8 Secondary Structure prediction

Introduction

Prediction of secondary structure is an important step towards the prediction of the whole 3D structure. It determines up to some point the global fold.

The driving forces can be assumed to be found in the local characteristics of the polypeptide chain.

3 main target classes:

- α-helix (H) : HB i-i+4
- β-sheet (E) :
- random coil (C)

All three classes have a similar frequency.

Other classifications:

- Turns (T)
- $3_{10} - helix(G, HBi - i + 3)$
- $\pi - helix(I, HBi - i + 5)$
- parallel/antiparallel sheets
- Bend (S)

Helix dipole moment

Alpha helix has an overall dipole caused by the dipoles of the carbonyl groups found in the peptide bond, all pointing along the helix axis, resulting in a positive dipole towards the N-terminus.

This overall dipole can destabilize the helix. That's why alpha helices are often capped by a N-terminal positively charged aminoacid.

This dipole is also of importance because the N-terminal positive charge can be often used to bind negative charged ligans, such as phosphates.

Amino Acid propensities

Aminoacids are observed at different frequencies in the different secondary structural element types.

These propensities can be used to predict secondary structure.

$$P_{i,s} = \frac{c_{i,s} / \sum_j c_{j,s}}{c_i / \sum_j c_j}$$

c=count
i=amino acid type
s=structural state (e.g. helix)

1=frequency as in reference
>1 increased
<1 decreased

Log(odds)

- $\log(odds) = \log\left(\frac{c_{i,S} / c_{!i,S}}{c_{i,!S} / c_{!i,!S}}\right)$ (! = "not")

Aminoacid classifications:

	Helix	Strand
Strong former	E A L	M V I
Former	H M Q W V F	C Y F Q L T W
Weak former	K I	A
Indifferent	D T S R C	R G D
Breaker	N Y	K S H N P
Strong breaker	P G	E

Different methods can make use of already determined propensities:

Chou-Fasman method

- Uses table of propensities derived from CD spectroscopy data of soluble, globular proteins.
- Likelihood for each aminoacid

Pseudocode:

```
Chou_fasman(sequence):
    assign all residues parameters
    for the whole sequence: #determine alpha helix
```

```
    indentify region where 4/6 have P(H)>100
        while(set of four has mean(P(H)) > 100):
            Extend alpha helix
for the whole sequence: #determine beta sheet
    indentify region where 3/5 have P(E)>100
        while(set of four has mean(P(E)) > 100):
            Extend beta sheet
    if average(PE of betasheet) > 105 and P(E) > P(H)::
        mark region as beta sheet
    else:
        discard
for the whole sequence: # determine turn
    p(t) = f(j)f(j+1)f(j+2)f(j+3) # Likelihood
    if p(t) > 0.000075 and average P(turn) > 100 in tetrapept
    and P(turn) > P(H) and P(E):
    tetrapet is a turn
```

Name	P(H)	P(E)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.07	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.11	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.05	0.117	0.128
Glutamic Acid	151	37	74	0.056	0.06	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.19	0.152
Histidine	100	87	95	0.14	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.07
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.12	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

Problem: it doesn't take into account the structure of the neighbors

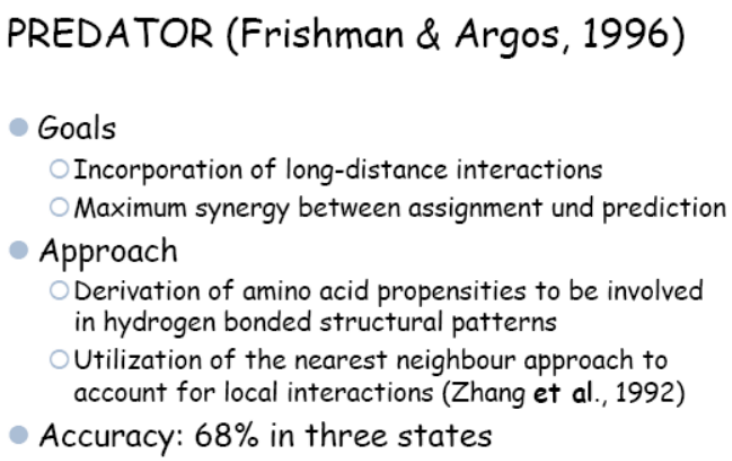
The gor method

- Built on Chou-Fasman values.
- One matrix for each feature
- Evaluate each residue plus 8 in each direction (sliding window of 17)

- Underpredicts beta strand

Supervised machine learning methods

- Train your algorithm on training dataset and evaluate on test dataset.
- k-nearest neighbor methods: Define a starting point as centroid, enclose close elements until k training examples are selected and label them by majority vote.
  - Application to proteins:
    - make a table of sequence windows from proteins with known structure
    - find 50 best aligments with this table
    - score frequencies of different structures in the middle position
    - Scan sequence for series of high scoring predictions
- Neural Networks can (and are often used) for protein predictions. The general idea of neural networks is to reproduce the structure of neural tissue: Dendrites receive the inputs and the neuron body integrates it in a single output (normally with help of an output sigmoidal function). The algorithm learns the weights of the different inputs, working as a linear transformation ans passes it to the next layer/output.



