

Sequence Feature Prediction

Diana Uskat, Verena Link

May 24, 2011

Outline

Introduction

Secondary Structure Prediction

Disordered Regions

Transmembrane Helices

Signal Peptides

GO Terms

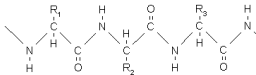
Summary

Sources

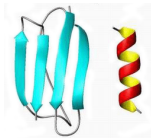
Introduction

Protein Structure

- ▶ Primary structure: amino acid sequence



- ▶ Secondary structure: α -helices, β -sheets and coils



- ▶ Tertiary structure: 3D-structure of a protein



α -helices

- ▶ H-bonds between the NH-group of an amino acid and the CO-group of the amino acid four residues earlier ($i+4$)
- ▶ Other rare forms are 3_{10} ($i+3$) and π -helices ($i+5$)

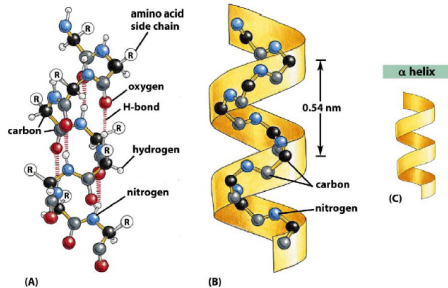


Figure 1: Alberts B, Johnson A, Lewis J. et al - Molecular Biology of the Cell. 4th edition., Garland Science (2002)

β -sheet

- ▶ H-bonds between the NH- and the CO-group may be located far apart in sequence
- ▶ Parallel and anti-parallel

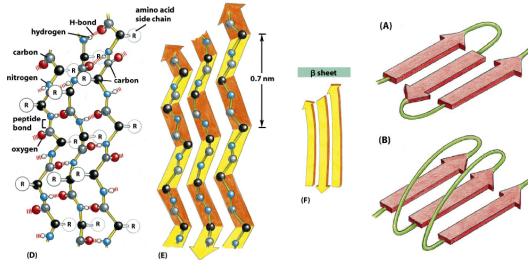


Figure 2: Alberts B, Johnson A, Lewis J. et al - Molecular Biology of the Cell. 4th edition., Garland Science (2002)

Secondary Structure Prediction

Secondary Structure Prediction

- ▶ Attempt to predict α -helices, β -sheets and coils based only on the primary structure informations
- ▶ Comparative modelling most reliable technique
- ▶ Gain: provide constraints for tertiary structure prediction
- ▶ Possible Method: *PSIPRED3.0*
- ▶ Determination of prediction success by comparing with DSSP

PSIPRED 3.0

- ▶ Uses neural networks with a single hidden layer and a feed-forward back-propagation architecture
- ▶ Split into three states:
 1. Generation of sequence profiles
 - ▶ Position-specific scoring matrix from PSI-BLAST as input for the neural network.
 2. Prediction of initial secondary structure
 - ▶ output layer where the units represent the three states of secondary structure (helix, strand or coil)
 3. Filtering of the predicted structure
 - ▶ Successive filtering of the outputs from the main network.

PSIPRED 3.0

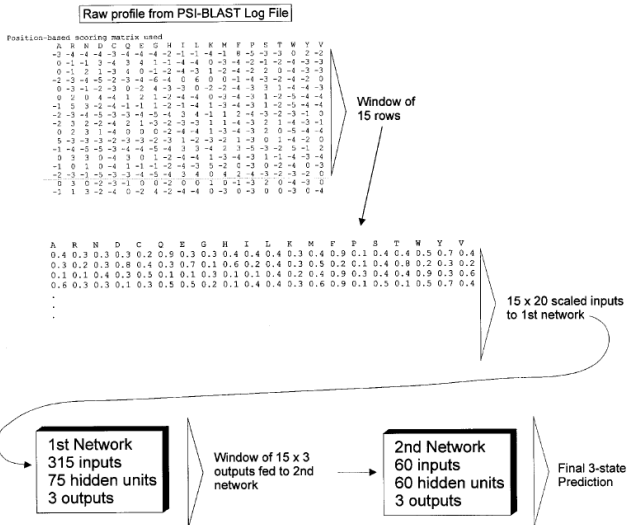


Figure 3: Jones - Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol.

