An Online Activity Introducing First Year Students to Genomics

Jonathan Visick
Steve Johnston

www.noctrl.edu/biology
Increased Need to Understand Genomics

- Relatively new area. Absolutely essential
- Central to following many recent ‘breakthrough’ papers not just molecular biology, but also evolution, ecology…
- Important for many general articles
- Often not taught until third or fourth year (if at all!)

Redefining Disease, Genes and All

May 6, 2008. By Andrew Pollack

Duchenne muscular dystrophy may not seem to have much in common with heart attacks. One is a rare inherited disease that primarily strikes boys. The other is a common cause of death in both men and women. To Atul J. Butte, they are surprisingly similar.

Dr. Butte, an assistant professor of medicine at Stanford, is among a growing band of researchers trying to redefine how diseases are classified – by looking not at their symptoms or physiological measurements, but at their genetic underpinnings. It turns out that a similar set of genes is active in boys with Duchenne and adults who have heart attacks.

The research is already starting to change nosology, as the field of disease classification is known. Seemingly dissimilar diseases are being lumped together. What were thought to be single diseases are being split into separate ailments. Just as they once mapped the human genome scientists are trying to map the "diseasome," the collection of all diseases and the genome, the collection of all diseases and the genes associated with them.

"We are now in a unique position in the history of medicine to define human disease precisely, uniquely and unequivocally," three scientists wrote of the new approach last year in the journal Molecular Systems Biology. Such research aims to do more than just satisfy some basic intellectual urge to organize and categorize. It also promises to improve treatments and public health.

Scientists are finding that two tumors that arise in the same part of the body and look the same on a pathologist's slide might be quite different in terms of what is occurring at the gene and protein level. Certain breast cancers are already being treated differently from others because of genetic markers like estrogen receptor and Her2, and also more complicated patterns of genetic activity.

"In the not too distant future, we will think about these diseases based on the molecular pathways that are aberrant, rather than the anatomical origin of the tumor," said Dr. Todd Golub, director of the cancer program at the Broad Institute in Cambridge, Mass.
Genomics can be highly intimidating!
Goals:

- Basic familiarity with
  - Sequence identification and retrieval at Genbank
  - BLAST
  - NCBI’s Gene Database
  - OMIM Database
  - Model Organisms
- Not expecting expertise or proficiency at this level
  - Improved literacy and decreased anxiety
- Skill of thinking about genomes
- Keep the students excited about biology
BIO102: Introduction to Cell Biology and Genetics

- Second introductory course
- Mostly first-year students
- Mostly Biology and Biochemistry majors
- Capped at 32 students
- Team-taught
- Presented near the end of the course
Online Activity

- Most existing tutorials are aimed at graduate students
- Still intimidating!

The challenge: design a tutorial that is approachable for first-year students but uses the actual databases

Welcome to the Cn3D 4.1 Tutorial!

Index

- Introduction
  - What does Cn3D do?
  - Downloading and Installing Cn3D
  - Document conventions
  - Literature references
- Retrieving individual structures (vMDM)
  - From an Entrez literature search
  - From an Entrez sequence neighbor
  - From a BLAST search
  - From a known PDB identifier
- Viewing individual structures in Cn3D
  - Basics of Cn3D controls
    - The structure window main menu
    - The style panel
  - Cn3D's sequence viewer
- Retrieving structure alignments (VAST)
- Viewing structure alignments in Cn3D
  - Cn3D's alignment viewer
  - Cn3D's alignment model
- Importing and conservation
  - Importing sequences and structures
  - Visualizing sequence conservation
- Annotating a structure
  - Backbone and termini labels
  - User-defined styles

To set up the exercises...

You will make a copy of the files needed for these exercises in your home directory. Open a terminal window, and, if you don't already have one, make a ~/tbes_work directory by typing at the User prompt:

```
mkdir ~/tbes_work
```

Make sure that you have a ~/tbes_work directory:

```
ls ~/tbes_work
```

Copy the needed directory, but instead of typing `tar zcvf`, type the location of the Summer School directory tree:

```
cp -r /mnt/odin/summer10b/materials/07-bioinformatics/filenbioinformatics/ ~/tbes_work/
```

For instance, if the materials are located at /mnt/odin, you will type:

```
cp -r /mnt/odin/summer10b/materials/07-bioinformatics/filenbioinformatics/ ~/tbes_work/
```

Check that you have the files in this directory by listing the contents:

```
ls -l ~/tbes_work/bioinformatics
```

```
ls -lR
```

In this tutorial, when we refer to ~/tbes_work/Bioinformatics/ and its subdirectories, we are referring to the copy which you have just made in your own home directory.
Online Activity

Overall Outline:

• Brief introduction to genomic analyses
• Brief introduction to model organisms
• Select a disease-associated protein in a model organism
• BLAST against the human proteome
• Gather basic gene information at NCBI’s Gene database
• Gather basic disease information at OMIM

Worksheet for evaluation
Bioinformatics

Fortunately, computers are very patient and don't mind such repetitive tasks as looking through billions and billions of nucleotides! **Bioinformatics** is the new science of using computers to find useful information in DNA and protein sequences. For example, you can instruct a computer to identify potential genes by looking for promoter, terminator, start and stop codon sequences. Once a gene is found, the genetic code can be used to investigate its protein product.

