Bioinformatics III
Structural Bioinformatics and Genome Analysis
Part Protein Secondary Structure Prediction

Sepp Hochreiter

Institute of Bioinformatics
Johannes Kepler University, Linz, Austria

Chapter 4

Protein Secondary Structure Prediction

Protein Secondary Structure Prediction

Traditional field in bioinformatics

First applications of neural networks (Sejnowski, NETalk)

Introduction

protein folding: first secondary structures are formed then 3D

secondary structure: most local information in primary sequence

prediction of secondary structure is important for
- design of de novo proteins
- homology detection
- model building methods (e.g. “Modeller”)
- determining structures with 2D NMR
- first step of ab initio structure prediction

However secondary structure: influenced by 3D interactions

Example:
amino acid subsequences forming a β-sheet
inserted at another place → α-helix

exchange of single amino acids can change the secondary structure

Assigning Secondary Structure to Measured Structures

training examples: from known structures

How to extract secondary structure?

DSSP

"Dictionary of Secondary Structure of Proteins" (DSSP)
Kabsch and Sander 1983 http://swift.cmbi.ru.nl/gv/dssp/

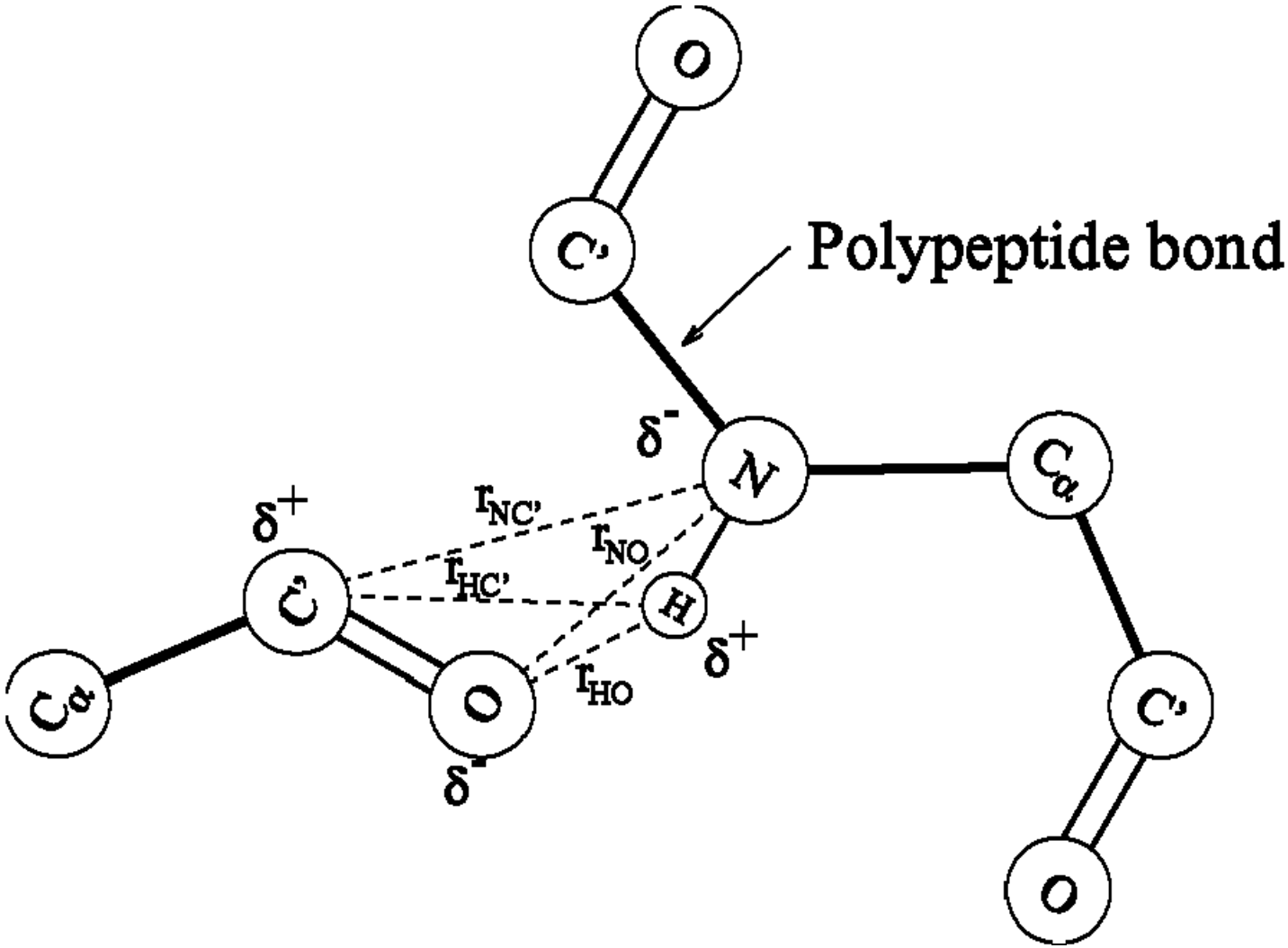
assigns sheet and helical structures based on
backbone-backbone hydrogen bonds

hydrogen bond: if bond energy is below -0.5 kcal/mol
from a Coulomb approximation

$$E = f\ \delta^+\ \delta^-\ \left(\frac{1}{r_{\mathrm{NO}}} + \frac{1}{r_{\mathrm{HC'}}} + \frac{1}{r_{\mathrm{HO}}} + \frac{1}{r_{\mathrm{NC'}}}\right)$$

f = 332 Å kcal/e^2 : normalizing constant
δ^+ and δ^-: polar charges in electron charges e
distances: next slide

DSSP



DSSP

unbroken structures

overlap: α-helix priority

α-helix: "H" → consecutive amino acids i → (i+4) hydrogen bonds
ends with two consecutive i → (i+4) hydrogen bonds

3_{10}-helix: "G" → amino acids i → (i+3) hydrogen bonds

π-helix: "I" → amino acids i → (i+5) hydrogen bonds

Turns: "T" → single helix hydrogen bonds

β-sheets: "E" → 2 hydrogen bonds or surrounded by hydrogen bonds

β-bridge: "B" → single amino acids with hydrogen bonds

bend: "S"

DSSP

Symbol	Meaning
H	α-helix
G	3_{10}-helix
I	π-helix
T	turn
E	β-sheet
B	β-bridge
S	bend
_	unassigned

Table 1: The secondary structure symbols assigned by DSSP.

