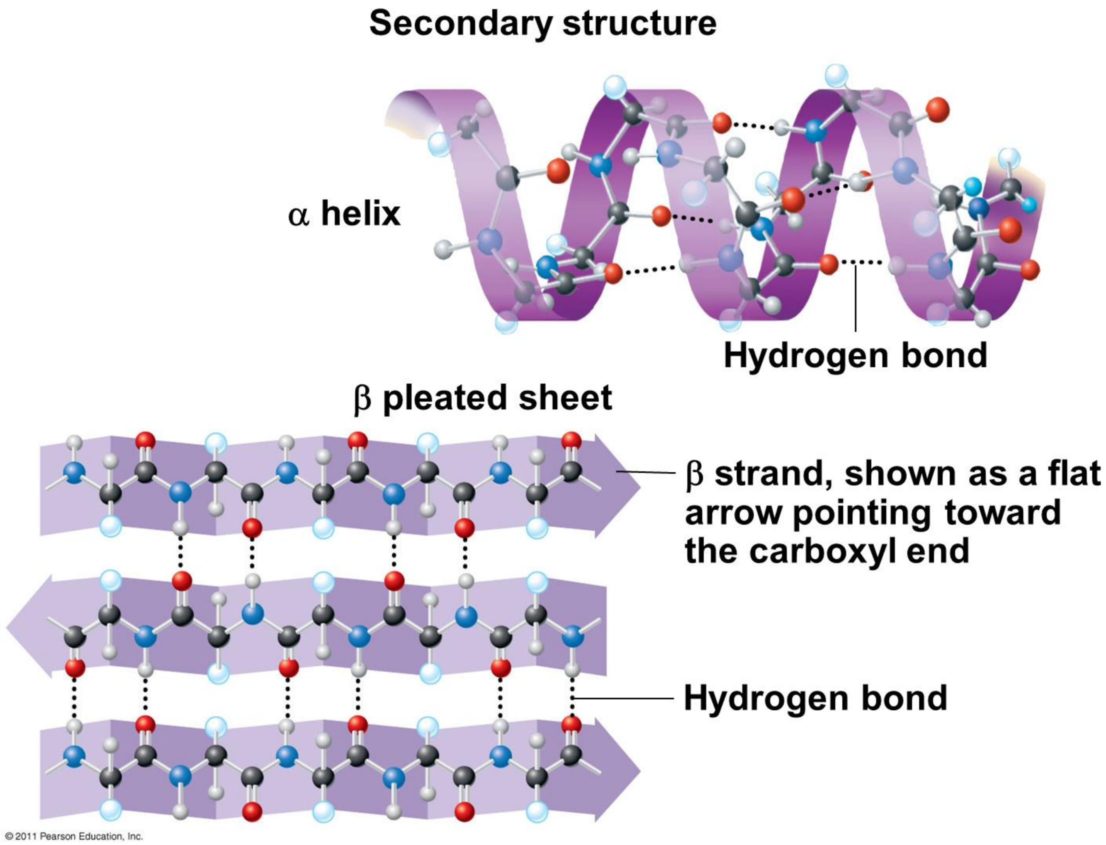
EECS730: Introduction to Bioinformatics

Lecture 12: Protein secondary structure prediction



© 2011 Pearson Education, Inc.

Some slides were adapted from Dr. Dong Xu (University of Missouri Columbia)

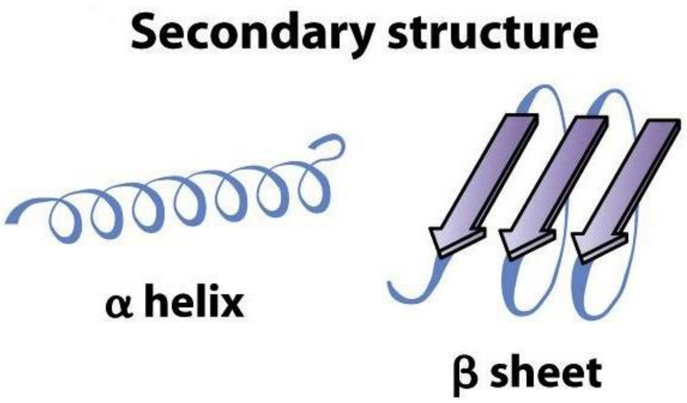
Structures in Protein

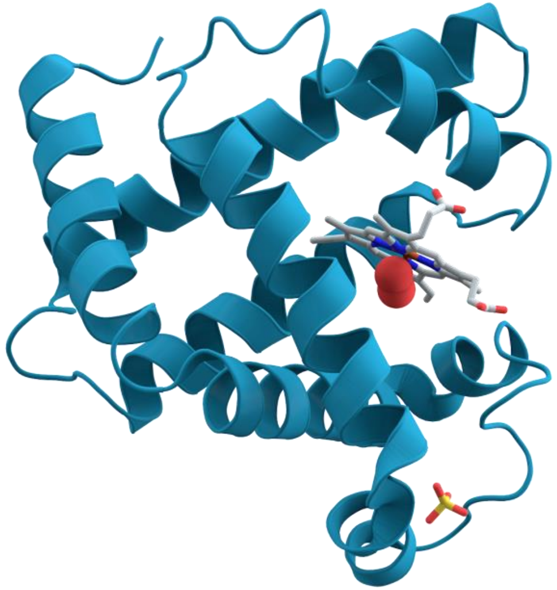
Language:

Letters → Words → Sentences

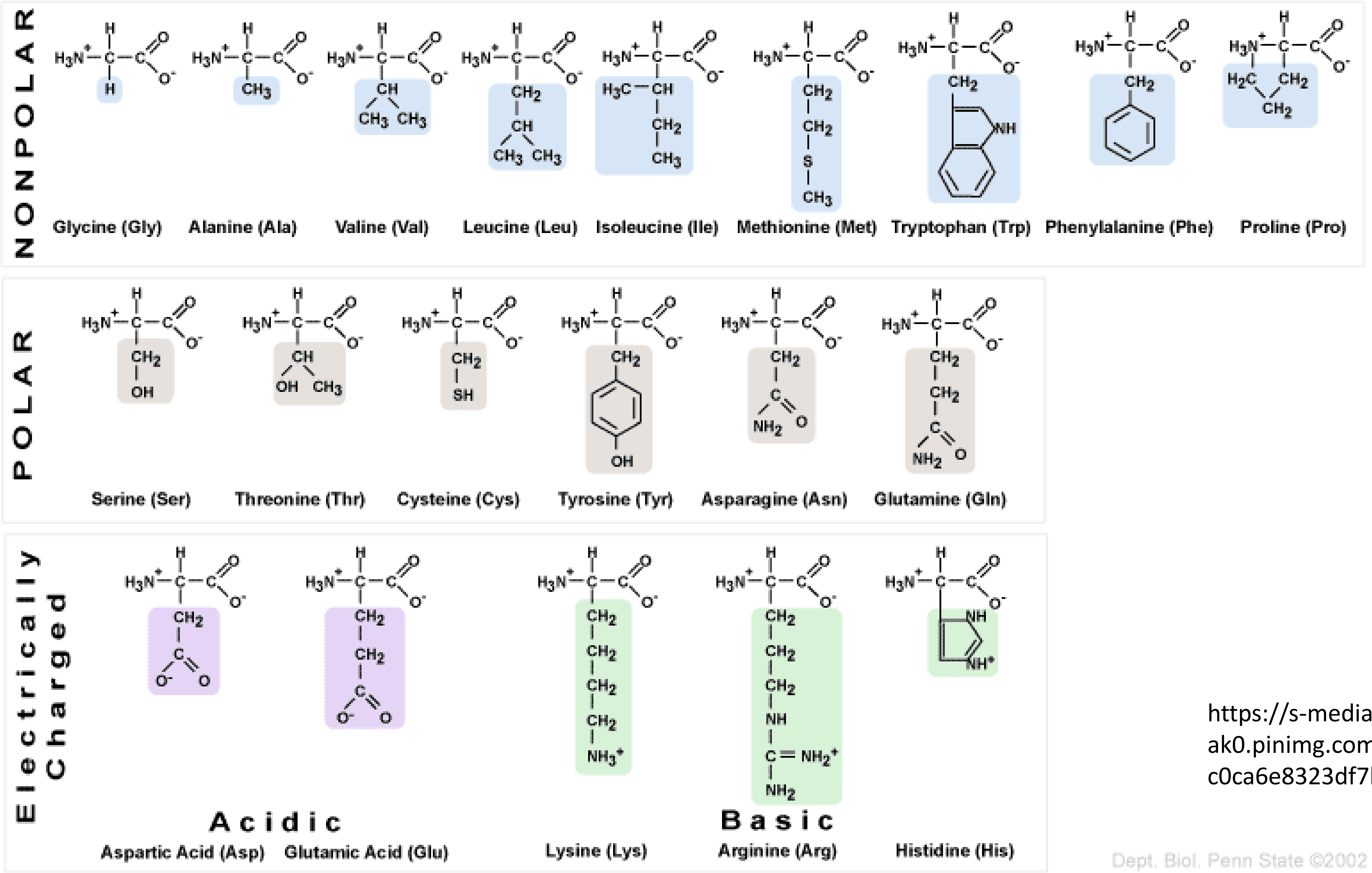
Protein:

Primary Structure → Secondary Structure →Tertiary Structure





Protein side chains



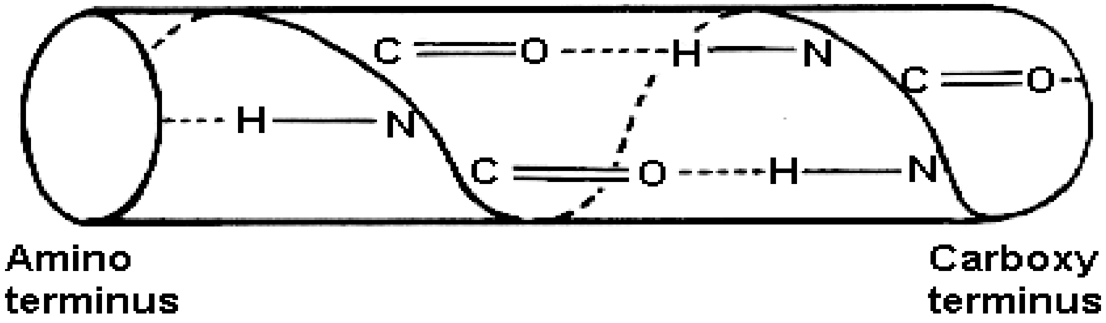
Dept. Biol. Penn State ©2002

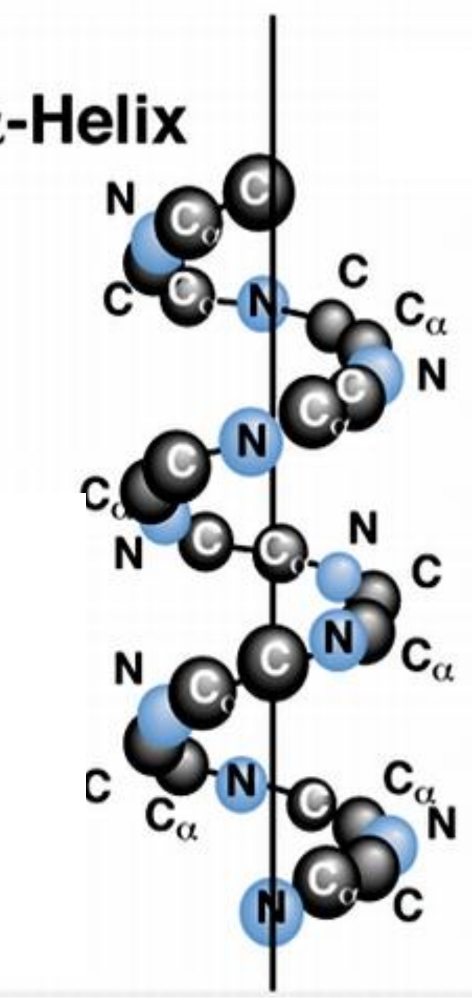
https://s-media-cache-ak0.pinimg.com/originals/ed/c0/ca/edc0ca6e8323df7bce06fd72ab5eca80.gif

α helix

- Single protein chain (local)
- Shape maintained by intramolecular H bonding between -C=O and H-N-

Toilet roll representation of the main chain hydrogen bonding in an alpha-helix.

