

PRACTICAL BIOINFORMATICS:
A COURSE FOR THE LIFE SCIENCES

HONORS THESIS

Presented to the Honors College of
Texas State University
in Partial Fulfillment
of the Requirements

for Graduation in the Honors College

by

Cody Anthony Hernandez

San Marcos, Texas
May 2015

PRACTICAL BIOINFORMATICS:
A COURSE FOR THE LIFE SCIENCES

Thesis Supervisor:

Kevin Lewis, Ph.D.
Department of Chemistry and Biochemistry

Approved:

Heather C. Galloway, Ph.D.
Dean, Honors College

COPYRIGHT

by

Cody Anthony Hernandez

2015

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

This work is protected by the Copyright Laws of the United States (Public Law 94-553, section 107). Consistent with fair use as defined in the Copyright Laws, brief quotations from this material are allowed with proper acknowledgment. Use of this material for financial gain without the author's express written permission is not allowed.

Duplication Permission

As the copyright holder of this work I, Cody Anthony Hernandez, authorize duplication of this work, in whole or in part, for educational or scholarly purposes only.

ACKNOWLEDGEMENTS

I would like to thank my committee for their continued support and encouragement: Dr. Kevin Lewis, my committee chair; and Dr. Heather C. Galloway. I am grateful to Dr. Lewis for his lectures over related topics, training in scientific writing, and mentorship through this project. I am also thankful to Dr. Galloway for her mentorship and encouragement over the past couple of years. In addition, I would like to thank Dr. Karen A. Lewis for her mentorship and discussions throughout this project.

Thank you to my parents: Mr. Tony A. Hernandez and Ms. Suzanne M. Rider. The countless times you have found a way to support me is far beyond what words can emulate. This journey has not been an easy one but it has been one well worth living and for that I am thankful. I would also like to thank Laura E. Pellerito for always sticking by me and encouraging me to continue pursuing my dreams, regardless of how unrealistic they seemed at the time. I greatly value your friendship and your belief in me over the past few years. I would also like to thank all of my family and friends whose support throughout my undergraduate career has been invaluable, especially John Hernandez, Eddie and Luna Hernandez, and Ann Berry. Without you my participation in various conferences and courses would not have been possible.

Finally, I would like to thank the two most influential people in my science career, David P. Anguiano and my research advisor at Trinity University Dr. Corina Maeder. To David– thank you for the training, encouragement, and countless discussions. Your relentless work ethic, patience, and passion for science has driven me to become who I am and for that I am forever grateful. To Corina– thank you for always believing in my abilities and never giving up on me. You allowed me the freedom to explore and the guidance to recover whenever I faltered. You pushed me to become the scientist that I am; you taught me that the sky is the limit and that we all are capable of accomplishing amazing things, you are my biggest fan. I hope that one day I can pay forward your mentorship.

TABLE OF CONTENTS

	Page
LIST OF TABLES.....	III
LIST OF FIGURE.....	III
ABSTRACT.....	V

	Page
CHAPTER 1:INTRODUCTION TO BIOINFORMATICS	1
1.1 Bioinformatics overview.....	1
1.2 The history of bioinformatics	2
1.3 Introducing the central dogma	3
1.4 Applied example of bioinformatics	4
1.5 An overview of large bioinformatics databases	6
1.5.1 National Center for Biotechnology Information (NCBI).....	7
1.5.2 Basic Local Alignment Search Tool (BLAST)	8
1.5.3 Gene Ontology Consortium (GOC).....	9
1.5.4 Saccharomyces Genome Database (SGD).....	10
1.6 Interpreting and organizing sequencing data.....	10
1.6.1 4Peaks.....	10
1.6.2 A Plasmid Editor (Ape).....	11
1.6.3 Serial Cloner.....	12
1.6.4 Multiple sequence comparison by log-expectation (MUSCLE).....	13
CHAPTER 2:THE STRUCTURE OF A COURSE IN PRACTICAL BIOINFORMATICS.....	15
Figure 2.1. Practical bioinformatics course outline.	15

2.1	A. Introduction to bioinformatics (~3-4 weeks)	15
2.1.1	NCBI exercise	16
2.1.2	SGD exercise	17
2.1.3	ExPASy exercise	17
2.1.4	Secondary Structure Characterization	18
2.2	B. Student projects and presentations (~6-8 weeks)	19
2.2.1	Student projects overview	20
2.3	An overview of molecular cloning	23
2.4	Project 1: Molecular Cloning	26
2.5	Assembling a plasmid tutorial	27
2.5.1	Picking the correct expression vector	27
2.5.2	Cloning software	28
2.5.3	ApE	29
2.5.4	Determining multiple cloning sites (MCS)	31
2.5.5	Helping students organize their project	32
2.6	Project 2: Structural characterization	32
2.7	Exploring Insulin using the Protein Database	33
2.7.1	Determining the structure	34
2.7.2	Determining the functional properties	34
	CHAPTER 3: EXPLORING THE CURRICULUM	36
	REFERENCES	38

LIST OF TABLES

	Page
TABLE 1.1. VARIOUS FIELDS UTILIZING BIOINFORMATICS	2
TABLE 1.2. STREAMLINED SEARCH ENGINES AVAILABLE ON NCBI.....	8
TABLE 2.1. EXPASY SEQUENCING ALIGNMENT PROGRAMS.	18
TABLE 2.2. SECONDARY STRUCTURAL CHARACTERIZATION PROGRAMS.....	19
TABLE 2.3. THE FEATURES OF SERIAL CLONER AND APE	29
TABLE 2.4. UNIQUE FEATURES OF THE APE PROGRAM.	30
TABLE 2.5. THE UNIQUE FEATURES OF THE SERIAL CLONER PROGRAM	31

LIST OF FIGURES

	Page
FIGURE 1.1. THE GROWTH OF GENBANK.....	3
FIGURE 1.2. THE CENTRAL DOGMA OF MOLECULAR BIOLOGY	4
FIGURE 1.3. GENOME WIDE ANALYSIS USING A DNA MICROARRAY.....	6
FIGURE 1.4. THE GENE ONTOLOGY CONSORTIUM (GOC).	9
FIGURE 1.5. THE 4PEAK'S INTERFACE	11
FIGURE 1.6. THE APE INTERFACE	12
FIGURE 1.7. THE SERIAL CLONER INTERFACE.	13
FIGURE 1.8. THE MUSCLE INTERFACE.....	14
FIGURE 2.2. THE POLYMERASE CHAIN REACTION (PCR)	23
FIGURE 2.3. AN OVERVIEW OF THE CLONING PROCESS	24
FIGURE 2.4. FEATURES OF THE PET15B EXPRESSION VECTOR	25
FIGURE 2.5. COMPONENTS OF A YEAST SHUTTLE VECTOR	26
FIGURE 2.6. THE ADDGENE REPOSITORY	28

FIGURE 2.7. RESTRICTION ENZYME CLEAVAGE FREQUENCY MAP IN APE	29
FIGURE 2.8. A VIRTUAL SUB-CLONING APPROACH USING SERIAL CLONER	30
FIGURE 2.9. RESTRICTION MAPPING IS SERIAL CLONER.	31
FIGURE 2.10. THE STRUCTURAL ANALYSIS OF FAST AND SLOW ACTING INSULIN	35
FIGURE 3.1. THE OUTLINE FOR A COURSE TAUGHT IN-CONCERT AND VARIOUS STUDENT RESPONSES TO AN EXIT SURVEY	38

ABSTRACT

Bioinformatics is the transformation of large amounts of data into useful knowledge that can be utilized for basic research science and biomedical applications. This field is one of the largest growing in science and is used as a tool in various areas of research. One of the main goals of bioinformaticians, aside from organizing data sets, is the normalization of datasets so that people from various disciplines can take advantage of the available data. Understanding the use and application of bioinformatics in research offers a considerable advantage to scientists. In addition, it makes post-baccalaureate students very competitive while applying for jobs or graduate school. This study specifically outlines a course in practical bioinformatics geared toward undergraduate students in the life sciences.

CHAPTER 1

INTRODUCTION TO BIOINFORMATICS

Students at Texas State University currently have an advantage over students at several other institutions due to their extensive training in laboratory applications in the life sciences. In addition to students' upper division teaching labs, students are also required to fulfill 1 year of undergraduate research in order to graduate with an American Chemical Society (ACS) certification. This training, in parallel with a course that could further polish these skills, could provide an invaluable foundation for students. Currently, very few institutions offer a course on a practical approach to bioinformatics. The majority of courses taught on bioinformatics require students to have experience with various coding languages and focus on the theory embodying the field of bioinformatics rather than the practical uses and extraction of data. A course aimed at practical approaches to bioinformatics could reinforce previous course material as well as help students with understanding future course topics.

1.1 Bioinformatics overview

The term bioinformatics, defined by Paulien Hogeweg and Ben Hesper, is “the study of informatic processes in biotic systems [1].” The definition is somewhat broad but applies specifically to assessing biological systems in a computationally intensive manner. Simply stated, bioinformatics uses mathematical and computational methods in order to help understand complex biological processes. The foundations for the complex algorithms stem from discrete mathematics, graph theory, system theory, and information theory. Some of the fields interested in bioinformatics are shown in Table 1.

Table 1.1. Various fields utilizing bioinformatics

•Antibiotic resistance	•Microbial applications	•Biological weapons defense
•Evolutionary studies	•Molecular medicine	•Improved nutritional quality
•Waste cleanup	•Personalized medicine	•Veterinary science
•Alternative energy	•Preventative medicine	•Biotechnology
•Crop improvement	•Gene therapy	•Climate change studies
•Forensic analysis	•Drug development	•Oil Industry

1.2 The history of bioinformatics

In 1972 the first gene was sequenced followed by the first genome in 1976 [2, 3]. It became evident shortly afterward that computer-assistance with analyzing, organizing, and interpreting the sequences would become necessary. The first computer-assisted analysis was completed in collaboration with several cryptologists from the National Security Agency (NSA) analyzing bacteriophages MS2 and PhiX174. Surprisingly, the first attempts to publish the findings of the computer-assisted analysis were rejected by several journals before being published in the Theoretical Journal of Biology in 1977 [4]. This method of analyzing data became known as bioinformatics, originally coined by Paulien Hogeweg and Ben Hesper in 1970 [5,6,7,8].

