Protein Sequence Analysis

Extraction of Structural (1D) Features from Sequence Alignments

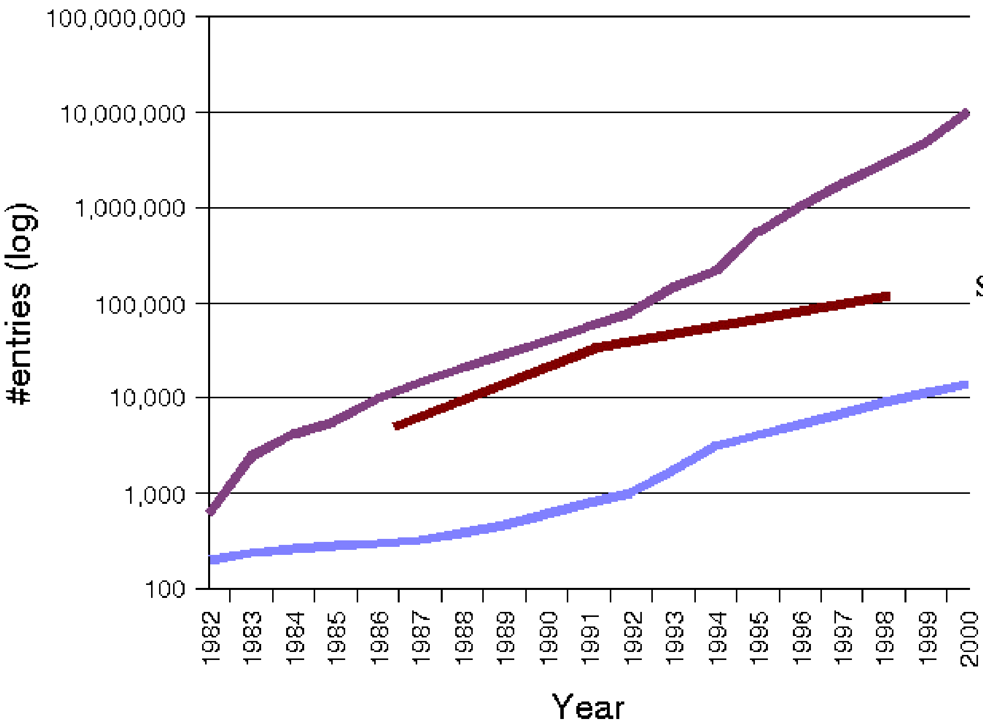
Florencio Pazos (CNB-CSIC)

Florencio Pazos Cabaleiro
Protein Design Group (CNB-CSIC)
pazos@cnb.uam.es

Protein Structure

MREYKLVVLGSGGVGKSALTVQFVQGIFVDEYDPTIEDSY
RKQVEVDCQQCMLEILDTAGTEQFTAMRDLYMKNGQGFAL
VYSITAQSTFNDLQDLREQILRVKDTEDVPMILVGNKCDL
EDERVVGKEQGQNLARQWCNCAFLESSAKSKINVNEIFYD
LVRQINR

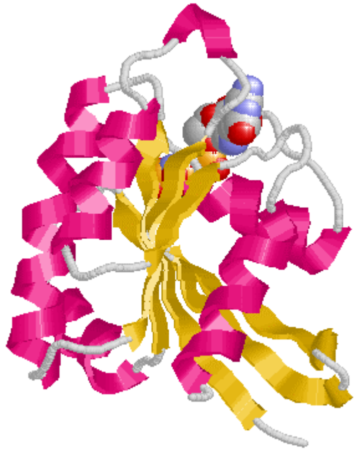
MLEILDTA**GTEQFT**AMRDLYMKNGQGFAL
VYSITAQSTFND**LQDLREQIL**RVKDTEDVPMIL
VGNKCDLEDERV



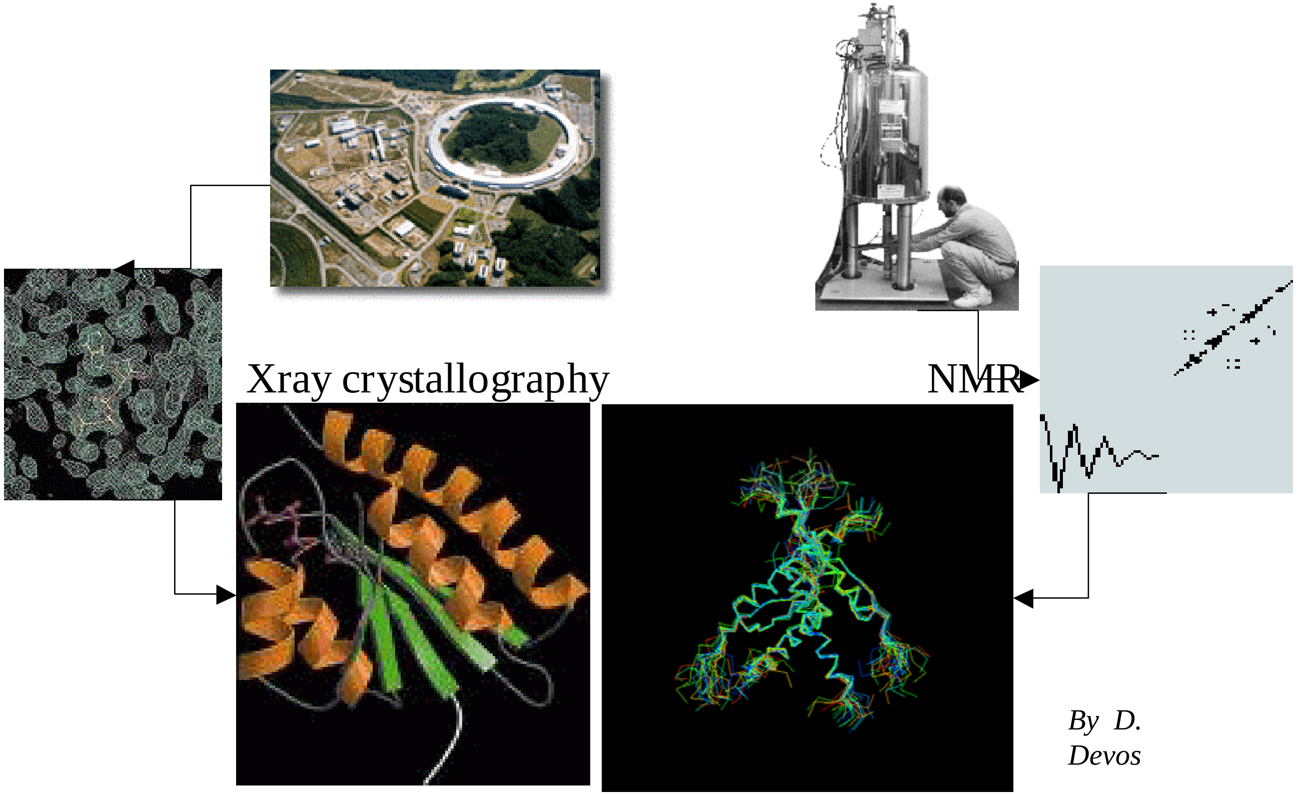
genbank

swissprot

PDB



Determining protein structures
Low-throughput ("traditional") approach

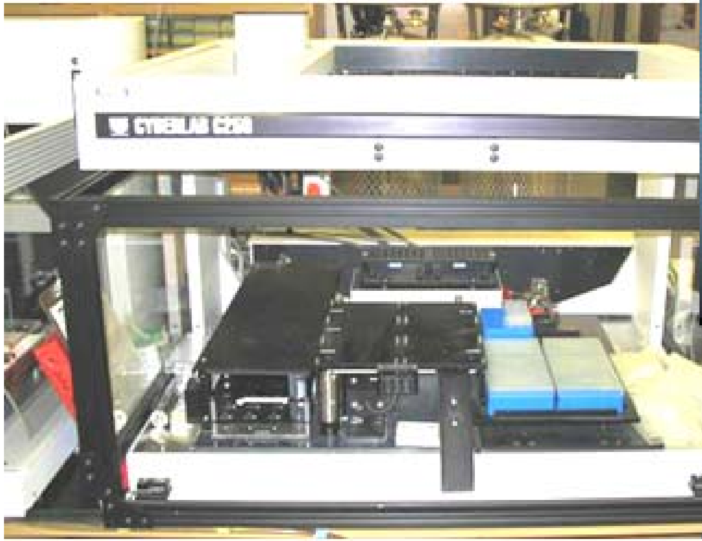


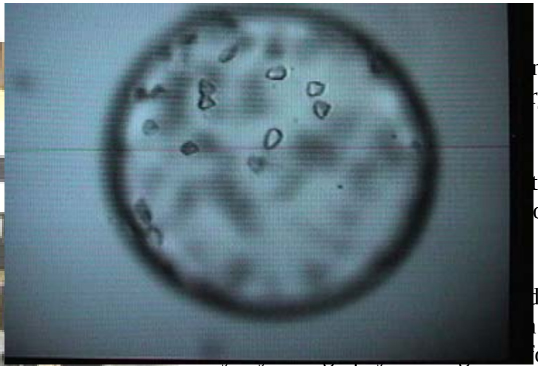
Xray crystallography

NMR

By D. Devos

Determining protein structures
High-throughput approach – Structural Genomics





nt of two minimal genomes, Mycoplasma
rystallography are being used for structural

ture determination of biologically important
ologically important proteins in Arabidopsis. The

d the main proteins of interest are signaling
Thermotoga maritima, and creating a high-
or structural determination.

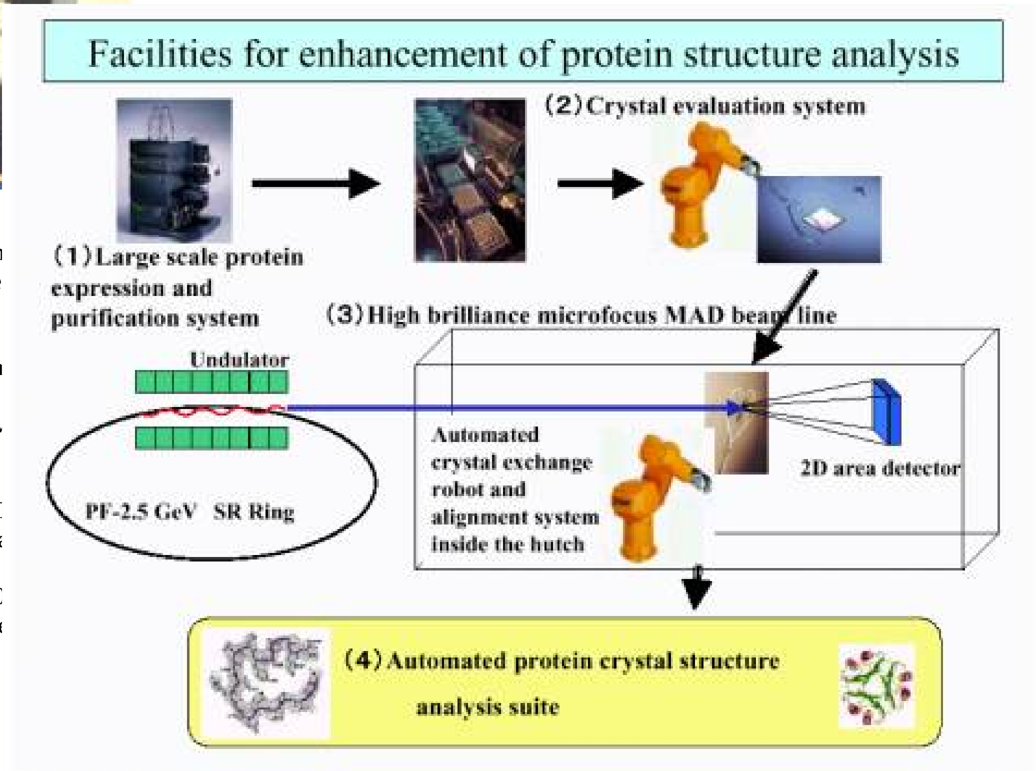
determination by X-ray crystallography.

•The Northeast Structural Genomics Consortium (NEGS)
The NEGS is focused on human proteins and proteins from eukaryotic n
also proteins that are interesting from a functional genomics perspective
spectroscopy.

