Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Vol. 33, 31–47, 2012

An Inquiry-based Bioinformatics Exercise incorporated Into a Newly Developed Molecular Biology Laboratory Course

Liane Chen[1] and Kathryn G. Zeiler[2]

University of British Columbia, [1]Department of Zoology, [2]Department of Botany, 6270 University Blvd. Vancouver B.C., CAN V6T 1Z4
(**lchen@zoology.ubc.ca**; **kathryn.zeiler@botany.ubc.ca**)

We have developed a research project that introduces students to bioinformatics tools and databases of the NCBI. Using these tools, students conduct research on genes of unknown function that have been identified through genomics research carried out by faculty members. Students analyze nucleotide and protein sequences, search for genes with similar sequences, and find conserved domains and structures, in order to develop hypotheses about the structure and function of an unknown gene product. Because poorly characterized genes are used, students must focus on the scientific process of analyzing data and synthesizing new ideas instead of searching previously published results.

Keywords: Bioinformatics, virtual lab, gene sequence, critical thinking, synthesis

Introduction

BIOL 341 (Introductory Molecular Biology Laboratory) is a third year laboratory course that was developed for the University of British Columbia (UBC). It is expected to have a large computer-based component, and must handle large enrolments. The new course material is currently being taught as part of BIOL 360 (Cellular Physiology Laboratory).

We have developed an assignment spanning several weeks that introduces students to bioinformatics tools in the context of scientific inquiry. Students conduct research on genes of unknown function, linked to genomics research carried out by UBC faculty members. The NCBI databases are used to analyze nucleotide and protein sequences, search for genes with similar sequences, and identify conserved domains and structures. These results allow students to perform more conventional literature searches to find indirect information about their gene, thus allowing them to develop hypotheses about the structure and function of their protein, and to propose experiments to test their hypotheses. Findings are written up as a research paper, but the assignment could also be adapted for oral and poster presentations.

Because poorly characterized genes are used, students cannot conduct literature searches on the genes themselves to find previously published results. The focus is not on obtaining a single perfect answer, but on the process of obtaining data, synthesizing new ideas, and finding the evidence to support their answers. In doing so, students also have the opportunity to add to the knowledge base. Additionally, this assignment provides the students with further practice in using scientific literature and should improve their technical writing skills.

Student Outline

The Bioinformatics Research Project

Learning objectives

Successful students will be able to:

- Search the NCBI databases for information that is relevant to their sequence or gene of interest.
- Conduct BLAST searches to find genes and proteins that are closely related to their own sequence of interest.
- Conduct sequence alignments to identify conserved regions within their sequence.
- Identify protein domains within their sequence of interest.
- Find a 3D protein or domain structure that could provide insight into the structure and function of their sequence of interest.
- Find scientific literature that could shed light on the function of their sequence of interest.
- Evaluate ideas for logic and reasoning.
- Identify ideas or strategies that could be applied to his/her own work.
- Offer constructive criticism on ideas, research strategies, writing, and presentation of data.
- Address the constructive criticism of peer reviewers and TAs, and use it to revise and improve their reports.
- Present their research findings in a journal article or poster presentation.

Part I. Introduction to the Bioinformatics project and the NCBI databases

We have several *C. elegans* and *S. cerevisiae* genes, and nobody knows what they do! It will be your term-long project to analyze a gene, propose a function, and propose a series of experiments to test your gene for this function. But first, you'll look at the NCBI databases, discover the types of information they contain, and explore whether they have any relevance to your gene.

Introduction

Researchers use large-scale genomic experiments to examine global changes in gene expression. Gene profiles for experimental and control conditions are compared, and differences in gene expression are examined further.

Thousands of genes and putative genes may be identified through this research. There is a further winnowing process to find the genes that are directly involved in the biological process of interest, and separate them from the genes that are indirectly involved. For example, in Meissner *et al.* (2009), genes identified through SAGE (serial analysis of gene expression) were further tested by RNAi (a method of blocking the expression of specific genes) to see which genes had a direct effect on muscle development.

Genome-sequencing projects and genome-wide experiments have led to the generation of a large body of. In addition to DNA sequences, there are data on protein structures, chromosome maps, genotypes and phenotypes, and global gene expression results. This data is often uploaded by researchers to public online databases for the use of other researchers. Thus, research may be conducted by using computers to analyze the data contained within these databases, long before any experiments are done at a lab bench.

For your Bioinformatics research project, you will be mining existing data for information about a little-known gene in order to develop a hypothesis about the structure and function of that gene. The genes that you will be studying may have some role in adaptation to high sugar stress (Erasmus *et al*., 2003), or in the development and organization of muscle (Meissner *et al*., 2009). This project will be completed in stages over the term, and will be done in pairs.

Complete the following exercises in your lab notebook.

Select a partner and a gene

1. Find a partner for your Bioinformatics project.
2. Each pair will be researching a different gene. Sign up for your gene on the list provided by the TA, and record the following information:
    - Your partner's name
    - Your gene sequence identifier
    - Your gene name (if present)

Create an NCBI account

Creating an account is necessary if you wish to save the results of your database searches for future use. Otherwise, you may find yourself repeating searches at a later date.

3. To create an account, go to **http://www.ncbi.nlm.nih.gov/guide/**. Click on "My NCBI" at the top right hand of the NCBI home page, and follow instructions to create a login and password. Since you will be working in pairs, you may wish to share a login and password. Record your login and password.

Introduction to NCBI databases

4. Go back to the NCBI website at **http://www.ncbi.nlm.nih.gov/guide/**
5. Type your selected gene into the search box and select "All databases."
6. Paste or copy this table into your notebook. Fill it in to indicate how many hits you detected for your gene in all the databases available.

Table 1. Number of pertinent records in the NCBI databases for the gene of interest.

