BINF 731

Protein Structure Analysis

http://binf.gmu.edu/vaisman/binf731/

Iosif Vaisman

2025

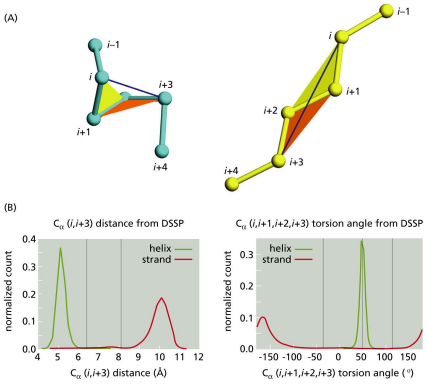
Secondary Structure: Computational Problems

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

Secondary Structure Conformations

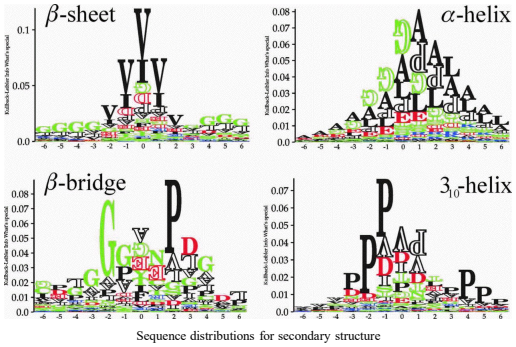
	ϕ	ψ
alpha helix	-57	-47
alpha-L	57	47
3-10 helix	-49	-26
π helix	-57	-80
type II helix	-79	150
β-sheet parallel	-119	113
β-sheet antiparallel	-139	135

Secondary Structure Assignment



Adopted from *Zvelebil, Baum, 2008*

Secondary Structure Assignment

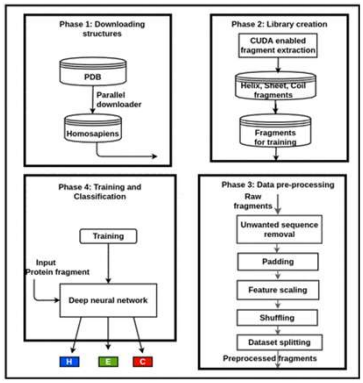


Sequence distributions for secondary structure

The number of aligned segments are: (A) 41803, (B) 4952, (C) 27320, (D) 1851 (707 non-homologous protein chains using DSSP).

C. Andersen & B. Rost, 2003

Secondary Structure Assignment



DLFSA algorithm (CNN based)