•The Southeast Collaboratory for Structural Genomics (SECSG)
The objective of the SECSG is to develop and test experimental and con
crystallography and NMR methods and to apply these strategies to scan
Homo sapiens and an ancestrally-related prokaryotic microorganism hav

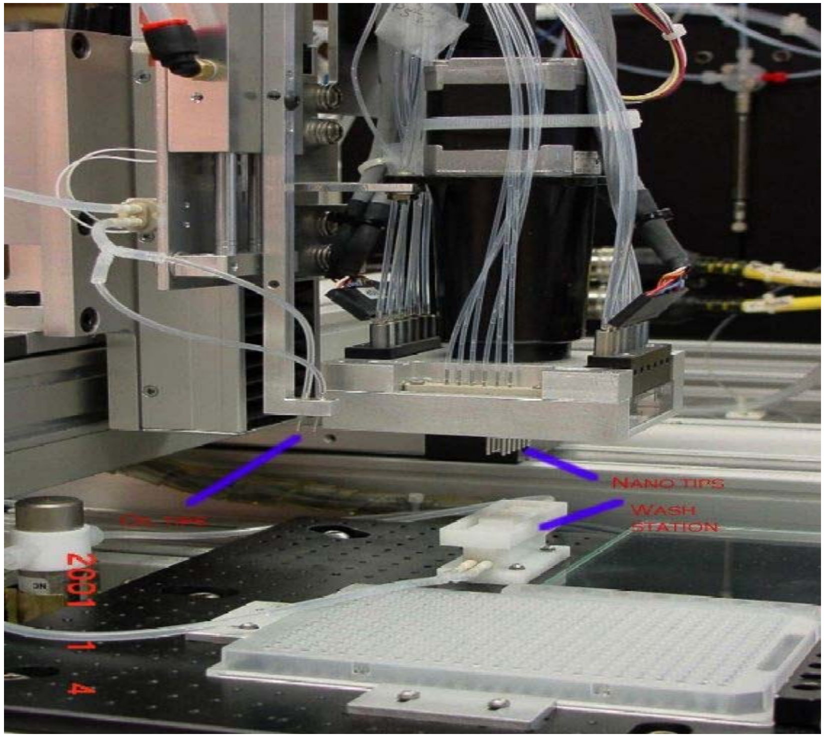
•Structural Genomics of Pathogenic Protozoa Consortium (SGPP)
The SGPP consortium aims to determine and analyze the structures of a
Trypanosoma brucei, *Trypanosoma cruzi* and *Plasmodium falciparum*. T
and malaria. X-ray crystallography is being used for structural determina

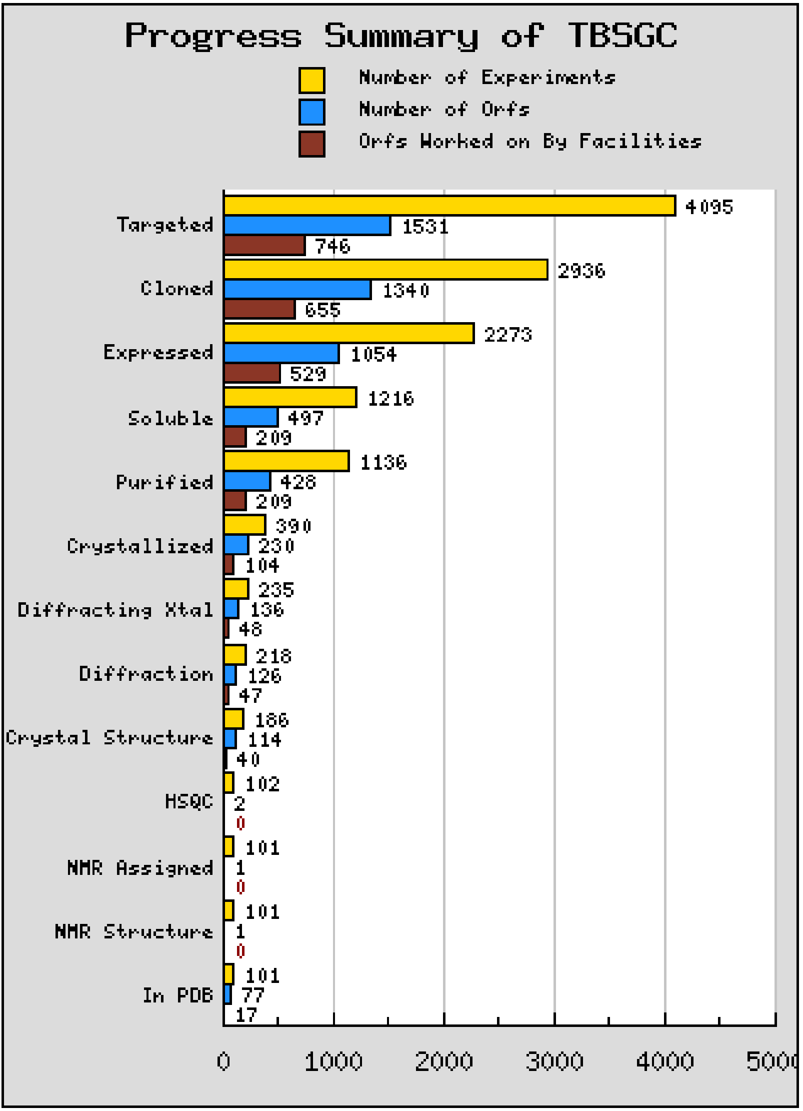
•The TB Structural Genomics Consortium (TB)
The goal of the TB consortium is to determine the structures of over 400
information that currently exists and that is generated by the project. The
protein structures are being determined using X-ray crystallography.


Facilities for enhancement of protein structure analysis
(1) Large scale protein expression and purification system
(2) Crystal evaluation system
(3) High brilliance microfocus MAD beam line
Undulator
PF-2.5 GeV SR Ring
Automated crystal exchange robot and alignment system inside the hutch
2D area detector
(4) Automated protein crystal structure analysis suite

Goldsmith-Fischman, S. and Honig, B. (2003) Structural genomics: Computational methods for structure analysis. *Protein Sci*, **12**, 1813-1821.

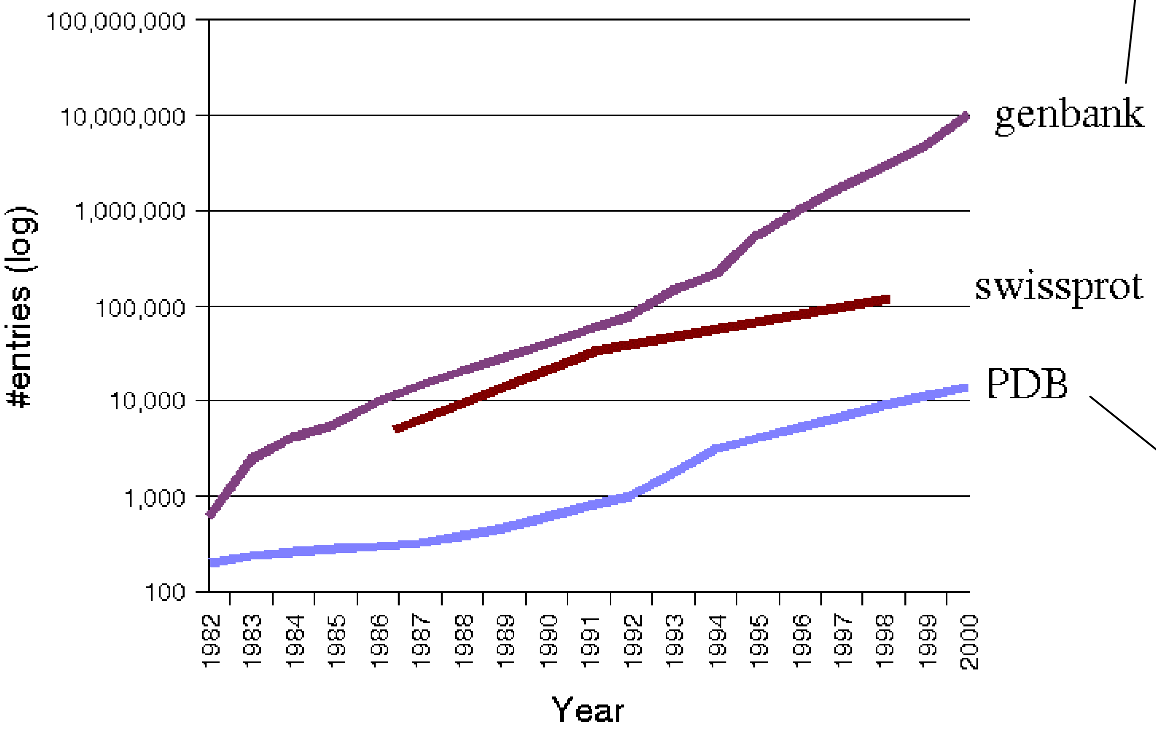
Structural Genomics





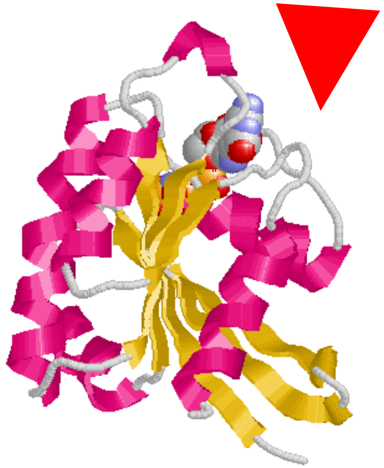
Vitkup, D., Melamud, E., Moult, J. and Sander, C. (2001) Completeness in structural genomics. *Nat Struct Biol*, **8**, 559-566.

Protein Structure

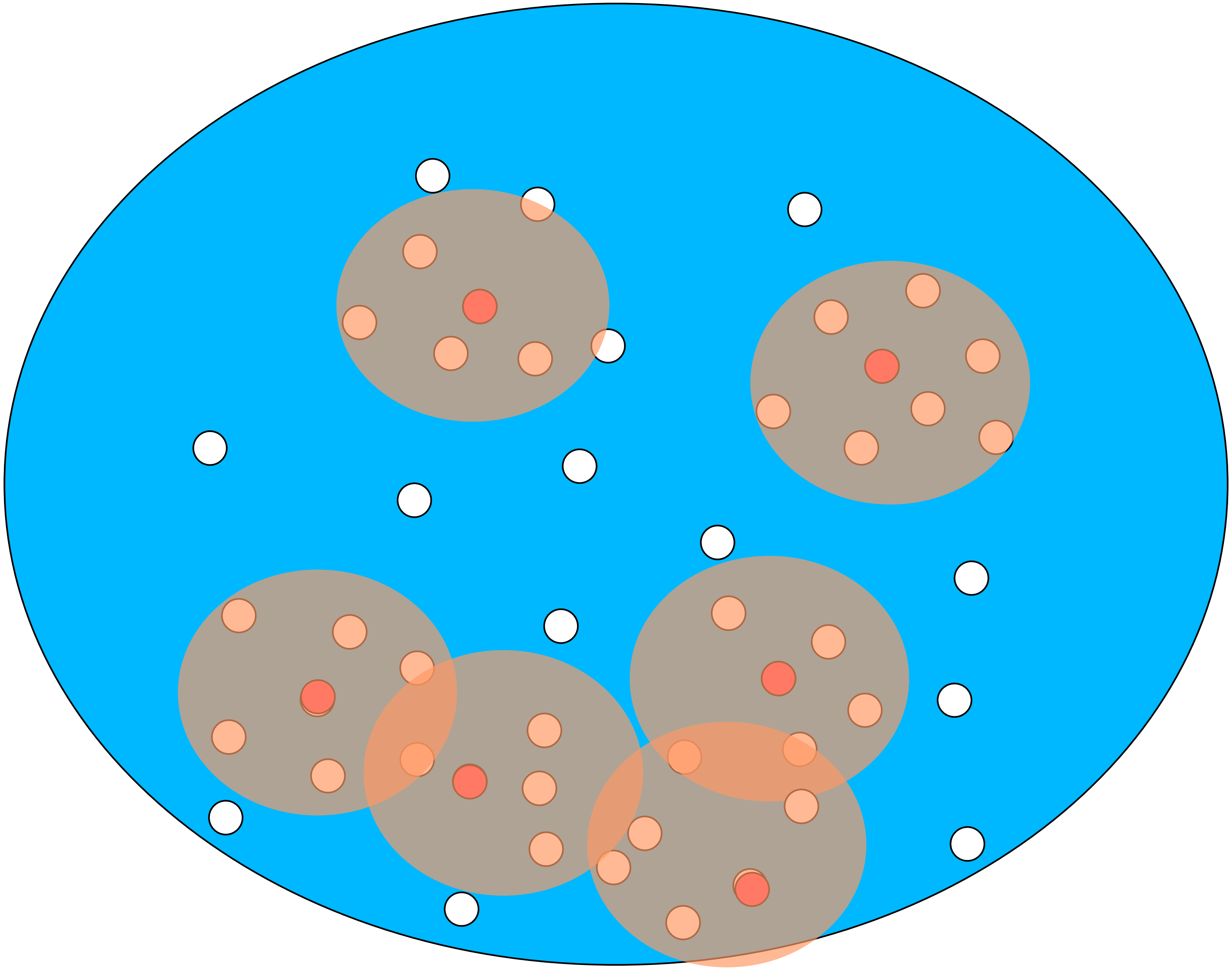


MREYKLVVLGSGGVGKSALTVQFVQGIFVDEYDPTIEDSY
RKQVEVDCQQCMLEILDTAGTEQFTAMRDLYMKNGQGFAL
VYSITAQSTFNDLQDLREQILRVKDTEDVPMILVGNKCDL
EDERVVGKEQGQNLARQWCNCAFLESSAKSKINVNEIFYD
LVRQINR

