NAVIGATING DATABASES

The role of bioinformatics in the biology curriculum.

The amount of information, with respect to biology, has exploded. Bioinformatics is the use of computer technology to manage this information. It is important that students have a working knowledge of how to navigate these databases. The need for such databases was as a result of the vast amount of information obtained from the Human Genome Project. Bioinformatics includes the disciplines of computer science, statistics, mathematics, and engineering.

University of Florida Center for Precollegiate Education and Training

NAVIGATING DATABASES

UF | Center for Precollegiate Education and Training
UNIVERSITY of FLORIDA

Author: Janet Bisogno, Ed.D

Thank you to the following who offered excellent review and suggestions:

Candace Roy
IB/AP Biology Teacher- Vanguard High School, Ocala Florida

The Berglund Lab at University of Florida
Andy Berglund, PhD
Melissa Hale
Carl Shotwell

This curriculum was developed as part of *Biomedical Explorations: Bench to Bedside*, which is supported by the Office Of The Director, National Institutes Of Health of the National Institutes of Health under Award Number R25 OD016551. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Please direct inquiries to Janet Bisogno at bisognoj@osceola.k12.fl.us or 407-873-2640

Last updated: 7/15/2016

Center for Precollegiate Education and Training

PO Box 112010 • Yon Hall, Room 331

Gainesville, FL 32611

Phone 352.392-2310• Fax 352.392-2311

This curriculum was developed in the laboratory of:

Andy Berglund, PhD

Professor
Department of Biochemistry and Molecular Biology
College of Medicine

The Berglund lab uses a broad range of approaches to study the molecular mechanisms of neurological diseases that are caused by microsatellite repeat expansions. For many of these diseases (myotonic dystrophy, ALS and ataxias), RNA processing (pre-mRNA splicing) pathways are negatively impacted with specific changes in pre-mRNA splicing proposed to lead to symptoms observed in affected individuals. We use biochemical, cellular and genomic assays to understand the mechanisms through which these diseases alter pre-mRNA splicing. The goal of our research is to use the results from these fundamental studies to identify innovative strategies to reduce or correct the improper pre-mRNA splicing that occurs in the disease state. For example, we have recently shown that small molecules can be used to rescue the mis-splicing in cell and mouse models of myotonic dystrophy.

TABLE OF CONTENTS

AUTHOR'S NOTE

As a teacher in the International Baccalaureate (IB) Program, teaching the understandings (what IB calls their standards) is vital as the exams students take are based on these understandings. One skill that students need to be able to do is to use databases. With the explosion of information after the Human Genome Project, navigating biological databases is important. While I utilized an activity that uses BLAST, I realized that what I was doing was inadequate but did not know what to do to make my lessons more relevant to the study of biology.

During the summer of 2016, I participated in a Summer Research Experience at University of Florida. Classroom teachers were placed into research labs with the goal of writing curriculum to take back to their classrooms. I saw very quickly how important these databases are to researchers and so decided to "up my game" and write curriculum that would give students an introduction to several of the databases being used by researchers.

I was unaware of the different types of databases that were available and what these databases could do. While I used BLAST to determine nucleotide sequences and identify the genus and species to which that nucleotide sequence belonged, I quickly realized I was barely scratching the surface of what these databases contained and could do.

I learned a tremendous amount of information and even with these lessons on bioinformatics, there is so much more. My hope is that because of their interaction with some of these databases, students will become aware of new science careers such as bioinformatics and proteomics. They are fascinating fields of science!

INTRODUCTION

The concepts included in this curriculum are frequently taught as individual units and at different times during the curriculum. This can make it difficult for students to understand how these concepts are interrelated. Transcription, translation, protein structure and function, and use of databases are used to show how these concepts are not stand alone processes. While working in the lab, it quickly became apparent that the use of databases is an integral and vital tool in research. As biological data continues to expand at rapid rates, the field of bioinformatics is critical to the organization and management of this data. Students need to enter college with a basic understanding of how to navigate these databases. Utilizing databases gives students the opportunity to see how concepts, usually taught as separate topics, are actually dependent upon the integration of these topics.

The protein used in these lessons is cytochrome c. This protein was chosen for several reasons. The primary reason is that it is a protein recommended by IB. It is a fairly small protein, 150 amino acids long and it has only three EXONS. Cytochrome c is an ancient protein and so is found in nearly all organisms which makes it easy to align human cytochrome c with other species.

My hope is that students will be able to search for a human gene of interest and be able to align that gene with two or more other species to see the differences. They will be able to see on which chromosome that gene is located as well as the position and size. The number of EXONS is easily determined. Utilizing the differences between the species, a cladogram can be examined. Students will also take a section of that gene to transcribe and translate. Molecular visualization software will allow students to see the three dimensional structure of the protein. Secondary structures, alpha helixes and beta pleated sheets are easily seen in these models, thus reinforcing the structure of proteins.

TIPS ABOUT THIS CURRICULUM

Lesson Plan Format: All lessons in this curriculum unit are formatted in the same manner. In each lesson you will find the following components:

KEY QUESTION(S): Identifies key questions the lesson will explore.

OVERALL TIME ESTIMATE: Indicates total amount of time needed for the lesson, including advanced preparation.

VOCABULARY: Lists key vocabulary terms used and defined in the lesson. Also collected in master vocabulary list.

LESSON SUMMARY: Provides a 1-2 sentence summary of what the lesson will cover and how this content will be covered. Also collected in one list.

STUDENT LEARNING OBJECTIVES: Focuses on what students will know, feel, or be able to do at the conclusion of the lesson.

UNDERSTANDINGS: Specific International Baccalaureate understandings, applications and skills and aims addressed in the lesson. Also collected in one list.

MATERIALS: Items needed to complete the lesson. Number required for different types of grouping formats (Per class, Per group of 3-4 students, Per pair, Per student) is also indicated.

BACKGROUND INFORMATION: Provides accurate, up-to-date information from reliable sources about the lesson topic.

ADVANCE PREPARATION: This section explains what needs to be done to get ready for the lesson.

PROCEDURE WITH TIME ESTIMATES: The procedure details the steps of implementation with suggested time estimates. The times will likely vary depending on the class.

ASSESSMENT SUGGESTIONS: There are assessment questions built into the lessons. As this is focused more on skills and not content summative assessment pieces are not included.

EXTENSIONS: (ACTIVITIES) There are many activities and reading sources available to augment and enhance the curriculum. They have been included. If you find additional ones that should be added, please let us know.

RESOURCES/REFERENCES: This curriculum is based heavily on internet source. As resources and references have been used in a lesson, their complete citation is included as well as a web link if available. All references and resources are also collected in one list.

STUDENT PAGES: Worksheets and handouts to be copied and distributed to the students.

TEACHER MASTERS: Versions of the student pages with answers or the activity materials for preparation.

Collaborative Learning: The lessons in this curriculum have been developed to engage students. There is no lecture component to these lessons.

Groups: The lessons are carried out in pairs but larger groups would work as well.

Inquiry-based: The lessons in the curriculum invite students to be engaged and ask questions. They navigate the databases with guidance from the teacher.

Technology: Due to the nature of bioinformatics, access to a computer with internet is mandatory. While it would be best for students to have a computer, the teacher could project the databases for students to see.

Content: These lessons assume the student has already learned about transcription and translation as well as the structure of proteins. The purpose of these lessons is to expose students to databases that can help reinforce what they already know. The larger picture is to allow students to explore other careers in science and have them realize the importance of bioinformatics in any science career they may choose.

Implementation notes**:** This curriculum should be modified and adapted to suit the needs of the teacher and students. To help make implementation easier in this first draft, notes have been included in lessons as needed.

Extensions: A natural extension would be to explore and "play" with the different databases. Students could search for other proteins and other organisms of interest.

Science Subject: IB and AP Biology

Grade and ability level: 9-12 students in advanced biology

Science concepts: DNA, transcription, translation, protein structure, protein function, bioinformatics, electron transport chain

LESSON SUMMARIES

LESSON ONE: Introducing Bioinformatics
Students will be introduced to bioinformatics by listening to a TED Talk given by a young researcher in the field of bioinformatics who uses the information found in these database in attempts to discover a cure for the disease of cystic fibrosis. What makes this video so compelling is that the young man reveals that he has cystic fibrosis and so presents the importance of bioinformatics to understanding diseases and therefore, finding cures. As students will be using the protein of Cytochrome C to navigate the various data bases, a review of cellular respiration will be presented. Students will review the structure of mitochondria, the electron transport chain and the role of cytochrome c.

LESSON TWO: Navigating the Databases- Part 1
Working in pairs, students will use the databases of NCBI, BLAST, In-Silico, EBI and Uniprot to find cytochrome c of human, chimpanzee, dog, and mouse. They will be able to locate the gene of interest on the chromosome, identify the size of the gene in base pairs as well as the number of exons. Students will then align sequences of two or more species to note the differences. Using these differences, students will examine a cladogram.

LESSON THREE: Navigating the Databases – Part 2
Using a short segment of the gene of interest, students will then transcribe to mRNA and then translate to amino acids. Students will then be able to use the nucleotide sequences to see the 3D structure of cytochrome C. Students will use Uniprot to see the 3D structure.

OPTIONAL: I-Tasser is another great program to see these structures but access requires the participant to sign up and open an account. Therefore, the teacher can project this program and demonstrate to the students what is available in I-Tasser. Jmol is another program to view molecular structure but must be downloaded onto the computer. This could be a concern in some schools so Uniprot was used instead.

LESSON SEQUENCING GUIDE

Since the classroom teacher knows his or her students best, the teacher should decide the sequencing of lessons. Below is a suggested pacing guide that can be used when planning to use this curriculum.

45 minute periods

Day 1	Day 2	Day 3
Lesson 1 Introducing Bioinformatics (45 minutes)	Lesson 2 Navigating the Databases-Part 1 (45 minutes)	Lesson 3 Navigating the Databases-Part 2 (45 minutes)

VOCABULARY

Bioinformatics: A field in science where complex biological data is collected, stored and analyzed.

Cytochrome C- A protein found in mitochondria and carries electrons from one molecule to the next. This is vital to cellular respiration.

DNA: Deoxyribonucleic acid is a nucleic acid containing the genetic instructions used in the development and functioning of all known living organisms.

RNA Processing: This occurs in eukaryotic cells and is when the primary transcript RNA is processed into mature RNA. The primary transcript contains both exons and introns and the introns are removed. A guanine cap is placed on the 5' and poly adenine tail is placed on the 3'.

Electron Transport Chain: A series of compounds found in the inner membrane of mitochondria and on the thylakoid membrane of chloroplasts. Their primary role is to pass electrons from one to another by means of redox reactions. This allows the pumping of protons across the membrane to create a concentration gradient. This then drives the production of ATP.

Exon: The term used to represent any part of a gene that becomes part of the processed mRNA after the introns have been removed.

Intron: A segment on a DNA or RNA that does not code for a protein and is removed in RNA processing.

Mitochondria: Rod-shaped organelles found in the cytoplasm of eukaryotic cells and is the site of the production of high energy compounds such as ATP.

Proton Gradient: This is created when proteins pump protons across a membrane against their concentration gradient. This is important in the processes of cellular respiration and photosynthesis.

Reading frame: A way of breaking a sequence of nucleotides in DNA or RNA into three letter codons, resulting in a possibility of three reading frames in mRNA and six in double-stranded DNA (since have forward and reverse).

Redox Reactions (Oxidation -Reduction): Chemical reactions where electrons are transferred from one molecule to another. The molecule donating or losing the electron is undergoing oxidation while the molecule receiving or gaining the electron is being reduced. These reactions occur in pairs.

