be critical while considering the development of a bioinformatics course. Efforts of an introductory bioinformatics course should focus on data manipulation and presentation: not necessarily on experimentally or theoretically generating new data.

1.2. Origins of Bioinformatics

Bearing in mind the conditions under which bioinformatics was realized can aid the development of a course by providing the historical context under which the principles of the field were developed. A course that maintains the principles of a field better prepares participants to function within that field.

Bioinformatics was a term originally coined by Paulien Hogeweg and Ben Hesper. The definition given by the pair was "the study of informatics processes in biotic systems," and was supposedly originally proposed in a Dutch article in 1970, but was published in a widely accessible article in 1978 [5]. It was established early on then that bioinformatics studies information: its storage, transmission, reception, analysis, and presentation. However, information in biology did not spontaneously appear in the 1970s. In a sense, it has always existed. What hadn't been prevalent was data and information that was easily quantifiable. After Alfred Sanger determined the amino acid sequence of insulin in 1951, such biological data was shown to exist although it was not easily produced or analyzed. Prior to computers, biologists would have to store, manipulate, and analyze this data by hand. Furthermore, proteins were, and still are, difficult to sequence. Proteins, while unique in structure from each other, often share many aspects: size, chemical composition, physical properties, etc. This makes them difficult to isolate and thus difficult to sequence. However, fundamentally speaking, bioinformatics can be said to have its genesis with the advent of protein sequencing.

The next big step forward for bioinformatics was the availability of computers. The algorithms and methods developed by hand could be entered into computers whereby the computers could then perform the analyses. As computers became smaller and more easily accessible, biologists began to use them more to analyze protein sequences.

Massive amounts of data did not become available for analysis until Alfred Sanger produced the first DNA sequence. DNA is easier to sequence for various reasons: fewer components (4 nucleotides versus 20 amino acids) and one macrostructure made DNA easier to isolate and read. After the process to sequence DNA was developed, biologists were able to generate data that could be processed in much higher volumes. Endeavors such as the human genome project further increased the data available for analysis and solidified the role of bioinformatics as a field devoted to the study of information within a biological system.

Returning to Hogeweg's and Hesper's original definition of bioinformatics, the principle of the field is to study the information contained and transmitted between biological systems. Protein and DNA sequences offered (relatively) easy means to quantify the data being stored and transmitted within biological systems and have been a prominent focus in bioinformatics. What can thus be gathered from the history of bioinformatics as it relates to course design is that data should be at the center of what is being taught. More specifically, an introductory course should introduce the concepts and tools required to analyze these data. Luckily, these data are widely available and easily accessible, as are many tools that can analyze the data.

1.3. Influences of Bioinformatics

As is evident in the definitions of bioinformatics offered up to this point, it is a diverse field of study that has many influences from various other fields. Understanding how each of these fields contributes to bioinformatics is a necessity for developing an effective course. Here, bioinformatics is deconstructed in such a way that the various fields that influence bioinformatics are examined in the context of designing an introductory course for applied bioinformatics.

1.4. Biology

As one could elucidate from the term, biology is at the core of bioinformatics. As bioinformatics is the study of information within a biological system, it stands to reason that understanding the biology behind the biological system is necessary. The biology component to bioinformatics provides the greater context for the information being analyzed. In essence, it gives the information being analyzed meaning. In order to successfully implement and use a tool of bioinformatics, one has to have an understanding of the systems that the information is found in.

Much of what bioinformatics consists of involves developing or using programs to analyze biological data that were produced by biologists. Currently, most bioinformatics designs tends to work in a one-way manner: programs are made and developed for analyzing biological data, but the information found within biological data does not necessarily contribute to processes required to develop or use the programs that analyzed the data. The trend is most often biologists seeking assistance from computer scientists and mathematicians to help analyze data as opposed to computer scientists and mathematicians seeking biologists' help to understand biological systems.

As it relates to the design of this course, it should be noted that some amount of biology will have to be taught to any

students who take the course. Since the course is at the introductory level, one cannot assume that every student who takes the course has a background in biology. To this end, the course will be designed such that the systems to be studied can be briefly described to students who may have no background in biology aside from a biology 101 course. Of the many topics in biology, the primary focus of this course will be those topics related to the central dogma of biology. As seen from the history of bioinformatics, topics such as genetics, genomics, and proteomics provide the most substantial amounts of data that can be analyzed computationally. Thus, an effective introductory course will be able to introduce these topics in a manner such that students enrolled in the course can develop an understanding of the topics sufficient enough to be able to understand the information being analyzed.

1.5. Computer Science and Engineering

Computer science is a field critical to the development of bioinformatics tools. At its core, computer science is the study of computing and includes various facets itself; most relevant to this course however is programming. Programming is the development of software: programs designed to accomplish specific tasks on computers. For bioinformatics, these tasks are ones that deal with analyzing biological data and transform it into useful information. This often requires a thorough understanding of mathematics, specifically in statistics. Many programs rely on statistical methods and employ various algorithms to analyze data to present information. Thus, in addition to programming competencies, a bioinformatician must be able to understand and develop algorithms that can perform various mathematical analyses on data. Not only must an individual be able to simply program, but they must be able to make the program useful.

Currently, many bioinformatics programs offer courses that involve programming, software development, and statistical analysis course work. Rarely is there a general use program or program designed for a task not related to biology that is inherently useful for analyzing biological data. And since computer scientists often do not have robust backgrounds in biology, the responsibility for designing such programs has tended to reside with bioinformaticians. Therefore, it is important for bioinformaticians to be trained in these areas. However, in the context of an introductory course intended to investigate the various applications of programs, much of the computer science related topics should not be included. The issue is that in order to develop useful programs for bioinformatics, a prerequisite is a solid understanding of programming. An introductory course such as the one being proposed cannot sufficiently teach the programming skills needed to produce useful bioinformatics programs.