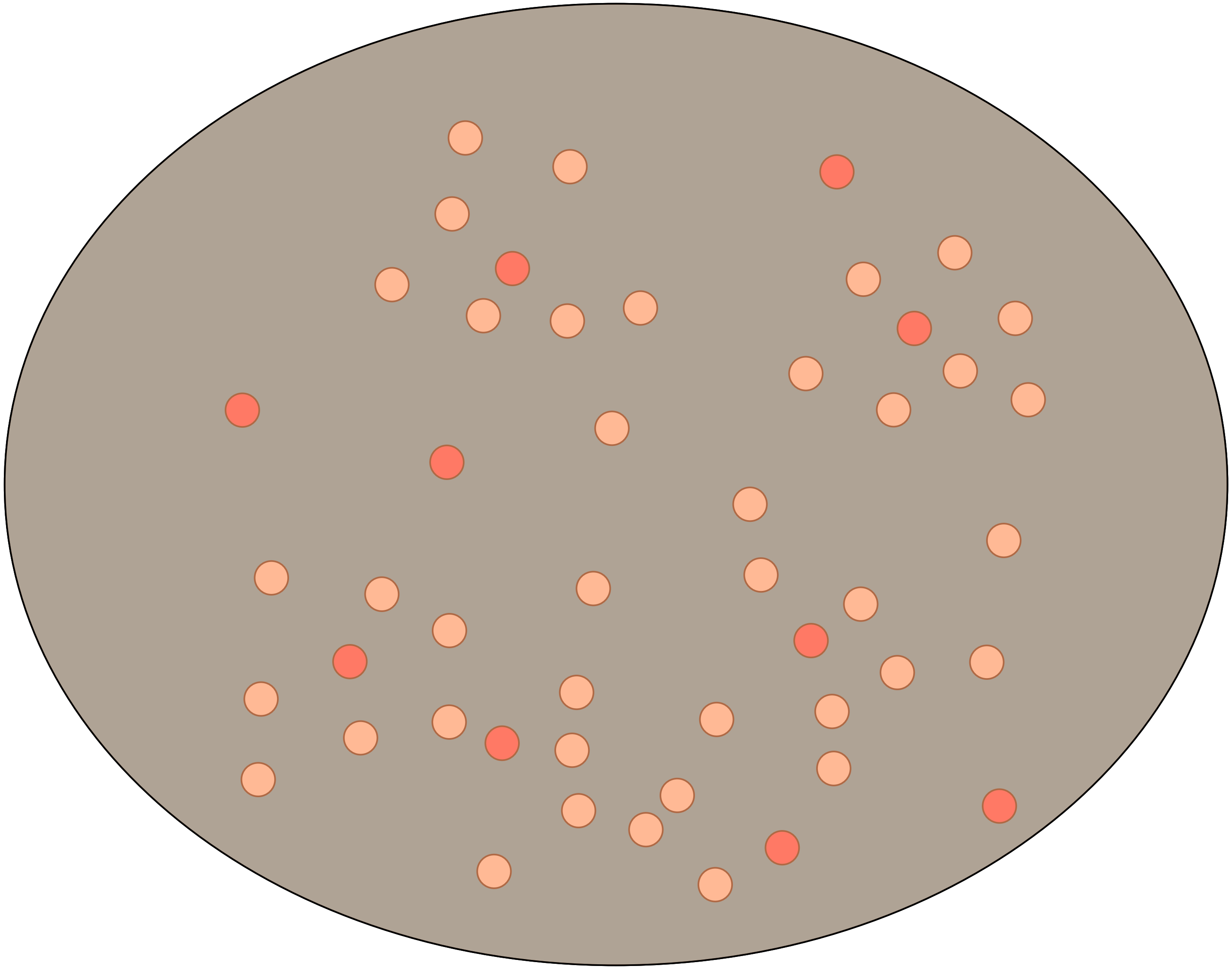
MLEILDTA**GTEQFT**AMRDLYMKNGQGFAL
VYSITAQSTFND**LQDLREQIL**RVKDTEDVPMIL
VGNKCDLEDERV



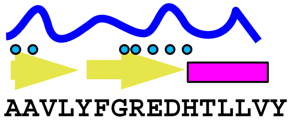
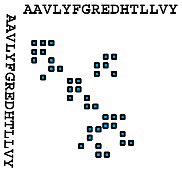
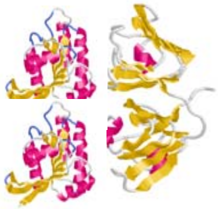
Structural Genomics and Protein Structure Prediction



Structural Genomics and Protein Structure Prediction



Protein Structure Prediction
Classification of Prediction Methods

Nivel estructura proteínas	Secundaria	--------	terciaria	cuaternaria
Representación de la proteína Uso de información extra	1D  	2D  	3D 	4D 
Ab Initio	pred. str. secundaria	mutaciones correlacionadas	- dinámica molecular - minimización de energía	*docking*
No Ab-Initio	pred. str. secundaria		- modelado por homología - *threading*	*docking con filtros*

Protein Structure Prediction

1D Characteristics

1D Characteristics: Features that can be represented by a single value associated to each residue (B. Rost).

These values can be labels representing “states”, like in secondary structure (H: helix, E: beta, …). They can also be continuous values (% accesible surface, …).

Some 1D characteristics:

Secondary structure

Solvent accessibility

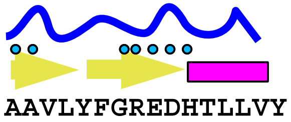
Post-transcriptional modifications

signal peptides

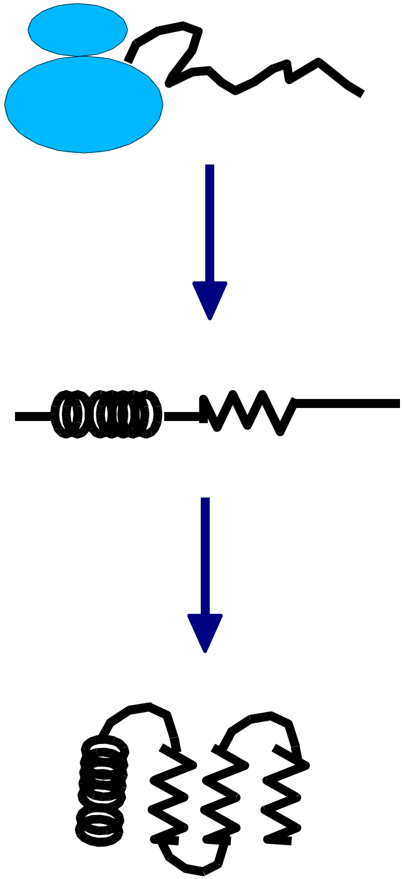
Coiled-coils

Unstructured regions

etc.



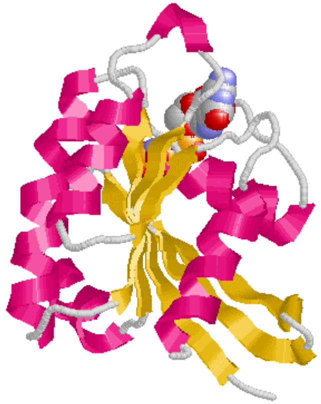
1D Characteristics - Secondary Structure