DSSP


```
PDB:1crn
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC    N-H-->O    O-->H-N    N-H-->O    O-->H-N
 ....
   15   15    V  H  >< S+     0   0   99   -4,-1.7   3,-1.3    2,-0.2   -2,-0.2
   16   16    C  H 3<>S+     0   0   18   -4,-2.5   5,-0.8    1,-0.3   -2,-0.2
   17   17    R  H 3X>S+     0   0   94   -4,-2.0   3,-1.6    1,-0.2   -1,-0.3
   18   18    L  T <<5S+     0   0  144   -3,-1.3   -1,-0.2   -4,-0.6   -2,-0.2
   19   19    P  T 3 5S-     0   0  107    0, 0.0   -1,-0.3    0, 0.0   -2,-0.1
   20   20    G  T < 5 +     0   0   53   -3,-1.6   -3,-0.2    1,-0.2   -2,-0.1
   21   21    T      <  -     0   0   37   -5,-0.8   -1,-0.2    1,-0.1    5,-0.1
   22   22    P    >>   -     0   0   81    0, 0.0    4,-2.2    0, 0.0    3,-0.7
   23   23    E  H 3> S+     0   0   70    1,-0.2    4,-2.5    2,-0.2    5,-0.1
   24   24    A  H 3> S+     0   0   63    1,-0.2    4,-1.7    2,-0.2   -1,-0.2
   25   25    I  H <> S+     0   0   99   -3,-0.7    4,-1.8    2,-0.2   -1,-0.2
   26   26    C  H  X S+     0   0    0   -4,-2.2    4,-1.9    2,-0.2    6,-0.4
   27   27    A  H  X S+     0   0   12   -4,-2.5    4,-2.7   -5,-0.2    5,-0.5
   28   28    T  H  < S+     0   0  120   -4,-1.7   -1,-0.2    1,-0.2   -2,-0.2
   29   29    Y  H  < S+     0   0  176   -4,-1.8   -1,-0.2   -5,-0.2   -2,-0.2
   30   30    T  H  < S-     0   0   24   -4,-1.9   -2,-0.2   -3,-0.2   -3,-0.2
   31   31    G  S  < S+     0   0   35   -4,-2.7   -3,-0.2    1,-0.4   -4,-0.1
   32   32    b      -       0   0    5   -5,-0.5   -1,-0.4   -6,-0.4    2,-0.3
   33   33    I  E     -A    3   0A  51  -30,-2.8  -30,-2.4   -3,-0.1    2,-0.5
   34   34    I  E     -A    2   0A  78   -2,-0.3  -32,-0.2  -32,-0.2    3, 0.0
 ....
```

β-sheet label
β-bridge 2 label
β-bridge 1 label
chirality assignments
bend assignments
π-helix hydrogen bonds
α-helix hydrogen bonds
3_{10}-helix hydrogen bonds
secondary structure synopsis

Output example for DSSP

DSSP

3 classes:

Symbol	conventional	newer
H	H	H
G	H	C
I	H	C
T	C	C
E	E	E
B	E	C
S	C	C
_	C	C

Table 1: The 8 secondary classes mapped to 3 classes.

DSSP


```
    HEADER    HYDROLASE   (SERINE PROTEINASE)         17-MAY-76   1EST
...
240  1  4  4  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS,
NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)                .
10891.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
162 67.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  ;
0  0.0   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES
84 35.0   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES;
...
26 10.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2)
30 12.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3)
10  4.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4)
...
#  RESIDUE AA STRUCTURE BP1 BP2  ACC    N-H-->O  O-->H-N  N-H-->O  O-->H-N
2   17   V  B 3   +A  182   0A   8  180,-2.5 180,-1.9   1,-0.2 134,-0.1

TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
-0.776 360.0   8.1 -84.5 125.5  -14.7   34.4   34.8


....;....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
|    |   |  | xxxxxxxxxxxxxxxxxxx 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
```

- ★ "# RESIDUE": two columns of residue numbers.
- ★ "AA": one letter amino acid code, lower case for SS-bridge CYS.
- ★ "S" (first column in STRUCTURE block): compromise summary of secondary structure,
- ★ "BP1 BP2": residue number of first and second bridge partner followed by one letter sheet label
- ★ "ACC": number of water molecules in contact with this residue *10. or residue water exposed surface in Angstrom^2.
- ★ "N-H–¿O" etc.: hydrogen bonds; e.g. -3,-1.4 means: if this residue is residue i then N-H of I is h-bonded to C=O of I-3 with an electrostatic H-bond energy of -1.4 kcal/mol. There are two columns for each type of H-bond, to allow for bifurcated H-bonds
- ★ "TCO": cosine of angle between C=O of residue I and C=O of residue I-1. For alpha-helices, TCO is near +1, for beta-sheets TCO is near -1. Not used for structure definition.
- ★ "KAPPA": virtual bond angle (bend angle) defined by the three C-alpha atoms of residues I-2,I,I+2. Used to define bend (structure code 'S').
- ★ "ALPHA": virtual torsion angle (dihedral angle) defined by the four C-alpha atoms of residues I-1,I,I+1,I+2. Used to define chirality (structure code '+' or '-').
- ★ "PHI PSI": IUPAC peptide backbone torsion angles
- ★ "X-CA Y-CA Z-CA": echo of C-alpha atom coordinates

STRIDE

STRuctural IDEntification method (STRIDE)
http://www.embl-heidelberg.de/argos/stride/stride_info.html

empirical hydrogen bond energy:

$$E = E_r \; E_t \; E_p$$

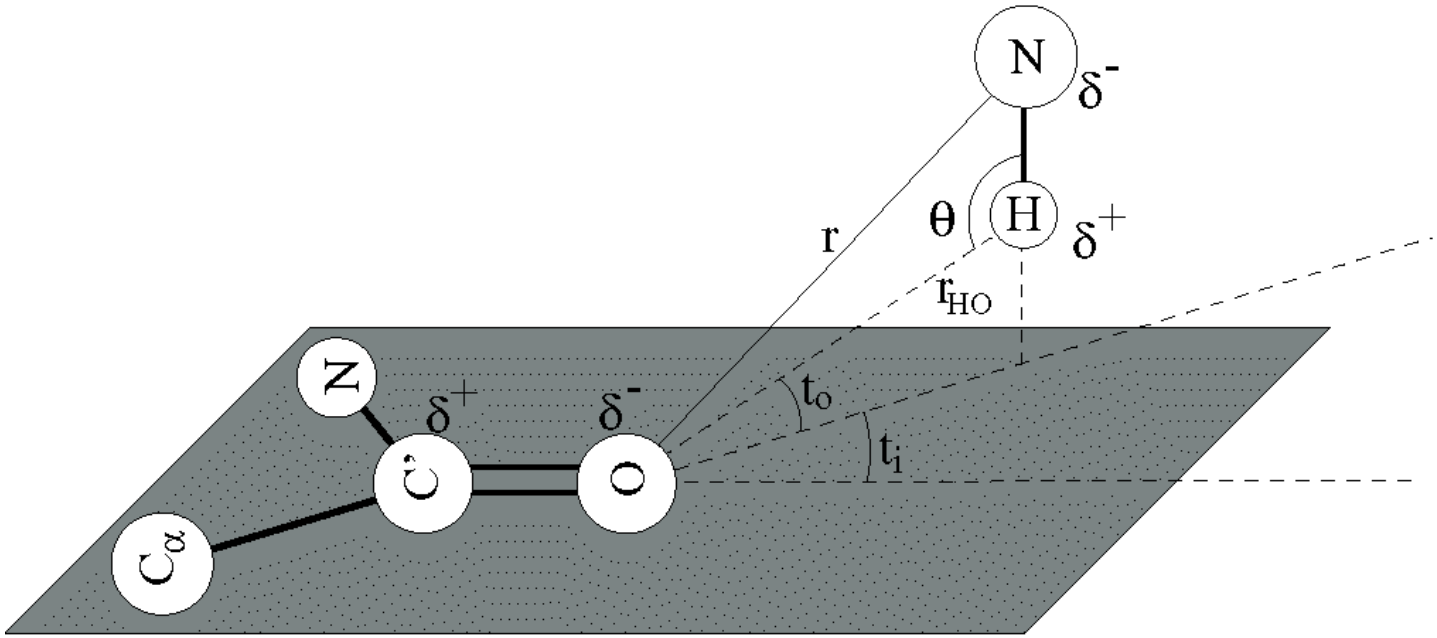
$$E_r = -2.8 \text{ kcal/mol} \left(\frac{43^6}{r_{\text{NO}}^6} - \frac{33^8}{r_{\text{NO}}^8} \right)$$

$$E_p = \cos^2 \theta$$

$$E_t = \begin{cases} (0.9 + 0.1 \sin(2\, t_i)) \cos t_o & \text{for} \quad 0° < t_i \leq 90° \\ K_1 \left(K_2 - \cos^2(t_i) \right) \cos t_o & \text{for} \quad 90° < t_i \leq 110° \\ 0 & \text{for} \quad 110° < t_i \end{cases}$$

energy E_r is similar to to the Lennard-Jones potential
(optimal distances of 3 Å for the backbone hydrogen bond)

STRIDE



STRIDE

STRIDE uses additionally to the energy the phi-psi torsion angles

→ Ramachandran plots

STRIDE agree better with experts than DSSP

Symbol	Meaning
H	α-helix
G	3_{10}-helix
I	π-helix
T	turn
E	β-sheet
B	β-bridge
C	unassigned

Table 1: The secondary structure symbols assigned by STRIDE.

