PREDICTION OF PROTEIN FEATURES

Beyond protein structure

(TM, signal/target peptides, coiled coils, conservation…)

- **N-terminal signals**
- **Transmembrane helices**
- **Solvent accessibility**
- **Coiled coils**
- **Low complexity**
- **Biased regions**

- N-terminal signals
- Transmembrane helices
- Solvent accessibility
- Coiled coils
- Low complexity
- Biased regions

N-terminal signals

Signal peptide

3-60 aa long

Direct the transport of a protein

From cytoplasm to: nucleus, nucleolus, mitochondrial matrix, endoplasmic reticulum, chloroplast, apoplast, peroxisome.

Often N-terminal
Nuclear localization signal is internal (K/R)

N-terminal are often cleaved by a peptidase

N-terminal signals

Secretory signal peptide 15-30 aa

Cleaved off after translocation

n-region: positive charge
h-region: hydrophobic region
c-region: polar region
(some conserved residues at pos -3
and -1 of cleavage site)

N-terminal signals

Secretory signal peptide 15-30 aa

Prokaryotes

Transport across plasma membrane

Gram-negative: Periplasmic space (extra mechanism needed for extracellular)

Gram-positive: extracellular

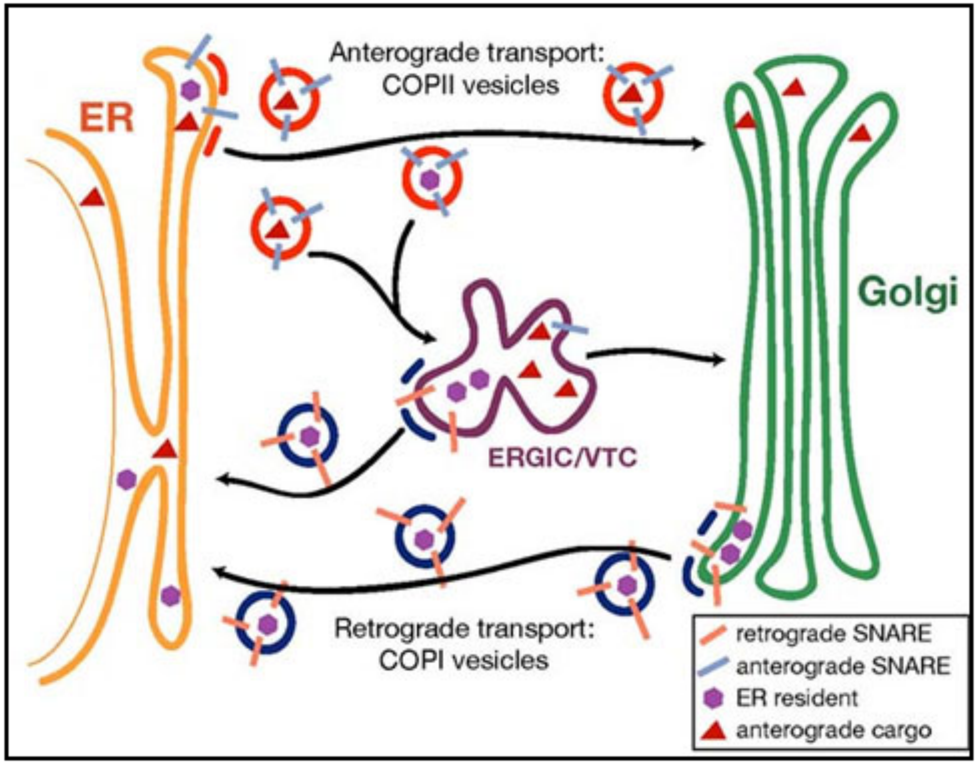
Eukaryotes

Transport across ER membrane. By default to the Golgi then to vesicles and secreted. (but there are signals for ER retention)

(and there are alternative pathways without signal peptide)

N-terminal signals

VTC: vesicular tubular clusters
(ER-Golgi intermediate compartment)



From Randy Schekman http://mcb.berkeley.edu/labs/schekman/

N-terminal signals

Targeting peptides

Cleaved off after translocation

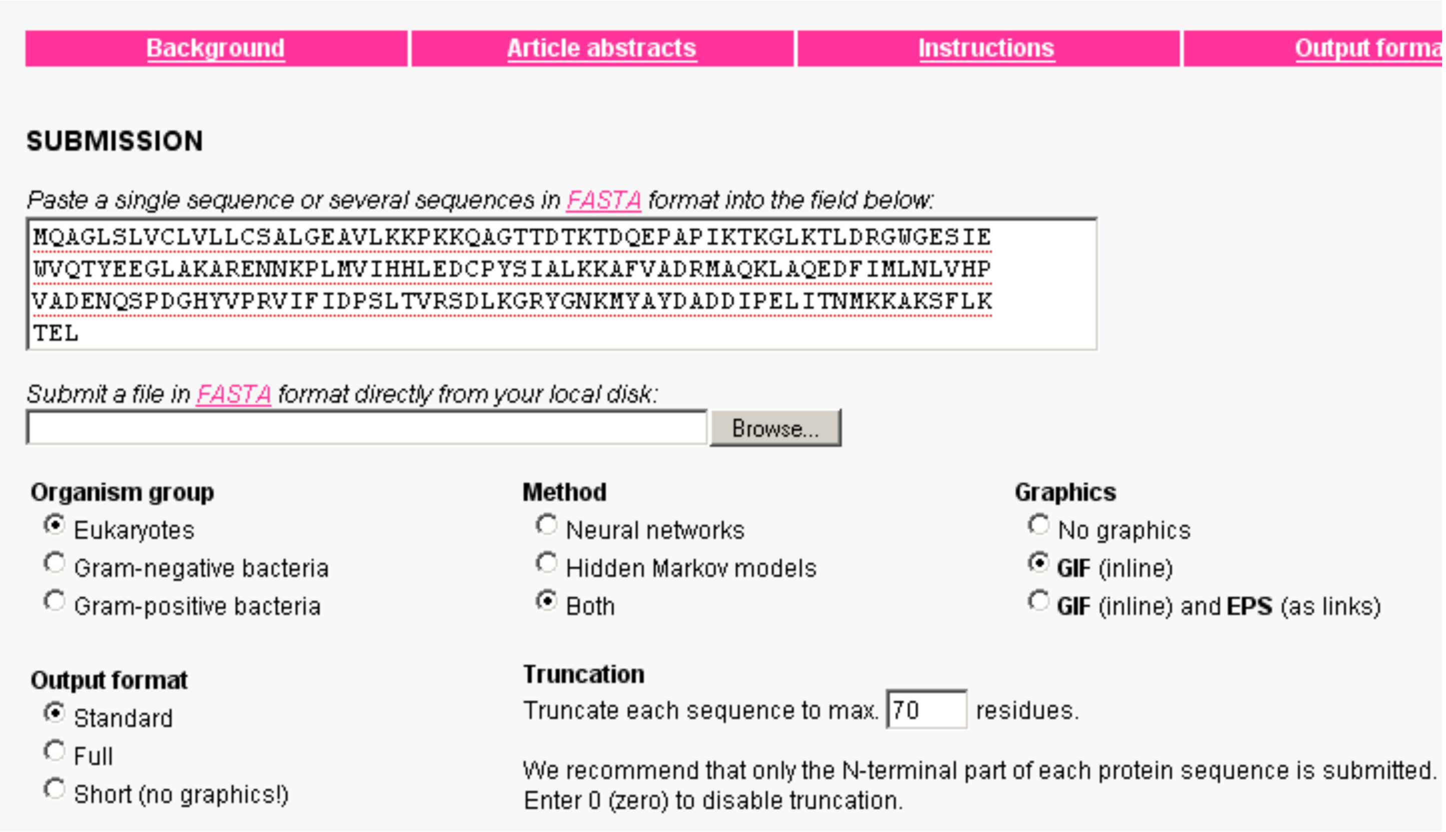
cTP chloroplast transit peptide

mTP mitochondrial targeting peptide

Some proteins are dually targeted to both chloroplasts and mitochondria using the same targeting sequence