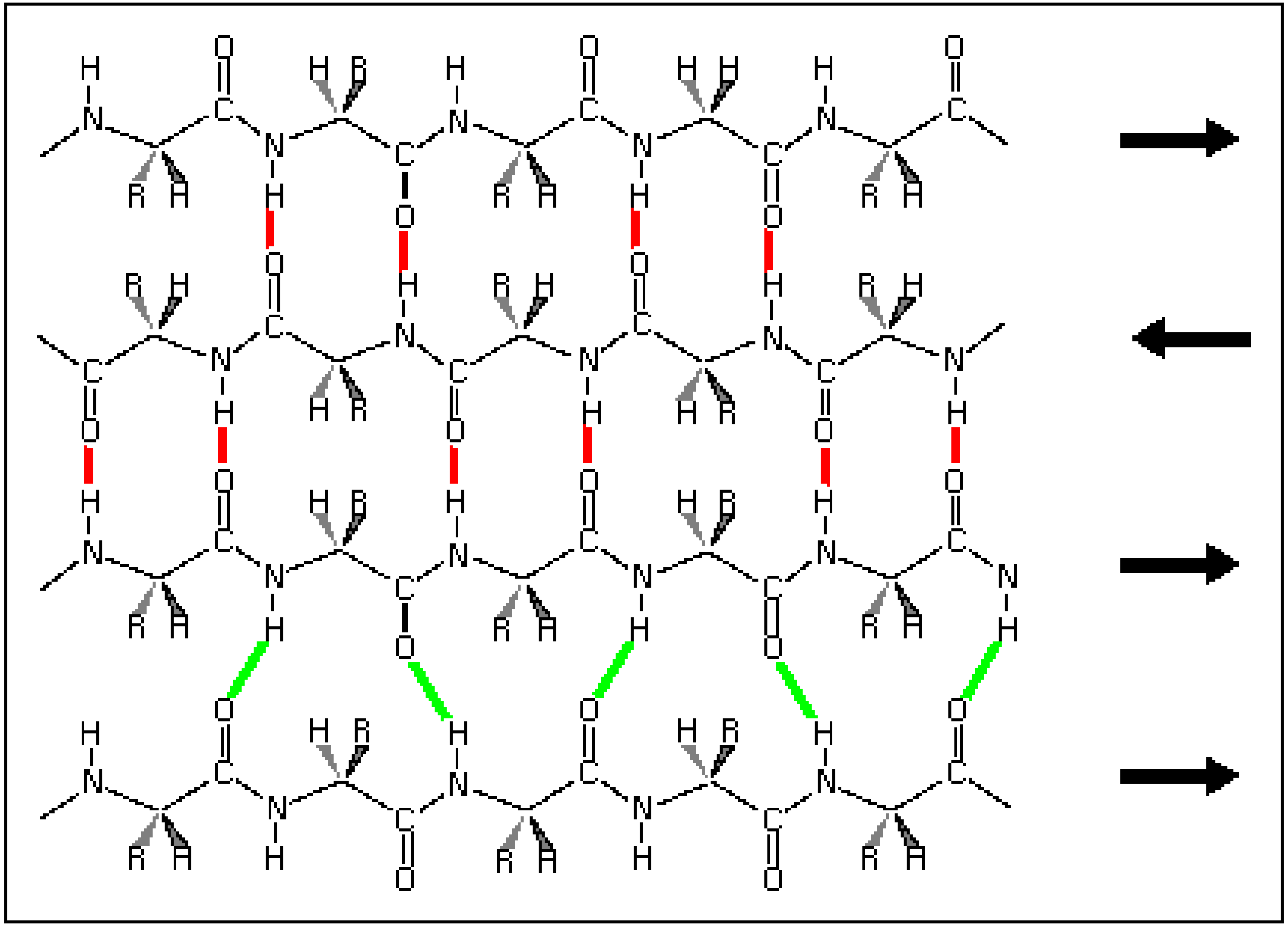




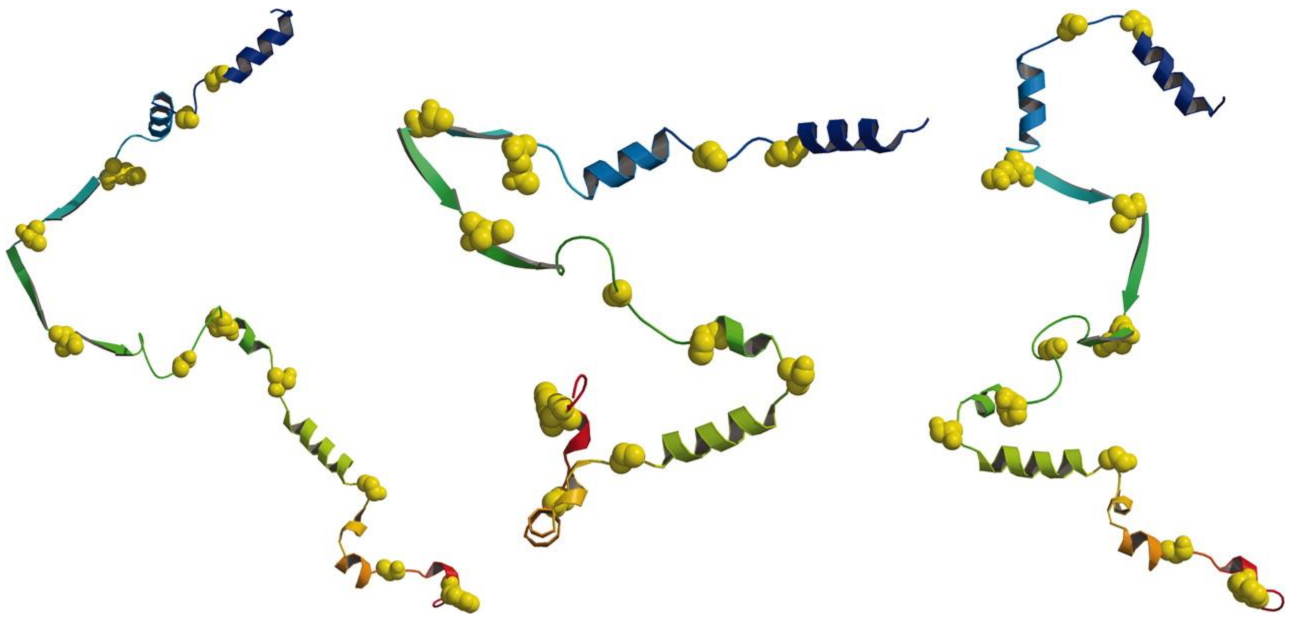
β sheet

- Several protein chains
- Shape maintained by intramolecular H bonding between chains
- Non-local on protein sequence

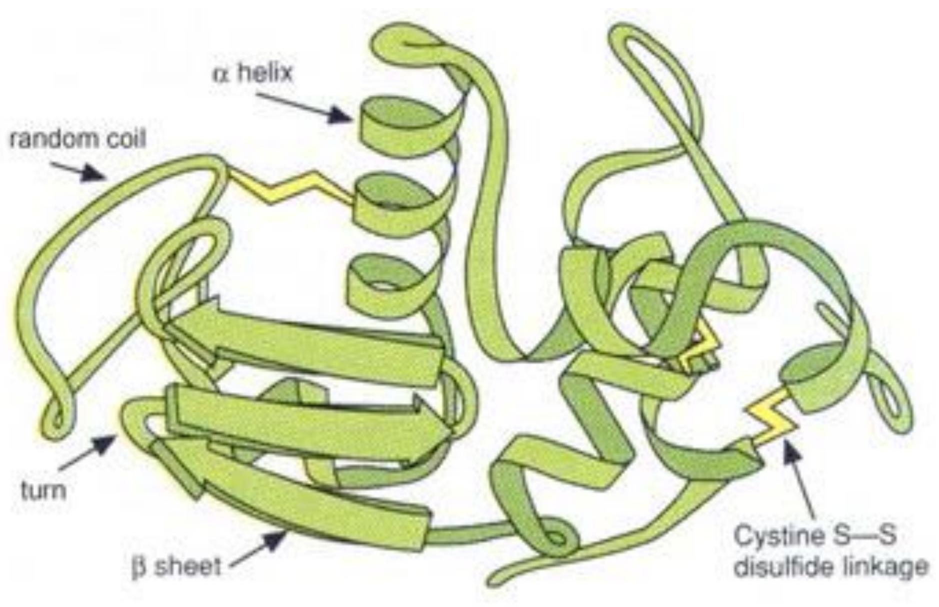
β -sheet (parallel, anti-parallel)



Random coil



http://www.pnas.org/content/101/34/12497/F3.large.jpg



https://getrevising.co.uk/revision-cards/biology_asf212ocr_specification_and_answers

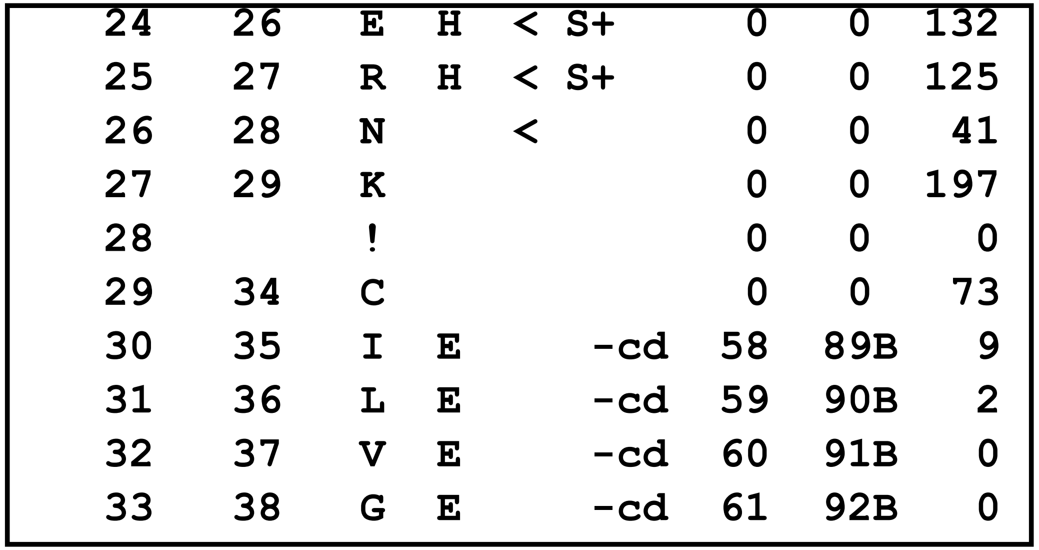
“A **random coil** is a polymer conformation where the monomer subunits are oriented **randomly** while still being bonded to adjacent units.” - Wikipedia

Classification of secondary structure

- Defining features
    - Dihedral angles
    - Hydrogen bonds
    - Geometry
- Assigned manually by experimentalists
- Automatic
    - DSSP (Kabsch & Sander,1983)
    - STRIDE (Frishman & Argos, 1995)
    - Continuum (Andersen et al.)