STRIDE

example for the STRIDE output

```
PDB:1crn
REM  |---Residue---|    |--Structure--|   |-Phi-|   |-Psi-|  |-Area-|      1CRN
....
ASG  VAL -   15   15    H    AlphaHelix    -69.24    -41.22      93.8      1CRN
ASG  CYS -   16   16    H    AlphaHelix    -56.67    -36.00      18.4      1CRN
ASG  ARG -   17   17    H    AlphaHelix    -77.07    -16.13      94.1      1CRN
ASG  LEU -   18   18    H    AlphaHelix    -53.21    -46.17     143.0      1CRN
ASG  PRO -   19   19    C         Coil     -77.19     -7.60     108.9      1CRN
ASG  GLY -   20   20    C         Coil     106.26      7.31      52.1      1CRN
ASG  THR -   21   21    C         Coil     -52.67    136.34      38.4      1CRN
ASG  PRO -   22   22    C         Coil     -56.98    146.62      81.9      1CRN
ASG  GLU -   23   23    H    AlphaHelix    -56.41    -36.19      60.9      1CRN
ASG  ALA -   24   24    H    AlphaHelix    -63.43    -34.86      61.3      1CRN
ASG  ILE -   25   25    H    AlphaHelix    -74.77    -37.89      98.2      1CRN
ASG  CYS -   26   26    H    AlphaHelix    -64.95    -31.69       0.0      1CRN
ASG  ALA -   27   27    H    AlphaHelix    -62.04    -54.03      11.6      1CRN
ASG  THR -   28   28    H    AlphaHelix    -68.78    -25.49     121.1      1CRN
ASG  TYR -   29   29    H    AlphaHelix    -67.59    -36.30     174.0      1CRN
ASG  THR -   30   30    H    AlphaHelix   -108.96    -18.47      23.4      1CRN
ASG  GLY -   31   31    C         Coil      91.82     -3.07      36.1      1CRN
ASG  CYS -   32   32    C         Coil     -69.52    164.38       4.6      1CRN
ASG  ILE -   33   33    E       Strand    -129.76    157.03      51.0      1CRN
ASG  ILE -   34   34    E       Strand    -111.56    129.59      78.0      1CRN
....
```

DEFINE and P-Curve

DEFINE: C_a-coordinate match with the optimal secondary structure

first perfect matches are found and then elongated (BLAST)

Problems appear with sheets: bend, curvature

P-Curve: differential geometry to calculate a helicoidal axis
Assignments: pattern matching

Prediction of Secondary Structure

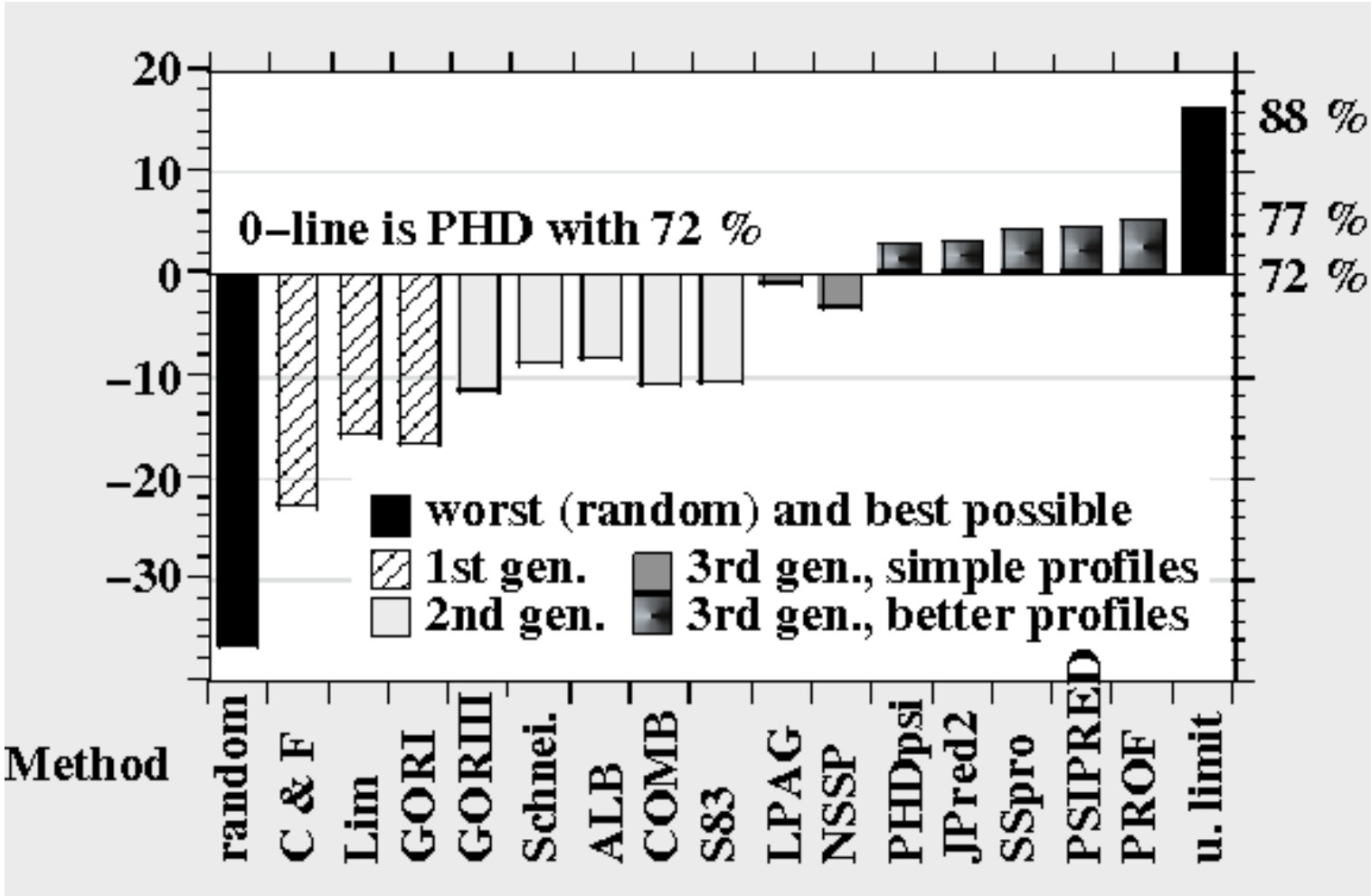
3 generations according to Burkhard Rost:

1. **Generation**: single residue statistics (70's)
   best performance: 60 %
   Examples:Chou-Fasman, GORI, Lim's approach

2. **Generation**: window over the current position (80's)
   best performance: 65 %
   Examples: GORIII, ALB, Schneider's approach

3. **Generation**: profile or position specific scoring matrix (PSSM) 90's
   best performance: 78 %
   Examples: PHD, Jpred2, SSPro and Porter, PSIPRED, PROF

Prediction of Secondary Structure



Chou-Fasman Method

Chou-Fasman method 1978:

Propensity: likelihood ratio of amino acid a and structure s

$$P_s(a) \qquad s = \alpha, \beta, t$$

$$P_s(a) = \frac{p(a,s)}{p(a)\ p(s)}$$

$p(a,s)$ estimated by the number of amino acid a in structure s divided by the number of all amino acids in the data base

$p(a)$ estimated by the number of amino acid a divided by the number of all amino acids in the data base

$p(s)$ estimated by the number of all amino acids in structure s divided by the number of all amino acids in the data base

Chou-Fasman Method

secondary structure: start a nucleation

- α-helices if four out of six residues have $P_\alpha > 1.03$
- β-strands if three out of five residues have $P_\beta > 1.00$

nucleation elongated until mean of 4 AA propensity is below threshold

Higher average propensity wins

Turns: four residues → probability $p(a \mid t, i)$ amino acid at i in turn
probabilities multiplied (independence) → joint probability

first probability must be larger than α-helix and β-strand
second probability larger than $7.5\ 10^{-5}$

Performance: 50-60% of accuracy

random guessing 33%

GOR Methods

GOR (Garnier-Osguthorpe-Robson): statistical principles

probability a in special SS depends on neighbors

frequency matrix $\boldsymbol{F}^s$ of 17 residue window for SS s

$F^s_{a,j}$ frequency of amino acid a at the jth position for SS s

$$P_s(a_l) = \sum_{j=l-8}^{l+8} F_{a_j,j}$$

maximal value → SS

GOR Methods

Garnier 1996: information theory

likelihood ratio: $P_s(a) = \frac{p(s \mid a)}{p(s)} \approx \frac{f_{s,a}}{f_s}$

mutual information between residue a and structure s

$$I(s,a) = H(s) + H(a) - H(s,a)$$

$$H(x) = -\sum_x p(x) \log p(x)$$

$$I(s,a) = \sum_{s,a} p(s,a) \log \frac{p(s,a)}{p(s)\ p(a)} = \mathbf{E}_{(s,a)} \log \frac{p(s,a)}{p(s)\ p(a)}$$

local information

$$\log \frac{p(s,a)}{p(s)\ p(a)} = \log \frac{p(s \mid a)}{p(s)}$$

GOR Methods

difference of the local information

$$I_{\text{loc}}(\Delta s; a) = \log \frac{p(s \mid a)}{p(s)} - \log \frac{p(\neg s \mid a)}{p(\neg s)} =$$

$$\log \frac{p(s \mid a)}{p(s)} - \log \frac{1 - p(s \mid a)}{1 - p(s)} =$$

$$\log \frac{p(s \mid a)}{1 - p(s \mid a)} + \log \frac{1 - p(s)}{p(s)}$$

$$I_{\text{loc}}(\Delta s; a) = \log \frac{f_{s,a}}{1 - f_{s,a}} + \log \frac{1 - f_s}{f_s}$$

Exponentiating:

$$\frac{p(s \mid a)}{1 - p(s \mid a)} = \frac{p(s)}{1 - p(s)} \exp\left(- I_{\text{loc}}(\Delta s; a)\right)$$

$I_{\text{loc}}(\Delta s; a)$ for 17 positions and probability ratios multiplied together

GOR Methods

single position specific influence of amino acids on SS

GORIII 1987: probability of SS conditioned on amino acid pairs
→ assuming less independence

conditioned on n amino acids requires to estimate
20^n variables $p(s \mid a_1, \ldots, a_n)$
BUT: data sets are not large enough to estimate these values

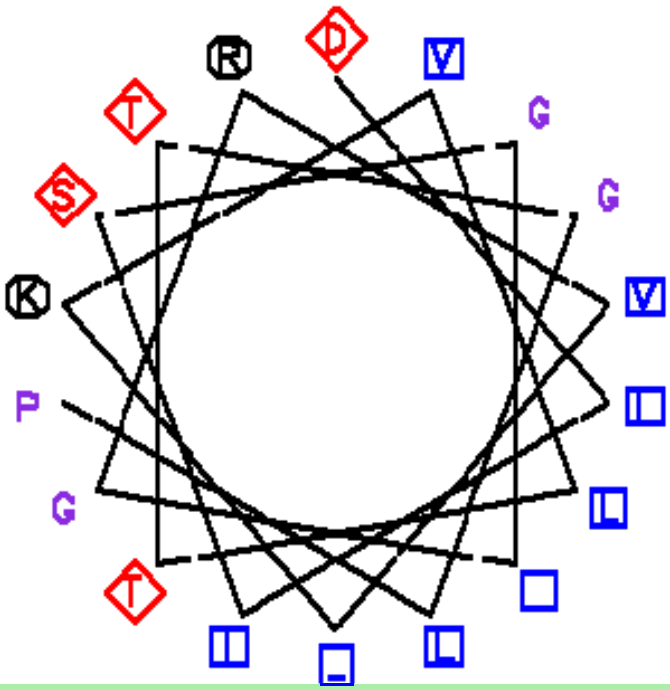
Lim's Method

stereochemical rules for prediction the secondary structure

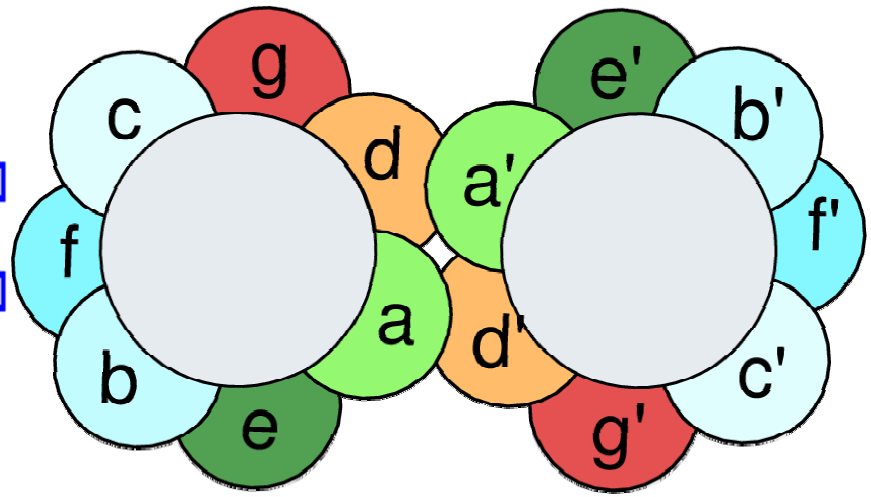
α-helix: hydrophobic side (faces internally of the protein)