In 1982 GenBank, the first large bioinformatics database was created [9]. Methods for analyzing the publically available data were published immediately after and quickly gained traction. Early pioneers for the computer-based analysis of sequencing information were Margaret Oakley Dayhoff and Elvin A. Kabat. Both of their methods of analyses for protein sequence analysis would later become the template for both RNA and DNA sequence analysis. Shortly after, GenBank became an open access database available to the public consisting of various nucleotide sequences and their respective protein sequences. Since then GenBank has grown exponentially and has aided in the Whole Genome Shotgun (WGS) project that began in the early 2000's (Figure 1.1).

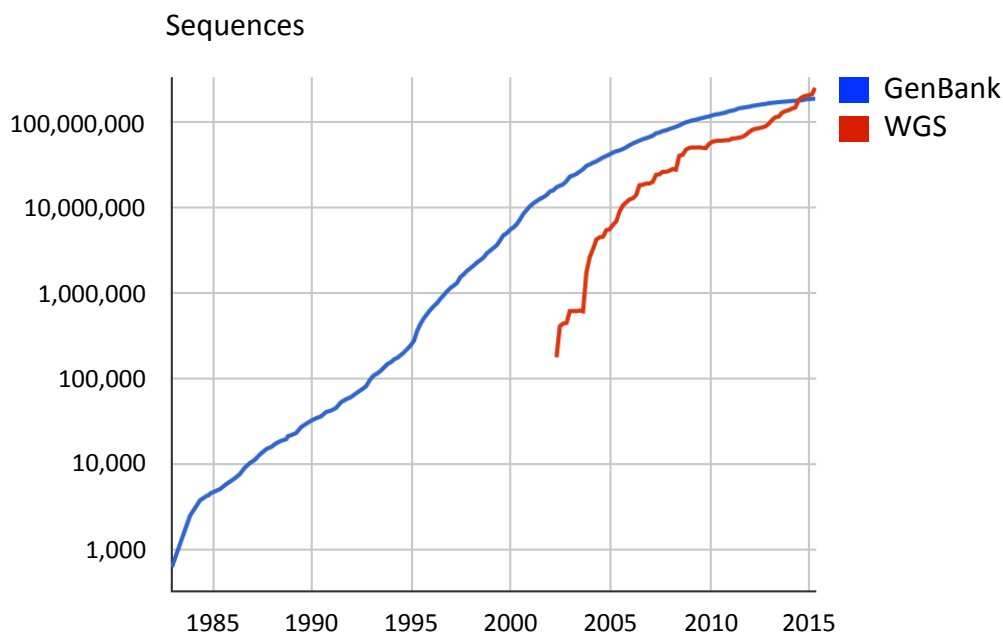


Figure 1.1. The growth of GenBank. After GenBank’s birth in 1982, the amount of sequences submitted each year has increased dramatically. This increase has also led to more complex computer based algorithms and various new forms of curation. This image was reprinted from www.ncbi.nlm.nih.gov/genbank.

1.3 Introducing the central dogma

Francis Crick first defined the central dogma in 1956 as the flow of information from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA) to Protein (Figure 1.2); implying that the flow of information from DNA was unidirectional [10]. DNA is composed of monomeric nucleotides consisting of monosaccharide deoxyribose and a phosphate group attached to a nucleobase — either Adenine (A), Cytosine (C), Guanine (G), or Thymine (T). These four monomeric nucleotides assemble into a polymer to form DNA. How these bases are arranged determines what the specific function they code for is. Information is then carried using RNA, similar in structure to DNA, to the ribosome.

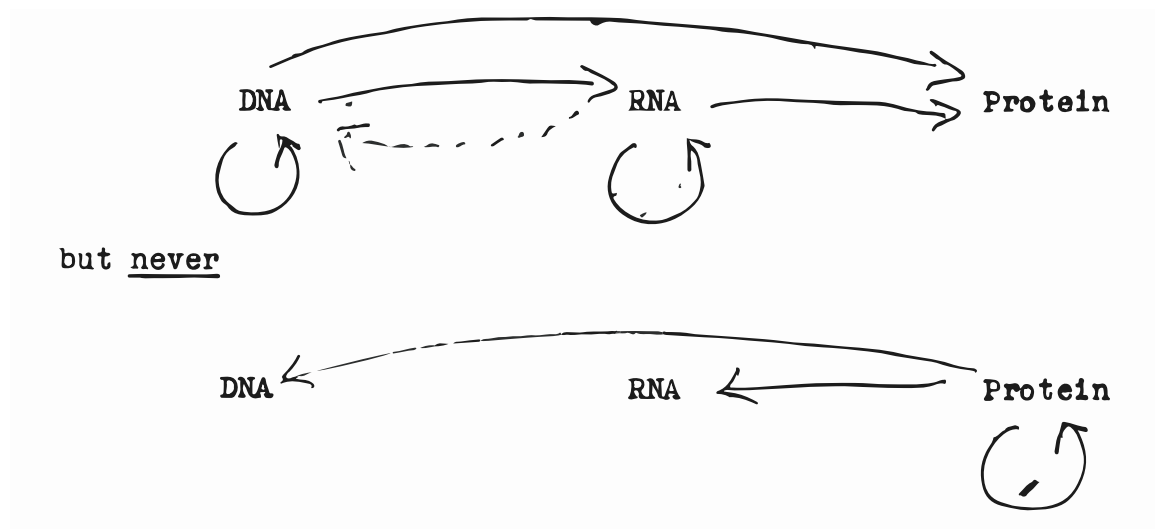


Figure 1.2. The central dogma of molecular biology. The flow of information in a biological system, the central dogma, is unidirectional. DNA is transcribed into RNA, which is then read by the ribosome and translated into protein. This image was reprinted from an unpublished acknowledgement made by Francis Crick, available at www.nih.ncbi.gov.

The ribosome is responsible for synthesizing protein, a polymer composed of the 20 essential amino acids. The sequence of amino acids in the polypeptide is determined by the sequence of the mature RNA transcript. The physical and chemical properties of the protein give rise to biological function. More specifically, each amino acid has its own physical and chemical properties and therefore how they are arranged form biochemical and biophysical clusters. These clusters can be identified using bioinformatic algorithms that computationally predict crucial regions and motifs that are necessary for function and stability of the polypeptide. This process applies for DNA and RNA as well and can be determined using sequenced based algorithms.

1.4 Applied example of bioinformatics

The development of bioinformatics has allowed for the rapid advancement of research and medicine over the past few decades. One of the largest advances has been the

development of genome wide analysis. This allows for several individuals that have a disease to be screened for the misregulation of specific genes. This provides further insight into the disease and in some cases can provide a target for gene therapy. A schematic of this is shown in Figure 1.3. The screening process begins by identifying a population of individuals with a disease, such as a specific type of cancer, and screening as many of these individuals as possible to allow for a statistically significant assessment of genes playing a role in the disease. Since the human genome consists of about 30 billion nucleotides (~30 thousand genes) this process creates an overwhelmingly large data set that can only reasonably be computed using computer-based algorithms. Without bioinformatics our ability to assess these large screens would be limited and our advances in biomedical research would be significantly hindered.

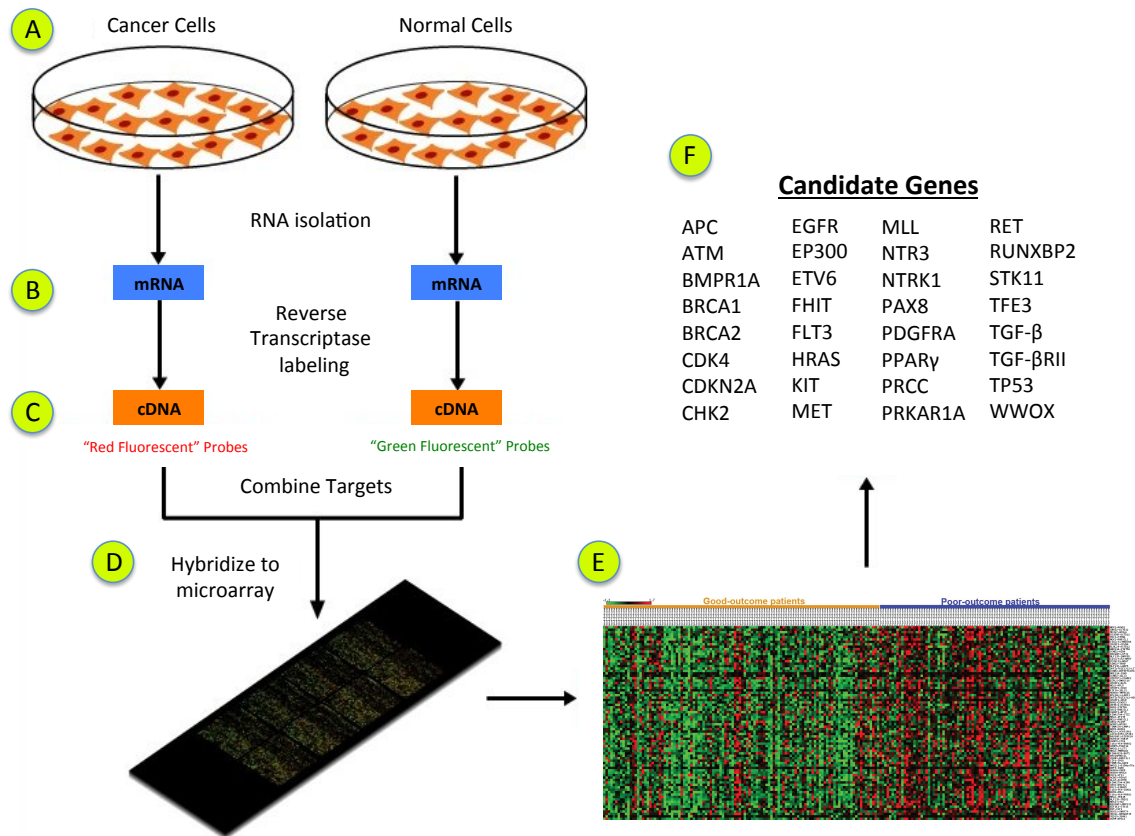


Figure 1.3. Genome wide analysis using a DNA microarray. After a population of individuals is identified that have cancer and A) their cells are extracted and cultured, their B) RNA is then isolated. Since expression is often correlated with the levels of RNA available for the ribosome the RNA can be C) reverse transcribed into chromosomal DNA (cDNA) and labeled using a fluorescent dye; this acts as a quantifiable source of RNA expressed in the cells. This is done for both cancerous and non-cancerous cells. The only difference is the fluorophores that label the cDNA. After the cancerous and non-cancerous cells have been labeled with red and green fluorophores, respectively, they are D) combined and washed over specific complementary oligos that immobilize them and allow for E) microarray analysis. Following bioinformatic analysis, a F) list of candidate genes can be suggested to play a role in the disease and begin to be targeted for potential gene therapy.