1.6. Information Technology

Information technology is generally understood to refer to the development and maintenance of systems of computers to store, retrieve, and send information. Essentially, it is a field that focuses on computer networks. As it relates to bioinformatics, this field has strong relations with developing systems appropriate for storing, retrieving, and sending biological data to and from scientists. These data are often stored in databases accessible by scientists across the world.

Fortunately, the global network required for such access does not necessarily need to be developed; the Internet serves as a foundation for many such systems. Databases maintained by the National Center for Biotechnology Information and the Universal Protein Resource serve as examples of institutions that utilize the Internet as a vehicle to disseminate the data they maintain. The challenge for bioinformaticians, then, lies in utilizing the Internet and ensuring any data or information gathered can be accessed over the Internet. This includes using universal data formats, communicating, and maintaining smaller networks that assist users in accessing larger networks.

For an introductory course in applied bioinformatics, the information technology component of bioinformatics shares the same fate as the computer science portion. The context of the course outlined here cannot aptly train students in information technology. In order to be able to do so, students will need to have a background in information technology and then be able to apply the skills from that background to the challenges facing biologists.

2. Purpose for an Applied Bioinformatics Course

As has been reviewed thus far, bioinformatics is a diverse field that integrates techniques from various other fields. The primary contributing fields are biology, computer science, and information technology. Currently at Indiana University South-east (IUS), there is no offered major in bioinformatics. The most similar programs are an informatics major with a minor in biology, or a biology major with a minor in informatics. While such programs offer training in both areas, they can lack integration between the two fields and do not introduce students to specific tools or techniques often employed in the field of bioinformatics. This course could serve to bridge such a gap between the two fields of informatics and biology.

2.1. The Case for Biology Majors

Biologists today are faced with a seemingly ever increasing amount of data to analyze: so much so that traditional methods for analysis have become impractical [6]. Computational means of data analysis are becoming more prominent in the field of biological research, especially in fields such as molecular biology and genomics. However, many current biology programs do not aptly prepare its students to work with the various programs that exist to analyze such data; many programs have not really changed in the last 50 years [7].

The course designed here could serve as a pedestal for training biology majors to successfully use and implement

existing tools designed to analyze biological data. It will have the greatest resonance with molecular and genetics studies as most of the data programs are designed to analyze come from such fields. As all students seeking a biology degree from IUS are required to take courses in molecular biology and genetics, the context for the tools to be used in this course already exist.

2.2. The Case for Undecided Majors

The primary purpose of this course will be to introduce students to the tools and processes used by bioinformaticians. As such, it stands to serve as an opportunity for students who are undecided in their majors to explore what bioinformaticians do. The work load of the course will be designed and taught in a way that students with minimal backgrounds in biology or informatics can take the course and still be successful. While IUS may not offer a degree in bioinformatics specifically, the course can still serve as a point of reference for students when they are enrolling in other classes to prepare them for future careers or studies.

2.3. The Case for High School Students

Often not considered when developing introductory courses for college students is whether or not the course could be adapted as an AP class for high school seniors or juniors. In the field of bioinformatics however, this should be an important consideration. Bednarski et al. note the importance of beginning bioinformatics training early in undergraduate students' careers [8]. Since AP courses are designed to enable high school students to receive college credits for completing the courses, high school students in AP classes could be considered undergraduates. Therefore, a course such as the one to be described here could serve as a useful vector for introducing high school students to bioinformatics, allowing them to experience a bioinformatics course allowing them to better consider further studies at the college level and beyond.

Currently however, the College Board does not offer an official AP course for bioinformatics. As such, there are two possibilities for implementing this course or a variation of this course to college-bound high school students: offer the course as a dual credit course on its own or integrate this course's content with an AP biology course. The former would require a sponsoring university, but the latter could be accomplished by willing teachers.

2.4. Course Goals

The goal of this course is to introduce students to a variety of bioinformatics tools and programs that can be used to analyze biological data. The data on which this course will focus will be genomic and proteomic data; students will participate in many exercises and projects that center on analyzing sequences of DNA, RNA, and amino acids that relate to functional proteins and their structures. While important to the overall understanding of what is being accomplished, students will not need to have backgrounds in biology or informatics to be successful. The knowledge required from these areas will be taught in tandem with the techniques the students will be shown throughout the course.

2.5. Course Structure

The course is designed to be carried out over a typical 16 week semester. In an effort to maximize the time students are exposed to and use the bioinformatics tools and programs presented in the course, most of the course work is projects. A general overview of the course is outlined below.

1. Introducing key biology concepts (2-3 weeks)
   a. Central dogma
   b. DNA – structure, function, genes, "junk DNA"
   c. RNA – structure, function, mRNA
   d. Proteins – levels of structure, classes
   e. Mutations
2. Databases (2-3 weeks)
   a. NCBI
      i PubMed
      ii GenBank
      iii Entrez/Gene
   b. Ensembl
   c. UniProt
   d. Protein Data Bank (PDB)
3. Sequence Alignment (2-3 weeks)
   a. BLAST
   b. Dot Plot
   c. Global Alignment
4. Structures (1 week)
   a. Exploring known structures
      i PDB
5. Phylogenies (0.5-1 week)
   a. Introduction to how programs build phylogenies
   b. What phylogenies portray
   c. Evolution basics
6. Projects (5-11 weeks)
   a. Comparing globulins
      i Human hemoglobin to mouse hemoglobin
      ii Mouse hemoglobin subunit to mouse myoglobin subunit
   b. Wild type versus mutant
      i Human wild type hemoglobin versus malaria affected

3. Recommended Contents for an Applied Bioinformatics Course

3.1. Introducing Key Biology Concepts

As mentioned previously, this course is to serve as an introduction to bioinformatics for college freshmen or high school seniors seeking to explore bioinformatics. As such, it is not necessarily the case that students of the course will have robust backgrounds in biology, if any background at all. Depending on how this course is actually implemented in a program, prerequisites could possibly be established. However, the goal of this course is to offer an initial perspective of

bioinformatics to any interested student, regardless of background in the biological sciences. As such, the initial portion of the course includes introductions to many of the key concepts that students will need to understand in order to effectively use the tools in later course work.