RNA: Ribonucleic acid is one of the three major macromolecules (along with DNA and proteins) that are essential for all known forms of life. Like DNA, RNA is made up of a long chain of components called nucleotides. Each nucleotide consists of a nucleobase, a ribose sugar, and a phosphate group. RNA directs the synthesis of proteins.

Transcription: DNA → RNA; During transcription, a DNA sequence is read by an RNA polymerase, which produces a complementary, antiparallel RNA strand. The RNA complement includes uracil (U) in all instances where thymine (T) would have occurred in a DNA complement.

Translation: RNA → Protein; In translation, messenger RNA (mRNA) produced by transcription is decoded by the ribosome to produce a specific amino acid chain, or polypeptide, that will later fold into an active protein.

UNDERSTANDINGS – IB BIOLOGY

UNDERSTANDINGS, APPLICATIONS AND SKILLS, AIMS	1	2A	2B
2.4 Proteins Understanding: The amino acid sequence determines the three-dimensional conformation of a protein		X	X
2.4 Proteins Aim 7: ICT can be used for molecular visualization of the structure of proteins			X
2.7 DNA Replication, Transcription, and Translation Understanding: Transcription is the synthesis of mRNA copied from the DNA base sequences by RNA polymerase.		X	
2.7 DNA Replication, Transcription, and Translation Understanding: Translation is the synthesis of polypeptides on ribosomes.		X	
2.7 DNA Replication, Transcription, and Translation Understanding: The amino acid sequence of polypeptides is determined by mRNA according to the genetic code.		X	
2.7 DNA Replication, Transcription, and Translation Skill: Use a table of mRNA codons and their corresponding amino acids to deduce the sequence of amino acids coded by a short mRNA strand of known base sequence.		X	
2.8 Cell Respiration Understanding: Cell Respiration is the controlled release of energy from organic compounds to produce ATP.	X		
3.1 Genes Understanding: A gene occupies a specific position on a chromosome.		X	
3.1 Genes Skill: Use of a database to determine differences in the base sequence of a gene in two species.		X	X
3.2 Chromosomes Skill: Use of databases to identify the locus of a human gene and its polypeptide product.		X	X
5.4 Cladistics Understanding: Cladograms are tree diagrams that show most probable sequence of divergence in clades.			X
8.2 Cell Respiration Understanding: Cell respiration involves the oxidation and reduction of electron carriers.	X		
8.2 Cell Respiration Understanding: Transfer of electrons between carriers in the electron transport chain in the membrane of the cristae is coupled to proton pumping.	X		
8.2 Cell Respiration Understanding: The structure of the mitochondrion is adapted to the function it performs.	X		
8.2 Cell Respiration Application and Skill: Annotation of a diagram of a mitochondrion to indicate the adaptations to its function.	X		
B.5 Bioinformatics Understanding: Databases allow scientists easy access to information.		X	X
B.5 Bioinformatics Understanding: The body of data stored in databases is increasing exponentially.		X	X
B.5 Bioinformatics Understanding: BLAST searches can identify similar sequences in different organisms.		X	X
B.5 Bioinformatics Understanding: Sequence alignment software allows comparison of sequences		X	X

UNDERSTANDINGS, APPLICATIONS AND SKILLS, AIMS	1	2A	2B
from different organisms.			
B.5 Bioinformatics BLASTn allows nucleotide sequence alignment while BLASTp allows protein alignment.		**X**	**X**
B.5 Bioinformatics Skill: Use of software to align two proteins		**X**	**X**
B.5 Bioinformatics Skill: Use of software to construct simple cladograms and phylograms of related organisms using DNA sequences.		**X**	**X**

BACKGROUND INFORMATION

Bioinformatics

The amount of information with respect to biology has exploded. Bioinformatics is a rising field in science and one that all students majoring in biological sciences will need to use and understand. Bioinformatics includes the use of computer technology to manage this information. It is important that students have a working knowledge of how to navigate these databases. The need for such databases was as a result of the vast amount of information obtained from the Human Genome Project. Bioinformatics includes the disciplines of computer science, statistics, mathematics, and engineering.

Mitochondria

Mitochondria are organelles found in the cytoplasm of eukaryotic organisms. Mitochondria have two membranes- an outer phospholipid bilayer and an inner phospholipid bilayer that is folded into cristae. The purpose of the folds is to increase the surface area for the electron transport chain and chemiosmosis. The inner membrane space is small, the purpose to allow a rapid increase in proton concentration gradients. Electrons are passed through the molecules on the inner membrane. This facilitates the pumping of hydrogen ions into the inner membrane space against their concentration gradient. Protons then pass passively through ATP Synthase cause phosphorylation of ADP and APT production.

Cytochrome C

Cytochrome C is considered an ancient protein – this means that it was developed early in the evolution of life and can be found in almost every living organism today. It has a vital function in the production of cellular energy in the form of ATP. It is found on the inner membrane, the cristae, of the mitochondria. Its role is to shuttle high energy electrons from cytochrome bc1 to cytochrome oxidase complex. It contains a heme group that binds to the electrons. The electrons are from the breakdown of glucose and fatty acids. The final electron acceptor is oxygen forming water. It is this protein that will be used to explore the databases.

LESSON ONE: INTRODUCTION TO BIOINFORMATICS

KEY QUESTION(S): What is Bioinformatics? What is the structure and function of the mitochondria?

OVERALL TIME ESTIMATE:

- Advanced Preparation: 40 minutes (30 minutes to assemble section packets, 10 minutes to open YouTube video on TedTalk)
- Student Procedure: 45 minutes

LEARNING STYLES: Visual and auditory

VOCABULARY:

Mitochondria: Rod-shaped organelles found in the cytoplasm of eukaryotic cells and is the site of the production of high energy compounds such as ATP.

Electron Transport Chain: A series of compounds found in the inner membrane of mitochondria and on the thylakoid membrane of chloroplasts. Their primary role is to pass electrons from one to another by means of redox reactions. This allows the pumping of protons across the membrane to create a concentration gradient. This then drives the production of ATP.

Cytochrome C: A protein found in mitochondria and carries electrons from one molecule to the next. This is vital to cellular respiration.

Redox Reactions (Oxidation -Reduction): Chemical reactions where electrons are transferred from one molecule to another. The molecule donating or losing the electron is undergoing oxidation while the molecule receiving or gaining the electron is being reduced. These reactions occur in pairs.

Proton Gradient : This is created when proteins pump protons across a membrane against their concentration gradient. This is important in the processes of cellular respiration and photosynthesis.

LESSON SUMMARY:

Students will be introduced to bioinformatics by listening to a TED Talk given by a young researcher in the field of bioinformatics who uses the information found in these databases in attempts to discover a cure for the disease of cystic fibrosis. What makes this video so compelling is that the young man reveals that he has cystic fibrosis and so presents the importance of bioinformatics to understanding diseases and therefore, finding cures. As students will be using the protein of Cytochrome C to navigate the various data bases, a review of cellular respiration will be presented. Students will review the structure of mitochondria, the electron transport chain and the role of cytochrome c.

STUDENT LEARNING OBJECTIVES:

The student will be able to...

1. Define bioinformatics and understand its use in biology.
2. Explain the structure and function of mitochondria.
3. Explain the role of cytochrome c in cellular respiration.

UNDERSTANDINGS:

2.8 Cell Respiration

Understanding: Cell respiration is the controlled release of energy from organic compounds to produce ATP.

8.2 Cell Respiration

Understanding: Cell respiration involves the oxidation and reduction of electron carriers.

Understanding: Transfer of electrons between carriers in the electron transport chain in the membrane of the cristae is coupled to proton pumping.

Understanding: The structure of the mitochondrion is adapted to the function it performs.

Application and Skill: Annotation of a diagram of a mitochondrion to indicate the adaptations to its function.

MATERIALS:

- YouTube video - https://www.youtube.com/watch?v=_eHz6qzTCfc
- Computer and projector to view video
- Copies of structure of mitochondrion, electron transport chain and cytochrome c (these can be projected on white board)

BACKGROUND INFORMATION:

This lesson assumes that cellular respiration and the structure of mitochondria have been taught. The purpose of this introductory lesson is to review the structure and function of mitochondria as well as the electron transport chain. The purpose of the review is because the protein cytochrome c is used to navigate the databases. The Ted Talk introduces the students to the field of bioinformatics. Teachers may want to view the video first and do some background research on bioinformatics.

ADVANCE PREPARATION:

1. View TedTalk (Transcript attached)
2. Copy structure of mitochondria, electron transport chain and cytochrome C if using as handouts to students.
3. If not using as handouts, have handouts on desk top of computer for easy access to project.

PROCEDURE AND DISCUSSION QUESTIONS WITH TIME ESTIMATES:

1. (12 min) Watch video from Ted Talks to introduce the field of Bioinformatics and its importance in the field of Biology. Found at https://www.youtube.com/watch?v=_eHz6qzTCfc. This TedTalk is given by a student, Spencer Hall. He is a student majoring in statistics. His plan is to attend graduate school studying statistics with a concentration in bioinformatics. Spencer has Cystic Fibrosis and believes that bioinformatics and statistical analysis of DNA could be used to cure deadly diseases.
2. (10 min) After the TedTalk, lead students in a discussion of their understanding of what bioinformatics is and its role in curing and treating diseases.
3. (20 min) Review the structure and function of mitochondria, electron transport chain and cytochrome C. Use the included diagrams of each of the above to lead the review.

REFLECTIVE QUESTIONS

1. Do you think Spencer Hall may have some bias towards the importance of bioinformatics in curing deadly disease? Explain your reasoning.
2. Draw a mitochondrion and annotate the structures and their function. Use IB draw rules.
3. Write a paragraph explaining how ATP is formed in cellular respiration.

EXTENSIONS:

1. Give students a blank copy of a mitochondrion to annotate – labeling structure and describing function.

RESOURCES/REFERENCES:

Bioplanet.com. (2013). What is bioinformatics?. Retrieved July 13, 2016 from http://www.bioplanet.com/what-is-bioinformatics.

Hall, S. (April 1, 2016).Bioinformatics: A way to deciphere DNA and cure life's deadliest diseases. Retrieved July 13, 2016 from https://www.youtube.com/watch?v=_eHz6qzTCf.

Reece, J., Urry, L., Cain, M., Wasserman, S., Minorski, P., & Jackson, R. (2014) Campbell AP Biology 10th ed. Pearson Education Inc.

Vossman - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=16469416

TEACHER PAGES: YOUTUBE TRANSCRIPT

Spencer Hall: Bioinformatics: A way to decipher DNA and cure life's deadliest diseases.

