Connecting human disease phenotype to genetic mutation and protein function: A modular data mining short course with an independent project sequence for lecture or lab

Janet M. Murray Ph.D., Heather E. Driscoll, and Kara Pivarski

Correspondence concerning this article should be addressed to Janet M. Murray Ph.D., Biology Department, 120A Marsh Life Science, University of Vermont, Burlington, VT 05405

Contact: Janet.Murray@uvm.edu

Synopsis

"Introduction to Data Mining" is a four-session online bioinformatics short course that covers topics including: literature searches, sequence databases, sequence similarity searches using BLAST, multiple sequence alignment, phylogeny reconstruction, protein structure databases, and 3D viewers. Each session is designed to familiarize undergraduate students with online databases and tools for use in scientific study. The module utilizes several online resources, databases, and search engines including: NCBI (NCBI Resource Coordinators, 2016), RCSB-PDB (Berman et al., 2000), Molviz.org (Sayle & Milner-White, 1995), as well as the open source PyMOL software (Schrodinger, 2015). The short course in its entirety is intended for mid- to upper-level undergraduates in a molecular biology, genetics, or biochemistry course. However, the modular design of the online course can be utilized to meet the needs of independent instructors and options are provided to adapt the materials for less advanced students. Although there are many data mining tutorials available, the unique strength of this educational module is the assignment of an independent project that necessitates the use of the data mining tools independently by each student, enhancing student familiarity and competence with the databases and tools that are introduced in the online tutorial (sessions 5 and 6). The resources for these projects are described and can be used separately from the online portion. The short course and independent research projects demonstrate the direct connection between genetic change, protein function, and human (clinical) phenotype.

Introduction

Biological information is represented in many forms: sequence data, structural data, and functional data. Data mining is an established tool for understanding function in genomics and proteomics. The advent of computational biology also offers new opportunities for those in chemistry, physics, mathematics, and computer science. This short course is an introduction to the concepts and principal databases of bioinformatics and structural biology/chemistry. The modular course is designed as a four-session online tutorial with integrated exercises and independent project assignments. The 5th and 6th sessions are designed for wrapping up and presenting the independent research projects. The short course in its entirety and each independent research project demonstrate the direct connection between genetic change, protein function, and human (clinical) phenotype.

The overall goal of this bioinformatics short course and independent research projects is to expose undergraduate science students to online databases and data mining techniques and increase their computational biology skills. This course should enable students to access and analyze sequence and structure data, explore phylogenetic relationships, create and edit images of protein molecules, generate a hypothesis as to the functional defect of a mutant protein associated with a given human disease phenotype, and present their results in several formats.

Approach/Method (Instructor Guidelines)

The tutorial and independent projects described below are designed for five, 2 – 2.5 hour lab or lecture periods and are targeted to upper-level undergraduate students. The final session (6) is a poster presentation forum held for students to present their independent research projects, which can be scheduled outside of the class period.

The "Introduction to Data Mining" short course was designed with three main goals; 1. to enhance learning about biological databases and data mining tools through repetitive use; 2. to build the tutorial in a modular format that creates flexibility for instructors; 3. to integrate independent research projects with bioinformatics concepts into a formal presentation given by each student. The course can be freely accessed at http://www.vgnoutreach.com/ and was developed and is maintained by the Vermont Genetics Network (VGN). VGN is the IDeA Network of Biomedical Research Excellence (INBRE) program in Vermont, supported by the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103449.

Enhancing learning through repetition

The online modules introduce students to online databases and bioinformatics tools through side-by-side text and images (see Table 1 for useful links to the online tools). Each tutorial describes the content and gives step-by-step instructions on how to access, search for, and interpret information from these sites. This is the student's first hands-on interaction with specific databases and their tools. There are two online exercises assigned each week. Each student must utilize the information and tools described in the tutorials to answer specific questions in the exercises. Although the tutorials are available to them as a reference, the exercises are unique from the examples given in the tutorial sessions and must be completed and submitted online to demonstrate the student's familiarization with the tools. Students are also given a weekly assignment related to their independent research project. Each assignment helps them gather information that will be useful in understanding, organizing, and delivering their final research project presentation. These assignments guarantee that each student utilizes the online

databases at least three times to gather distinct information. Over the years we have noted that most students repeat information gathering for their research project when they are putting together their final product. This adds another independent interaction with most databases and tools.