biochemical insights → high explanatory

Helical wheel



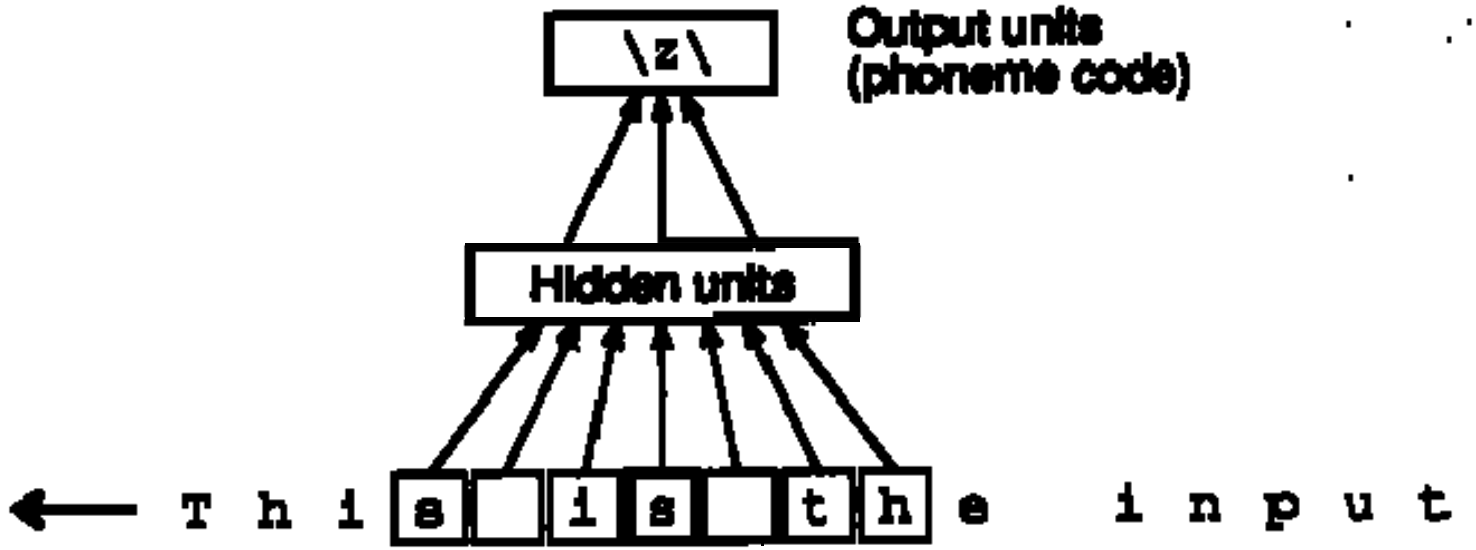
Coiled-coil



Neural Networks

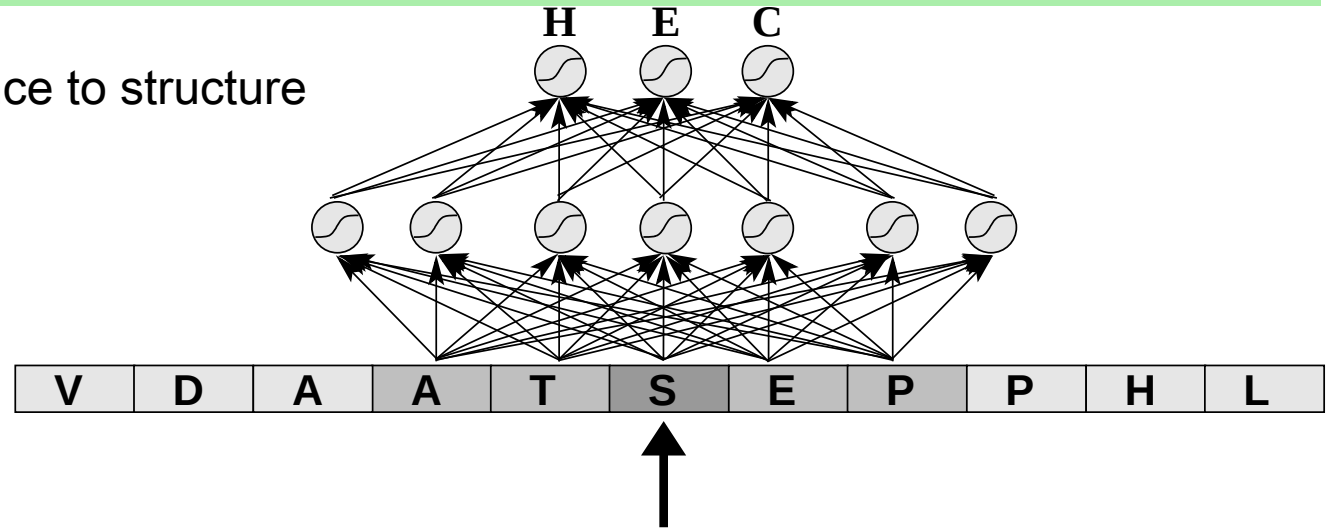
1988 Qian and Sejnowski with NETTalk architecture

64.3% accuracy

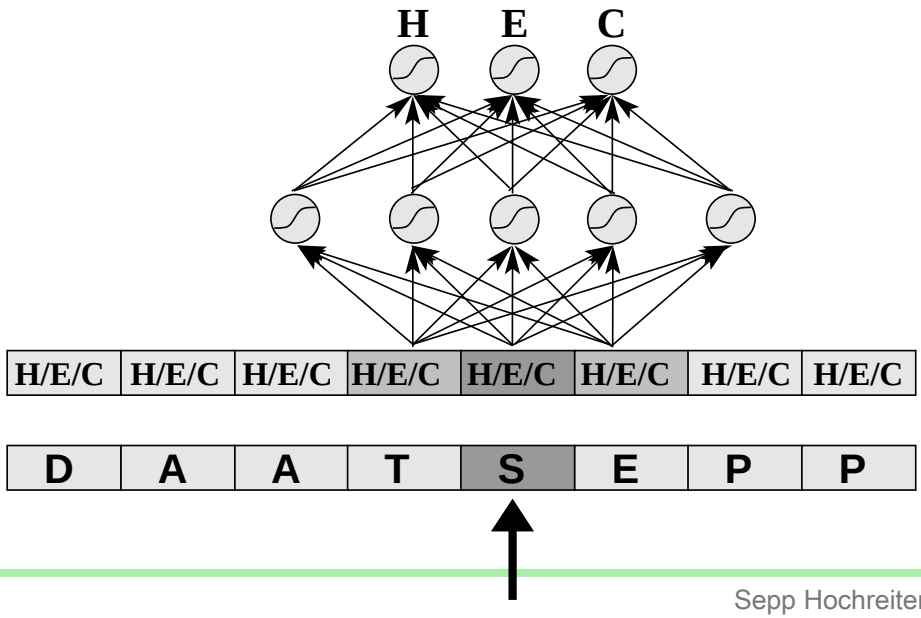


Neural Networks

Sequence to structure



Structure to structure



PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro

1993 “Profile network from HeiDelberg” (PHD) introduced by Burkart Rost was a breakthrough: accuracy 70.2 %

important novelty:

profiles (multiple alignments) and position specific scoring matrices (PSSMs) instead of the primary sequence

→ averaged over many sequences which are very similar to the query sequence

→ averaged over the same structure

Helix and or β-sheet regions: better recognized because average gives hydrophobic or hydrophilic regions

PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro

Another important fact: long range information is included

similar local PSSM → similar sequences are used for alignment
→ similar global structure → long range interactions fit

