ANALYZING YOUR PROTEIN USING BIOINFORMATICS TOOLS

Predicting simple physicochemical properties

Predicting protein structure and function
Protein biochemistry using a computer

**Major tools:**


Swiss EMBnet: [www.embnet.org](http://www.embnet.org)

All you need to know is the sequence of your protein
ExPASy

ExPASy server is a World-leading resource for protein information.

It is a home of UniProtKB/Swiss-Prot sequence database, which is a manually annotated protein knowledge-base established in 1986 and is now maintained by Swiss Institute of Bioinformatics.

The annotations include the description of the following items:

1. Function(s) of the protein
2. Post-translational modification(s). For example carbohydrates, phosphorylation, etc.
3. Domains and sites. For example calcium binding regions, ATP-binding sites, etc.
4. Secondary structure
5. Quaternary structure. For example homodimer, heterotrimer, etc.
6. Similarities to other proteins
7. Disease(s) associated with deficiencie(s) in the protein
8. Sequence conflicts, variants, etc.

ExPASy also contains a suite of programs which help to analyze your protein on the basis of its sequence
Predicting the main physicochemical properties of a protein

The program ProtParam, a component of the ExPASy server, estimates many basic physicochemical properties of a polypeptide on the basis of its sequence.

Point your browser to: www.expasy.org/tools/

Click the ProtParam link
Either put the accession number of your protein or paste the amino acid sequence. Then click the ‘Compute parameters’ button.
Limit the sequence you want to analyze (the coordinates of the N- and C-termini). The program will analyze the entire sequence by default.
ProtParam

- sequence and its limits
- Molecular weight
- isoelectric point (pI)
- amino acid composition
- atomic composition
- chemical formulae
- extinction coefficient
- estimated half life

A crude estimate of protein stability. If below 40, the protein is stable.
Predicting peptide produced by cutting a protein with specific proteases

Cutting a protein by proteases is often used when you want to study only a portion of your protein. It is also a critical step for preparing a protein for mass-spectrometry analysis. The program PeptideCutter from the ExPASy server (www.expasy.org/tools/) predicts the size of peptides produced from the protein by action of proteases with different specificity.
**PeptideCutter**

1. Put either the accession number or the amino acid sequence.
2. Select one protease or a combination of the protease enzymes.
3. Click the ‘Perform’ button.
PeptideCutter

your sequence

positions of cleavage sites within the protein

sequence of the peptide

length of the peptide

mol weight of the peptide

Available enzymes

Trypsin

Positions of cleavage sites

14 1 157.782
17 1 316.350
25 8 970.137
29 4 561.657
33 4 509.658
41 8 961.086
47 6 690.813
65 18 2323.588
68 3 386.496
70 2 205.244
72 2 259.349
78 6 716.805
87 9 1056.141
94 7 706.751
103 9 1015.157
114 11 1335.433
130 16 1611.907
141 11 1258.374
144 3 401.468
154 10 1188.370
157 6 288.347
163 6 557.691
205 42 4650.403
206 1 174.203
207 1 174.203
221 14 1385.809
242 21 2263.504
250 8 927.085
251 1 146.189
262 11 1206.848
263 1 174.203
267 4 551.640
274 7 746.750
305 28 2976.364
307 2 562.039
310 18 1812.104
311 10 1878.934
Primary Structure Analysis

A “sliding window” approach allows to average the property of a segment of a protein.

MRGATYWFFKEDRKPPLLMTCSGTYPOIVVQNERKDKTHPLOKDFVNMOGGFKRYF

Choose a window of N amino acids and average the property of amino acids within the amino acid residues within the window. Then slide the window by one amino acid and do it again. Plot the results.

The properties can include hydrophobicity, charge, propensity to form an alpha helix, etc.
Prediction of transmembrane domains in proteins

The knowledge of whether your protein has transmembrane domains tells you a lot about possible biological functions of the protein, its physicochemical properties and its behavior during purification. The presence of a transmembrane domain at the N-terminus suggests that the protein is secreted. The presence of several transmembrane domains suggests that it is a membrane protein.

Many transmembrane domain prediction algorithms are based on the sliding window approach which trace the hydrophobicity profile.

Na+/H+ antiporter of Arabidopsis thaliana
Using ProtScale for prediction of transmembrane domains

Point your browser to: www.expasy.org/tools/

ProtScale - Amino acid scale representation (Hydropobicity, other conformational parameters, etc.)