In 1799, while constructing a fort in Egypt, a group of French soldiers came across an ancient Egyptian tablet containing hieroglyphics on it. What made the discovery of this tablet so groundbreaking was not just the fact that it contained hieroglyphics, because by this point quite a few stones with that script had already been discovered. It was what came with the hieroglyphics that made the stones so important, Greek. Because, while for over a thousand years no one had been able to read the ancient Egyptian script everyone trained in classical languages could read Greek. This discovery, forty years later, enabled the French scholar?, Polian to find a way to code the meaning of the hieroglyphics. The symbols which had been silent for over a millennium became a new window into ancient Egyptian life and culture. Today, in the first two decades of the 21st century we are faced with a challenge and an opportunity similar to that of the Rosetta Stone. It comes from a language which dwarfs any human language in the immense number of volumes written in it and that's the language of DNA. DNA contains the blueprints for every biological process of every living thing on earth and even a small defect in the genetic code can cause devastating health consequences. Understanding how genes work, both individually and with each other, is crucial to the development of 21st century personalized medicine. The problem is that even when we actually have an organism's DNA in front of us, it's often not easy to tell what an individual gene does. Even more difficult can be teasing out the subtle influences multiple genes have on each other. But answering those kinds of questions is foundational to attacking genetically influenced diseases at the root cause. Forty years ago, even if we had had all the genetic information that we have today, which by the way we didn't, most of it has been found in the last twenty years but even if we had, we could not have even begun to sift through it. An army of people would never have had time to sift through the enormous amount of genetic data needed to answer health-related questions. Today, however, a new sub-field of biology is being developed that is making what formerly seemed impossible into a reality. That field is bioinformatics. Bioinformatics is the application of statistics and information theory to genetic data. It allows us to sift through an enormous amount of that data in a way that would be impossible to do manually. Let me give you an example of how bioinformatics can be useful. It is a very common classic statistical procedure called the two sample t-test and its purpose is that if we have two groups and we want to know if the average number of individuals with a certain feature in this group is the same as the average number of individuals with that feature in this group. We run the test to find out. So for example if I have two jars of marbles red marbles and blue marbles in each jar and I want to know if the average number of red marbles and a handful from this jar is going to be the same as the average number of red marbles in a handful from this jar. I can run the test by taking a sample of maybe thirty marbles from this jar, thirty marbles from this jar and then counting up the number of times red marbles appear in each handful. Once I have those two numbers, I plug them back into the ugly equation you just saw and what I get back is the probability that I have this number of red marbles in this hand and this number of red marbles in this hand we just got, if both jars have the same number of red marbles. That may seem arcane but let me give you an example of how we can use it in bioinformatics. Suppose we want to find out whether a particular gene is associated with thyroid cancer. We can take a sample of, maybe, the genomes of a hundred people with the cancer and a hundred healthy people and then count the number of times a gene appears in each group. Once we have those two numbers, just like with the marbles, we plug them back into the equation and what we get back, remember, is the probability that that gene appears this many times in the cancer group and this many times in the healthy group if you're equally likely to have the gene, whether you have the cancer or not and if that probability is sufficiently low we usually use the cutoff point, of maybe 5 percent, and then we can conclude that it's very likely that gene does have some sort of association with the cancer. This is a paradigm example of what makes bioinformatics so useful. We started with an ocean of genetic data before genomes of two hundred people and using statistics have narrowed our focus down to just one gene that can then be further studied by geneticists. Now. I have given you a little bit of an idea of how we can use bioinformatics but you may still be wondering why specifically genetic understanding of how disease works is important. And to answer that question I want to take one particularly cruel condition

as an example. Cystic fibrosis is a genetically inherited mutation that causes progressive lung disease and the way that works is everybody's lungs, CF or not have a thin mucous lining, the purpose of which is to catch bacteria and other irritants so they can be coughed out of the lungs and infection doesn't start. In CF patients the mucous lining is much thicker, which means it does what it's supposed to in catching the bacteria but it keeps them in the lungs. The very thing which is intended to stop an infection becomes a breeding ground for those same bacteria. Over time the thickened mucus clogs the airways, inflammation spreads throughout the lungs and in eighty percent of CF patients this leads to death by lung failure. Now there are two reasons why I picked CF as my example for how we can use Bioinformatics. The first, on a more optimistic note, that I'll get to in just a minute, is that the past few years have seen some major advances in how cystic fibrosis is treated and those advances were based upon a deep understanding of how the mutated CFTR gene works. The second however is that when I'm standing up here in front of all of you talking about bioinformatics and disease research, do not think that this is just some sort of interesting intellectual exercise that you'll hear and go home and forget about. It's not for me and it shouldn't be for you either because I am speaking to you first and foremost, not as a statistics major and first and foremost not as someone who wants to study bioinformatics in graduate school but as someone who has cystic fibrosis. And on my behalf and on behalf of the other one hundred thousand people worldwide with CF, I am telling you, our lives depend on our ability to better understand how the mutated gene works. Prior to about four years ago all the existing treatments for CF could be boiled down to two things. Clearing mucous out of the lungs and killing the bacteria that have gotten into the lungs. These were both helpful but neither of them actually address the root cause of the condition, which is the thickened mucous. Thickened mucous in CF lungs is caused by defective chloride channels in the lung cells. So if you can imagine cystic fibrosis is like a giant hole in your bathtub wall with water gushing out of it. Everything we've had so far has basically been ways of just using larger buckets to scoop the water out faster so the tub doesn't flood. This works for a little while. It's been effective in bringing up the CF lifespan from about 10 years in 1970 to thirty-seven years today but as long as the water is still gushing freely out on the wall, as long as the chloride channels are still malfunctioning the problem isn't really fixed. In the last four years, two new CF drugs, Kalydeco and Orkambi have been developed that partially repair the defective chloride channels in the cells. They're not cures but they're the first step towards patching over the hole and the only way we were ever able to develop such drugs was by understanding deeply and intricately how the mutated gene works. The decoding of hieroglyphics was no doubt an important milestone in our understanding of ancient Egypt but however fascinating such an intellectual pursuit may be, it does not begin to compare in importance and urgency to race to translate genetic information into the material for new disease treatments. Lives are being lost every day to heritable and genetically influenced diseases while we possess the information needed to save them, walked away right in front of us. If we are to combat the worldwide rise in cancer incidence, the untold suffering in the Global South from numerous virally caused illnesses and conditions like cystic fibrosis which has destroyed millions of lives through its long history, then we must continue to expand our genetic Rosetta Stone and we must use that stone to better understand diseases at the root causes. I challenge my generation to be one that will be remembered as one which made full use of its Mathematical and biological advances to save lives and end suffering. The Egyptian culture which gave us the hieroglyphics has passed into history and all that remains of their civilization are the monuments they have left to us. The day will come when we too, like the Egyptians, are nothing but a thing of the past to those who come after and they will know that we were the first generation entrusted with bioinformatics with all its potential for human health. And when that day comes, when we are nothing but a memory to those who come after, when all that is left of us on earth is what we have done, let this be our monument to them. That we pushed medicine into undreamed of frontiers, that we left behind a better understanding of disease that had ever been possible before and that we took up every weapon bioinformatics has given us to continue this fight against human disease.

Thank you.

TEACHER PAGES: SUMMARY QUESTIONS AND ANSWERS LESSON ONE

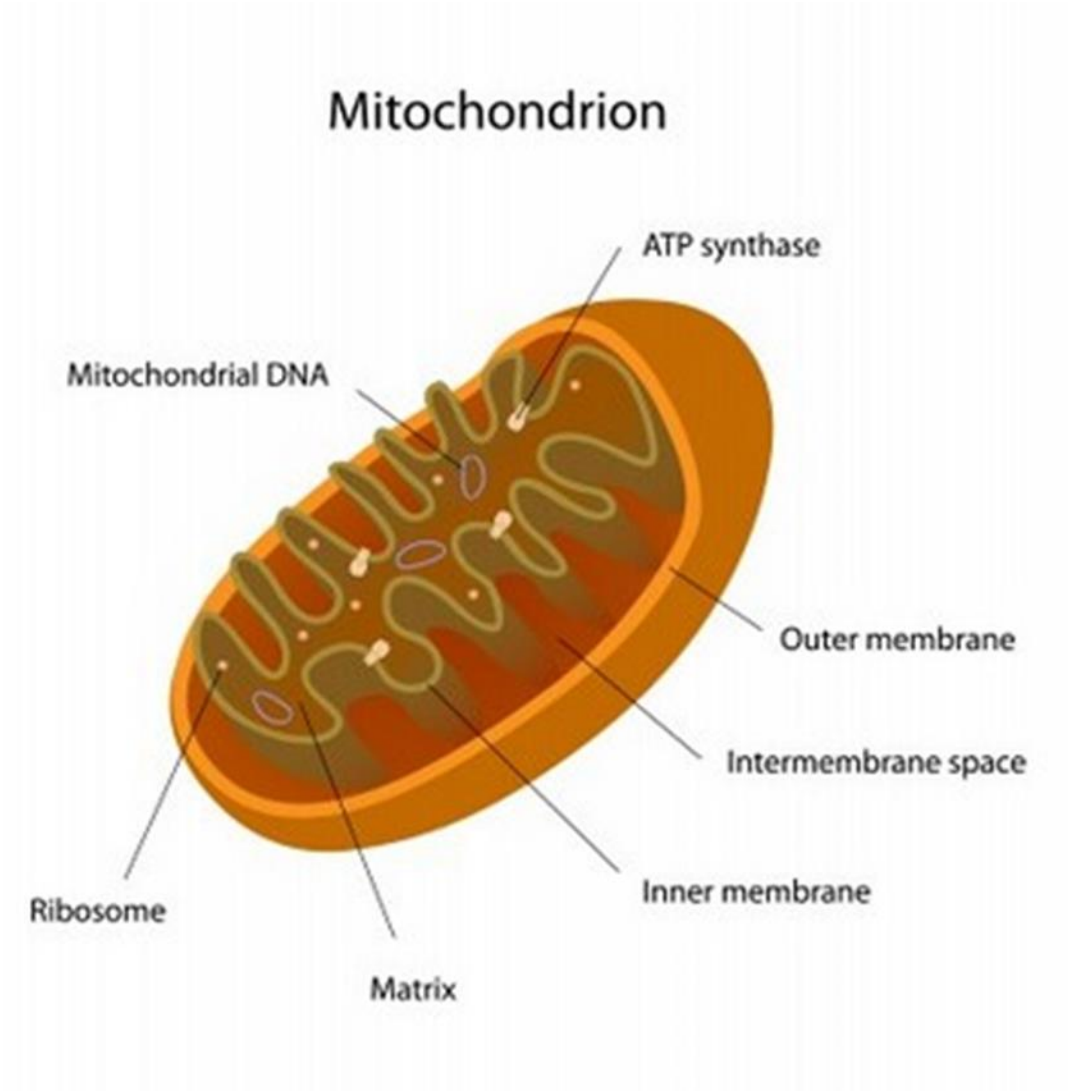
SUMMARY QUESTIONS

1. Do you think Spencer Hall may have some bias towards the importance of bioinformatics in curing deadly disease? Explain your reasoning. (This would be a good time to bring up bias in science. Discussions could involve whether or not bias is always a negative thing. What are some things scientists do to decrease bias? Is bias inherent in all we do?)
2. Draw a mitochondrion and annotate the structures and their function. Use IB draw rules. (Be sure to include two phospholipid bilayers with the inner one folded into cristae. The cristae need to be thin, both should be labeled. Other structures should include the intermembrane space, matrix, ribosomes, DNA. Key words IB moderators will be looking for – ETC is on inner membrane – makes energy in the form of ATP. Do not just say energy – no point will be given. The answer must include ATP. The purpose of the folding of inner membrane is to increase surface area for increased ATP production. The intermembrane space is small so that a proton gradient can be established rapidly. Remember also that the lines representing inner and outer membranes must be continuous with no breaks and they cannot touch each other – no points will be given if they are not drawn like this. Lines must be straight and touch the object being identified. Labels must be horizontal. Do not cross lines.)
3. Write a paragraph explaining how ATP is formed in cellular respiration. (This assumes an 8 mark question. Points should be awarded for the following: Glucose is converted to two molecules of pyruvate in glycolysis, pyruvate enters mitochondria and converted to acetyl CoA by decarboxylation. NADH and CO_2 if formed. Fatty acids can be converted to acetyl CoA. Acetyl groups enter Krebs cycle. FAD/NAD+ accepts hydrogen for from NADH and $FADH_2$. These two molecules donate electrons (cannot accept donates H+)) to electron transport chain. Electrons release energy as they pass along the chain. Oxygen is the final electron acceptor and produces water. Protons are pumped across the inner membrane and build up a proton gradient. Protons flow into the matrix through ATPase and ATP is produced. Produces 36/38 ATP per molecule of glucose.)