1.5 An overview of large bioinformatics databases

Most projects and questions in the life sciences are first explored and studied using bioinformatics databases before proposing a specific question. Several databases exist for these searches and have been streamlined according to independent interests. This includes

literature related to specific genes, nucleotide sequences, structural characteristics, phylogenetic relationships, and much more. The utilization and curation of these databases have become critically important for the advancement of medicine and biomedical research.

1.5.1 National Center for Biotechnology Information (NCBI)

Similar to Google, a search in the National Center for Biotechnology Information (NCBI) will generate an abundance of information associated with whatever gene or organism that one are interested in. It is the headquarters for bioinformatics information and resources in the biological sciences. A search for *Escherichia coli* (*E. coli*) generates over 20 million hits that are separated into the following sections: literature, genes, health, proteins, genomes, and chemicals. This can be overwhelming and therefore a more specific search can be done using specific search criteria such as those shown in Table 1.2. Needless to say, this is typically the starting point for biological inquiries or gaining information on a new project or question [11]. In summary, NCBI is a database of databases that provides genomic and biomedical information.

Table 1.2. Streamlined search engines available on NCBI.

Database	Analysis
CDD	Conserved Domain Database; Protein sequences conserved in evolution and alignments of known domain to 3-d structures in MMDB database.
GenBank	NIH annotated DNA sequences.
Gene	Genes components and their specific characteristics.
Genome	Whole genomes of over 1000 organisms as well as genome sequences in progress. Includes viruses and other non-living genomes.
NCBI Education Page	Information and tutorial sessions for NCBI affiliated databases.
Nucleotide Database	A collection of sequences from several databases including PDB and others.
OMIM	Online Mendelian Inheritance in Man; Human genes and disease.
Protein Database	Protein database searches GenPept, RefSeq, Swiss-Prot, PIR, PRF, and PDB.
PubMed	Biomedical literature database, provides abstracts and link to full text on PubMed Central or other websites.
SRA	Sequence Read Archive; Next generation sequencing database.
Structure	3D structure of macromolecular molecules.
Taxonomy	Phylogenetic trees and lineages for over 160,000 organisms. New taxomic information added every day.
UniGene	Transcriptome analysis.

1.5.2 Basic Local Alignment Search Tool (BLAST)

The BLAST algorithm was first published in the *Journal of Biology* in 1990 and has since evolved into a family of pipelined programs [12]. The BLAST database facilitates the search for primary biological sequences composed of either nucleotides or amino acids that are similar in sequence to a known nucleic acid or amino acid sequence. The program is useful in the mapping of DNA and subsequent identification of its corresponding species. In addition, it locates specific domains in proteins, and establishes phylogenic relationships. Finally, it compares sequences and can be used to help annotate one organism's genome or proteome in relation to another.

1.5.3 Gene Ontology Consortium (GOC)

The GOC project began in 1998 with the goal of consistently describing the products of certain genes across various databases [17]. The project originally included the FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD) but has since grown to include most of the major genome repositories for plants, animals, and microbes. The three main goals of the GOC are the following: i) documenting the ontologies, ii) annotating the ontologies, and iii) developing new software to streamline documenting and annotating ontologies. In short, the database normalizes methods for explaining the product of specific genes in a few short words and contains a side bar to help determine specific products in various biological systems. For example, a quick search for Lactate Dehydrogenase (LDH) shows the various subunits affiliated with the protein complex, a direct annotation of its function, and a filter for gene ontology (Figure 1.4). Once the filter is selected it shows different ways of interpreting the gene product along with the respective references.

The screenshot displays the GOC search interface. On the left, a 'Free-text filtering' box contains the search term 'LDH'. Below it, a sidebar lists filters: 'document_category: bioentity', 'No current user filters.', and a list of filter categories: Source, Type, PANTHER family, Taxon, Direct annotation, and Inferred annotation. The main area, titled 'Found entities', shows a table of results. The table has columns: Acc, Name, Taxon, PANTHER family, Type, Source, Direct annotation, and Synonyms. Three results are visible, all for 'L-lactate dehydrogenase' from 'Homo sapiens'.

Acc	Name	Taxon	PANTHER family	Type	Source	Direct annotation	Synonyms
LDHC	L-lactate dehydrogenase	Homo sapiens		protein	UniProtKB	L-lactate dehydrogenase activity carbohydrate metabolic process more...	G3XAP5_HUMAN LDHC hCG_15532
LDHB	L-lactate dehydrogenase	Homo sapiens		protein	UniProtKB	L-lactate dehydrogenase activity carbohydrate metabolic process more...	A8MW50_HUMAN LDHB
LDHA	L-lactate dehydrogenase A chain	Homo sapiens		protein	UniProtKB	carboxylic acid metabolic process oxidoreductase activity, acting on the CH-OH group of donors.	F5GXC7_HUMAN LDHA

Figure 1.4. The Gene Ontology Consortium (GOC). The GOC is a pipelined search method for quickly determining function, taxon, and various gene aliases. This image was screenshoted from www.geneontology.org

1.5.4 Saccharomyces Genome Database (SGD)

The *Saccharomyces* genome database provides genome specific information about *Saccharomyces cerevisiae* (*S. cerevisiae*), also known as budding yeast. The SGD utilizes a comprehensive approach to understand the *Saccharomyces* genome using general search criteria and primary sequence criteria such as nucleotide or amino acid sequences. Site curators maintain the database manually by extracting primary literature information and reporting it using phenotype annotations and gene ontology. The database provides a powerful platform for the transversion between simple eukaryotes and higher order eukaryotic organisms. Needless to say, this is a powerful tool for helping zone in on *Saccharomyces* specific information as well as primary literature associated with this model organism.

1.6 Interpreting and organizing sequencing data

The sequencing of DNA and interpretation of sequencing data has become invaluable to the life sciences and is an integral component contributing to the molecular biology revolution that has taken place over the past few decades. With the ability to sequence DNA came the need to organize and efficiently interpret the sequencing results. Often times biologists work with large sequences of DNA that can range from billions of nucleotides, as in the human genome project, to a couple of hundred nucleotides. The ability to annotate, log, and edit these sequences is important for the investigator and has been pipelined by several databases and bioinformatics tools.

1.6.1 4Peaks

The majority of sequencing data is now transferred back and forth between sequencing facilities and researchers in various file formats that must be translated from raw sequencing data into interpretable DNA sequences. One program that does this is 4Peaks

(Figure 1.5). The program can convert the most common sequencing file formats, such as ab1., scf., and ctf., to its respective nucleotide sequence. In addition, it can search for sequencing motifs (e.g. T7 promoter), translate the DNA sequence to its protein sequence, and can also do a direct BLAST search on the sequence. The program won the Apple Design Award in 2004 and two more in the years that followed.

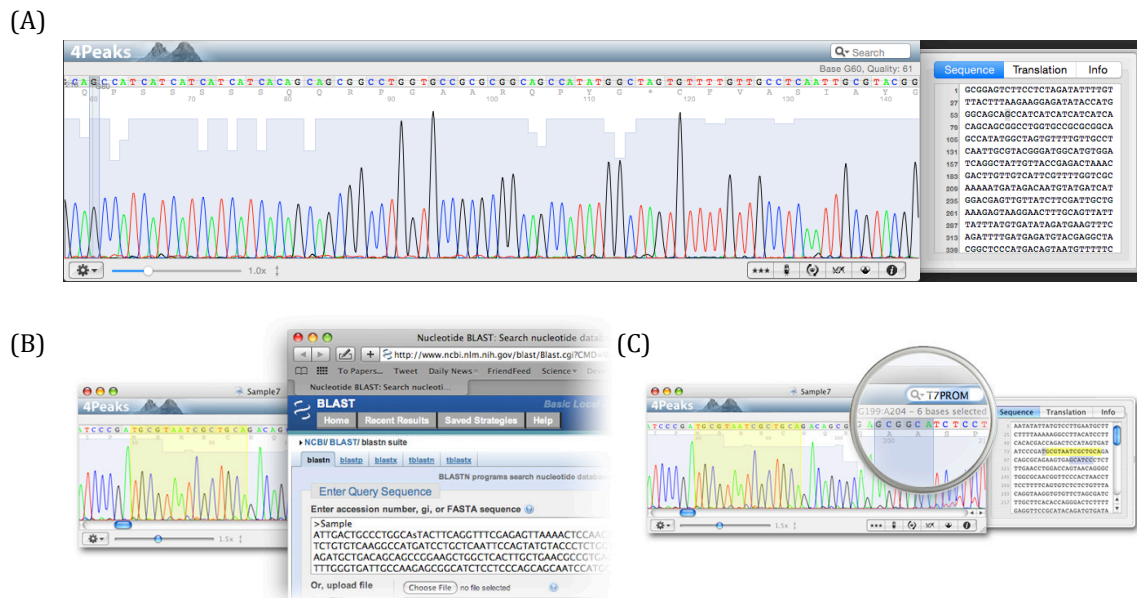


Figure 1.5. The 4Peak's interface. These images are screenshots from the 4Peaks program depicting the (A) chromatograph, (B) BLAST function, (C) search for DNA motif (e.g. T7 promoter). This image is a compilation of screenshots from the 4peaks program.

