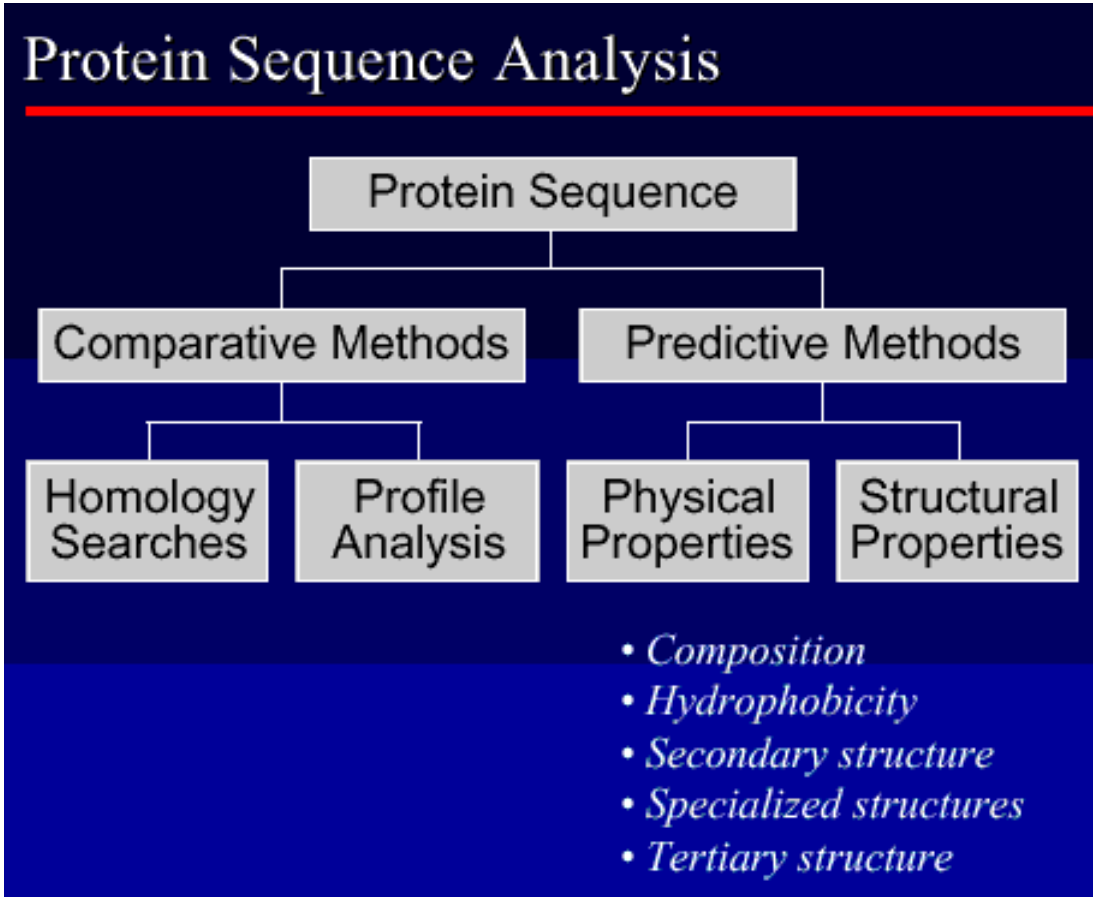
Working with a single protein sequence



Doing biochemistry on a computer

ExPASy: www.expasy.ch

Swiss EMBnet: www.ch.embnet.org

A. 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

B. 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

C. Predicting post-translational modifications

http://expasy.org/tools/

NetPhos: prediction of Ser, Thr and Tyr phosphorylation
NetAcet: prediction of N-acetyltransferase A substrates
SignalP: prediction of signal peptide cleavage sites
MITOPROT: prediction of mitochondrial trageting sequences
NetNGlyc: prediction of N-glycosylation sites in human proteins

D. Topology prediction

PSORT: prediction of subcellular localization
TMpred: prediction of transmembrane regions

E. 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

F. Secondary structure prediction

http://expasy.org/tools/
http://www.biogem.org/tool/chou-fasman/
http://www.compbio.dundee.ac.uk/www-jpred/

Assignments

Perform A~F predictions with the human & a bacterial protein