ProtScale

- De novo repeat detection in protein sequences
- Searches a protein sequence for repeats
- REPRO - De novo repeat detection in protein sequences
- TRUST - De novo repeat detection in protein sequences
- XSTREAM - De novo tandem repeat detection and architecture modeling in protein sequences

- SAPS - Statistical analysis of protein sequences at EMBnet-CH [Also available at EBI]
- Coils - Prediction of coiled coil regions in proteins (Lupas's method) at EMBnet-CH [Also available at PBIL]
- Paircoil - Prediction of coiled coil regions in proteins (Berger's method)
- Paircoil2 - Prediction of the parallel coiled coil fold from sequence using pairwise residue probabilities with the Paircoil algorithm.
- Multicoll - Prediction of two- and three-stranded coiled coils
- 2ZIP - Prediction of Leucine Zippers

- PESTfind - Identification of PEST regions at EMBnet Austria

- HLA_Bind - Prediction of MHC type I (HLA) peptide binding
- PEPVAC - Prediction of superpep MHC binders
- RANKPEP - Prediction of peptide MHC binding
- SYFPEITHI - Prediction of MHC type I and II peptide binding

- ProtScale - Amino acid scale representation (Hydropobicity, other conformational parameters, etc.)
- Drawhca - Draw an HCA (Hydrophobic Cluster Analysis) plot of a protein sequence
- Peptide Builder
- Protein Colourer - Tool for coloring your amino acid sequence
- Three To One and One To Three - Tools to convert a three-letter coded amino acid sequence to single letter code and vice versa.
- Three-one-letter amino acid converter - Tool which converts amino acid codes from three-letter to one-letter and vice versa.
- Colorseq - Tool to highlight (in red) a selected set of residues in a protein sequence
- HelixWheel / HelixDraw - Representations of a protein fragment as a helical wheel

- RandSeq - Random protein sequence generator
Using ProtScale for prediction of transmembrane domains

Accession number or sequence

Select property

Select the size of the sliding window (19 is optimal for transmembrane domains)

Submit
Using ProtScale for prediction of transmembrane domains

Click the ‘range’ to see the hydrophobicity profile

Putative TM domains
Using ProtScale for prediction of transmembrane domains

Putative TM domains

Cut-off line
(1.6 for Kyte & Doolittle algorithm)

A “good” signal should be robust and should ‘show-up’ if you use different algorithms.
Looking for motifs in your protein

Many specific biological functions are associated with relatively short amino acid sequence motifs in proteins.

Finding such motifs tells you a lot about your protein. For example, posttranslational modifications are often associated with specific sequence motifs.

However, one needs to remember that sequences resembling short motifs can appear simply by chance. Therefore, finding a motif in your protein does not necessarily mean that your protein HAS a corresponding function, but only that it MAY have that function.

The PROSITE database which is a part of ExPASy site contains the information of all the known motifs in proteins. The ScanProsite server helps to look for the known domains in your protein.

PROSITE operates with two types of motifs:

- **Patterns** – specific, relatively short sequence motifs.
- **Profiles** – takes into account the entire protein to predict the motifs.
ScanProsite: finding known patterns and predicting posttranslational modifications

Point your browser to: www.expasy.org/tools/
ScanProsite

Accession number or sequence

uncheck

check

click

START THE SCAN reset
ScanProsite

ScanProsite Results Viewer

This view shows ScanProsite results together with ProRule-based predicted intra-domain features (help). show hits of frequently occurring signatures; do not scan for profiles

Hits for all PROSITE (release 20.56) motifs on sequence P12259 [UniProtKB/Swiss-Prot (release 57.10 of 03-Nov-08: 512205 ent)]

found: 121 hits in 1 sequence

P12259 FAS_HUMAN (2004 sa) ReName: Full=Coagulation factor V; AllName: Full=Activated protein C cofactor; AllName: Full=Proaccelerin, labile factor; Contains: ReName: Full=Coagulation factor V heavy chain; Contains: ReName: Full=Coagulation factor V light chain; Flags: Precursor; Homo sapiens (Human)

the sequence of your protein

pointing cursor on the pattern highlights it in the sequence

pattern documentation

pattern name

pattern location

pattern 3D structure
ScanProsite: 3D of the pattern

Click pattern 3D structure

Your cluster in a protein with the known 3D structure
ScanProsite: modification sites

If you scroll down the ScanProsite Results page, you come to the section called ‘hits by patterns with a high probability of occurrence’. It lists very short patterns which can be associated with posttranslational modifications.