1.6.2 A Plasmid Editor (Ape)

ApE is a free bioinformatics tool used to help organize plasmid DNA sequences [14]. Similar to 4peaks, it can also read ABI sequencing, trace files, and produce a chromatograph of the sequencing spectra. The highlights of the program include the user-friendly interface that does not require prior-knowledge of coding language but can be downloaded in an open source format. The graphic map, as shown in Figure 1.6, can be

The figure displays the circular map and linear sequence of the pDONR221-1-2 destination vector. The circular map shows the following features:

- Amp^r 8900..9559**: Ampicillin resistance gene.
- Kan/neoR 7893..8684**: Kanamycin/Neomycin resistance gene.
- LacZ alpha 7072..7140**: LacZ alpha gene.
- M13-pet 7001..6984**: M13-pet gene.
- 17 6985..6958**: A small region.
- 424 (pKAS1.4 u47GFPntx)**: A 424 bp region.
- 10506 bp**: The total size of the vector.
- LacP 177..199**: Lac promoter.
- M13-rev 205..225**: M13-rev gene.
- unc-47 promoter 252..1445**: unc-47 promoter.
- transsplice acceptor 1435..1445**: Transsplice acceptor.
- unc-47 1446..1919**: unc-47 gene.
- GFP(65T) 1920..2786**: GFP(65T) gene.
- unc-54 3'UTR 2926..3341**: unc-54 3'UTR.
- unc-47 3342..5257**: unc-47 gene.

The linear sequence shows the DNA code with annotations for features like the 424 psKAS1.4 u47GFPntx, the GFP(65T) gene, and the unc-54 3'UTR. The sequence is flanked by 1482 and 4762 bp markers.

1.6.3 Serial Cloner

12

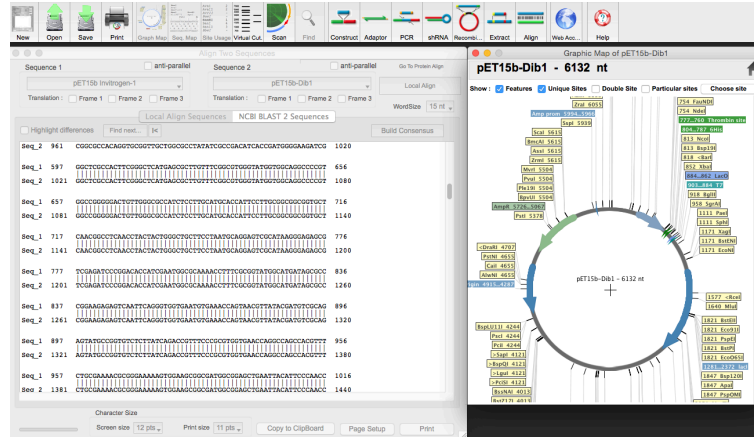


Figure 1.7. The Serial Cloner interface. These images are screenshots from the Serial Cloner program, displaying the (A) DNA sequence alignment and a (B) Vector map.

1.6.4 Multiple sequence comparison by log-expectation (MUSCLE)

After retrieving sequencing data it is often useful to compare known sequences to other known sequences of interest. A simple way to do this is by using the MUSCLE program. This program can align two or more sequences, either DNA or protein, and identify regions that are the same and regions that are different (Figure 1.7). The user has the option of doing this interactively or by email. Some of the features include a graphic representation output of the sequence alignment or a CLUSTAL format.

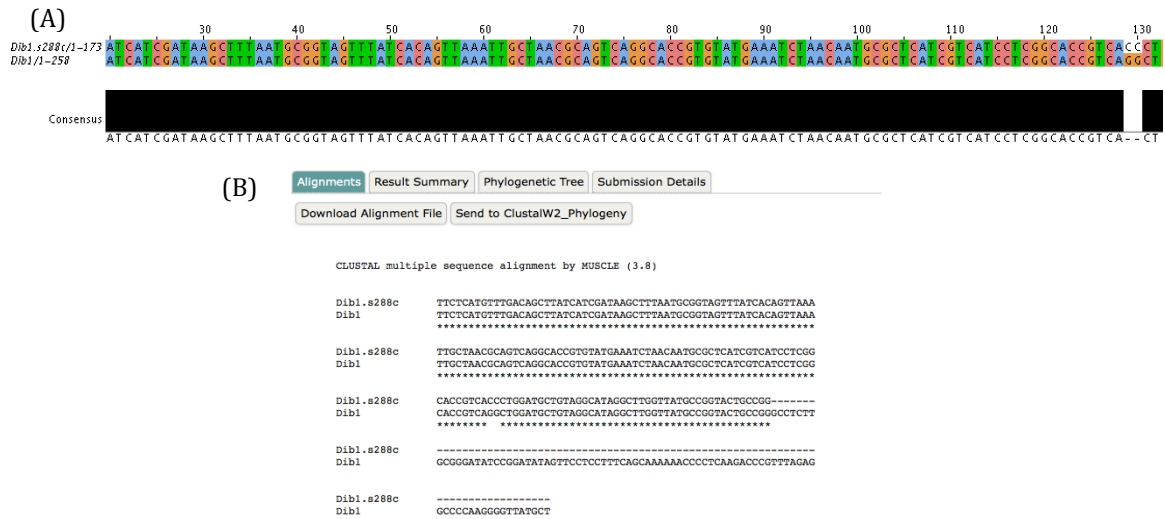


Figure 1.8. The MUSCLE interface. (A) Jalview visualization function, (B) CLUSTAL format. These images are screenshots from the MUSCLE multiple sequence alignment program.

CHAPTER 2

THE STRUCTURE OF A COURSE IN PRACTICAL BIOINFORMATICS

Needs some transition material here from ch. 1 to 2 and explain the diagram below.

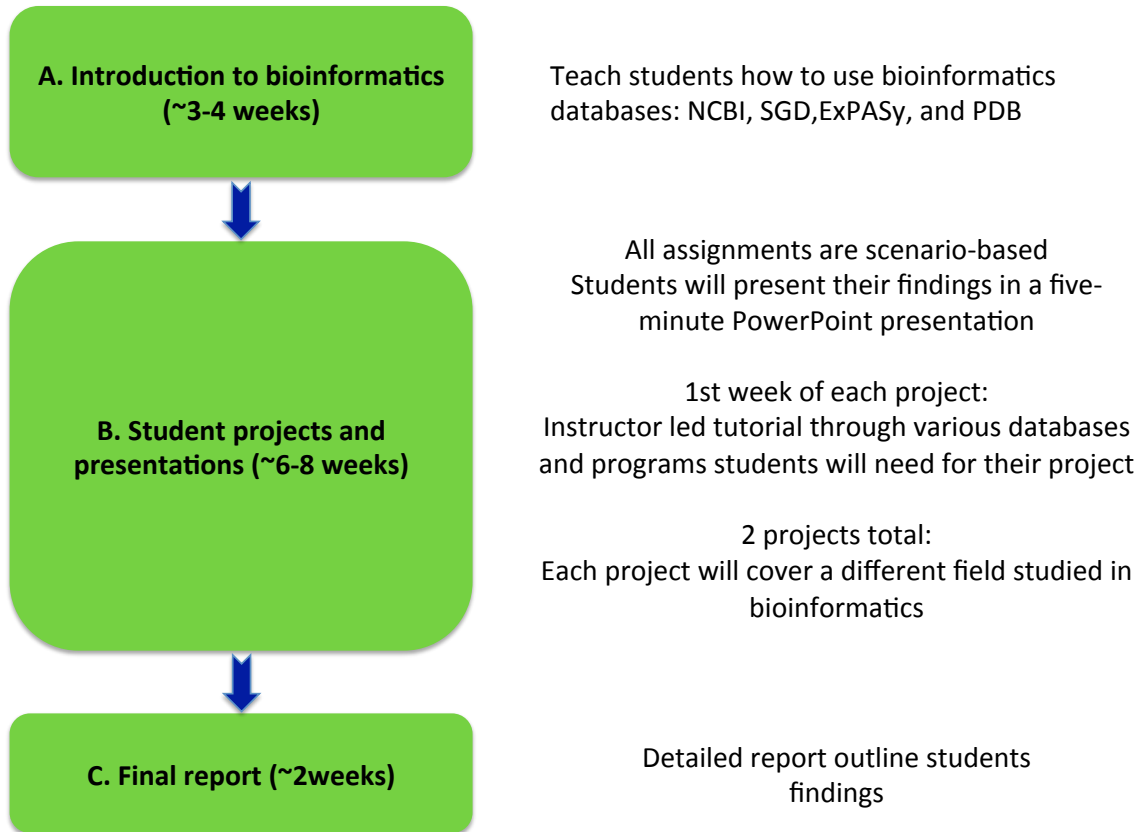


Figure 2.1. Practical bioinformatics course outline.

2.1 A. Introduction to bioinformatics (~3-4 weeks)

As previously noted, NCBI is a database of databases and ExPASy is a database of programs. Several programs and databases will be explored during the tutorial part of the student projects and therefore an introduction to NCBI, the database of databases, should be explored first to familiarize students with the headquarters of bioinformatics inquiry. In addition, it may also be useful to explore the SGD, a large but much more interactive database. This could help students understand that bioinformatics approaches can be

streamlined according to specific organisms. After students are familiar with the bioinformatics terminology the ExPASy database, the database of programs, can be explored to show students the various resources and programs available for bioinformatics analysis.