J. V. Antony et al., 2021

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
    |    |   |  | xxxxxxxxxxxxxxxxxxxx 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      (example from 1EST)
   38   50   N  T 3  S+     0   0   34
   39   51   W  E <   -KL  36  98C   6
```

Secondary Structure Assignment

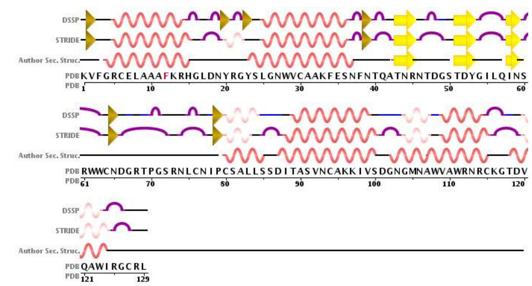
#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC
1	2 A	T		0	0	77
2	3 A	T	E -A	34	0A	21
3	4 A	a	E -A	33	0A	0
4	5 A	b	-	0	0	0
5	6 A	P	S S+	0	0	52
6	7 A	S	S > S-	0	0	48
7	8 A	I	H > S+	0	0	123
8	9 A	V	H > S+	0	0	98
9	10 A	A	H > S+	0	0	6
10	11 A	R	H X S+	0	0	55

B = residue in isolated β-bridge
E = extended strand, in β-ladder
G = 3-helix (3/10 helix)
H = α-helix
I = 5-helix (π helix)
T = hydrogen bonded turn
S = bend

Figure 1. Essentials of the DSSP file. Left: a small part of a secondary structure description. From left to right the columns contain the sequential number of the amino acid, its PDB number, the chain identifier, the amino acid sequence (with paired cysteines replaced by pairs of lower case characters), the actual secondary structure assignment (in red), a description of the type of turn encountered, in case of β-sheets the partner β-strand(s) and the sequential strand identifier, the solvent accessibility of the residue in square Ångströms. The lines further contain information (data not shown) about the geometry of the hydrogen bond(s) that were used to assign the secondary structure, local backbone angles and torsion angles and the coordinates of the Cα. Right: the meaning of the most used (red) column from DSSP files: the secondary structure assignments. Most people convert B, S and T simply into loop (which is a blank in DSSP), sometimes the G is converted into a H, and the I (π helix) is so rare that people tend to just forget about it.

Joosten et al., 2010

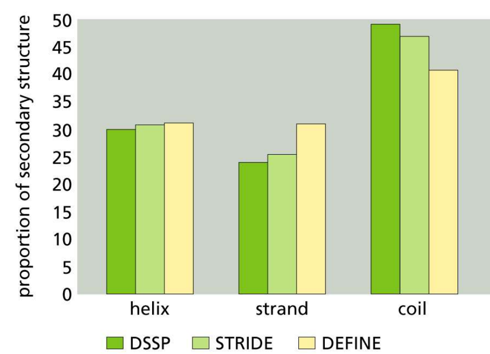
Secondary Structure Assignment



Author Sec. Struc. Legend
E: beta strand
H: alpha helix

Comparison of DSSP, Stride and author's assignments

Secondary Structure Assignment



Adopted from *Zvelebil, Baum, 2008*

Secondary Structure Assignment

a)

DSSP \ DLFSA	Helix	Sheet	Coil
Helix	0.886	0.048	0.065
Sheet	0.034	0.889	0.076
Coil	0.031	0.062	0.898

b)

STRIDE \ DLFSA	Helix	Sheet	Coil
Helix	0.896	0.040	0.064
Sheet	0.041	0.883	0.075
Coil	0.064	0.040	0.888

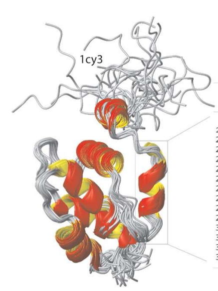
c)

PDB \ DLFSA	Helix	Sheet	Coil
Helix	0.894	0.041	0.065
Sheet	0.033	0.890	0.075
Coil	0.031	0.062	0.898

