Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 25

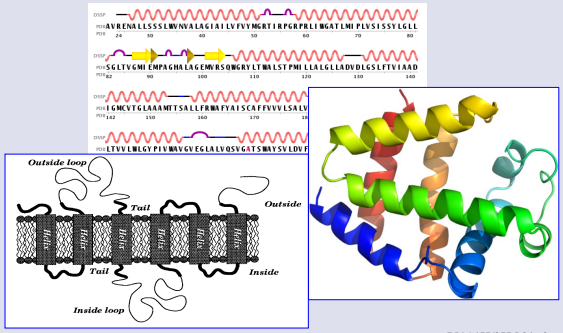
1

TODAY'S TOPICS

- ➢ Protein structures
  - • Protein structure databases (PDB, SCOP, CATH)
- ➢ Secondary structure prediction
  - • Prediction statistics

2

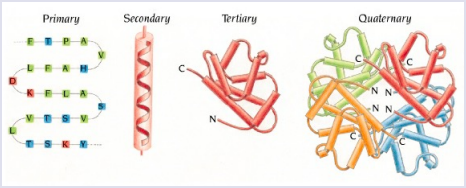
Prediction of protein structures



3

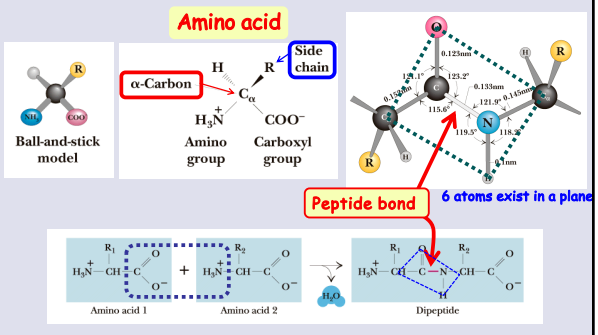
Hierarchical nature of protein architecture

- **Primary structure**: amino acid sequence
- **Secondary structure**: alpha helices, beta sheet, etc. (hydrogen bonding pattern of main chain):
- **Tertiary structure** (the assembly and interactions of the helices and sheets)
- **Quaternary structure** (the assembly of the monomers)



4

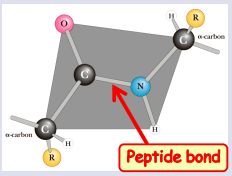
Proteins = polypeptide chains



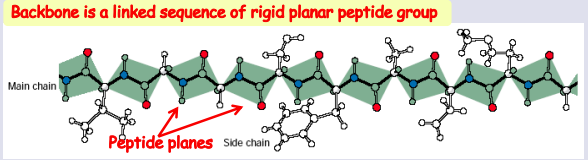
5

Proteins = polypeptide chains

- **Backbone (main chain)**: atoms that participate in peptide bonds
  - ➜ ignores the side chains
  - ➜ 6 atoms (O, C, N, H, and C_α) of the peptide group lie in a plane

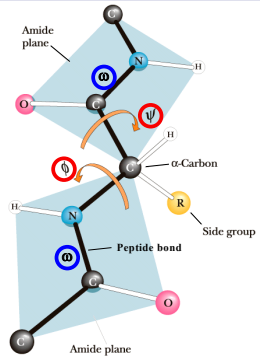


Backbone is a linked sequence of rigid planar peptide group



6

Torsion angles



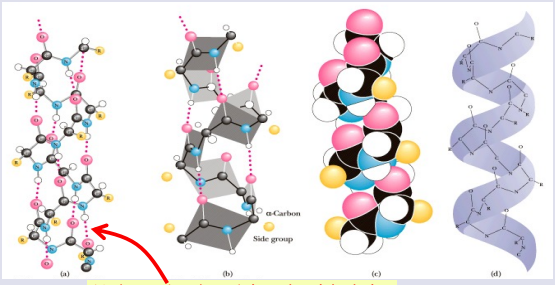
- The conformation of the backbone can be described by the torsion angles (or dihedral angles or rotation angles): ϕ and ψ
  - → The entire path of the backbone is known if ϕ and ψ are specified
- Some values of ϕ and ψ are more likely than others
  - → Due to steric interference between non-bonded atoms

BIOS477/877 L25 - 7

7

Secondary structure

➢ Alpha helix: ϕ = -60 and ψ = -45

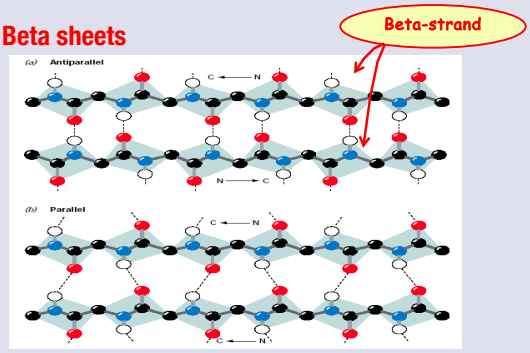


BIOS477/877 L25 - 8

8

Secondary structure

➢ Beta sheets



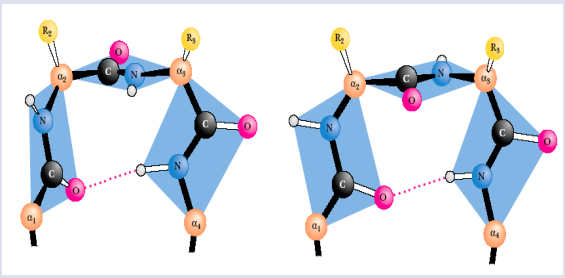
BIOS477/877 L25 - 9

9

Secondary structure

➢ Beta turns

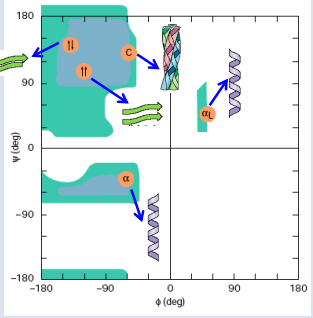
- Proline and glycine occur frequently in beta turns