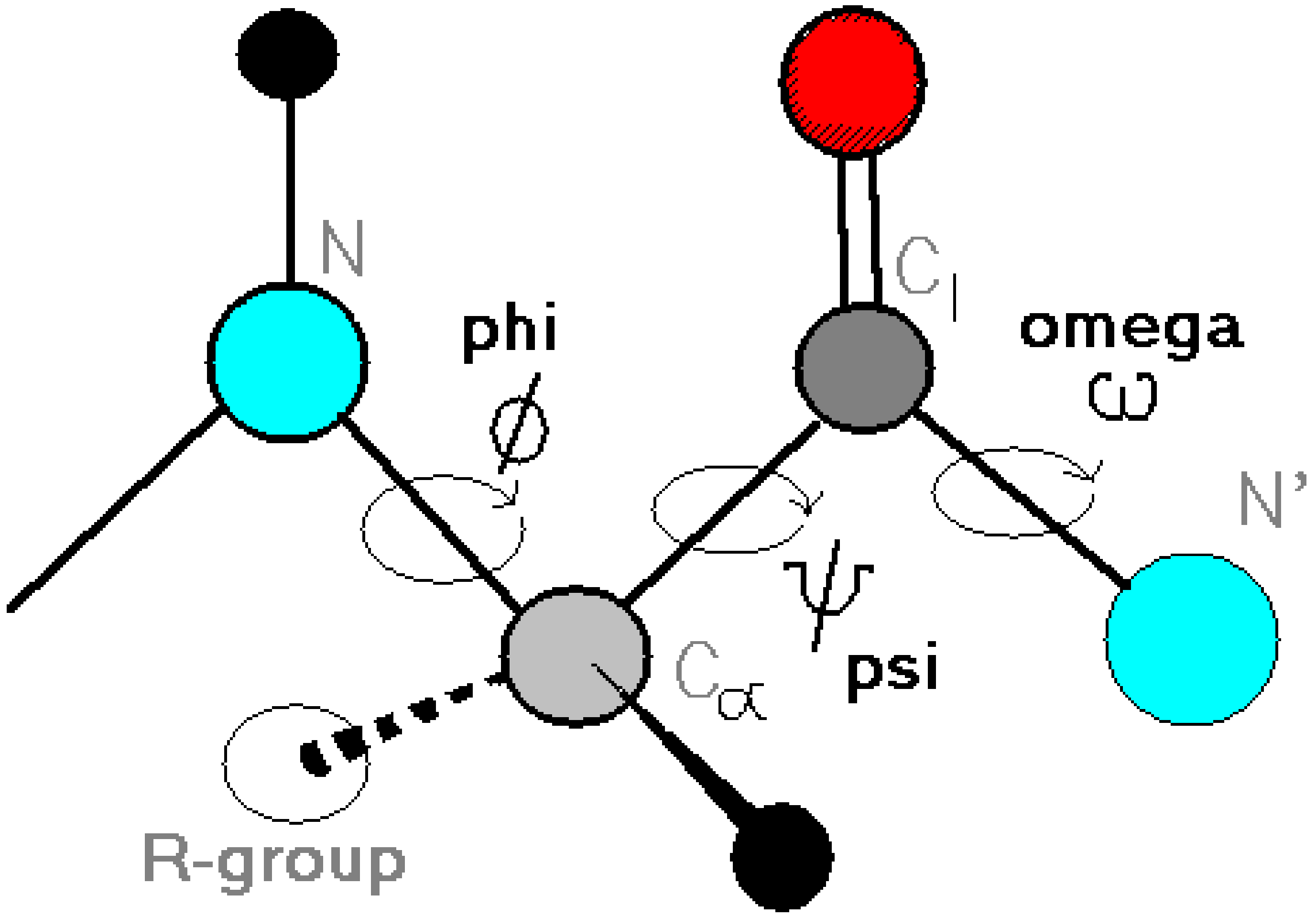
Classification

- Eight states from DSSP
  - H: α–helix
  - G: 3_{10} helix
  - I: π-helix
  - E: β–strand
  - B: bridge
  - T: β–turn
  - S: bend
  - C: coil
- CASP Standard
  - H = (H, G, I), E = (E, B), C = (C, T, S)

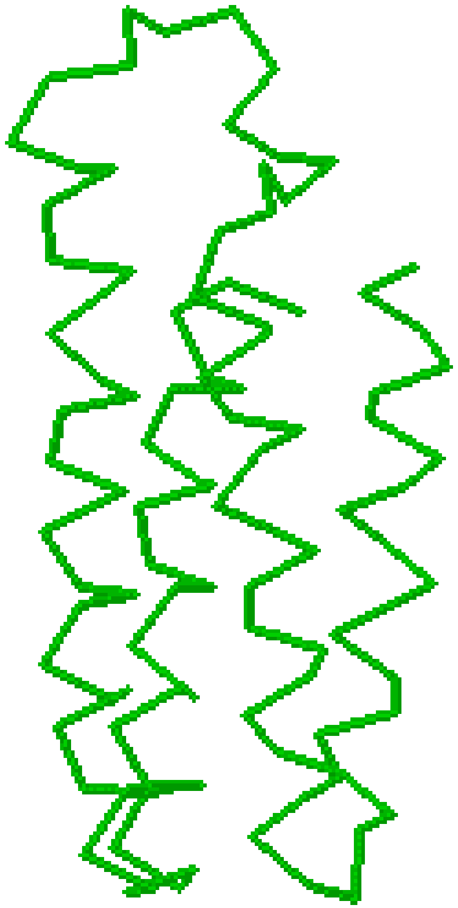


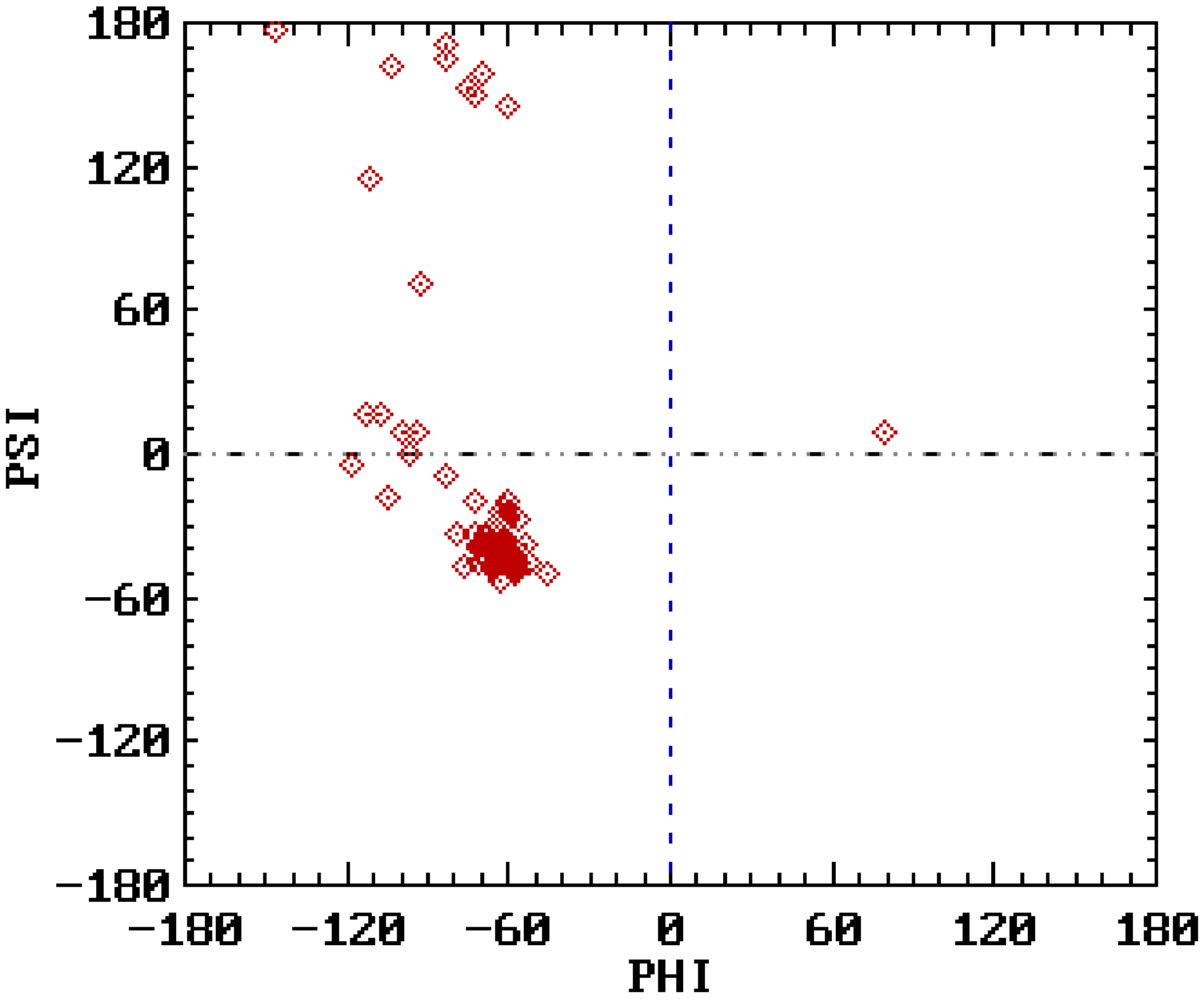
```
   24   26   E  H  < S+        0    0  132
   25   27   R  H  < S+        0    0  125
   26   28   N     <           0    0   41
   27   29   K                 0    0  197
   28        !                 0    0    0
   29   34   C                 0    0   73
   30   35   I  E      -cd    58  89B    9
   31   36   L  E      -cd    59  90B    2
   32   37   V  E      -cd    60  91B    0
   33   38   G  E      -cd    61  92B    0
```