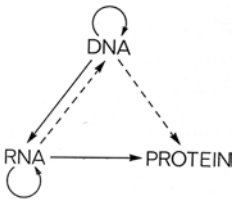

Figure 1. *Image reprinted from [9]. This depicts the central dogma adjusted for reverse transcription (RNA to DNA), replicating RNA, and DNA to amino acid translation.*

Much of the course work for this proposed course centers around what is known as the central dogma of biology. The central dogma postulates that information, specifically the information contained within DNA, RNA, and proteins, is transferred in usually one way: DNA to RNA to amino acids which become functional proteins. As more was discovered in the field of molecular biology, the central dogma was adjusted to better describe the evidence that researchers were gathering. For example, after the initial proposal of the central dogma, viruses that used RNA as a template for synthesizing DNA were discovered, a process which was not described by the central dogma at the time. By 1970 however, a more complete central dogma was developed. Figure 1 depicts the central dogma as it better describes more current evidence. Solid arrows represent typical flows of information, while dotted arrows represent possible flows of information [9]. This depiction captures the notion that RNA can be used to make DNA, DNA can be directly translated to an amino acid sequence, and RNA can replicate itself.

Ensuring that students have a firm understanding of the concepts behind the central dogma will be vital to their success. An instructor should emphasize this topic throughout the course, and be sure to include it in any lectures presented to the students. Additional topics that students should understand well and thus, should be taught, include the structure of DNA, RNA (specifically mRNA and its modifications), proteins, and genes, as well as the functions of each as they relate to the central dogma.

3.2. Databases

Being able to search, retrieve, and analyze data and information from a variety of databases is a crucial component to bioinformatics research. The databases used throughout this proposed course should be introduced early on in order to maximize students' exposure to and familiarity with these tools, as they will be referenced frequently throughout this course. The databases chosen to be used in this course consist of some of the most referenced and frequently updated databases used in bioinformatics. The information contained within these databases is heavily referenced by many publications throughout the world. It was determined that any student who continues to pursue bioinformatics would inevitably find themselves referencing these databases, so they were chosen to be used for this course.

Specific topics that should be covered during the databases topic would include how to search a database, how to prioritize results, and potentially converting file formats. Ensuring that database queries are succinct is an important skill that students should be taught. Providing too much detail in a query could negatively limit results while being too general could return too many results. Instructing students on how to use Boolean operators as well as how to appropriately use symbols to narrow or broaden results, such as those used to truncate terms or operate as a wildcard, should be a focus area during this topic. Another skill to highlight during this topic should be to show students how to skim results to determine whether or not they are relevant. Fortunately, the databases recommended for this course contain academic material such as journal articles which tend to have descriptive titles and abstracts. However, many students could potentially not know how an academic paper is structured, so this material should be taught. Students should also be taught how to convert file formats. Generally, most of the tools and databases recommended for use in this course use the FASTA format (Figure 2), but some students may encounter other formats. To prepare for this, an instructor should introduce the FASTA format and compare and contrast it to others, as well as how or where the formats can be converted.

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Figure 2. *An example of the FASTA format. The first line describes the sequence, while all other lines are the sequence.*

3.3. Sequence Alignment

One of the key techniques that is used in bioinformatics is the sequence alignment. A sequence alignment establishes a degree of similarity between a sequence of interest and other available sequences, either in a pairwise or database wide manner. This is a key step when determining the function of a particular gene or protein. The topics presented here represent the technique of sequence alignment itself as opposed to specific programs. For example, the NCBI website provides a BLAST program[2] (Figure 3) for use over

2 Screen capture from https://blast.ncbi.nlm.nih.gov/Blast.cgi on 3/28/2017.

the Internet, but it merely employs the BLAST algorithm. Other programs and websites will also accomplish a BLAST.

During this topic, the primary focus should be to differentiate between local, global, and pairwise alignments. Local alignments employ algorithms to compare specific sections of a sequence to other specific sections of another sequence, while global alignments compare the entirety of one sequence to the entirety of another. Pairwise alignments are used to compare two specific sequences. For this proposed course, students will perform local and pairwise alignments on the sequences for their projects, but introducing each of the types of alignments and having students be able to differentiate between them could assist them if they continue to pursue bioinformatics.