BIOS477/877 L25 - 10

10

Ramachandran plot



- The sterically allowed values for ϕ and ψ can be calculated (colored regions)

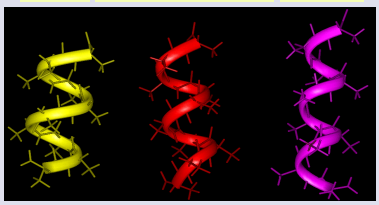
α: right-handed alpha helix
α_L: left-handed alpha helix
↑↑: parallel beta sheet
↑↓: antiparallel beta sheet
C: collagen helix

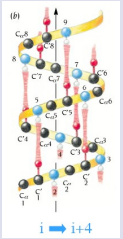
BIOS477/877 L25 - 11

11

Secondary structure conformation

π-helix Righ-handed α-helix 3_{10}-helix





	Φ	Ψ	H-bond pattern
Right handed α-helix	-57	-47	i + 4
π-helix	-57	-70	i + 5
3_{10} helix	-49	-26	i + 3
Parallel β-sheet	-119	113	
Antiparallel β-sheet	-139	135	

BIOS477/877 L25 - 12

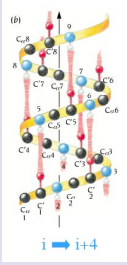
12

Secondary structure assignment

➢ **DSSP: Dictionary of Secondary Structure of Proteins**
https://pdb-redo.eu/dssp
- to standardize secondary structure assignment
- a database of secondary structure assignments for all protein entries in PDB

Kabsch and Sander (1983)

Based on the backbone-backbone hydrogen bonds (*e.g.*, α if two consecutive amino acids have i → i+4 hydrogen bonds) and geometrical features extracted from x-ray coordinates

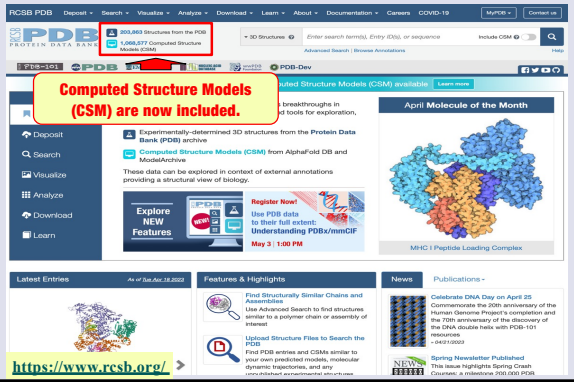


➢ **STRIDE: Secondary Structural Identification method**
https://github.com/MDAnalysis/stride Frishman & Argos (1995)

➜ uses empirically derived hydrogen bond energy and ϕ - ψ torsion angle criteria

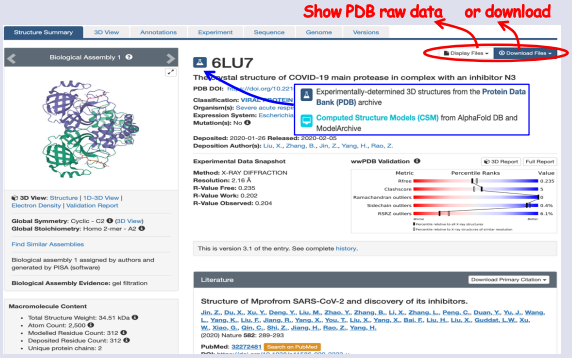
13

PDB: Protein Data Bank



14

PDB: Protein Data Bank

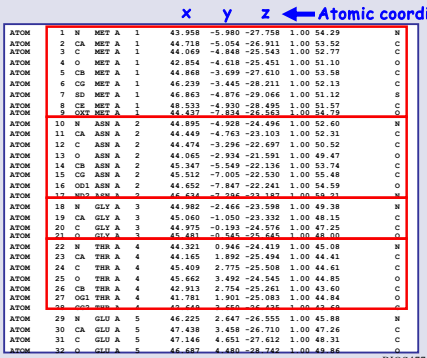


15

PDB format: http://www.wwpdb.org/docs.html

16

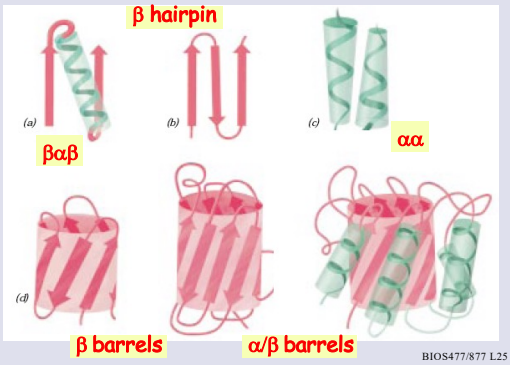
PDB format: http://www.wwpdb.org/docs.html



http://www.rcsb.org/pdb/

17

Databases for Protein Structure Classification



18

SCOP: Structural Classification of Proteins

http://scop.berkeley.edu/ (SCOPe: SCOP extended)

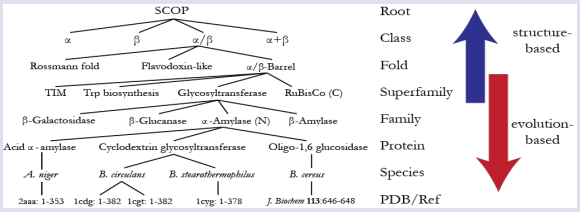
http://scop.mrc-lmb.cam.ac.uk/ (SCOP 2, not available?)