EXTENSIONS

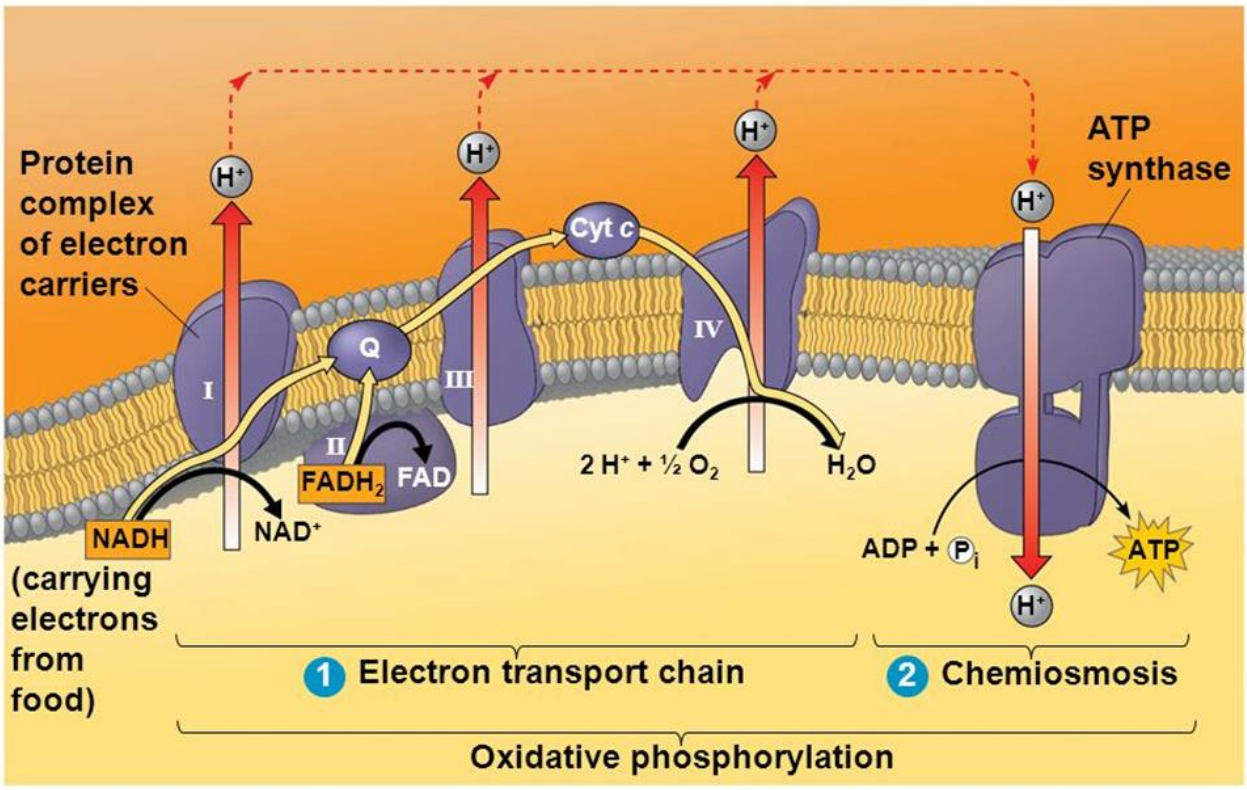
1. Give students a blank sheet of paper to draw, label and annotate a mitochondrion. Use above to score.

TEACHER PAGES: MITOCHONDRION, ELECTRON TRANSPORT CHAIN AND MOLECULAR STRUCTURE OF CYTOCHROME C FOR REVIEW

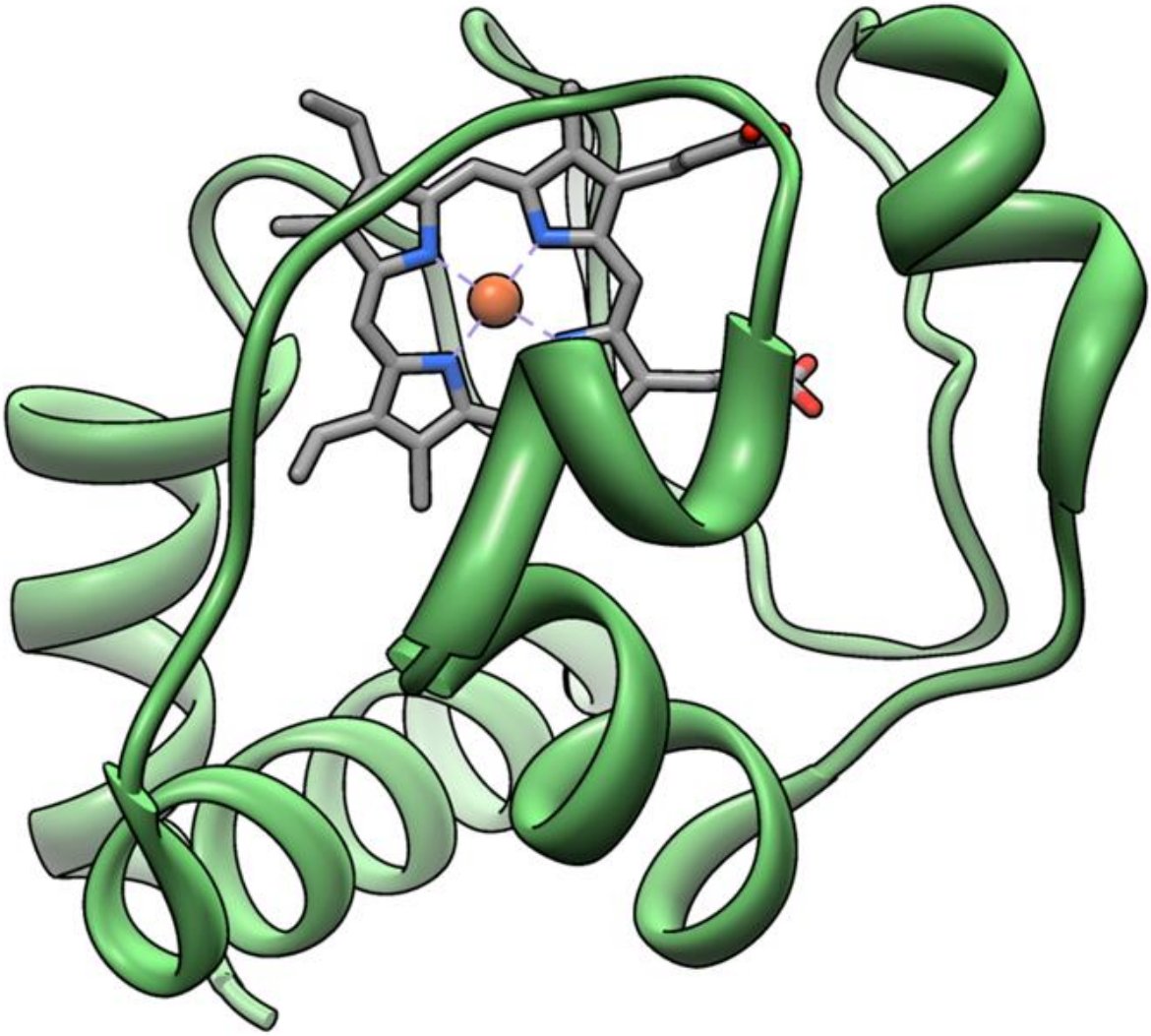


ELECTRON TRANSPORT CHAIN

Figure 9.15



MOLECULAR STRUCTURE OF CYTOCHROME C



STUDENT WORKSHEET: SUMMARY QUESTIONS

Name_____Date_____Period_____

Answer the following questions after the review.

1. Do you think Spencer Hall may have some bias towards the importance of bioinformatics in curing deadly disease? Explain your reasoning.

2. Draw a mitochondrion and annotate the structures and their function. Use IB draw rules. (5 marks)

3. Write a paragraph explaining how ATP is formed in cellular respiration. (8 marks)

LESSON TWO: NAVIGATING DATABASES

OVERALL TIME ESTIMATE:

- Advanced Preparation: 30 minutes
- Student Procedure: 90 minutes (2 days)

LEARNING STYLES: Visual, auditory, and kinesthetic.

VOCABULARY:

Bioinformatics: A field in science where complex biological data is collected, stored and analyzed

Cytochrome C- A protein found in mitochondria and carries electrons from one molecule to the next. This is vital to cellular respiration.

Exon: The term used to represent any part of a gene that becomes part of the processed mRNA after the introns have been removed.

Intron: A segment on a DNA or RNA that does not code for a protein and is removed in RNA processing.

RNA: Ribonucleic acid is one of the three major macromolecules (along with DNA and proteins) that are essential for all known forms of life. Like DNA, RNA is made up of a long chain of components called nucleotides. Each nucleotide consists of a nucleobase, a ribose sugar, and a phosphate group. RNA directs the synthesis of proteins.

LESSON SUMMARY:

Working in pairs, students will use the databases of NCBI, BLAST, In-Silico, EBI and Uniprot to find cytochrome c of human, chimpanzee and mouse. They will be able to locate the gene of interest on the chromosome, identify the size of the gene in base pairs as well as the number of exons. Students will then align sequences of two or more species to note the differences. Using these differences, students will create a cladogram.

STUDENT LEARNING OBJECTIVES:

The student will be able to...

1. Use the database from UCSC to find a gene of interest on two species, *Homo sapiens* and *Mus muculus.*
2. Using a cladogram, determine the closest living relatives of both species.
3. Using the gene of interest, determine on which chromosome and position the gene is located, determine how many base pairs are included in the gene and how many EXONS and INTRONS are located on that gene.
4. Determine which species has the larger gene.
5. Using one EXON, determine the base pairs in that EXON.
6. Using the database from NCBI, find the DNA sequence for a gene of interest.
7. Using the database from In-silico, convert the DNA sequence from the gene of interest to mRNA.
8. Using the database Emboss, convert the mRNA sequence from the gene of interest to amino acids.
9. Identify the amino acids found in the gene of interest by using an Amino Acid chart and the IUPAC code for each.
10. Use BLAST to align two or more sequences and determine the percentage that are the same.

11. Examine the alignment to find the differences between the two species.
12. Examine a cladogram from the differences in DNA sequences between several organisms.
13. Use Uniprot to see the molecular structure of cytochrome c.
14. Use the above databases to find different genes from different organisms to examine as above.

STANDARDS: See table on page 9-10.

2.4 Proteins
Understanding: The amino acid sequence determines the three-dimensional conformation of a protein
2.4 Proteins
Aim 7: ICT can be used for molecular visualization of the structure of proteins
3.1 Genes
Understanding: A gene occupies a specific position on a chromosome.
3.1 Genes
Skill: Use of a database to determine differences in the base sequence of a gene in two species.
3.2 Chromosomes
Skill: Use of databases to identify the locus of a human gene and its polypeptide product.
5.4 Cladistics
Understanding: Cladograms are tree diagrams that show most probable sequence of divergence in clades.
B.5 Bioinformatics
Understanding: Databases allow scientists easy access to information.
B.5 Bioinformatics
Understanding: The body of data stored in databases is increasing exponentially.
B.5 Bioinformatics
Understanding: BLAST searches can identify similar sequences in different organisms.
B.5 Bioinformatics
Understanding: Sequence alignment software allows comparison of sequences from different organisms.
B.5 Bioinformatics
BLASTn allows nucleotide sequence alignment while BLASTp allows protein alignment.