N-terminal signals

Søren Brunak http://www.cbs.dtu.dk/services/SignalP/



SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:

```
MQAGLSLVCLVLLCSALGEAVLKKPKKQAGTTDTKTDQEPAPIKTKGLKTLDRGWGESIE
WVQTYEEGLAKARENNKPLMVIHHLEDCPYSIALKKAFVADRMAQKLAQEDFIMLNLVHP
VADENQSPDGHYVPRVIFIDPSLTVRSDLKGRYGNKMYAYDADDIPELITNMKKAKSFLK
TEL
```

Submit a file in FASTA format directly from your local disk:

Organism group

- ● Eukaryotes
- ○ Gram-negative bacteria
- ○ Gram-positive bacteria

Method

- ○ Neural networks
- ○ Hidden Markov models
- ● Both

Graphics

- ○ No graphics
- ● **GIF** (inline)
- ○ **GIF** (inline) and **EPS** (as links)

Output format

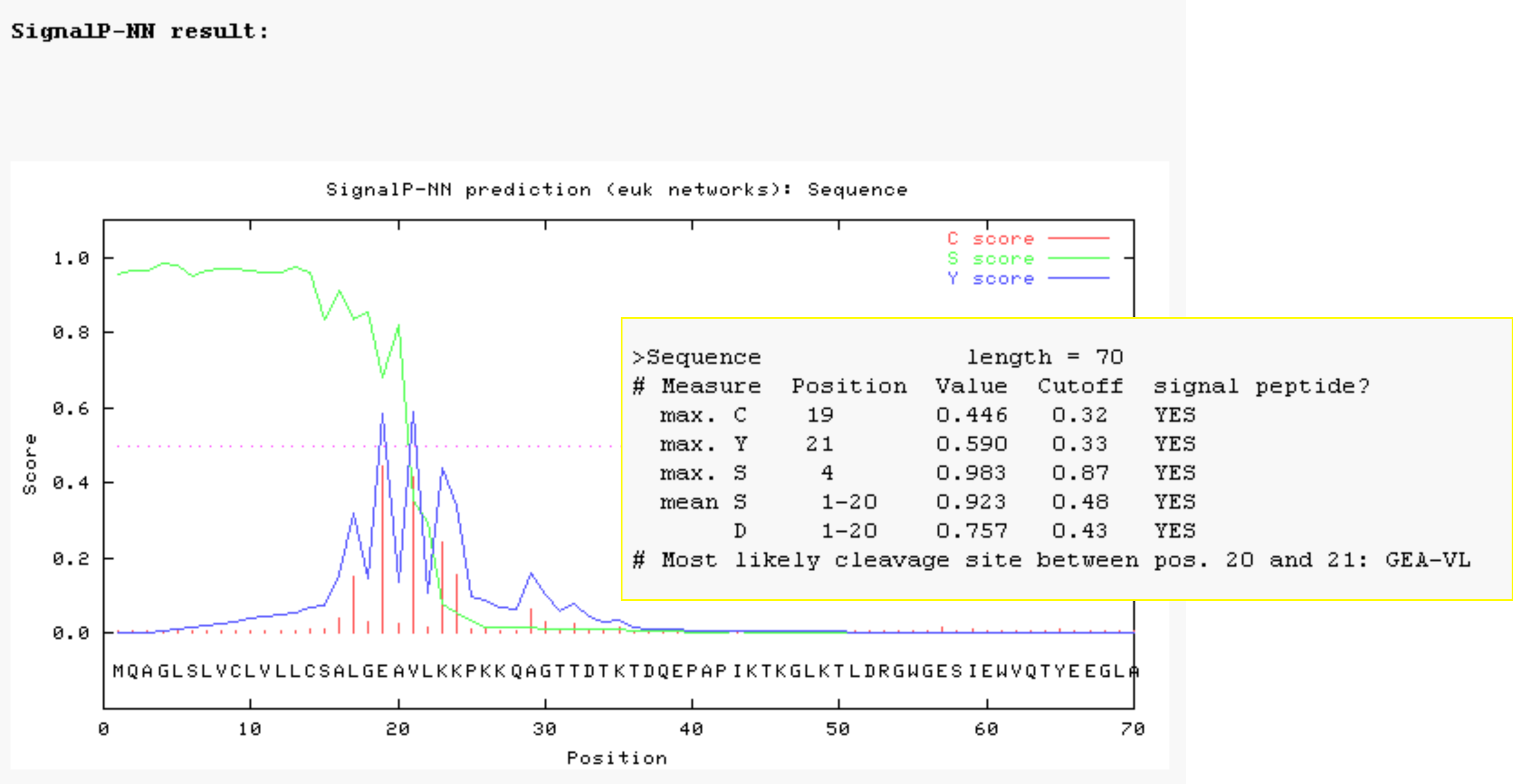
- ● Standard
- ○ Full
- ○ Short (no graphics!)

Truncation

Truncate each sequence to max. 70 residues.

We recommend that only the N-terminal part of each protein sequence is submitted. Enter 0 (zero) to disable truncation.

