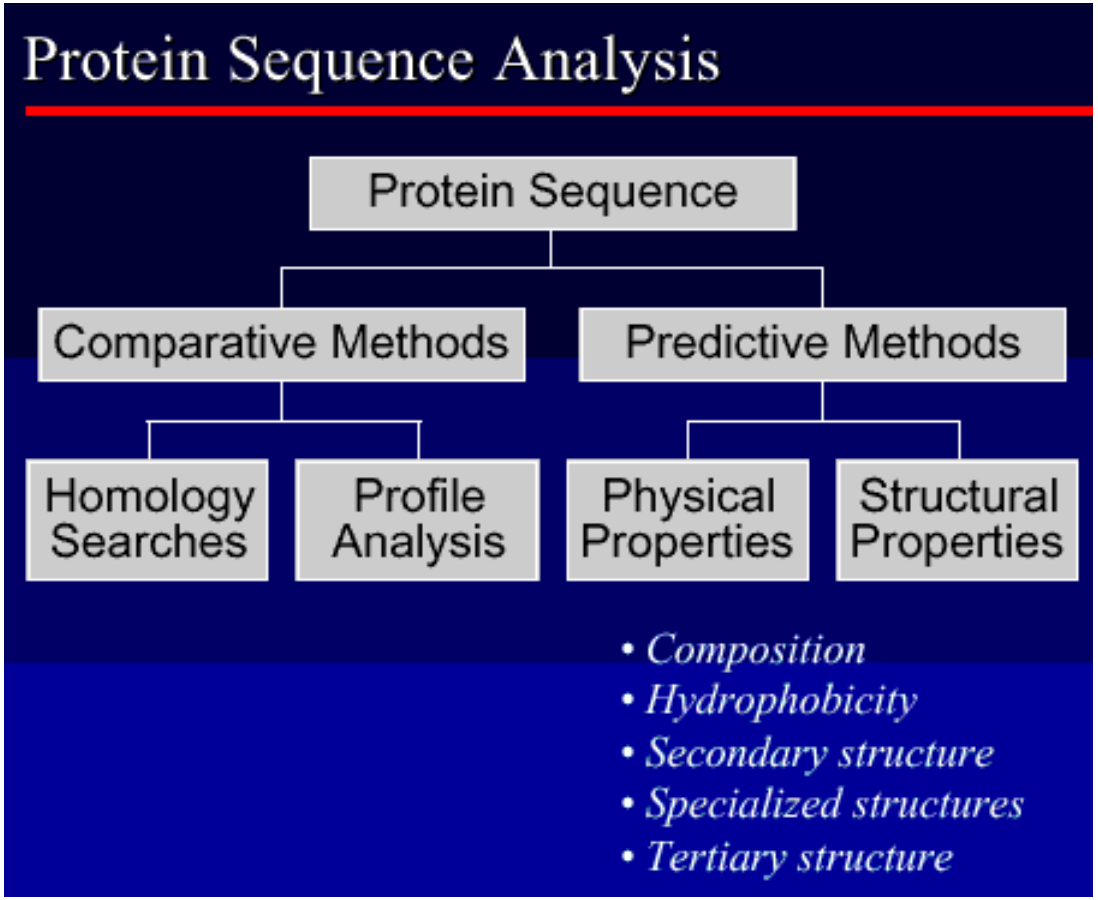
Working with a single protein sequence



Doing biochemistry on a computer

ExPASy: www.expasy.ch

Swiss EMBnet: www.ch.embnet.org

Physico-chemical properties of a protein

Primary structure analysis: www.expasy.ch/tools/#primary

Click the ProtParam link

Primary structure analysis

- ProtParam - Physico-chemical parameters of a protein sequence (amino-acid and atomic compositions, isoelectric point, extinction coefficient, etc.)
- Compute pI/Mw - Compute the theoretical isoelectric point (*pI*) and molecular weight (*Mw*) from a UniProt Knowledgebase entry or for a user sequence
- ScanSite pI/Mw - Compute the theoretical *pI* and *Mw*, and multiple phosphorylation states
- MW, pI, Titration curve - Computes *pI*, composition and allows to see a titration curve

- Radar - De novo repeat detection in protein sequences
- REP - Searches a protein sequence for repeats
- REPRO - De novo repeat detection in protein sequences
- TRUST - De novo repeat detection 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)
- Multicoil - 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 supertypic MHC binders
- RANKPEP - Prediction of peptide MHC binding
- SYFPEITHI - Prediction of MHC type I and II peptide binding

- ProtScale - Amino acid scale representation (Hydrophobicity, other conformational parameters, etc.)
- Drawhca - Draw an HCA (Hydrophobic Cluster Analysis) plot of a protein sequence
- Protein Colourer - Tool for coloring your amino acid sequence
- Three To One - Tool to convert a three-letter coded amino acid sequence to single letter code
- 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