Most of the information about the human genome and other genomes has been made freely available through various databases on the World Wide Web. That means you have access to the same wealth of information that can be accessed by a big-time researcher!

The purpose of this tutorial is to familiarize you with the kinds of data that are out there and some tools you can use to "mine" useful information from them. We will investigate some known genes, but you could use the same tools to do real-world research and find out more about genes whose roles are not yet understood!

Model organisms: the genes we share...

Perhaps you’ve wondered why we talk about bacteria, peas, yeast, *Drosophila*, mice and other organisms in class and study them in lab. Aren’t humans what we’re really interested in?

Well, there are other organisms on earth...but even if humans are your main interest, we have a lot to learn from other living things. **Model organisms** are very often used to get information that would be difficult or impossible to obtain from direct human research. All living organisms have to solve the same basic problems, so what we learn about flies, worms, mice or bacteria can pay off later in human applications.

One good use of model organisms is in unravelling how human genetic diseases work.

There are 100s of human genetic diseases, each caused by inheriting particular alleles of particular genes. Of course, we don't have "disease genes" in our genome, so a gene that causes a genetic disease is really a normal gene that has a normal function in our cells. But, if a mutation occurs, the resulting allele could fail to make a functional protein or make a protein that functions differently. If that change has a noticeable negative effect on some function of the human body, we say we have a genetic disease.
Below are several proteins that—in their mutant forms—can cause genetic disease, and the model organism in which each one is found. Choose one that you would like to study for this exercise.

**Choose your gene...**
- Dystrophin protein - *Urosophila melanogaster*
- CFTR protein - *Caenorhabditis elegans*
- RAD14 protein - *Saccharomyces cerevisiae*
- RecQ protein - *Escherichia coli*
- Ferroportin 1 protein - *Danio rerio*
- Tyrosinase - *Mus musculus*
This is the entry for your protein in Genbank, one of the major databases of genomic information. Every gene in every sequenced organism has an entry like this.

At the top, you can see an accession number that identifies your gene, then information about the organism and some references.

If you scroll down, you'll see the sequence of the protein. We could also look up the nucleotide sequence of the DNA, but today we'll work with the protein sequence.

The amino-acid sequence may not look very familiar to you, since it uses a code in which one letter represents one amino acid, rather than the three-letter code you've used. Click here to open a new window with a genetic code table you can use for reference.

Continue...
<table>
<thead>
<tr>
<th>Region</th>
<th>191..216</th>
</tr>
</thead>
<tbody>
<tr>
<td>/gene=&quot;RAD14&quot;</td>
<td></td>
</tr>
<tr>
<td>/locus_tag=&quot;YMR201C&quot;</td>
<td></td>
</tr>
<tr>
<td>/region_name=&quot;Zinc finger region&quot;</td>
<td></td>
</tr>
<tr>
<td>/experiment=&quot;experimental evidence, no additional data recorded&quot;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site</th>
<th>207</th>
</tr>
</thead>
<tbody>
<tr>
<td>/gene=&quot;RAD14&quot;</td>
<td></td>
</tr>
<tr>
<td>/locus_tag=&quot;YMR201C&quot;</td>
<td></td>
</tr>
<tr>
<td>/site_type=&quot;mutagenized&quot;</td>
<td></td>
</tr>
<tr>
<td>/experiment=&quot;experimental evidence, no additional data recorded&quot;</td>
<td></td>
</tr>
<tr>
<td>/note=&quot;M→H: In RAD14-2; loss of recognition of cyclobutane pyrimidine dimers.&quot;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site</th>
<th>216</th>
</tr>
</thead>
<tbody>
<tr>
<td>/gene=&quot;RAD14&quot;</td>
<td></td>
</tr>
<tr>
<td>/locus_tag=&quot;YMR201C&quot;</td>
<td></td>
</tr>
<tr>
<td>/site_type=&quot;mutagenized&quot;</td>
<td></td>
</tr>
<tr>
<td>/experiment=&quot;experimental evidence, no additional data recorded&quot;</td>
<td></td>
</tr>
<tr>
<td>/note=&quot;Y→C: In RAD14-2; loss of recognition of cyclobutane pyrimidine dimers.&quot;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>222..273</th>
</tr>
</thead>
<tbody>
<tr>
<td>/gene=&quot;RAD14&quot;</td>
<td></td>
</tr>
<tr>
<td>/locus_tag=&quot;YMR201C&quot;</td>
<td></td>
</tr>
<tr>
<td>/region_name=&quot;XPA_C&quot;</td>
<td></td>
</tr>
<tr>
<td>/note=&quot;XPA protein C-terminus; pfam05181&quot;</td>
<td></td>
</tr>
<tr>
<td>/do_xref=&quot;PDB:65742&quot;</td>
<td></td>
</tr>
</tbody>
</table>