(1999)

- ▶ Database of secondary structure assignments for all PDB-entries, NO PREDICTION!
- ▶ Defines the secondary structure by given atomic coordinates in PDB-format
- ▶ Main idea: algorithm based mainly on H-bonding:
- ▶ 'n-turns': H-bond between the CO of residue i and the NH of residue $i+n$ where $n = 3, 4, 5$
=> α -Helices: repeating 4-turns
- ▶ 'bridges': H-bond between residues not located to each other in sequence
=> β -Sheet: repeating bridges

Disordered Regions

Amino acid properties

- ▶ hydrophobic amino acids: Ala, Ile, Leu, Met, Phe, Pro, Trp, Tyr, Val
- ▶ hydrophilic amino acids: Asn, Cys, Gln, Gly, Ser, Thr
- ▶ neutral/positively charged amino acids: Lys, Arg, His
- ▶ neutral/negatively charged amino acids: Cys, Tyr, Asp, Glu

Disordered Regions

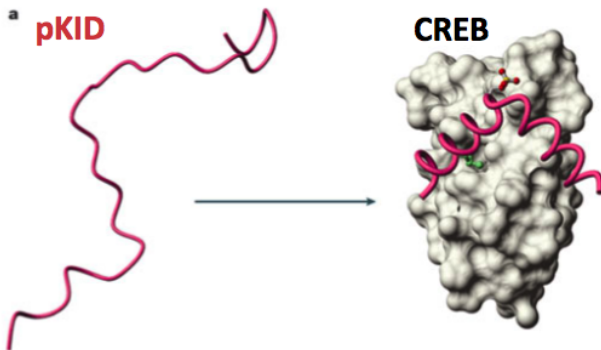


Figure 4: Dyson H & Wright P - Intrinsically unstructured proteins and their functions, Nat Rev Mol Cell Biol (2005)

Disordered Regions

- ▶ Long regions without regular secondary structure
- ▶ Dynamically flexible (distinct from loops)
- ▶ Adopt regular structure only upon binding to substrates or other proteins
- ▶ Conserved
- ▶ Over-represented in regulatory functions
- ▶ Functionally very important

Disordered Regions - Prediction

- ▶ Invisible in electron density maps or unfolded by CD measurement
- ▶ Usually: depletion of hydrophobic and bulky amino acids
- ▶ Large solvent accessibility => prevalence of polar and charged amino acids
- ▶ High percentage of proline

Methods - MD (Meta-Disorder)

- ▶ Combines different methods (NORSnet, PROFbval, Ucon and DISOPRED2)
- ▶ Use sequence profiles
- ▶ Use other useful features (solvent accessibility, secondary structure, low complexity regions)

Methods - DISOPRED

- ▶ Knowledge-based method
- ▶ Based on a neuronal network
- ▶ Residues appear in sequence records
- ▶ Coordinates are missing from the electronic density map

Methods - NORSp

- ▶ NORs: segments of > 70 consecutive residues with $< 12\%$ of these residues in helix, strand and coiled-coil regions
- ▶ At least: 10 adjacent residues exposed to solvent
- ▶ Prediction by merging predictions of secondary structure, transmembrane helices and coiled-coil regions

Transmembrane Helices

Transmembrane Helices

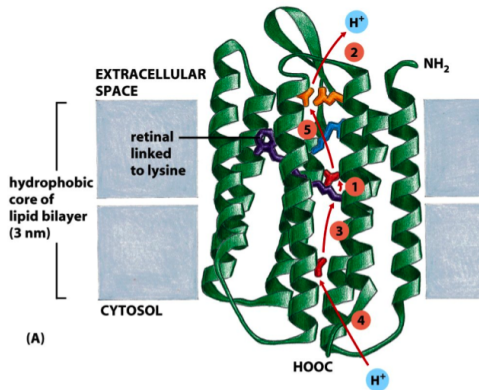


Figure 5: Alberts B, Johnson A, Lewis J. et al - Molecular Biology of the Cell. 4th edition., Garland Science (2002)