Dihedral angles

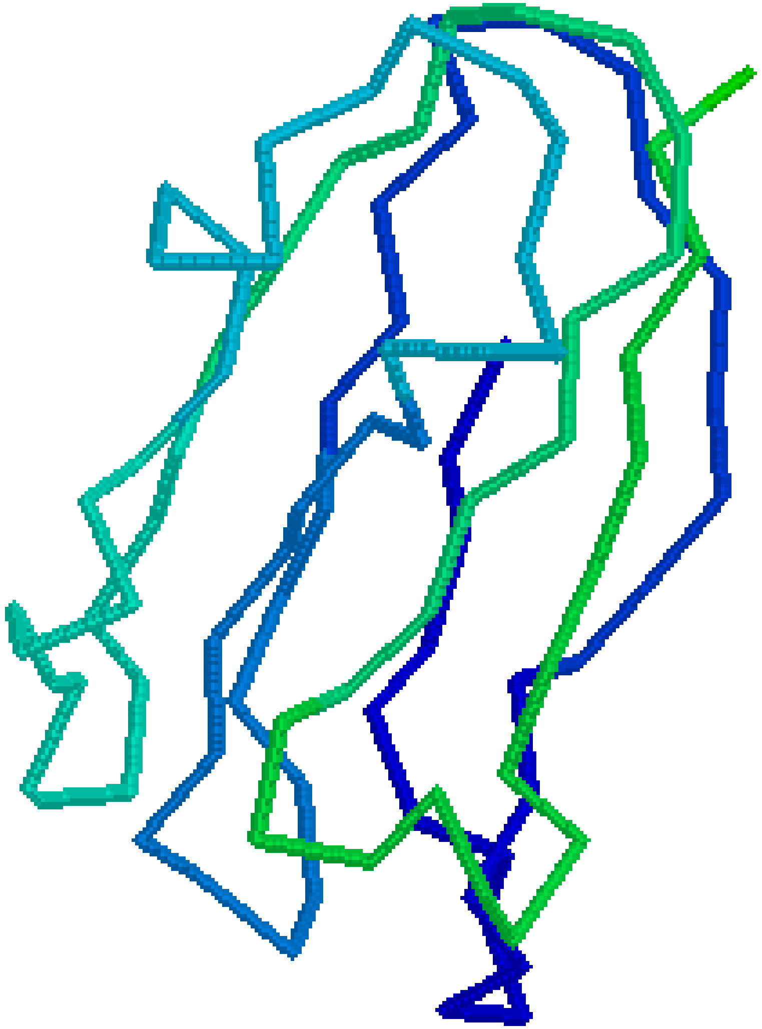


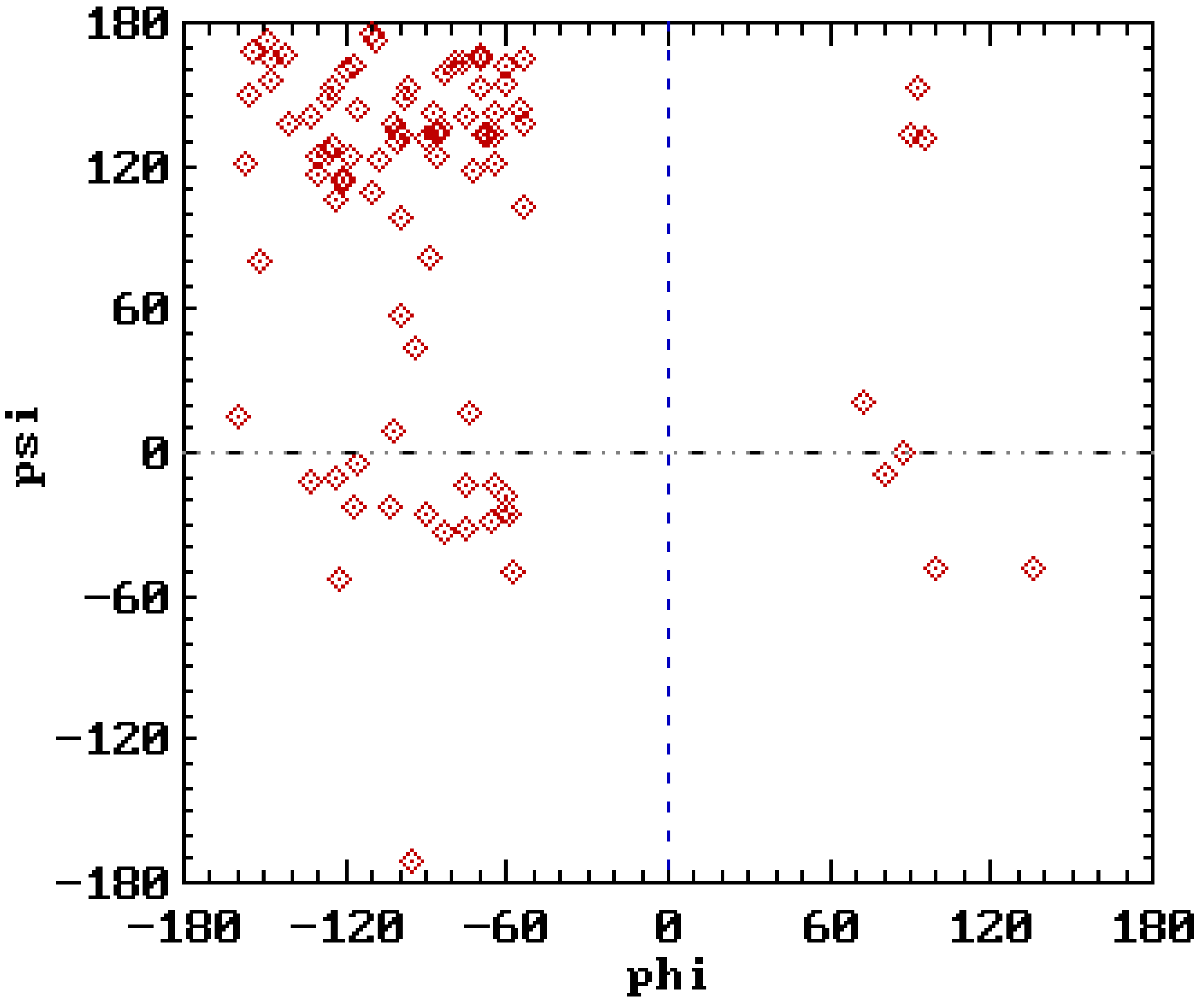
Ramachandran plot (alpha)





Ramachandran plot (beta)





Protein secondary structure prediction

Given a protein sequence (primary structure)

GHWIATRGQLIREAYEDYRHFSSECPFIP

Predict its secondary structure content

(C=Coils H=Alpha Helix E=Beta Strands)

CEEEEECHHHHHHHHHHHCCCHHCCCCCC

Protein secondary structure prediction

- An easier problem than 3D structure prediction (more than 40 years of history).
- Accurate secondary structure prediction can be an important information for the tertiary structure prediction
- Protein function prediction
- Protein classification
- Predicting structural change

Naïve way

- You can always predict protein secondary structure by pairwise sequence alignment

- Similar to the non-coding RNA sequence-structure alignment

- We are going to focus on scenarios where no homology can be detected (no good alignment can be computed)

- *De novo* prediction

Summary of methods

- Statistical method
  - Chou-Fasman method, GOR I-IV
- Nearest neighbors
  - NNSSP, SSPAL
- Neural network
  - PHD, Psi-Pred, J-Pred
- Support vector machine (SVM)
- HMM

Measure

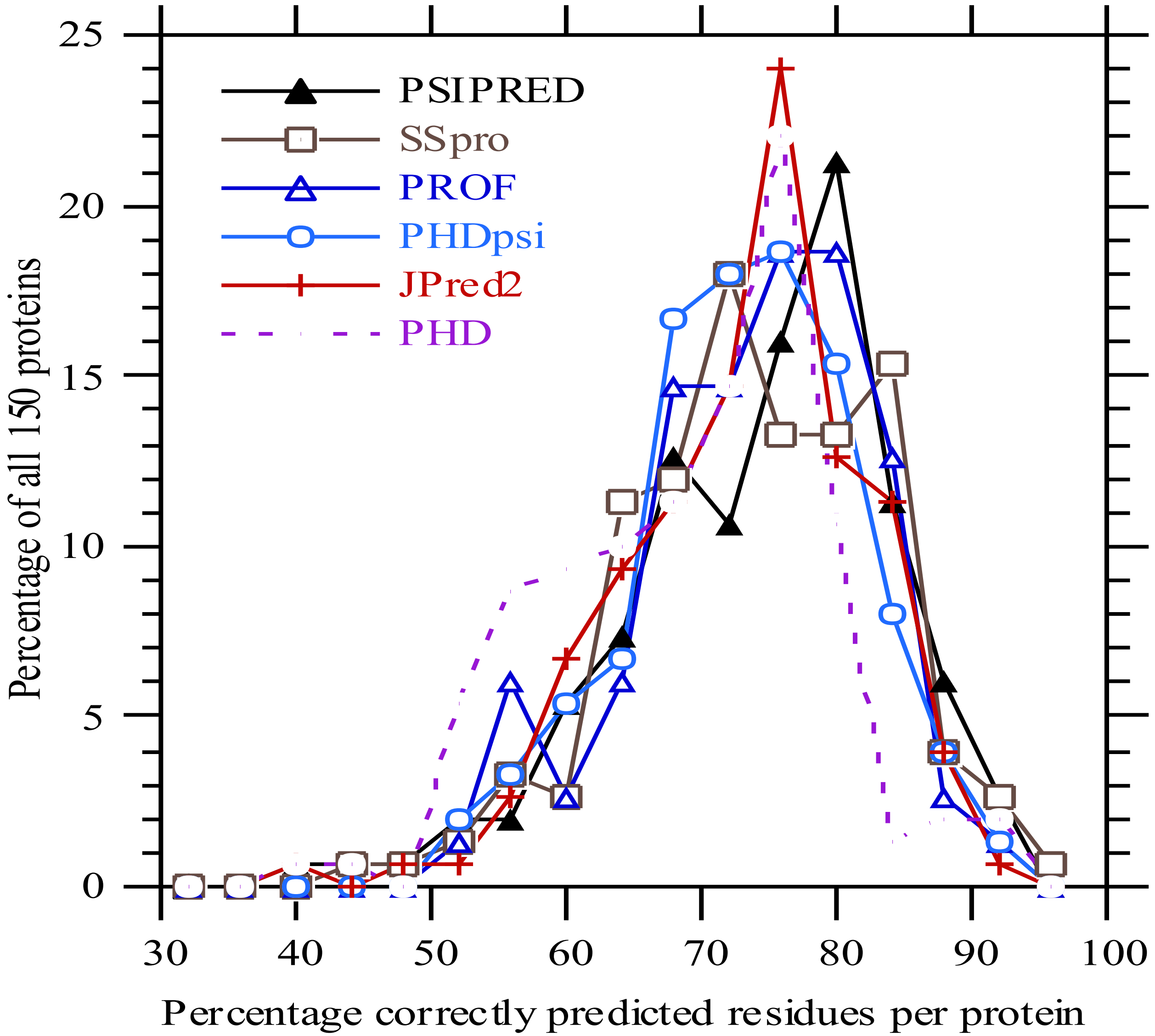
Three-state prediction accuracy: Q_3

$$Q_3 = \frac{\text{correctly predicted residues}}{\text{number of residues}}$$

A prediction of all loop: Q_3 ~ 40%

Accuracy

1974 Chou & Fasman	~50-53%
1978 Garnier	63%
1987 Zvelebil	66%
1988 Qian & Sejnowski	64.3%
1993 Rost & Sander	70.8-72.0%
1997 Frishman & Argos	<75%
1999 Cuff & Barton	72.9%
1999 Jones	76.5%
2000 Petersen et al.	77.9%



Assumptions

- The entire information for forming secondary structure is contained in the primary sequence.
- Side groups of residues will determine structure.
- Examining windows of 13 - 17 residues is sufficient to predict structure.
- Basis for window size selection:
  - α-helices 5 – 40 residues long
  - β-strands 5 – 10 residues long

Chou-Fasman Method

From PDB database, calculate the propensity for a given amino acid to adopt a certain ss-type

$$P_{\alpha}^{i} = \frac{P(\alpha \,|\, aa_i)}{p(\alpha)} = \frac{p(\alpha, aa_i)}{p(\alpha)\, p(aa_i)}$$