2.1.1 NCBI exercise

Exploring databases with students through a list of practical applications could be a very useful technique for exploring the database. In order to explore NCBI, one exercise could be determining what Lactate Dehydrogenase (LDH) is by using a PubMed search and scrolling through abstracts. A follow up could be retrieving the nucleotide sequence using BLAST. The last exercise could be going through the GOC to determine what the products of the LDH reaction would be. A goal for the exercises could be to have students form groups of 3 and work together to answer the following questions:

1. What is a local alignment?
2. What is a global alignment?
3. What is the difference between the two?
4. If I needed to find a specific nucleotide sequence for a gene where is one place that I could find it?
5. I have determined a specific gene of interest in humans and want to know which model organism has the highest sequence homology to it. Where could I find this?
6. If I wanted to see the product of my gene after it is translated to protein and active in the cell where would I go?

2.1.2 SGD exercise

The SGD contains a relatively friendly user-interface that allows users to access a wealth of information as well as specific information depending on the interest of the inquirer. Students could be given the following worksheet and told to answer the 10 questions using the SGD database before a formal lecture is given over the database:

1. What is the length of the amino acid sequence?
2. What is the molecular weight (KDa) of the protein?
3. What is the P.I. of the protein?
4. What is the extinction coefficient of the protein?
5. What is function of *ACT1*?
6. What other aliases does *ACT1* have?
7. Where is the location of *ACT1* on a yeast chromosome?
8. What are the subfeatures of *ACT1*?
9. What is the composition of each individual amino acid?
10. What are 4 genes that share phenotype similarities with *ACT1*?

In addition, it may be useful if the students who find each of these can show how they found them specifically. The idea behind the exercise would be to see whether or not students are beginning to build on their vocabulary from the NCBI exercise as well as too promote student engagement and participation.

2.1.3 ExPASy exercise

ExPASy contains an integrative portal that allows for streamlined approaches to answering questions using bioinformatics. One useful exercise for this database may be analyzing DNA, RNA, or protein sequences. Sequence alignment tools vary in alignment

parameters and often times offer different types of information. A pairwise sequence alignment is often used to identify specific regions of similarity between two proteins that could be linked functionally, structurally, or evolutionarily. Alternatively, a multiple sequence alignment is used to compare sequences of similar length with the goal of determining homologous relationships. For example, if an inquirer were to sequence DNA from a wild type strain and a multiple mutant strains of the same organism they would use a multiple sequence alignment tool since the DNA sequence is similar between the two. Although, if the DNA sequence is not characterized previous to the search then a pairwise sequence alignment would be utilized. As an exercise students could go through the programs listed in Table 2.1 and note their differences as well as give an example:

Table 2.1. ExPASy sequencing alignment programs.

Program	Function	Website
MUSCLE	Basic MSA	http://www.ebi.ac.uk/Tools/msa/muscle/
Strap	MSA (structure and sequence)	http://www.bioinformatics.org/strap/
T-Coffee	MSA (structure and sequence)	http://www.ebi.ac.uk/Tools/msa/tcoffee/
PRATT	Protein MSA	http://web.expasy.org/pratt/
MaxAlign	Gap removal from alignments	http://www.cbs.dtu.dk/services/MaxAlign/
WebLogo	Homology sequence Logo	http://weblogo.berkeley.edu/
WU BLAST	PSA-Local	http://www.ebi.ac.uk/Tools/sss/wublast/

After the first part of the class, students can go around the classroom and discuss their examples as well as their findings for each program.

2.1.4 Secondary Structure Characterization

In addition to sequence alignment, it can also be useful to introduce secondary structural characterization of proteins. The two most common secondary structures, alpha helices and beta sheets, can be predicted using various programs in order to help suggest

protein function and structural dynamics. In addition, several motifs exist that can also lend insight into determining biological function. For example, coiled coil protein motifs are often observed in transcriptional regulatory proteins and therefore predicting this motif provides a clue to the structural dynamics as well as the biological function of the protein. It may be useful for groups of students to be assigned individual motifs to be briefly explained at the beginning of class. Following students' familiarity with the protein motifs, the various databases shown in Table 2.2 could be explored.

Table 2.2. Secondary structural characterization programs

Program	Function	Website
PROF	Secondary Structure Prediction System	http://www.aber.ac.uk/~phiwww/prof/
PORTER	Protein Secondary Structure Prediction	http://distill.ucd.ie/porter/
SOPMA	Secondary structure prediction	https://npsa-prabi.ibcp.fr/cgi-bin.html
SOSUI	Classification and Secondary Structure Prediction	http://harrier.nagahama-i-bio.ac.jp/sosui/
GOR	Protein secondary structure prediction	https://npsa-prabi.ibcp.html
Jpred	Secondary Structure Prediction Server	http://www.compbio.dundee.ac.uk/www-jpred/
MARCOIL	Coiled-coils prediction	http://toolkit.tuebingen.mpg.de/marcoil
APSSP	Advanced Protein Secondary Structure Prediction	http://imtech.res.in/raghava/apssp/
CFSSP	Protein secondary structure prediction	http://www.biogem.org/tool/chou-fasman/
Poodle	Prediction of disordered protein regions	http://mbs.cbrc.jp/poodle/poodle.html
NetTurnP	Prediction of Beta-turn regions in proteins	http://www.cbs.dtu.dk/services/NetTurnP/

2.2 B. Student projects and presentations (~6-8 weeks)

Following students' introduction to bioinformatics will be the second part of the course, the project portion. The goal of this part is to give students hands-on learning experience, allowing them to explore their personal interests as well as gain exposure to other students' interests. This could help students develop their interests and begin a proactive approach toward future career paths. Student projects will be voluntarily proposed

after the introduction part of the bioinformatics course. This will allow them to receive the proper training on the databases before picking their project.

2.2.1 Student projects overview

Before the start of the course it may be useful to explain to students that part of the course is dependent upon a project in which they must choose one gene to investigate for two projects. The first project will begin with students giving a brief overview (elevator pitch) of which genes they picked and the reasons why they picked them. The goal of this exercise is to help students learn to give quick 25-45 second overviews of their project without exhausting their listener, a skill that is critical for the future workforce as well as communication in science. Following this will be the first tutorial in molecular cloning. The first part will begin with an overview of molecular cloning with a concentration on assembling a plasmid using various databases. The following documents could be handed out to students during the first day of class to help them begin planning for their project:

Project overview

For your first and second project you will investigate a specific gene of interest using various bioinformatics and computational approaches. Whichever gene you pick will be assigned to you for the rest of the semester; make sure that you are interested in the gene you pick and allow yourself enough time to research it. A list of examples is listed below. The gene can be from any organism and should be relevant to the fields listed below.

Must be relevant to these fields using bioinformatics:

- | | | |
|------------------------|-------------------------|------------------------------|
| •Antibiotic resistance | •Microbial applications | •Biological weapons defense |
| •Evolutionary studies | •Molecular medicine | •Improve nutritional quality |
| •Waste cleanup | •Personalized medicine | •Veterinary science |
| •Alternative energy | •Preventative medicine | •Biotechnology |
| •Crop improvement | •Gene therapy | •Climate change studies |
| •Forensic analysis | •Drug development | •Oil Industry |

Turn in a short proposal for your project that follows this format:

Gene name:

Accession number:

Why is this gene interesting to you:

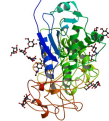
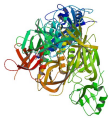
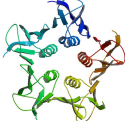
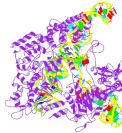
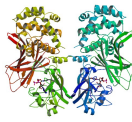
Which field you intend on relating your study too and why is it relevant to this field:

Nucleotide sequence of your gene (Color the coding regions red, the non-coding regions blue):

Abstract from NCBI-PubMed explaining gene function:

An example of genes that may be interesting to you are listed on the next page→

Example Genes:

Protein	Function	Structure	PDB ID
Ricin	In addition to being one of the deadliest toxins in the world it is also, in complex with specific antibodies, a potential cure for cancer		2AAI
Pertussis toxin	An extremely potent toxin that causes lethality by permanently turning on G-proteins, leading to intestinal flooding		1PRT
Botulinum Toxin	The most acutely lethal toxin known; commonly used for botox and medical purposes		1C48
Cas9	Used for genetic engineering in complex with CRISPR guide RNA. Recently used to cure latent HIV infection		4008
Diphtheria Toxin	Causes the diphtheria disease in humanS by de-activating EF-2 and therefore inhibiting protein synthesis		1MDT

Thoughts for the project:

- You will continue using this gene throughout the rest of the semesters' exercises and therefore you may want to spend time thinking about which gene you pick or which field you are interested in.

2.3 An overview of molecular cloning

Molecular cloning is an important tool for the study of specific functions in the biological sciences. One of the first steps in molecular cloning consists of identifying and isolating a gene of interest followed by amplification of the gene via polymerase chain reaction (PCR) (Figure 2.2). The next step involves inserting the gene into a cloning vector that is compatible with the replication criteria for the model organism (Figure 2.3). Cloning vectors are pieces of DNA that contain desirable and well characterized features such as antibiotic resistance for selection of transformed bacteria or selection markers that allow for selection on specific medium types, such as *HIS3* markers in yeast. Vectors come in a variety of different genotypes and sizes based on the specific areas of research.