MATERIALS:

1. Computer – one per two or three students.
2. Internet Access
3. Copy of student worksheets with questions
4. Copy of sequences used. These can be posted where students can have access to cut and paste-this is in case students are unable to navigate the databases.

BACKGROUND INFORMATION: Background information needed for this assignment is at the beginning of the guide.

ADVANCE PREPARATION:

1. Teachers should have visited each database and have them open on the desktop of the teacher computer.
2. Be prepared to project each website as students navigate through them and showing them key features of each website.
3. Place copies of sequences where students can cut and paste if needed (Teacher WebPage, Edmodo, Moodle)
4. Make copies of student worksheets.

PROCEDURE AND DISCUSSION QUESTIONS WITH TIME ESTIMATES:

Part A: Using UCSC Genome Browser to find a gene of interest and on which chromosome it is located (20 minutes).

1. Go to www.Genome.ucsc.edu/index.html
2. On the home page, click Genomes located in the task bar at the top.
3. Be sure that "Human" is clicked under popular species at the left.
4. Examining the cladogram below answer the following question.

a) What are the three animals most similar to humans? (Chimpanzee, Bonobo, Gorilla)

5. Click Genome Browser located in the task bar at the top-located next to Genomes
6. Search "CYCS *homo sapiens*"
7. Click on the first one – CYCS *Homo sapiens* cytochrome c, somatic
8. At the top, under the search bar, you will see a pictorial representation of a chromosome. The red line shows the position of the gene for cytochrome c. To the left of the chromosome you will see the position number and above the search bar you will find the number of the beginning bp and the ending bp. The size is there as well in bp.

b) On what chromosomes is Cytochrome C located? (7)
c) On what arm of the chromosome is Cytochrome C located? (p)
d) What is the position? (p 15.3)
e) What is the starting position? (25,120,091)
f) What is the ending position? (25,125,361)
g) How many base pairs are found in Cytochrome C? (5,271 bp)

9. You will look at the very first dark line. There should be a dark box with white lettering to the left of the line. The letters are CYCS.
10. Placing your cursor over each of the larger, black boxes will tell you the EXON number.
11. Placing your cursor over the longer, thinner lines will tell you the INTRON number.

h) How many EXONS are found in Cytochrome C? (3)
i) How many INTRONS are found in Cytochrome C? (2)

12. Zoom out (on the right side of the tool bar at the top) to be sure you see all exons. Try 1.5 – you may have to do this more than once.
13. Place the cursor at the beginning of EXON 1 and highlight that exon. (This is a bit tricky – your cursor needs to be at the top of the chart – above the little numbers that are above EXON 3). Click "zoom out" when the box appears. You may have to do this more than once – you are looking for the base pairs found in this EXON.
14. You will see the base pairs for EXON 1. Below the base pairs you will see the IUPAC code or the amino acids found in this protein.
15. Go back to Genomes and click on Mouse. Examining the cladogram underneath, what are the four animals most closely related to the mouse? (rat, Chinese hamster, kangaroo rat, squirrel)
16. Go to Genome Browser and search for CYCS *mus musculus*. (you must use scientific names in order to search for genes for specific animal)

j) On which chromosome is the gene for cytochrome c located? (6)
k) On what arm of the chromosome is t located (q)
l) What is the starting position? (50,561,522) The ending position? (50,567,389)
m) How many base pairs does it contain? (5868 bp)

n) Which gene is larger, the human or mouse? (mouse)
o) How many base pairs larger? (597)

EXTENSION: Have students search for other animals in the genome and answer the same questions – they can compare two non-human species.

Part B: Using NCBI to find DNA sequences of gene of interest (20 minutes)

1. Go to https://www.ncbi.nlm.nih.gov
2. Click on Nucleotide – you will find this on the right side of the page under Popular Resources
3. In the search bar at the top, be sure the left sided box is set to "Nucleotide." In the search bar type "cytochrome c *homo sapiens*
4. Click on number 5: *Homo sapiens* cytochrome c, somatic (CYCS) and click GenBank found underneath. Scroll down to find the summary of the gene.

a) For what does cytochrome encode? (a small heme protein)
b) Where is cytochrome c found? (in inner membrane of mitochondria)
c) What is the function of cytochrome c? (accepts electrons from cytochrome b and transfers them to the cytochrome oxidase complex)
d) What is another function? (initiates apoptosis)
e) What disease is caused by a mutation in cytochrome c? (autosomal dominant nonsyndromic thrombocytopenia)

5. Scroll to the top of page and click FASTA

6. This is the DNA sequence for that protein. Copy and paste it into a word document. Be sure to include the first line beginning with >gi. Label this DNA sequence-Cytochrome C *Homo sapiens*

Part C: Using In-silico to convert DNA to mRNA (5 minutes)

1. Go to http://in-silico.net/tools/biology/sequence_conversion
2. In the search box, copy and paste the DNA sequence of cytochrome c you saved from the previous activity.
3. Click DNA -> RNA
4. This now will be your mRNA sequence from your DNA sequence.
5. Copy and paste this into the same word document. Label this mRNA sequence Cytochrome C *Homo sapiens.*

Part D: Using Emboss to convert mRNA to amino acids (30 minutes)

1. Go to http://www.ebi.ac.uk/tools/st/emboss_transeq/
2. Copy and paste the mRNA you saved. DO NOT INCLUDE THE BEGINNING LINE STARTING WITH >GI.
3. You do not need to use your email or change any other parameters.
4. Click submit.
5. This will be your amino acid sequence for cytochrome C.
6. Copy and paste this into the same word document and label amino acid sequence cytochrome c *Homo sapiens*.

a) Using the amino acid chart, determine what amino acids are found in the first line of your protein. The *means that amino acid may vary – just ignore it for now.
(VAASSAARQGAGARSEFGCTYTGT*ARTGVSLDLESGDVRLRSGSVRCASD*KEN*IWVM)(Valine, Alanine, Alanine, Serine, Serine, Alanine, Alanine, Arginine, Glutamine, Glycine, Alanine, Glycine, Arginine, Serine, Glutamic Acid, Phenylalanine, Glycine, Cysteine, Threonine, Tyrosine, Threonine, Glycine, Threonine- *ignore-, Alanine, Arginine, Threonine, Glycine, Valine, Serine, Leucine, Aspartic Acid, Leucine, Glutamic Acid, Serine, Glycine, Aspartic Acid, Valine, Arginine, Leucine, Arginine, Serine, Glycine, Serine, Valine, Arginine, Cysteine, Alanine, Serine, Aspartic Acid, * , Lysine, Glutamic Acid, Asparagine, Isoleucine, Tryptophan, Valine, Methionine)

Part E: Using Blast- Part A (20 minutes)

1. Go to http://ncbi.nlm.nih.gov
2. On the right hand side, click on Nucleotide
3. In the search bar, type CYCS *Pan troglodytes* (remember you must use scientific names – using chimpanzee will result in an error message)
4. Click on the number three, *Pan troglodytes* cytochrome c, somatic and click FASTA below
5. Copy and paste the sequence into the same word document where you placed the human sequences. Label it cytochrome c *Pan troglodytes*.
6. Go to http://blast.ncbi.nlm.gov/Blast.cgi
7. Click on "align two or more sequences." You will find this beneath the search box. Clicking on this will cause a second box to open. Clear the search boxes to be sure you are only searching for what you want.
8. In the first box, copy and paste the *Homo sapiens* cytochrome c sequence your saved earlier. In the second box, copy and paste the *Pan troglodyte's* cytochrome sequence you saved earlier.
9. Click on BLAST at the bottom of the page. Be patient, this may take a few seconds.
10. Scroll to the bottom of the page and you will see the alignment of the two sequences. The first line is what you put into the first box -*Homo sapiens* (called Query). The second line is what you put in your second box-*Pan troglodytes* (called Subject).

a) What percentage of the two strands are identical? (99%)
b) Looking at the first five lines of comparison, how many differences do you find? (4) If you examine the third line, Query 291, subject 121, how many differences do you see? (1)
c) What is it? (Human with a C and chimp with a T)

Part F: Using Blast- Part B (20 minutes)

1. Now let us compare more than two sequences. Go to http://ncbi.nlm.nih.gov.
2. Search for cytochrome c *Canis familiaris*. Pick number 1 and FASTA.
3. Copy and paste the sequence into your word document labeling it cytochrome c, *Canis familiaris*.
4. Repeat the search as above using *Mus musculus.* Choose number three, FASTA and then copy and paste the sequence into your word document. Label it cytochrome c, *Mus musculus*.
5. Go to BLAST, http://blast.ncbi.nlm.nih.gov/Blast.cgi
6. Click on Nucleotide. This is on the left side of the page.
7. Under the first box, you will see "Align two or more sequences." Click the box to the left of this and a second box should appear.

8. Go back to your word document and copy and paste the DNA sequence for cytochrome c for *Homo sapiens* into the first box. This is your Query. IT IS VERY IMPORTANT YOU GET THE WHOLE SEQUENCE AND IT BEGINS WITH THE >gi.
9. From your word document, copy and paste the DNA sequence for cytochrome c for *Pan troglodytes* in the second box. IT IS VERY IMPORTANT YOU GET THE WHOLE SEQUENCE AND IT BEGINS WITH THE >gi.
10. You can then copy and paste the other sequences for the other organisms in the next line in the second box. BLAST looks for the >gi to tell it that this is a new sequence. All three organisms can go in the second box-*Pan troglodytes, Mus musculus, Canis familiaris*.
11. Click BLAST. Be patient, it may take a few seconds.
12. Scroll down and look at the first box. Remember Query is what you put in the first box – in this case Homo sapiens. The subject is identified at the top of the box.

a) Looking at the comparison between *Homo sapiens* and *Canis familiaris*, what percentage is identical? (81%)
b) Looking at the comparison between *Homo sapiens* and *Pan troglodytes*, what is percentage is identical? (99%)
c) Looking at the comparison between *Homo sapiens* and *Mus musculus*, what is percentage is identical? (85%)

13. Look at the box above the first alignment sequences. You should see "Sequences producing significant alignments. Click the boxes next to each description and then click on "Distance Tree of Results' located at the top of the box.
13. You should now see the phylogenic tree.

d) Which species is most closely related to the *Canis familiaris*? (*Mus musculus*)
e) Beginning with the species most closely related to *Homo sapiens*, list the species in the tree from most closely related to furthest. (*Pan troglodytes, Mus musculus, Canis familiaris)*

Part G: Molecular Visualization (15 minutes)

1. Go to the following website. http://www.uniprot.org/
2. In the search bar at the top, type CYCS (the abbreviation for cytochrome c).
3. Choose the first one and click on the entry number (P99999).
4. On the left hand side, you will see a table – the top word is "function." Above the table, click on the word "none" and then click only structure.

   a) How many alpha helixes are present in cytochrome c? (5)
   b) How many beta pleated sheets are in cytochrome c? (3)

5. Look under 3D structure data bases and under Entry, click the first light blue letters/numbers -1J3S.
6. Double click on the picture of the molecular structure of cytochrome c.
7. Click through the pictures looking at the protein from different angles. You should clearly see the alpha helixes, beta pleated sheets as well as the heme portion of the molecule.

References

EMBOSS Transeq. (n.d.). Retrieved July 13, 2016, from http://www.ebi.ac.uk/tools/st/emboss_transeq/.