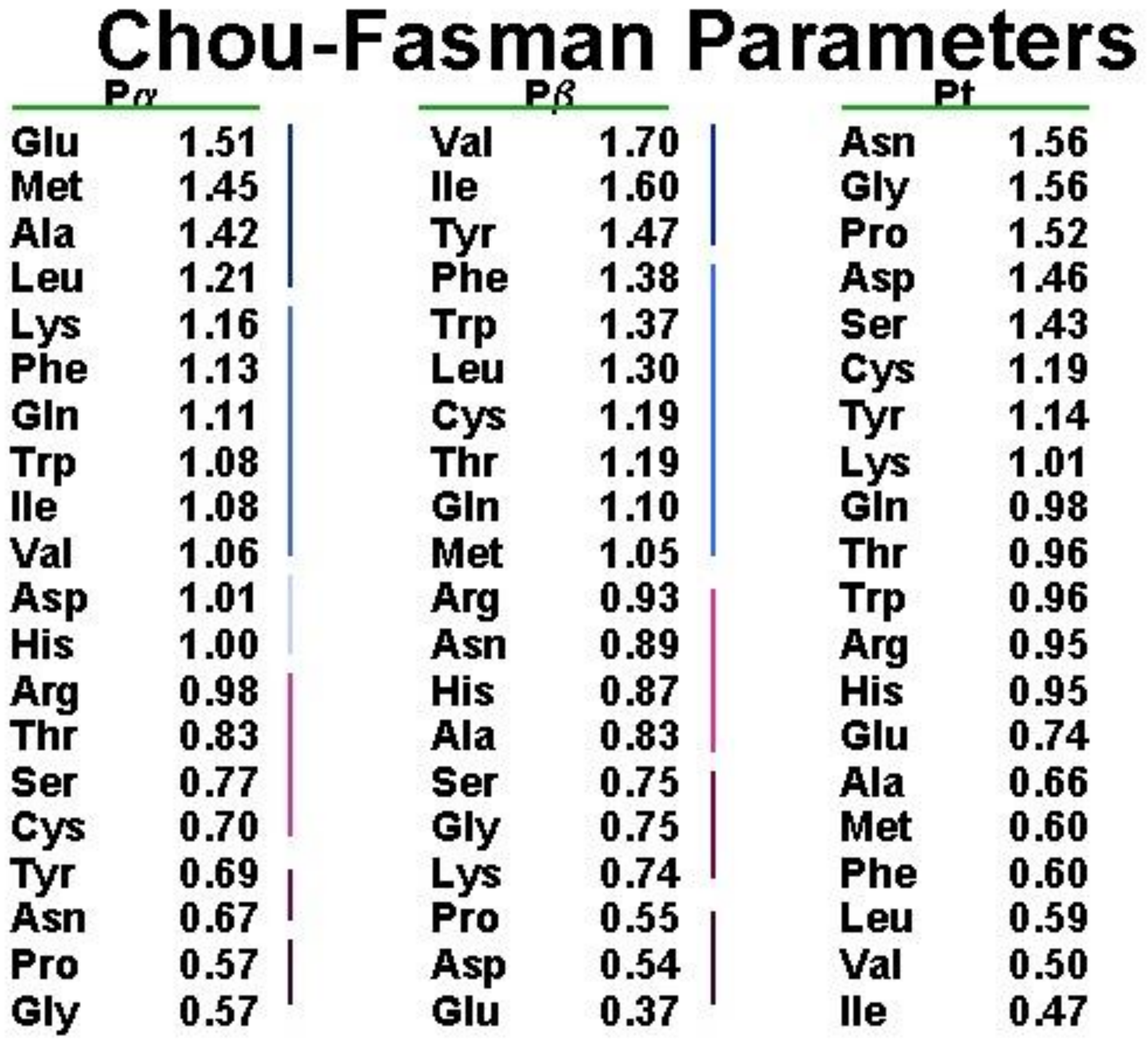
Example:

#Ala=2,000, #residues=20,000, #helix=4,000, #Ala in helix=500

$P(\alpha, aa_i)$ = 500/20,000, $p(\alpha)$ = 4,000/20,000, $p(aa_i)$ = 2,000/20,000

P = 500 / (4,000/10) = 1.25

Chou-Fasman Method



Chou-Fasman Parameters

$P\alpha$		$P\beta$		Pt	
Glu	1.51	Val	1.70	Asn	1.56
Met	1.45	Ile	1.60	Gly	1.56
Ala	1.42	Tyr	1.47	Pro	1.52
Leu	1.21	Phe	1.38	Asp	1.46
Lys	1.16	Trp	1.37	Ser	1.43
Phe	1.13	Leu	1.30	Cys	1.19
Gln	1.11	Cys	1.19	Tyr	1.14
Trp	1.08	Thr	1.19	Lys	1.01
Ile	1.08	Gln	1.10	Gln	0.98
Val	1.06	Met	1.05	Thr	0.96
Asp	1.01	Arg	0.93	Trp	0.96
His	1.00	Asn	0.89	Arg	0.95
Arg	0.98	His	0.87	His	0.95
Thr	0.83	Ala	0.83	Glu	0.74
Ser	0.77	Ser	0.75	Ala	0.66
Cys	0.70	Gly	0.75	Met	0.60
Tyr	0.69	Lys	0.74	Phe	0.60
Asn	0.67	Pro	0.55	Leu	0.59
Pro	0.57	Asp	0.54	Val	0.50
Gly	0.57	Glu	0.37	Ile	0.47

Chou-Fasman Method

Helix, Strand

1. Scan for window of 6 residues where average score > 1 (4 residues for helix and 3 residues for strand)
2. Propagate in both directions until 4 (or 3) residue window with mean propensity < 1
3. Move forward and repeat

Conflict solution

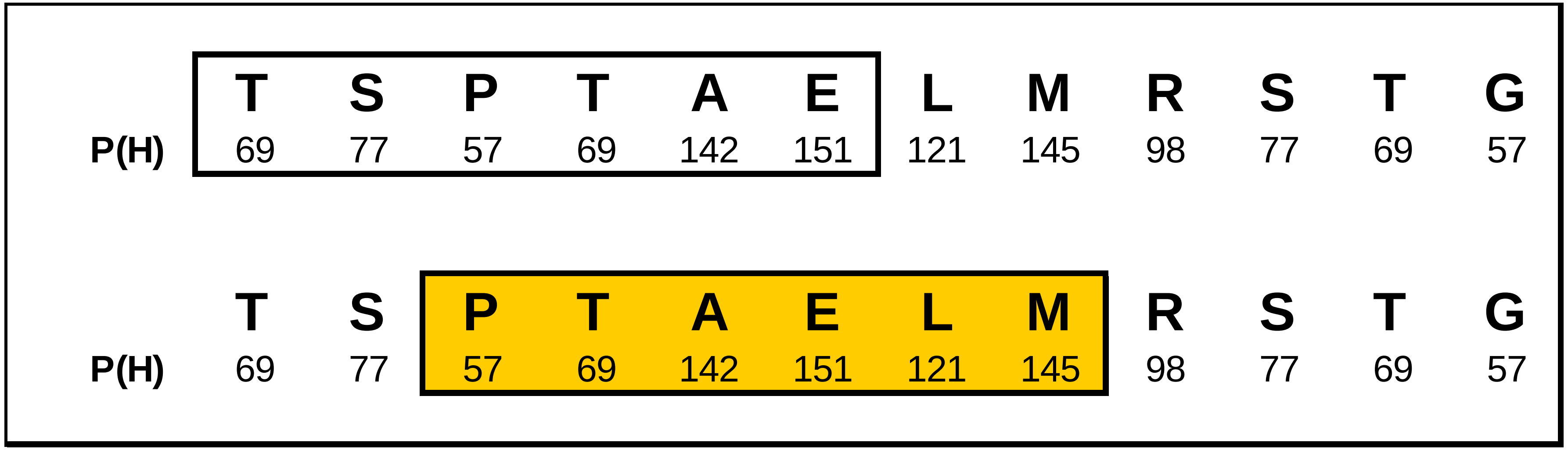
Any region containing overlapping alpha-helical and beta-strand assignments are taken to be helical if the average P(helix) > P(strand). It is a beta strand if the average P(strand) > P(helix).

Accuracy: ~50% → ~60%

GHWIATRGQLIREAYEDYRHFSSECPFIP

Initialization

Identify regions where 4/6 have a P(H) >1.00 "alpha-helix nucleus"

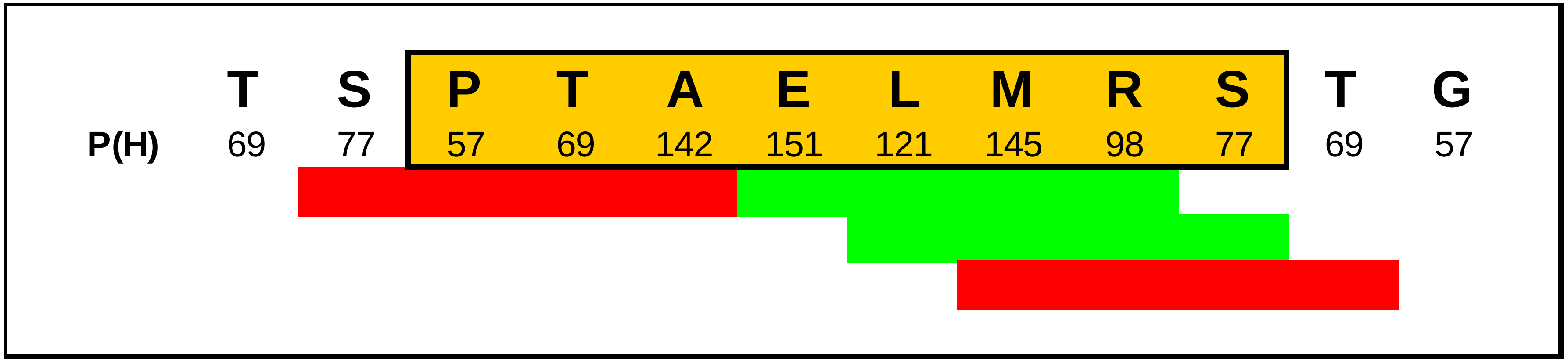


	T	S	P	T	A	E	L	M	R	S	T	G
P(H)	69	77	57	69	142	151	121	145	98	77	69	57

	T	S	P	T	A	E	L	M	R	S	T	G
P(H)	69	77	57	69	142	151	121	145	98	77	69	57

Extension

Extend helix in both directions until a set of four residues have an average P(H) <1.00.



	T	S	P	T	A	E	L	M	R	S	T	G
P(H)	69	77	57	69	142	151	121	145	98	77	69	57

Nearest Neighbor Method

- Predict secondary structure of the central residue of a given segment from homologous segments (neighbors)
  - (i) From database, find some number of the closest sequences to a subsequence defined by a window around the central residue
  - (ii) Compute *K* best non-intersecting local alignments of a query sequence with each sequence.

- Use *max* $(n_{\alpha}, n_{\beta}, n_c)$ for neighbor consensus or *max*$(s_{\alpha}, s_{\beta}, s_c)$ for consensus sequence hits

Environment preference score

Each amino acid has a preference to a specific structural environments.