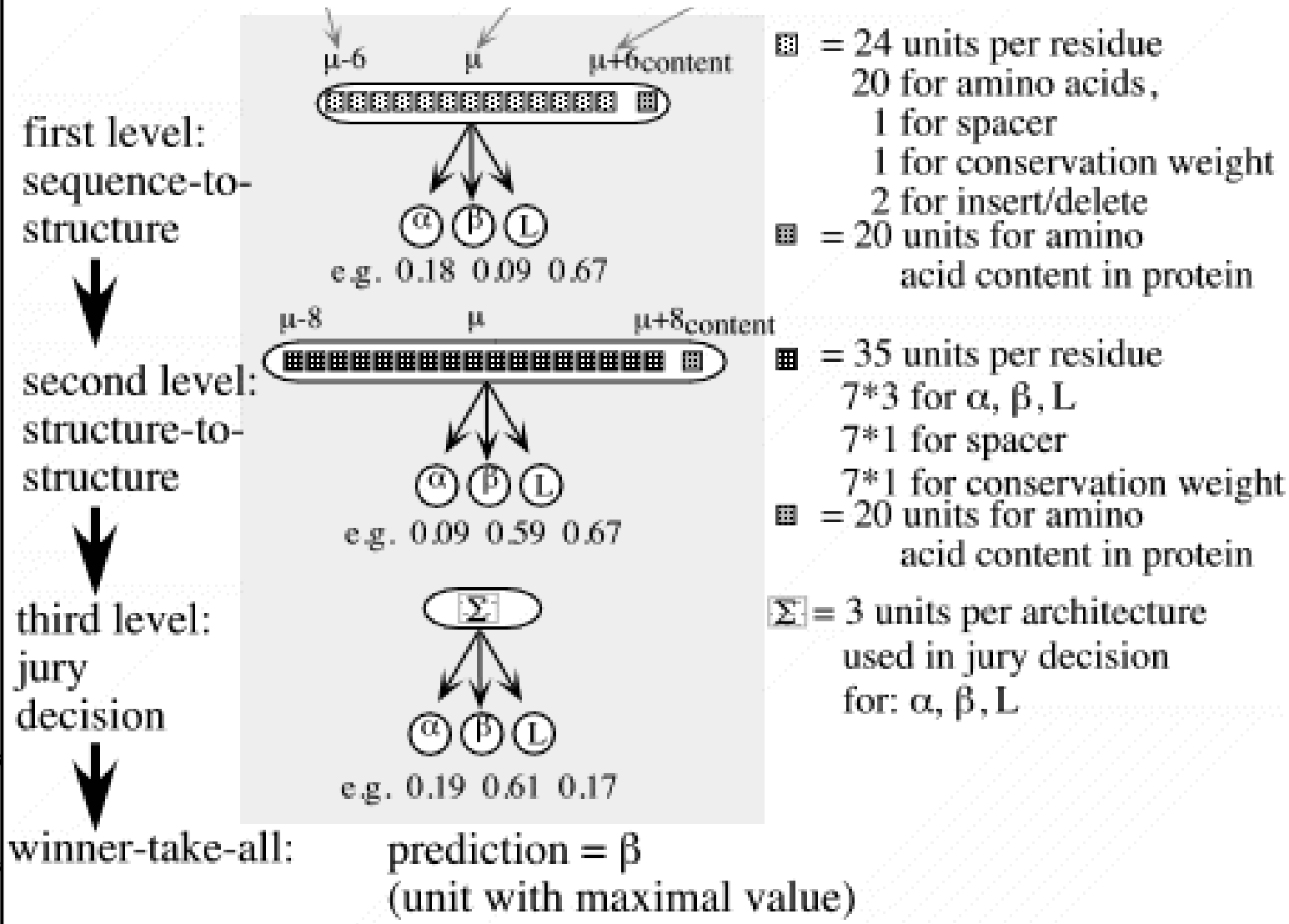
E.g.: strand needs partner → global structure ensures partner

3 levels: sequence-to-structure network
structure-to-structure network
HHHHHCHHHH → HHHHHHHHHH
voting procedure
correct for biases like underestimating β-sheets

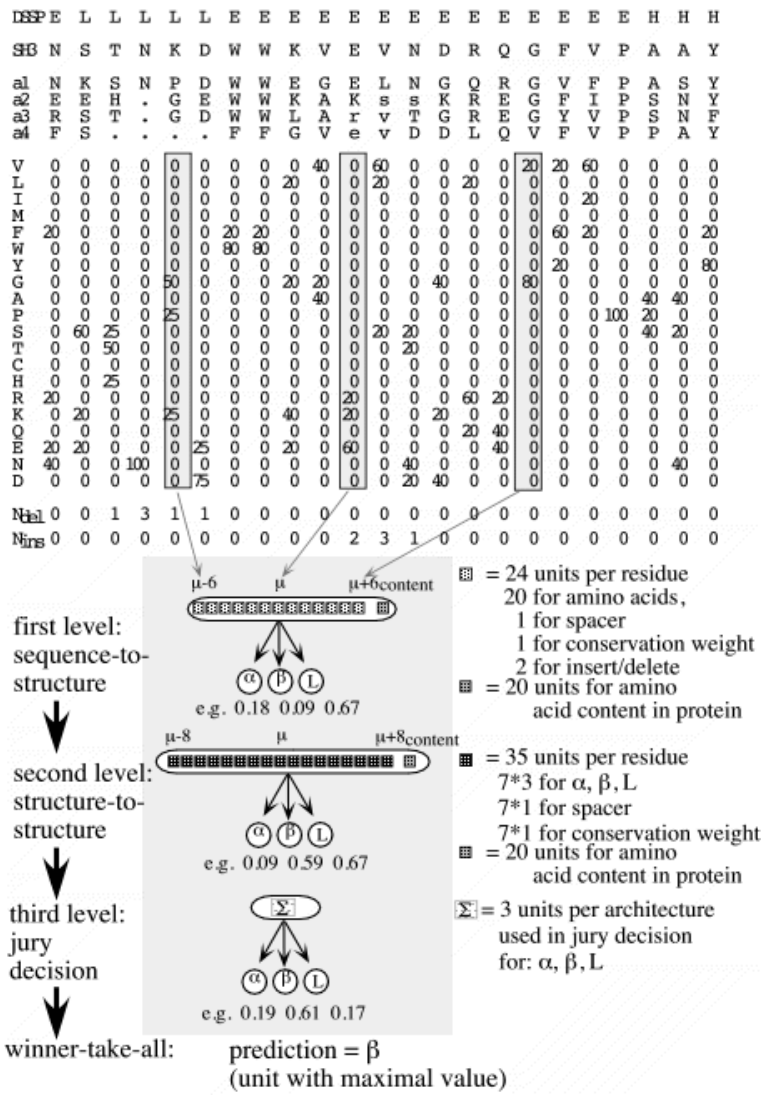
PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro

	DSSP	E	L	L	L	L	L	E	E	E	E	E	E	E	E	E	E	E	E	E	H	H	H
	SH3	N	S	T	N	K	D	W	W	K	V	E	V	N	D	R	Q	G	F	V	P	A	A	Y
	a1	N	K	S	N	P	D	W	W	E	G	E	L	N	G	Q	R	G	V	F	P	A	S	Y
	a2	E	E	H	.	G	E	W	W	K	A	K	s	s	K	R	E	G	F	I	P	S	N	Y
	a3	R	S	T	.	G	D	W	W	L	A	r	v	T	G	R	E	G	Y	V	P	S	N	F
	a4	F	S	.	.	.	.	F	F	G	V	e	v	D	D	L	Q	V	F	V	P	P	A	Y
	V	0	0	0	0	0	0	0	0	0	40	0	60	0	0	0	0	20	20	60	0	0	0	0
	L	0	0	0	0	0	0	0	0	20	0	0	20	0	0	20	0	0	0	0	0	0	0	0
	I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	F	20	0	0	0	0	0	20	20	0	0	0	0	0	0	0	0	0	60	20	0	0	0	20
	W	0	0	0	0	0	0	80	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	80
	G	0	0	0	0	50	0	0	0	20	20	0	0	0	40	0	0	80	0	0	0	0	0	0
	A	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	40	40	0
	P	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	20	0	0
	S	0	60	25	0	0	0	0	0	0	0	0	20	20	0	0	0	0	0	0	0	40	20	0
	T	0	0	50	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	H	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	R	20	0	0	0	0	0	0	0	0	0	20	0	0	0	60	20	0	0	0	0	0	0	0
	K	0	20	0	0	25	0	0	0	40	0	20	0	0	20	0	0	0	0	0	0	0	0	0
	Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	40	0	0	0	0	0	0	0
	E	20	20	0	0	0	25	0	0	20	0	60	0	0	0	0	40	0	0	0	0	0	0	0
	N	40	0	0	100	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	40	0
	D	0	0	0	0	0	75	0	0	0	0	0	0	20	40	0	0	0	0	0	0	0	0	0
	Nhel	0	0	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Nins	0	0	0	0	0	0	0	0	0	0	2	3	1	0	0	0	0	0	0	0	0	0	0

PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro



PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro



PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro

PHD: accuracy 70.2% for profiles
vs. 63% for primary sequence

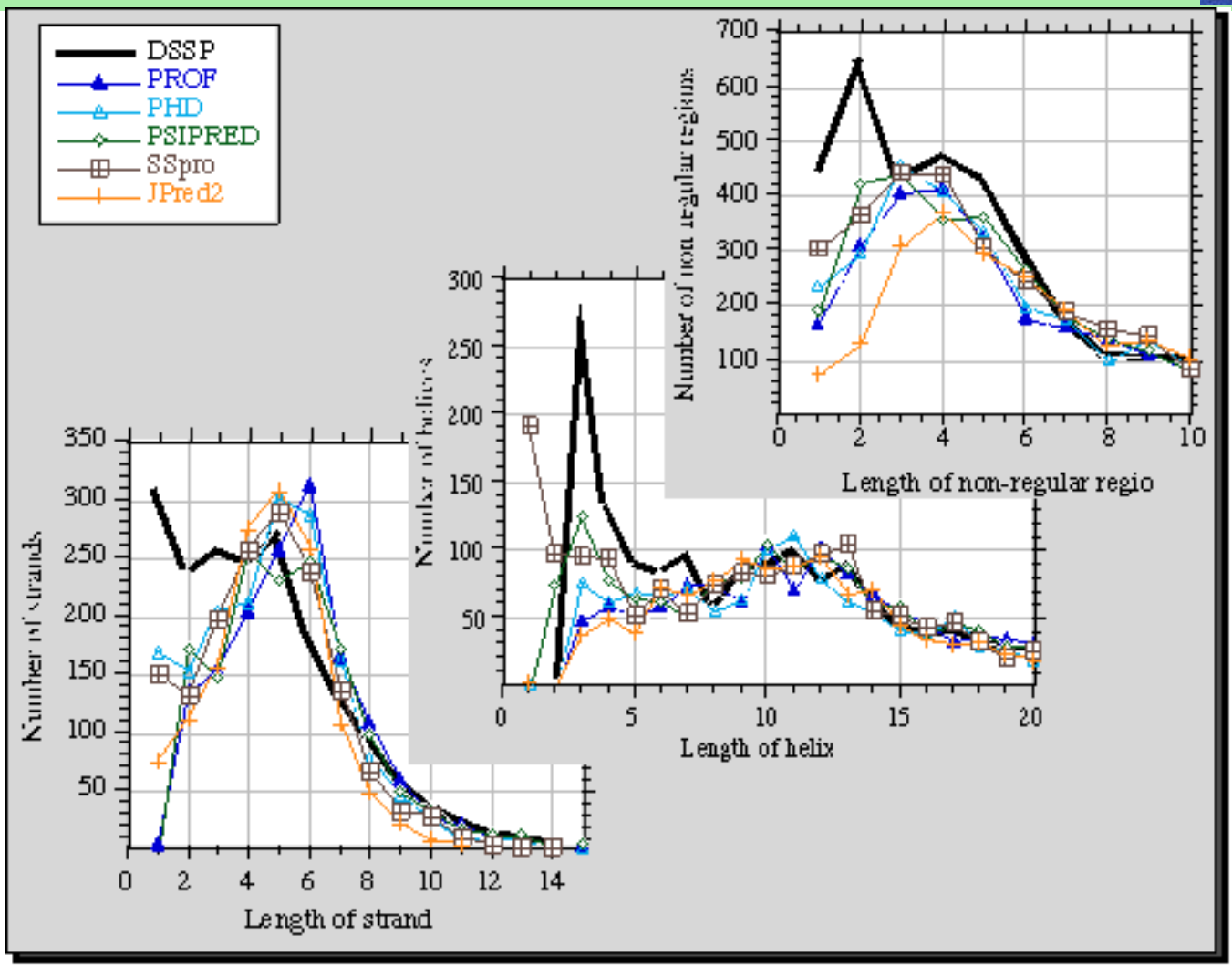
PSIPRED: based on PSI-BLAST PSSM

Other methods:
NSSP
PHDpsi
JNet
SSpro
PROF
JPred2

SSpro: based on a recurrent neural network

Currently the best performing method is an extension of SSPro:
Porter

PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro

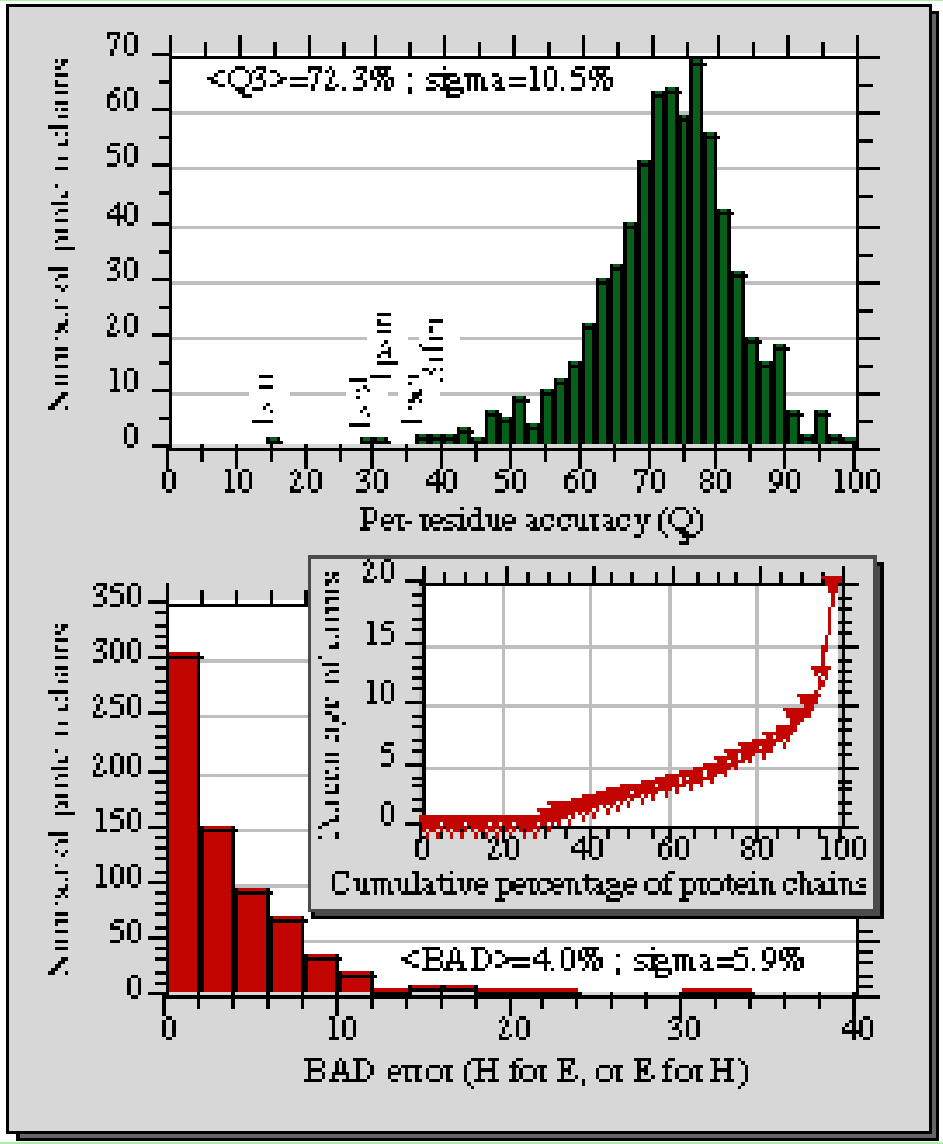


Distribution of segment length and predicted segment length

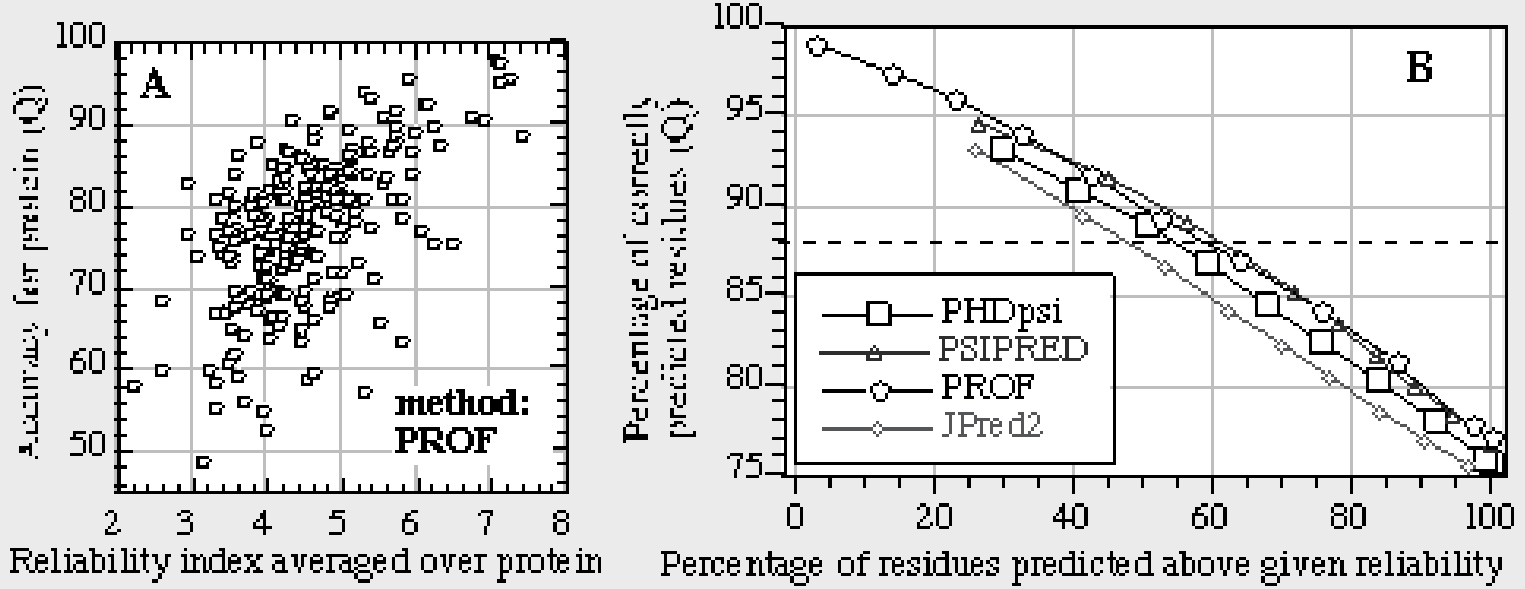
PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro

Secondary structure prediction for different methods from 1st generation to 3rd generation

PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro



PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro



Reliability and accuracy are plotted against each other.

If residues are predicted with higher reliability then the accuracy is higher.

Evaluating Secondary Structure Prediction

Models: constructed on training data and tested on test data

data must be based on known structures (PDB)

Non-Homologous Test Sequences

test set: independent identical distributed (iid)

nature does not sample proteins iid

selection bias stems from the experimenter

correct for the non-iid sampling:

proteins which are similar to the training set are removed (30% to 40% of mutual identity)

identity between pairs of test sequences is below the threshold (some sequences types are very often in the test)

Secondary Structure Classes

how the secondary structure has been assigned: DSSP or STRIDE ?
(differ only at the end of structural elements)

3 classes H,E, and C : how the mapping is done?

Symbol	conventional	newer	with turn
H	H	H	H
G	H	C	C
I	H	C	C
T	C	C	T
E	E	E	E
B	E	C	C
S	C	C	C
_	C	C	C

Quality Measures

binary classification task with a positive class (+1) and a negative class (-1) with N test examples

1. TP: true positive - positive correctly predicted;
2. FN: false negative - positive incorrectly predicted;
3. FP: false positive - negative incorrectly predicted;
4. TN: true negative - negative correctly predicted.

true	predicted		total
	+1	-1	
+1	TP	FN	TP + FN
-1	FP	TN	FP + TN
total	TP + FP	FN + TN	N

Table 1: Confusion matrix.

Quality Measures

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{N}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{balanced error} = 0.5\,(\text{specificity} + \text{sensitivity})$$