- Classification of protein domains of known structure according to their evolutionary and structural relationships
  - Protein domain: fundamental unit of classification
    - ➜ an evolutionary unit observed in nature either in isolation or in more than one context in multidomain proteins
    - ➜ hierarchically classified into: families, superfamilies, folds, and classes

(in SCOP2, classification is network-like but not hierarchical)

19

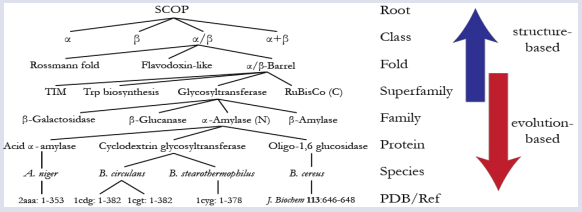
SCOP: Structural Classification of Proteins



- Family: Proteins having a common evolutionary origin
  - ➜ Amino acid identities of 30 % and greater (detectable using BLAST, PSI-BLAST, HMMER, *etc.*)
- Superfamily: Families having a probable common evolutionary origin
  - ➜ Lower identities but whose structures and functional features suggest common evolutionary origin

20

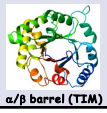
SCOP: Structural Classification of Proteins



- Fold: Same major secondary structures in same arrangement with the same topological connections

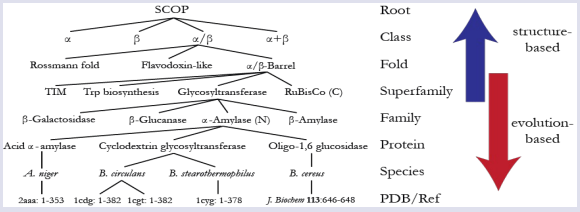

Rossmann fold


Flavodoxin-like fold


α/β barrel (TIM)

21

SCOP: Structural Classification of Proteins

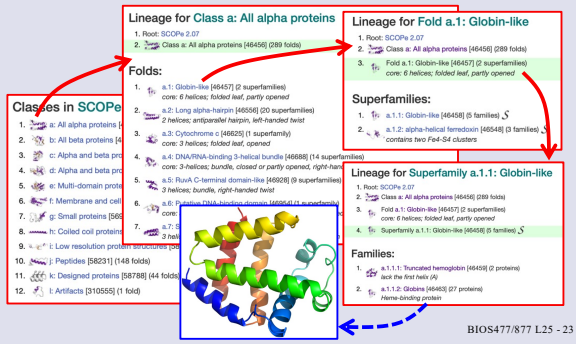


- IUPRs (Intrinsically Unstructured Protein Region): non-globular structures (IUPRs and Types are introduced in SCOP 2)
  - ➜ Folds and IUPRs are grouped into 5 structural Classes: 1) all α, 2) all β, 3) α and β (α/β; interspersed α and β), 4) α plus β (α+β; segregated α and β), and 5) small proteins
  - ➜ Folds and IUPRs are also grouped based on 4 protein Types: soluble, membrane, fibrous, or intrinsically disordere

22

SCOP: Structural Classification of Proteins

http://scop.berkeley.edu/



23

SCOP: Structural Classification of Proteins

"The SCOP classification of proteins has been constructed manually by visual inspection and comparison of structures, but with the assistance of tools"

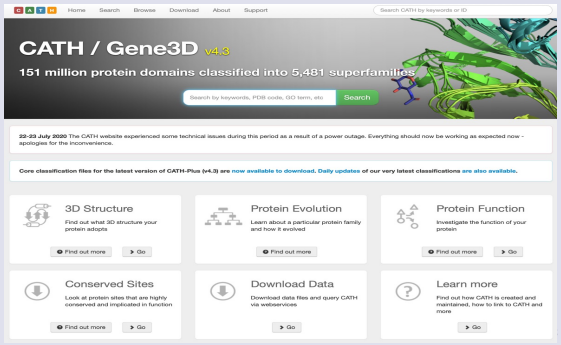
SCOPe 2.08-stable statistics:
106976 PDB entries (released/updated prior to 2021-07-28). 344851 Domains. 1 Literature reference.

Class	Number of folds	Number of superfamilies	Number of families
a: All alpha proteins	290	519	1089
b: All beta proteins	180	375	993
c: Alpha and beta proteins (a/b)	148	247	1003
d: Alpha and beta proteins (a+b)	396	580	1387
e: Multi-domain proteins (alpha and beta)	74	74	128
f: Membrane and cell surface proteins and peptides	69	131	204
g: Small proteins	100	141	280
Totals	**1257** (26 new)	**2067** (42 new)	**5084** (88 new)

24

CATH Protein Structure Classification

http://www.cathdb.info/



25

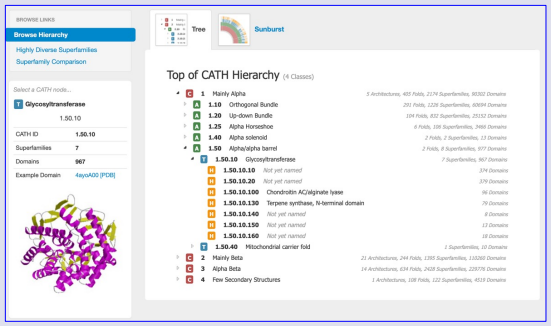
CATH Protein Structure Classification

- **Class, C-level**: determined according to the secondary structure composition and packing within the structure
  - ➜ mainly-α, mainly-β, α–β, and few secondary structures
- **Architecture, A-level**: the overall shape of the domain structure
  - ➜ determined by the orientations of the secondary structures
  - ➜ done by manually
- **Topology (Fold family), T-level**: depending on both the overall shape and connectivity of the secondary structures
  - ➜ done by the structure comparison algorithm (SSAP)
- **Homologous Superfamily, H-level**: depending on sequence similarity and structural comparison
  - ➜ sequence identity >= 20 ~ 35%
- **Sequence families, S-level, or Domain**: based on sequence similarity