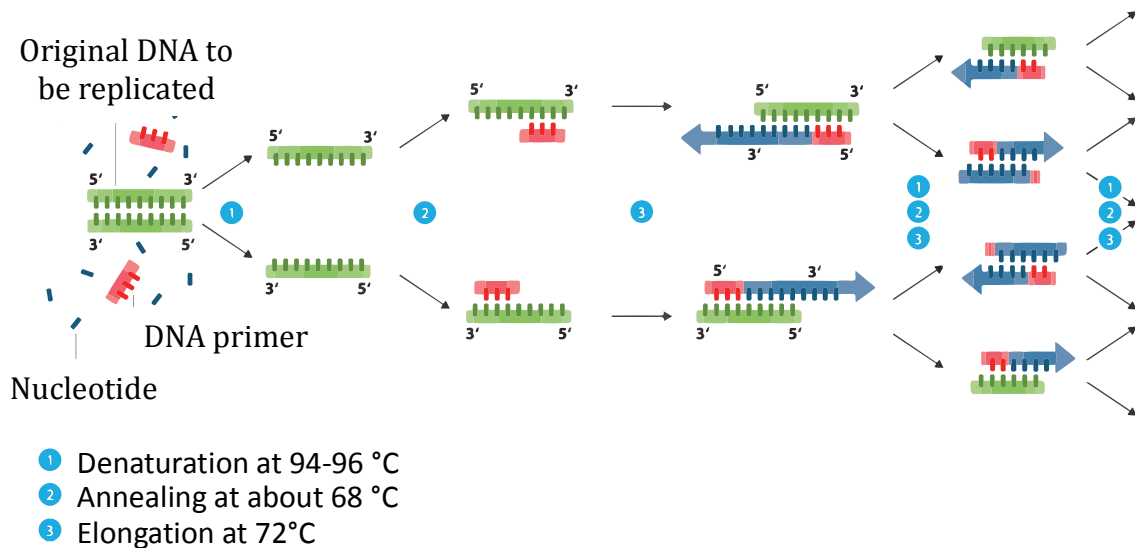
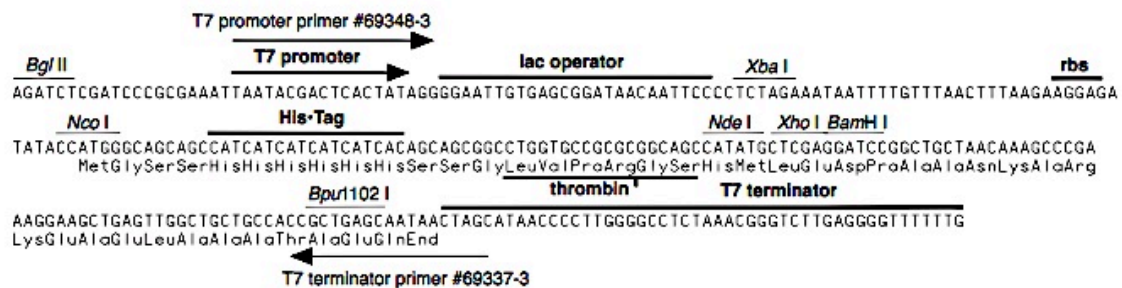


Figure 2.2. The Polymerase Chain Reaction (PCR). DNA templates of interest can be amplified using PCR by (1) Denaturing the DNA and separating it into single stranded DNA (ssDNA). (2) Annealing complementary primers to the ssDNA follows the denaturation process before DNA is finally (3) elongated in the final step of the PCR cycle. The idea for this figure came the work of Enzoklop under a creative common license.



pET-15b cloning/expression region

Figure 2.4. Features of the pET15b expression vector. The multiple cloning region downstream of the T7 promoter can be selectively expressed using a compound similar to allolactose, IPTG. The vector also contains a His-tag that allows for the successful purification of the protein coded for by the cloned gene. This image was reprinted from the Invitrogen website.

Shuttle vectors are also used in molecular cloning. An example would be shuttle vectors that can transverse between bacteria and yeast. These vectors are typically propagated in *E. coli* following ligation of the gene in order to attain high enough concentrations of the plasmid before transforming the vector into yeast (Figure 2.5). A specific example of this vector would be the pRS300 vector series, described by Robert S. Sikorski and Philip Hieter [16]. They constructed a vector that could propagate in *E.coli* and then be transferred, after purification from *E .coli*, to *S. cerevisiae*. These plasmids, in addition to having an origin of replication specific to *E. coli* (OriC) also contain features specific for *S. cerevisiae*: an origin of replication specific for yeast, a yeast centromere (*CEN*), and a selectable marker such as *HIS3* or *URA3*. Both types of vectors, expression and shuttle, can be used as tools to answer different questions but together represent the fundamental steps in choosing a vector for molecular cloning.

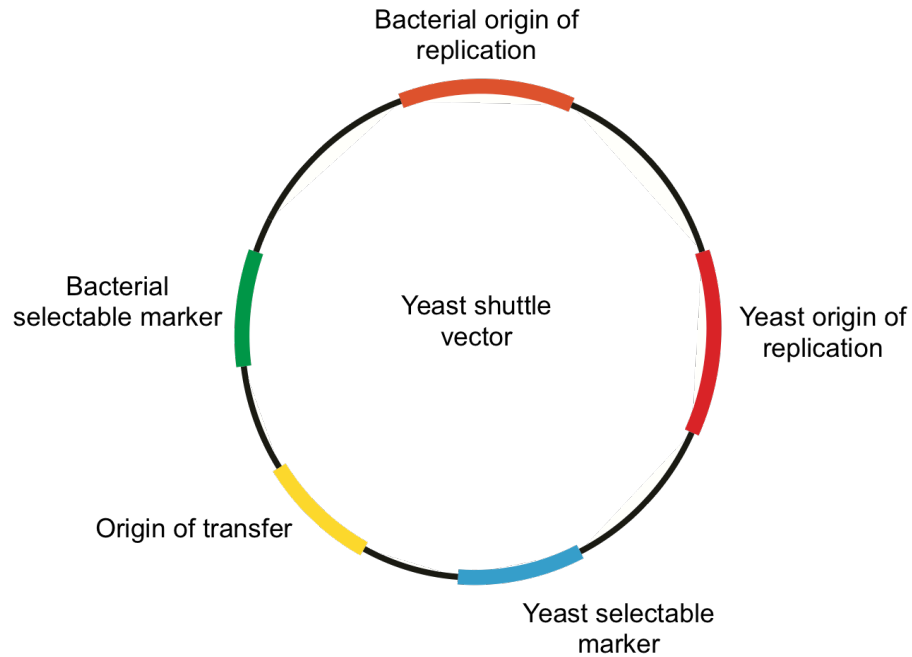


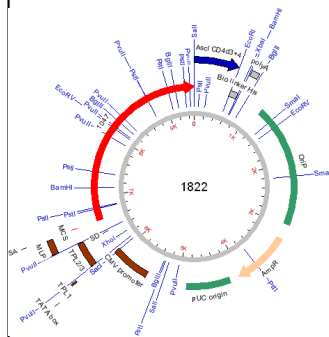
Figure 2.5. Components of a yeast shuttle vector. Multiple origins of replication must be present in order to propagate the plasmid from *E.coli* to yeast. In addition, an origin of transfer must also be present in order for the plasmid to be passed on other bacteria. The selectable marker allows for the screening and selection of plasmid containing organisms.

2.4 Project 1: Molecular Cloning

As previously mentioned, the projects will be scenario based. Several scenarios will exist based on the student's projects, but an example of a scenario that would be given to a student who picked insulin would be:

Project 1: Molecular Cloning

You are applying for a grant to the Bill and Melinda Gates foundation in order to begin mass-producing insulin. You must first outline how you will successfully produce the insulin protein using bacteria. Create a ~5 minute PowerPoint presentation that outlines how you intend to do this. Make sure to include the following:



- i. Accession number of the insulin gene
- ii. Which part of the insulin gene sequence used
- iii. Which expression vector you used
- iv. Which affinity tag you used
- v. A vector map with annotations of the specific genes

2.5 Assembling a plasmid tutorial

2.5.1 Picking the correct expression vector

Addgene, a plasmid repository, offers an expression guide for various organisms and uses a graphically rich vector representation to illustrate specific features (Figure 2.6).

Explaining these annotations is crucial for student's projects since they are going to need to choose a vector, explain the reasoning for choosing the vector, and be able to describe the specific annotations. The repository distinguishes between organism-specific vectors and

cell-specific vectors. Keeping track of the vectors that are available as well as learning about new vector constructs that could be used to streamline certain cloning approaches is essential for molecular biology. The repository offers a training/tutorial in molecular cloning as well that focuses on explaining the process experimentally using actual volumes and ratios rather than just theoretical principles.

A. Choose by:

- Species-specific expression
- Epitope tag or fusion protein
- Selectable markers
- Viral expression and packaging
- Reporters, shRNA expression, transgenics and genome modification

Species-Specific Expression

If you want to drive expression of your favorite gene, you will need a plasmid with a promoter that will be functional in your host organism.

Host	Relevant Promoters	Representative Empty Backbones
Mammalian	CMV, SV40, EF1a, CAG	<ul style="list-style-type: none"> pSG5, Flag HA - Transient expression pBABE-puro - Retroviral expression pWPXL - Lentiviral expression pBI-MCS-EGFP - Tet-inducible Transient expression - Gradia Lab plasmids for mammalian expression, Ligation Independent Cloning (LIC)
Bacteria	Lac, T7, araBAD	<ul style="list-style-type: none"> pBAD LIC cloning vector - Bacterial expression (see all expression vectors from the Gradia lab) p15TV-L - His-tagged bacterial expression vector (see tagged expression vectors from the SGC) pPro18 - Propionate-inducible expression vector (see entire pPro collection) Arrowsmith Lab Plasmids - More LIC bacterial cloning vectors pTD plasmid series - Expression with Strep, His or optimized YFP tags for purification or localization experiments
Yeast	GAL4, PGK, ADH1, ADE2, TRP1	<ul style="list-style-type: none"> Gateway destination vectors - Inducible expression, fusion proteins, and more Sandmeyer Lab Plasmids - Set of yeast expression vectors with various markers, promoters, etc

B.