Flexible and customizable online module

The bioinformatics short course has been ported to the Moodle™, a web-based content management system (https://moodle.org/), and is updated each summer with new information, links, and databases. Link and database changes are also checked for updates before each semester. A unique version of the course is built for each instructor when requested so they may utilize all the tools available in Moodle™ (grading, blog, etc.), have access to all instructor materials, as well as gain the ability to edit the "standard" module to best fit their needs. We have had instructors deliver the modules over six weeks in laboratory sessions as designed, others have expanded the length of the delivery to introduce concepts throughout a semester in shorter class periods and some have utilized a subset of the tutorials for their students. For example, one instructor wanted to introduce their early undergraduates to the NCBI PubMed (NCBI Resource Coordinators, 2016; Muin & Fontelo, 2006; Muin, Fontelo, & Ackerman, 2006) for literature searches. The instructor hid the rest of the modules and utilized the tutorials that taught Boolean logic associated with database searches and how to use NCBI PubMed effectively to retrieve scientific literature. We have also seen a section of the module used to introduce students conducting undergraduate research to NCBI BLAST (Altschul et al., 1990; Park, Sheetlin, Ma, Madden, & Spouge, 2012) tools.

The delivery of the module typically follows one of two formats. In the first, the instructor briefly reviews the main concepts for their students and the students spend the majority of their in-class time completing the tutorials and begin working on the exercises. They complete the online exercises within a couple days and the presentation assignment before the next scheduled session. In the following class the exercises are reviewed and new concepts are briefly introduced before students begin the tutorials. The second format separates each weekly module into two class sessions. Each weekly module introduces two main concepts/tools. Either one concept is introduced and its related exercise is completed in each session or all concepts are introduced and the tutorials are completed in the first class period and the related exercises are completed in the second. Both

scenarios for the second format work well with students that require more background information and more support during the tutorial/exercises. We have found that having students working side-by-side on the exercises helps them better learn the tools. The "Introduction to Data Mining" short course is also designed to be useful for learner-centered instruction, e.g. flipped- or inverted-classroom learning (King, 1993; Walvoord & Anderson, 2009). In this third format, the students complete the tutorials independently before class and then work on the exercises and their presentation assignments in the presence of other students and their instructor(s). This works very well as long as all students review the tutorials before class. Of note, the short course has also been independently assigned to interns over the summer to learn basic data mining tools with only minimal interaction necessary from an instructor or faculty mentor.

Integration of independent research projects

Although there are many data mining tutorials available, the unique strength of this educational module is the independent project assignment and the resources provided for each disease phenotype a student can choose for their research. The projects are designed to engage students in a research project that requires the independent use of the tools that they have learned throughout the module. In the final project poster presentation, students describe a specific disease, the gene and protein associated with the disease, the molecular defect, and hypothesize as to why the molecular change could lead to the clinical phenotype. The research project ties independent concepts and tools from the module together to help students develop a greater overall understanding of the direct connection between genetic change, protein function, and human (clinical) phenotype. The assembly of 40 disease phenotypes with single amino acid mutations and corresponding x-ray crystal structures could be used in many formats. (More information regarding this resource is available in the specific outlines and resources section - see Human Disease Phenotype List).

Introduction of Short Course and Independent Projects

Each week students learn bioinformatics concepts and data mining through tutorials in Moodle™. Students complete exercises related to the material presented in the tutorials and upload them to Moodle™. A presentation assignment relating to their independent project (disease phenotype) is due before the next class session. Readings and supplemental materials are provided in a section at the beginning of each week in Moodle™. The content in these sections is intended to help students review, or learn for the first time, concepts that will be helpful, and often critical, for successfully completing the tutorials and exercises in this module.

Moodle™ provides many resources for instructors including announcements and forums, chat resources, and grading tools. This open source software has been available for over 10 years and has many online resources.

Overview material and resources for instructors and students is presented at the beginning of the short course. This includes a course description, course outline, information on computer requirements, and links to all the websites and documents required each week. A biology review is provided for students and instructors describing background information with which students should be familiar as they move through the tutorials. An overview of the independent research project is provided with a list of suggested disease phenotypes and an example of a completed poster presentation. A separate downloadable document is available to instructors, which describes the OMIM® data, allelic variants, and structural information for each disease phenotype. More information regarding this resource is provided in Weeks 5 and 6 below.