26

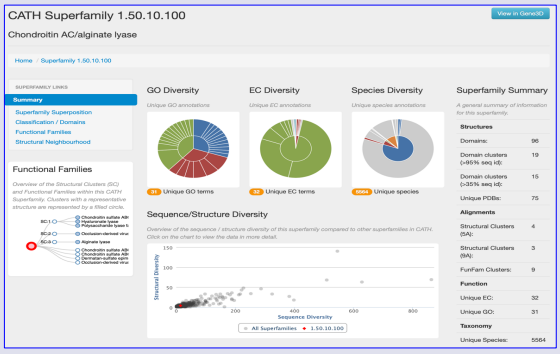
CATH Protein Structure Classification

http://www.cathdb.info/



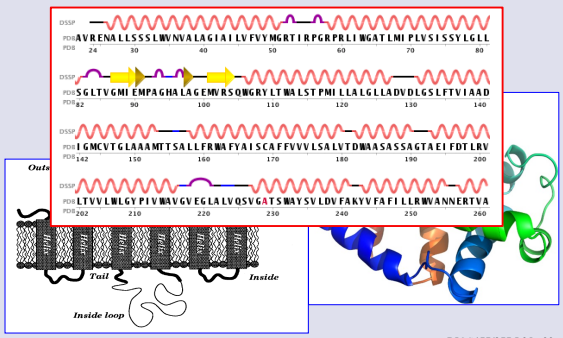
27

CATH Protein Structure Classification



28

Secondary structure prediction



29

Secondary structure prediction

➢ Various amino acids stabilize or destabilize alpha helix

Table 6.1
Helix-Forming and Helix-Breaking Behavior of the Amino Acids

Amino Acid		Helix Behavior*	
A	Ala	H	(I)
C	Cys	Variable	
D	Asp	Variable	
E	Glu	H	
F	Phe	H	
G	Gly	I	(B)
H	His	H	(I)
I	Ile	H	(C)
K	Lys	Variable	
L	Leu	H	
M	Met	H	
N	Asn	C	(I)
P	Pro	B	
Q	Gln	H	(I)
R	Arg	H	(I)
S	Ser	C	(B)
T	Thr	Variable	
V	Val	Variable	
W	Typ	H	(C)
Y	Tyr	H	(C)

*H = helix former; I = indifferent; B = helix breaker; C = random coil; ( ) = secondary tendency.

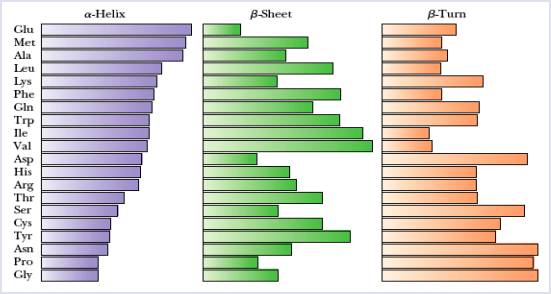
H: helix forming amino acid

Proline: helix breaker

30

Secondary structure prediction

Relative frequency of occurrence of amino acids in α-helices, β-sheets, and β-turns



31

Secondary structure prediction

➢ **Chou-Fasman method** Chou and Fasman (1974)

http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=misc1

- The first generation, simple statistical method
- Based on the property (propensity) of a single residue
- Propensities are calculated from a set of proteins whose structures are known

➜ Propensity for α-helix, β-sheet, or turn
= log{Pr(amino acid *i* in α-helix , β-strand, or turn) / Pr(amino acid *i*)}

32

Secondary structure prediction

- Conformational propensity table used in Chou-Fasman method

Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Ala	142	83	66	0.06	0.076	0.035	0.058
Arg	98	93	95	0.070	0.106	0.099	0.085
Asp	101	54	146	0.147	0.110	0.179	0.081
Asn	67	89	156	0.161	0.083	0.191	0.091
Cys	70	119	119	0.149	0.050	0.117	0.128
Glu	151	37	74	0.056	0.060	0.077	0.064
Gln	111	110	98	0.074	0.098	0.037	0.098
Gly	57	75	156	0.102	0.085	0.190	0.152
His	100	87	95	0.140	0.047	0.093	0.054
Ile	108	160	47	0.043	0.034	0.013	0.056
Leu	121	130	59	0.061	0.025	0.036	0.070
Lys	114	74	101	0.055	0.115	0.072	0.095
Met	145	105	60	0.068	0.082	0.014	0.055
Phe	113	138	60	0.059	0.041	0.065	0.065
Pro	57	55	152	0.102	0.301	0.034	0.068
Ser	77	75	143	0.120	0.139	0.125	0.106
Thr	83	119	96	0.086	0.108	0.065	0.079
Trp	108	137	96	0.077	0.013	0.064	0.167
Tyr	69	147	114	0.082	0.065	0.114	0.125
Val	106	170	50	0.062	0.048	0.028	0.053

β-turn

f(1), ..., f(4): Freq. of each AA at each position in β-turn

33

Secondary structure prediction

➢ **Chou-Fasman method** Chou and Fasman (1974)

http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=misc1

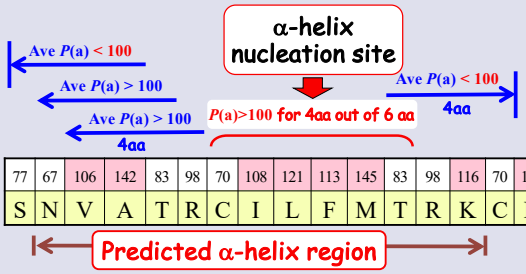
1. Identification of helix and sheet nuclei:
   helix: 4 out of 6 aa w/ high helix propensity ($P > 100$)
   strand: 3 out of 5 aa w/ high strand propensity ($P > 100$)
2. Propagation in both directions until 4-aa Average (P) < 100
3. Helix if $\Sigma P(a) > \Sigma P(b)$, strand if $\Sigma P(b) > \Sigma P(a)$