N-terminal signals

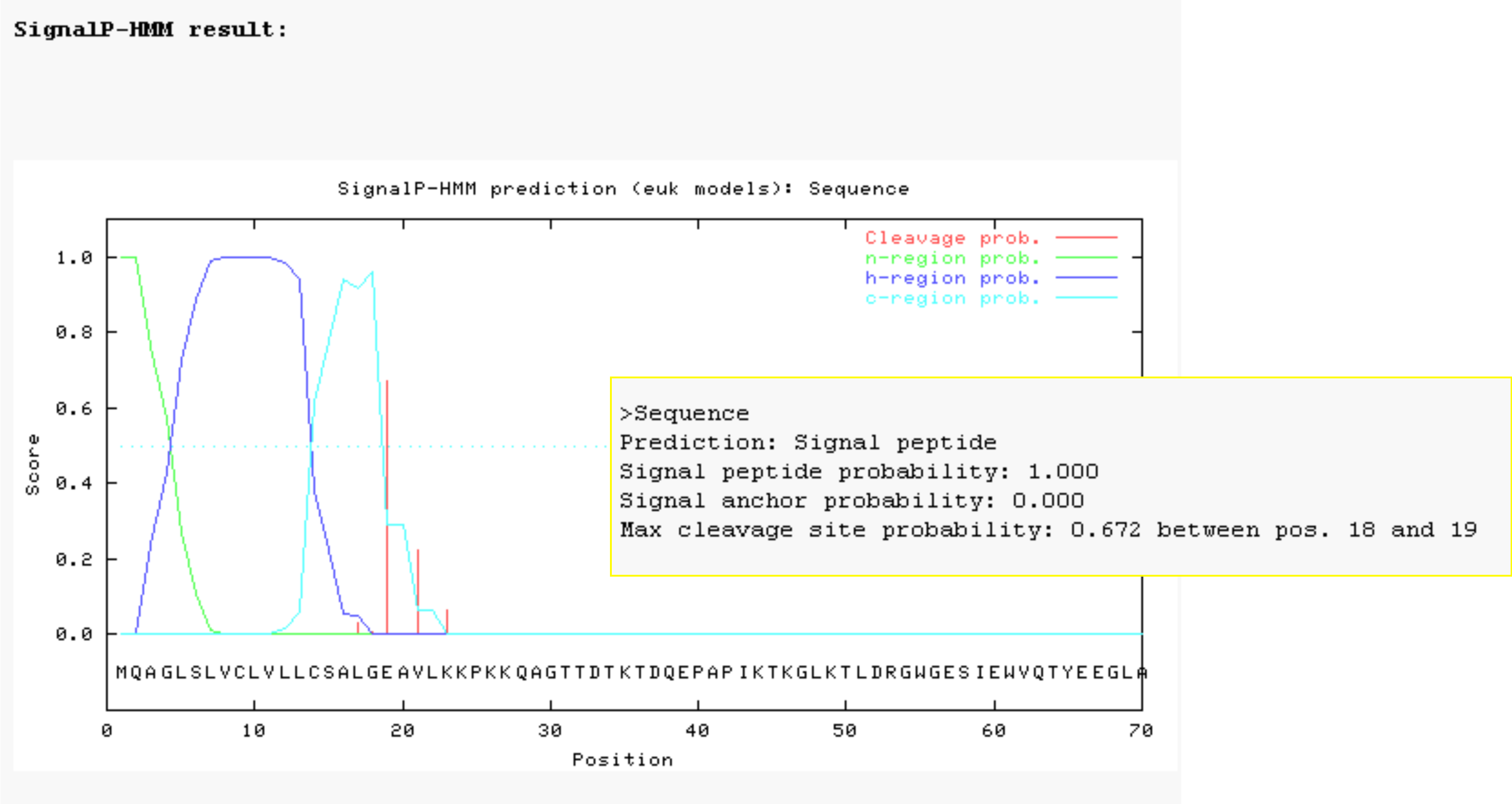


SignalP-NN result:

SignalP-NN prediction (euk networks): Sequence
C score
S score
Y score
Score
1.0
0.8
0.6
0.4
0.2
0.0
MQAGLSLVCLVLLCSALGEAVLKKPKKQAGTTDTKTDQEPAPIKTKGLKTLDRGWGESIEWVQTYEEGLA
0
10
20
30
40
50
60
70
Position

```
>Sequence                    length = 70
# Measure  Position  Value   Cutoff   signal peptide?
  max. C    19       0.446   0.32     YES
  max. Y    21       0.590   0.33     YES
  max. S     4       0.983   0.87     YES
  mean S     1-20    0.923   0.48     YES
       D     1-20    0.757   0.43     YES
# Most likely cleavage site between pos. 20 and 21: GEA-VL
```

N-terminal signals


SignalP-HMM result:

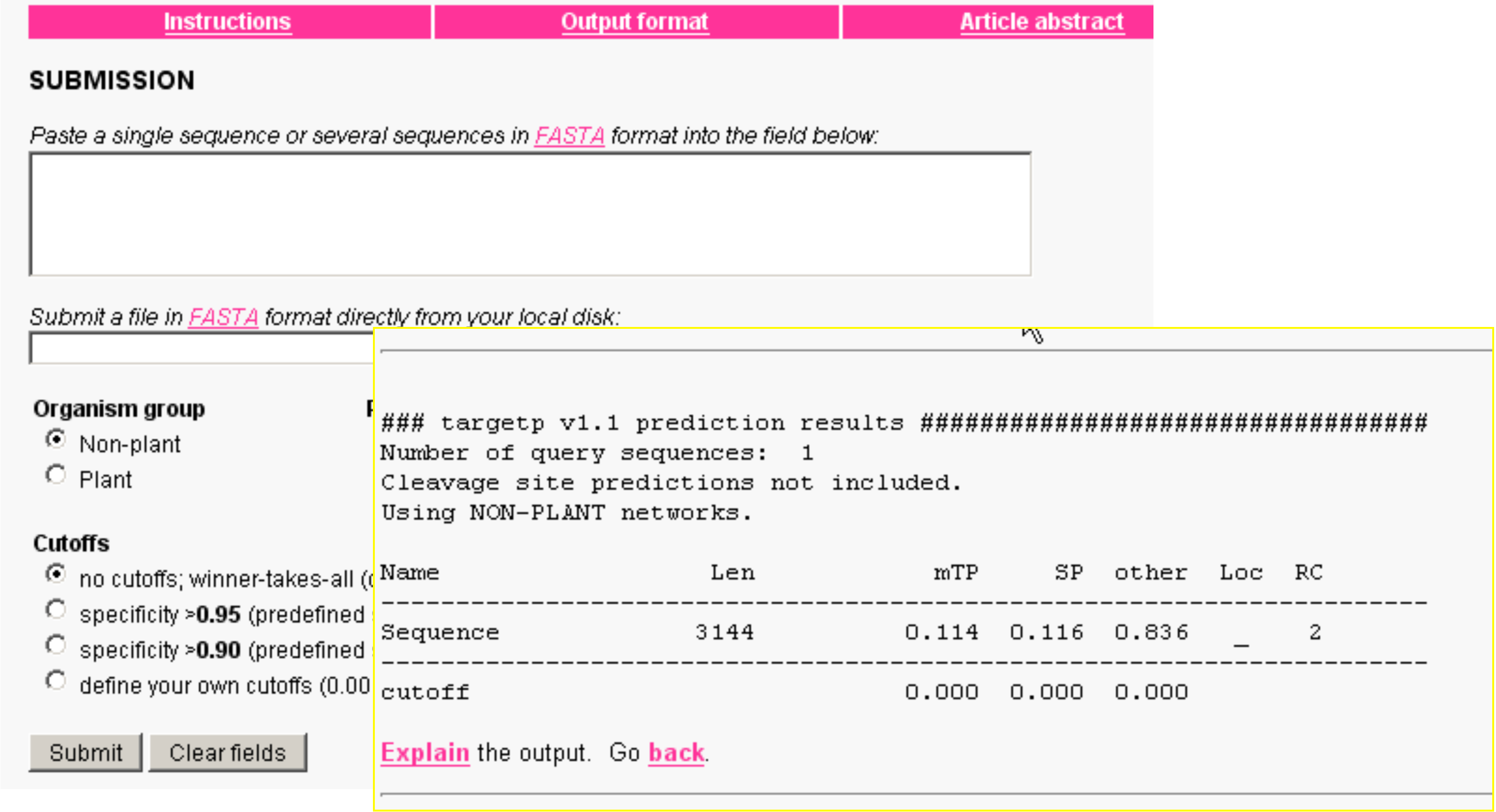
SignalP-HMM prediction (euk models): Sequence
Cleavage prob.
n-region prob.
h-region prob.
c-region prob.
Score
1.0
0.8
0.6
0.4
0.2
0.0
MQAGLSLVCLVLLCSALGEAVLKKPKKQAGTTDTKTDQEPAPIKTKGLKTLDRGWGESIEWVQTYEEGLA
0
10
20
30
40
50
60
70
Position

```
>Sequence
Prediction: Signal peptide
Signal peptide probability: 1.000
Signal anchor probability: 0.000
Max cleavage site probability: 0.672 between pos. 18 and 19
```

N-terminal signals

Soren Brunak http://www.cbs.dtu.dk/services/TargetP/

- Mostly based on signal peptides



SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:

Submit a file in FASTA format directly from your local disk:

Organism group

● Non-plant
○ Plant

Cutoffs

● no cutoffs; winner-takes-all (
○ specificity >**0.95** (predefined
○ specificity >**0.90** (predefined
○ define your own cutoffs (0.00