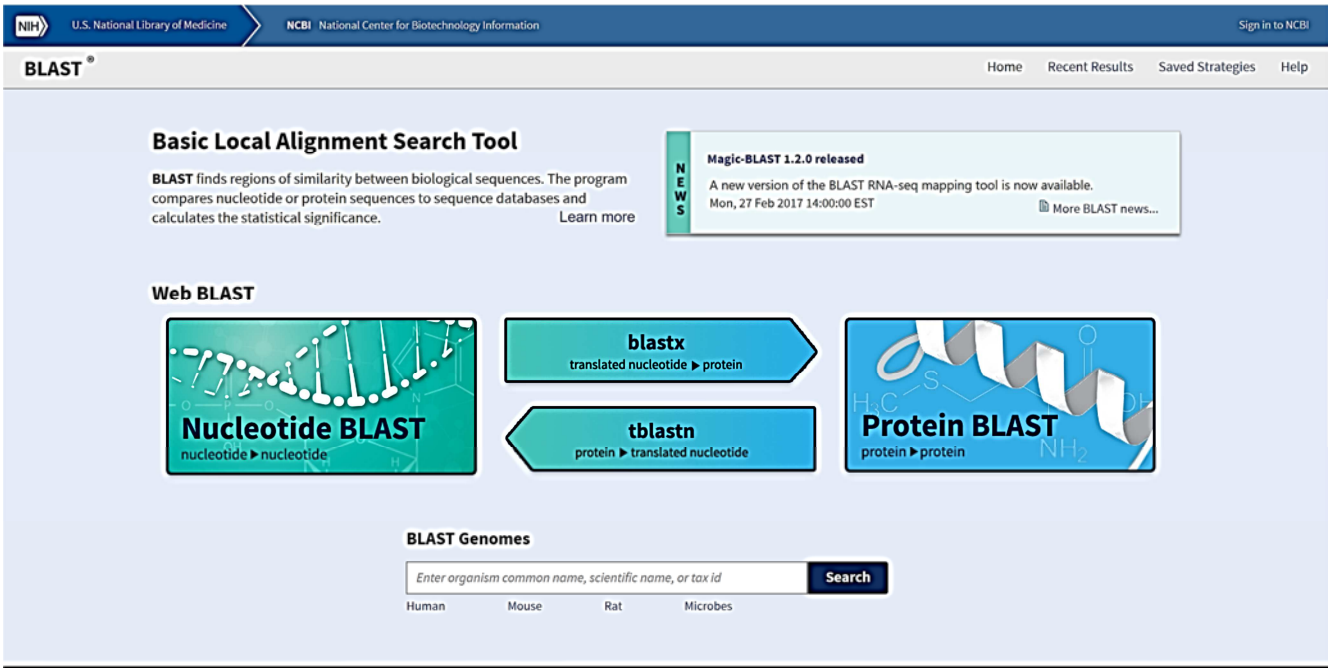

Figure 3. *Homepage of the NCBI BLAST web tool.*

3.4. Structures

Visualizing the structure of a protein is often the end goal of many bioinformatics related endeavors. If a protein's function has not been determined in a laboratory setting, the structure can often be derived via sequence comparison to offer a strong prediction of the function of the protein as form often dictates function. As such, it is important to introduce students to a tool they can use to view the structures of known proteins. What is key to note here is that this portion of the course will only enable students to visualize known protein structures. The methods for predicting protein structures from amino acid sequences are too complex and complicated to be accomplished given the introductory scope of this course.

3.5. Phylogenies

Another common objective of bioinformatics endeavors is to generate meaningful phylogenies. Phylogenies provide visual representations of the relatedness between varying classification groups of organisms. By analyzing DNA and protein sequences as well as morphological data, the evolutionary relatedness of organisms can be predicted. Developing and understanding phylogenies is an important skill to have in classifying organism to establish evolutionary relationships and has implications in developing vaccines and treatments for diseases. Developing phylogenies requires a fair understanding of advanced statistics and more advanced sequence alignments, so creating one has been excluded from this proposed course. However, students should still be given an overview of how phylogenies are created and how the process relates to bioinformatics. Phylogenies can differ in levels of complexity (Figure 4). Comparing and contrasting how information is presented in different kinds of phylogenies could offer students some insight as to how bioinformaticians determine how to present various information.