➜ Turn prediction:
i) $p(t) > 0.000075$, ii) $P(turn) > 100$, and
iii) $P(a) < P(turn) > P(b)$, where $p(t) = f(j)f(j+1)f(j+2)f(j+3)$

34

Secondary structure prediction

➢ **Chou-Fasman method**



35

Secondary structure prediction

➢ **Garnier-Osguthorpe-Robson (GOR)** Garnier *et al.* (1978)

http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=misc1

- Likelihood of a secondary structure state depends on the neighboring residue
  ➜ Chou-Fasman method assumes that each amino acid individually influences secondary structure
- Uses principles of information theory Garnier *et al.* (1996)
- Algorithm
  ➜ For each residue, 8 N-terminal and 8 C-terminal positions are considered (a window size = 17)
  ➜ The frequencies with which each amino acid occupies each of the 17 window positions in helices, strands, and turns are compiled (a 17 x 20 scoring matrix) *e.g.*, *P*(Ala at pos 1 | alpha helix) → Lod score
  ➜ The probabilities that each residue in a target sequence will be involved in a helix, strand, or turn are calculated

36

Secondary structure prediction

➢ Chou-Fasman *vs.* GOR

- Chou-Fasman method was originally claimed at 70-80 % accuracy
- GOR method was originally claimed at 64 % accuracy

➔ Later studies described a lower accuracy: ~ 55% accuracy for both methods!

37

Secondary structure prediction

➢ Chou-Fasman *vs.* GOR (example)

Chou-Fasman

```
        MGRRVPALRQLLVLAVLLLKPSQLQSRELSGSRCPEPCDCAPDGALRCPGPRAGLARLSL
 helix     <----------------------->                              <-------
strand     EEEEEEEEEEEEEEEEEEE                                    EEEEEE
 turns   T                  TT   TT   TT   T      T       TT

        TYLPVKVIPSQAFRGLNEVVKIEISQSDSLERIEANAFDNLLNLSELLIQNTKNLLYIEP
 helix ----------------------------------------------------------------
strand EEEEEEEEEEEEEEEEEEEEEEEEE                EEEEEEEEEEEEEEEEEEE
 turns          TT              T T                          T T      T
```

GOR

```
        MGRRVPALRQLLVLAVLLLKPSQLQSRELSGSRCPEPCDCAPDGALRCPGPRAGLARLSL
 helix       HHHHHHHHHHHHH    H
strand    EEE                 EEEE                    EEE    EEEEEEEE
 turns TTT                 T      TTTTTTTTTTTTT TTTT    T   T
 coil                   CC                  C        CC

        TYLPVKVIPSQAFRGLNEVVKIEISQSDSLERIEANAFDNLLNLSELLIQNTKNLLYIEP
 helix              HHHH HHH      HHHHHHHHHHHHHHHHHHHHH
strand EEEEEEEEEEEE          E                                EEEE
 turns                                                    TT      T
 coil              CCCCC         CCCCC                        CC
```

(N-terminal of LSHR_RAT Lutropin-choriogonadotropic hormone receptor)

38

Secondary structure prediction

➢ Neural Network methods

- It requires many solved data to train the programs to distinguish patterns found in known secondary structures from other patterns not located in these structures
- Inputs can be other than sequence information
- It can capture higher order correlations
- Risk of overtraining (or overfitting; fits perfectly with the training data, but not other sequences)

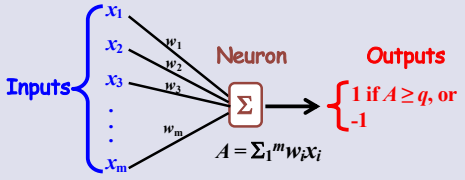
(Neural networks are used also in gene prediction)

39

Secondary structure prediction

➢ Neural Network methods

- Perceptron (threshold unit)

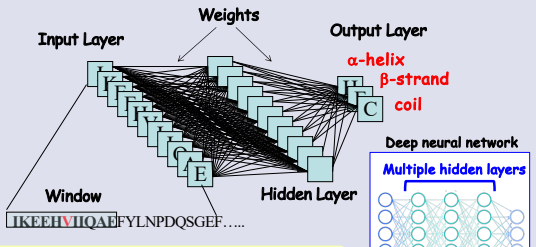


- Using a training dataset, w_i's can be optimized.
  → A set of known input-output pairs
- Neural networks are built from many perceptrons.

40

Secondary structure prediction

➢ Neural Network methods



Various information from surrounding amino acids can be used to predict the secondary structure of the central amino acid

41

Secondary structure prediction

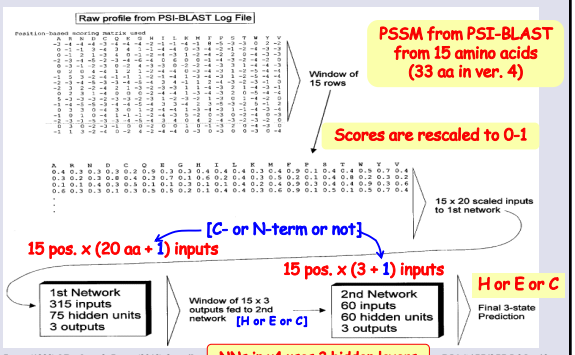
➢ Prediction accuracy has been increased by incorporating

- More sophisticated algorithms (neural network, nearest neighbor, etc.)
- Multiple alignment
- Position specific scoring matrix (PSI-BLAST)

➔ Current accuracy level: 75~80%
(82~85% by deep learning methods)