$$\text{Matthews corr.} = \frac{\text{TP TN} - \text{FP FN}}{\sqrt{(\text{TP} + \text{FN})\,(\text{TP} + \text{FP})\,(\text{FN} + \text{TN})\,(\text{FP} + \text{TN})}}$$

$$\text{weight of evidence} = \log \frac{\text{TP TN}}{\text{FP FN}}$$

Receiver Operating Characteristic curve (ROC):
plots sensitivity vs. (1 - specificity) for a binary classifier system
plot of the fraction of true positives vs. the fraction of false positives

Quality Measures

Measures used in secondary structure prediction

Q_3 accuracy – percent correct from all

sub-classifier accuracies: Q_H Q_E Q_C

β-sheets: most problems

quality measure is the segment overlap (SOV for Segment OVerlap)

$$\mathrm{SOV}_s = \frac{1}{N_s} \sum_{S_1 \cap S_2 \in s} \frac{\min \mathrm{ov}(S_1, S_2) + \delta(S_1, S_2)}{\max \mathrm{ov}(S_1, S_2)} \mathrm{length}(S_1)$$

- S_1 and S_2 : observed and predicted SS segments of $s = H, E, C$
- Length(S_1) is the number of residues in the segment S_1
- min ov(S_1,S_2): is the length of actual overlap
- max ov(S_1,S_2): is the length of total extend of S_1 and S_2
- $\delta(S_1, S_2)$ measure of disagreement

Quality Measures

$$\delta(S_1, S_2) = \min \left\{ \begin{array}{l} \max \mathrm{ov}(S_1, S_2) - \min \mathrm{ov}(S_1, S_2) \\ \min \mathrm{ov}(S_1, S_2) \\ \mathrm{length}(S_1)/2 \\ \mathrm{length}(S_2)/2 \end{array} \right\}$$

$S_1 \cap S_2 \in s$ pairs of segments have at least one residue in state s

N_s number of residues form sequence 1 in state s:

$$N_s = \sum_{S_1 \cap S_2 \in s} \mathrm{length}(S_1) + \sum_{S_1 \cap S_2 \notin s, S_1 \in s} \mathrm{length}(S_1)$$

overall segment overlap

$$\mathrm{SOV} = \frac{1}{N} \sum_{s \in \{H,E,C\}} \sum_{S_1 \cap S_2 \in s} \frac{\min \mathrm{ov}(S_1, S_2) + \delta(S_1, S_2)}{\max \mathrm{ov}(S_1, S_2)} \mathrm{length}(S_1)$$

$$N = \sum_{s \in \{H,E,C\}} N_s = N_H + N_E + N_C$$

Quality Measures

support vector machine approaches:
higher SOV value with state of the art values Q_3

tested the support vector approach and obtained
Q_3=78.84% and SOV=77.85%

PROFS Q_3=76.51%

PSIPRED Q_3=77.75 % and SOV=67.36%

According to the diploma thesis of Sebastian Drescher

Problems in Quality Comparisons

new method is compared to the results of previous methods
newer methods use newer versions of

- data bases like the NR
- software like PSI-BLAST
- the original data set

newer methods know

- about the pitfalls other methods
-
- performance threshold to achieve