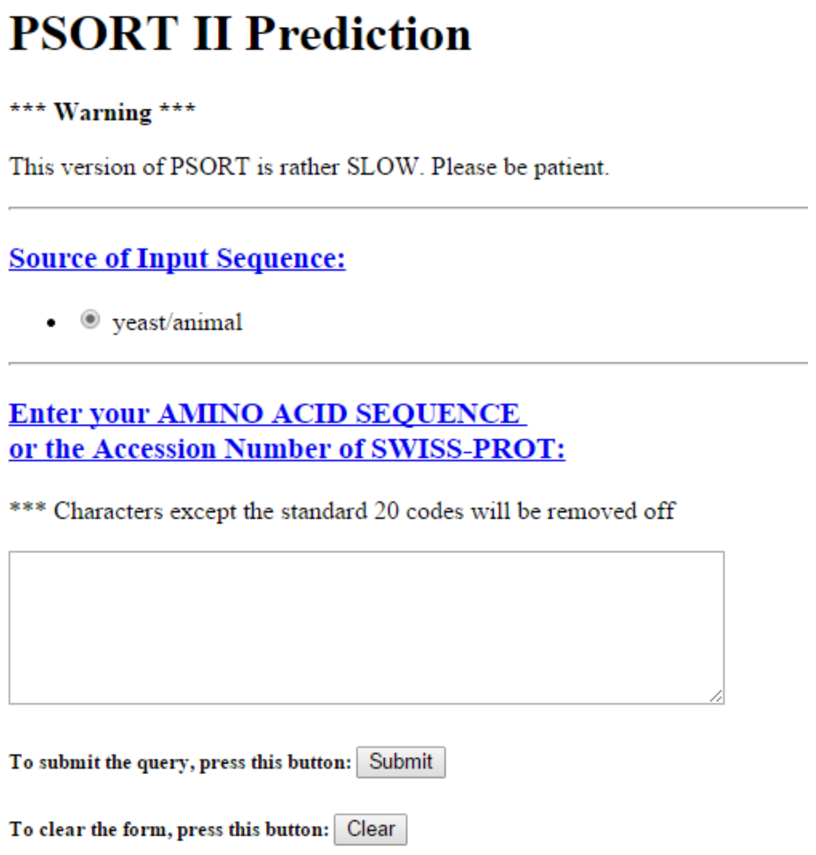
```
### targetp v1.1 prediction results ##################################
Number of query sequences:  1
Cleavage site predictions not included.
Using NON-PLANT networks.

Name                  Len            mTP     SP  other  Loc  RC
----------------------------------------------------------------------
Sequence             3144          0.114  0.116  0.836   _    2
----------------------------------------------------------------------
cutoff                             0.000  0.000  0.000
```

Explain the output. Go back.

PSORT
Prediction of subcellular location

http://psort.hgc.jp/form2.html



PSORT II Prediction

***** Warning *****

This version of PSORT is rather SLOW. Please be patient.

Source of Input Sequence:

- ● yeast/animal

**Enter your AMINO ACID SEQUENCE
or the Accession Number of SWISS-PROT:**

*** Characters except the standard 20 codes will be removed off

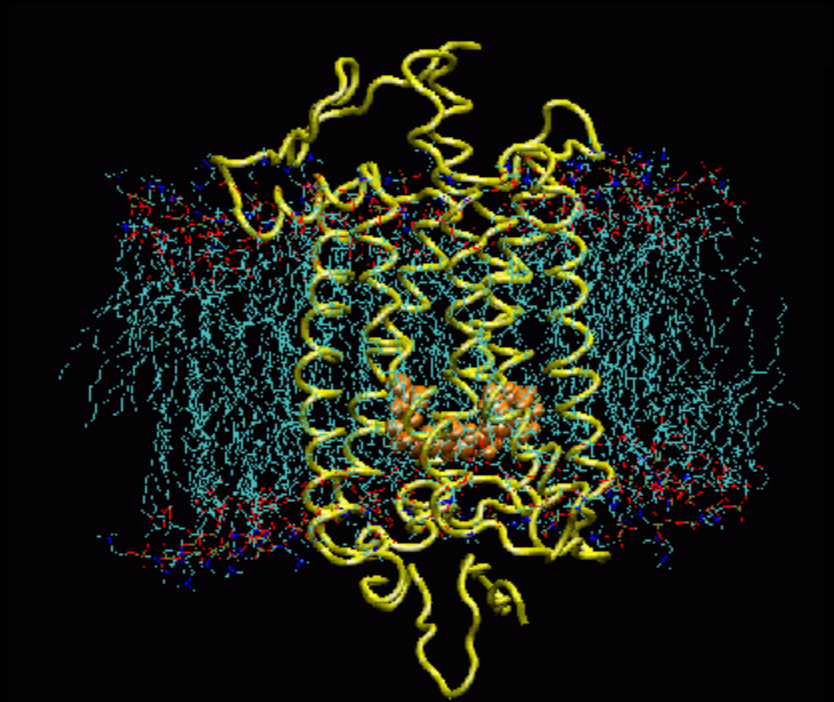
To submit the query, press this button: Submit

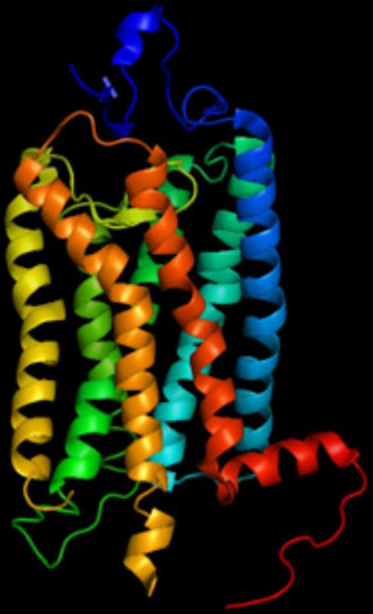
To clear the form, press this button: Clear

- **N-terminal signals**
- **Transmembrane helices**
- **Solvent accessibility**
- **Coiled coils**
- **Low complexity**
- **Biased regions**

Transmembrane helices

Rhodopsin: sensitive to light





7 TM helices

Left from	http://www.ks.uiuc.edu/Research/rhodopsin/
Right from	http://ocw.mit.edu/

Transmembrane helices

Hydrophobic helices of approx. 20 residues that traverse the cell membrane perpendicular to its surface