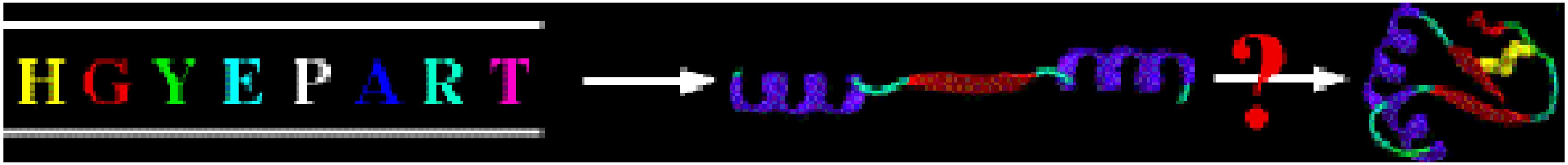


Primary structure

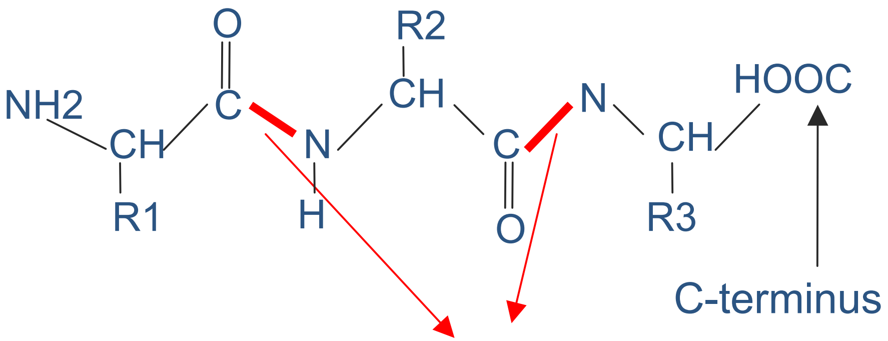
Secondary structure

Tertiary structure





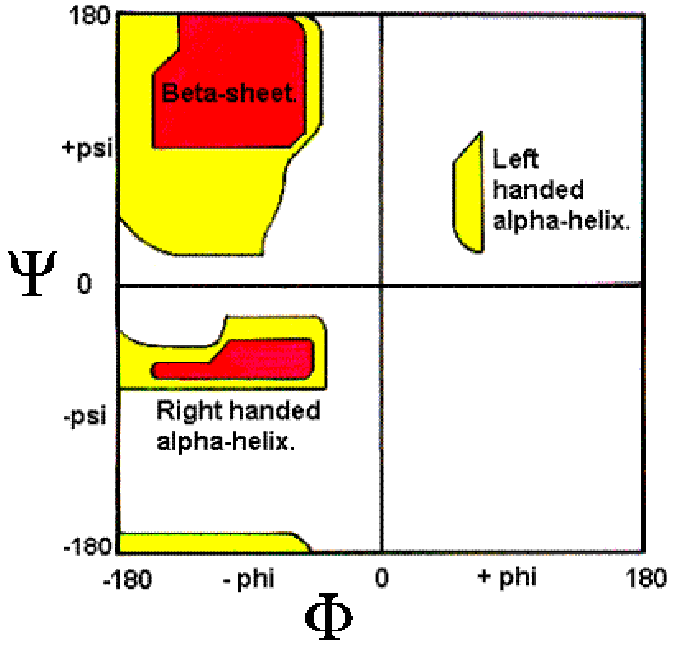
1D Characteristics
Secondary Structure

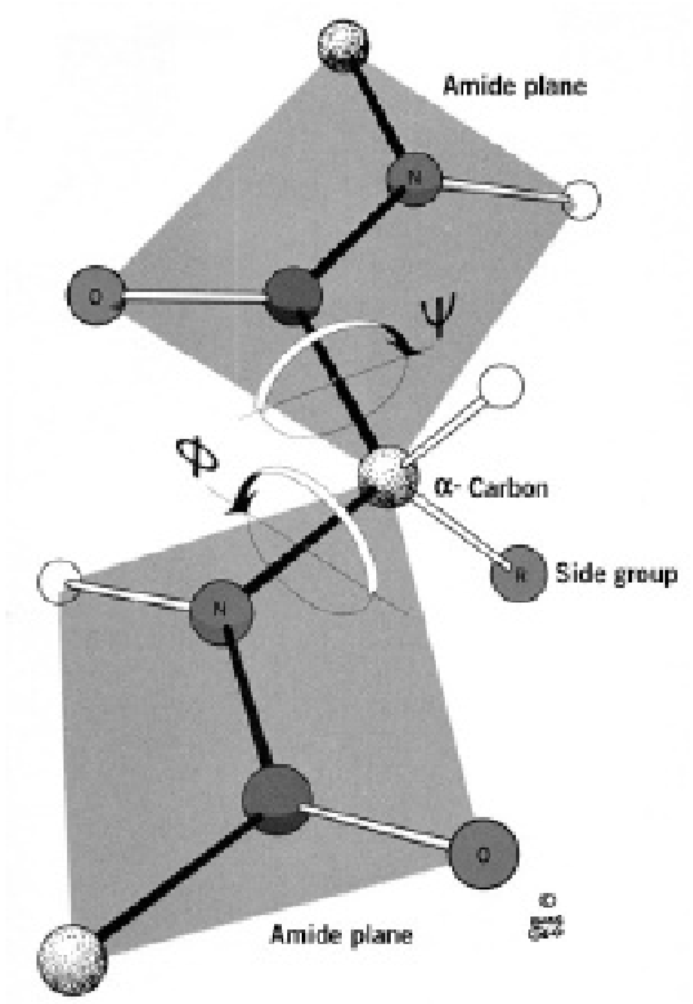


C-terminus

Peptide bonds

The Ramachandran Plot.





1D Characteristics
Secondary Structure

```
  1 ASKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTT
      TTGGGGSSEEEEEEEEEEEETTEEEEEEEEEEEETTTTEEEEEEEETT

 51 GKLPVPWPTLVTTFSYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTIFF
    SS SS GGGGHHHHSSS GGG B  GGGGGG HHHHTTTT EEEEEEEEE

101 KDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNV
    TTS EEEEEEEEEEETTEEEEEEEEEEE   TTSTTTTT B S    EEE

151 YIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHY
    EEEEEGGGTEEEEEEEEEEEETTS EEEEEEEEEEEESSSS      SEE

201 LSTQSALSKDPNEKRDHMVLLEFVTAAGIT HGMDELYK
    EEEEEEEE  TT  SSEEEEEEEEEEES
```

Definition: T=hydrogen bond turn, H=helix, G=310 helix, I=phi helix, B=residue in isolated beta bridge, E=strand, and S=bend

Prediction: H/E/T (3 states only)

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Secondary Structure
First-generation Methods

Statistical methods simply based on the tendency of each aminoacid to form each type of secondary structure.

- Chou & Fasman en 1974, proposed the first method. They calculated the tendencies from the **15 structures** solved. Later, this method showed a reliability of 57% when tested on 62 proteins. (=> close to random)
- Garnier (1978), calculated these probabilities for pairs of residues, improving the reliability (~60%)

Chou, P.Y. and Fasman, G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 222-244/225.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97-120.

Secondary Structure
First-generation Methods

Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.110	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic Acid	151	037	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.120	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

Glu, Met Ala y Leu : strong tendency to form helix.
Val, Ile y Tyr: strong tendency to form strand.

Chou, P.Y. and Fasman, G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 222-244/225.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97-120.

Secondary Structure
Second-generation Methods

- Their main characteristic is the usage of a window of adjacent residues, so that context information is used for the prediction.
- Many algorithms (fed with this contextual information) were used (Neural networks, graph theory, rule-based systems, multivariate analysis, …)
- This innovation improve the accuracy close to 70%.

- Limitations
  - Accuracy (< 70% - 3 states -)
  - Low accuracies for β–strands.
  - Tendency to predict short secondary structure elements (both α and β).
  - Due to:
  - The number of known structures (for training) is still low and they do not cover the space of sequences.
  - Long range interactions (residues far apart in the sequence but close in 3D) are not taken into account.

Garnier, J. and Robson, B. (1989) The GOR method for predicting secondary structure in proteins. In D., F.G. (ed.), *Prediction of protein structure and the principles of protein conformation*. Plenum Press, New York, pp. 417-465

Secondary Structure
Third-generation Methods

Initiated by Levin (~69%) and Rost & Sander (PHD 72%)

- The main novelty is the inclusion of evolutionary information in the form of multiple sequence alignments (profiles – Levin, 1993).
- The problem with the bad predictions for β-strands is solved by balancing the training set since 3D structures contain more α than β (Rost y Sander, 1994)
- For methods based on NN, a second network is used to smooth the predictions and avoid short elements.
- This breaks the 70% limit.

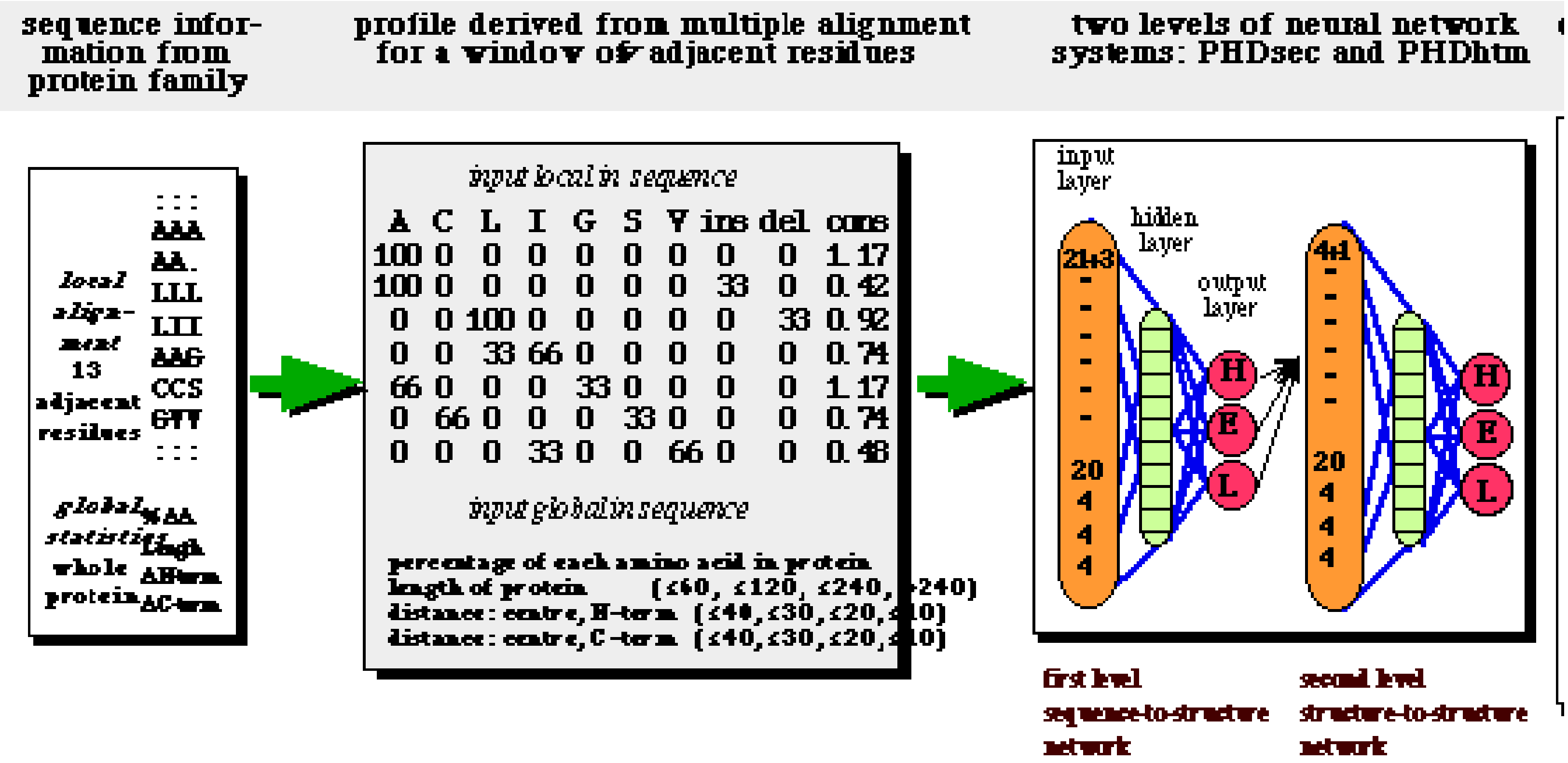
Levin JM, Pascarella S, Argos P, Garnier J. (1993). Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng.* **6(8):**849-54.

Rost, B. and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A*, **90**, 7558-7562.

Rost, B., Sander, C. and Schneider, R. (1994) PHD - A mail server for protein secondary structure prediction. *Comp. Applic. Biosci.*, **10**, 53-60.

Secondary Structure
Third-generation Methods

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

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)

:::
AAA
AA.
local alignment 13 adjacent residues
LLL
LII
AAG
CCS
GVV
:::
global statistics whole protein
%AA
length
AN-term
AC-term
input layer
hidden layer
output layer
21+3
20
4
4
4
H
E
L
4+1
20
4
4
4
H
E
L
first level sequence-to-structure network
second level structure-to-structure network

Rost, B. and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A*, **90**, 7558-7562.

Rost, B., Sander, C. and Schneider, R. (1994) PHD - A mail server for protein secondary structure prediction. *Comp. Applic. Biosci.*, **10**, 53-60.

Secondary Structure
Third-generation Methods

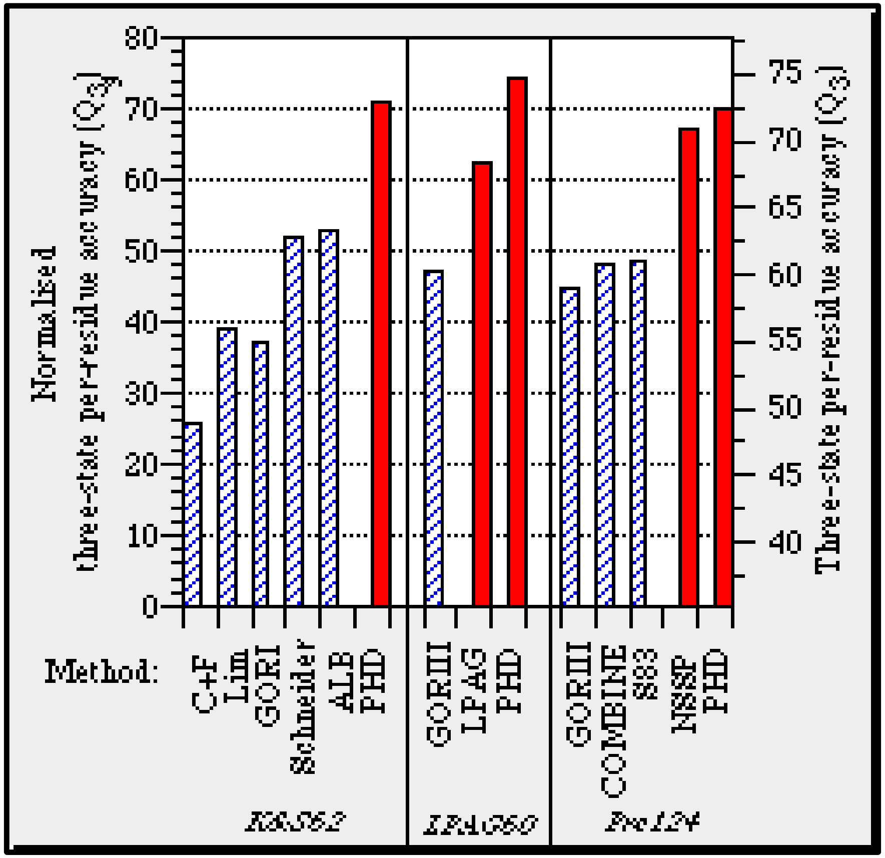
- Most forthcoming methods followed PHD's srategy, improving the results basically by improving the input multiple sequence alignment (including remote homologues (PSI-BLAST), filtering, ...). *PSIPRED* (1999) ~77%, HMMs used by Kevin Karplus *et al*. in *SAMT99sec* (1999).