Specific Outlines and Resources

Each weekly session begins with optional resources (supplemental material) that review the scientific background for the tutorial material that will be presented. Then, specific concepts/databases are introduced, followed by two important data mining tools with exercises provided to help students demonstrate competency of the material. An independent research project assignment intended to help students gather information necessary for their final presentation is provided. A guide for obtaining this information and an example disease phenotype is included for each weekly presentation assignment.

Table 1.	
Bioinformatics Database/Tool	**Website**
Introduction to Data Mining	http://www.vgnoutreach.com/
Moodle™	https://moodle.org/
NCBI (National Center for Biotechnology Information)	https://www.ncbi.nlm.nih.gov/
OMIM® (Online Mendelian Inheritance in Man)	https://www.omim.org/
NCBI/PubMed	https://www.ncbi.nlm.nih.gov/pubmed/
Medline	https://www.nlm.nih.gov/pubs/factsheets/medline.html
NCBI/Genbank	https://www.ncbi.nlm.nih.gov/genbank/
NCBI/RefSeq (Reference Sequence Database)	https://www.ncbi.nlm.nih.gov/refseq/
NCBI/BLAST (Basic Local Alignment Search Tool)	https://blast.ncbi.nlm.nih.gov/Blast.cgi
NCBI/CDD (Conserved Domain Database)	https://www.ncbi.nlm.nih.gov/cdd
MolViz (Molecular Visualization Resources)	http://www.umass.edu/microbio/chime/
JSmol information	http://wiki.jmol.org/index.php/JSmol
NCBI/MMDB (Molecular Model Database)	https://www.ncbi.nlm.nih.gov/structure/
RSCB-PDB (Research Collaboratory for Structural Bioinformatics - Protein Data Bank)	http://www.rcsb.org/
PyMOL	https://www.pymol.org/
NCBI/Ensemble/dbSNP (Short Genetic Variation)	https://www.ncbi.nlm.nih.gov/projects/SNP/
ExAc (Exome Aggregation Consortium)	http://exac.broadinstitute.org/
NCBI/ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/

Week 1

Students are acquainted with the concepts of bioinformatics and databases. NCBI databases are introduced with information on

Global Cross-database searches within NCBI and background on Boolean operators and phase searching. The main databases introduced in week 1 are Online Mendelian Inheritance in Man (OMIM®) (Amberger, Bocchini, Schiettecatte, Scott, & Hamosh, 2015; Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005) and NCBI PubMed (Muin & Fontelo, 2006; Muin et al., 2006). The OMIM® database provides clinical information related to human disease phenotypes as well as information associated with disease implicated genes, the mutant alleles, and their phenotypes. NCBI PubMed is a search tool for Medline (https://www.nlm.nih.gov/pubs/factsheets/medline.html) and allows students to access primary literature and reviews related to their inquiries. The logic behind effective database searches is stressed in the tutorials and related exercises.

Independent Research Project Assignment Week 1

After unique disease phenotypes are assigned to each student, the students begin gathering the first information required for their project. We ask students to provide a brief description of the disease along with the OMIM® identifiers for the disease and the causative gene. Students also utilize PubMed to start to identify primary literature, and most importantly at this stage, reviews related to the disease phenotype.

It is important in week one to discuss the types of alleles that will be of interest to the students' projects, namely single amino acid substitutions within the protein that lead to a disease phenotype. Single amino acid substitutions result in translation of full length proteins and these substitutions can be introduced into wild-type structures utilizing the 3D protein modeling tools that are introduced to the students. These mutations may be very interesting if they occur in conserved domains within the protein structure. It should also be noted that for some of these proteins only certain regions of the protein may have been crystallized for x-ray analysis of the structure. They will need to focus on the mutations associated with a crystallized portion of the structure for their final mutant structure analysis.

Week 2

During week 2 students are introduced to the NCBI nucleotide sequence databases Genbank (Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2005)and RefSeq (O'Leary et al., 2016). In these tutorials students learn how to dissect and interpret the information in Genbank records, aka GenBank flat files. The exercises are designed to further familiarize the students with this file type and help them demonstrate their ability to find and interpret information contained within them. **Note**: the example in the second tutorial contains a gene sequence in reverse direction.

Independent Research Project Assignment 2

In this assignment students identify nucleotide files, both genomic and mRNA, as well as the protein file associated with their disease phenotype. They are asked to provide specific information within those files, including intron and exon locations and sizes, the length of the coding region as well as the 5' and 3' untranslated regions, length of the protein and the protein sequence.