42

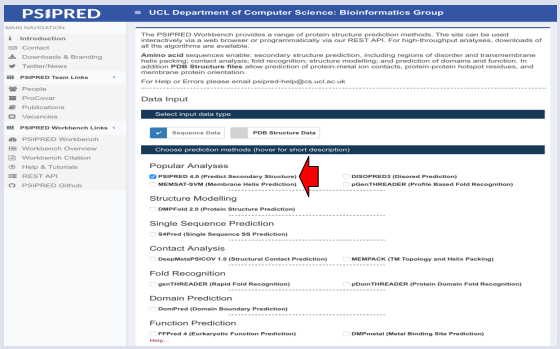
PSIPRED



43

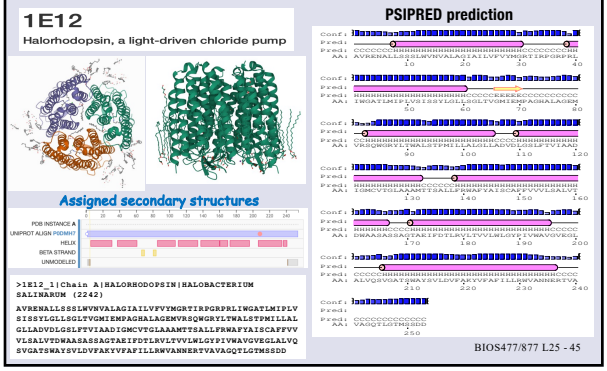
PSIPRED Protein Sequence Analysis Workbench

http://bioinf.cs.ucl.ac.uk/psipred/



44

Secondary structure prediction

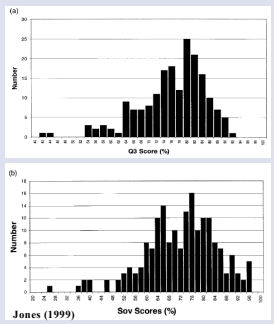


BIOS477/877 L25 - 45

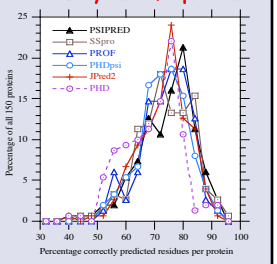
45

PSIPRED prediction accuracy

Distribution of prediction accuracy from 187 PDB proteins (structure is known)



Accuracy varies for proteins



Q3=84.2% in ver 4 Buchan & Jones (2019)

BIOS477/877 L25 - 46

46

Prediction statistics (general)

Accuracy = *(TP+TN)/(TP+FN+FP+TN)*
Error = 1 - Accuracy
Sensitivity = *TP/(TP+FN)* [=*TP*/(actual positives)]
Specificity = *TN/(TN+FP)* [=*TN*/(actual negatives)]

True positive rate = Sensitivity (or Recall)
False positive rate = *FP/(TN+FP)* = 1 - Specificity
Precision (or Positive predictive value) = *TP/(TP+FP)*

TP: True positive, *TN*: True negative, *FP*: False positive, *FN*: False negative

Confusion matrix		predicted	
		+	-
actual	+	TP	FN
	-	FP	TN

• Mathews correlation coefficients: *CC* (or *MCC*)

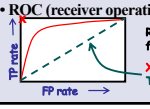
$$CC = \frac{(TP \bullet TN - FP \bullet FN)}{\sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)}}$$

If everything is correct, *CC* = 1, if all predictions are wrong, *CC* = -1

• F-measure (*F* or *F1*): the harmonic mean of precision and recall

$F = 2TP/(2TP+FN+FP)$ 0 < *F* < 1

• ROC (receiver operating characteristic) plot



ROC plot can be obtained using different threshold scores for identifying positive and negative predictions

X: if prediction is perfect (TP=1, FP=0)
The worst case scenario (random choice prediction)

BIOS477/877 L25 - 47

47

Prediction statistics (secondary structures)

➤ Accuracy of prediction

• Three-state per-residue accuracy:

$$Q_3 = 100 \times \frac{(P_\alpha + P_\beta + P_T)}{N_{res}}$$

p_α, p_β, p_T: numbers of residues correctly predicted as α-helix, β-strand, or turn (loop)
N_{res}: the total number of residues identified (known) as α-helix, β-strand, or turn (loop)
$N_{res} = N_\alpha + N_\beta + N_T$

➜ Per-state accuracy: *e.g.*, $Q_\alpha = 100 \times p_\alpha / N_\alpha$

• Mathews correlation coefficients: C_α, C_β, C_T

$$C_\alpha = \frac{(p_\alpha n_\alpha - u_\alpha o_\alpha)}{\sqrt{(n_\alpha + u_\alpha)(n_\alpha + o_\alpha)(p_\alpha + u_\alpha)(p_\alpha + o_\alpha)}}$$

If everything is correct, $C_\alpha = 1$
If all predictions are wrong, $C_\alpha = -1$

p_α: number of correct positive predictions
n_α: number of correct negative predictions
o_α: number of false positives
u_α: number of false negatives

BIOS477/877 L25 - 48

48

Prediction statistics (secondary structures)

➢ **Accuracy of prediction**

- **Segment overlap value: SOV** Zemla *et al.* (1999)

$$SOV = (1/N)\Sigma_x[\{minov(S_1;S_2) + \delta(S_1;S_2)\}/maxov(S_1;S_2)] \text{ X } len(S_1)$$

S_1, S_2: observed and predicted secondary structure segments
N: the total number of residues
minov: the actual overlap, maxov: the total extent of the segment

$$\delta(S_1;S_2) = \min\begin{cases} maxov(S_1;S_2) - minov(S_1;S_2), \\ minov(S_1;S_2), \\ int\{0.5 \text{ x } len(S_1)\}, \text{ or} \\ int\{0.5 \text{ x } len(S_2)\} \end{cases}$$