- The other main strategy is con combine predictions coming from different methods (consensus methods). *Jpred2* (Cuff y Barton, 2000).

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**, 195-202.

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics*. **14(10):**892-3.

Secondary Structure Prediction



1st generation methods: Chou & Fasman, Lim, GORI

2nd generation methods : Schneider, ALB, GORIII

3rd generation methods: LPAG, COMBINE, S83, NSSP, PHD

76-78%

Accuracy limit?

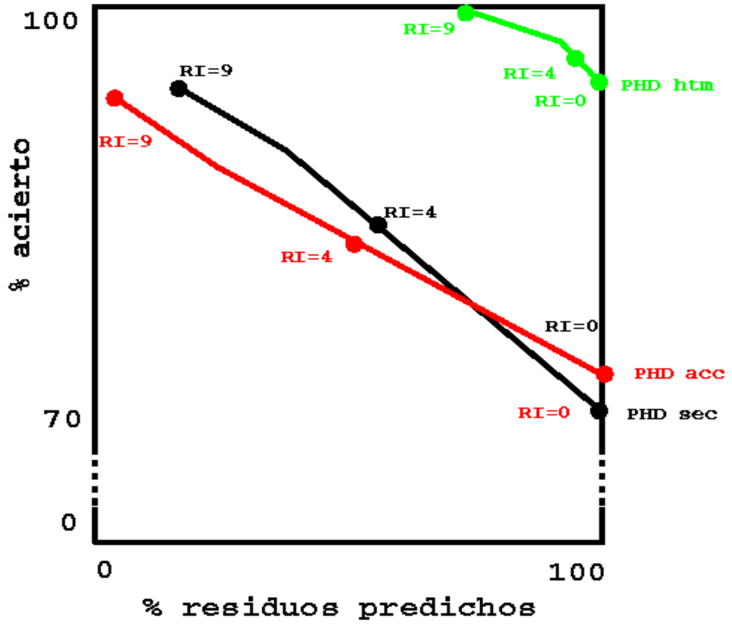
- Limit in the definition of secondary structure
 (DSSP vs. others)
- Limit in the local information

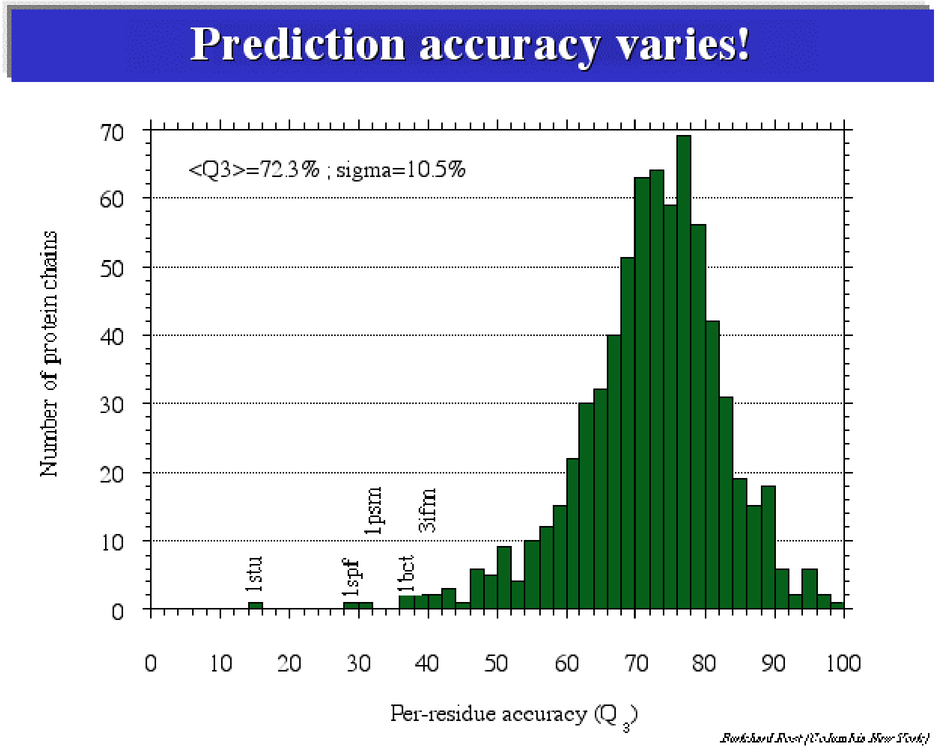
Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Secondary Structure Prediction
Things to take into account

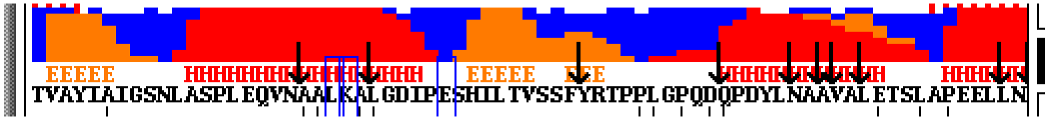
Balance accuracy/coverage

Results vary from one protein to another



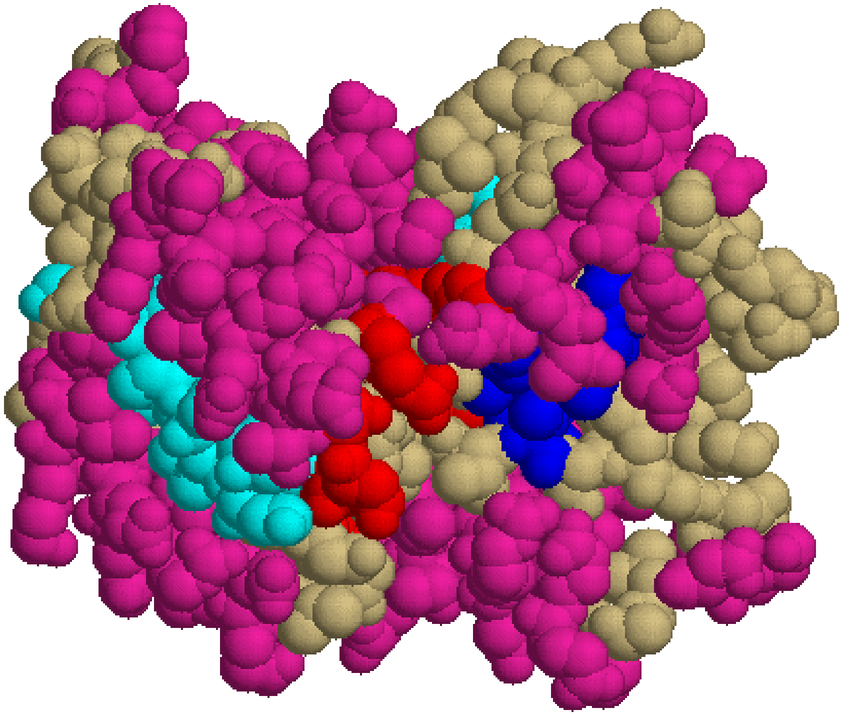


Burkhard Rost (Columbia New York)

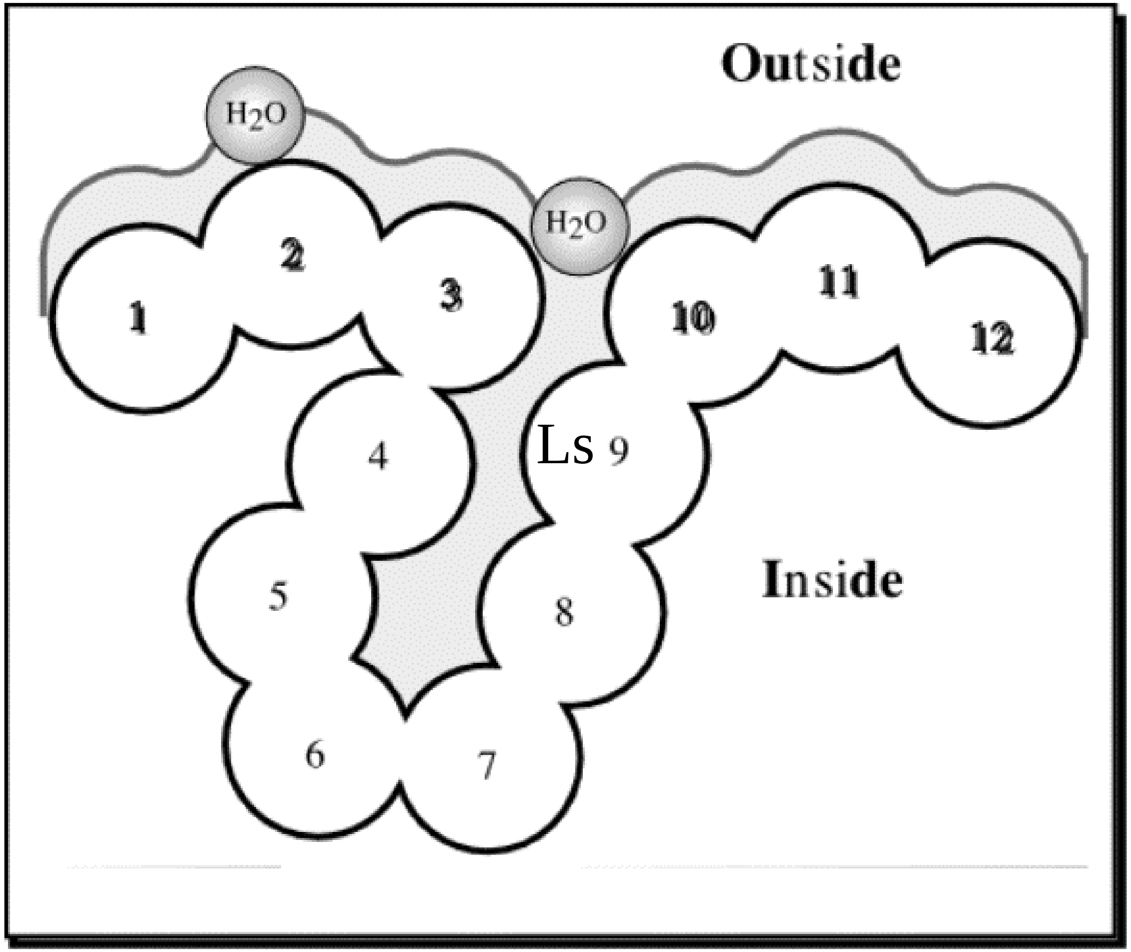

EEEEE
HHHHHHHHHHHHHHHHH
EEEEE
EEE
HHHHHHHHHHHH
HHHHHHH
TVAYIAIGSNLASPLEQVNAALKALGDIPESHILTVSSFYRTPPLGPQDQPDYLNAAVALETSLAPEELLN

1D Methods
Solvent Accessibility

- Model discrimination
- Functional sites / binding sites
- Mutant design, protein labeling, etc.



1D Methods
Solvent Accessibility



Programs for defining accessibility report (from the 3D structure) the accessible surface of each residue in Å^2.