CAPITAL LETTERS: match with the database pattern
low case letters: no match

Be careful with the short patterns: they can be misleading. Myristate is a fatty acid attached to the N-terminus. None of these hits is at the N-terminus. So all these matches are false.
Finding protein domains

Protein domain is an independent folding unit. It is a portion of protein that keeps its shape if removed from the protein.

Domains are portable: new proteins are often formed by combining domains of other proteins.

Domains can be associated with a specific function: RNA-binding domains, DNA-binding domains, ATPase domains, metal-binding domains, etc.

Understanding which domains the protein is built of helps to predict a possible function of an unknown protein.
Finding protein domains

There are several different databases that contain information about protein domains. All are slightly different. Some have been built manually – more accurate but are likely to be incomplete. Some databases have been generated automatically – most likely all-inclusive, but have a lot of noise (junk information – falsely predicted domains).

<table>
<thead>
<tr>
<th>Name</th>
<th>Web address</th>
<th>Number of domains</th>
<th>Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PfamA (IP)</td>
<td><a href="http://www.sanger.ac.uk/Software/Pfam">www.sanger.ac.uk/Software/Pfam</a></td>
<td>7973</td>
<td>Manual</td>
</tr>
<tr>
<td>PRINTs (IP)</td>
<td><a href="http://www.bioinf.ma.ac.uk/dbbrosers/PRINTS">www.bioinf.ma.ac.uk/dbbrosers/PRINTS</a></td>
<td>1900</td>
<td>Manual</td>
</tr>
<tr>
<td>PRODOM (IP)</td>
<td>protein.toulouse.inra.fr/prodom/current/html/home/php</td>
<td>736000</td>
<td>Automatic</td>
</tr>
<tr>
<td>SMART (IP)</td>
<td>smart.embl-heidelberg.de</td>
<td>685</td>
<td>Manual</td>
</tr>
<tr>
<td>TIGRFAM (IP)</td>
<td><a href="http://www.tigr.org/TIGRFAMs">www.tigr.org/TIGRFAMs</a></td>
<td>2453</td>
<td>Manual</td>
</tr>
<tr>
<td>BLOCKs</td>
<td>blocks.fhcrc.org/</td>
<td>12542</td>
<td>Automatic</td>
</tr>
</tbody>
</table>

Some algorithms can search many different databases. For example, InterProScan (IP) searches many different domain collections (but not all).
Finding protein domains

InterProScan server looks for the previously known domains in your new protein. It compares your sequence with InterPro, a domain database that incorporates most of the domain collections. Point your browser to: www.ebi.ac.uk/InterProScan/

Select databases to search (deselect the largest collections, like ProDom),

Paste your sequence

Submit the search request
Finding protein domains

Type of diagnostic (family or domain)

Hyperlink to domain documentation (Summarizes info from different databases: this is where you find a lot of info about functions of this domain)

Link to domain entry in the corresponding database (Pfam, in this case)

Location of the domain in the protein sequence
Finding a Radical SAM domain in your protein suggests that the protein uses the radical chemistry to catalyze the reaction (of methyl transfer).

This means that the protein will require S-adenosyl methionine (SAM) for its activity; that it most likely uses [4Fe-4S] cluster and that the purification and activity studies of the protein need to be carried out in anaerobic conditions.
Finding protein domains with CD server

The CD (Conserved Domains) server of the NCBI searches for domains in proteins and assigns a score which helps to discriminate between ‘good’ and spurious matches.

Point your browser to: www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
Finding protein domains with CD server

Domains found in your protein.
Red bars – from SMART database

Hovering over the feature opens an info window

Clicking on the [+] opens the alignment with the consensus domain

The score (the E-value)
Finding protein domains with CD server

Hovering over the feature opens an info window.
Finding protein domains with CD server

Clicking on the [+] opens the alignment with the consensus domain.