- Main locations: core, cap, loop

Transmembrane Helices

- ▶ Flanking caps: different amino acid composition for cytosolic & non-cytosolic side
- ▶ Many polar and charged residues -> contact to phosphate groups of the lipids
- ▶ Helix: surrounding a core region with 5 – 25 amino acids length
- ▶ On different sides of the membrane: different amino acids distribution
- ▶ Loop: cytosolic side: positively charged => non-cytosolic side: negatively charged (prevalence!)

Methods - TMHMM

- ▶ TransMembrane Hidden Markov Model
 - ▶ 3 main locations
 - ▶ 7 different states

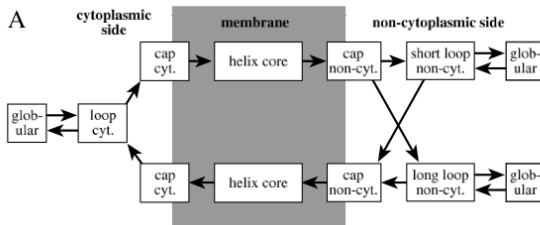


Figure 6: Sonnhammer, Heijne, Krogh - A hidden Markov model for predicting transmembrane helices in protein sequences, Proc Int Conf Intell Syst Mol Biol.(1998)

Signal Peptides

Signal Peptides

- ▶ Proteins are synthesized in the cytosol => transport to target component
- ▶ Mostly: N-terminal targeting sequences (proteolytically removed during or after the entry)

Signal Peptides

- ▶ Final destination:
 - ▶ Mitochondria:
 - ▶ Arg, Ala, Ser: over-represented
 - ▶ Asp, Glu: rare
 - ▶ Chloroplast:
 - ▶ Low content of acidic residues
 - ▶ Over-representation of hydroxylated residues

Signal Peptides

- ▶ Secretory pathway
 - ▶ 3 regions:
 - ▶ Positively charged n-region
 - ▶ Hydrophobic h-region
 - ▶ Prokaryotes: **Leu** & **Ala** in equal amounts
 - ▶ Eukaryotes: dominated by **Leu**
 - ▶ Polar c-region

Methods - SignalP

- ▶ Recognition of cleavage site
- ▶ Classification of amino acids as belonging to the signal peptide or not
- ▶ Challenge by prediction: Signal anchors often have similarities with signal cleavage sites after their transmembrane region

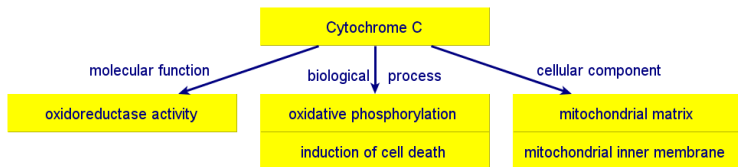
Combined Method - Phobius

- ▶ HMM for transmembrane protein and signal peptides
- ▶ Combination of TMHMM and SignalP
 - ▶ Three different start states
 - ▶ Globular as TMHMM
 - ▶ N-region for Signal peptide prediction

GO Terms

GO Terms

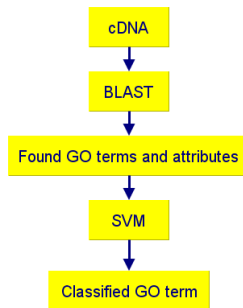
- ▶ Consistent nomenclature of gene products in different databases according to their biological content
- ▶ Covers three domains: cellular components, molecular function and biological process