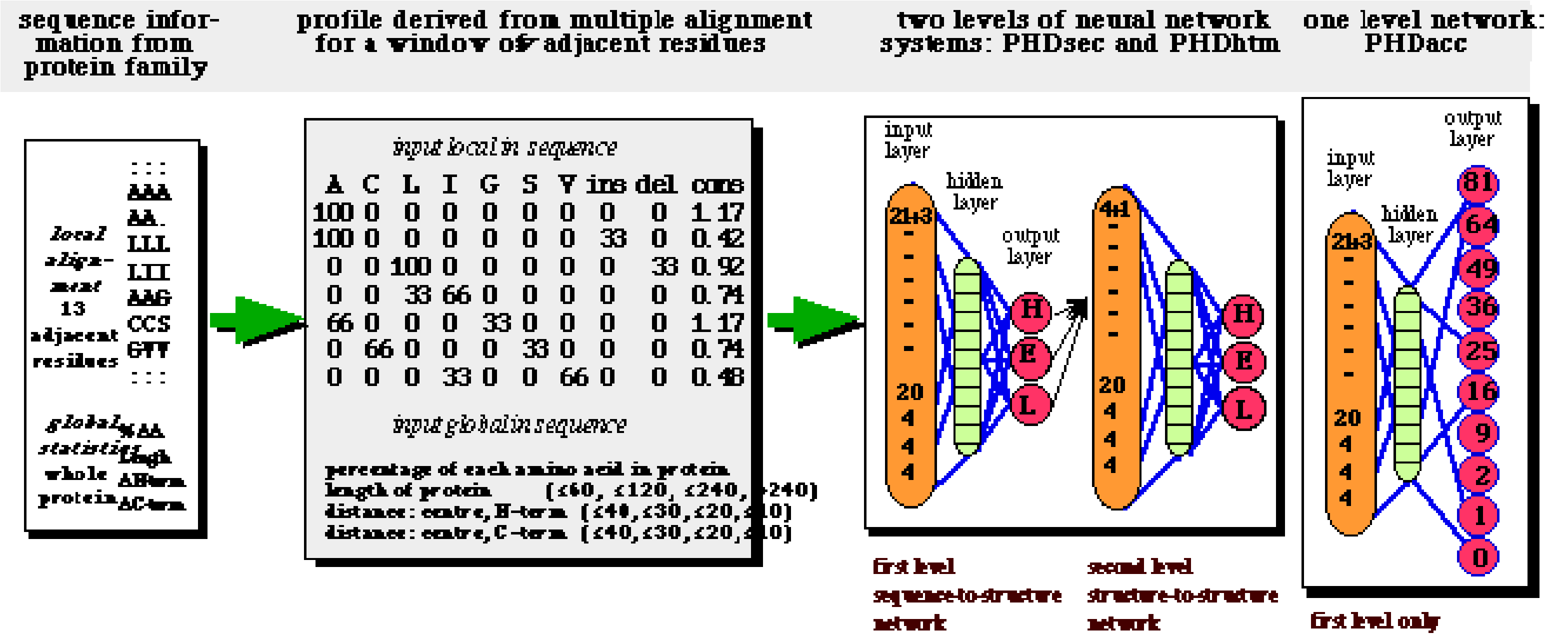
Most prediction methods reduce the problem by considering only 2 states: buried (rel. accs. <16%, abs <50 Å^2) and exposed (rel. accs. >= 16%, abs >=50 Å^2).

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

1D Methods
Solvent Accessibility

- Same “history” as secondary structure: frequencies (tendencies) -> windows -> neural networks + evolutionary information (alns.) / consensus.

- Usually the programs are the same, with small adaptations of the NN for the representation of accessibility.



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
: : :
AAA
AA.
local alignment 13 adjacent residues
LLL
LII
AAG
CCS
GVV
: : :
global statistics whole protein
%AA
length
ΔN-term
ΔC-term
input local in sequence
input global in sequence
input layer
hidden layer
output layer
21*3
20
4
4
4
H
E
L
4*1
81
64
49
36
25
16
9
2
1
0
first level sequence-to-structure network
second level structure-to-structure network
first level only

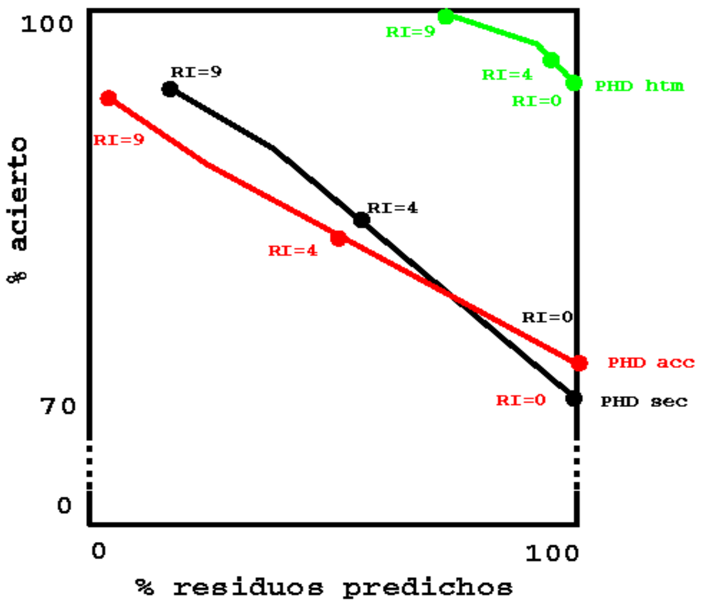
A	C	L	I	G	S	V	ins	del	cons
100	0	0	0	0	0	0	0	0	1.17
100	0	0	0	0	0	0	33	0	0.42
0	0	100	0	0	0	0	0	33	0.92
0	0	33	66	0	0	0	0	0	0.74
66	0	0	0	33	0	0	0	0	1.17
0	66	0	0	0	33	0	0	0	0.74
0	0	0	33	0	0	66	0	0	0.48

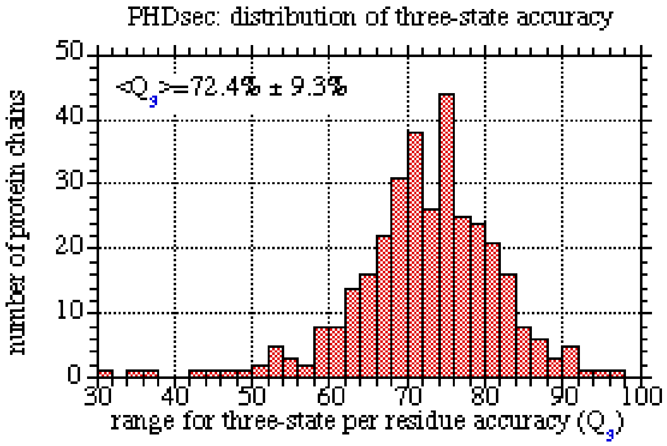
percentage of each amino acid in protein
length of protein (≤60, ≤120, ≤240, >240)
distance: centre, N-term (≤40, ≤30, ≤20, ≤10)
distance: centre, C-term (≤40, ≤30, ≤20, ≤10)

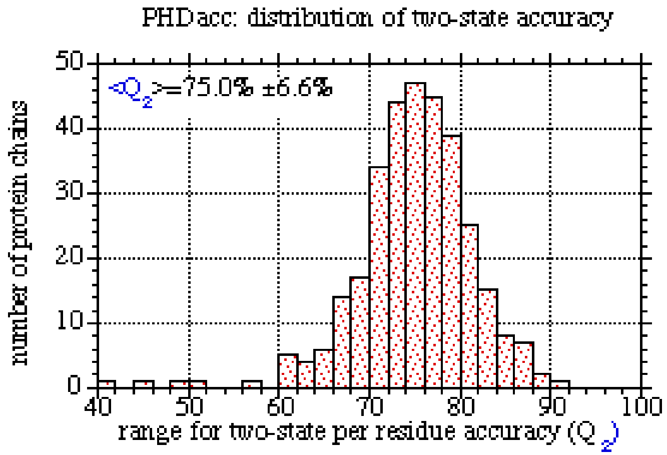
Rost, B. and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A*, **90**, 7558-7562.

Rost, B., Sander, C. and Schneider, R. (1994) PHD - A mail server for protein secondary structure prediction. *Comp. Applic. Biosci.*, **10**, 53-60.

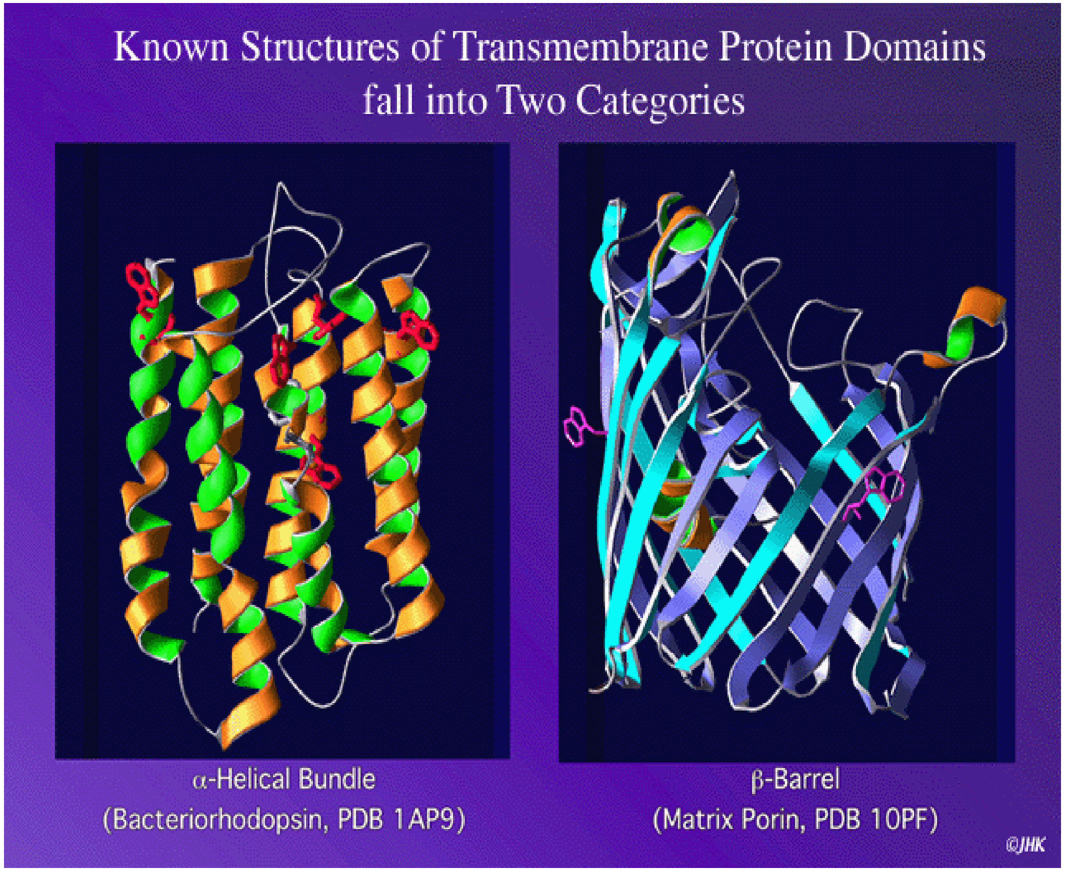
1D Methods
Solvent Accessibility







1D Methods
Transmembrane segments



Known Structures of Transmembrane Protein Domains fall into Two Categories

α-Helical Bundle
(Bacteriorhodopsin, PDB 1AP9)

β-Barrel
(Matrix Porin, PDB 1OPF)

©JHK

-Difficult to crystalize. Few structures

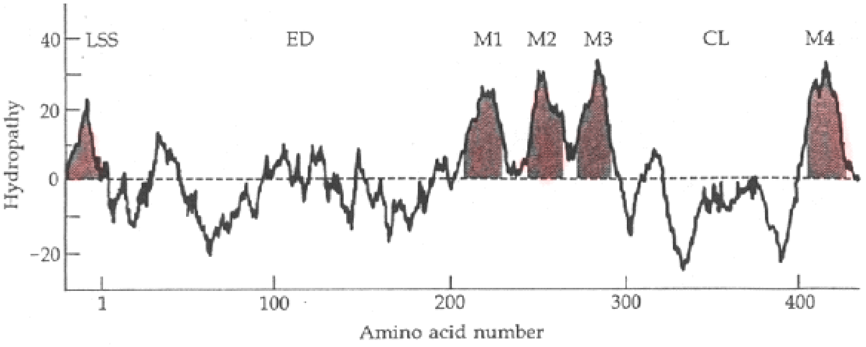
- Preliminary information on domains, functional areas, etc.

