National Center for Biotechnology Information (NCBI)
Bioinformatics

1. What subjects and publication types are included in the database?

   Genome data – from any species. There are 38 databases covering different aspects of the subject. There are three search engines – Entrez searches all 38 databases; BLAST & VAST search sequence information found in the databases.

2. What are the coverage dates?

   1988 - present

3. What Boolean or Proximity operators are used by the database?

   AND     OR     NOT
   However, each database handles a space differently. In some, a space between terms is an implied AND, while in others, a space between terms is an implied OR. So click on the Details tab within each database’s results page to determine how that database interpreted your query.

4. If truncation is allowed then what symbols are used?

   * = unlimited number of characters
   The system allows you to do left and right truncation.

5. If wildcards are allowed then what symbols are used?

   Not available

6. How do you search for a phrase?

   Enclose your phrase in quotes.

7. Any unique features?

   Cn3D – software that allows you to see and manipulate the structure of a protein.
   BLAST – searches a base pair or amino acid sequence to find sequences that are closely related to your search.
   VAST – searches the structure of the protein to find other proteins that have a similar structure.
   PubChem Compound – contains validated chemical depiction information provided to describe substances in PubChem Substance.
   PubChem Substance – contains descriptions of chemical samples, from a variety of sources, and links to PubMed citations, protein 3D structures, and biological screening results. Also searchable via Reaxys.
   NCBI Entrez – a unified search engine for all 38 NCBI databases.
   Each database has a different set of stop words, i.e., words that it ignores.
The practice will focus on four databases:
- Nucleotide
- Protein
- OMIM
- Structure

And one viewer program
- Cn3D

As well as show two other databases
- PubChem Compound / PubChem Substance

And show two search engines
- VAST
- BLAST


Otherwise follow these directions:
1. Go to the Chemistry Library web page (chemistry.library.nd.edu)
2. Click on more... under the Databases section in the Search & Find tab
3. Scroll down to NCBI - Entrez

NCBI Entrez – OMIM

When do I search OMIM?
Search OMIM when looking for a possible genetic cause of a human disease such as genetic mutations (a.k.a. allelic variants).

1. Roughly 1 million people in the United States have Parkinson’s Disease (PD). There is no known gene mutation producing this disease but it does “run in families.” Michael J. Fox, Rev. Billy Graham, and Muhammad Ali are well known PD sufferers. What can be found at NCBI regarding this disease?

Type Parkinson into the All Databases search box at the top of the page.
(Entrez cross-database search results page appears.)

How many OMIM hits do you get? _____
(You should get 225-275)

Click on OMIM.
You are now looking at the first 20 hits. Each hit has at most three lines of information.

1st line gives the **record number** and **disease name(s)**.

2nd line may provide additional information regarding the disease.
   Some records don’t have this second line.

3rd line gives the **gene map locus** or the location on a chromosome.
   Some records don’t have this third line.

On the right side of the results page is a **Links** link that provides rapid access to other databases about your specific protein or gene. There may be other links.

Click the **Links** link for **record number #168601**. How many links appear? ____
   (don’t select any of the links)  (You should get 15-25 links)

Click on the **Limits** tab. This tab is near the top of the page.
   You will see four boxes with a title in the grey bar area of each box.

2. In the **MIM Number Prefix:** box select **gene with known sequence**
   In the **Only Records with:** box select **Allelic Variants**
   Check those boxes and click GO.

   How many hits did you get? ____
   (You should get 50-100.)

Select **record number #602544**.

3. You are now looking at the record for the PARK2 gene. On the right side is the Table of Contents for this article as well as links to other databases.

Click on the **Table of Contents** link then **Allelic Variants** heading.

   How many allelic variants (or genetic mutations) are there for the PARK2 gene? ____
   (There should be 20-25.)

4. Each variant lists the phenotype and mutation (this includes gene name and mutation location).

   A. Look at the third item in the list. Where is the gene mutation? ____
      (Hint: At what position in the sequence was THR replaced by ARG?)

   B. How many other mutations happen at the same position in the sequence? ____
NCBI – Protein

When do I search Protein?
Search NCBI Protein anytime you need an amino acid sequence of a protein.

Click the browser back button until you return to the NCBI OMIM page (with the NCBI logo in the upper left corner). Then select All Databases from the black horizontal bar.