In-silico. (n.d.). Retrieved July 13, 2016, from http://in-silico.net.

National Center for Biotechnology Information. (n.d.). Retrieved July 13, 2016, from http://www.ncbi.nlm.nih.gov.

The UniProt Consortium,UniProt: a hub for protein information, Nucleic Acids Res. 43: D204-D212 (2015). Retrieved July 13, 2016 from www.uniprot.org.

UCSC Genome Browser: Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006. Retrieved July 13, 2016 from www.genome.ucsc.edu.

EXTENSIONS:

ACTIVITIES:

- Students can use the programs above using organisms of their own choosing. It is important that they use scientific names.
- Students can also use other proteins. Smaller proteins are better than large ones. Some examples would be lactase, insulin, catalase, myosin.

TEACHER NOTES:

1. When searching for proteins, be sure you do not use Predicted. Try to find the proteins that is somatic. You may need to scroll down a to find it. Also, look at the size. Try not to pick proteins that are really large as they will take a long time to download and BLAST.
2. Another common mistake is to choose the "Chromosome." This will download the whole chromosome. Don't' do this – it will take an extremely long time and will not give you what is needed in this lesson. All the proteins included should only take a few seconds to download. If it is taking longer than that, the protein is too large and you did not choose the correct one.
3. When cutting and pasting sequences, be sure you get the first line of the sequence – beginning with >gi EXCEPT for EMBOSS. If you get an error message while using EMBOSS, it is probably due to the first line – delete it and begin with only the actual sequence.
4. Always use scientific names. While some will take common names, others will not. They all take scientific names.
5. When aligning two or more sequences in BLAST, you may encounter an error message – "Not enough likeness." This means that the differences are too great. For example, Human and fruit fly. You can change some things to help. At the bottom of the page, under the BLAST button is Algorithm parameters. Click on the + and change the Expect Threshold. Default is 10 but you can up to 100 or 1000. Also, under Program Selection (found above the BLAST button), click "More dissimilar sequences (discontiguous megablast). This will then allow you to see the differences.
6. Let the students "play." They cannot really do any harm and remember – The back button is your friend!

STUDENT PAGE: NAVIGATING DATABASES-PROCEDURES AND QUESTSIONS

Part A: Using UCSC Genome Browser to find a gene of interest and on which chromosome it is located.

1. Go to www.Genome.ucsc.edu/index.html
2. On the home page, click Genomes located in the task bar at the top.
3. Be sure that "Human" is clicked under popular species at the left.
4. Examining the cladogram below answer the following question.

    a) What are the three animals most similar to humans?

_____

_____

5. Click Genome Browser located in the task bar at the top-located next to Genomes
6. Search "CYCS *Homo sapiens*"
7. Click on the first one – CYCS *Homo sapiens* cytochrome c, somatic
8. At the top, under the search bar, you will see a pictorial representation of a chromosome. The red line shows the position of the gene for cytochrome c. To the left of the chromosome you will see the position number and above the search bar you will find the number of the beginning bp and the ending bp. The size is there as well in bp.

    b) On what chromosomes is Cytochrome C located? _____

    c) On what arm of the chromosome is Cytochrome C located? _____

    d) What is the position? _____

    e) What is the starting position? _____

    f) What is the ending position? _____

    g) How many base pairs are found in Cytochrome C? _____

9. You will look at the very first dark line. There should be a dark box with white lettering to the left of the line. The letters are CYCS.
10. Placing your cursor over each of the larger, black boxes will tell you the EXON number.
11. Placing your cursor over the longer, thinner lines will tell you the INTRON number.

    h) How many EXONS are found in Cytochrome C? _____

    i) How many INTRONS are found in Cytochrome C? _____

12. Zoom out (on the right side of the tool bar at the top) to be sure you see all exons. Try 1.5 – you may have to do this more than once.

13. Place the cursor at the beginning of EXON 1 and highlight that exon. (This is a bit tricky – your cursor needs to be at the top of the chart – above the little numbers that are above EXON 3). Click "zoom out" when the box appears. You may have to do this more than once – you are looking for the base pairs found in this EXON.

14. You will see the base pairs for EXON 1. Below the base pairs you will see the IUPAC code or the amino acids found in this protein.

15. Go back to Genomes and click on Mouse.

    k) Examining the cladogram underneath, what are the four animals most closely related to the mouse?

_____

_____

16. Go to Genome Browser and search for CYCS *Mus musculus*. (you must use scientific names in order to search for genes for specific animal)

    l) On which chromosome is the gene for cytochrome c located? _____

    m) On what arm of the chromosome is it located? _____

    n) What is the starting position? _____The ending position? _____

    o) How many base pairs does it contain? _____

    p) Which gene is larger, the human or mouse? _____

    o) How many base pairs larger? _____

Part B: Using NCBI to find DNA sequences of gene of interest

1. Go to https://www.ncbi.nlm.nih.gov

2. Click on Nucleotide – you will find this on the right side of the page under Popular Resources

3. In the search bar at the top, be sure the left sided box is set to "Nucleotide." In the search bar type "cytochrome c *Homo sapiens*

4. Click on number 5: *Homo sapie*ns cytochrome c, somatic (CYCS) and click GenBank found underneath. Scroll down to find the summary of the gene.

    a) For what does cytochrome encode? _____

    b) Where is cytochrome c found? _____

c) What is the function of cytochrome c?

_____

_____

d) What is another function? _____

e) What disease is caused by a mutation in cytochrome c?

_____

_____

5. Scroll to the top of page and click FASTA

6. This is the DNA sequence for that protein. Copy and paste it into a word document. Be sure to include the first line beginning with >gi. Label this DNA sequence-Cytochrome C *Homo sapiens*

Part C: Using In-silico to convert DNA to mRNA

1. Go to http://in-silico.net/tools/biology/sequence_conversion
2. In the search box, copy and paste the DNA sequence of cytochrome c you saved from the previous activity.
3. Click DNA -> RNA
4. This now will be your mRNA sequence from your DNA sequence.
5. Copy and paste this into the same word document. Label this mRNA sequence Cytochrome C *Homo sapiens.*

Part D: Using Emboss to convert mRNA to amino acids

1. Go to http://www.ebi.ac.uk/tools/st/emboss_transeq/
2. Copy and paste the mRNA you saved. DO NOT INCLUDE THE BEGINNING LINE STARTING WITH >GI. You do not need to use your email or change any other parameters.
3. Click submit.
4. This will be your amino acid sequence for cytochrome C.
5. Copy and paste this into the same word document and label amino acid sequence cytochrome c *Homo sapiens*.

a) Using the amino acid chart, determine what amino acids are found in the first line of your protein. The *means that amino acid may vary – just ignore it for now.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Part E: Using Blast- Part A (20 minutes)

1. Go to http://ncbi.nlm.nih.gov
2. On the right hand side, click on Nucleotide
3. In the search bar, type CYCS *Pan troglodytes* (remember you must use scientific names – using chimpanzee will result in an error message)
4. Click on the number three, *Pan troglodytes* cytochrome c, somatic and click FASTA below
5. Copy and paste the sequence into the same word document where you placed the human sequences. Label it cytochrome c *Pan troglodytes*.
6. Go to http://blast.ncbi.nlm.gov/Blast.cgi

7. Click on "align two or more sequences." You will find this beneath the search box. Clicking on this will cause a second box to open. Clear the search boxes to be sure you are only searching for what you want.
8. In the first box, copy and paste the *Homo sapiens* cytochrome c sequence your saved earlier. In the second box, copy and paste the *Pan troglodyte's* cytochrome sequence you saved earlier.
9. Click on BLAST at the bottom of the page. Be patient, this may take a few seconds.
10. Scroll to the bottom of the page and you will see the alignment of the two sequences. The first line is what you put into the first box -*Homo sapiens* (called Query). The second line is what you put in your second box-*Pan troglodytes* (called Subject).

    a) What percentage of the two strands are identical? _____

    b) Looking at the first five lines of comparison, how many differences do you find? (4) If you examine the third line, Query 291, subject 121, how many differences do you see? ____

    c) What is it? _____

Part F: Using Blast- Part B

1. Now let us compare more than two sequences. Go to http://ncbi.nlm.nih.gov.
2. Search for cytochrome c *Canis familiaris*. Pick number 1 and FASTA.
3. Copy and paste the sequence into your word document labeling it cytochrome c, *canis familiaris.*
4. Repeat the search as above using *Mus musculus*. Choose number three, FASTA and then copy and paste the sequence into your word document. Label it cytochrome c, *Mus musculus*.
5. Go to BLAST, http://blast.ncbi.nlm.nih.gov/Blast.cgi
6. Click on Nucleotide. This is on the left side of the page.
7. Under the first box, you will see "Align two or more sequences." Click the box to the left of this and a second box should appear.
8. Go back to your word document and copy and paste the DNA sequence for cytochrome c for *Homo sapiens* into the first box. This is your Query. IT IS VERY IMPORTANT YOU GET THE WHOLE SEQUENCE AND IT BEGINS WITH THE >gi.
9. From your word document, copy and paste the DNA sequence for cytochrome c for *Pan troglodytes* in the second box. IT IS VERY IMPORTANT YOU GET THE WHOLE SEQUENCE AND IT BEGINS WITH THE >gi.
10. You can then copy and paste the other sequences for the other organisms in the next line in the second box. BLAST looks for the >gi to tell it that this is a new sequence. All three organisms can go in the second box-*Pan troglodytes, Mus musculus, Canis familiaris*.

11. Click BLAST. Be patient, it may take a few seconds.

12. Scroll down and look at the first box. Remember Query is what you put in the first box – in this case *Homo sapiens*. The subject is identified at the top of the box.

    a) Looking at the comparison between Homo sapiens and *Canis familiaris*, what percentage is identical? _____

    b) Looking at the comparison between *Homo sapiens* and *Pan troglodytes*, what is percentage is identical? _____

    c) Looking at the comparison between *Homo sapiens* and *Mus musculus*, what is percentage is identical? _____

13. Look at the box above the first alignment sequences. You should see "Sequences producing significant alignments. Click the boxes next to each description and then click on "Distance Tree of Results' located at the top of the box.

13. You should now see the phylogenic tree.

    d) Which species is most closely related to the *Canis familiaris*? _____

    e) Beginning with the species most closely related to *Homo sapiens*, list the species in the tree from most closely related to furthest.

_____

_____

Part G: Molecular Visualization

1. Go to the following website. http://www.uniprot.org/
2. In the search bar at the top, type CYCS (the abbreviation for cytochrome c).
3. Choose the first one and click on the entry number (P99999).
4. On the left hand side you will see a table – the top word is "function." Above the table, click on the word "none" and then click only structure.

    a) How many alpha helixes are present in cytochrome c? _____

    b) How many beta pleated sheets are in cytochrome c? _____

5. Look under 3D structure data bases and under Entry, click the first light blue letters/numbers -1J3S.
6. Double click on the picture of the molecular structure of cytochrome c.
7. Click through the pictures looking at the protein from different angles. You should clearly see the alpha helixes, beta pleated sheets as well as the heme portion of the molecule.