Prediction Of Function And GO Terms

- ▶ Homology-based Methods:
 - ▶ Gopet
 - ▶ MultiPfam2GO
- ▶ Ab initio methods:
 - ▶ LOCnet
 - ▶ ProtFun 2.2 Server

- ▶ Homology-based method
- ▶ Gain: Assignment of uncharacterised cDNA sequences to GO molecular function term
- ▶ Uses Blast against GO-mapped protein databases
- ▶ Uses SVM for the classification



MultiPfam2GO

- ▶ Gain: Assignment of multidomain combinations to GO terms
- ▶ Uses a Naïve Bayesian network to classify all subsets of the multidomain.
- ▶ Assignment of a GO term only if all subsets of the multiple-domain-set classify one GO term with a P-Value < 0.001
- ▶ Example:



- ▶ Ab-initio method for the prediction of sub-cellular localization (GO cellular components term)
- ▶ Only for eukaryotic and prokaryotic proteins.
- ▶ Uses neural networks which consist of three layers
- ▶ trained on global features
 - ▶ amino acid composition
 - ▶ evolutionary information from sequence profiles
 - ▶ predicted secondary structure composition
 - ▶ composition of predicted surface accessible residues
- ▶ Sorts proteins into one of four classes:
 - Extracellular
 - Nuclear
 - Cytoplasmic
 - Mitochondrial

ProtFun 2.2 Server

- ▶ Ab-initio sequence-based method
- ▶ Gain: Assignment of orphan proteins to functional classes (GO terms)
- ▶ Integrates relevant features which are more directly related to the linear sequence of amino acids
- ▶ -> queries a large number of other feature prediction servers (PSIPred, TMHMM,)
- ▶ Uses an ensemble of five different neural networks (three layer feed-forward)

Summary

Summary

- ▶ Following features can be predicted:
 - ▶ Secondary Structure & Disordered Regions
 - ▶ Transmembrane Helices
 - ▶ Signal Peptides & Protein location
 - ▶ GO Terms and function
- ▶ mostly less data available (only small trainings set, risk for over-fitting)
- ▶ the predictions can only give a general idea
- ▶ further research is necessary to improve the methods

Sources

- ▶ Alberts B, Johnson A, Lewis J. et al - Molecular Biology of the Cell. 4th edition., Garland Science (2002)
- ▶ Jones - Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. (1999)
- ▶ Dyson H & Wright P - Intrinsically unstructured proteins and their functions, Nat Rev Mol Cell Biol (2005)
- ▶ Sonnhammer, Heijne, Krogh - A hidden Markov model for predicting transmembrane helices in protein sequences, Proc Int Conf Intell Syst Mol Biol.(1998)
- ▶ Emanuelsson et al. - Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence, J. Mol. Biol (2000)
- ▶ Nielsen et al. - Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, Protein Engineering (1997)
- ▶ Ward et al. - The DISOPRED server for the prediction of protein disorder, Bioinformatics Applications note (2004)
- ▶ Käll et al. - An HMM posterior decoder for sequence feature prediction that includes homology information, Bioinformatics (2005)
- ▶ Käll et al. - Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server, Nucl. A. Res. (2007)
- ▶ Kahsay et al. - An Improved Hidden Markov Model for Transmembrane Protein Detection and Topology Prediction and Its Applications to Complete Genomes, Bioinformatics (2005)
- ▶ Liu, Rost - NORSp: predictions of long regions without regular secondary structure, Nucl. A. Res. (2003)
- ▶ Vinayagam et al. - GOPET: a tool for automated predictions of Gene Ontology terms., BMC Bioinformatics (2006)
- ▶ Schlessinger et al. - Improved Disorder Prediction by Combination of Orthogonal Approaches, PLOS (2009)
- ▶ Forslund, Sonnhammer - Predicting protein function from domain content, Bioinformatics (2008)
- ▶ Jensen et al. - Ab initio prediction of human orphan protein function from post-translational modifications and localization features., J. Mol. Biol. (2002)
- ▶ from <http://www.geneontology.org/GO.doc.shtml>