# hits	Database		# hits	Database
	PubMed			Books
	PubMed Central			OMIM
	Site Search			OMIA
	Nucleotide			dbGaP
	EST			UniGene
	GSS			CDD
	Protein			3d Domains
	Genome			UniSTS
	Structure			PopSet
	Taxonomy			Geo Profiles
	SNP			GEO DataSets
	dbVar			Cancer Chromosomes
	Gene			PubChem BioAssay
	SRA			PubChem Compound
	BioSystems			PubChem Substance
	HomoloGene			Protein Clusters
	GENSAT			Peptidome
	Probe			
	Genome Project			

By this point, you have probably noticed that there are many databases available at the NCBI. Several will not contain any information about your sequence of interest, nor would they contain the type of data that will be useful at this stage of your research.

7. Select 4 of the databases that did not receive any hits in your search, and describe the kind of information that the database provides.

Research your gene

Note: You may copy relevant records or web links into a word file as part of your research notes. Instead of printing out these files, make note of these files and what they contain in your lab notebook, and be prepared to hand in these files if requested to do so.

These NCBI databases are more likely to provide useful information towards your project:

- Nucleotide
- Protein
- Structure
- CDD
- PubMed

8. Search each of these databases for your gene of interest. For each database that returns a hit, verify that the information is specific for your gene, and describe what the search has told you about your gene.

   You will probably have noticed that for each record that you retrieve, the NCBI will provide links to related information and tools on the right hand of the screen. Some will take you directly to databases already listed here. Others will take you to different groups within NCBI.

9. Select a link, and describe the kind of information it provides. Was the information directly relevant to your gene?

There are other scientific databases outside the NCBI. Many of the NCBI records are cross-referenced with the information contained within these databases, and we blinks may also be provided for easy access. For example:

- *Wormbase* (**http://www.wormbase.org/**) is a website with links to genetic databases and tools for *C. elegans*.
- *The Saccharomyces Genome Database* (**http://www.yeastgenome.org/**) is an online database for *S. cerevisiae* data.

10. Search one of these databases for information about your unknown gene. What kind of information does this database describe, and what does it say about your gene?

Any ideas so far?

11. Recall that you are trying to come up with ideas about what your unknown gene might do, and why its activity may be important for dealing with sugar stress or muscle development. Based on the information that you have found so far, what ideas do you have, and what might you want to explore further? Brainstorm as many ideas as possible-- you can decide later about which ideas are worth pursuing. Write these ideas down in your lab notebook.

Formal assignment: Create your first Bioinformatics figures

Note: When writing up your formal figures, report or poster, information from your research notes MUST be rewritten in your own words. Images and sequences may be copied and modified, as long as they are properly acknowledged and cited.

1. Create a figure based on the results of a nucleotide, protein, CDD, or 3D Domain search. This figure should highlight the key features of your gene. You should create one figure for one set of results, and your bioinformatics partner should create a second figure based on the results of a different database search.
2. Include a figure number and a descriptive caption that helps the reader interpret the figure.
3. Include a written description of what is shown by the data in the figure, plus the direct inferences of that data—this should be part of the Results section of your final report or poster.
4. Include any larger interpretations you may have about your gene or gene product, based on the data you have—this may become part of the Discussion section of your final report or poster.

Make a duplicate copy of the entire figure, caption, and written lines, for peer review. This question is due next week.
Note: You may end up making additional figures of your results from the other databases for your final report.

References

Meissner, B., Warner, A., Wong, K., Dube, N., Lorch, A., *et al*. (2009). An integrated strategy to study muscle development and myofilament structure in *Caenorhabditis elegans*. *PLoS Genet*. 5(6): e1000537. doi:10.1371/jornal.pgen.1000537

Erasmus, D. J., van der Merwe, G. K., & van Vuuren, H. J. J. (2003). Genome-wide expression analyses: Metabolic adaptation of *Saccharomyces cerevisiae* to high sugar stress. *FEMS Yeast Research* 3: 375-399.

Of note

Meissner *et al*. (2009) is a report on research conducted by Dr. Moerman's lab. Don Moerman is a UBC Professor in the Department of Zoology. For more information about his research, check out his department web page and his home page:

- **http://www.zoology.ubc.ca/person/moerman**
- **http://www.zoology.ubc.ca/~dgmweb/people_0don.htm**

Erasmus *et al*. (2003) is a research project conducted by Dr. van Vuuren's lab. Hennie J. J. van Vuuren is a UBC Professor and Eagles Chair in Food Biotechnology, the Director of the Wine Research Centre, and an associate member of the Michael Smith Laboratories (at UBC – Point Grey campus). For more information about his research, check out his web page:
http://www.landfood.ubc.ca/wine/vanvuuren/vanvuuren.html

Part II. BLAST, sequence alignments, and search strategies

There is very little research on my gene.... what next?

Having difficulty in finding specific information about your gene? That's not a surprise, as little is currently known about them. However, you can make educated guesses about what your gene does by analyzing it for similarities to genes and domains that happen to be well known.

This is a summary of useful questions to ask yourself when researching a little-known sequence. We will go through these in more detail below.

1. Sequence similarities
    - Is this unknown gene similar in nucleotide and amino acid sequence to known genes? What are the functions of these known genes, and how closely do the sequences match up to your gene?

2. Sequence alignments
    - When you match up the sequences of related genes to your selected gene, are there areas that match up more closely? (i.e., regions where sequences are more conserved)?

3. Protein domains
    - Can any protein domains be identified?
    - What are the functions of these domains?
    - What known proteins contain these domains, and how do they use these domains?
    - If no domain has been identified, can you still find regions within your gene that are more highly conserved? Conserved regions are more likely to be important to gene function than highly variable regions.

4. Protein structure
    - Can you find a 3D structure of conserved domains/similar proteins?
    - How does the 3D structure relate to the function of the protein?
    - Which part of your unknown and homologous proteins should be similar?

5. Literature searches
    - Have you identified key words in your analyses that could lead you to research that is related to your gene?
        - Are there closely related genes?
        - Can you find research on the functional domains and sequence motifs within your gene?