Structural variables:

- secondary structure, solvent accessibility

Non-redundant protein structure database: FSSP

$$S(i, j) = \log \frac{p(aa_i \mid E_j)}{p(aa_i)} = \log \frac{p(aa_i, E_j)}{p(aa_i)\,p(E_j)}$$

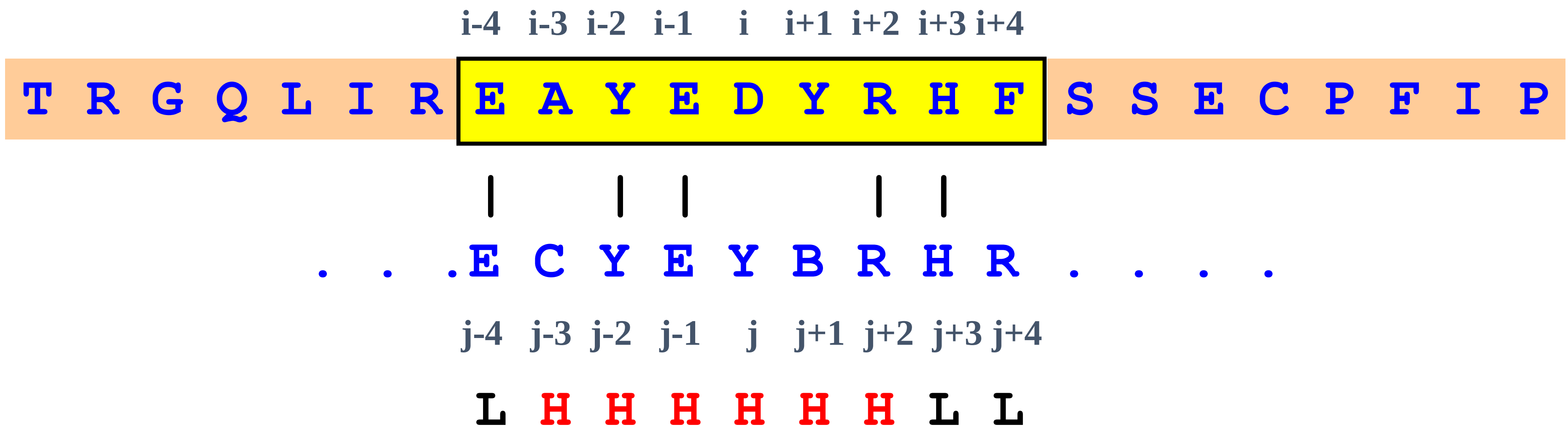
Scoring matrix

	Helix			Sheet			Loop		
	Buried	Inter	Exposed	Buried	Inter	Exposed	Buried	Inter	Exposed
ALA	-0.578	-0.119	-0.160	0.010	0.583	0.921	0.023	0.218	0.368
ARG	0.997	-0.507	-0.488	1.267	-0.345	-0.580	0.930	-0.005	-0.032
ASN	0.819	0.090	-0.007	0.844	0.221	0.046	0.030	-0.322	-0.487
ASP	1.050	0.172	-0.426	1.145	0.322	0.061	0.308	-0.224	-0.541
CYS	-0.360	0.333	1.831	-0.671	0.003	1.216	-0.690	-0.225	1.216
GLN	1.047	-0.294	-0.939	1.452	0.139	-0.555	1.326	0.486	-0.244
GLU	0.670	-0.313	-0.721	0.999	0.031	-0.494	0.845	0.248	-0.144
GLY	0.414	0.932	0.969	0.177	0.565	0.989	-0.562	-0.299	-0.601
HIS	0.479	-0.223	0.136	0.306	-0.343	-0.014	0.019	-0.285	0.051
ILE	-0.551	0.087	1.248	-0.875	-0.182	0.500	-0.166	0.384	1.336
LEU	-0.744	-0.218	0.940	-0.411	0.179	0.900	-0.205	0.169	1.217
LYS	1.863	-0.045	-0.865	2.109	-0.017	-0.901	1.925	0.474	-0.498
MET	-0.641	-0.183	0.779	-0.269	0.197	0.658	-0.228	0.113	0.714
PHE	-0.491	0.057	1.364	-0.649	-0.200	0.776	-0.375	-0.001	1.251
PRO	1.090	0.705	0.236	1.249	0.695	0.145	-0.412	-0.491	-0.641
SER	0.350	0.260	-0.020	0.303	0.058	-0.075	-0.173	-0.210	-0.228
THR	0.291	0.215	0.304	0.156	-0.382	-0.584	-0.012	-0.103	-0.125
TRP	-0.379	-0.363	1.178	-0.270	-0.477	0.682	-0.220	-0.099	1.267
TYR	-0.111	-0.292	0.942	-0.267	-0.691	0.292	-0.015	-0.176	0.946
VAL	-0.374	0.236	1.144	-0.912	-0.334	0.089	-0.030	0.309	0.998

Distance between *k*-mers

Alignment score is the sum of score in a window of length *l*:

$$Score(i, j) = \sum_{k=-l/2}^{l/2} [M(i+k, j+k) + cS(i+k, j+k)]$$



i-4 i-3 i-2 i-1 i i+1 i+2 i+3 i+4

T R G Q L I R E A Y E D Y R H F S S E C P F I P

. . . E C Y E Y B R H R . . . .

j-4 j-3 j-2 j-1 j j+1 j+2 j+3 j+4

L H H H H H H L L

Inference based on neighbors

1	-	L	H	H	H	H	H	H	L	L	-	S_1
2	-	L	L	H	H	H	H	H	L	L	-	S_2
3	-	L	E	E	E	E	E	E	L	L	-	S_3
4	-	L	E	E	E	E	E	E	L	L	-	S_4
n	-	L	L	L	L	E	E	E	E	E	-	S_n
n+1	-	H	H	H	L	L	L	E	E	E	-	S_{n+1}

:

- *max* (n_α, n_β, n_L) or *max* $(\Sigma s_\alpha, \Sigma s_\beta, \Sigma s_L)$

Incorporating evolutionary information

- ❑ “All naturally evolved proteins with more than 35% pairwise identical residues over more than 100 aligned residues have similar structures.”
- ❑ Stability of structure w.r.t. sequence divergence (<12% difference in secondary structure).
- ❑ Position-specific sequence profile, containing crucial information on evolution of protein family, can help secondary structure prediction (increase information content).
- ❑ Gaps rarely occur in helix and strand.
- ❑ ~1.4%/year increase in Q3 due to database growth at the beginning.

Evolution information

- ❑ Sequence-profile alignment.
- ❑ Compare a sequence against protein family.
- ❑ More specific.
- ❑ BLAST vs. PSI-BLAST.
- ❑ Look up PSSM instead of PAM or BLOSUM.

$$Score(i,j)=\sum_{k=-l/2}^{l/2}[PSSM(j+k,i+k)+cS(i+k,j+k)]$$

Achieved accuracy ~75%

PSIPRED (Neuron networks)

- D. Jones, J. Mol. Boil. **292,** 195 (1999).
- Method : Neural network
- Input data : PSSM generated by PSI-BLAST
- Bigger and better sequence database
  - Combining several database and data filtering
- Training and test sets preparation
  - No sequence & structural homologues between training and test sets by PSI-BLAST (mimicking realistic situation).

PSIPRED

- PSI-BLAST (iterative sequence-profile alignment)

- Searching the target sequencing against protein database and generates profile

- The profile contains evolutionary information

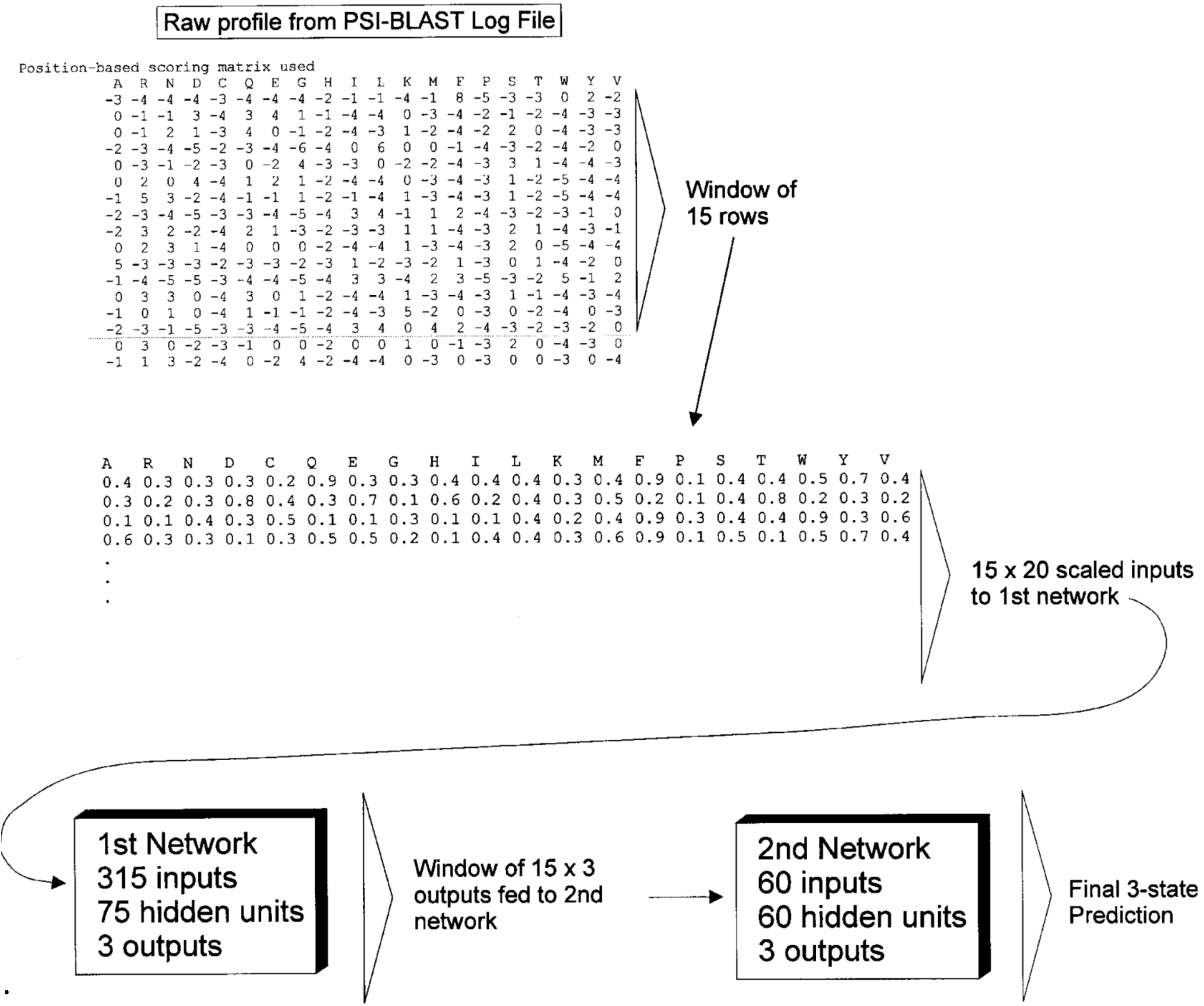
- Use profile of proteins with known secondary structure as training for neuron network

PSIPRED

- A window of 15 amino acid residues was found to be optimal.
- The first input layer comprises 315 input units, divided into 15 groups of 21 units. The extra unit per amino acid is used to indicate where the window spans either the N or C terminus of the protein chain.
- A large hidden layer of 75 units was used for the first network, with another three units making the output layer where the units represent the three-states of secondary structure (helix, strand or coil).
- A second network has an input layer comprising just 60 input units, divided into 15 groups of four. Again the extra input in each group is used to indicate that the window spans a chain terminus.
- A smaller hidden layer of 60 units was used for the second network.