TEACHER PAGES: AMINO ACID CHART AND SEQUENCES

AMINO ACID	ABBREVIATION	IUPAC SYMBOL
Alanine	Ala	A
Cysteine	Cys	C
Aspartic Acid	Asp	D
Glutamic Acid	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histadine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

Retrieved and adapted from http://www.sigmaaldrich.com/life-science/metabolomics/learning-center/amino-acid-reference-chart.html

SEQUENCES FOR LESSON TWO

Cytochrome C -*Homo sapiens*

>gi|300863084|ref|NM_018947.5| Homo sapiens cytochrome c, somatic (CYCS), mRNA

GTAGCCGCCAGCTCGGCCGCACGTCAGGGCGCGGGAGCGCGGAGCGAGTTTGGTTGCACTTACACCGGTA

CTTAAGCGCGGACCGGCGTGTCCTTGGACTTAGAGAGTGGGGACGTCCGGCTTCGGAGCGGGAGTGTTCG

TTGTGCCAGCGACTAAAAAGAGAATTAAATATGGGTGATGTTGAGAAAGGCAAGAAGATTTTTATTATGA

AGTGTTCCCAGTGCCACACCGTTGAAAAGGGAGGCAAGCACAAGACTGGGCCAAATCTCCATGGTCTCTT

TGGGCGGAAGACAGGTCAGGCCCCTGGATACTCTTACACAGCCGCCAATAAGAACAAAGGCATCATCTGG

GGAGAGGATACACTGATGGAGTATTTGGAGAATCCCAAGAAGTACATCCCTGGAACAAAAATGATCTTTG

TCGGCATTAAGAAGAAGGAAGAAAGGGCAGACTTAATAGCTTATCTCAAAAAAGCTACTAATGAGTAATA

ATTGGCCACTGCCTTATTTATTACAAAACAGAAATGTCTCATGACTTTTTTATGTGTACCATCCTTTAAT

AGATCTCATACACCAGAATTCAGATCATGAATGACTGACAGAATATTTTGTTGGGCAGTCCTGATTTAAA

ACTAAGACTGGCTTGTGGTTAAATGAATATGTTCAGTTTTTGAATTTTAATAGTAACTCCAATTCAGTAA

ATGGTATCACTGTTTACCCCTTTTAAAGATATGATTAGACTTCGTTAGTAATGTTCAACTTTTCACAAAG

ATGGTGAGTGCCATCTTAAAACTTACTGGAGATTGGTTTTATATTTAGATTTATATAACTGGTTATGTGA

ATATATTTAAATACTGGGGAAATTGCTTCACTGTCTTAGAACCAAGCAAGATTCACCTGTGTTTTGTGTT

CATGTTCATTTGCCTCTTAAAGGCAAGGGTTGAAGATAAATAAGGTAGCAATGTCTATAGTTTTGGCCTT

AACTATGCCAATCTAATTATAATTCCCTGTATTTAAAATGGTTTCTTTTACTTATTGAAAGGCATTTTAG

TGTGGTTTATGTGTAATATTAAAGATTATTCAACACCTCTCACATCTTACAGATCTATAAGGTCACATGC

TTTTAAAATAGTAGCAAGTTAAACTTCACTCTTGAATTCTTTACAATCTAAGTCAAACTAAGTTATAATT

TAGGATTGTCTTTAAACAGCCATTCAGAAACAAAACTGTAGAACTGTGTATTTGATTGGGAATGGTGCTT

TTGCCAACTTAAAAGGATTAAAGTAACGGAGATATACACAAATTTTAAAATTATGTGTGATCACAAGACT

AAAGATAATTAAAAAGAAAACCACAGATCATGACTTTTTGACTGTGCTTGATTTCATGACTGATGCACAA

ATTTTAATGATTAAAAAGTGCAGGAGCCCTAAATGTCAGTGCAGCAGCCCTAAATGTCAGTGCAGCAGTG

TTAACCAGTCATGGTGCTAGATTGTTTACTTGGTTTTCTAGGACTGCCTCAACTAGAATAACACTTCACT

AATTGACTCTTAGTTTCTTTGCTCAGATTGAGAACTGCAGCATTTATGCCAGACATGGACAGAGGAATGC

CTGTGGTCATAGTTTTGTGATGTGTAACAGTGTATAATTACATACTGAATTATTTCATGCATAGTCTGTG

CCATACACATTTAGAGTAGTCCTTGGAGATTTTATGGAGATGGTGAGCACAAGGTAAGTCATAAAGAATA

ATGAGAAAATAAATCTATGCTGGTGCAGCTGAGAACTGTATCTTTGTGGGACAGTGAGAAGACTGAGAAG

ATGTGAATCCATGGTCTCAAAGGTGATAGGGACGATTAGATAGGTGTTTTAAGGCCTGAAAGCAATTTAT

AACATATGAGTCTTATTTTTATTTATAGAAATGTGGAAAGCTTGCTGTAATTCATATTTGAAGTCCTAGT

CTGAGTTCTGGTGGGGAATTTAAAAATGCATCCTGGAAATCCTTTAAAGATTTCAGACTTTGAAAGGCCT

TGTAGCAGAGGACTTGGTGACTGTATAAAGTTAGTGGTATTCAGGGACAGTGTAGCAAGTAGCTTACAAG

GGGACAATTCTGGACTAATGAGAAAGACCTGAAGTGAAGGCTAGAGAGTTGATTTTTTTTTTTTTTGGCA

TCCTGGAAATGATACAGGAAACATATTAAGATAGATACAGAAATGTGTTCAACCTTCCATCTTGGCTAGT

TGTGGCGTTTAGTTTGTTTTTTGAGACATGGTCACGCTGTGTCGCCAAGGCTGGAGTGCAGTGGTGCGAT

CTCGGCTGGCTGCAACCTCTATTTCCCAGGCTCAAGCGATTCTCTCACTTCAGCTTCCCAAGTAGCTGGG

ACTACAGGTGTTCGCCACCATGCCCAGCTAATTTTTTTGTAGAGATGGAGTTTTGCCATGTGGTCCCAGG

CTGGTCTCAAACTCCTGAGCTCAAGCAATCCGTCCACTTGCCTTGGCTCCCCAAAGTGCTGGGATTACAG

GCGTGAGCCACCAGGCCCTGCCTGGTTTTCAAATTCAGAAATCTTATTATTTAACCCAGAAGTAATCAGC

CCAGTAGTAACTTAGGTTTAATTTTTTTTCAGGTTTAAAATTTTTCTCATTTATTTTTTCTGAGACGGAG

TTTCGCCCTTTTCGCCCAGGCTGAGTACAGTGGTGCAATCTCACTGCAACCTCCGCCTTCCAGTTGCAAG

TGATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGATTACAGGCACCCGCCACCACGCCTGGCTAATTTTT

GTATTTTTAGTGGAGATGGTGTTTCACCATGTTGGCCAGACTGGTCTTGGACTCCTGACCTCGTGATCCA

CCCACCTTGGCCTCCCAAAGTTCTAGGATTACAGGTGTGAGCCACCACGTCCGGCCAATTTTTCTCATTT

CTATGCCTCCTATATTAAGGTCTGTGTTGGCACAGATGAGTAACTGCCATGTTCTAGGTCAGTTATACCC

AAGCACTTCTGGTGGTTTAAAATGTGATTCTGTAACTTTTTTATTTTTATTTTTTTGAGATAATTTCACT

CTTGTTGCCCAGGCTGGAGTGCAATGGCGTGATCGCTGCTCACCGCAACCTCCGCCTTCCAGGTTCAAGC

GATTCTCCTGACTCAGCCTCTCAAGTAGCTGGGATTACAAGCATGCGCCACCATGCCCAGCTTATTTTGT

GTTTTTAGTAGAGACAGGGTTTCTCCATGCTGGACAGGCTGGTCTTGAACTCCCGACGTCAGGTGATCTA

CCTGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACCACGCCTGGCCAATTATGTAATTTT

TTAAAAAGGACATTTCTATCAGGGATATATACCTTCAGAAATAAGGAAATAGGGGAAAAAAAGAGCACTA

TAAACCACATGTTTTCATTTCTAGTGCTTCGCTGTAAGTGGCTAGGTTGGTAGAATCAAAAACAAGGGCC

AGATGTATTTAAGGGGTATTCAGATGCCACCTACATGCTTATTTTGTCTAGAACAGTGCTGTCTAATAGA

ACTTTCTGTGACGATGGATATTTTGTAGACTTTTGCTGTCCAGTGTGGTAGCCACTAACCACATGTGGCT

GTTAAGCCCTTGAAATATAGCTAGTGTGACTAGAAAGTATTTTATTTTAAATTTACATAGGCACAAGTGG

CTAGTGGCTACTGTATTGACATTCTGGGTCTAGGACTAGAACCAGTGGTCTGTAACAAAAGTACTTTCTC

TTTACTCTATTAAATCTAGAATTAGCCGGGCATGGTCGCTCATGCCTGTAATCCCAGCACTTTGGGAGGC

CAAGGCAGGCAGATCACTTGAGGTCAGGCGTTTGAGACCAGCCTGGTCAACATGGCGAAACCCTGTCTCT

ACAAAAAACATAAAAATTAGCCAGGTGTGGTGGTGGGCACCTGTAATCTCAGCTACTTGGGAGGCTGAGG

CACAAGAATCACTTGAACTTGGGAGGTGGAGGTTGCAGTGAGCCAAGATTGTGCCACTGCACTCAAGCCT

GGGTGACGAGTGAAACTGTCTCCAAAAAAAAAAAAAAATCTAGAATTCTTGGAAGTACATTATATTGCCC

TTCAGAATAGATTCCAGTTCCTGTTGTGCTCACCTTTATAATTTTACCATAAGTTTTACCTATTCGTAAG

TTGGCAGTTTTAGATAGATAACATTCTGGTGGTAGCTAGGGATTTACCTTTTGTCATCCTTTTCCTGCAC

TTCTCTTGAATTCCTTTATAGATGTACAGTTTTGCTTTAACCACTGAAGATTGCTGTAAATTATAAAGGG

TGTGATAGAATCCACATGGCTGTCAAGAAGGAGATCTTACCAAGGACAGTTGACTGACTAGTCTCAGATT

GTTTCATATCATTTATACTTGGGTAAGAGTAAACTAGATAACTGGGCGTCGTGGTGCACACCTGTAGTCC

CAGCCACTCTGAGGCAGGAGGACTGCTTGTGCCCAGAAGTTCGAGGCTGCAGTGTAGCTGTGATTGTGCT

ACTGCACTCCAGCCTGGGCAACAGATAAAGGAAACTCCATCTCTTTTAAAAAAAAAAAAAGTGGTCTGGG

TGCAGTAGGTCATGCCTGCAATCCCAGCACTTTGGGAGGCCAAGGCAGGCAGATCACCTGAGGTCAGGAG

CTTGAGACCAGCCTGGCCAACATGGTGAAACCCCATCTCTACCAAAAATATAAAAACTAGCTGGGCGTGG

TGGCGGCACCTGTAATTGCAGCTATTTGAGAGGCTGAGGCAGGAGAATCGCTTGAACCTGGGAGATGGGG

GTTGCAGTGAGCCAAGACCGCCCCATCGCACTCCAGCCTGGGCAACAATAGTGAAACTCCGTCTCAAAAA

GAAAAAAAGTTTCCTTAGAATGGAAAATATTCATTCATGAGCTCTTTTGGCAATCCGTCATCAGTATATT

CTGAAAACCAATAAGATGTTGCCAAGTTGGGGGCGAGAGCTATGTAATGCAAGGCATATGCCTGATGAAG

TATACAAATACACCTGACCAGAAACTTTGTCTCCCACATAAGTCTCTTCTAGGCACTGTCGGGGTACATA

CTGAGCTGCTGCTTTGGCTGTATTTTCTGTGCCTCAGAATAACCATTGTCCTGGTGTTCATATCCTTAGA

GTTCAGTACAAAATGTTGGATATCCATTTAATAGGTTCCAGGTTATCTTAGTTGGAGTTTGGGGTATTTG

AAAACGTCATGCCTTCAGGCTATCATTTCCCTCAGAAAGCTAAGTAAATTTACTGCATTCATTTCTCAAA

GAGTAAAAGTGCAGGTTGTATGTGTCTATGAACATTTAAACATGTTAAAATGTTAAATTTAACATTTTAA

ATTTAAACATTTAAATATGTCTGTAACTTGAACAGTGTAGTTTCAGAAAGGACCACTGGGCTAGTGTAAT

GCAGAAAATGCTGGGTCTAGGATTAGGAGAAAATTGTGTTTAGTGTGTATCAATAAACAGCCCGTGGACC

CAATCTGAAAAAAA

Cytochrome C – *Pan troglodytes*

>gi|115392118|ref|NM_001071821.1| Pan troglodytes cytochrome c, somatic (CYCS), mRNA