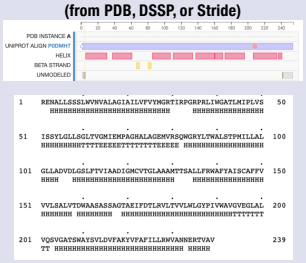
➜ attempts to capture segment prediction

49

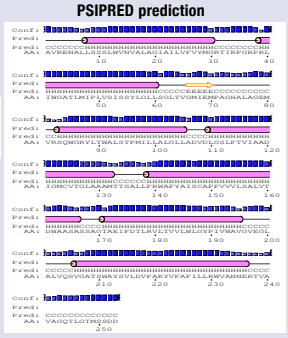
Evaluating secondary structure prediction

1E12
Halorhodopsin, a light-driven chloride pump

Secondary structure assignment
(from PDB, DSSP, or Stride)



PSIPRED prediction



50

Evaluating secondary structure prediction

➢ **Accuracy of prediction (an example)**

Q_α or $Q_H = 100 \times p_\alpha / N_\alpha$: per-residue accuracy for α-helix

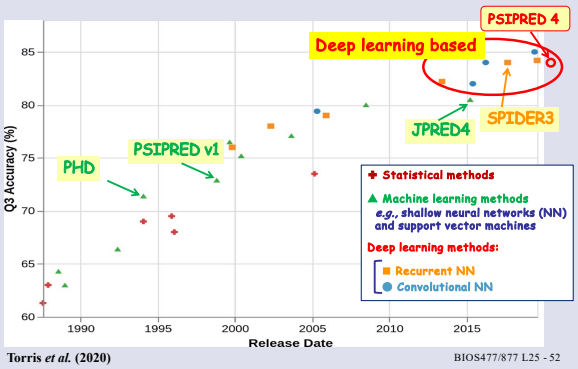
p_α: number of residues correctly predicted as helix.

N_α: the total number of residues identified (known) as helix (in DSSP).

p_α=171, N_α=186, Q_H=100x171/186 = 91.9%

51

Secondary structure prediction

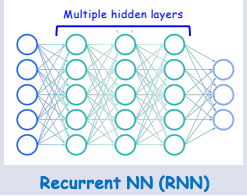


Torris *et al.* (2020)

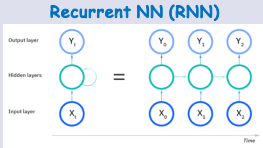
52

Deep Neural Networks

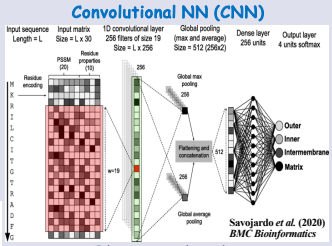
Multi-layer feedfoward NN



Recurrent NN (RNN)



Convolutional NN (CNN)



Savojardo *et al.* (2020) *BMC Bioinformatics*

Filters move along the sequence generating feature maps

Outputs depend on the prior elements

[IBM cloud education]

53

Sixty-five years of the long march in protein secondary structure prediction: the final stretch?

Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal and Yaoqi Zhou

Briefings in Bioinformatics, 19(3), 2018, 482-494

Data set Method	TS115		CASP12		Server location
	Q3	P-value*	Q3	P-value*	
Jpred4	0.771[b]	0.0007	0.751	0.04	http://www.compbio.dundee.ac.uk/jpred4/index.html
SPINE X	0.801	0.0002	0.769	0.006	http://sparks-lab.org/SPINE-X/
PSIPRED 3.3	0.802	0.12	0.780	0.19	http://distillf.ucd.ie/porterpalealе/
SCORPION	0.817	0.45	0.805	0.44	Stand-alone version from http://hpcr.cs.odu.edu/c3scorpion/
SPIDER2	0.819	NA	0.798	NA	http://sparks-lab.org/server/SPIDER2/
PORTER 4.0	0.820	0.17	0.798	0.67	http://distillf.ucd.ie/porterpalealе/
DeepCNF	0.823	0.01	0.821	0.14	http://raptorx2.uchicago.edu/StructurePropertyPred/predict/

Note.
*Paired t-test from SPIDER2.

SCORPION: Singlet/doublet/triplet prob. + PSSM
SPIDER2: Iterative learning of secondary structure, backbone torsion angles, and solvent accessible surface area using deep NN (three layers) with PSSM + physico-chemical properties
DeepCNF: Deep convolutional NN with PSSM + 78 AA related features

PSIPRED 4: Q3=84.2%; PSSM + 2 hidden layer NNs

54

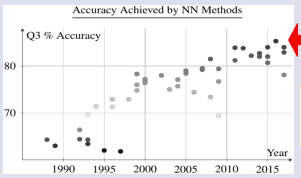
Protein secondary structure prediction using neural networks and deep learning: A review

Computational Biology and Chemistry 81 (2019) 1–8

Wafaa Wardah[a], M.G.M. Khan[a], Alok Sharma[b,c], Mahmood A. Rashid[c,d,*]

The major periodically relevant state-of-the-art methods are shown along with the types of feature values they employed in their networks.