Secondary structure prediction from multiple sequences

- Prerequisite: exponential growth of the protein sequence databank
- The majority of sequences have at least one homolog
- Standard approach
  - Database search to find related sequences
  - Multiple alignment
  - Extraction of sequence variation patterns
- 6-7% improvement of the prediction accuracy
- Main drawback: total reliance on the alignment quality

PSI-PRED (Jones, 1999)

- Position-specific scoring matrix by PSI-BLAST to query sequence
- Screen the sequence by overlapping windows of 15 aa
- Submit the input of a 15 aa window to a neural net
- Train the neural net parameters on a training set
- Test on a test set

Physical approach towards helix prediction(AGADIR)

Based on Helix-coil transition theory, general for polymers but often used for proteins. Tries to capture the difference in energy between a coil random structure and an α-helix.

$\Delta G_{helical-segment} = \Delta G_{Int} + \Delta G_{Hbond} + \Delta G_{SD} + \Delta G_{nonH} + \Delta G_{dipole}$

where:

ΔG_{int} are the intrinsic tendencies of the residues to adopt helix conformation.

ΔG_{Hbond} are the contributions of main chain and i, i+4 hydrogen bonds

ΔG_{SD} Sums the net contributions with respects to the random coil state of all side chain interactions.

ΔG_{nonH} Captures the contribution to stability of N and C terminal residues.

ΔG_{dipole} represents the iteraction of charged groups with the helix macrodipole

Trans-membrane element prediction

Trans membrane proteins constitute 30% of all proteins in a cell, and receptors are an important target for pharmaceutical industry.

Aminoacids are differentially hydrophobic, and that's often used for prediction of transmembrane elements.

Structurally, they tend to have charged residues flanking hydrophobic segments, and the positively charged extrem tend to face towards the cytoplasm (weaker in Archaea).

Sometimes they also have amphipathic α-helix after the hydrophobic region to interect both with the environment and the cell layer. They tend to have a repetitive structure of charged and hydrophobic residues alternating with a repeat distance corresponding to the period of the structure. This can be seen in a helical wheel plot and hydrophobic moments (a vectorized representation of the hydrophobicity of the sequence).

The starting point are propensity tables for the different aminoacids representing how likely is for a given aminoacid to interact with water.

Both for flanking and hydrophobic region are calculated.

$$P_m^i = \frac{n_{i,seg} / n_{all,seg}}{n_{i,total} / n_{all,total}} \qquad P_e^i = \frac{n_{i,edge} / n_{all,edge}}{n_{i,total} / n_{all,total}}$$

There are different likelihood tables:

- Kyte-Doolittle hydropathy.
- Hopp-Woods hydrophilicity.
- Eisenberg et al. normalized consensus. Basic hydrophilibity plot: Calculate average hydropathy over a window and slide window until the entire sequence has been analyzed

Post-processing

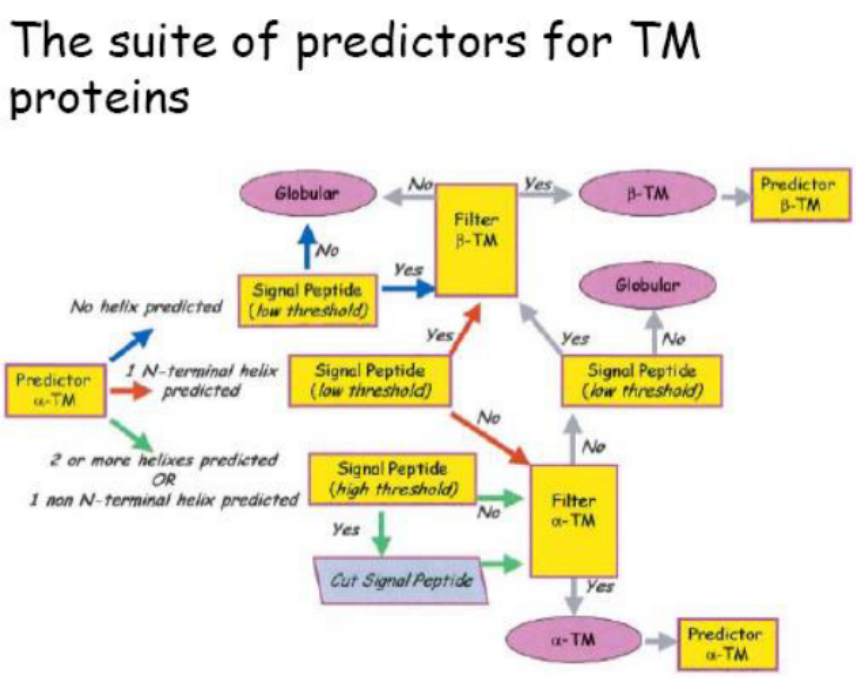
- Eight or more consecutive positions with P(m) > 1.23 are assigned to TM-helices
- Elongated while P(m) > 1.17 and residues <= 21
- Start and end points of TM-segment are set with P(e) > 1.08
- Split sequences if they are sufficiently long to contain multiple TM-helices
- Long helices that can not be spilt are shortened

Persson and Argos - Results

- Tested on 28 families with 126 TM-segments
- Only 5 segments were predicted wrong
- 96% correct

Also markov-chain models are used to solve this problem, and neural networks.

β-barrel element prediction is mostly based on hydropathy analysis and similarity search.



performance assessment

Biggest databases result in more accurate predictions.

- Qindex: Percentage of residues correctly predicted as α-helix,coil... the score is high even for random predictions. $Q_3 = \frac{N_{predicted}}{N_{observed}} \cdot 100$