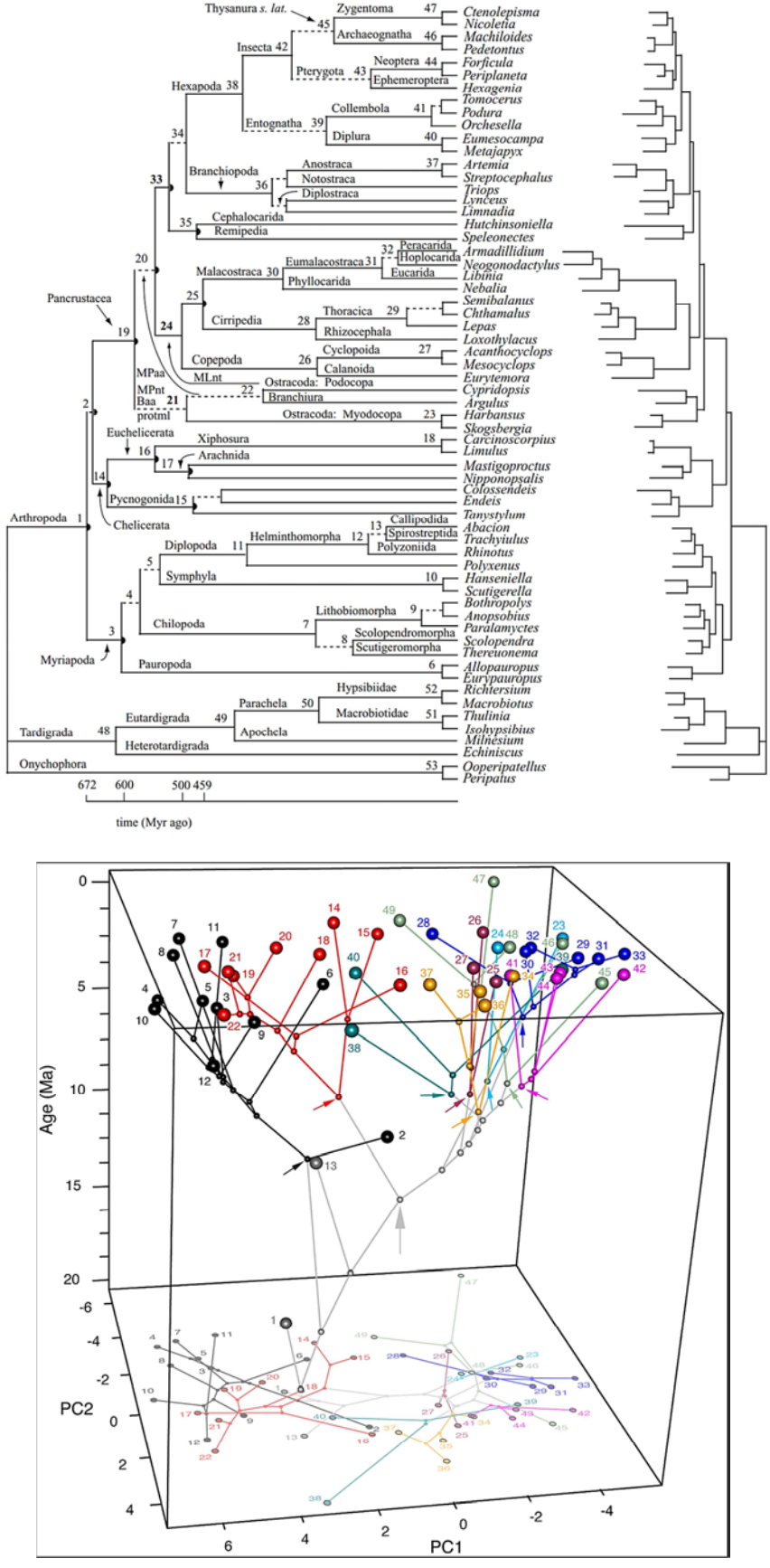

Figure 4. *Two examples of phylogenies. Top - A more traditional phylogeny, showing the relatedness of organisms based on one trait or sequence with time included (image reprinted from [10]). Bottom - A more advanced phylogeny, showing the relatedness of organism based on two traits and/or sequences against time (reprinted from [11]).*

3.6. Projects

The projects described here are intended to be accomplished successfully by any undergraduate freshman or high school senior. They are designed such that students are exposed to and work with the various tools described in the first half of the course and are still able to comprehend the biological concepts being studied. There is a danger that the course work here can become too technical for the students for which this course is designed, so any teachers, instructors, or professors presiding over this course should expect to be very involved in these projects. Additionally, because of the technical nature of the work that will be completed, the projects are meant to be completed in class. This way, a teacher or professor is able to closely monitor and offer assistance when students require it. Furthermore, the having the work completed in a classroom setting allows flexibility in delivering the information and completing the tasks. It is recommended that most of the information previously described be delivered with supplemental assignments such as worksheets and brief exercises and to assign the projects afterwards so that students are able to completely focus on the projects, but the possibility remains that some professors and teachers might want to assign the projects as the information pertaining to them is delivered in class.

Comparing Globulins: Human hemoglobin to mouse hemoglobin

The goal of this project is to demonstrate to students how sequences for proteins that accomplish similar or identical tasks between species is conserved. A similar exercise was described by Offner [12]. In the exercise described by Offner, students compared the DNA sequences of the beta hemoglobin subunits between humans and chimpanzees by implementing a BLAST analysis on a downloaded DNA sequence for the human hemoglobin beta subunit and selecting to BLAST it against only the chimpanzee genome. Offner then has students compare the sequences and answer some questions about the results. Using this exercise as a framework, this project will accomplish much of the same task: students will be asked to compare the DNA sequences of the human hemoglobin protein to that of the mouse hemoglobin protein.

It is recommended that this project be more research driven on the students' part. This differs from Offner's exercise which generally provides all of the information for the students. For this project, have students perform a PubMed search for any articles that could provide them with some back ground information on the structure of hemoglobin. Direct them towards articles that describe the protein as consisting of four subunits and that these subunits are encoded by different genes. Then, walk the students through determining which gene sequence should be used in the comparison. Although it is recommended that the alpha subunit genes be used for this project as they tend to be more conserved and stable [13], using any of the genes associated with the hemoglobin should accomplish the goal of the project. Similar to the Offner exercise, work with students to determine the introns and exons of the sequences (which can be accomplished by going to the specific GenBank entry for either specific sequence and clicking on "mRNA" – the mRNA sequence will become highlighted) so that they can compare similarities and differences between the two DNA types.

```
Query  30    ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAA  89
             || |||||| | |  ||||||||||  ||||   ||||||||||| |||   |  |||||
Sbjct  346   ACACTTCTGATTCTGACAGACTCAGGAAGAA---ACCATGGTGCTCTCTGGGGAAGACAA  402

Query  90    GACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGA  149
              | |||| ||||||| |||||||| ||| | || |  || | ||  || ||||| || ||
Sbjct  403   AAGCAACATCAAGGCTGCCTGGGGGAAGATTGGTGGCCATGGTGCTGAATATGGAGCTGA  462

Query  150   GGCCCTGGAGAGGTGAG----GCTCCCTCCCCTGCTCCGACC-CGGGCTCC----TCGCC  200
              |||||||| |||||||    |  || |   |||    || | | || |||    | | |
Sbjct  463   AGCCCTGGAAAGGTGAGAACAGGACCTTGATCTGTAAGGATCACAGGATCCAATATGGAC  522

Query  201   C--GCCCGGACCCACAGGCCACCCTC-AACCGTCCTGGCCCCGGACCCAAACCCCACCCC  257
             |  || |   | ||  || || | || |||  | ||   |   ||||  |||  | |  |
Sbjct  523   CTGGCACTCGCTCAGTGGGCAGCTTCTAACTATGCTTTTCTGTGACCTCAACTTCTCTTC  582

Query  258   TCACTCTGCTTCTCCCCGCAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCC  317
             || |    ||||||||   ||||||||      ||||||||||||||||||||||||||
Sbjct  583   TCTC----CTTCTCCC---AGGATGTTTGCTAGCTTCCCCACCACCAAGACCTACTTCCC  635

Query  318   GCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGA  377
              ||||| ||  | |||||||||||||||||||| ||||| |||||||||||||| |||||
Sbjct  636   TCACTTTGATGTAAGCCACGGCTCTGCCCAGGTCAAGGGTCACGGCAAGAAGGTCGCCGA  695
```