5. Type parkin into the search box.

How many hits do you get in the Protein database? __________
(You should get 12,000-13,000)

Click on Protein to get the results list.

You are now looking at the first 20 hits. Each hit has four lines of information.
1st line gives the protein name; [organism name]
2nd line gives the number of amino acids in the protein
3rd line gives NCBI record information.
4th line provides links to related information

6. Many of the records have not been curated (that is they have not been checked for accuracy.) In NCBI a curated record is called RefSeq.

Do either A or B:
A. In the Filter your results: section, click on the RefSeq (413) link.

OR

B. Click on the Limits link. Then select RefSeq in the Source Database: area. Search.

How many RefSeq records are there for the parkin protein? ______
(You should get between 400-450.)

To the right of the list shows the top organisms with entries for this protein.

7. How many organisms have 5+ records for the parkin protein? ______
Look for the Top Organisms box. You may need to click the More... link.
(You should get 10-20.)

How many of these protein sequences are from humans? ______
(You should get 75-100)

Click on the # for Homo sapiens
Select record NP_004553.2

You are now looking at an NCBI Protein record. The record is divided into four sections:

1. **Protein identification information** (LOCUS to SOURCE ORGANISM) including the number of amino acids in the protein or gene.

2. **Literature references** (REFERENCE)

3. **Information about the protein** (COMMENT to FEATURES) including chromosome location, number of amino acids in the sequence and sometimes, but not in this record, there are allelic variants listed.

4. **Sequence** (ORIGIN) the sequence is in a standardized form – 60 amino acids per line and a space after ten amino acids. The number preceding each line is the number for the first amino acid in that line of the sequence.

8. How many amino acids are in this protein? _____
   Hint: Look in the LOCUS section.

9. On which chromosome is this gene found? _____
   Hint: Look in the FEATURES – source section.

Also note that the record gives the location on the chromosome – q25.2-q27 (q arm, region 25.2 to 27).
This information was also found in OMIM at the top of the record.

10. Turn back to question #4. What was your answer to question #4A? _______
    (What is the position of the gene mutation?)

    List the five amino acids in the sequence following that position. _______________
    (Hint: Go to the bottom of the record.)
NCBI – Nucleotide

When do I search Nucleotide?
Search Nucleotide anytime you need a nucleotide sequence, nucleic acid sequence, or base pair sequence

11. On the right side of the Protein page there are several sections. Look for the Related information section. Then click the Nucleotide link.

How many hits did you get? ________
(You should get between 5-10.)

Each of the entries has four lines of information.
1st line gives the definition for the sequence
2nd line gives the number of base pairs
3rd line gives the record number information
4th line provides links to related information

Select record NM_004562.2

You are now looking at the NCBI Nucleotide record. The record has the same four sections as the NCBI Protein record.

12. How many base pairs are in this mRNA sequence? __________
(Hint: Look in the LOCUS section.)

13. List the 10 base pairs beginning with 2311. ________________
(Should begin with a and end with t)
NCBI – Structure

When do I search Structure?
Search NCBI Structure anytime you want to view the structure of a protein, or you want to see how many chains the protein is composed of, or you want to see if any other protein is structurally similar to your protein.

14. Type 1iyf into the search box. Select Structure from the drop-down menu.

1IYF is the PDBid number for part of the parkin protein.

15. What is the MMDB record number? ______________________

16. How many residues (aka amino acids) are in 1iyf? ____________
(See the bar graph near the bottom of the page – Click Show annotation.)

17. What is the name of the domain families – specific hits? ________________

This is also known as the “Conserved Domain” that 1iyf may have evolved from. The domain proteins represent the building blocks for other proteins in all organisms. See your answer to Q7 for how many other organisms use this same domain.

18. Click the Domain Families link. How many residues are different between 1iyf and the conserved domain protein?
(Click the + sign in front of the conserved domain protein in #17.)

The blue letters represent differences between the two proteins.

Click the browser’s back button to return to the NCBI Structure Summary MMDB page. Click View Structure button to the right of the large picture of the protein.

This should launch the Cn3D software. Two windows will open. The window with the white background (Alignment/Sequence Viewer) shows the amino acid sequence of the protein in the window with the black background.

Notice that the color of the letters in the Alignment/Sequence Viewer window matches the color in other window.