PSIPRED

- Window size = 15
- Two networks
- **Accuracy ~76%**



Raw profile from PSI-BLAST Log File

```
Position-based scoring matrix used
                  A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
                 -3 -4 -4 -4 -3 -4 -4 -4 -2 -1 -1 -4 -1  8 -5 -3 -3  0  2 -2
                  0 -1 -1  3 -4  3  4  1 -1 -4 -4  0 -3 -4 -2 -1 -2 -4 -3 -3
                  0 -1  2  1 -3  4  0 -1 -2 -4 -3  1 -2 -4 -2  2  0 -4 -3 -3
                 -2 -3 -4 -5 -2 -3 -4 -6 -4  0  6  0  0 -1 -4 -3 -2 -4 -2  0
                  0 -3 -1 -2 -3  0 -2  4 -3 -3  0 -2 -2 -4 -3  3  1 -4 -4 -3
                  0  2  0  4 -4  1  2  1 -2 -4 -4  0 -3 -4 -3  1 -2 -5 -4 -4
                 -1  5  3 -2 -4 -1 -1  1 -2 -1 -4  1 -3 -4 -3  1 -2 -5 -4 -4
                 -2 -3 -4 -5 -3 -3 -4 -5 -4  3  4 -1  1  2 -4 -3 -2 -3 -1  0
                 -2  3  2 -2 -4  2  1 -3 -2 -3 -3  1  1 -4 -3  2  1 -4 -3 -1
                  0  2  3  1 -4  0  0  0 -2 -4 -4  1 -3 -4 -3  2  0 -5 -4 -4
                  5 -3 -3 -3 -2 -3 -3 -2 -3  1 -2 -3 -2  1 -3  0  1 -4 -2  0
                 -1 -4 -5 -5 -3 -4 -4 -5 -4  3  3 -4  2  3 -5 -3 -2  5 -1  2
                  0  3  3  0 -4  3  0  1 -2 -4 -4  1 -3 -4 -3  1 -1 -4 -3 -4
                 -1  0  1  0 -4  1 -1 -1 -2 -4 -3  5 -2  0 -3  0 -2 -4  0 -3
                 -2 -3 -1 -5 -3 -3 -4 -5 -4  3  4  0  4  2 -4 -3 -2 -3 -2  0
                  0  3  0 -2 -3 -1  0  0 -2  0  0  1  0 -1 -3  2  0 -4 -3  0
                 -1  1  3 -2 -4  0 -2  4 -2 -4 -4  0 -3  0 -3  0  0 -3  0 -4
```

Window of 15 rows

```
A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
0.4 0.3 0.3 0.3 0.2 0.9 0.3 0.3 0.4 0.4 0.4 0.3 0.4 0.9 0.1 0.4 0.4 0.5 0.7 0.4
0.3 0.2 0.3 0.8 0.4 0.3 0.7 0.1 0.6 0.2 0.4 0.3 0.5 0.2 0.1 0.4 0.8 0.2 0.3 0.2
0.1 0.1 0.4 0.3 0.5 0.1 0.1 0.3 0.1 0.1 0.4 0.2 0.4 0.9 0.3 0.4 0.4 0.9 0.3 0.6
0.6 0.3 0.3 0.1 0.3 0.5 0.5 0.2 0.1 0.4 0.4 0.3 0.6 0.9 0.1 0.5 0.1 0.5 0.7 0.4
.
.
.
```

15 x 20 scaled inputs to 1st network

1st Network
315 inputs
75 hidden units
3 outputs

Window of 15 x 3 outputs fed to 2nd network

2nd Network
60 inputs
60 hidden units
3 outputs

Final 3-state Prediction

D. Jones, J. Mol. Boil. **292,** 195 (1999).

SVM

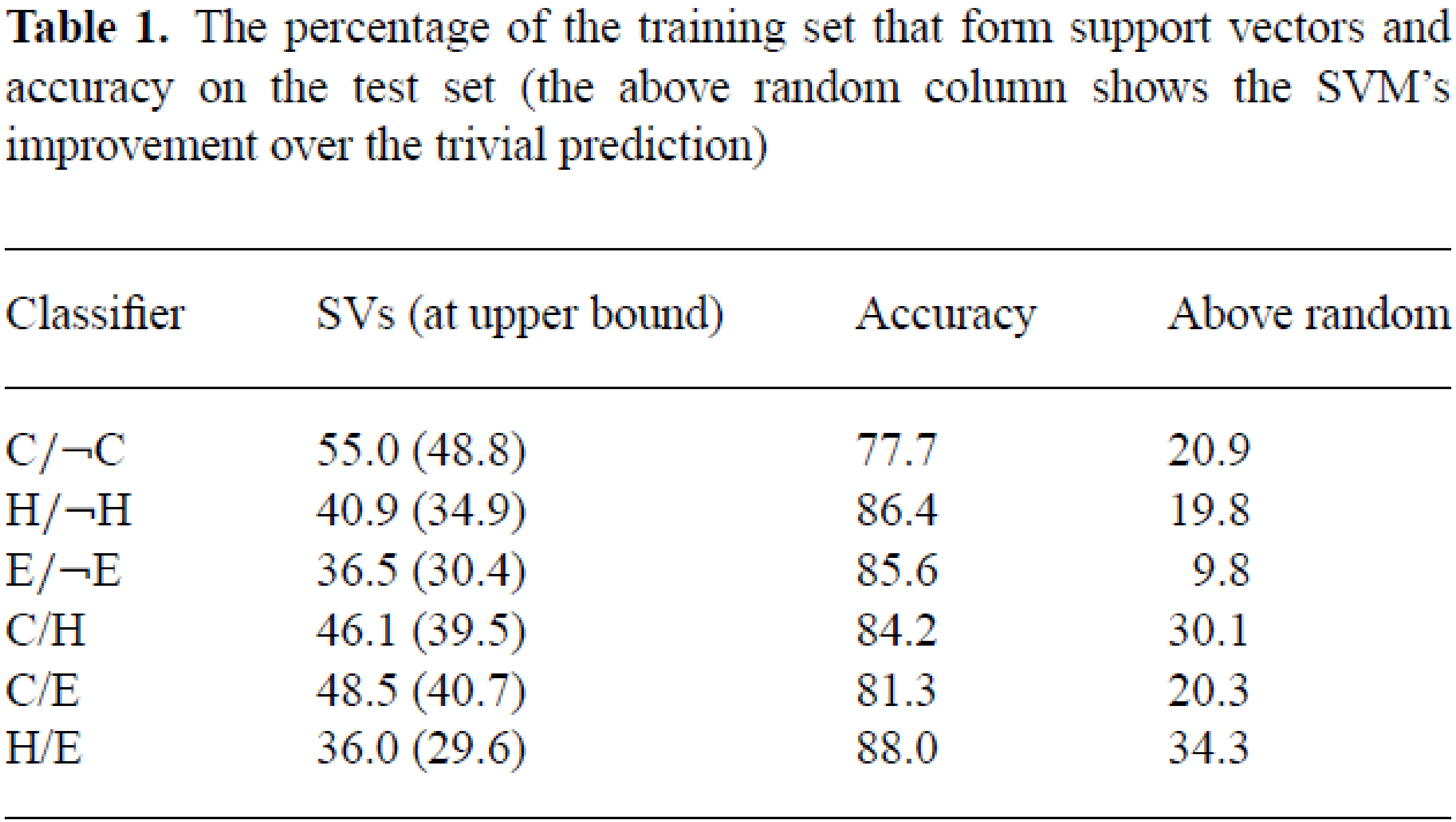


Table 1. The percentage of the training set that form support vectors and accuracy on the test set (the above random column shows the SVM's improvement over the trivial prediction)

Classifier	SVs (at upper bound)	Accuracy	Above random
C/¬C	55.0 (48.8)	77.7	20.9
H/¬H	40.9 (34.9)	86.4	19.8
E/¬E	36.5 (30.4)	85.6	9.8
C/H	46.1 (39.5)	84.2	30.1
C/E	48.5 (40.7)	81.3	20.3
H/E	36.0 (29.6)	88.0	34.3

$$K(\mathbf{x}, \mathbf{z}) = \left(\frac{\mathbf{x} \cdot \mathbf{z} + 1}{50} \right)^2$$

Ward et al. 2003, Bioinformatics

SVM

- The inputs from each sequence appear in the form of a 20 ×M position-specific scoring matrix from three iterations of a PSI-BLAST search, where M is the length of the target sequence. The scoring matrix for a window of 15 positions, centered on the target residue, is used as the input to the SVM.
- In cases where the window extends beyond the protein termini, ‘empty’ attributes are filled with zeros

Ward et al. 2003, Bioinformatics

SVM cont.