Transmembrane helices

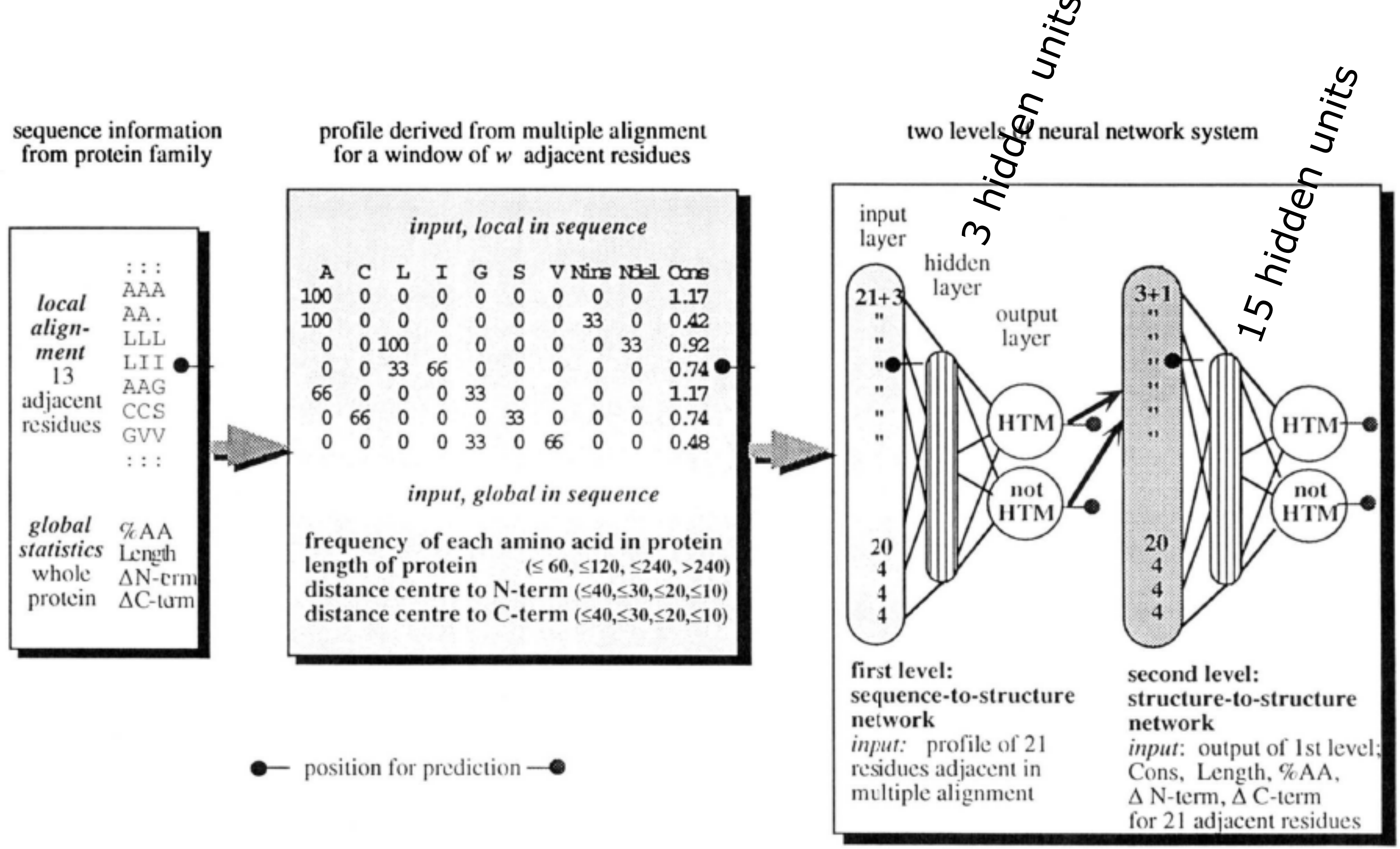
Methods for prediction use:

- hydrophobicity analyses

- the preponderance of positively charged residues on the cytoplasmic side of the transmembrane segment (positive inside rule)

- multiple sequence alignments

Transmembrane helices


3 hidden units
15 hidden units
sequence information
from protein family
local
align-
ment
13
adjacent
residues
:::
AAA
AA.
LLL
LII
AAG
CCS
GVV
:::
global
statistics
whole
protein
%AA
Length
ΔN-term
ΔC-term
profile derived from multiple alignment
for a window of w adjacent residues
input, local in sequence

A	C	L	I	G	S	V	Nins	Ndel	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	0	33	0	66	0	0	0.48

input, global in sequence
frequency of each amino acid in protein
length of protein (≤ 60, ≤120, ≤240, >240)
distance centre to N-term (≤40,≤30,≤20,≤10)
distance centre to C-term (≤40,≤30,≤20,≤10)
position for prediction
two levels of neural network system
input layer
hidden layer
output layer
21+3
"
20
4
4
4
HTM
not HTM
3+1
20
4
4
4
HTM
not HTM
first level:
sequence-to-structure network
input: profile of 21 residues adjacent in multiple alignment
second level:
structure-to-structure network
input: output of 1st level; Cons, Length, %AA, Δ N-term, Δ C-term for 21 adjacent residues