ATGGGTGATGTTGAGAAAGGCAAGAAGATTTTTATTATGAAGTGTTCCCAGTGCCATACCGTTGAAAAGG

GAGGCAAGCACAAGACTGGGCCAAATCTCCATGGTCTCTTCGGGCGGAAGACAGGTCAGGCCCCTGGATA

TTCTTACACAGCCGCCAATAAGAACAAAGGCATCATCTGGGGAGAGGATACACTGATGGAGTATTTGGAG

AATCCCAAGAAGTACATCCCTGGAACAAAAATGATATTTGTCGGCATTAAGAAGAAGGAAGAAAGGGCAG

ACTTAATAGCTTATCTCAAAAAAGCTACTAATGAGTAA

Cytochrome C-*Mus Musculus*

>gi|347943556|gb|JF919281.1| Mus musculus cytochrome c (CYCS) gene, complete cds

GTCTTCGAGTCCGAACGTTCGTGGTGTTGACCAGCCCGGAACGGTGAGCGCGCGACGGGGATGCGGACCG

GGAACGGGGCGAGGAGGCCGGGGCAGGCCTAACCTACAAAGCCATGCAGAACTGTGGCACGGCTGGAGGC

GGCGTGGGTGTAGGACCGGGGCCCGGAGTTCTAGAAGGAACTCGCCCGCCGCCCCTTCTGAGAGCAGAAG

CTGCTGTTCCTTCAGGACTGGGCCAGTTCTTTTCCTGATTAAGCATCTGGGAGGGTGGGTTTGTAAAGTC

GTAGTCACCCCTGGGATCTCAACGTGAGAGTGAGACGGTTCCTTCCCCAGTCCTCCCTCGCCTTCCCGTG

CCGGACTAAAGCAAGAGAGAAGAGACTTAACCTCCTAATGAGACCGGAGCAGGCGGAAGCACTTAGGATC

ACCCCCAGCCTCCCTTATCTTTGGAAGTGACTTTAATCGGCAGCACTGACAGCAGTTTAGCATTCTGAAA

TGTTAAGAGTTGTAGTGGTTTTGCTTTAATGAGTCCGGTTTCATATTTGTGACTTTTTGACCTTGCCTTC

TTCGGAAACTGGGGTTTGAGCTGTGGGTTTCCATTTCCCTTGGCATGGACCTGTCTTGAGCACAGACTTG

CTTTAAAGGAGAGTCTCCTAGGTGATGTATTTGCTATTGAAGAATATTGCAAGTGGACTGAGTTAGCACC

CTGCCTGGTGTTAAGAGGACATAAGAATAGCGGTTTGAATGGTCATTGGGATCCGTAGTACGTTTTACTG

TTGGATCTTTTCCCTTTTAGAATTAAAAATGGGTGATGTTGAAAAAGGCAAGAAGATTTTTGTTCAGAAG

TGTGCCCAGTGCCACACTGTGGAAAAGGGAGGCAAGCATAAGACTGGACCAAATCTCCACGGTCTGTTCG

GGCGGAAGACAGGCCAGGCTGCTGGATTCTCTTACACAGATGCCAACAAGAACAAAGGTAACGGGGGGAG

CTGCTGTCAGCCACAGCACAGGTTGCTTGGGTTAACCAGTGCAGAATTACCAGGTGTGTAAAACACTTAA

CCTCTGCATCTCTTTCTGTTTAGGCATCACCTGGGGAGAGGATACCCTGATGGAGTATTTGGAGAATCCC

AAAAAGTACATCCCTGGAACAAAAATGATCTTCGCTGGAATTAAGAAGAAGGGAGAAAGGGCAGACCTAA

TAGCTTATCTTAAAAAGGCTACTAATGAGTAATTCCACTGCCTTATTTATTACAAAACAAATGTCTCATG

GCTTTTAATGTACACCATAATTTAATTCACACACCAAATTCAGATCATGAATGGCTAGCAATGTTTTTGT

TGGACAGTCCTGATTTAAGTAAAACTGACTTGTCATAAAGTG

Cytochrome C-*Canis familiaris*

>gi|308081834|ref|NM_001197045.1| Canis lupus familiaris cytochrome c, somatic (CYCS), mRNA

GACGGATCGAGTTCGGTTGCACCGACACCGGTACTTAGGTGCGGAGCGGCGTGTCCTTGGACTTAGAGAG

CGGGACGTCCGGCTTGCGAGCGGGCTTCTTCGTTTCGCGCGCGCGACAGAAAGGCGATTGAAAAATGGGT

GATGTTGAGAAGGGCAAGAAGATTTTTGTTCAGAAGTGTGCGCAGTGCCATACTGTGGAAAAGGGAGGCA

AGCACAAGACTGGGCCAAACCTCCACGGTTTATTTGGGCGGAAGACCGGTCAGGCCCCTGGATTTTCTTA

CACGGATGCCAACAAGAACAAAGGCATCACCTGGGGAGAGGAGACCCTGATGGAGTATTTGGAGAATCCC

AAGAAGTACATCCCTGGAACAAAAATGATCTTCGCTGGCATTAAGAAGACAGGGGAAAGAGCAGACTTAA

TAGCTTATCTCAAAAAAGCTACTAAGGAGTAATAGTTGGCTATCGCCTTATTTATTACAAAACAAAAATG

TCTCATGACTTTTTTTTATGTGTACCATATTTAAATTGATCTCATTGACCAGAATTCAGATCATGAGTGG

CTGATAGAAGGTTTCTCTTGGACAGTCCTGATTTAAATAAGATTGGCTTGTGGTTAAATGAATATGATCA

GTTTTTTGAACTTTGATAGTAATTCTGGTTCAGCAAGTGCTCTCACTGTTTTCCCCTTCTAAAAAATATG

ATTGGAATTGATTAGAGATGTGTAGCTTTTCACAGAGATGGTAAATGCCACCTCTAAATTTATAGATTGG

TTTTATATTTAGATTTATATAACGGGTTATATTAATATATTTAAATACTGAGGAAATCCCTTCACTCTCA

GAACAGCAAGACTCAGCTGTGTTTCAAGTTGTGTTCCCTAGCCTGTTAAAGGCAATGGCTGAAGATAAGC

TGGCAGTGTCCACTTTATCTTTTTGGTCTTAACTATGCCAATTTAATTAAAATTCTTTGCATCTAAAATG

TTGCCTTTTGCTAATTGAAAGGCATTTTAGTGTGGTTTGTGCGTAATATTAAAATTTTAACACTTGGGTT

TTATACACCTGTAAGGTCAGACGCTTTTAAAATTTAGTGTGAGGAGATACAGCAGCAGACTTGGCTTGTG

AATTCTTGACTAAGACTACGTTATTATTCAGGATTGTCTTCTAAACAGCTATTCAGTGACACAAATGTGG

AATTAACCATGTATGTATGATTGGTAATGGTGCTTTTGCCAACTCCTTAGAAAGGTTAAAATAGAGGAGA

TCCGTCATCAGCACAAATTTGTACATTATGTGATAAGGTTTAAGATAATTAAAAACAAAGTCACAGATCA

TEACHER FEEDBACK FORM

Thank you for reviewing *Navigating Databases* curriculum. Please review the entire curriculum and then complete the questions below. Comments and suggestions are greatly appreciated!

Teacher name: _____

Date reviewed: _____ Email: _____

School: _____ Grade Level Taught: _____

Part I: Evaluation of the Entire Curriculum

SECTION A**:** For each item below, please indicate your response to each question as it relates to the curriculum **overall** by marking Strongly Agree (SA), Agree (A), Undecided (U), Disagree (D), or Strongly Disagree (SD).

	SA	A	U	D	SD
Are the experimental procedures appropriate for your students?					
Are the topics addressed important for your course objectives?					
Are the topics addressed relevant to your students' lives?					
Are the topic addressed interesting to your students?					
Is the depth of coverage of topics appropriate?					
Is the overall quality of the curriculum satisfactory?					
Is the content in the curriculum properly sequenced?					
Is the content in the curriculum adaptable for a range of student ability levels?					

Part II: Please provide additional comments pertaining to the curriculum.

_____
_____
_____
_____
_____
_____

Part II: Evaluation of individual lessons

Section A: Please make comments and suggestions regarding each lesson.

Lesson 1: Introduction to Bioinformatics	
Lesson 2: Navigating Databases Part A: Using USCS Genome Browser	
Lesson 2: Navigating Databases Part B: Using NCBI	
Lesson 2: Navigating Databases Part C: Using In-silico	
Lesson 2: Navigating Databases Part D: Using Emboss	
Lesson 2: Navigating Databases Part E: Using BLAST Part A	
Lesson 2: Navigating Databases Part F: Using BLAST Part B	
Lesson 2: Navigating Databases Part G: Molecular Visualization	

CONTENT AREA EXPERT EVALUATION

Thank you for reviewing Navigating Databases curriculum. Please review the entire curriculum and then complete the questions below. Comments and suggestions are greatly appreciated!

Reviewer Name: _____

Date Reviewed: _____ Department/Division: _____

Employer: _____ Email: _____

Job Title: _____

Part I: For each item below, please indicate your response to each question as it relates to the curriculum **overall** by marking Yes (Y), No (N), Undecided (U).

	Y	N	U
1. Is the science content in the curriculum accurate?			
2. Is the science content in the curriculum current?			
3. Is the science content in the curriculum related to major biological concepts?			
4. Is the science content in the curriculum important for science literacy?			
5. Is the science content coverage in the curriculum thorough and complete?			
6. Is the content in the curriculum properly sequenced for a novice?			
7. Is the overall quality of the curriculum satisfactory?			
8. Are there additional concepts that should be included? If yes, please elaborate below.			

Part II: Please include below any comments or suggestions about the curriculum.

1. General comments about the overall curriculum _____

(Part II continued): Please include any comments or suggestions about the curriculum

2. COMMENTS REGARDING INDIVIDUAL LESSONS

Lesson 1: Introduction to Bioinformatics	
Lesson 2: Navigating Databases Part A: Using USCS Genome Browser	
Lesson 2: Navigating Databases Part B: Using NCBI	
Lesson 2: Navigating Databases Part C: Using In-silico	
Lesson 2: Navigating Databases Part D: Using Emboss	
Lesson 2: Navigating Databases Part E: Using BLAST Part A	
Lesson 2: Navigating Databases Part F: Using BLAST Part B	
Lesson 2: Navigating Databases Part G: Molecular Visualization	

STUDENT FEEDBACK FORM

Student name: _____ Date: _____ Student grade level: _____
_____ Circle one: Male Female

School name: _____ Teacher's name: _____
Subject: _____

Section A: Evaluation of individual lessons- For each question below, please indicate your response for each **specific lesson** by marking High, Moderate, Low,