Week 3

Students begin to focus on protein sequences in week 3. They learn about NCBI BLAST (Basic Local Alignment Search Tool) (Park et al., 2012), conduct a protein BLAST (BLASTP) search and learn how to interpret the data that is incorporated into the BLAST "Hit List" page, including a search of the Conserved Domain Database (CDD) (Marchler-Bauer et al., 2017) that accompanies the "hit list". The second part of week 3 focuses on multiple sequence alignments and phylogenetic analysis using a distance-based method. In this section students conduct a protein multiple alignment using COBALT (Constraint-based Multiple Alignment Tool) (Papadopoluous et al. 2007) and construct phylogenetic trees a neighbor-joining phylogeny (Saitou and Nei, 1987) in the NCBI web interface. The first exercise requires students to conduct another protein BLAST search and interpret the data received. The second exercise requires a third protein BLAST search and selection of the top 10-15 protein hits for phylogenetic analysis using tools in NCBI. These assignments, similar to all assigned exercises, are submitted in Moodle™.

Independent Research Project Assignment 3

Students begin to explore the amino acid sequence and structure of the protein associated with their disease phenotype. Using NCBI BLAST and the phylogenetic tools introduced in the tutorial and exercises, they determine if there are conserved domains within their protein of interest and if these domains, and this protein, have been conserved through evolution.
Once this exercise is completed it is a good time to provide information regarding the x-ray structure(s) associated with their independent research project. Using this information they can determine if there are certain domains of their protein that have been crystallized for analysis and begin to limit the mutant alleles that they can use for their project. The structures and associated alleles are provided in the disease information resource provided to faculty (see Human Disease Phenotype List).

Week 4

Week 4 is focused primarily on protein structures and databases. However, the first tutorial introduces the JSmol 3D structure viewer (Hanson, Prilusky, Renjian, Nakane, & Sussman, 2013) focusing on DNA structure. This tutorial utilizes the Molecular Visualization Resources (MolVIZ) website (Marmorstein, Carey, Ptashne, & Harrison, 1992; Sayle & Milner-White, 1995). The first exercise focuses on this tool and the analysis of the DNA structure. The NCBI Molecular Modeling Database (MMDB) (Madej et al., 2014) is then introduced, followed by the Research Collaboratory for Structural Bioinformatics - Protein Database (RCSB - PDB) (Berman et al., 2000). Students learn how to search for protein structures, obtain information regarding a protein, and view the structure using JSmol in RSCB-PDB.

The next tutorial introduces PyMOL 3D protein viewer (Schrodinger, 2015). This is open source software that can be downloaded for academic use. Students utilize this software to examine the protein structure they initially viewed in JSmol. They are exposed to important tools provided in PyMOL to examine and alter the structure. An additional PyMOL tutorial providing information about additional tools is presented for those who are interested in learning about structure analysis. The tools necessary for the exercise assignment and each student's independent research project assignment are provided within the first PyMOL tutorial. In the exercise a second protein structure is manipulated in PyMOL and students must answer questions to demonstrate their understanding of protein structure. Students are guided through making a single amino acid mutation and must hypothesize what structural and functional implications this mutation may have. They will need to form a similar hypothesis either independently or based on primary literature for their independent research project.
If this week's material is divided into two sessions we suggest that the PyMOL section is taught separately. This software has many features and can take some time to learn. The tutorial presents the steps students will take to examine their protein structure and make a single amino acid mutation associated with their disease phenotype and allele of interest.

Independent Research Project Assignment 4

Students explore the protein structure(s) associated with their assigned research project. They provide wild-type and mutant structures showing the area of the protein that is changed by the mutation they focus on and a hypothesis as to the structural

and functional defect associated with the amino acid mutation. Some mutant x-ray structures are available for direct comparison, while most require the incorporation of the amino acid change into the wild-type structure. This information is available to the instructor in the disease resource. Students should access primary literature in order to help form or support their hypothesis. More time is given for students to complete this assignment with instructor support in week 5.

Week 5

In week 5 students are given the opportunity to work on their independent projects and discuss any questions they may have concerning their disease phenotype with their instructor(s). In this session, most students are trying to finalize the wild-type and mutant structures they will be presenting on their posters and may require guidance utilizing PyMOL.

Between weeks 5 and 6 students will need to print their poster materials. Many institutions have poster printers available and instructors should be sure to book sufficient time for printing of the posters. If a poster printer is not available Microsoft Publisher® allows for printing of the poster material in sections on standard printers. The sections then need to be aligned and taped together to complete the poster. We have brought together students in advance of their poster session to complete the assembly process. We utilized a large lab/classroom with several paper cutters, tape and other materials to help them complete this process when it is necessary.