BLAST

http://blast.ncbi.nlm.nih.gov/Blast.cgi

The Basic Local Alignment Search Tool (BLAST) is a web-based program that allows researchers to compare nucleotide and amino acid sequences, line up regions that match, and calculate the statistical significance of these matches. Closely related genes are more likely to have matches in sequence than distantly related genes. Thus, BLAST results can suggest evolutionary relationships between genes. Because nucleotide and amino acid sequences that are critical to protein function are more likely to be evolutionarily conserved than non-critical sequences, BLAST results can also be used to identify greater detail about the putative gene product. Thus, BLAST results may indicate the location of protein domains, reveal key amino acids that are critical for the structure and function of these protein domains, suggest a structure and function for an unknown gene, and indicate whether an unknown gene could be considered a member of a gene family with similar structures and functions.

There are different types of BLAST searches that can be applied:

- **Nucleotide blast**: This will take a nucleotide sequence and search it against other nucleotide sequences. This search is useful when you want to identify genes that are similar to your target nucleotide sequence.
- **Protein blast:** This will take an amino acid sequence and search it against other amino acid sequences. This search is useful when you want to identify proteins that are similar to your target protein sequence.

A nucleotide or protein BLAST search will probably be enough for you to find genes and proteins that are related to your poorly characterized gene. The other types of BLAST searches would be broader and less stringent—good for finding more distantly related genes and proteins.

blastx: This will take an amino acid sequence and search it against translated nucleotide sequences. Translations are done in all three reading frames. This search is broader than Nucleotide or Protein Blast, because it finds all possible nucleotide sequences that code for proteins that are similar to your target protein.

tblastn: This will translate a nucleotide sequence in all three reading frames and search it against amino acid sequences. This search is broader than Nucleotide or Protein Blast, because it finds proteins that are similar to any protein that could be coded by your target nucleotide sequence.

tblastx: This will translate a nucleotide sequence in all three reading frames, and search it against translated nucleotide sequences. This is a very broad search, because it translates your target nucleotide sequence in all three frames, and finds other nucleotide sequences that could be translated in any frame to produce similar proteins.

In order to carry out a sequence search or an alignment, sequences must be entered in a specific format:

Fasta format

Use ">" plus any title or labeling information in the first line. Add sequence in the second line. There should be no spaces or lines in the middle of your sequence. Sequences should be in single-letter code for nucleic acids or amino acids.

```
>YAR075W
MVVFKNIGHIITKALALGSSTVMMGGMLAGTTESPGEYLYQDGKRLKAYRGMGSIDAMQKTGTKGNASTSRYFSESDSVL-
VAQGVSGAVVDKGSIKKFIPYLYNGLQHSCQDIGCRSLTLLKENVQSGKVRFEFRTASAQLEGGVNNLHSYEKRLHN
```

Bare sequence

Sequences may be entered on their own:

```
MVVFKNIGHIITKALALGSSTVMMGGMLAGTTESPGEYLYQDGKRLKAYRGMGSIDAMQKTGTKGNASTSRYFSESDSVL-
VAQGVSGAVVDKGSIKKFIPYLYNGLQHSCQDIGCRSLTLLKENVQSGKVRFEFRTASAQLEGGVNNLHSYEKRLHN
```

Sequences may also be entered with numbers and spaces, e.g., cut and pasted from a Genbank report:

```
      1  atggtggtgt tcaaaaacat tggtcatatt attaccaaag ctttggctct tggttcttct
     61  actgttatga tgggtggtat gttggccggt actaccgaat caccaggtga atatctctat
    121  caagatggta aaagattgaa ggcgtatcgt ggtatgggct ccattgacgc catgcaaaag
    181  actggtacca aaggtaatgc atctacctcc cgttactttt ccgaatcaga cagtgttttg
    241  gtcgcacaag gtgtctctgg cgctgtcgtt gacaaaggat ccattaagaa atttattccg
    301  tacttgtaca atggattaca acattcttgt caagacatcg gctgtaggtc gttaacttta
    361  ctaaaggaaa atgtccaaag cggtaaagtt agatttgaat tcagaaccgc ttctgctcaa
    421  ctagaaggtg gtgttaataa cttacattcc tacgaaaaac gtttacataa ctga
```

Gene identifiers

These can be accession numbers, gi numbers, or NCBI sequence identifiers. These identifiers must be entered in the required format, or they will not be recognized; e.g., from a Nucleotide search on YAR075W:

```
LOCUS       AY692623                 474 bp    DNA     linear   PLN 24-APR-2007
DEFINITION  Saccharomyces cerevisiae clone FLH114659.01X YAR075W gene, complete
            cds.
ACCESSION   AY692623
VERSION     AY692623.1  GI:51012696
```

For more details on sequence format and on the single letter code used for nucleotide and amino acid sequences, refer to **http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml**

Complete the following exercises in your lab notebook, and save supplementary information to a word document.
Note: You may copy relevant records or web links into a word file as part of your research notes. Instead of printing out these files, make note of these files and what they contain in your lab notebook, and be prepared to hand in these files if requested to do so.

BLAST exercises

1. Find the Nucleotide or Protein record for your gene. Cut and paste the species name, the sequence identifier, the name of the sequence, and the sequence itself, into a document for easy access.
2. Start a new BLAST search. Select the Nucleotide BLAST if you are going to use the nucleotide sequence, and the Protein BLAST if you are going to use the protein sequence.
3. Enter the sequence identifier (or the sequence in appropriate format) into the "Enter Query Sequence" box.

   You may change the settings that modify the stringency of your search (i.e., how closely the search should match up to your sequence) and which select the sequence databases to be searched (e.g. which organism to search), but the default setting should also produce results.
4. Click on the "BLAST" button at the bottom of the screen to start your search. Look at the sequences that match up with your gene of interest, and click on the links to their records.
5. Which sequences have been identified as genes or gene products? Is anything known about those genes
6. Cut and paste the sequence information (name, sequence identifier and sequence) into a document for your sequence alignment activity.

   Note: Pair-wise alignments between your gene and the related sequences are all listed near the bottom of the page. However, it is sometimes simpler to look only at alignments between specific sequences of interest.