Figure 5. *Section of a nucleotide sequence alignment between the alpha subunit of hemoglobin for a human (Query) and a house mouse (Subject).*

In addition to DNA sequence, have students compare the amino acid sequences in a similar manner (Figure 6). Have the students browse through the Uniprot database to locate the amino acid sequences and perform another BLAST with the amino acid sequences related to the DNA sequence of the subunit the class determined to use. Students can then visually examine the structures of the two subunits by searching for their structures on the PDB.

```
Score          Expect  Method                        Identities      Positives       Gaps
253 bits(645)  3e-88   Compositional matrix adjust.  122/142(86%)   131/142(92%)    0/142(0%)

Query  1    MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG  60
            MVLS  DK+N+KAAWGK+G H  EYGAEALERMF SFPTTKTYFPHFD+SHGSAQVKGHG
Sbjct  1    MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG  60

Query  61   KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP  120
            KKVADAL +A  H+DD+P ALSALSDLHAHKLRVDPVNFKLLSHCLLVTLA+H PA+FTP
Sbjct  61   KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP  120

Query  121  AVHASLDKFLASVSTVLTSKYR  142
            AVHASLDKFLASVSTVLTSKYR
Sbjct  121  AVHASLDKFLASVSTVLTSKYR  142
```

Figure 6. *Section of an amino acid sequence alignment between the alpha subunit of hemoglobin for a human (Query) and a house mouse (Subject).*

A formal lab report is suggested for this project. Writing a formal lab report will be a good introduction to the research writing process encountered by many bioinformaticians and will provide an excellent experience should students choose to continue their education in the field of biology or bioinformatics. A key question that the teacher or professor should have students consider is having the students consider why the sequences are similar. Answers should include references to the fact that the protein accomplishes the same important task in both species and that any deviation in this sequence could prove fatal. Additionally, have students address the question of how the DNA sequences can be more different than the amino acid sequences. Answers to this question should include a discussion on how DNA encodes for proteins using codons and that many codons encode for the same amino acids. Since the target audience is high school seniors and college freshmen, a teacher or professor may consider guiding the students as they choose references and offer a concrete rubric and example for the report.

Comparing Globulins: Mouse hemoglobin to mouse myoglobin

This project very closely mirrors the previously described project. The difference lies in the goal of the project. This project aims to demonstrate how proteins with similar tasks show similarities within the same species. To illustrate, this project asks students to perform an exercise similar to the previous exercise, except they will compare the DNA and amino acid sequences of a mouse hemoglobin subunit and a myoglobin protein.

```
c77050668-770  32892 GACCCTGGATAAGTTTGACAAGTTCAAGAACTTGAAGTCAGAGGAAGATA  32941
                     |||.||       |.|||    |||       |||   ||||
32283672-3228      1 GACACT-------TCTGA----TTC-------TGA---CAGA--------     21

c77050668-770  32942 TGAAGGGCTCAGAGGACCTGAAG-AAGCATGGTTGCACCGTGCTCACAGC  32990
                           |||||       |||| ||.||||        |||||     |
32283672-3228     22 -------CTCAG-------GAAGAAACCATG--------GTGCT-----C     44

c77050668-770  32991 CCTGGGTACCATCCTGAAGA-AGAAGGGACAACAT----GCTGCC-----  33030
                     .|||||         ||||| |.|||   ||||||    ||||||
32283672-3228     45 TCTGGG---------GAAGACAAAAG---CAACATCAAGGCTGCCTGGGG     82

c77050668-770  33031 -GAGAT------CCAGCCTCTAGCCCAATCACACGCCACCAAGCACAAGA  33073
                      .||||      |||   |...||..|||        |...|||..|||
32283672-3228     83 GAAGATTGGTGGCCA---TGGTGCTGAAT--------ATGGAGCTGAAG-    120

c77050668-770  33074 TCCC------GGTCA----AGTACCTGGA--GGTAGGCGGCCACAGCAAG  33111
                      |||      |||.|    ||.||||.||  .|||.| |..|||||.|
32283672-3228    121 -CCCTGGAAAGGTGAGAACAGGACCTTGATCTGTAAG-GATCACAGGA--    166

c77050668-770  33112 TCTCCAGGGCAGAGATATAAATCCCAGCTTAGCCACTCAATACGAGTGGC  33161
                       ||||        ||||..|.|       .|.|||||..|  .|||||.
32283672-3228    167 --TCCA--------ATATGGACC-------TGGCACTCGCT--CAGTGGG    197

c77050668-770  33162 CTGCTTTCTCCCCACTAAGC-TTTCT-----CCCCAGTTCTCCTCTATAA  33205
                     |.||||    |..|||||.|| |||||     |.|.|.|||||.|||.|
32283672-3228    198 CAGCTT----CTAACTATGCTTTTCTGTGACCTCAACTTCTCTTCTCT--    241

c77050668-770  33206 CCTAC-CCCAGCCCTGTTCCCCTGAGGGTGCTGAGGAGCCACACAGCATC  33254
                     |||.| ||||| ..||||          ||||            |||.||
32283672-3228    242 CCTTCTCCCAG-GATGTT----------TGCT------------AGCTTC    268


HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHF---     47
                     |.||..:...:...|||:....|.:|.|.|..:|.:.|.|...|..|
MYG_MOUSE          1 MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNL     50

HBA_MOUSE         48 ---DVSHGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRV     94
                        :...||..:|.||..|..||.:...........:..|:..||.|.::
MYG_MOUSE         51 KSEEDMKGSEDLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKI    100

HBA_MOUSE         95 DPVNF-KLLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142
                     ||.: :.:|..:.:..|...|..||......::.|.|......:.:||:
MYG_MOUSE        101 -PVKYLEFISEIIIEVLKKRHSGDFGADAQGAMSKALELFRNDIAAKYK    148
```