Week 6

The final session is dedicated to the presentations of the students' independent research projects. Each poster should include a discussion on the following aspects of the phenotype:

- Clinical Features / Genotype / Phenotype
- Molecular Genetics / Gene Function
- Protein Function / Biochemistry / Allelic Variants
- Phylogenetic analysis of the protein and its evolution
- A 3-D visualization of the protein which explains its structure and function.
- An explanation of how and why the defective variant results in the aberrant phenotype including a description of the changes in primary, secondary, and tertiary structure of the protein which result in aberrant function
- A bibliography of all resources

Although students have accumulated these data throughout the course, we find that many of the students rebuild some of their figures for their poster. Many comment about how much easier it is to regenerate the data even after weeks of not utilizing the tools. Students also find that the incorporation of all the assignments into a single presentation allows them to connect the information. They gain a better understanding of the course tools and how the gathered data fit together to create a detailed analysis of a molecular defect associated with a clinical human phenotype.

The duration of the poster session events vary depending on the size of the class, the timing within the semester, and the selected location. In smaller classes students have taken turns presenting their poster to the entire class. Larger classes tend to divide the session into two presentation times so that half the students are presenting and students have a chance to view others' posters. We have also had students present their posters as part of a larger research event held at their institution. We have had instructors set up poster presentations in common areas and provide refreshments to attract other faculty and students. Many students invite their advisors, other faculty, and their friends to their poster presentation. This is a great way for students to practice presenting scientific information to different audiences.

Independent Research Project Information and Instructor Resources

The final project for this module is a poster presentation on a human genetic disease or condition from the provided list of suggested phenotypes. Students periodically would like to research a human disease phenotype that is not on the list. We try to accommodate the students if appropriate allelic variants and a suitable structure are available. These new disease phenotypes can then be added to the disease list.

Currently, the disease list contains 40 disease phenotypes including Porphyria: Congenital Erythropoietic, which is used to create the presentation assignment examples. The Human Disease Phenotype List is available as a separate document. Below is an example of the information for each disease phenotype provided for the instructor (Figure 1). The title of the disease is provided followed by the associated gene name and OMIM® IDs for the gene and the disease. Mutant alleles appropriate for student research projects are presented in the table. A wild-type PDB structure ID is provided with the amino acid sequence represented in the crystal structure listed. If there are mutant structures available, they are provided in this area or in the table (PDB structure). Any information regarding differences in amino acid numbering between the allele and the structure are also given. For Porphyria: Congenital Erythropoietic only the wild-type structure is available and there is no discrepancy between the amino acid numbering for the allele listed and the structure. The number of available mutations varies for each disease phenotype.

Figure 1. Sample Disease Phenotype Information for Instructors

Porphyria: Congenital Erythropoitic

Gene – UROPORPHYRINOGEN III SYNTHASE; UROS

OMIM - 606938 gene
OMIM - (263700) disease

Alleles –Mutant alleles in table below

Structures
WT structure in PDB (1JR2) aa 1-260
Create mutant structure using PyMol software.

Number ▲	Phenotype	Mutation	dbSNP	ExAC	ClinVar	PDB Structure
.0001	PORPHYRIA, CONGENITAL ERYTHROPOIETIC	UROS, CYS73ARG	[rs121908012]	[rs121908012]	[RCV000003948]	
.0002	PORPHYRIA, CONGENITAL ERYTHROPOIETIC	UROS, PRO53LEU	[rs121908013]	-	[RCV000003949]	
.0003	PORPHYRIA, CONGENITAL ERYTHROPOIETIC	UROS, ALA66VAL	[rs28941774]	-	[RCV000003950]	
.0004	PORPHYRIA, CONGENITAL ERYTHROPOIETIC	UROS, THR62ALA	[rs28941775]	[rs28941775]	[RCV000003951]	
.0005	PORPHYRIA, CONGENITAL ERYTHROPOIETIC	UROS, THR228MET	[rs121908014]	[rs121908014]	[RCV000003952]	
.0006	PORPHYRIA, CONGENITAL ERYTHROPOIETIC	UROS, LEU4PHE	[rs121908015]	[rs121908015]	[RCV000003953]	
.0009	PORPHYRIA, CONGENITAL ERYTHROPOIETIC	UROS, VAL82PHE	[rs121908016]	[rs121908016]	[RCV000003956]	
.0010	PORPHYRIA, CONGENITAL ERYTHROPOIETIC	UROS, GLY188ARG	[rs121908017]	-	[RCV000003957]	

Figure 1. Disease phenotype information is provided for each research project assignment. This example is for Porphyria: Congenital Erythropoietic and includes the OMIM® database IDs, an allele list, and structure IDs for the PBD database.