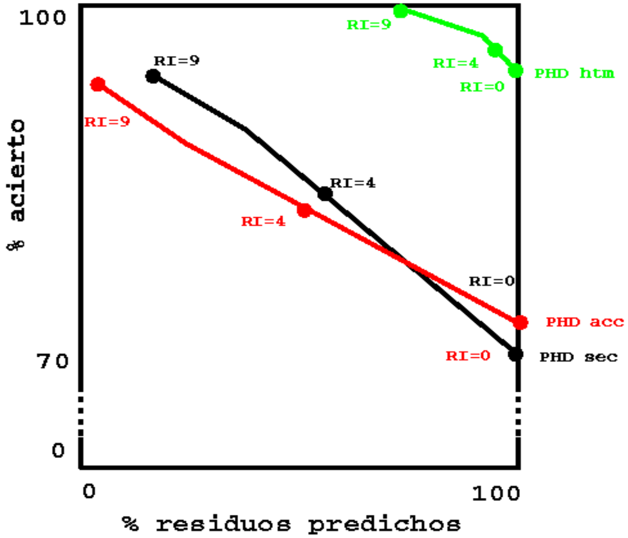
1D Methods
Transmembrane helices

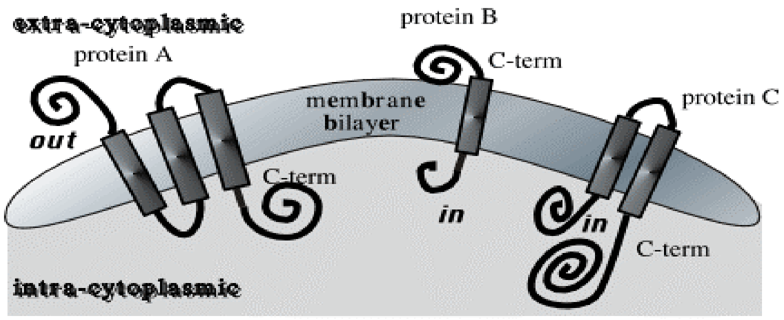
- 20-30 residues.
- hydrophobic.
- charged cytoplasmic loops, ...

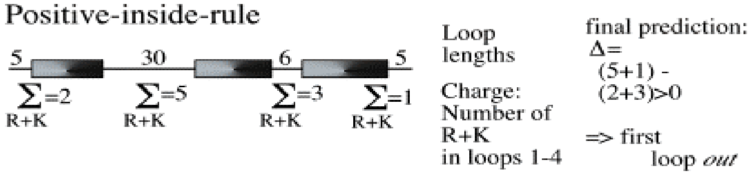
Clear characteristics => easy to “learn”

Same NN as for sec. and acc.









1D Methods
Transmembrane helices

MEMSAT - *http://bioinf.cs.ucl.ac.uk/psipred/*

TMAP - http://www.mbb.ki.se/tmap/index.html

TopPred2 - http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html

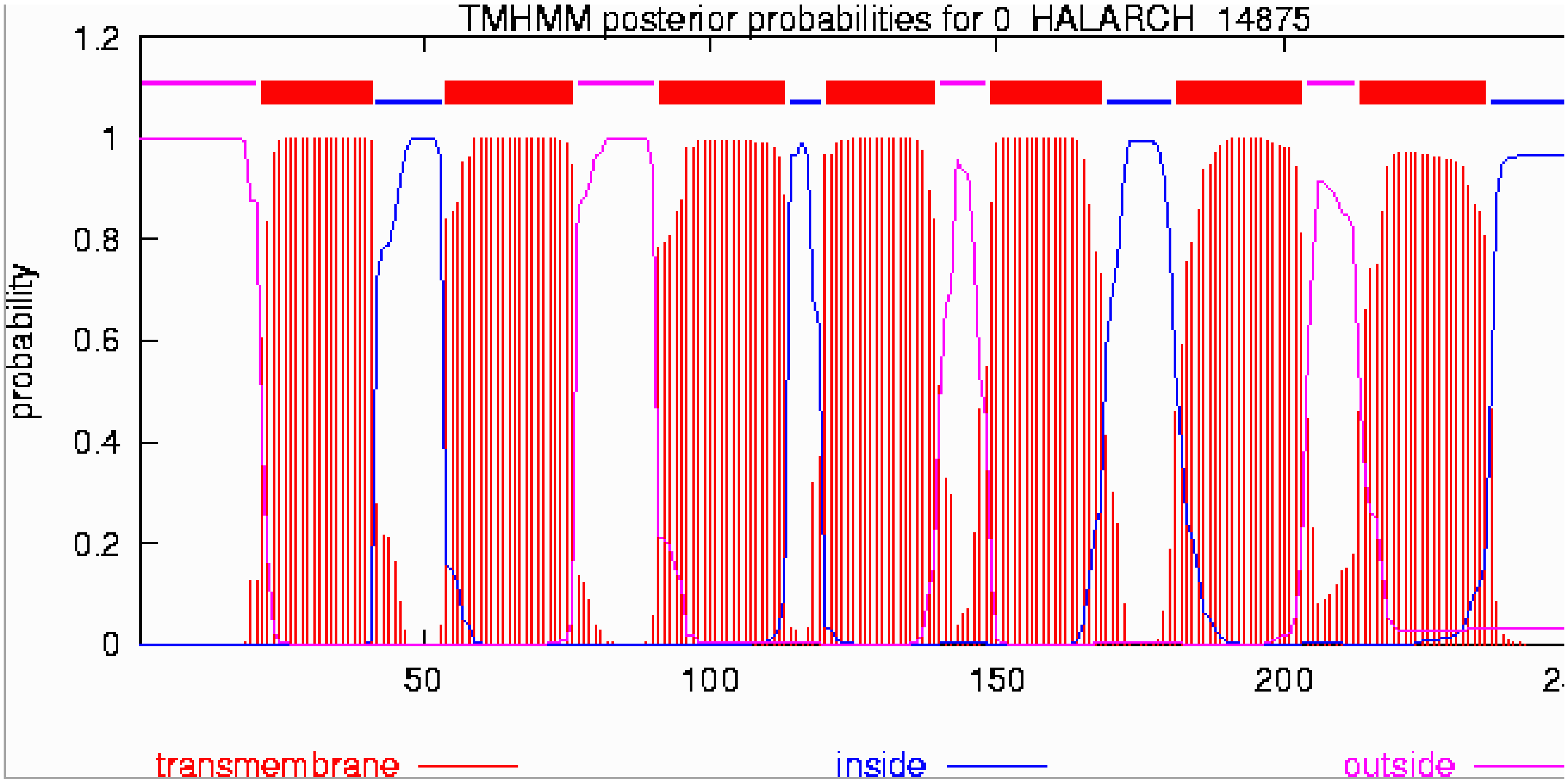
HMMTOP - http://www.enzim.hu/hmmtop/

PHDhtm - *http://www.embl-heidelberg.de/predictprotein/*

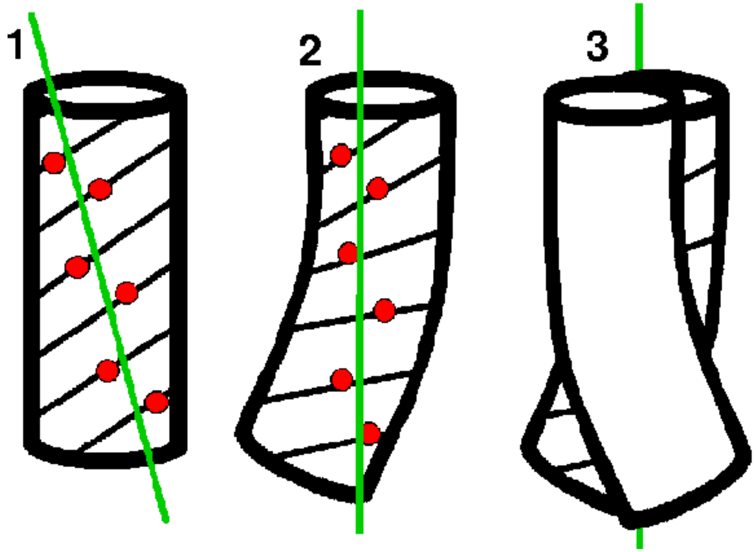
DAS - *http://www.enzim.hu/DAS/DAS.html*

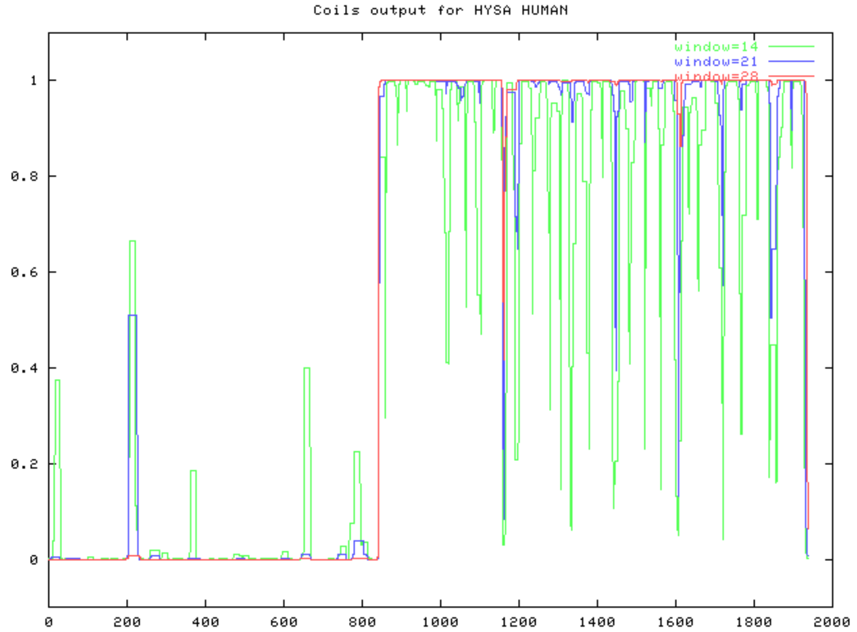
TMHMM - *http://www.cbs.dtu.dk/services/TMHMM/*

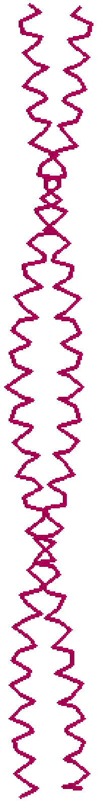
1D Methods
Transmembrane helices



Coiled-coils



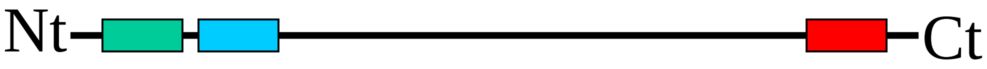


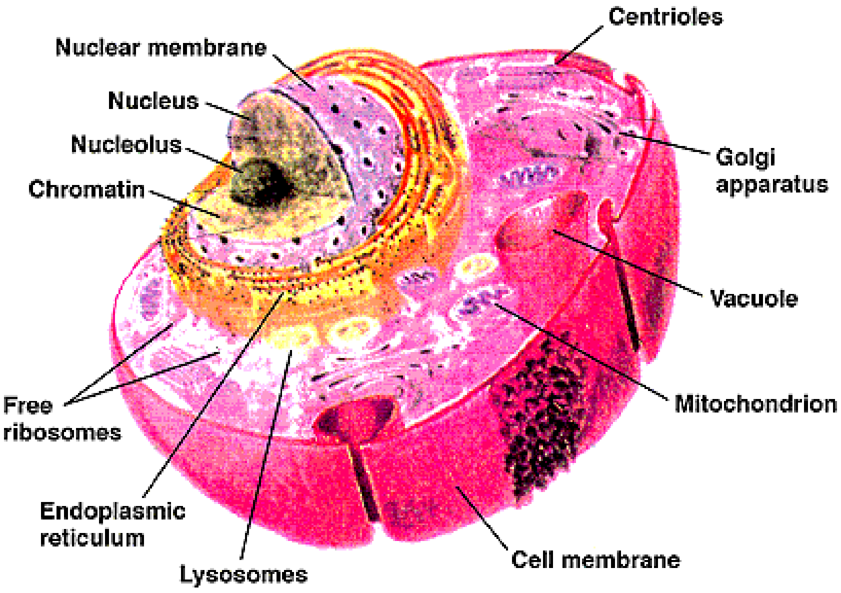


[**a**bc**d**efg]$_n$

Lupas, A., Dyke, M.v. and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162-1164.