Filter to keep helix length in 17-25 range

Rost et al (1995) *Protein Science*

Transmembrane helices TMHMM

Søren Brunak http://www.cbs.dtu.dk/services/TMHMM/

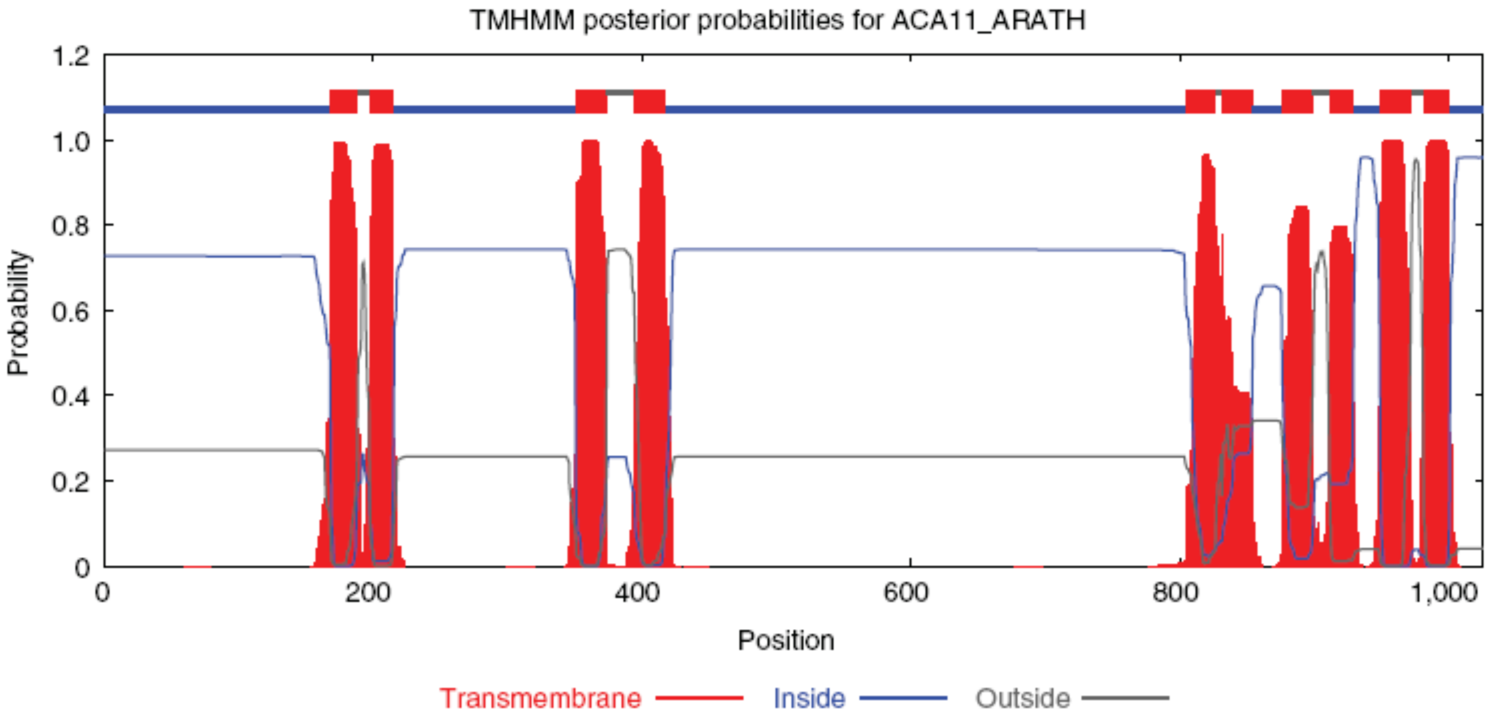


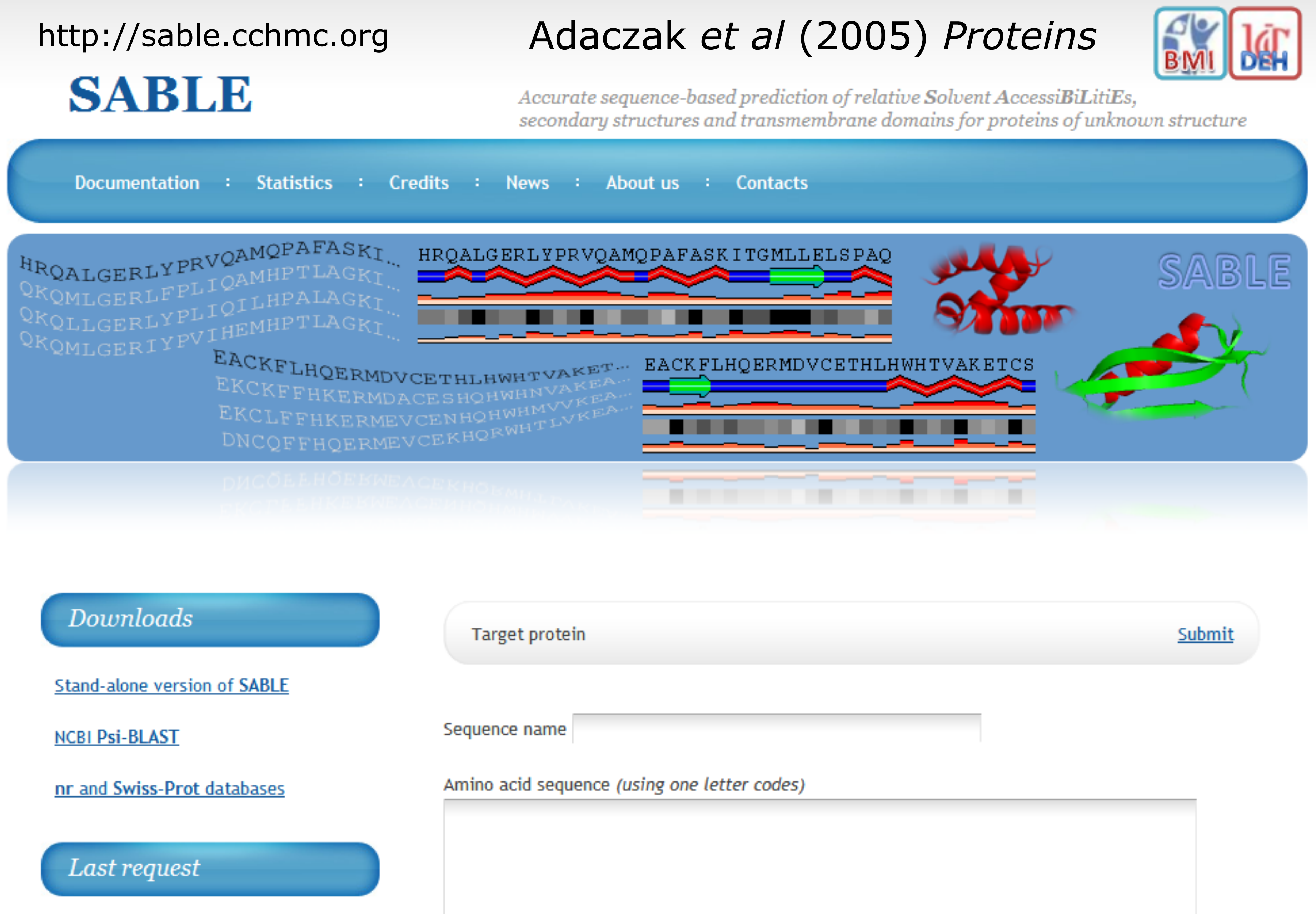
Figure 6 | The graphical output of TMHMM, showing the posterior probabilities for transmembrane, inside (i.e., cytoplasmic), and outside (i.e., lumenal or exterior) regions. In this example (*Arabidopsis thaliana* putative calcium-transporting ATPase isoform 11), ten transmembrane regions are predicted.

- **N-terminal signals**
- **Transmembrane helices**
- **Solvent accessibility**
- **Coiled coils**
- **Low complexity**
- **Biased regions**

Solvent accessibility

http://sable.cchmc.org

Adaczak *et al* (2005) *Proteins*



SABLE

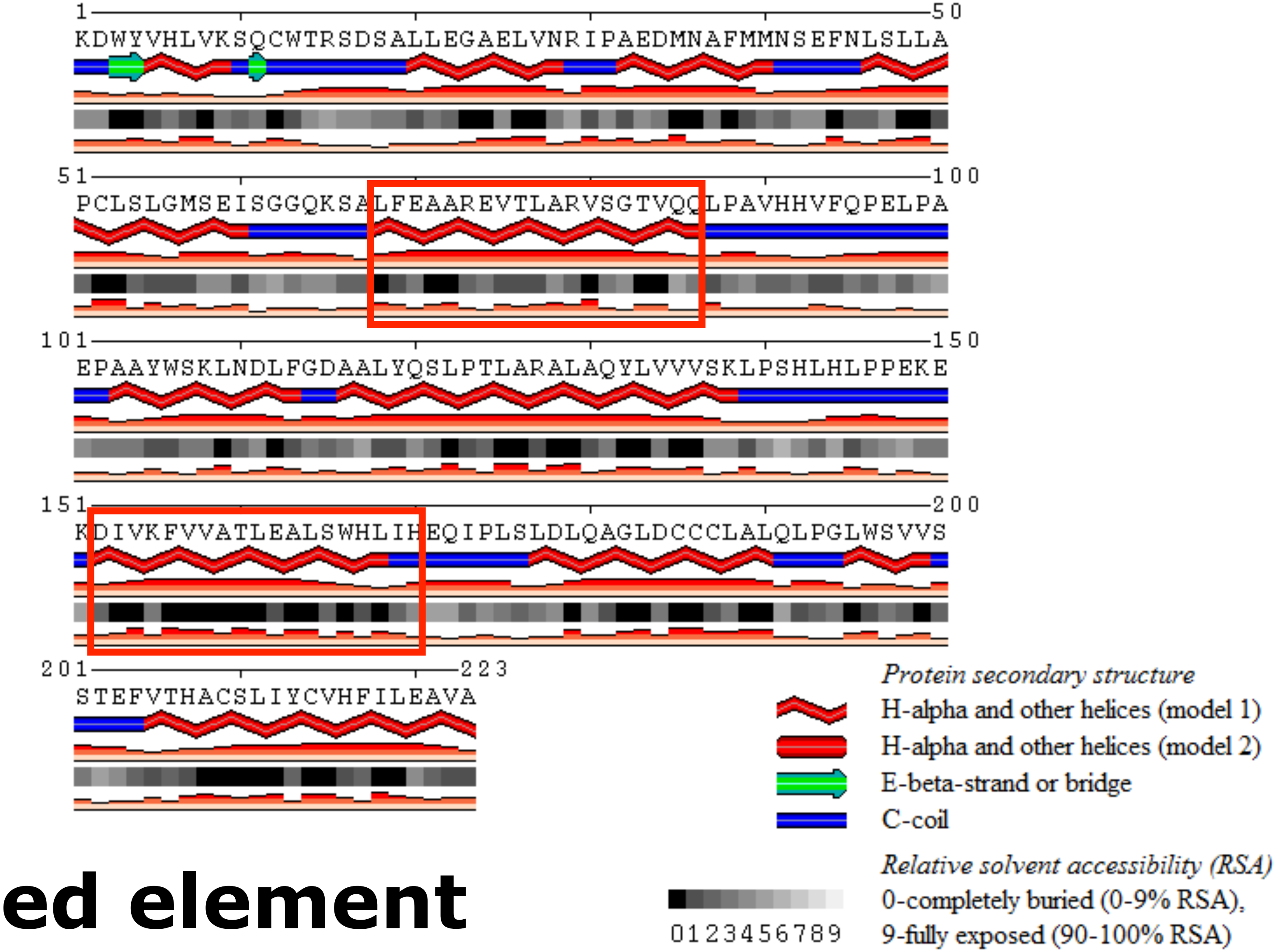
Accurate sequence-based prediction of relative Solvent AccessiBiLitiEs, secondary structures and transmembrane domains for proteins of unknown structure