Sequence Alignments

Alignments with BLAST

Once you have used your sequence in a BLAST search, select 2-3 of your related sequences and align them with your original sequence. If possible, choose sequences that are from different species. Alignments may be done with nucleotide or amino acid sequences.

7. Start a new BLAST search. Select the Nucleotide BLAST if you are going to use the nucleotide sequence, and the Protein BLAST if you are going to use the protein sequence.
8. Enter the sequence identifier for your gene (or the sequence in appropriate format) into the "Enter Query Sequence" box.
9. Place a check in the box labeled "Align two or more sequences."
10. Enter the sequence identifier for your related sequences into the "Enter Subject Sequence" box. Keep sequence identifiers separated by a blank line.

You may change the settings that modify the stringency of your search (i.e., how closely the search should match up to your sequence), but the default setting should also produce results.

11. Click on the "BLAST" button at the bottom of the screen to start your alignment.

Interpreting sequence alignments

The alignment below is between nucleotide sequences. Gaps in the sequence alignment appear as dashes, and horizontal lines indicate the nucleotides that match between sequences.

```
>ref|NM_001179347.1|   Saccharomyces cerevisiae S288c Imd2p (IMD2), mRNA
Length=1572

GENE ID: 856626 IMD2 | Imd2p [Saccharomyces cerevisiae S288c]
(Over 10 PubMed links)

 Score =  750 bits (406),  Expect = 0.0
 Identities = 453/475 (95%), Gaps = 6/475 (1%)
 Strand=Plus/Plus

Query  1     ATGGTGGTGTTCAAAAACATTGGTCATATTATTACCAAAGCTTTGGCTCTTGGTTCTTCT  60
             ||||||||||||  ||||||||||||  | ||||||||||||||||||||||||||||||
Sbjct  1103  ATGGTGGTGTTC-AAAACATTGGTC--A-TATTACCAAAGCTTTGGCTCTTGGTTCTTCT  1158

Query  61    ACTGTTATGATGGGTGGTATGTTGGCCGGTACTACCGAATCACCAGGTGAATATCTCTAT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  1159  ACTGTTATGATGGGTGGTATGTTGGCCGGTACTACCGAATCACCAGGTGAATATTTCTAT  1218

Query  121   CAAGATGGTAAAAGATTGAAGGCGTATCGTGGTATGGGCTCCATTGACGCCATGCAAAAG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1219  CAAGATGGTAAAAGATTGAAGGCGTATCGTGGTATGGGCTCCATTGACGCCATGCAAAAG  1278
```

The pair-wise alignment below is between amino acid sequences. The middle sequence indicates the matches between query and subject sequences. Spaces in the middle sequence indicate mismatches and "+" indicates amino acids that may be different but which are similar in type. Dashes in the query or subject sequence indicate that there is a gap in the sequence, inserted to improve alignment.

```
>lcl|48742 test 1
Length=522

 Score =  269 bits (687),  Expect = 6e-77, Method: Compositional matrix adjust.
 Identities = 128/152 (84%), Positives = 136/152 (89%), Gaps = 1/152 (0%)

Query  6    NIGHIITKALALGSSTVMMGGMLAGTTESPGEYLYQDGKRLKAYRGMGSIDAMQKTGTKG  65
            NIGHI+ KALALG+S VMMGGMLAGTTESPGEY +QDGKRLK YRGMGS+DAMQKT  KG
Sbjct  372  NIGHIV-KALALGASCVMMGGMLAGTTESPGEYFFQDGKRLKTYRGMGSVDAMQKTDKKG  430
```

12. How closely does your sequence align with the related sequences?
13. Are there regions within your gene that match up more closely than others? This could indicate a conserved region that is important for the function of your gene.

    A sequence alignment can be used to infer evolutionary relationships between genes.

14. When your alignment is done, click on "Distance Tree of Results" to see how closely your sequences may be related. Copy and save this image—you may want to use this in your final report.

Multiple sequence alignments with COBALT

BLAST can be used to align both nucleotide and protein sequences, but it will only align paired sequences. The Constraint-Based multiple Alignment Tool (COBALT) is a tool that aligns only protein sequences, but it will align several protein sequences together. This multiple alignment can be more useful for identifying domains and regions that are highly conserved between proteins.

If you have protein sequences that have already been aligned by BLAST, you can click on "Multiple Alignment" to align all sequences together. This link will automatically run your sequences through the Constraint-Based multiple Alignment Tool (COBALT). Alternatively, you can enter your protein sequences directly into COBALT.

http://www.ncbi.nlm.nih.gov/tools/cobalt/

```
48740  1    ------------------------------------------------------------MVVFKNIGHIITKALALGSSTV  22
48742  309  TREQAASLIQAGCDGLRIGMGSGSICITQEVMACGRPQGTAVYNVTKFANQFGVPCMADGGIGNIGHI-VKALALGASCV  387
48743  320  TREQAAQLIAAGADGLRIGMGSGSICITQEVMAVGRPQGTAVYAVAEFASRFGIPCIADGGIGNIGHI-AKALALGASAV  398
```

Unlike BLAST, COBALT does not highlight mismatched sequences. Like BLAST, it inserts gaps (indicated by "-") in order to optimize the alignment between sequences.

15. Are there regions within your multiple alignment that match up more closely between sequences than other regions? This could indicate a conserved region that is important for the function of your gene.

Protein domains

A protein domain is a region of polypeptide that folds into a specific structure with a specific function. The structure and function of a domain is determined by the amino acid sequence of that domain, which is in turn encoded in the nucleotide sequence of the gene. Thus, nucleotide and amino acid sequences can be analyzed for domains.

You may have already seen protein domains listed within the Nucleotide or Protein records of your sequence. The following protein domain search tools may also be linked to these sequence records:

Conserved Domain Search Service: A tool that analyzes the amino acid sequence of your gene product for protein motifs and domains. It may also identify protein families that your gene may belong to. Results are linked to the Conserved Domain Database.