Clicking on the domain ID opens an info page which among other things shows the structure of the domain in a similar protein.
Finding protein domains with CD server
Predicting protein secondary structure

The main secondary structure elements of the proteins are:

- α-helices
- β-sheets (or β-strands)
- random coils

Predicting the secondary structure elements of the protein is the first step towards understanding its structure and eventually, function. Knowing the secondary structure of a protein helps to predict its folding, exposed residues, antigenic regions, plan subcloning experiments, etc.

Modern prediction algorithms use the Hidden Markov Models to predict secondary structure elements in proteins.

PSIPRED is one of the most accurate servers for predicting protein secondary structure.
Predicting protein secondary structure

Point your browser to: bioinf.cs.ucl.ac.uk/psipred/

1. Paste the protein sequence
2. Put your e-mail (has to be an ‘institutional’ address, .edu, not .gmail, .yahoo or the likes)
3. Name your sequence
4. Click the “Predict” button
Predicting protein secondary structure

Within few minutes you will receive an e-mail with the information:

'H' is for α-helical
'E' is for 'extended' (β-strands)
'C-' is for 'random coil'

Confidence level (9 – high, 0 – poor)

At the bottom of the e-mail is a linked to the graphic file (see the next slide)
Predicting protein secondary structure
Predicting your protein

The PredictProtein is one of the most comprehensive sites for predicting various features in the protein besides its secondary structure. The URL address is www.sdsc.edu/predictprotein/

The server is very busy and you might need to wait long time for the results.

The default information you obtain includes:

✔ A secondary structure (H – helical, E – extended, C – random coil)

✔ A prediction of solvent accessibility of different residues

✔ A prediction of transmembrane helices

✔ A prediction of globular regions in the protein

✔ A description of PROSITE motifs

✔ A description of putative domains

✔ Multiple alignment of your sequence with the homologous sequences in the database

✔ A prediction of disulphide bonds
3D structure

All the information about 3D structures of proteins and other biological molecules and complexes are stored in one World-wide database:
Protein Data Bank (PDB)

The information about 3D structure of a molecule consists of coordinates of each (or most) atoms which defines a precise position of an atom in an imaginary space. The positions of the atoms relative to each other define the structure of the molecule.

When information about the 3D (usually crystallographic) structure is deposited in the PDB database, each structure is assigned an accession number (1YI2, 2QAL, etc.).

In order to obtain and view the structure of a protein, you need to know its pdb accession number. It can be either found in the paper describing the structure, or retrieved by searching the PDB database.
3D structure

Point your browser to: www.rcsb.org/pdb/

put either a search term (for example, a protein name) or a pdb number
3D structure

In the list of suggested structures, find the one you want to explore. Then click on its PDB accession number.
3D structure

If you want just to see an image of the protein, simply click on the image to see its enlarge version.

Click on the viewer option if you want to explore the 3D structure in a separate viewer.

Click Jmol if you want to explore the structure in the browser viewer.
can rotate and explore in the browser window
Click the ‘Download Files’ button to get the sequence (FASTA format) or the PDB file with atomic coordinates. You can use the pdb file to view and explore the protein structure in the third-party structure-rendering programs.
Predicting the 3D structure of a new protein

If you found (identified, described) a new protein but have not crystallized it, you want to be able to predict its 3D structure from its sequence alone. A simple way to address this question is to look for a homologous protein with the known structure in the PDB database.

Let us assume that you have identified and sequenced the gene of the TolB protein from *Ricketssia canorii* and you want to know how the protein folds.

(We will cheat and simply retrieve the tolB sequence of *Ricketssia canorii* from the database: a) log into the NCBI site (www.ncbi.nlm.nih.gov); b) choose ‘Protein’ from the dropdown menu; c) in the search window, type identifier NP_360043; d) when the protein sequence is retrieved, change Display format to FASTA; e) select the sequence and copy it to the clipboard).
Predicting the 3D structure of a new protein

Run protein BLAST to find in the pPDB database structures of the proteins homologous to yours.

*Rickettsia canorii* TolB sequence (NP_360043)

Select PDB database

Blast
Predicting the 3D structure of a new protein

several good hits with high similarity to our sequence. Bars show that homology covers the entire sequence.

Go to the PDB database and retrieve the structures (preferably, the one with the highest resolution)
The structure of your protein should be very similar to this one.
You can further model the structure of your protein using energy minimization algorithms.

One of such servers is SWISS_MODEL:
swissmodel.expasy.org
Predicting the 3D structure of a new protein