Protein Structure Prediction: Secondary Structure

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Protein structure - Mozilla {Build ID: 2004031616}

File Edit View Go Bookmarks Tools Window Help Debug QA

Back Forward Reload Stop http://www.cmbi.kun.nl/gv/dssp/ Search Print

Home Bookmarks Release Notes Plug-ins Extensions Support Mozilla Community Drop_Off Locator

Centre for Molecular and Biomolecular Informatics

University of Nijmegen, Toernooiveld 1, P.O. Box 9010, 6500 GL Nijmegen, +31 (0)24-3653391,
postmaster@cmbi.kun.nl

Main Search

The DSSP database

The DSSP program was designed by Wolfgang Kabsch and Chris Sander to standardize secondary structure assignment. The DSSP database is a database of secondary structure assignments (and much more) for all protein entries in the Protein Data Bank (PDB).

Information

- The help document for the DSSP program.
- The DSSP article scanned in, or as PDF file.
- The license form for an academic DSSP source code.
- The license form for a commercial DSSP source code.
- A bill for commercial users.
- **AFTER** faxing the license form to the FAX number indicated at the form (+31 (0)24 3652977) you can extract the DSSP distribution by clicking here or from the anonymous FTP area of ftp.cmbi.kun.nl. Do a cd to pub/molbio/software and download dsspcmbi.zip . In any case, type **unzip dsspcmbi.zip** to unpack, then look at README.TXT.
- Precompiled executables are also available for Linux and Windows. (The Windows .exe file was compiled under Linux using Mingw32, has never seen a Windows environment and should thus be virus-free. Download the source if you want to be 100% sure.) Under Windows the DSSP output does not make it to the console, so redirect it to a file instead: dsspcmbi source.pdb destination.dssp >messages.txt
- Several changes have been made to the DSSP program to solve problems with recent PDB files. These are documented in the source code.
- Commercial users are requested to transfer Euro 1000 to account number of the "Stichting WHAT IF" no. 54.83.62.262 at the ABN-AMRO in Nijmegen. Please mention DSSP. Please transfer the money before down-loading the software.
- We have a version of the PDBFINDER with the secondary structure according to DSSP indicated as 1-letter code strings. Look at the example. You can download the entire file from ftp.cmbi.kun.nl/pub/molbio/data/pdbfinder2/PDBFIND2.TXT.gz.

start lect3 Protein structure - M... 7:48 PM

Secondary Structure Assignment

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).

Secondary Structure Prediction

Given the sequence of amino acids of a protein, what is its secondary structure?

Primary structure: GHWIATRGQLIREAYEDYRHFSSECPFIP

Secondary structure: CEEEEECHHHHHHHHHHHCCCHHCCCCCC

Notation: H: Helix E: Strand C: Coil

Conformational Preferences of Amino Acids

Amino acid	Preference α-helix	β-strand	Reverse turn
Glu	**1.59**	0.52	1.01
Ala	**1.41**	0.72	0.82
Leu	**1.34**	1.22	0.57
Met	**1.30**	1.14	0.52
Gln	**1.27**	0.98	0.84
Lys	**1.23**	0.69	1.07
Arg	**1.21**	0.84	0.90
His	**1.05**	0.80	0.81
Val	0.90	**1.87**	0.41
Ile	1.09	**1.67**	0.47
Tyr	0.74	**1.45**	0.76
Cys	0.66	**1.40**	0.54
Trp	1.02	**1.35**	0.65
Phe	1.16	**1.33**	0.59
Thr	0.76	**1.17**	0.90
Gly	0.43	0.58	**1.77**
Asn	0.76	0.48	**1.34**
Pro	0.34	0.31	**1.32**
Ser	0.57	0.96	**1.22**
Asp	0.99	0.39	**1.24**

Helical Preference.

Strand Preference.

Turn Preference.