BMI
DEH
Documentation : Statistics : Credits : News : About us : Contacts
HRQALGERLYPRVQAMQPAFASKI...
QKQMLGERLFPLIQAMHPTLAGKI...
QKQLLGERLYPLIQILHPALAGKI...
QKQMLGERIYPVIHEMHPTLAGKI...
HRQALGERLYPRVQAMQPAFASKITGMLLELSPAQ
EACKFLHQERMDVCETHLHWHTVAKET...
EKCKFFHKERMDACESHQHWHNVAKEA...
EKCLFFHKERMEVCENHQHWHMVVKEA...
DNCQFFHQERMEVCEKHQRWHTLVKEA...
EACKFLHQERMDVCETHLHWHTVAKETCS
SABLE
Downloads
Stand-alone version of SABLE
NCBI Psi-BLAST
nr and Swiss-Prot databases
Last request
Target protein
Submit

Sequence name

Amino acid sequence *(using one letter codes)* accuracy up to 88.9%

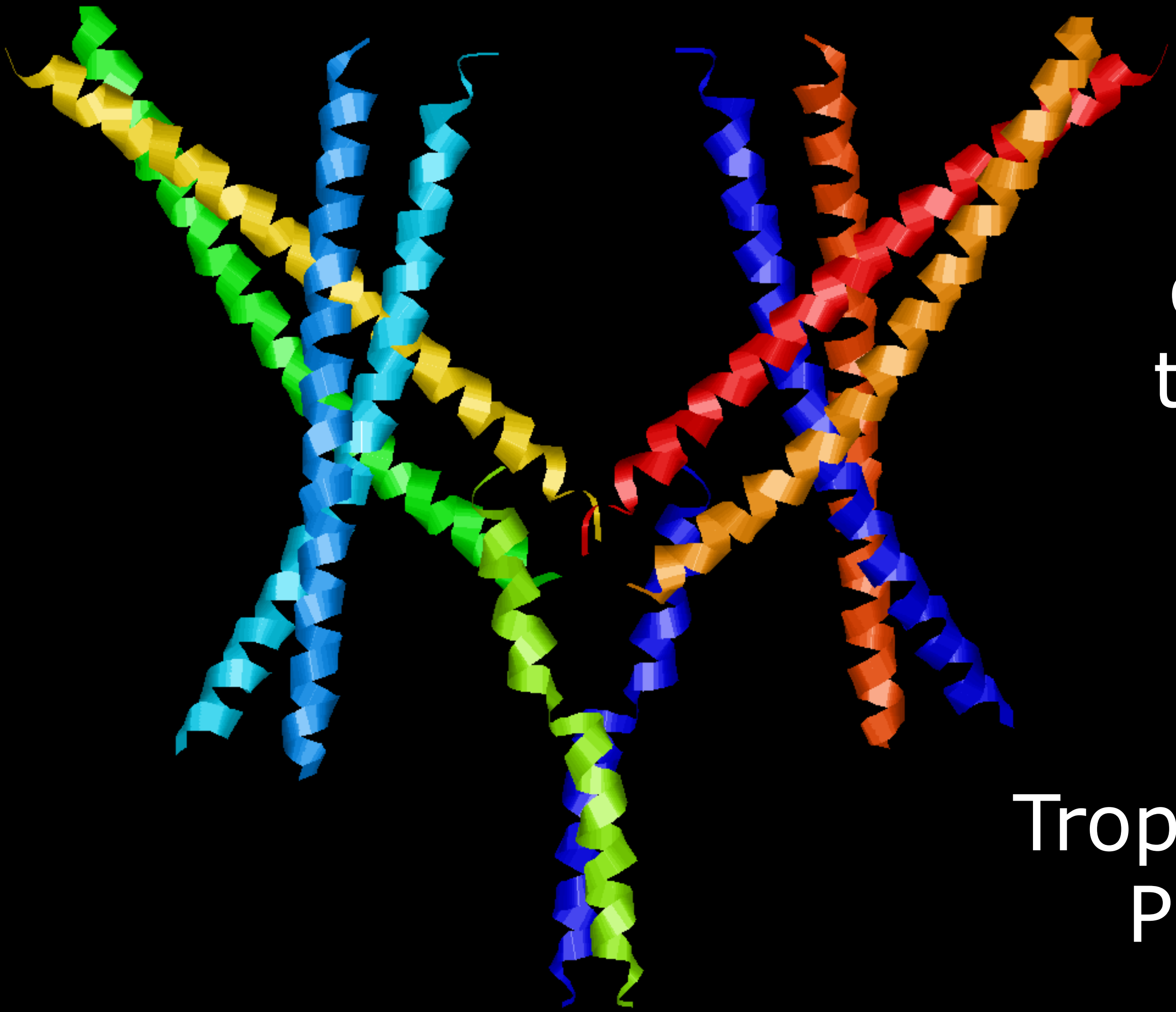
Amphipathic alpha helix


1
50
KDWYVHLVKSQCWTRSDSALLEGAELVNRIPAEDMNAFMMNSEFNLSLLA
51
100
PCLSLGMSEISGGQKSALFEAAREVTLARVSGTVQQLPAVHHVFQPELPA
101
150
EPAAYWSKLNDLFGDAALYQSLPTLARALAQYLVVVSKLPSHLHLPPEKE
151
200
KDIVKFVVATLEALSWHLIHEQIPLSLDLQAGLDCCCLALQLPGLWSVVS
201
223
STEFVTHACSLIYCVHFILEAVA
Protein secondary structure
H-alpha and other helices (model 1)
H-alpha and other helices (model 2)
E-beta-strand or bridge
C-coil
Relative solvent accessibility (RSA)
0123456789
0-completely buried (0-9% RSA),
9-fully exposed (90-100% RSA)

Buried element

- N-terminal signals
- Transmembrane helices
- Solvent accessibility
- Coiled coils
- Low complexity
- Biased regions