ORIGIN

1 mTpeqkakle ankiaierl rkrqiilssdq lariesrnrp lktrplavts gsnrddnas
61 avhvpnhgg psalantnta ttslygsuvv dgskrdasvl dkrptdrirp sirkqdyiel
121 dfatmoung gyninpckgL psdftddcsef esefgskckq clqdkwkegl erkmlyens
181 ppehiskopk cisciniem dpvhdyvflk qvckqoskeh pekyallkt ekseydyflit
241 pelnedilfr ilknpnphsg tfarmqlyfr cvveaafkk wqgeelde wprreekgk
291 rrekkyekki kemrktraq eynrlrekk hkgahihhs dpvdggedid gytqiqrro
361 dgleeteed i

We could spend a lot of time finding out what's known about your protein in the model organism, we really want to learn about human genetic disease, right?

So, let's find out if there is a similar gene in humans! Proteins that have similar amino-acid sequences or genes that have similar DNA sequences often have similar functions.

To find out if any protein encoded in the human genome is similar to your protein, we'll use a tool called BLAST.

Start by highlighting the amino-acid sequence of the protein and copying it to the clipboard.

Continue... Previous
BLAST (Basic Local Alignment Search Tool) is software that compares DNA or protein sequences. A BLAST search will compare your protein's amino-acid sequence to every other protein in the entire Genbank database!

When BLAST compares two proteins, it looks for the same amino acid in the same place, or similar amino acids in the same place. For example, if one protein has Val and the other has Ile, both are hydrophobic, so they'd be considered similar.

BLAST then calculates an overall similarity score for the two proteins and moves on to compare your protein to the next one in the database. Amazingly, it can compare your protein with every one of the other proteins in the entire database in a matter of minutes!
To do a BLAST search for proteins similar to yours, paste your amino-acid sequence into the large box labeled "Search".

Since today we only care about human proteins that are similar, look under Choose Search Set and in the Organism box, type "Homo sapiens" (or choose from the list that appears as you type). BLAST will then only compare your sequence with proteins from the human genome.

Then click the BLAST button.

Continue...
None of the text is readable, the page contains a screenshot of a scientific software interface with graphs, tables, and data.
Now we're getting somewhere! This page has a wealth of information about your gene.

Now is a good time to look at your worksheet and see what questions you might be able to answer by using the information here. Note that you have to look around and think about what you're seeing and reading to answer some of the questions.

To answer the questions about the length of the DNA and RNA for your gene, look at the graphical view of your gene not too far down the page. Open the Try our new Sequence Viewer link in a new tab (it's hard to get back from otherwise), and then set Theme to NCBI Fancy.

When you're ready to move on to learning about the disease, click Continue below.

Continue...

Previous
**XPA GENE; XPA**

Alternative titles; symbols

XPA COMPLEMENTING GENE; XPAC

Gene map locus 9q22.3

TEXT

DESCRIPTION

The XPA gene encodes a protein involved in DNA excision repair (Tanaka et al).

CLONING

Tanaka et al. (1989, 1990) cloned a mouse gene that restored UV light-resistant pigmentosum complementation group A (XPA; 273700). No correction was obtained. 

Tanaka et al. (1990) cloned a cDNA corresponding to the human XPA gene. The protein product has a molecular mass of 31 kD. The human protein is 95% similar to the mouse protein, consistent with a DNA binding protein. Two mRNAs, 1.3 and 1.4 kb, were identified by Northern analysis. These proteins encoded an N-terminal extension of the XPA protein that could be translated from an ATG signal that partially restores DNA repair defect in XPA cells. A 1.0-1.1-kb XPA cDNA conferred UV resistance to several cells derived from patients with XP but not with other XP complementation groups.
Major Questions on the Worksheet

- Number of amino acids in the model organism protein
- For the human gene:
  - Name and chromosomal location
  - Size (in base-pairs) of DNA and processed mRNA
  - Number of exons and introns
- Brief description of the human disease
- Normal function of the human protein
- Nature of the disease-causing allele(s)
  - Dominant or recessive? Autosomal or sex-linked?
- How does this allele contribute to the disease?
- One peer-reviewed reference
Assessment and Evaluation

• No rigorous, formal assessment has been done
• Many students report being fascinated
• Generates lots of student questions
• Easily expandable
• Effective review and integration of key concepts
• Frequent changes to databases
Available at:

http://depts.noctrl.edu/biology/courses/102/genome/genome.htm

(optimized for use with Mozilla. Doesn’t work well with Internet Explorer)

Comments and suggestions are always welcome!

Thank you to…

Jon Visick
BIO102 Students

www.noctrl.edu/biology