Protein Bioinformatics
Part I: Access to information

260.655
April 6, 2006
Jonathan Pevsner, Ph.D.
pevsner@kennedykrieger.org

Outline

[1] Proteins at NCBI
  • RefSeq accession numbers
  • Cn3D to visualize structures

[2] The Protein Data Bank (PDB)

[3] UniProt

[4] ExPASy (Expert Protein Analysis System)
  • DeepView, the Swiss-Pdb Viewer.

Central dogma of molecular biology

DNA → RNA → protein

genome → transcriptome → proteome

Central dogma of bioinformatics and genomics
There are three major public DNA databases

- **EMBL** Housed at EBI European Bioinformatics Institute
- **GenBank** Housed at NCBI National Center for Biotechnology Information
- **DDBJ** Housed in Japan

Accession numbers are labels for sequences

NCBI includes databases (such as GenBank) that contain information on DNA, RNA, or protein sequences. You may want to acquire information beginning with a query such as the name of a protein of interest, or the raw nucleotides comprising a DNA sequence of interest.

DNA sequences and other molecular data are tagged with accession numbers that are used to identify a sequence or other record relevant to molecular data.
What is an accession number?

An accession number is label that used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples (all for retinol-binding protein, RBP4):

- X02775  GenBank genomic DNA sequence
- NT_030059  Genomic contig
- R6757948  dbSNP (single nucleotide polymorphism)
- N91765.1  An expressed sequence tag (1 of 328)
- NM_006744  RefSeq DNA sequence (from a transcript)
- NP_007636  RefSeq protein
- AAC02945  GenBank protein
- Q28369  SwissProt protein
- 1KT7  Protein Data Bank structure record

Accessing protein sequences via Entrez

Entrez Gene with RefSeq

Entrez Gene is a great starting point: it collects key information on each gene/protein from major databases. It covers all major organisms.

RefSeq provides a curated, optimal accession number for each DNA (NM_006744) or protein (NP_007635)

Example #1. Sean mentioned silk fibroin. How do you find its sequence?

From the NCBI home page, type "silk fibroin" and hit "Go"
Example #2. Find the sequence of myoglobin

From the NCBI home page, type “myoglobin” and hit “Go”
1. Entrez Gene entries offer a wealth of information, including links to RefSeq entries and to the Human Protein Reference Database (HPRD; Akhilesh)

2. HomoloGene provides access to RefSeq identifiers of a protein family, and offers links (domains, pairwise alignments)

3. Entrez Protein shows 922 myoglobins (too many), including 47 RefSeq (still a lot).
You can try scrolling through the RefSeq list, or apply "Limits".

As another approach, click "TaxBrowser"…

Enter the name of the organism you are interested in…

Follow a link of interest…

Now click protein…

You now can view all sperm whale proteins…
…and restrict the output to sperm whale myoglobin.

4. Structure provides links to myoglobin structures

Access to PDB through NCBI

- Molecular Modeling DataBase (MMDB)
- Cn3D ("see in 3D" or three dimensions): structure visualization software
- Vector Alignment Search Tool (VAST): view multiple structures
You can limit the output to particular species, e.g. with the command human[organism].

Click 2MM1 to enter MMDB, the Molecular Modeling Database.
Overlay two or more structures with VAST at NCBI

Click "Chain"...

Click one or more boxes then "View 3D Structure"...
Overlay two or more structures with VAST at NCBI

Access the Conserved Domain database at NCBI

Click “globin”…
The Protein Data Bank (PDB)

- PDB is the principal repository for protein structures
- Established in 1971
- Currently contains over 35,000 structure entities

Updated 3/06
PDB holdings (September, 2005)

<table>
<thead>
<tr>
<th>Category</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins, peptides</td>
<td>29,876</td>
</tr>
<tr>
<td>Protein/nucl. complexes</td>
<td>1,338</td>
</tr>
<tr>
<td>Nucleic acids</td>
<td>1,500</td>
</tr>
<tr>
<td>Carbohydrates</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>32,727</td>
</tr>
</tbody>
</table>

Search for keyword DNAC yields mouse zinc finger binding proteins
UniProt (Universal Protein Resource) at www.uniprot.org

UniProt combines information in Swiss-Prot, TrEMBL, and PIR. UniProt is comprised of three components:

- The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information.
- The **UniProt Reference Clusters (UniRef)** databases combine closely related sequences into a single record.
- The **UniProt Archive (UniParc)** is a comprehensive repository, reflecting the history of all protein sequences.

[3] UniProt
Example: search for *E. coli* DnaC at NCBI

Approach: NCBI ➔ TaxBrowser ➔ *E. coli* ➔ proteins ➔ dnac ➔ RefSeq ➔ three entries shown here.

Example: search for *E. coli* DnaC at UniProt

202 entries found
Example: search for *E. coli* DnaC at UniProt

Add input box (organism, coli) and 20 entries found

The UniProt entry links to Entrez Gene (but not RefSeq)

ExPASy to access protein and DNA sequences

ExPASy sequence retrieval system
(ExPASy = Expert Protein Analysis System)

Visit http://www.expasy.ch/

ExPASy Proteomics Server

When you search the ExPASy database, you are now querying the UniProt Knowledgebase.

► UniProtKB/Swiss-Prot, a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.

UniProtKB/Swiss-Prot Release 49.3 of 21-Mar-2006: 212,425 entries (More statistics)
ExPASy to access protein and DNA sequences

When you search the ExPASy database, you are now querying the UniProt Knowledgebase.

UniProtKB/TrEMBL; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

UniProtKB/TrEMBL Release 32.3 of 21-Mar-2006: 2,666,963 entries

Example: find human tyrosinase at ExPASy

From the ExPASy home, click Swiss-Prot → SRS → Start → Continue
Example: find human tyrosinase at ExPASy

Note the lack of RefSeq and multiple accession numbers
Choose the right protein by inspection (P14679)
Example: find human tyrosinase at ExPASy

Example: find human tyrosinase at NCBI

Human tyrosinase: blastp against pdb to find known structures
Human tyrosinase: blastp against pdb

Go to www.pdb.org and enter 2AHL...

...click Swiss PDB viewer
View protein structures with DeepView

- move (translate)
- rotate
- measure dihedral angles (ω, ϕ, ψ) from a selected atom.
- measure bond angles (pick center atom, then two more atoms)
- click, select two atoms, determine distance in angstroms
- identify an atom (and the group to which it belongs). Type: CA, CB, O  
  Group: LYS116, etc.  
  x, y, z atom coordinates
► Go to window ➔ control panel.

► Shift/click to select the first five amino acid residues of myoglobin. They should appear red.
► Click “labl” (i.e. label)(see arrow, above right). Those five residues now have a “v”.
► Inspect the display panel; those five residues are labeled.

Download and practice using DeepView! Try using myoglobin.

The ExPASy download site includes a helpful web-based tutorial
http://www.usm.maine.edu/~rhodes/SPVTut/