P00533: 1-1210
Number of amino acids: 1210
Molecular weight: 134277.4
Theoretical pI: 6.26
Amino acid composition: Ala (A) 72 6.0% Arg (R) 60 5.0% Asn (N) 66 5.5% Asp (D) 61 5.0% Cys (C) 60 5.0% Gln (Q) 49 4.0% Glu (E) 77 6.4% Gly (G) 85 7.0% His (H) 31 2.6% Ile (I) 69 5.7% Leu (L) 111 9.2% Lys (K) 66 5.5% Met (M) 25 2.1% Phe (F) 36 3.0% Pro (P) 75 6.2% Ser (S) 84 6.9% Thr (T) 64 5.3% Trp (W) 13 1.1% Tyr (Y) 36 3.0% Val (V) 70 5.8% Asx (B) 0 0.0% Glx (Z) 0 0.0% Xaa (X) 0 0.0%
Total number of negatively charged residues (Asp + Glu): 138
Total number of positively charged residues (Arg + Lys): 126
Atomic composition: Carbon C 5875 Hydrogen H 9284 Nitrogen N 1646 Oxygen O 1786 Sulfur S 85
Formula: $C_{5875}H_{9284}N_{1646}O_{1786}S_{85}$
Total number of atoms: 18676
Extinction coefficients: Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm. Ext. coefficient 128890 Abs 0.1% (=1 g/l) 0.960, assuming ALL Cys residues appear as half cystines Ext. coefficient 125140 Abs 0.1% (=1 g/l) 0.932, assuming NO Cys residues appear as half cystines
Estimated half-life: The N-terminal of the sequence considered is M (Met). The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro). >20 hours (yeast, in vivo). >10 hours (Escherichia coli, in vivo).
Instability index: The instability index (II) is computed to be 44.59 This classifies the protein as unstable.
Aliphatic index: 80.74
Grand average of hydropathicity (GRAVY): -0.316

Digesting a protein in a computer

http://www.expasy.org/tools/peptidecutter/

Separate the domains in your protein
Identify potential post-translational modification by mass spectrometry
Remove a tag protein when you express a fusion protein
Make sure that the protein you are cloning isn’t sensitive to some endogenous proteases

Primary structure analysis

1. Hydrophobic regions: membrane spanning

   http://www.expasy.ch/tools/protscale.html

   Hphob. / Kyte & Doolittle: the recommended threshold value is 1.6

   Compare with another scale

   TMHMM: http://www.cbs.dtu.dk/services/TMHMM/

2. Coiled-coil regions: potential protein-protein interaction

   http://www.ch.embnet.org/software/COILS_form.html

3. Hydrophilic stretches: looping out at the surface

   FESTfind:

   (http://www.bioinformatrix.com/net/modules.php?name=Web_Links)

Predicting post-translational modifications

PROSITE patterns (functional residues)
Small well-conserved segments

PKA phosphorylation
[RK] – x – [ST]: ex. RGT, KCS, KET

prokaryotic C4 Zn-finger
C-[DES] – x – C – x(3) – I – x(3) – R – x(4) – P – x(4) – C – x(2) – C

Scan prosite

http://www.expasy.ch/tools/scanprosite/
read PDOC
be careful with species information
how to remove false positives
how to find genuine negatives
everything is not in PROCITE

Finding domains

Domains: Independent globular folding units

A portion of protein that can be active on its own

Results from an alignment between the profile (domain) and your sequence

The main domain collections

PROCITE-Profile	www.expasy.ch/procite
PfamA	www.sanger.ac.uk/Software/Pfam
PfamB	www.sanger.ac.uk/Software/Pfam
PRINTSs	www.bioinf.man.ac.uk/dbbrosers/PRINTS
PRODOM	prodes.roulouse.inra.fr/prodom/doc
SMART	smart.embl~heidelberg.de
TIGRFAM	www.tigr.org/TIGRFAMs
BLOCKs	www.blocks.fhcrc.org

Finding domains with InterProScan

http://www.ebi.ac.uk/InterProScan/

integration of many databases

Finding domains with the CD server

CD: conserved domains server at NCBI

come along with a score

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

deselect Low Complexity check box

sequences that contain repeated residues are low-complexity sequences

an amino acid is over represented in many interesting domains

ex. Leucine zippers, glycine-rich domains

ragged ends indicate partial matches: mostly insignificant

different colors from different domains

E-values need to be below 0.01 to mean something

Assignments

Search hydrophobic regions in the human protein

Perform the following analysis with the most and the least bacterial homologous proteins

Secondary structure prediction (Jpred)

Domain pattern (InterProScan)