Coiled coils



dimers
trimers

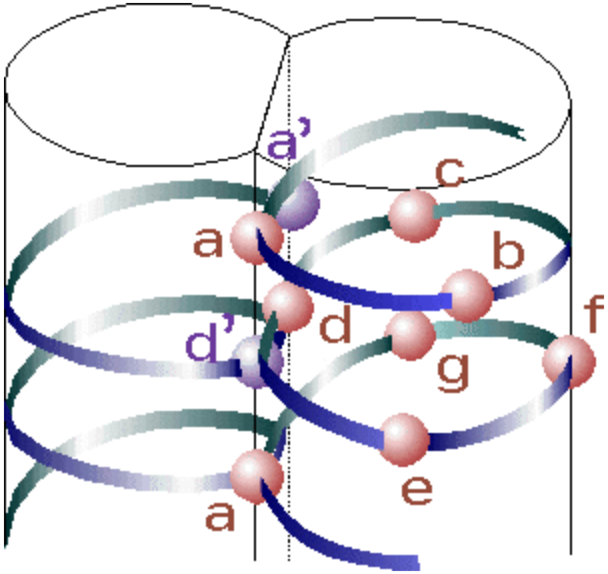
Tropomyosin
PDB:2Z5I

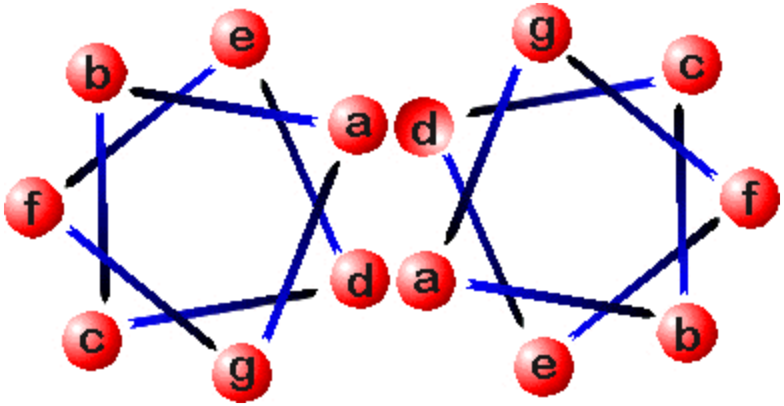
Coiled coils

Heptad repeat: a – b – c – d – e – f – g

H P P H C P C

H = hydrophobic; P = polar; C = charged

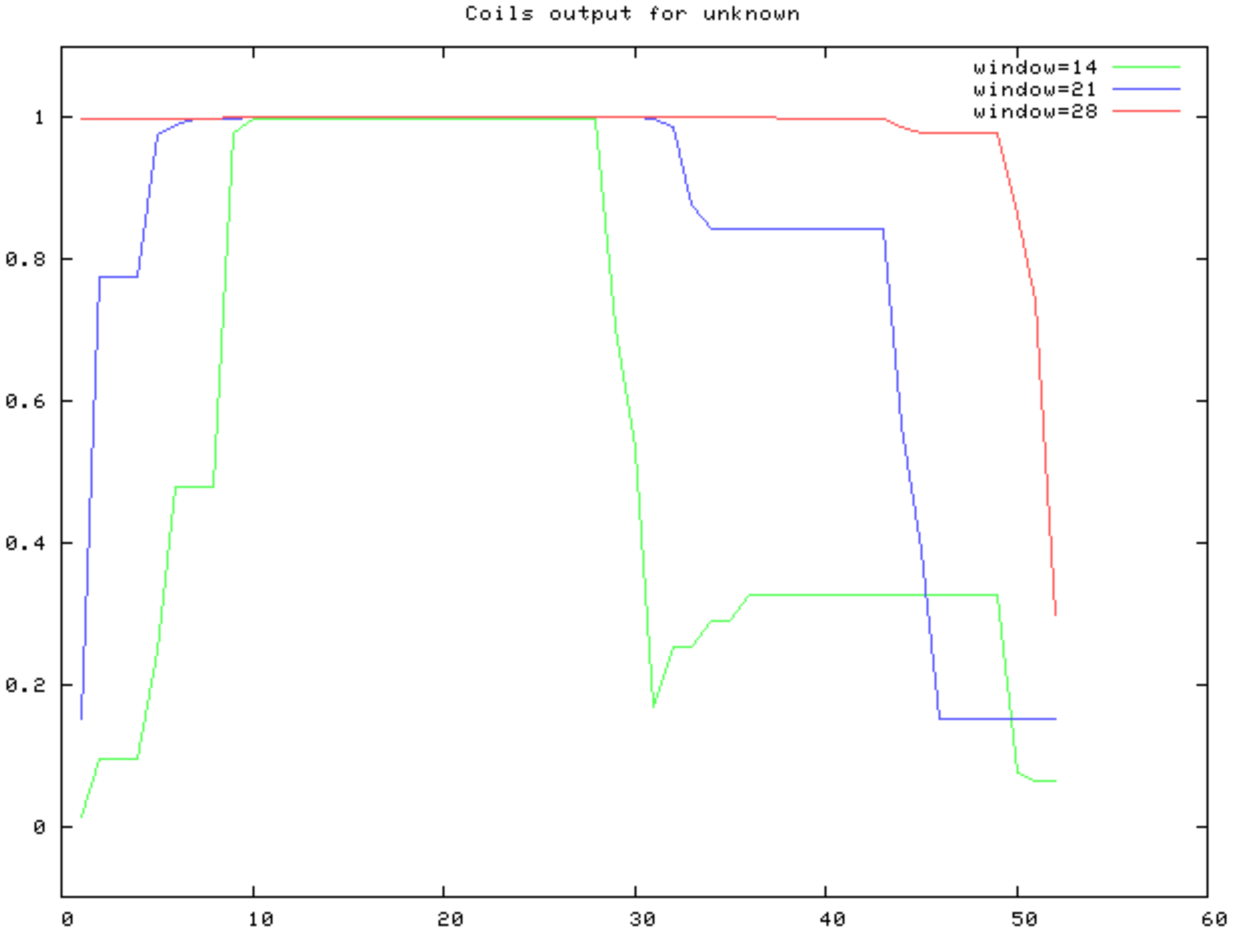




Source: http://cis.poly.edu/~jps/coilcoil.html

Coiled coils

Andrei Lupas   http://www.ch.embnet.org/software/COILS_form.html



Lupas *et al* (1991) *Science*

Exercise 1/4

Predict TM alpha-helices with TMHMM

- Here you can see the entry in the UniProt database for a short fly protein of unknown function: http://www.uniprot.org/uniprot/Q28WW9

- Obtain the sequence of this protein from here: http://www.uniprot.org/uniprot/Q28WW9.fasta

- Run the sequence in TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) and check the output.
- How many TM helices are predicted for this protein? What is the predicted orientation of the protein?

Exercise 2/4

Predict secondary structure with Jpred

- Let’s predict the secondary structure of the little transmembrane protein using a multiple sequence alignment with homologs.

- Load littleMSA_fasta.txt on JalView

- Calculate secondary structure prediction using Web Service > Secondary Structure Prediction > Jnet
(Do not select any sequences when doing this so that the alignment is used)

- Select the menu Colour and option Clustalx to view the amino acids by property.

- Can you see the TM region (hydrophobic residues are coloured blue)?

- What type of structure was predicted for that region? There is a C-terminal proline rich region. Is that region predicted to be structured? Is that region conserved?

Exercise 3/4

Sequence conservation on 3D

- Load in JalView a multiple sequence alignment of plant ferredoxins ferredoxins2_fasta.txt.

- Select FER1_SPIOL. Right click on FER1_SPIOL. Select structure > Associate structure with sequences > discover PDB ids.

- Now again, right click on FER1_SPIOL > 3D Structure data. Select 1a70 and click View. This will open a window where you can view its structure (PDB 1A70). The viewer is Jmol. Try rotating the structure.

- The sequence is connected to the structure. Mouse over the sequence and see how the corresponding amino acid is highlighted in the 3D view. Click on the 3D view and the amino acid will be highlighted in the alignment.

- Apply color (BLOSUM62) in the alignment window. Then in the 3D view option View > color by, then choose the option that uses the alignment.

- Hint: If in the structure window you apply colour then you will loose the interactivity. You have to go to the view option and apply Color by… option.

Exercise 4/4

Overlap a 2nd structure

- Now do the same with FER1_MAIZE. Use 3B2F. Say that you want to add it to the view. The two 3Ds will be overlapped.

- Use view in the 3D view to select and deselect chains to view. View > Select chain > click out 3B2F:B

- Are any significant differences between these two structures?

- What is the most conserved region of ferredoxin? Is it structured?

- In the alignment apply Colour > Zappo. This will colour all residues according to residue type. Find a position in a loop where these two ferredoxins have a different amino acid.
(Hint: you can clear the labels by deselecting a chain to view and selecting it again)