Lab Project 1: Introduction to NCBI Database Search: Protein and Gene Sequences

Due: October 15, 2014

Protein Accession Numbers
Mesafint Fanuel: NP_578001  Tiffani Gardner: NP_598001

The purpose of this lab project is to get familiar with the NCBI website and know how to search for a particular protein and its corresponding gene as well as study their characteristics.

Introduction to NCBI: The National Center for Biotechnology Information (NCBI) maintained by the US National Library of Medicine and National Institutes of Health is one of the world’s most important resources and repositories for biological data. URL: http://www.ncbi.nlm.nih.gov/. This fantastic online resource provides an extensive network of databases cataloging an ever-growing wealth of genetic, medical, and biochemical information from all walks and crawls of life. Entire genomes, from viruses to humans, are compiled, organized, and cross-referenced within these networks, such that surfing the genome can be almost as easy as surfing the web.

Gquery: The primary portal for accessing data at NCBI is called GQuery. To move to the GQuery portal – select All Databases from the navigation bar at the top of the NCBI start page, by clicking “Search” on the empty field. First, scan down the assortment of databases queried through GQuery. You will notice there is everything from the biomedical literature at PubMed to nucleotide databases, taxonomy databases, protein structure databases, and expression profile databases.

Record Identification: Records in the NCBI database are identified using several identifiers. We will use the accession number to retrieve the most recently reported amino acid sequence for a protein and its corresponding gene nucleotide sequence. An Accession number is a unique identifier for a particular sequence record. Even if the sequence record gets updated, the accession number remains the same. However, to uniquely identify an earlier sequence and its updated versions, an accession number is accompanied by a version number. Version numbers follow the Accession number and indicate the revision history of that entry starting with 1 and increasing with each revision. The standard format is Accession.Version.

A GI number (GenInfo Identifier – sometimes written in lower case, "gi") is simply a series of digits that are assigned consecutively to each sequence record processed by NCBI. The GI system of identifiers runs parallel to the accession.version system; therefore, if the DNA or protein sequence changes in any way, it will receive a new GI number.

The Accession number will always give you the most up to date information on a record, while the GI number will always take you back to a specific record. There are times when you want the most current information, and other times when you want to point to a particular piece of information from a particular point in time (e.g. a particular record that you did an analysis with), even if more information has been subsequently added.

Example: When a new entry is submitted to GenBank (a NCBI database), it will be assigned an accession number (say AF000001). Since this is the first version the Accession will be appended with ".1", so it will look like AF000001.1. At the same time it will be given a GI number (say GI:1234567). Now imagine that the researcher who originally submitted the record wants to update the information. The updated record will keep the same Accession
number, but increase in version number (AF000001.2), which the new record will be given a completely new GI number (say GI:9876543).

**Protein and Gene to be searched:** In this lab project description, I will explain the procedure of searching the NCBI database using a sample protein and its corresponding gene. For your lab project submission, you will repeat this procedure for the protein and gene assigned to you. I will use a gene from *Arabidopsis Thaliana*, a small flowering plant (it is like the fruit fly of the plant world) that has a rapid life cycle and requires little space to grow. The protein product of this gene is recorded under accession number NP_565676.

1. Go to the Gquery browser (as mentioned above) and type the accession number assigned to you in the search space. You will get a list of records corresponding to this. We are interested in the protein and gene records for this accession number.

![](image1)

2. Click on the Protein link. I get a page as shown below. You would get a different page depending on the accession number assigned to you.

![](image2)

**Questions to Answer:**
Include relevant screenshots for your reports, as well as show the entries in the web page as you talk (during recording)

Q1) How many amino acids are part of the protein assigned to you?

Q2) What is the version number and GI number for the record assigned to you? If its version number is greater than 1, what is the GI number of the sequence that was replaced by your assigned sequence?
Q3) What is the molecular mass of your protein?

Q4) In your protein page, you may see one or more regions with characteristic domain names (like RING, ARM, etc). If present, these regions are indicated with the starting and ending sequence numbers of amino acids that constitute them. You could also retrieve them from the “Conserved Domains” link under the Related Information section on the right side.

Identify the domain regions of your protein (name, brief info, etc) and their starting/ending amino acid sequence numbers.

Q5) What is the encoding mRNA for your protein (list its name, accession number, version number and GI) as well as the number of bases constituting it?

Q6) Show the snapshot of the Amino Acid sequence constituting your protein?

Q7) Are any “identical proteins” reported in NCBI for the protein assigned to you? If so, give their Accession number and GI numbers? If so, list them.

Q8) What are the top 5 proteins with sequences related to yours? How many amino acids constitute these proteins and give their Accession #s/ GI #s? How many of these top 5 proteins do you think have amino acid length close to that of your protein?

Questions to Answer:
Include relevant screenshots for your reports, as well as show the entries in the web page as you talk (during recording)

Q9) Identify the name of the gene, its symbol, description and ID.

Q10) What is the chromosome number in the organism that this gene is present? Locate the starting and ending nucleotide sequence numbers of this gene in that chromosome.

Q11) How many exons are part of this gene sequence? What are their starting and ending nucleotide sequence numbers?

Q12) Locate the genes (names, their starting/ending nucleotide sequence numbers in the chromosome) in the neighborhood of the gene assigned to you? Also, show the NCBI graphics that illustrates this.

Q13) How many nucleotide bases are part of the CDS region of the gene? Identify their starting/ending nucleotide sequence numbers? Compare these with those observed for the mRNA nucleotide sequence and explain the reasons for the overlap as well as non-overlap in certain areas.

Q14) What are the functions and processes that your gene/protein is associated with? List them and explain with a sentence or two for each.

Q15) Select the FASTA format under the Genomic regions to display your gene’s nucleotide sequence in text format. Take a screenshot of this. We will use this format for future lab projects.

Submission:
Submit a report featuring your answers for questions 1 through 15, along with the necessary screenshots. Also, record a video (using a Desktop recording software) of your search/query process and identification of the answers for questions 1 through 15. Upload your video through Googledrive and submit the report in class.