Conformational Preferences of Amino Acids

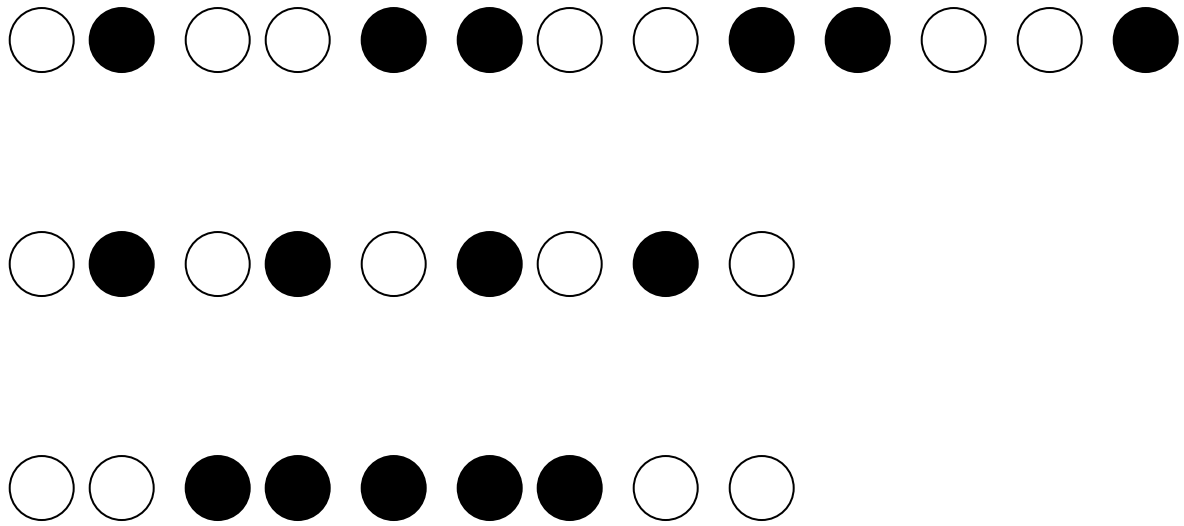
Amino acid	Preference α-helix	β-strand	Reverse turn
Glu	**1.59**	0.52	1.01
Ala	**1.41**	0.72	0.82
Leu	**1.34**	1.22	0.57
Met	**1.30**	1.14	0.52
Gln	**1.27**	0.98	0.84
Lys	**1.23**	0.69	1.07
Arg	**1.21**	0.84	0.90
His	**1.05**	0.80	0.81
Val	0.90	**1.87**	0.41
Ile	1.09	**1.67**	0.47
Tyr	0.74	**1.45**	0.76
Cys	0.66	**1.40**	0.54
Trp	1.02	**1.35**	0.65
Phe	1.16	**1.33**	0.59
Thr	0.76	**1.17**	0.90
Gly	0.43	0.58	**1.77**
Asn	0.76	0.48	**1.34**
Pro	0.34	0.31	**1.32**
Ser	0.57	0.96	**1.22**
Asp	0.99	0.39	**1.24**

Extended flexible side chains.

Bulky side chains, beta-branched.

Restricted conformations, side chain – main chain interactions.

Secondary Structure Prediction



Helix

Edge strand

Buried strand

By eye!

A Little Bit of History…

The early methods for secondary structure prediction suffered from lack of data, and were usually performed on single sequences.

1974: Chou and Fasman.

Propensities of formation based upon frequency of occurrence, rule based.

1974: Lim.

Theory based on chemical side-chain properties, very complex rules.

1978: Garnier, Osguthorpe, Robson.

Sliding window, consensus approach.

The prediction accuracy for all of those methods were roughly 50-55%.

Measures for Prediction Accuracy

The standard measure for prediction accuracy is (still) the Q3 measure. It is simply the proportion (in percent) of all amino acids that have correct matches for the three states C, E, H.

In recent years, the segment overlap measure (SOV) has been used more extensively. It aims for measuring how well secondary structure elements have been predicted rather than individual residues.

Rost et al (1994), JMB 235, pp 13-26.

Automated Methods

The availability of large families of homologous sequences together with advances in computing techniques has pushed the prediction accuracy well above 70%. Most methods are available as web servers. They include:

PHD

`http://www.embl-heidelberg.de/predictprotein/predictprotein.html`

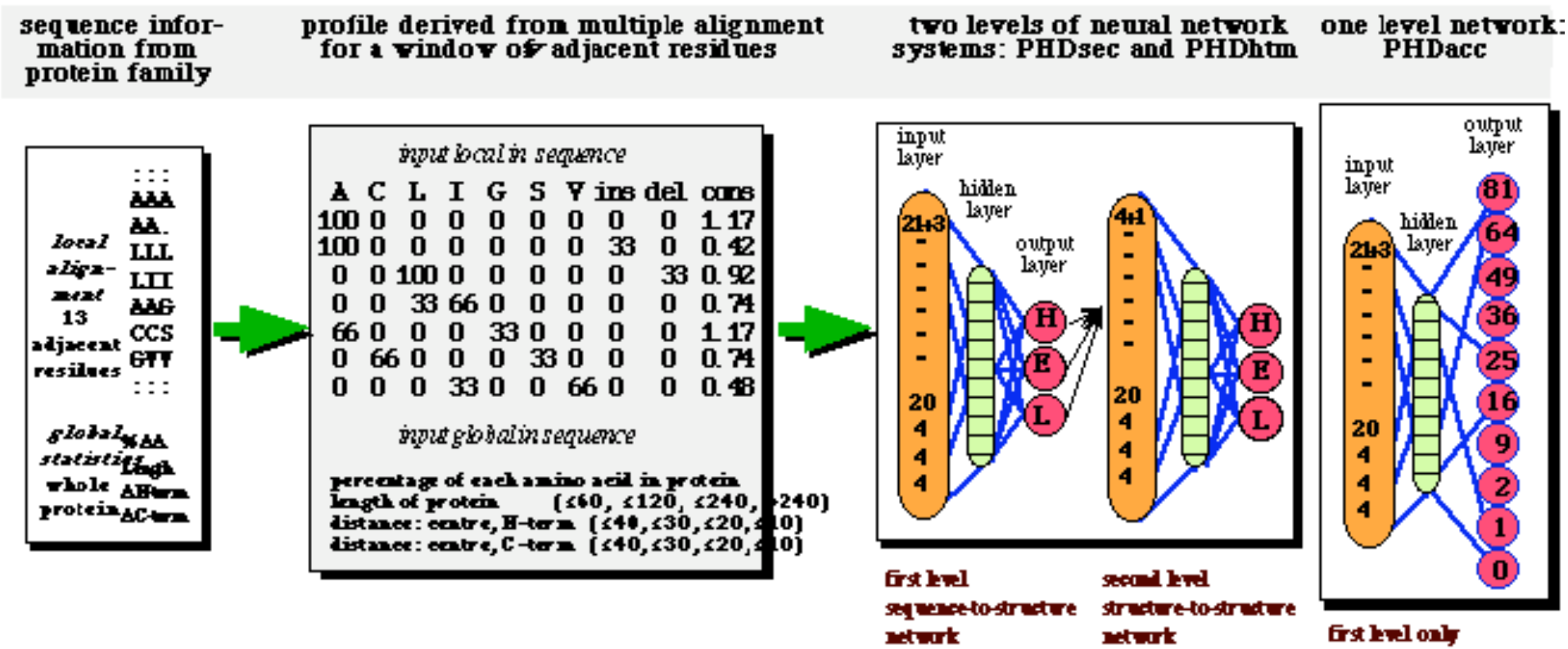
PSI-PRED

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

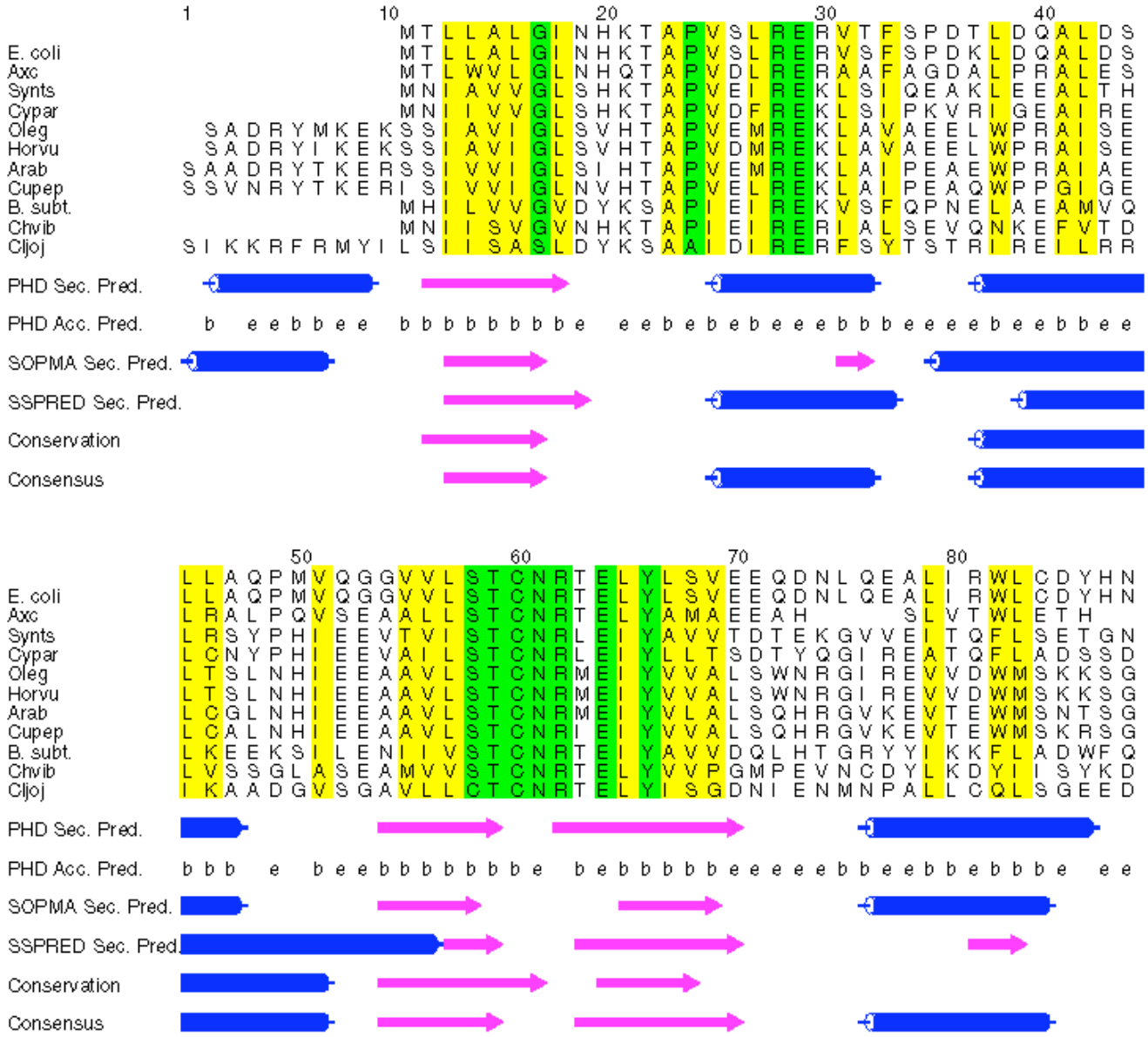
JPRED

`http://www.compbio.dundee.ac.uk/~www-jpred/`

PHD


sequence information from protein family
profile derived from multiple alignment for a window of adjacent residues
two levels of neural network systems: PHDsec and PHDhtm
one level network: PHDacc
local alignment 13 adjacent residues
AAA
AA.
LLL
LII
AAG
CCS
GVV
global statistics whole protein
%AA
Length
AN-term
AC-term
input local in sequence
A C L I G S V ins del cons
100 0 0 0 0 0 0 0 0 1.17
100 0 0 0 0 0 0 33 0 0.42
0 0 100 0 0 0 0 0 33 0.92
0 0 33 66 0 0 0 0 0 0.74
66 0 0 0 33 0 0 0 0 1.17
0 66 0 0 0 33 0 0 0 0.74
0 0 0 33 0 0 66 0 0 0.48
input global in sequence
percentage of each amino acid in protein
length of protein (≤60, ≤120, ≤240, >240)
distance: centre, N-term (≤40, ≤30, ≤20, ≤10)
distance: centre, C-term (≤40, ≤30, ≤20, ≤10)
input layer
hidden layer
output layer
21+3
20
4
4
4
H
E
L
4+1
first level sequence-to-structure network
second level structure-to-structure network
81
64
49
36
25
16
9
2
1
0
first level only

Consensus


1 10 20 30 40
E. coli
Axc
Synts
Cypar
Oleg
Horvu
Arab
Cupep
B. subt.
Chvib
Cljoj
PHD Sec. Pred.
PHD Acc. Pred.
SOPMA Sec. Pred.
SSPRED Sec. Pred.
Conservation
Consensus
50 60 70 80
E. coli
Axc
Synts
Cypar
Oleg
Horvu
Arab
Cupep
B. subt.
Chvib
Cljoj
PHD Sec. Pred.
PHD Acc. Pred.
SOPMA Sec. Pred.
SSPRED Sec. Pred.
Conservation
Consensus