Figure 7. *Section of an EMBOSS Water nucleotide sequence alignment (left) and an amino acid sequence alignment (right) between the alpha subunit of hemoglobin and the myoglobin of a house mouse..*

The protocol should closely follow that of the previous globulin project. In fact, students may use some of the materials from the previous project: if they still have the DNA and amino acid sequences as well as the structure of the mouse hemoglobin subunit from the previous project, they need not spend the time searching for them again. A deviance from the last project lies in the fact the students will be performing a pairwise alignment. Instead of using the BLAST tool from the NCBI website, have students acquire the specific sequences they wish to compare and perform a pairwise alignment using the Water local alignment tool (Figure 7) available from the European Bioinformatics Institute[3] or using a similar pairwise alignment tool. The Water local alignment has options for both nucleotide and amino acid sequences, so students can perform both alignments there. In addition to the alignments, also have students view the structures of the proteins using the PDB.

One consideration to make with this project lies in the fact that hemoglobin is a tetramer with two different subunits whereas myoglobin is a dimer with two identical subunits. A teacher or professor could consider having the students compare the sequences and structure of a subunit of myoglobin to both subunits of hemoglobin and ask the students to determine which subunit myoglobin is most similar too. Alternatively, the project can be completed by comparing myoglobin[4] (Figure 8) to only one subunit of hemoglobin that the teacher or professor decides.

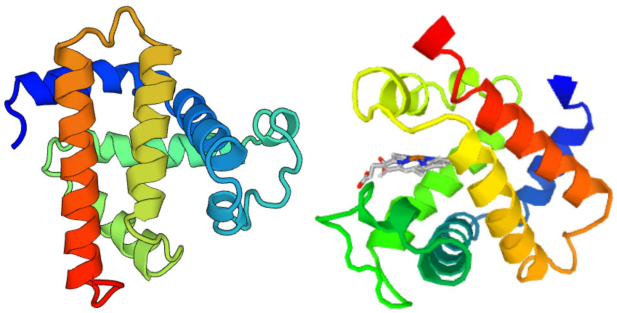

Figure 8. *3D structures of a subunit of myoglobin (left) and the alpha subunit of hemoglobin [14] (right) for a house mouse.*

Similar to the previous project, it is recommended to have students complete a lab report for this project. The formality of the lab report can be decided by the teacher or professor of the class. While a formal lab report would continue to offer research writing experience to students, a less formal lab report could be appropriate given that this project is similar to the last. However, either kind of report should discuss that, given how the amino acid sequences are fairly different, how can they accomplish such a similar task? Answers should clearly discuss that many amino acids have similar properties and both proteins have a similar shape and employ similar methods for binding oxygen using heme groups.

3 http://www. ebi.ac.uk/Tools/psa/

4 Where the myoglobin is retrieved from https://swissmodel.expasy.org/repository /uniprot//P04247 on 3/28/2017

Wild type versus mutant

The goal of this project is to demonstrate to students how a change in the sequence of amino acids of a protein can have drastic and detrimental effects. This will be accomplished by having students analyze the sequences and structures of normal, wild type hemoglobin in humans to the hemoglobin that is affected by sickle cell anemia (Figure 9). The project can be pitched as an investigative scenario by which students are presented with a description of the disease and are told to determine its cause.

A teacher or professor could have students locate the DNA and amino acid sequences of both the wild type and mutant variants, or they could be provided as part of the scenario. Another possibility is that students be provided only with the amino acid sequence of the beta subunit of hemoglobin and are told to determine if the cause of the disease can be found within it. After doing this, students should follow a similar protocol that was used in the previous project where they compared hemoglobin to myoglobin. A teacher or professor should guide the students to determine to perform a pairwise alignment between the given amino acid sequence and the normal amino acid sequence. From this, students should be able to pinpoint the difference in the two sequences. The scenario should then be structured to ask the students how this could have happened. A teacher or professor should guide the students toward analyzing the genes of the two sequences. Students should then be provided the DNA sequence of the mutated amino acid sequence and told to search for the same gene for the non-mutated amino acid sequence. To do this, students should perform a BLAST on the DNA sequence of the affected amino acid sequence and determine that the gene encodes the beta subunit of hemoglobin and that there is a mutation in the DNA that leads to the difference in amino acids. Then, have students use the PDB to look at the structures of a normal and affected hemoglobin.

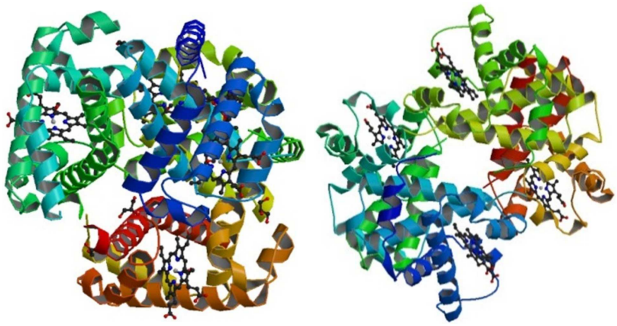

Figure 9. *Images of hemoglobin affected by sickle cell disease (left) as shown in [15] and normal hemoglobin (right) as shown in [16].*

A formal lab report or research paper is recommended for this project. Students should be asked to address several aspects of the project, but discussions should generally surround the central dogma and how genes in DNA ultimately affect proteins which determine the traits of an organism. Additionally, have students write about sickle cell disease: describing what the disease is, its negative and positive

impacts, and how this mutation ultimately leads to this disease.

