

Introduction to Bioinformatics

Iosif Vaisman

Email: ivaisman@gmu.edu

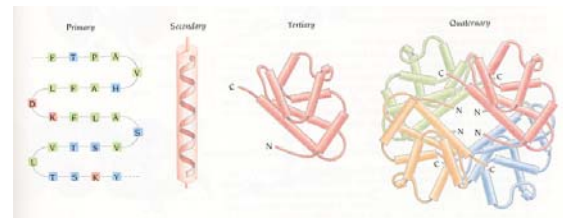
Secondary Structure: Computational Problems

Secondary structure characterization
Secondary structure assignment
Secondary structure prediction
Protein structure classification

Secondary Structure: Computational Problems

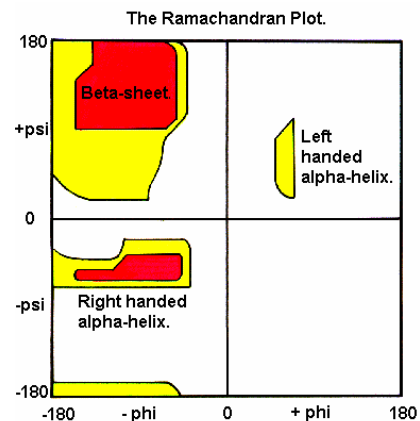
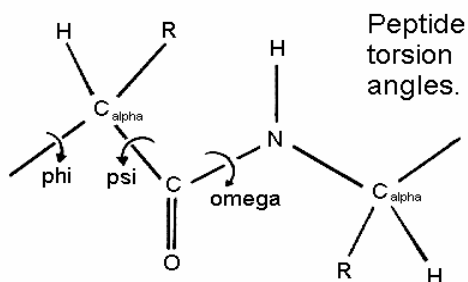
Secondary structure characterization
Secondary structure assignment
Secondary structure prediction
Protein structure classification

Protein Structure Hierarchy



- Primary - the sequence of amino acid residues
- Secondary - ordered regions of primary sequence (helices, beta-sheets, turns)
- Tertiary - the three-dimensional fold of a protein subunit
- Quaternary - the arrangement of subunits in oligomers.

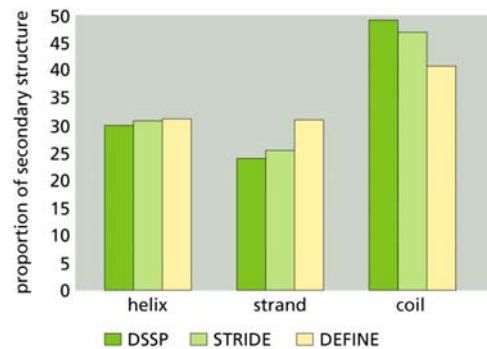
Secondary Structure Conformations



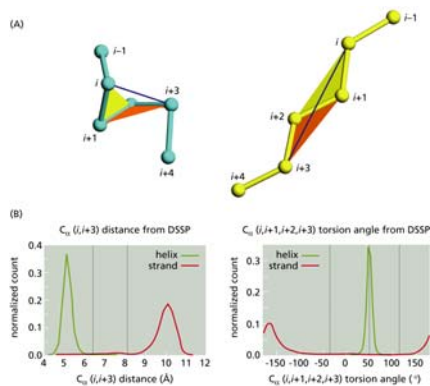
Secondary Structure Conformations

	ϕ	ψ
alpha helix	-57	-47
alpha-L	57	47
3-10 helix	-49	-26
π helix	-57	-80
type II helix	-79	150
β -sheet parallel	-119	113
β -sheet antiparallel	-139	135

Secondary Structure Assignment



Secondary Structure Assignment



Secondary Structure Assignment

```

.....1.....2.....3.....4.....5.....6.....7..
--- sequential resnumber, including chain breaks as extra residues
--- original PDB resname, not nec. sequential, may contain letters
--- amino acid sequence in one letter code
|
|  -- secondary structure summary based on columns 19-38
|  xxxxxxxxxxxxxxxxxxxxxxxx recommend columns for secstruc details
|  -- 3-turns/helix
|  -- 4-turns/helix
|  -- 5-turns/helix
|  -- geometrical bend
|  -- chirality
|  -- beta bridge label
|  -- beta bridge label
|  -- beta bridge partner resnum
|  -- beta bridge partner resnum
|  -- beta sheet label
|  -- solvent accessibility
|
# RESIDUE AA STRUCTURE BP1 BP2 ACC
35 47 I E + 0 0 2
36 48 R E > S- K 0 39C 97
37 49 Q T 3 S+ 0 0 86
38 50 N T 3 S+ 0 0 34
39 51 W E < -KL 36 98C 6
  