Place your cursor anywhere on the structure of the protein. Click and hold and drag the cursor around to move the protein structure.
19. How many β sheets (flat arrows or lines) do you see? ______

How many α helices (cylinders) do you see? ______

20. Find one of the ends of the protein and double click on it.

What color does it change to? ______

What happens in the sequence in the Sequence/Alignment Viewer window?

21. How many amino acids are within 5 Angstroms of the aspartic acid (d) that is in position PDB 39 of the protein sequence? ______

(You should get 10-15 residues.)

In the Sequence/Alignment Viewer window
1. Hover the tip of the cursor over the sequence. In the lower left you will see some numbers.
2. Move the cursor until PDB 39 appears.
3. Click on the aspartic acid (d) that is in position PDB 39 of the sequence. The letter d is now highlighted yellow as is the residue.
4. Go to the Select menu – Select by Distance.
5. Enter a distance of 5 Angstroms. Check Select protein residues (uncheck everything else).
6. Count the number of highlighted amino acids in the sequence/alignment viewer

22. For fun (Optional) – change the view and coloring of the protein.

The Style menu – Rendering shortcuts changes the view of the protein. Select tube (or anything else.)

The Style menu – Coloring shortcuts changes the coloring of the protein. Select charge (or anything else.)
NCBI – VAST

**When do I do a VAST search?**
Search VAST (Vector Alignment Search Tool) anytime you want to see proteins that have a similar structure to another protein.

Go back to record MMDB 22283 in the browser. Click Show annotation near the bottom of the page. Click the grey bar to the right of 1 Protein.

You are now looking at results of a VAST search. The four bars immediately below the yellow section are for the 1IYF protein and the domain families, just like you saw in the Structure record.

The remaining bars represent the most structurally similar proteins in descending order. To the left of those bars are the PDBids for those proteins.

23. What protein has the greatest structural similarity to 1IYF? ________________

Click the check box to the left of 1V6E then click the View 3D Alignment button. (gray means there is no structural similarity between the proteins, blue means there is structural similarity but different amino acids, and, red means the proteins are structurally similar and the amino acids are the same.)

24. What taxonomy (genus species) does 1V6E come from? ________________

Go back to the VAST search results page. Click on the 1V6E link. Look for Source organism: in the upper right part of the page.

What is the common name? ____________________
(click on the genus species name)

NCBI – BLAST

**When do I do a BLAST search?**
Search BLAST (Basic Local Alignment Search Tool) anytime you want to find regions of local similarity between nucleotide sequences or protein sequences. The program compares the sequence you entered to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
25. For Fun (Optional) – What is the name of the protein that has the most amino acids in common with 1IYF?

Go back to record MMDB 22283.

1. Click on a link in the Molecules and interactions section to go to the NCBI Protein record.
2. Scroll to the bottom of the record.
3. Highlight and Copy the entire sequence (all 81 aa).
4. Click on the NCBI logo in upper left corner of the page.
5. Click on the BLAST link in the Popular Resources section.
6. In the Basic BLAST section click on the protein blast link.
7. In the Enter Query Sequence area paste the sequence into the big search box – delete any line numbers.
8. Click the BLAST button
9. WAIT – about 60 seconds. The screen will refresh with an answer.

The results page has three sections after the information about your search and your sequence.

1. The first section (Graphic Summary) has a chart showing your sequence at the top then 50 line representing 50 sequences that most closely match the sequence.

2. The second section (Descriptions) lists the name of the sequence and the Score and E value (the higher the score and the lower the E value the better.) There are also links to other NCBI databases: S – Structure; G – Gene; M – Map View; U – Unigene.

   Notice that the first answer is 1iyf – that is the protein sequence we BLASTed.

   The 15th answer is NP_004553.2 – that is the record we retrieved in question 7 when we limited the parkin protein search to RefSeq records.

3. The third section (Alignments) compares your sequence with each answer. Each answer has three sequence lines. The top amino acid sequence is yours. The bottom sequence is a closely matched sequence. And the middle line shows any amino acid that is in both sequences at the same point. Scroll down to the last answer to see more clearly how this works. There may also be links to PubMed articles.

Finding your compound in NCBI:

Click on the NCBI logo in the upper left and search for your compound in All Databases.

26. Is your compound mentioned in OMIM? (You may need to deselect the Limits from Q2)
   If yes, then name one genetic disorder it is associated with.

27. Is your compound listed in PubChem Compound or PubChem Substance?