Performance ~77%

Ward et al. 2003, Bioinformatics

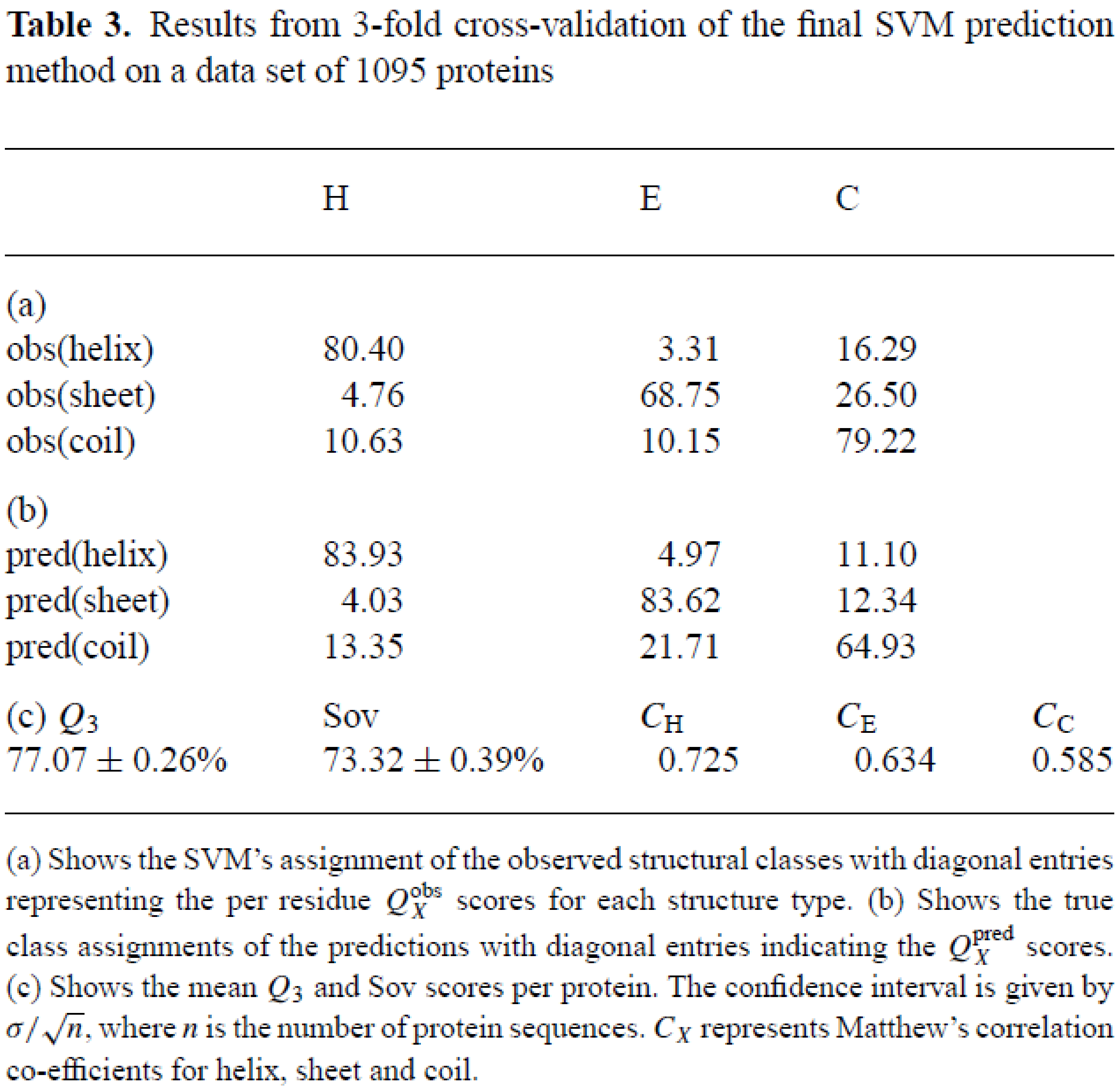
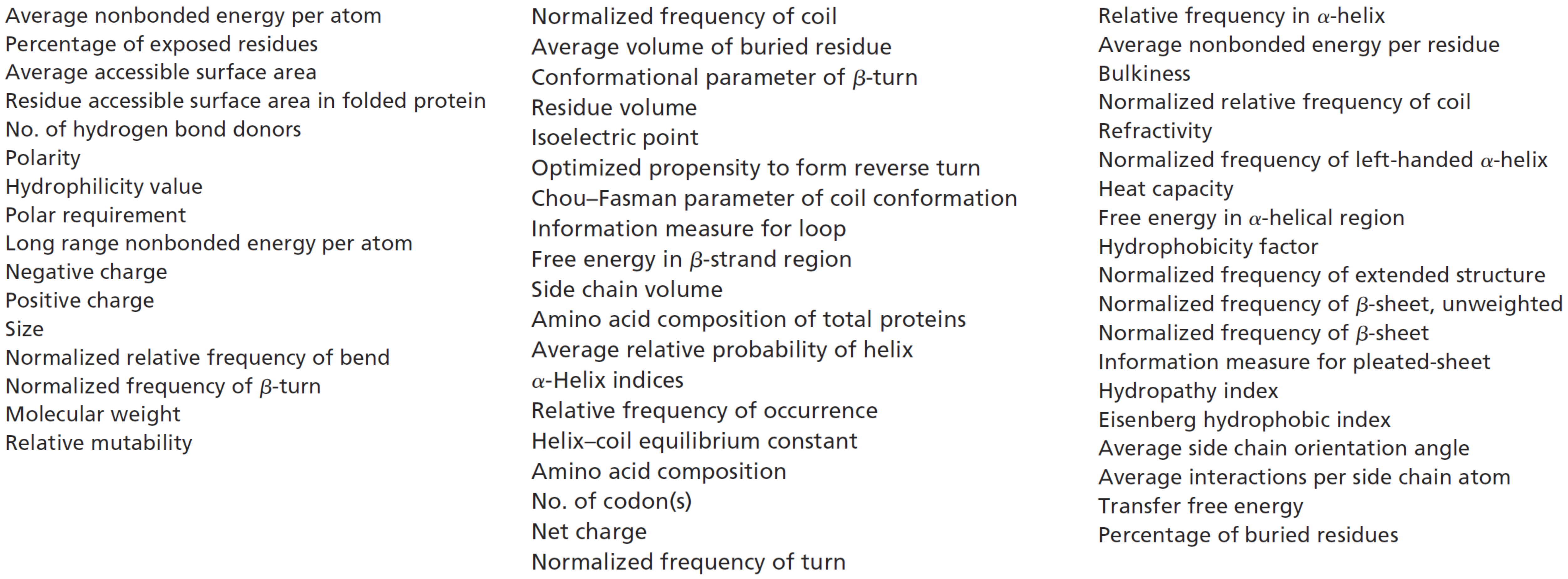


Table 3. Results from 3-fold cross-validation of the final SVM prediction method on a data set of 1095 proteins

	H	E	C	
(a)				
obs(helix)	80.40	3.31	16.29	
obs(sheet)	4.76	68.75	26.50	
obs(coil)	10.63	10.15	79.22	
(b)				
pred(helix)	83.93	4.97	11.10	
pred(sheet)	4.03	83.62	12.34	
pred(coil)	13.35	21.71	64.93	
(c) Q_3	Sov	C_{H}	C_{E}	C_{C}
77.07 ± 0.26%	73.32 ± 0.39%	0.725	0.634	0.585

(a) Shows the SVM's assignment of the observed structural classes with diagonal entries representing the per residue Q_X^{obs} scores for each structure type. (b) Shows the true class assignments of the predictions with diagonal entries indicating the Q_X^{pred} scores. (c) Shows the mean Q_3 and Sov scores per protein. The confidence interval is given by $\sigma/\sqrt{n}$, where n is the number of protein sequences. C_X represents Matthew's correlation co-efficients for helix, sheet and coil.

Sequence features other than PSSM



Average nonbonded energy per atom
Percentage of exposed residues
Average accessible surface area
Residue accessible surface area in folded protein
No. of hydrogen bond donors
Polarity
Hydrophilicity value
Polar requirement
Long range nonbonded energy per atom
Negative charge
Positive charge
Size
Normalized relative frequency of bend
Normalized frequency of β-turn
Molecular weight
Relative mutability

Normalized frequency of coil
Average volume of buried residue
Conformational parameter of β-turn
Residue volume
Isoelectric point
Optimized propensity to form reverse turn
Chou–Fasman parameter of coil conformation
Information measure for loop
Free energy in β-strand region
Side chain volume
Amino acid composition of total proteins
Average relative probability of helix
α-Helix indices
Relative frequency of occurrence
Helix–coil equilibrium constant
Amino acid composition
No. of codon(s)
Net charge
Normalized frequency of turn

Relative frequency in α-helix
Average nonbonded energy per residue
Bulkiness
Normalized relative frequency of coil
Refractivity
Normalized frequency of left-handed α-helix
Heat capacity
Free energy in α-helical region
Hydrophobicity factor
Normalized frequency of extended structure
Normalized frequency of β-sheet, unweighted
Normalized frequency of β-sheet
Information measure for pleated-sheet
Hydropathy index
Eisenberg hydrophobic index
Average side chain orientation angle
Average interactions per side chain atom
Transfer free energy
Percentage of buried residues

Atchley et al., 2005, PNAS

Deep learning network

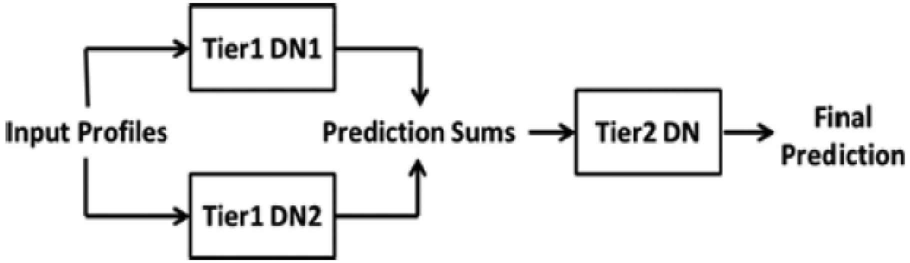
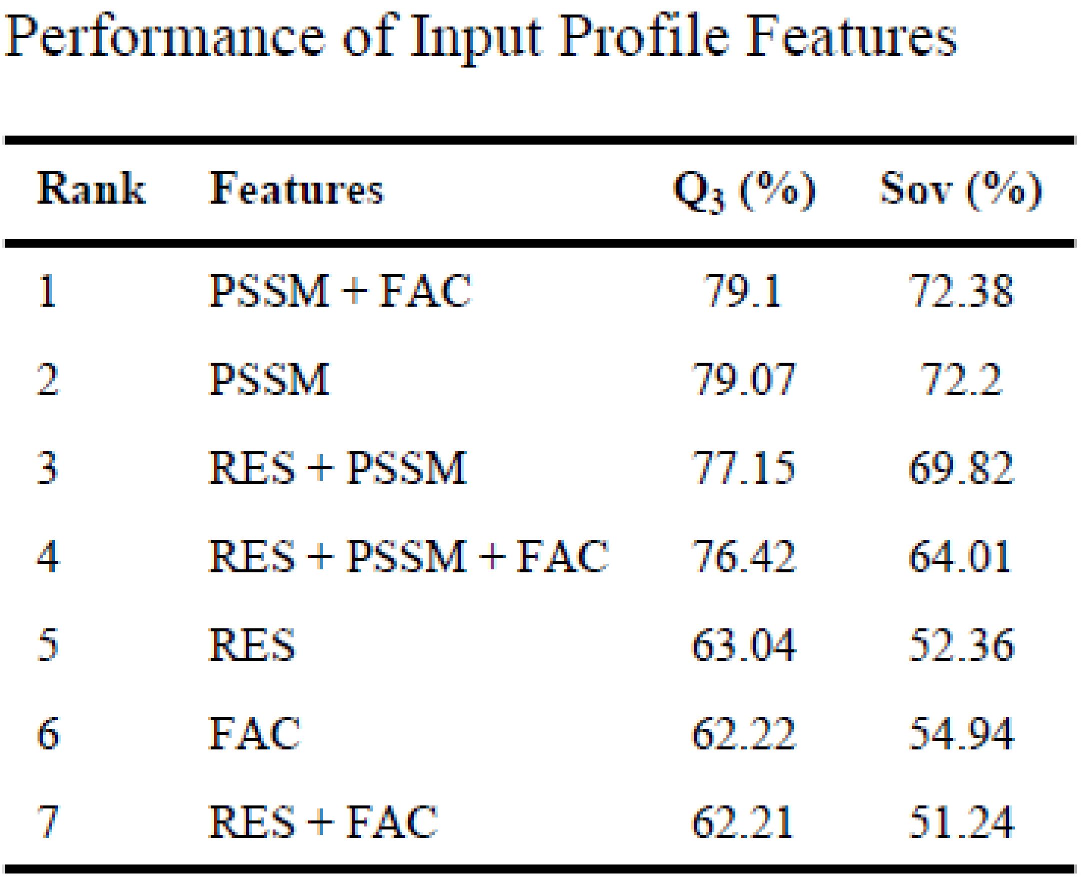


Fig. 2.
Block diagram showing the DNSS secondary structure prediction workflow.



Performance of Input Profile Features

Rank	Features	Q_3 (%)	Sov (%)
1	PSSM + FAC	79.1	72.38
2	PSSM	79.07	72.2
3	RES + PSSM	77.15	69.82
4	RES + PSSM + FAC	76.42	64.01
5	RES	63.04	52.36
6	FAC	62.22	54.94
7	RES + FAC	62.21	51.24

Spencer et al. 2015, ACM TCBB

Summary

- “However, secondary structure prediction has failed to appreciably improve upon the state-of-the-art 80% accuracy. As noted, recent methods have improved upon this accuracy by a small margin, but we must question how important it is to tweak secondary structure prediction tools to generate such a small improvement in accuracy. It is looking more and more like secondary structure prediction scores may not significantly improve until the discovery of features that can benefit the prediction process over and above the contribution of the sequence profiles alone.”

Spencer et al. 2015, ACM TCBB