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

CDART (Conserved Domain Architecture Retrieval Tool): Analyzes protein sequences for protein domains. It presents a schematic diagram illustrating the relative size and location of the domains on your protein sequence, and it lists other proteins that contain similar domains. Its results are linked to the Conserved Domains tool.

http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps

16. Select either the Conserved Domain Search Service or the CDART tool. Enter the amino acid sequence of your gene product in FASTA format, within the query box of this tool.
17. Make note of the protein domains that are found in your gene.

Protein Structure

The crystal structures of several proteins and protein domains are available in a number of the NCBI databases. While these databases may not contain the protein encoded by your gene of interest, it may have images of the domains encoded by your gene, or images of homologous proteins.

Structure: A molecular modeling database that contains 3D structures of proteins that have been derived experimentally through X-ray crystallography and Nuclear Magnetic Resonance (NMR).

http://www.ncbi.nlm.nih.gov/sites/entrez?db=structure

Conserved Domain Database (CDD): Contains conserved domains that have been identified through multiple sequence alignments. It will retrieve 3D structures as well as domain and superfamily information.

http://www.ncbi.nlm.nih.gov/cdd

18. Using Structure or CDD, search for a relevant protein domain, or for a protein that is similar in sequence to your gene product.
19. Click on individual search results. Are they relevant to your gene of interest, to the homologous proteins found in your earlier searches, or to the domains contained within your gene?

Both databases allow the retrieval of 3D images that may be viewed and manipulated with Cn3D, a free program which may be downloaded from the NCBI Resource pages. This program may be used to rotate images, and to highlight structures of interest by highlighting the relevant sequences.

http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml

20. Click on the thumbnail image to open the structure within the Cn3D viewer. Rotate the image to see the structure from all angles.
21. Save one representative image for your report. Describe the relationship between this structure and your gene of interest—how does the structure of this protein or protein domain relate to the function of your gene?

Literature searches

You may not have found journal articles that are directly related to your gene of interest-- not surprising, since little is known about your gene. However, there may be a lot of published research on other genes and gene products that resemble the gene that you are studying, on the protein domains that you have found within your gene product, or on the protein superfamilies that your gene product may belong to.

22. Do a literature search on some of the closely related genes that were identified through your BLAST search. Select three of the most useful references on related genes and list them here. Describe the content of each reference with a sentence.
23. Do a literature search on some of the protein domains or protein families that were identified in your analyses. Select three of the most useful references on the conserved domains and list them here. Describe the content of each reference with a sentence.

    (Note: You may want to make note of more references. The more research you conduct, the more likely you are to develop a strong hypothesis and experimental plan for your final report.)

Any ideas?

24. With your partner, brainstorm ideas about the function of your gene, based on your research to date.
25. Look at last week's ideas, and see which ideas should be kept and which ideas have been ruled out by your new research.
26. Identify the strongest one or two ideas, and support them with the relevant results (e.g., the presence of a domain may suggest a particular activity). If they are concrete enough, you may start writing down potential hypotheses.

Formal Assignment: Create your next Bioinformatics figure and brainstorm ideas about your gene.

Note: When writing up your formal figures, report or poster, information from your research notes MUST be rewritten in your own words. Images and sequences may be copied and modified, as long as they are properly acknowledged and cited.

1. Create a figure with appropriate labels, title, and figure caption, for any two search results from your BLAST search, sequence alignment, and protein or domain structure searches (i.e., you and your partner will each create one figure).
2. Include a written description of what is shown by the data in the figure, plus the direct inferences of that data—this should be part of the Results section of your final report or poster.
3. Include any larger interpretations you may have about your gene or gene product, based on the data you have—this may become part of the Discussion section of your final report or poster.

Make a duplicate copy—this will be submitted next week for peer review, and TA marking.

Note: You may wish to generate additional figures and paragraphs for the remaining search results, as they will also add to your final report.

References

National Center for Biotechnology Information. (n. d.). Basic Local Alignment Search Tool. Retrieved from **http://blast.ncbi.nlm.nih.gov/Blast.cgi**

National Center for Biotechnology Information. (n. d.). CDART: Conserved domain architecture retrieval tool. Retrieved from **http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps**

National Center for Biotechnology Information. (n. d.). Conserved Domains. Retrieved from **http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi**

National Center for Biotechnology Information. (n. d.). Constraint-based multiple alignment tool. Retrieved from **http://www.ncbi.nlm.nih.gov/tools/cobalt/**

National Center for Biotechnology Information. (n. d.). Structure. Download Cn3D 4.1 for PC, Mac and Unix. Retrieved from **http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml**

National Center for Biotechnology Information. (n. d.). Welcome to NCBI. Retrieved from **http://www.ncbi.nlm.nih.gov/guide/**

Materials

- Computers: One internet-capable computer per student or student pair, plus one computer for instructor demonstrations. The computers and internet connection should be robust enough to download and run the Cn3D software from the NCBI database, and run multiple searches and programs simultaneously.
- One LCD projector and screen with computer hookup, for instructor demonstrations.
- Microsoft Office (Word; PowerPoint) or Microsoft-compatible programs (e.g. Open Office; Google Docs) to copy data and take notes throughout the Bioinformatics exercises, and to create figures and write the final report.

Notes for the Instructor

The Bioinformatics lab does not require much set-up in the lab itself, but will require preparation. Preparation falls into 2 categories: A pre-screen of the genes to assign to students, and teaching resources to help students complete the assignment.

Selection of poorly characterized genes

The preparation begins with selection of the genes to assign to students. Source papers for poorly characterized genes with suspected biological function may be found in published genome-wide experiments—typically microarray experiments that examine global changes in gene expression. These genes may be found in the paper itself, or may be located in supplemental information. It is critical to pre-screen genes before assigning them to students, in order to ensure that there are limited or no direct publications about the gene, and that there are enough BLAST and domain search results for students to have a reasonably good chance at proposing and researching a gene function. For example, a quick search of the putative, uncharacterized gene product of YDR133C, a yeast sequence, reveals no conserved domains, and poor homology to identifiable proteins. In contrast, YGR001C (another yeast sequence) is found to contain a domain that identifies it as an N6-adenine methyltransferase, and a BLAST search reveals close homology to other N6-adenine methyltransferases.