The table of alleles provided is generated from the allelic variants table present in the OMIM® gene/locus record for uroporphyrinogen III synthase (UROS) (https://www.omim.org/allelicVariant/606938). The OMIM® record presents each allele separately with clinical information and references in list view as well as in table view. The table view is the starting template for the information provided to instructors. Initially, all alleles that are not single amino acid mutations, including those that result in early termination of the protein, are removed from the table. (One exception is noted for Cystic Fibrosis; CF as there is a structure provided for the single aa 508 deletion, which is the primary cause of CF.) This is followed by the removal of the alleles that are not directly related to the disease phenotype described.

The first column of the table in figure 1 contains the allele number, for example .001 in the table above corresponds to 606938.001 (allele #1 associated with Uroporphyrinogen III synthase; UROS, OMIM® ID - 606938). This is followed by a phenotypic description in column 2 and the amino acid substitution associated with the mutation in column 3. Column 4 (dbSNP) provides a link to specific variant information in Ensembl (Zerbino, D.R. et al., 2017). "Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation." Column 5 provides a link to The Exome Aggregation Consortium (ExAC)(Karczewski et al., 2017; Lek et al., 2016), "… a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community." This database provides population genetics for the allele variant when available. Column 6 provides a link to the NCBI Clinical Variant database (ClinVar) (Landrum et al., 2016; Landrum et al., 2014). "ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. ClinVar thus facilitates access to and communication about the relationships asserted between human variation and observed health status, and the history of that interpretation." The first 6 columns are found in the table view of the allelic variant list for each gene. The final column has been added and lists the PDB structure ID number for the mutant structure or a particular wild-type structure if there are many to start with for a specific disease. Mutant structures are highlighted in green when provided. In the example in figure 1, only the wild-type structure is available, so no mutant structures are present in the PDB structure column.

Related genes and diseases within the instructor list

There are several genes that are associated with multiple disease phenotypes, as well as similar diseases that are caused by mutations in different genes. (see Table 2 and Table 3, respectively). Five genes are available within the resource list that cause multiple disease phenotypes (Table 2). SERPINA1 is especially interesting as Alpha-1 Antitrypsin Deficiency is caused by loss of function of the SEPINA1 protein while Antithrombin Pittsburgh is caused by a mutation that introduces a new function to the protein. Mutations in the other four genes in Table 2 can result in different clinical results. Although some of the clinical manifestation may be similar, the diseases are considered and treated as distinct diseases/syndromes. Some of these differences are due to amino acid mutations occurring in different parts of the protein affecting function in distinct ways. Comparison of distinct molecular mutation and phenotypic differences can result in more in-depth discussions regarding functional domains, phenotype penetrance, and structure/function relationships within a protein.

Table 2. Genes Associated with Multiple Disease Phenotypes	
Gene	**Diseases**
SERPIN PEPTIDASE INHIBITOR - SERPINA1	Alpha-1 Antitrypsin Deficiency Antithrombin Pittsburgh
FIBROBLAST GROWTH FACTOR RECEPTOR 2; FGFR2	Apert Syndrome Crouzon Syndrome Ladd Syndrome Pfeiffer Syndrome
LAMIN A/C; LMNA	Emery-Dreifuss Muscular Dystrophy Lipodystrophy, Familial Partial Type 2 Mandibuloacral Dysplasia with Type A Lipodystrophy
PRION PROTEIN; PRNP	Creutzfekdt-Jakob Disease Fatal Familial Insomnia Gerstmann-Straussler Disease
PHOSPHATASE AND TENSIN HOMOLOG; PTEN	Cowden Disease 1 Macrocephaly/autism Syndrome

Table 3 provides information for three disease phenotypes that can be caused by a single mutation in more than one gene. Comparison of these independent research projects can lead to discussions of molecular defects of distinct proteins within the same pathway, or comparison of proteins that have similar molecular functions resulting in similar disease phenotypes when altered.