Sequences from Depositing Scientist:

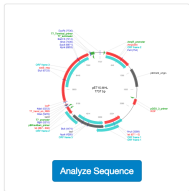
Full (1)

Full Sequences from Depositor (1)

```
>Author sequence
tctctgaacgcaaaagggcgtggtgagcgtctttttataggttaagtcgataaataaggttctt
tagcgtcaggtggaacttttggggaagtgtgoggaacccctattgtttttttaaatacatt
caattatgtatcgtctgagcaataaacctgtataaagcttcaataatctgaaagagagat
ggatctcaaacatttcgtgtgacgtctatctccctttttgggaatttgccttctgttgcac
ccgaagagcgtgtggaagttaaaagctgtgaagctcagtttgggtgcacagctgttgcacag
atctcacaagcgtgaagctcgttgcaggttttgcgcgcgaagagcttttccatgagcagctttaa
agttcgtatgtggtggtggtatctatcccgctgaacgcgcgcgaagagcagcagcagcagc
tattctcagatgacttgggttgactctccacgcgcgaagagcagcagcagcagcagcagcagc
gagatctatgagctggtggtggtggtggtggtggtggtggtggtggtggtggtggtggtggtg
agggcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagc
cggagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagc
tcgcaaatatcaactggtggtggtggtggtggtggtggtggtggtggtggtggtggtggtggtg
ggatcaagcttgcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagc
ggcgtgagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagc
ttatctcaagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagc
atgctgaagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagc
```

Sequences from Addgene:

Partial (1)



Analyze Sequence

Figure 2.6. The AddGene repository. A) The plasmid repository offers an organism specific expression guide and B) sequence-specific vector map. This image is a compilation of screenshots from the AddGene website (www.addgene.org).

2.5.2 Cloning software

Both ApE and Serial Cloner are useful programs for organizing DNA sequencing files and can also be used in organizing cloning projects and visualizing the construction of a plasmid during cloning. Both programs offer graphically rich displays for visualization in addition to several other shared features shown in Table 2.3.

Table 2.3. The features of Serial Cloner and ApE

Serial Cloner and ApE features
Show DNA sequence and text map
Aligns two DNA sequences
Translates DNA sequence and aligns
Restriction digest map-linear or circular
Hyperlinks graphic to corresponding DNA sequence
Virtual: cutter, PCR, cloning
Full graphical map
Imports DNA strider format files
Extract specific fragments

2.5.3 ApE

Similar to the NEB cutter program that helps determine restriction enzyme sites (www.neb.com), ApE also predicts unique sites and provides information about the frequency at which the specific sequence appears in the sequence, shown in Figure 2.7. Similarly, the primer find feature can also help find unique primer regions based on the users pre-set parameters and specific criteria. A summary of ApE's tools is shown in Table 2.4.

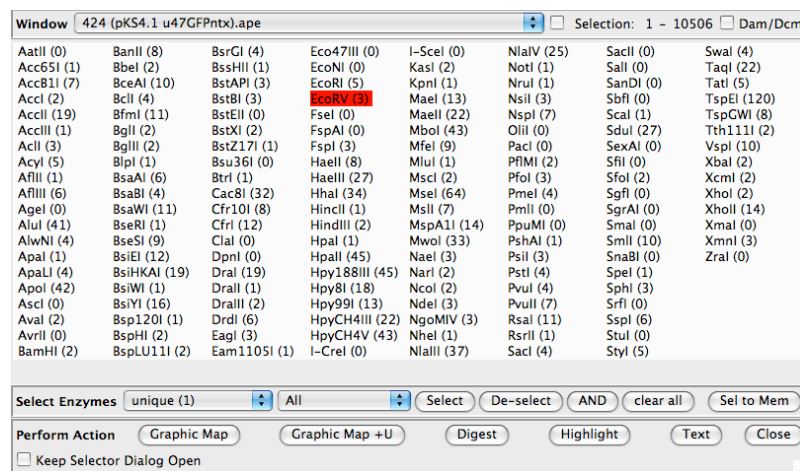


Figure 2.7. Restriction enzyme cleavage frequency map in ApE. The restriction enzyme table for selection of unique sites on a DNA sequence is a feature of the Ape program. In addition to the table shown is the ability to utilize catalog information and insert new restriction enzyme sites. This image was screenshot from the ApE program.

Table 2.4. Unique features of the ApE program.

ApE
Allows users to define new enzymes by name and recognition site
Multiple ABI sequence chromatographs alignments
Finds translationally silent restriction sites

2.4.2 Serial Cloner

During sub-cloning projects feedback is given; only ligating DNA fragments if they have compatible ends after restriction digest. The user interface, shown in Figure 2.8 is easy to use and help is provided in tutorial videos listed on the website. A list of the features unique to serial cloner is provided in Table 2.5.

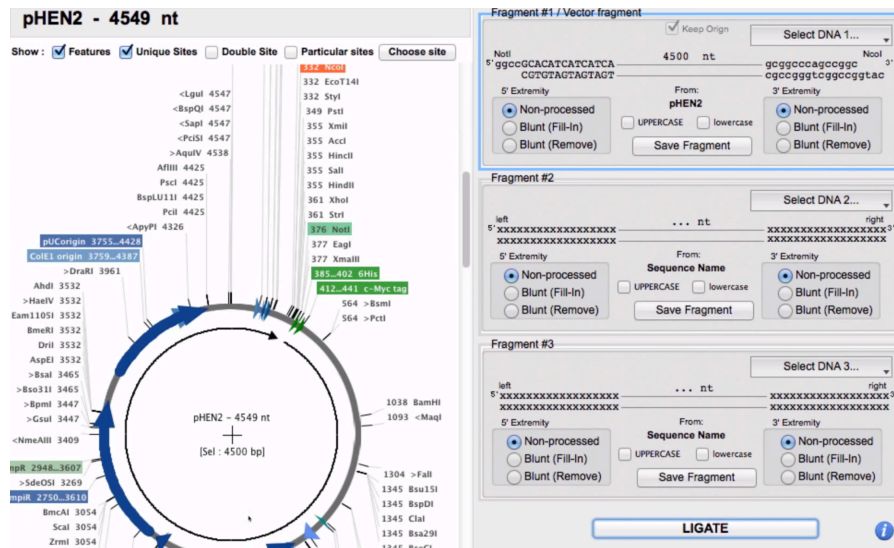


Figure 2.8. A virtual sub-cloning approach using Serial Cloner. The program allows you to manually join sequences together and provides feedback during the cloning process by only allowing complementary sequences to be ligated. In addition, the interface is graphically rich and provides useful information in terms of plasmid features and unique restriction enzyme sites. These images were screenshot from the Serial Cloner program.

Table 2.5. The unique features of the Serial Cloner program

Serial Cloner
Can align sequences using local algorithm or BLAST2Seq
Ligate Fragments- Only ligates if correctly matched
Imports DNA files in DNA Strider, Vector NTI, MacVector
Scans the DNA for features and automatically annotates
Numerically select fragments, find restriction sites, ORF or any nucleotide or peptide sequence,
Calculates Tm, GC content, and molecular weight of resulting peptides
Adaptor synthesis
shRNA Set-up

2.5.4 Determining multiple cloning sites (MCS)

Being able to properly identify the MCS is a crucial part of the cloning process (Figure 2.9) . In addition, the directionality of the insert into the vector backbone is also crucial for expression of the correct coding sequence of the protein. Both of these can be identified, modeled, and verified using Serial Cloner. Therefore, a walk through tutorial in serial cloner is an important part of the student's tutorial before they begin working on cloning their plasmid into a vector.

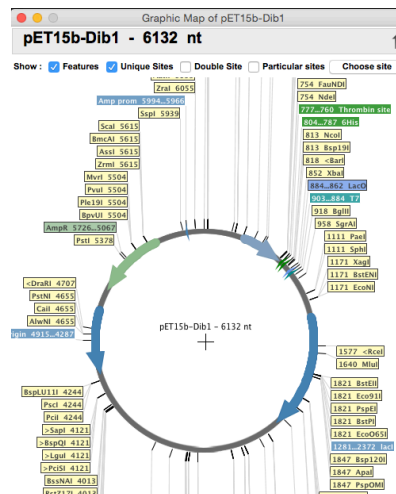


Figure 2.9. Restriction mapping is Serial Cloner.

2.5.5 Helping students organize their project

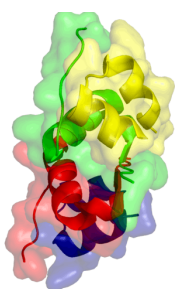
For most students this will be the first time they are ever exposed to a molecular cloning project and therefore walking them through an example of how to organize the project may be helpful. For example, guiding students through a PubMed search for a primary literature article that shows how a gene of interest was cloned.

2.6 Project 2: Structural characterization

The structural characterization and subsequent profiling of specific proteins is useful when studying protein function or its dynamics in a system of proteins. One way to do this is by looking at a 3-D structure of the protein and investigating its structural properties. This is the focus of the second project. For example, as an extension of the previous example, a student who picked insulin would receive the following:

Project 2: Structural Characterization

Create a short, ~5 minute power point presentation for your investors that outlines the secondary and 3-dimensional structural characteristics of insulin protein. Make sure you include the following:



(Polyview 3-D, 3E7Y)

- i. Specific secondary-structural characteristics.
- ii. Important regions for catalytic activity/function
- iii. Protein Data Bank ID and author
- iv. Citation for the primary literature article
- v. Organism of origin
- vi. Method of structure determination

2.7 Exploring Insulin using the Protein Database

The protein database (PDB) is the largest repository for proteins structures and is a very interactive website for exploring various biomolecules structure. Being able to utilize this database for structural insight holds importance for teaching students the importance of certain protein motifs as well as for investigating structural dynamics. One way of integrating the PDB into the course curriculum could be by adding it to the original cloning project that was completed in project 1. This could help students develop ownership of their project as they would be able to see how to molecular assemble a gene into a plasmid for expression as

well as the structure of the protein once it is expressed. An example of this process would be the cloning and subsequent structural analysis of human insulin.

2.7.1 Determining the structure

The PDB is a large database and walking through it with students would be necessary before they began using the database for their project. An example of how to navigate through PDB is available on their website in an interactive tutorial listed in their resources section.

2.7.2 Determining the functional properties

Determining the functional properties of insulin can be streamlined if students are capable of navigating through the databases originally introduced in the beginning of the course. This will require students to engage in primary literature to determine what structural properties their proteins contain. For example, a literature search on insulin would tell you that it is commonly used for the treatment of patients with diabetes. Furthermore, it contains only alpha helices and can exist in either monomeric or hexameric form depending on hydrophobic surface interactions. One of the main focuses researchers have been concerned with is how to provide fast acting insulin. Slow acting insulin, the hexameric form, and complexes with zinc and must equilibrate to its monomeric form before it can begin binding to its receptor. Therefore, a goal for researchers in the field was to engineer fast acting insulin that did not alter the host receptor binding but rather just limited the formation of the hexameric units of insulin. Researchers accomplished this by genetically engineering a mutated form of insulin, as shown in Figure 2.10 at the C-terminal tail of the B-chain.