A cursory approach to screening genes involves the following steps:

1. Conduct a Pubmed search to check for the number of publications that are specific for the gene.
2. Run a Protein search and select a record that is specific for your gene. Check this record for interesting annotations about domains and putative functions. (The Nucleotide search can be skipped as it will contain similar annotations.)
3. While in the Protein record, follow the link to run a BLAST search.
4. While in the BLAST results, check the list of homologous sequences for identified or putative proteins, and see if these proteins can be found on PubMed without too much difficulty.
5. While in the BLAST results, click on the domain search to identify motifs, domains, and protein superfamilies. Check if these can also be found on PubMed.
6. Once in Domains, click on a record to see if there is a 3D structure available.

Screening also provides a snapshot of the results that students would be expected to get in their searches. An example of selected sequences and their search results are shown in Table 1.

The instructor will need to decide whether it is important for students to gain useful results from all featured Bioinformatics tools, or if the goal is to gain enough clues to propose a function. The instructor will also need to decide on the scope of the assignment—Will the students be required only to glean information from the featured NCBI tools, or will they need to conduct further literature searches to support arguments about the proposed functions of their genes? If further research is required, it will be important to evaluate how difficult it might be for a novice to piece together information from the literature. For example, the yeast sequence YAL065C does not contain any recognized domains, but a BLAST search reveals close homology to a number of flocculin proteins—enough of a clue for a novice to research. In contrast, the yeast sequence YGL101W is identified as a metal-dependent phosphohydrolase with homology to other metal-dependent phosphohydrolases, but a PubMed search reveals no review articles for this family of proteins, and domain-specific information may be buried within papers about seemingly unrelated enzymes.

Student support

This assignment is meant to introduce students to some of the more commonly used tools within the NCBI. The NCBI databases are vast and the tools are all interconnected, and this can be overwhelming to novices, so it is critical that the instructor and TAs be comfortable with the procedures outlined in the assignment. Care will also be needed to keep students centered on the featured tools. For each tool, students will need to know what details may be ignored for now, and what details should be examined more closely. For example, a BLAST search may be left in its default settings, but the Maximum Identity and E value will tell the student how closely an aligned sequence matches with their unknown gene, and the probability that this match might have occurred by chance.

Ideally, one will be able to demonstrate the activities using sample sequences on a computer projected to the entire class. Students should have assistance on hand as they at-

Table 1. Sample Bioinformatics search results for poorly characterized genes. Genes were selected from Erasmus *et al*. (2003).

Sequence	Pubmed	Protein	Domains	BLAST	Other
YLR183C	No; indirect	ACCESSION NP_013284	Forkhead associated domain; phosphopeptide binding site	Plm2p; S-adenosyl-methioine synthetase; transcription factor Tos4 (40% identity)	3D structure for domains or homologs
YBL032W (Hek2/ KHD1)	Indirect; several through hek2/KHD1	ACCESSION AAT92833	K homology RNA-binding domain, PCBP_like; nucleic acid binding region; G-X-X-G motif	RNA binding protein; KH domain RNA binding protein	3D structures in Conserved Domain Database (CDD)
YBR028C (YPK3)	no	ACCESSION P38070	Protein kinase; Catalytic domain of AGC family Protein Ser/Thr Kinases; ATP binding site; Proton acceptor; activation loop; Protein kinase C terminal domain;	protein kinase; cAMP-dependent protein kinase substrate; ser/thr kinase psk1	Domains and homologs found in CDD

tempt the same activities on their own computers. Detailed pictorial instructions or video clips would help students navigate the NCBI tools outside of the lab. For example, Figure 1 illustrates the features of a BLAST search that would provide useful information to a student. However, some flexibility should be built into these instructions as the website layout can change without notice.

It may also be necessary to review basic concepts in molecular biology and biochemistry, so that students are able to make the appropriate inferences from their search results. Students may struggle with the concept of protein domains and the relationship between sequence, structure, and function. They may also have trouble with the search tools if they confuse nucleotide sequences with protein sequences, and vice versa. For example, students may try to input a nucleotide sequence into a protein sequence search for domains.

Students will also need to be shown how to present the information that they find. Many will create basic figures by cutting and pasting images or text directly from the NCBI, without any thought of whether the image is useful, without editing or processing the figure to make it easier for the reader to interpret, and without explanation of why the figure is important. For example, see Figure 2.

The final report

This activity has been used to generate a research paper in the style of a journal article. However, the Bioinformatics research project can be adapted for multiple uses. It could be shortened to a series of assignment questions that ask students to propose functions for their genes. Students could present their assigned sequences in short oral presentations or in posters, allowing them to share their findings with their peers. If the Bioinformatics project involves sequences that have been identified by a researcher in the department, that researcher could be invited to give a talk about the original study, and then invited to the student posters on the individual genes. The genes themselves could be used in virtual cloning exercises. For example, students could design PCR primers to detect or clone their sequences.

Student reports are graded for the accuracy of the analysis, the depth of research, the ability to synthesize results, and the flow and logic of the reasoning. Table 2 indicates the range of student results that have been seen in these reports.