```
a) 5ZGG.pdb (Seq:1-34, HOMOSAPIENS)
   T R G T   T D   N L I   P V Y   C S I L A A   V V V G L V   A Y I A F K   R W N S   Sequence
   ---------------------------------------------------------------------------------------------
   C C C C | H H | H H H | H H H | H H H H H H | H H H H H H | H H H H H H | C C C C   By DLFSA
   - - C C   C C   - - H   H H H   H H H H H H   H H H H H H   H H H H H H   H H H -   By DSSP
   C C C C   C C   C H H   H H H   H H H H H H   H H H H H H   H H H H H H   H H H -   By STRIDE
   C C C C   C C   C H H   H H H   H H H H H H   H H H H H H   H H H H H H   H H H H   By PDB
   C C C C   C C   C C H   H H H   H H H H H H   H H H H H H   H H H H H H   H H H C   By P-SEA
   C C C C   C H   H H H   H C C   C H H H H H   H H H H H H   H H H H H H   H H C C   By STICK
   C C C C   C C   C C H   H H H   H H H H H H   H H H H H H   H H H H H H   H H H C   By SST
   C H H C   C C   C C H   H H H   H H H H H H   H H H H H H   H H H H H H   H H H C   By PCASSO
```

J. V. Antony et al., 2021

Secondary Structure Assignment



	(a) DSSP assignment for models 1-23	(b) DSSPcont assignment for model 1							
	1 5 10 15 20 23	G	H	I	T	E	B	S	L
20 Val	LLLLLLLLLLLLLLLLLLLLLLL	0	0	0	32	0	0	0	68
21 Ser	TLLHTLLTLLLLHLLLHLLTLLT	0	0	0	100	0	0	0	0
22 Glu	TTTHTTTTHTSSHTSTHSSTSTT	0	0	0	100	0	0	0	0
23 Glu	STTHHTTHHTSSHTHTHSSSHTH	0	32	0	21	0	0	47	0
24 Ser	HHHHHHHHHHHHHHHHHHHHHHH	0	99	0	1	0	0	0	0
25 Leu	HHHHHHHHHHHHHHHHHHHHHHH	0	99	0	1	0	0	0	0
26 Asn	HHHHHHHHHHHHHHHHHHHHHHH	0	99	0	1	0	0	0	0
27 Lys	HHHHHHHHHHHHHHHHHHHHHHH	1	99	0	0	0	0	0	0
28 Val	HTHTHHTHHTHHHTHHHHHTHHH	1	87	0	12	0	0	0	0
29 Arg	TTTTTTTTHTTTTTTTTTTTTTT	1	53	0	46	0	0	0	0
30 Asn	SSTSSSSSSTSSSSSTSTSSSTS	0	0	0	10	0	0	90	0
31 Arg	SSTSTSSSSTTSSTTTTTTSSTS	0	0	0	32	0	0	68	0
32 Glu	SSTSTLSLSTTSSTTTTTTLSLL	0	0	0	32	0	0	68	0
33 Glu	LLLLLLLLLLLLLLLLLLLLLLL	0	0	0	0	0	0	0	100

1cy3 structure from: Rothemund et al. & Sonnichesen, 1999, Structure, 7, 1325

DSSPcont assignment for 1c3y fragment. The variations between the secondary structure assignments for different NMR models of the same protein illustrate the impact of fluctuations on structure

P. Carter et al., 2003

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:
– LLEEEELLLLHHHHHHHHHHLHHHL...

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

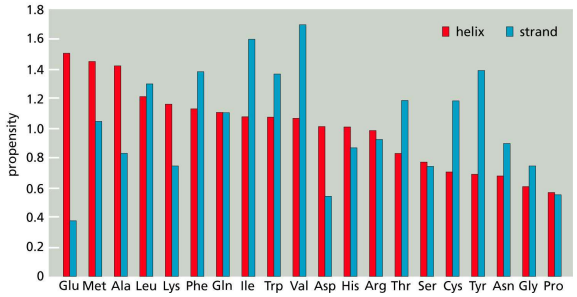