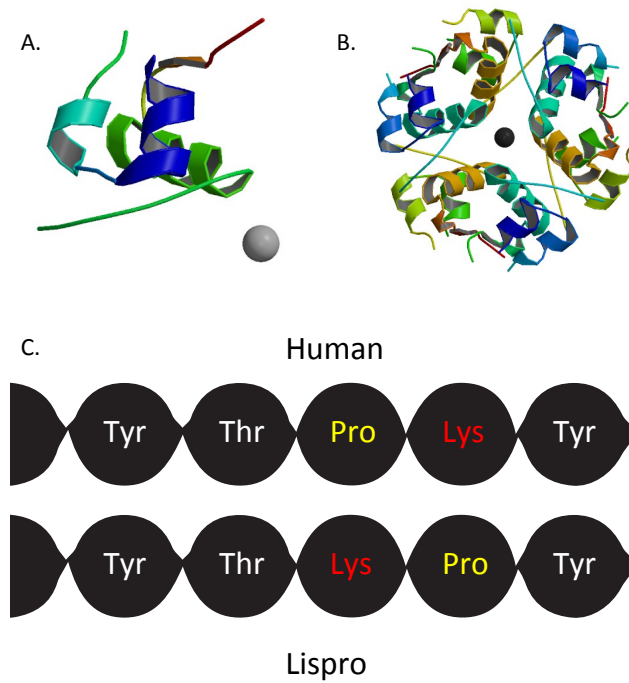


Figure 2.10. The structural analysis of fast and slow acting insulin. The A) monomeric form is fast acting insulin (PDB: 4ins) and the slow acting form is B) hexameric (PDB: 1aio). By altering the C) C-terminal of the B-chain of insulin it is possible to deliver fast acting insulin to patients with diabetes. This form of insulin is a prescription drug under the alias Insulin Lispro.

CHAPTER 3

EXPLORING THE CURRICULUM

The course proposed in Practical Bioinformatics, originally proposed to be integrated into the core curriculum for biochemistry majors at Texas State, cannot actually be integrated into the course curriculum. Current accreditation requirements for the ACS certification do not allow for another course to be integrated. Although, there are several other avenues in which the course could be extended to biochemistry and other life science majors. One option would be having the course available as an elective for students to take as a minor. The other would be having the course taught as an honors course, which would allow students to count their credits toward a minor in honors or toward graduating in the honors college. The final option, originally proposed and tested by Goodman and co-workers, would be teaching the course in concert between computer scientist and biochemist [18]. In addition to Goodman and co-workers original outline, it may also be useful to include mathematicians in the curriculum. Their outline and results of course evaluations are shown in Figure 3.1.

Mathematicians often have an overlapping knowledge of computer science, an interest in graph theory, and other pure mathematical concepts. The knowledge and creative capabilities that are associated with this background are the basis of much of the bioinformatics databases and therefore their role in adding to the diversity of the course could be critical. Computer science (CS), often interested in the theory associated with computational algorithms could prove insightful into writing programs or scripts that could be utilized in open source formatting databases such as NCBI. This is critical because it

would allow the use of open-source databases. Using closed-sourced databases limit the user to what the databases have decided to represent rather than what is actually being asked. Finally, biochemist and other life science majors would provide their background in understanding biological processes to pose a question or identify a problem that could be further be explored as a group. If all three-subject areas were able to collaborate in a project like this it could provide invaluable experience, exposure to real-world collaboration, and potentially a publication. All of these skills and outcomes could provide a significant advantage in helping students with their future career as well as provide faculty with additional publications.

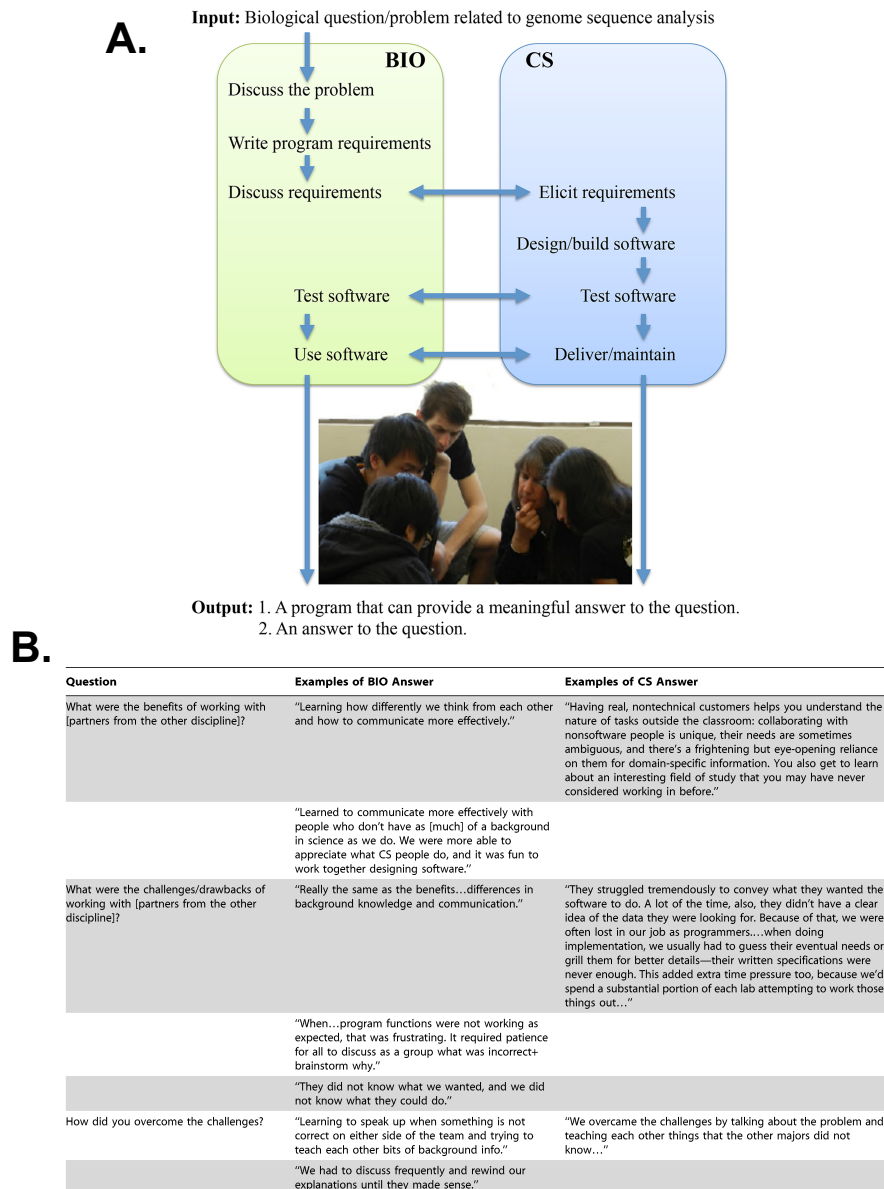


Figure 3.1. The A) outline for a course taught in-concert and various B) student responses to an exit survey. This image was reprinted from [18].

REFERENCES

- [1] Hogeweg, P. *PLoS Computational biology* **2011**, 7, e1002021
- [2] Jou, W. M.; Haegeman, G.; Ysebaert, M.; and Fiers, W. *Nature* **1972**, 237, 82-88.
- [3] Sanger, F.; Nicklen, S.; and Coulson, A. R. *Proc. Natl. Acad. Sci.* **1977**, 74, 5463-5467.
- [4] Hewitt, C. *Artificial Intelligence* **1977**, 323–364.
- [5] Hogeweg, P.; and Hesper, B. *Computers in biology and medicine* **1978**, 8, 319-327.
- [6] Hogeweg, P. *Simulation* **1978**, 31, 90–96.
- [7] Hesper, B.; Hogeweg, P.; *Kameleon* **1970**, 1, 28–29.
- [8] Hogeweg, P.; *PLoS Computational biology* **2011**, 7, e1002021.
- [9] Bilofsky, H. S.; and Christian, B. *Nucleic acids research* **1988**, 16, 1861-1863.
- [10] Agostino, M. *Taylor & Francis* **2012**, 1, 246.
- [11] Geer, L. Y.; Marchler-Bauer, A.; Geer, R. C.; Han, L.; He, J.; He, S.; ... and Bryant, S. H. *Nucleic acids research* **2009**, gkp858.
- [12] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; and Lipman, D. J. *Journal of molecular biology* **1990**, 215, 403-410.
- [13] Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; ... and Sherlock, G. *Nature genetics* **2000**, 25, 25-29.
- [14] Paradis, E.; Claude, J.; and Strimmer, K. *Bioinformatics* **2004**, 20, 289-290
- [15] Serial Cloner. Vers. 2.1. **2015**, web.
- [16] Schwab, R.; Ossowski, S.; Riester, M.; Warthmann, N.; and Weigel, D. *Plant Cell*. **2006** 18, 1121-33.
- [17] Cherry, J. M.; Hong, E. L.; Amundsen, C.; Balakrishnan, R.; Binkley G.; Chan, E. T.; Christie, K. R.; Costanzo, M. C.; Dwight, S. S.; Engel, S.R.; Fisk, D. G.; Hirschman, J. E.; Hitz, B. C.; Karra, K.; Krieger, C. J.; Miyasato, S. R.; Nash, R. S.; Park, J.; Skrzypek, M. S.; Simison, M.; Weng, S.; and Wong, E. D. *Nucleic Acids Res* **2012**. 40, D700-5.
- [18] Goodman, A. L.; and Dekhtyar, A. *PLoS computational biology* **2014**, 10, e1003896.