4. Conclusion

The course proposed in this paper is an attempt at introducing students to bioinformatics at an early stage in their academic development. A need for such a course is supported by bioinformatics education research and researchers opinions [2], [3], [6], [7], [8], [12]. This course has been designed considering students with little biology, computer science, and information technology backgrounds with a goal of highlighting the kinds of tasks that bioinformaticians can perform. With regards to Indiana University Southeast specifically, this course has been structured to serve as a potential link between the informatics and biology programs, while concurrently serving as an exploratory elective for undecided majors. Since the university has established computer science and biology programs, collaboration between them could yield a successful implementation of this course or of a variation of this course. For high school AP students, this course could present an otherwise unknown field of study to potential undergraduate students. As bioinformatics is not an official AP course offered by the College Board, this course could instead be offered as a dual credit course in conjunction with a participating university. In this manner, Indiana University Southeast could develop this course completely and use it as a community outreach opportunity by working with local high schools to offer a dual credit course.

Prior to any actual implementation though, several pilot studies are recommended to be conducted. A formal set of unit and lesson plans should be developed and distributed to instructors and teachers willing to participate in a study. Data on student performance, satisfaction, and future decisions/success (for example, number of students who formally pursue bioinformatics) are all metrics that could be used to gauge the effectiveness of this course. Similarly, data should be gathered from the teachers/instructors who taught the course lessons. These data will potentially be opinion based, such as whether or not the instructor believed the content of the course to be relevant and whether or not they felt that students were able to understand the material could be used to determine the success or effectiveness of this course.

Pending results from pilot studies, the content of this course proposal should be adjusted to address the needs of the students and instructors. Regardless of whether or not a course is presented exactly as outlined in this proposal, a similar course that introduces bioinformatics in an applied manner should be seriously considered in an effort to address the educational needs of the fields of bioinformatics, biology, bioengineering and biotechnology.

References

[1] US Department of Energy Human Genome Program. 2003. Genomics and Its Impact on Science and Society. http://web.ornl.gov/sci/techresources/Human_Genome/publicat/primer2001/primer11.pdf.

[2] Sczyrba A., Konermann S., and Giegrach R. 2008. Two interactive Bioinformatics courses at the Bielefeld University Bioinformatics Server. *Briefings in Bioinformatics*. 9 (3), pp. 243-249.

[3] Iratxeta-Perez, C., Andrade-Navarro, M. A., and Wren, J. D. 2006. Evolving research trends in bioinformatics. *Briefings in Bioinformatics*. 8 (2), pp. 88-95.

[4] National Institute of Health. 2000. *NIH Working Definition of Bioinformatics and Computational Biology*.

[5] Hogeweg, P. 2011. The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology*. 7 (3), e1002021.

[6] Burhans, D. T., DeJongh, M., Doom, T. E., and LeBlanc, M. 2004. Bioinformatics in the Undergraduate Curriculum: Opportunities for Computer Science Educators. In *Proceedings of the 35th SIGCSE technical symposium on Computer science education* (SIGCSE '04). ACM, New York, NY, USA, 229-230. DOI=http://dx.doi.org/10.1145/971300.971381.

[7] Pevzner, P. and Shamir R. 2009. Computing has Changed Biology – Biology Education Must Catch Up. *Science*. 325 pp. 541-542.

[8] Bednarski, A. E., Elgin, S. C. R., and Pakrasi, H. B. 2005. An Inquiry into Protein Structure and Genetic Disease: Introducing Undergraduates to Bioinformatics in a Large Introductory Course. *Cell Biology Education*. 4 (3). pp. 207-220.

[9] Crick, F. 1970. The Central Dogma of Molecular Biology. *Nature*. 227. pp. 561-563.

[10] Reiger, J. C., Shultz, J. W., and Kambic, R. E. 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Procedings of the Royal Society B*. 272. pp. 395-401.

[11] Sakamoto, M. and Ruta, M. 2012. Convergence and Divergence in the Evolution of Cat Skulls: Temporal and Spatial Patterns of Morphological Diversity. *PLoS ONE*. 7 (7). e39752.

[12] Offner, S. 2010. Using the NCBI Genome Databases to Compare the Genes for Human & Chimpanzee Beta Hemoglobin. *The American Biology Teacher*. 72 (4) pp. 252-256.

[13] Hardison, R. C. 2012. Evolution of Hemoglobin and its Genes. *Cold Spring Harbor Perspectives in Medicine*. 2 (12) a011627.

[14] Sundaresan, S. S., Ramesh, P., and Ponnuswamy, M. N. 2009. Crystal Structure of hemoglobin from mouse (*Mus musculus*) and 2.8. To be published. doi: 10.2210/pdb3hrw/pdb.

[15] Oksenberg, D., Dufu, K., Patel, M. P., Chuang, C., Li, Z., Xu, Q., Silva-Garcia, A., Zhou, C., Hutchaleelaha, A., Patskovska, L., Patskovsky, Y., Almo, S. C., Sinha, U., Metcalf, B. W., Archer, D. R. 2016. GBT440 increases haemoglobin oxygen affinity, reduces sickling and prolongs RBS half-life in a murine model of sickle cell disease. *British Journal of Haematology*. 175 (1) pp. 141-153.

[16] Khoshouei, M., Radjainia M., Baumeister, W., and Danev, R. 2016. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *bioRxiv*. Preprint.