Acknowledgements

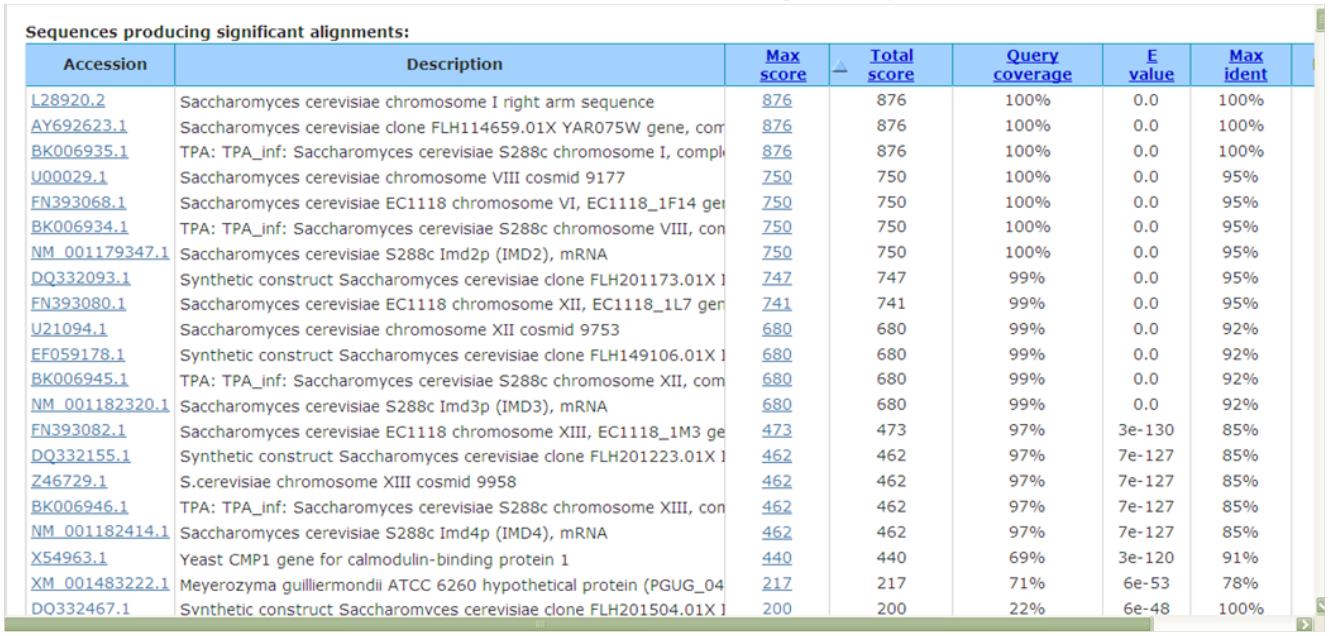
We would like to thank our BIOL 448 students who tested this activity to let us see what the common issues and difficulties would be, as well as the 2010W class of BIOL 360, who were the first students to do this research project as part of their course.

Pay attention to:

• **E value:** The probability that this sequence matches the query sequence entirely by chance. The smaller the probability, the stronger the match.

• **Max Ident:** Maximal Identity; The highest percentage of matching sequence

Click on **Accession** numbers to view the records for the aligned sequences.



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
L28920.2	Saccharomyces cerevisiae chromosome I right arm sequence	876	876	100%	0.0	100%
AY692623.1	Saccharomyces cerevisiae clone FLH114659.01X YAR075W gene, com	876	876	100%	0.0	100%
BK006935.1	TPA: TPA_inf: Saccharomyces cerevisiae S288c chromosome I, compl	876	876	100%	0.0	100%
U00029.1	Saccharomyces cerevisiae chromosome VIII cosmid 9177	750	750	100%	0.0	95%
FN393068.1	Saccharomyces cerevisiae EC1118 chromosome VI, EC1118_1F14 ge	750	750	100%	0.0	95%
BK006934.1	TPA: TPA_inf: Saccharomyces cerevisiae S288c chromosome VIII, con	750	750	100%	0.0	95%
NM_001179347.1	Saccharomyces cerevisiae S288c Imd2p (IMD2), mRNA	750	750	100%	0.0	95%
DQ332093.1	Synthetic construct Saccharomyces cerevisiae clone FLH201173.01X I	747	747	99%	0.0	95%
FN393080.1	Saccharomyces cerevisiae EC1118 chromosome XII, EC1118_1L7 ge	741	741	99%	0.0	95%
U21094.1	Saccharomyces cerevisiae chromosome XII cosmid 9753	680	680	99%	0.0	92%
EF059178.1	Synthetic construct Saccharomyces cerevisiae clone FLH149106.01X I	680	680	99%	0.0	92%
BK006945.1	TPA: TPA_inf: Saccharomyces cerevisiae S288c chromosome XII, com	680	680	99%	0.0	92%
NM_001182320.1	Saccharomyces cerevisiae S288c Imd3p (IMD3), mRNA	680	680	99%	0.0	92%
FN393082.1	Saccharomyces cerevisiae EC1118 chromosome XIII, EC1118_1M3 ge	473	473	97%	3e-130	85%
DQ332155.1	Synthetic construct Saccharomyces cerevisiae clone FLH201223.01X I	462	462	97%	7e-127	85%
Z46729.1	S.cerevisiae chromosome XIII cosmid 9958	462	462	97%	7e-127	85%
BK006946.1	TPA: TPA_inf: Saccharomyces cerevisiae S288c chromosome XIII, con	462	462	97%	7e-127	85%
NM_001182414.1	Saccharomyces cerevisiae S288c Imd4p (IMD4), mRNA	462	462	97%	7e-127	85%
X54963.1	Yeast CMP1 gene for calmodulin-binding protein 1	440	440	69%	3e-120	91%
XM_001483222.1	Meyerozyma guilliermondii ATCC 6260 hypothetical protein (PGUG_04	217	217	71%	6e-53	78%
DQ332467.1	Synthetic construct Saccharomyces cerevisiae clone FLH201504.01X I	200	200	22%	6e-48	100%

Note: YAR075W has a 100% identity and E=0 for itself and for the chromosome that it's found on– it's a perfect match for itself (Max Ident = 100%), and there is no chance that this could have happened at random (E value = 0).

Figure 1. Pictoral guide to reading BLAST results. This was one of a series of figures used to highlight features of the BLAST records that the student would find useful.