Chou-Fasman algorithm:

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 Parameters



Adopted from *Zvelebil, Baum, 2008*

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(turn) > 1.00`
3) `P(a) < P(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

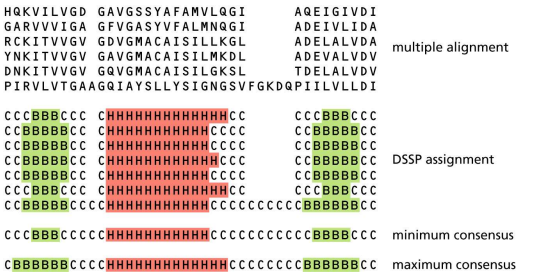
Likelihood of a secondary structure state depends on the neighboring residues:

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

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

Accuracy for a single sequence - 60%
Accuracy for an alignment - 65%

Evolutionary Information



Adopted from *Zvelebil, Baum, 2008*

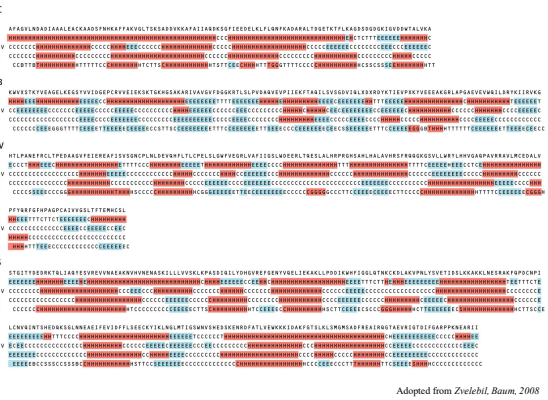
Evolutionary Methods

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

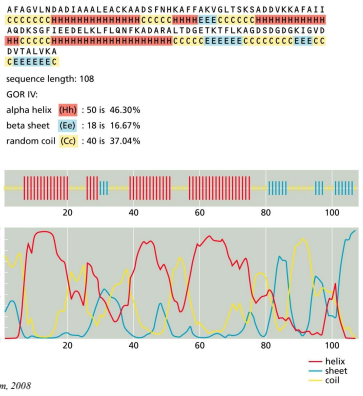
Additional information used for prediction
- mutation statistics
- residue position in sequence
- sequence length

Evolutionary Methods



Adopted from *Zvelebil, Baum, 2008*

Evolutionary Methods



Adopted from *Zvelebil, Baum, 2008*

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

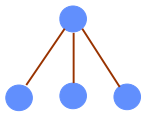
Stereochemical Methods

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	-	+	+	-

Neural Networks

Perceptron



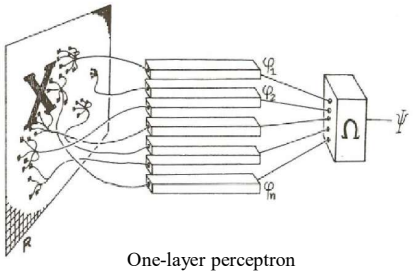
Output layer

Input layer

$$Y = \begin{cases} 1 \text{ if } \Sigma\, w_i i_i > \Theta \\ 0 \text{ otherwise} \end{cases}$$

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

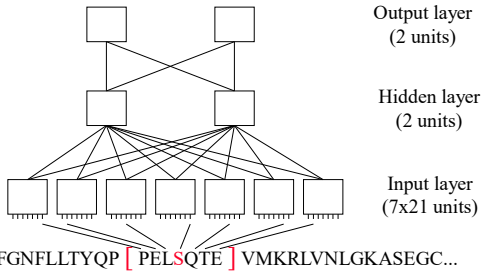
Neural Networks



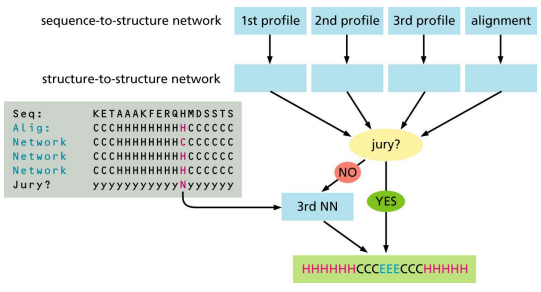
One-layer perceptron

M. L. Minsky & S. Papert, 1969

Neural Networks Methods

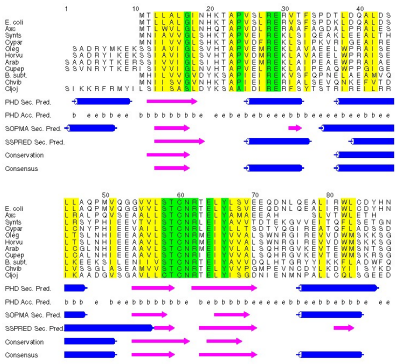


Jnet Algorithm

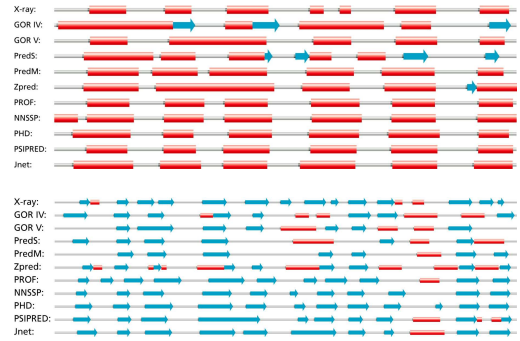


Adopted from *Zvelebil, Baum, 2008*

Jnet Algorithm

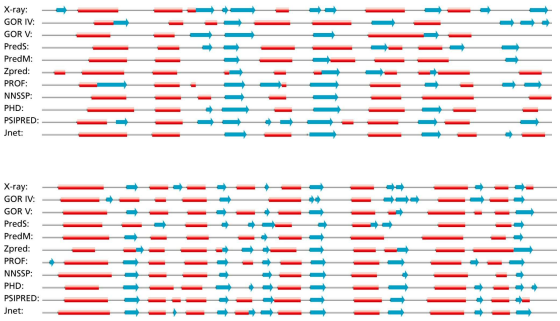


Accuracy of prediction



Adopted from *Zvelebil, Baum, 2008*

Accuracy of prediction



Adopted from *Zvelebil, Baum, 2008*

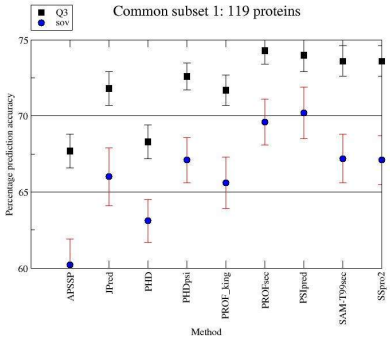
Accuracy of Prediction

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

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

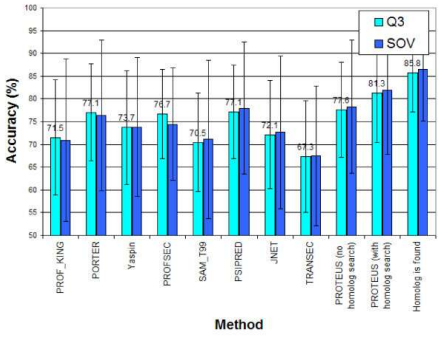
Range: 50-85%

Accuracy of prediction



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

Accuracy of prediction



S. Montgomerie et al., 2006