```

(example from 1EST)

Secondary Structure: Computational Problems

Secondary structure characterization

Secondary structure assignment

Secondary structure prediction

Protein structure classification

Secondary Structure Prediction

Three-state model: helix, strand, coil

Given a protein sequence:

– NWVLSTAADMQGVVTDGMASGLDKD . . .

Predict a secondary structure sequence:

– LLEEEELLLLHHHHHHHHHHHLLHHLL . . .

Methods:

- statistical
- stereochemical

Accuracy: 50-85%

Statistical Methods

Residue conformational preferences:

Glu, Ala, Leu, Met, Gln, Lys, Arg - helix
Val, Ile, Tyr, Cys, Trp, Phe, Thr - strand
Gly, Asn, Pro, Ser, Asp - turn

Chou-Fasman algorithm:

Identification of helix and sheet "nuclei"
Propagation until termination criteria met

Chou-Fasman Parameters

Name	P(a)	P(b)	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	0.070	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.110	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic Acid	151	37	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	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	0.070
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	0.120	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

Chou-Fasman Algorithm

Identification of helix and sheet "nuclei"

helix - 4 out of 6 residues with high helix propensity ($P > 100$)

sheet - 3 out of 5 residues with high sheet propensity ($P > 100$)

Propagation until termination criteria met

Turn prediction

1) $p(t) > 0.000075$

2) $P(\text{turn}) > 1.00$

3) $P(a) < P(\text{turn}) > P(b)$

where $p(t) = f(j)f(j+1)f(j+2)f(j+3)$

P.Y. Chou, G.D. Fasman, *Biochemistry*, 1974, 13, 211-222

Garnier - Osguthorpe - Robson (GOR) Algorithm

Likelihood of a secondary structure state depends on the neighboring residues:

$$L(S_j) = \sum (S_j; R_{j+m})$$

Window size - $[j-8; j+8]$ residues

Accuracy for a single sequence - 60%

Accuracy for an alignment - 65%

Evolutionary Methods

Taking into account related sequences helps in identification of "structurally important" residues.

Algorithm:

find similar sequences

construct multiple alignment

use alignment profile for secondary structure prediction

Additional information used for prediction

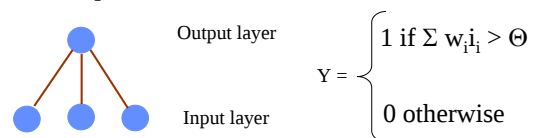
mutation statistics

residue position in sequence

sequence length

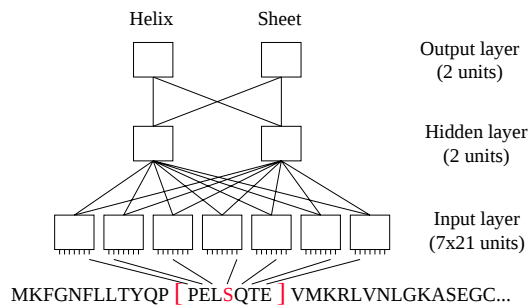
Neural Networks

Perceptron



Learning process: $\Delta w_i = (T_p - Y_p) i_{pi}$

Neural Networks Methods



Stereochemical Methods

Patterns of hydrophobic and hydrophilic residues in secondary structure elements:

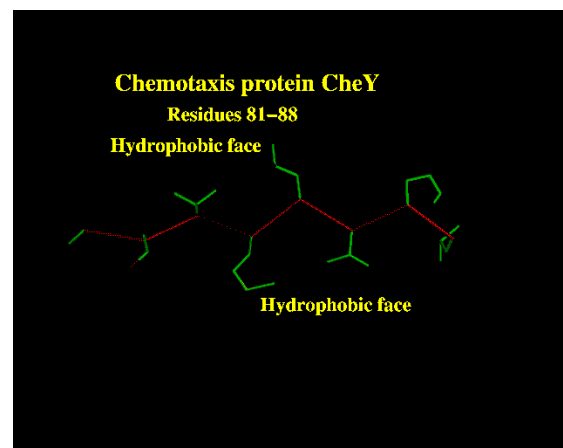
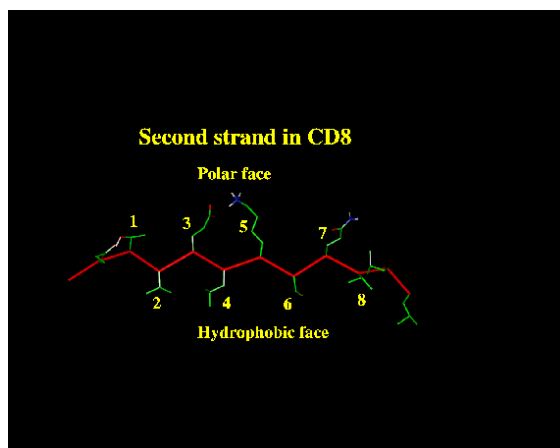
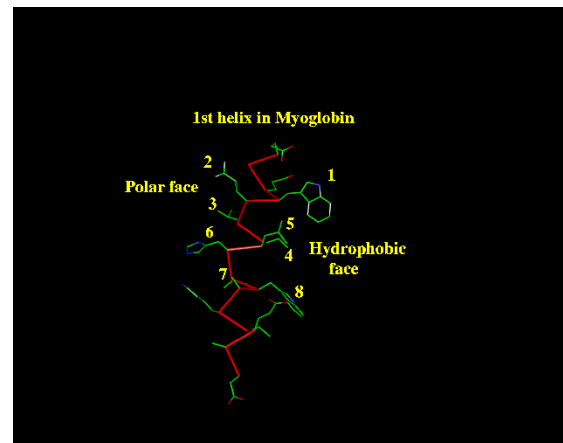
- segregation of hydrophobic and hydrophilic residues
- hydrophobic residues in the positions 1-2-5 and 1-4-5
- oppositely charged polar residues in the positions 1-5 and 1-4 (e.g. Glu (i), Lys (i+4))

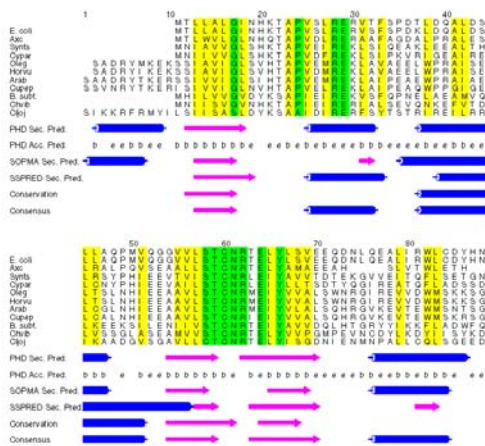
Definitions of hydrophobic and hydrophilic residues (hydrophobicity scales) are ambiguous

Stereochemical Methods

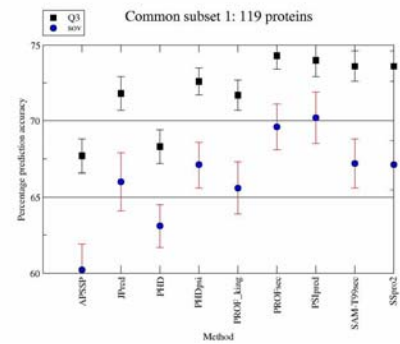
Hydropathic correlations in helices and sheets

		F-F	F-L	L-F	L-L
α	$i, i+2$	-	+	+	-
	$i, i+3$	+	-	-	+
	$i, i+4$	+	-	-	+
	$i, i+5$	-	+	+	-
β	$i, i+1$	-	+	+	-
	$i, i+2$	+	-	-	+
	$i, i+3$	-	+	+	-





Accuracy of prediction



EVA (<http://cubic.bioc.columbia.edu/eva/>)

Accuracy of Prediction

$$Q_3 = \frac{PH + PE + PC}{N}$$

$$W = \log \frac{TP \times TN}{FP \times FN}$$

Range: 50-85%