Table 3. Disease Phenotypes Caused by Single Mutations in Related Genes

Disease	**Type**	**Gene**
Hereditary nonpolyposis Colorectal Cancer	Lynch Syndrome, Type 1	MutS, E. COLI, HOMOLOG OF, 2; MSH2
Hereditary nonpolyposis Colorectal Cancer	Type 2	MutL, E. COLI, HOMOLOG OF, 1; MLH1
Neurofibromatosis	Type 1	NEUROFIBROMIN 1; NF1
Neurofibromatosis	Type 2	MERLIN, SCHWANNOMIN; SCH
Porphyria	Acute Intermittent	PORPHOBILINOGEN DEAMINASE; PBGD
Porphyria	Congenital Erythropoitic	UROPORPHYRINOGEN III SYNTHASE; UROS
Porphyria	Cutanea Tarda and Hepatoerythropoietic	UROPORPHYRINOGEN DECARBOXYLASE; UROD

Student research project presentations should be utilized to point out the different types of defects that can result from a single amino acid mutation. This can include the formation of an unstable protein, a stable protein that interacts differently with associated molecules or other proteins, changes in catalytic domains from mutation within or near the catalytic site, or structural changes affecting the catalytic site caused by distant mutations. The major take home message is that proper protein folding and retention of necessary molecular interaction is what determines protein function.

Student Concerns

Many students pick a disease with which they have some familiarity. This can include familial ties, therefore, information regarding specific clinical outcomes could raise personal concerns with specific students. We have not found this to be an issue with several hundred students. The information to which the student would be exposed to is publically available and upper-level undergraduates with specific interest in a disease have already gathered basic knowledge of prognosis and clinical outcomes. We do advise instructors to be aware of these concerns and to offer information regarding resources to students that experience anxiety around their personal ties with their independent project.

Student Evaluation

Students can be evaluated in several ways. The tutorial provides exercises for each session that tests student ability to navigate the online tools and to interpret the recovered data. These exercises are submitted online for instructor evaluation. Each instructor can also adapt the questions or add other assessment material to the module site. This could be very useful if the module is broken down into smaller segments (see *Flexible and customizable online module,* within the Approach/Method section). The tutorial also assigns independent project assignments that can be assessed to determine student progress and understanding of the specific disease at the clinical, nucleic acid, and protein level.

The final independent project presentation can be completed in several ways, including written reports and/or oral presentations. We have found that presentation of the material in a poster format to be the most effective. The student must clearly and concisely present the information to their fellow students, instructors, and any invited attendees. This must be done in both a written/visual form as well as engaging with the attendees to describe their findings and answer any questions. We commonly open the poster session to faculty, staff, and students at the undergraduate college, allowing students to further cultivate their presentation skills.

Justification:

The 'Introduction to Data Mining' short course and independent research project demonstrate the direct connection between genetic change, protein function, and human (clinical) phenotype. The tutorials cover topics including, literature searches, sequence databases, BLAST, multiple sequence alignment and phylogeny construction, protein structure databases, and 3D protein viewers. Student will learn how to access the databases, gather and interpret the data presented, and distill the information into a presentable format.

The core competencies addressed include increasing a student's ability to: 1) locate, read, and comprehend primary literature research papers on genetics topics, 2) critique complex data sets and use bioinformatics tools to access and assess genetics data, 3) communicate research result effectively, including writing research papers and giving presentations, 4) effectively explain genetics concepts to different audiences, and 5) tap into the interdisciplinary nature of science.

Specific core concepts supported for student learning at a sophisticated level include: understanding the genomic structure of genes, differentiating between a gene and an allele, being able to explain how the genetic code leads to transcription and translation, describing how changes within triplet repeats can alter gene function and create a phenotype, distinguishing between loss-of-function and gain-of-function mutations and their potential phenotypic consequences, predicting the most likely effects of changes in protein activity due to changes in protein structure, interpreting bioinformatics data to compare homologous genes/proteins in different species and inferring relative degrees of evolutionary relatedness, using comparative data from multiple species to identify which regions of a protein are critical for function, and utilizing bioinformatics tools to analyze gene structure, genetic variants and resulting protein structural changes.

Undergraduate students are expected to strengthen their scientific writing and presentation skills by generating a scientific poster and participating in a poster session. All students are expected to answer specific questions within the online tutorial using correct scientific language as well as demonstrate their ability to communicate scientific concepts. The independent research projects on human (clinical) phenotypes can lead to a discussion of genetic testing including what should be tested, who

should be tested, and the potential social impacts of testing.