Figure 2 (next page). Examples of student figures for Nucleotide BLAST results **A.** Graphical representation of Nucleotide BLAST results for YAR075W. On the BLAST site, each red bar is hyperlinked to a Nucleotide record, and the name of the sequence appears when the mouse pointer hovers over bar. However, this information is lost when the graphic is copied out of context in a student figure. **B**. Nucleotide BLAST results for YAR075W. Students may also take a screen shot of the BLAST table of results, without defining the terms, without noticing that the sequence descriptions are cut off in the figure, and without editing out unnecessary information. **C.** Table of Nucleotide BLAST results for YAR075W. This table takes selected sequences from the original BLAST results and places them in a table, along with key information that would be useful for the reader.

A

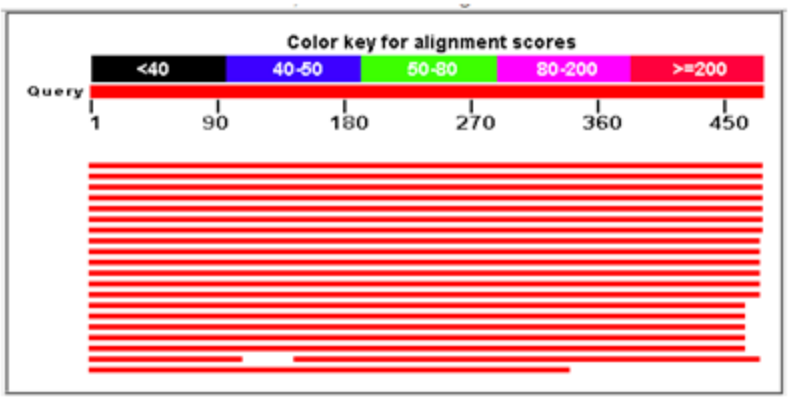


Figure 1. Nucleotide BLAST search of YAR075W. These sequences show close homology to YAR075W.

B

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
BK006935.1	TPA: TPA_inf: Saccharomyces cerevisiae S288c chromosome I, compl	876	876	100%	0.0	100%
AY692623.1	Saccharomyces cerevisiae clone FLH114659.01X YAR075W gene, com	876	876	100%	0.0	100%
L28920.2	Saccharomyces cerevisiae chromosome I right arm sequence	876	876	100%	0.0	100%

Figure 2. Nucleotide BLAST search of YAR075W. These sequences show close homology to YAR075W.

C

Table 1: Nucleotide BLAST search of YAR075W. Selected sequences are shown.

Accession number	Description	E value	Maximum identity
NM_001182320	IMD3 (Inosine monophosphate dehydrogenase 3) from S. cerevisiae	0	95%
X54963	Yeast CMP1 gene for calmodulin-binding protein 1	4×10^{21}	91%
NM_001182414	IMD4 (inosine monophosphatase dehydrogenase 4) from S. cerevisiae	8×10^{127}	85%

Literature Cited

Erasmus, D. J., van der Merwe, G. K., & van Vuuren, H. J. J. (2003). Genome-wide expression analyses: Metabolic adaptation of *Saccharomyces cerevisiae* to high sugar stress. *FEMS Yeast Research* 3: 375-399.

Meissner, B., Warner, A., Wong, K., Dube, N., Lorch, A., *et al*. (2009). An integrated strategy to study muscle development and myofilament structure in *Caenorhabditis elegans*. *PLoS Genet.* 5(6): e1000537. Doi:10.1371/jornal.pgen.1000537

National Center for Biotechnology Information. (n. d.). Basic Local Alignment Search Tool. Retrieved from **http://blast.ncbi.nlm.nih.gov/Blast.cgi**

National Center for Biotechnology Information. (n. d.). CD-ART: Conserved domain architecture retrieval tool. Retrieved from **http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps**

National Center for Biotechnology Information. (n. d.). Conserved Domains. Retrieved from **http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi**

National Center for Biotechnology Information. (n. d.). Constraint-based multiple alignment tool. Retrieved from **http://www.ncbi.nlm.nih.gov/tools/cobalt/**

National Center for Biotechnology Information. (n. d.). Structure. Download Cn3D 4.1 for PC, Mac and Unix. Retrieved from **http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml**

National Center for Biotechnology Information. (n. d.). Welcome to NCBI. Retrieved from **http://www.ncbi.nlm.nih.gov/guide/**

About the Authors

Dr. Liane Chen is an Instructor in the Departments of Zoology and Botany, at the University of British Columbia. She currently teaches courses and labs in cellular and molecular biology. This Bioinformatics research project came out of an overly ambitious idea that she pitched in her first academic interview ever—it included a lab to clone these rare genes—and she has retained an interest in exploring and developing effective teaching methods.

Dr. Kathryn G. Zeiler is an Instructor in the Departments of Zoology and Botany at the University of British Columbia. She has taught a wide range of college-level biology courses including cell biology, molecular biology, microbiology, botany, field biology, field botany, and learning communities between biology and composition as well as between biology and geology. Kathryn has developed and adapted many biology courses and currently works with Liane to develop and deploy upper-level cell and molecular biology lab courses, along with other interesting teaching and learning activities.

Table 2. Student results by grade.

Grade	Report characteristics
A	There is a clear flow of logic from bioinformatics analysis to hypothesis to proposed research. Specific evidence is given, and is well-supported by the scientific literature. Figures are informative and well-explained.
B	Reasonable analysis, hypothesis, and proposed research. There are occasional gaps in the logic, and evidence and experiments can be vague or general, but there is support from the literature. Figures are generally OK.
C	The bioinformatics exercises are satisfactory. A general hypothesis and experimental plan is given but is poorly supported by further research. Figures consist mainly of rough cut and paste images, and are not well explained.
D	The exercises are minimal and the ideas in the report are inconsistent. Figures consist mainly of rough cut and paste images, and are not well explained.

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with biology learning and teaching in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/**

Papers published in *Tested Studies for Laboratory Teaching: Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

Chen, L., and Zeiler, K.G. 2012. Biology Laboratory Course. *Tested Studies for Laboratory Teaching,* Volume 33 (K. McMa-hon, Editor). Proceedings of the 33rd Conference of the Association for Biology Laboratory Education (ABLE), 390 pages. http://www.ableweb.org/volumes/vol-33/?art=3