Neural network method	Accuracy (Q3)	Seq info	Evo info	Physico chem info
Qian &Sejnowski 1988 (Qian and Sejnowski, 1988)	64.3%	√		
PHD 1994 (Rost et al., 1994)	71.4%	√	√	
PSIPRED 1997 (Jones, 1999)	76.5%	√	√	
JPRED3 2008 (Cole et al., 2007)	81.5%	√	√	
SPIDER3 2017 (Heffernan et al., 2017)	84%	√	√	√



Evolutionary or profile information = MSA

Limit of secondary structure prediction accuracy:

estimated to be 88% (Rost 2001)

55

Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks

Jack Hanson[1,*], Kuldip Paliwal[1], Thomas Litfin[2], Yuedong Yang[3] and Yaoqi Zhou[2,4,*]

Bioinformatics, *35*, 2019, 2403-10

Table 2. Test performance of several recently developed predictors alongside SPOT-1D on the latest PDB structures (TEST2018) (250 proteins)

Predictor	SS3	SOV3	PSS3	SS8	SOV8	PSS8	ASA	HSEa-U	HSEa-D	CN	θ	τ	ϕ	ψ
SPIDER-3-Single	72.57	64.08	$<1\times10^{-10}$	59.81	57.86	$<1\times10^{-10}$	0.570	0.603	0.533	0.619	11.07	45.39	23.77	43.05
RaptorX DeepCNF	81.62	66.58	$<1\times10^{-10}$	70.43	65.66	$<1\times10^{-10}$	–	–	–	–	–	–	21.01	35.95
PSRSM	81.94	74.22	$<1\times10^{-10}$	–	–	–	–	–	–	–	–	–	–	–
SPIDER-3	83.84	73.89	$<1\times10^{-10}$	–	–	–	0.768	0.764	0.716	–	7.73	29.62	18.38	28.10
PORTER-5	84.10	74.04	$<1\times10^{-10}$	73.22	70.27	$<9.89\times10^{-9}$	–	–	–	–	–	–	–	–
MUFOLD	84.78	77.56	$<2.73\times10^{-8}$	73.66	71.34	$<2.15\times10^{-9}$	–	–	–	–	–	–	17.78	27.24
NetSurfP-2.0	85.31	78.58	$<2.20\times10^{-3}$	73.81	71.14	$<3.64\times10^{-7}$	0.801	–	–	–	–	–	17.90	26.63
SPOT-1D-base	85.66	78.77	$<1.08\times10^{-2}$	74.26	71.45	$<1.33\times10^{-4}$	0.799	0.812	0.775	0.837	7.03	26.86	17.15	25.41
SPOT-1D	86.18	79.00	–	75.41	73.30	–	0.803	0.814	0.779	0.841	6.91	25.94	16.89	24.87

SS3=Q3, SS8=Q8
ASA: Solvent accessible surface area
CN: Contact numbers
HSE: Half-sphere exposure

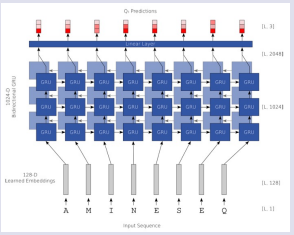
Backbone torsion angles (mean absolute error)

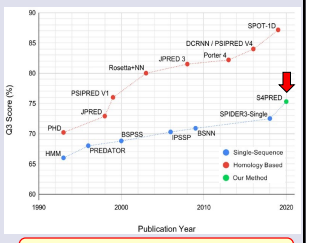
56

Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework

Lewis Moffat[1,2,*] and David T. Jones[1,2,*]

Bioinformatics, *37*, 2021, 3744-51





Uses bidirectional gated recurrent unit (GRU) with 1024 hidden dimensions in each direction

S4PRED (= next version of PSIPRED-single) achieved 75% accuracy without using profiles

57

Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment

Jaspreet Singh, Kuldip Paliwal, Thomas Litfin, Jaswinder Singh & Yaoqi Zhou

Scientific Reports 12, Article number: 7607 (2022) | Cite this article

	Model	SS3	SS8	ASA	HSE-u	HSE-d	CN	ψ	ϕ	θ	τ
No MSA	SPIDER3-Single	72.57	59.81	0.647	0.523	0.487	0.547	43.05	23.78	11.07	45.38
No MSA	ProteinUnet	72.57	60.30	0.620	0.537	0.510	0.545	42.93	23.42	10.28	44.94
No MSA	SPOT-1D-Single	74.28	72.17	0.665	0.573	0.563	0.585	40.58	22.16	9.35	42.32
	NetSurfP-2.0(profile)*	85.35	73.48	0.783	–	–	–	26.63	17.90	–	–
	SPOT-1D (profile)	86.18	75.41	0.787	0.732	0.737	0.777	24.87	16.88	6.91	25.94
No MSA	SPOT-1D-LM (This work)	86.74	76.47	0.814	0.759	0.761	0.690	23.74	15.99	6.46	24.60

*NetSurfP-3.0: no MSA, LM + deep learning is also now available

Backbone torsion angles (mean absolute error)

SS3=Q3, SS8=Q8
ASA: Solvent accessible surface area
HSE: Half-sphere exposure; CN: Contact numbers
LM: language model (natural language processing)

58

Secondary structure prediction

➢ Prediction program servers

- PSIPRED Server http://bioinf.cs.ucl.ac.uk/psipred/
- MPI Bioinformatics Toolkits (Quick2D) https://toolkit.tuebingen.mpg.de
- PredictProtein Server http://www.predictprotein.org/

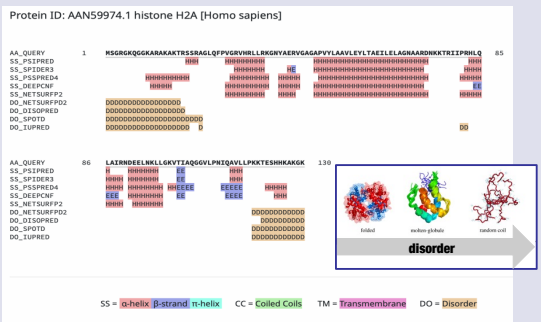
[Recent review papers (on Canvas)]

Meng and Kurgan (2016); Jiang *et al.* (2017); Oldfield *et al.* (2019); Torris *et al.* (2020); Ismi *et al.* (2022)

59

Secondary structure prediction

MPI Bioinformatics Toolkits (Quick2D) https://toolkit.tuebingen.mpg.de



60