Acknowledgements

We would like to thank Dr. William Barnes, who joined the Vermont Genetics Network (Outreach Core) at the University of Vermont on sabbatical in 2006-2007 from Clarion University of Pennsylvania and inspired the initial development of our data mining module prototype. Research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103449. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIGMS or NIH.

References

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(®)), an online catalog of human genes and genetic disorders. *Nucleic Acids Research, 43*(Database issue), D789-D798. doi:10.1093/nar/gku1205

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2005). GenBank. *Nucleic Acids Research, 33*(Database Issue), D34-D38. doi:10.1093/nar/gki063

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235-242. doi:10.1093/nar/28.1.235

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research, 33*(Database Issue), D514-D517. doi:10.1093/nar/gki033

Hanson, R. M., Prilusky, J., Renjian, Z., Nakane, T., & Sussman, J. L. (2013). JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. *Israel Journal of Chemistry, 53*(3-4), 207-216. doi:10.1002/ijch.201300024

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., . . . MacArthur, D. G. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Research, 45*(Database issue), D840-D845. doi:10.1093/nar/gkw971

King, A. 1993. "From Sage on the Stage to Guide on the Side." College Teaching Vol. 41, No. 1 (Winter), pp. 30-35.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research, 42*(Database issue), D980-D985. doi:10.1093/nar/gkt1113

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., . . . Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research, 44*(Database issue), D862-D868. doi:10.1093/nar/gkv1222

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-291. doi:10.1038/nature19057 http://www.nature.com/nature/journal/v536/n7616/abs/nature19057.html - supplementary-information

Madej, T., Lanczycki, C. J., Zhang, D., Thiessen, P. A., Geer, R. C., Marchler-Bauer, A., & Bryant, S. H. (2014). MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Research, 42*(Database issue), D297-D303. doi:10.1093/nar/gkt1208

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., . . . Bryant, S. H. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research, 45*(Database issue), D200-D203. doi:10.1093/nar/gkw1129

Marmorstein, R., Carey, M., Ptashne, M., & Harrison, S. C. (1992). DNA recognition by GAL4: structure of a protein-DNA complex. *Nature, 356*(6368), 408-414.

Muin, M., & Fontelo, P. (2006). Technical development of PubMed Interact: an improved interface for MEDLINE/PubMed

searches. *BMC Medical Informatics and Decision Making, 6*, 36-36. doi:10.1186/1472-6947-6-36
Muin, M., Fontelo, P., & Ackerman, M. (2006). PubMed Interact: an Interactive Search Application for MEDLINE/PubMed. *AMIA Annual Symposium Proceedings, 2006*, 1039-1039.
NCBI Resource Coordinators (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 44*(Database issue), D7-D19. doi:10.1093/nar/gkv1290
O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research, 44*(Database issue), D733-D745. doi:10.1093/nar/gkv1189
Papadopoulous, J.S., Agarwala, R.; COBALT: constraint-based alignment tool for multiple protein sequences, *Bioinfomatics,* Volume 23, Issue 9, 1 May 2007, Pages 1073-1079, https://doi.org/10.1093/bioinformatics/btm076
Park, Y., Sheetlin, S., Ma, N., Madden, T. L., & Spouge, J. L. (2012). New finite-size correction for local alignment score distributions. *BMC Research Notes, 5*, 286-286. doi:10.1186/1756-0500-5-286
Saitou, N. Nei, M.; the neighbor-joining method: a new method for reconstructing phylogenetic trees., Molecular Biology and Evolution, Voluen 4, Issue 4, 1 July 1987, Pages 406-425, https://doi.org/10.1093/oxfordjournals.molbev.a040454
Sayle, R. A., & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends in Biochemical Sciences, 20*(9), 374-376. doi:http://dx.doi.org/10.1016/S0968-0004(00)89080-5
Schrodinger, LLC. (2015). *The PyMOL Molecular Graphics System, Version 1.8*.
Shameer, K., & Sowdhamini, R. (2007). IWS: Integrated web server for protein sequence and structure analysis. *Bioinformation, 2*(3), 86-90.
Walvoord B.E., Anderson V.J. 2009. Effective Grading: A Tool for Learning and Assessment in College. 2 Edition. San Francisco, CA: Jossey-Bass. p. 272.
Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bai, J., Billis, K., Cummins, C., Gall, A., Giron, C.G. and P. Flicek (2017) Ensemble 2018. Nucleic Acids Res., doi: 10.1093/nar/gkx1098