Sorting signals - *PSORT*





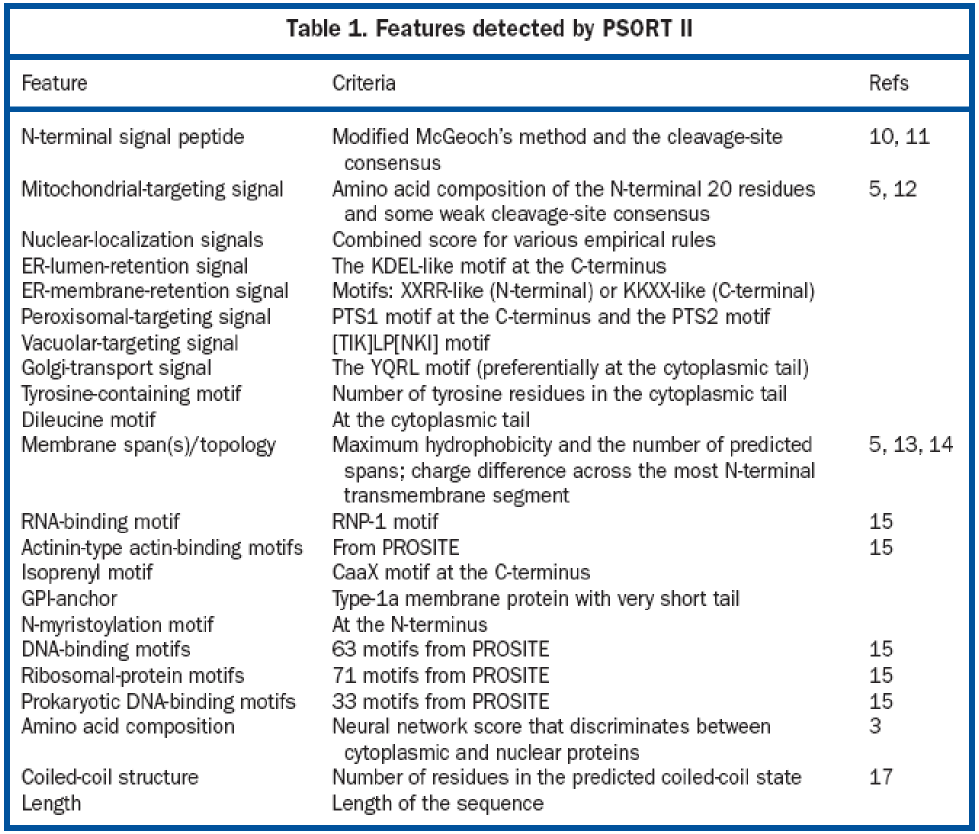
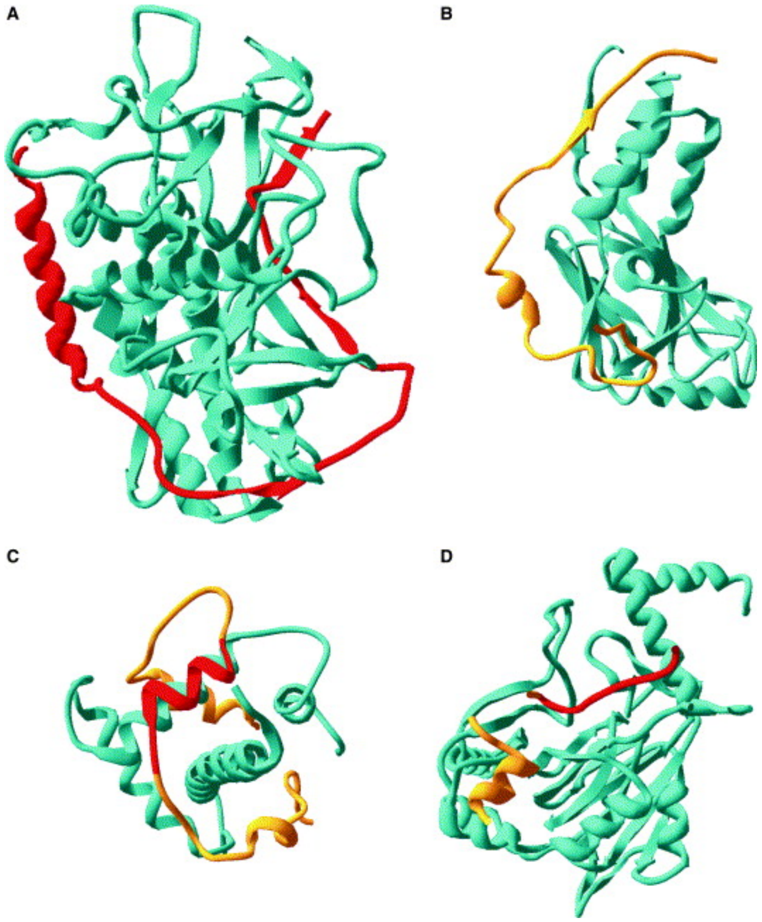


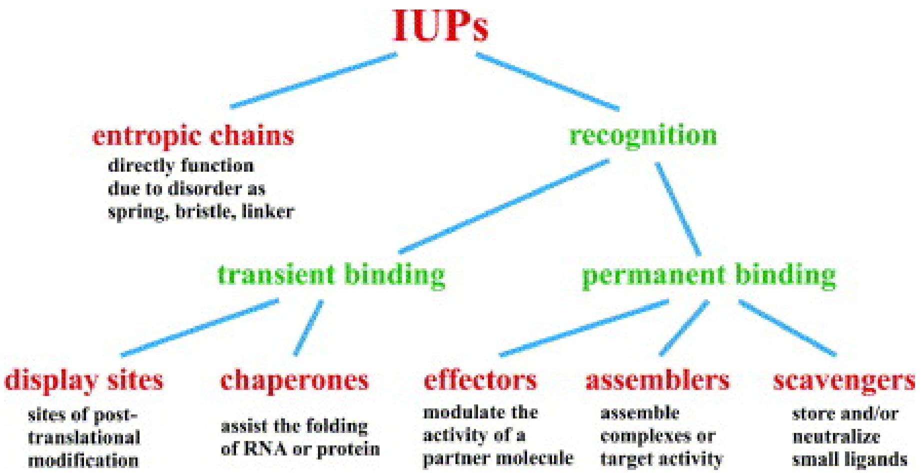
Table 1. Features detected by PSORT II

Feature	Criteria	Refs
N-terminal signal peptide	Modified McGeoch's method and the cleavage-site consensus	10, 11
Mitochondrial-targeting signal	Amino acid composition of the N-terminal 20 residues and some weak cleavage-site consensus	5, 12
Nuclear-localization signals	Combined score for various empirical rules	
ER-lumen-retention signal	The KDEL-like motif at the C-terminus	
ER-membrane-retention signal	Motifs: XXRR-like (N-terminal) or KKXX-like (C-terminal)	
Peroxisomal-targeting signal	PTS1 motif at the C-terminus and the PTS2 motif	
Vacuolar-targeting signal	[TIK]LP[NKI] motif	
Golgi-transport signal	The YQRL motif (preferentially at the cytoplasmic tail)	
Tyrosine-containing motif	Number of tyrosine residues in the cytoplasmic tail	
Dileucine motif	At the cytoplasmic tail	
Membrane span(s)/topology	Maximum hydrophobicity and the number of predicted spans; charge difference across the most N-terminal transmembrane segment	5, 13, 14
RNA-binding motif	RNP-1 motif	15
Actinin-type actin-binding motifs	From PROSITE	15
Isoprenyl motif	CaaX motif at the C-terminus	
GPI-anchor	Type-1a membrane protein with very short tail	
N-myristoylation motif	At the N-terminus	
DNA-binding motifs	63 motifs from PROSITE	15
Ribosomal-protein motifs	71 motifs from PROSITE	15
Prokaryotic DNA-binding motifs	33 motifs from PROSITE	15
Amino acid composition	Neural network score that discriminates between cytoplasmic and nuclear proteins	3
Coiled-coil structure	Number of residues in the predicted coiled-coil state	17
Length	Length of the sequence	

Nakai, K & Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*. **24(1):**34-6

Unstructured proteins and protein regions





IUPs

entropic chains
directly function
due to disorder as
spring, bristle, linker

recognition

transient binding

permanent binding

display sites
sites of post-
translational
modification

chaperones
assist the folding
of RNA or protein

effectors
modulate the
activity of a
partner molecule

assemblers
assemble
complexes or
target activity

scavengers
store and/or
neutralize
small ligands

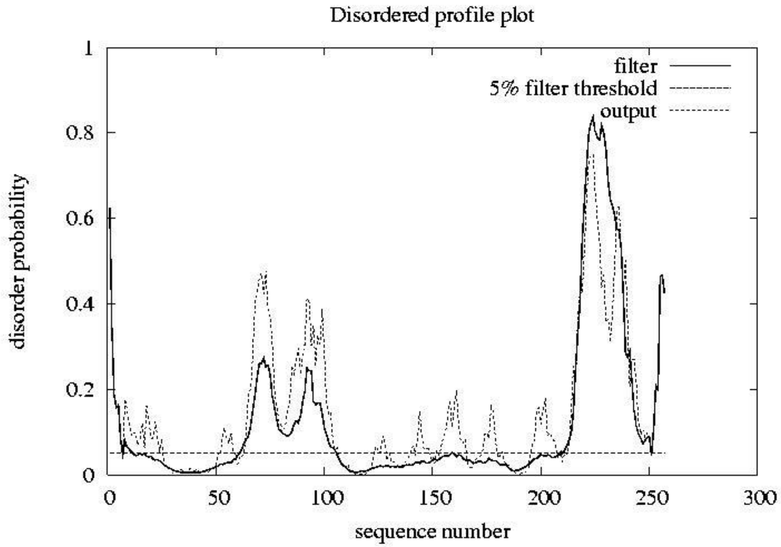
Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett*, **579**, 3346-3354.

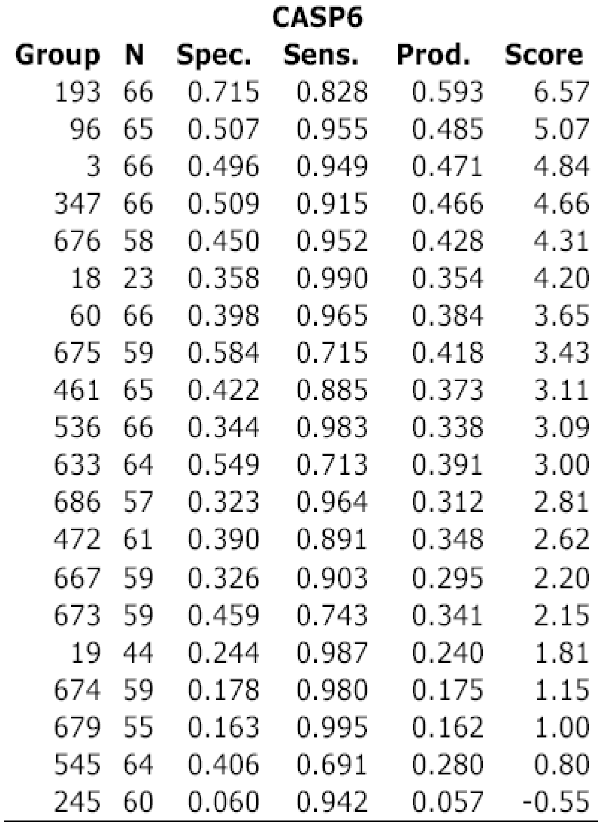
Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. Flavors of protein disorder. *Proteins* **52**, 573-84. (2003).

Unstructured regions
Prediction Methods

<u>Compositionally biased regions</u>. Wootton et al (*SEG*).

<u>Specific for disorder.</u> 003 Jones UCL (David Jones, University College London) support vector machines (*DISOPRED*)





			CASP6		
Group	**N**	**Spec.**	**Sens.**	**Prod.**	**Score**
193	66	0.715	0.828	0.593	6.57
96	65	0.507	0.955	0.485	5.07
3	66	0.496	0.949	0.471	4.84
347	66	0.509	0.915	0.466	4.66
676	58	0.450	0.952	0.428	4.31
18	23	0.358	0.990	0.354	4.20
60	66	0.398	0.965	0.384	3.65
675	59	0.584	0.715	0.418	3.43
461	65	0.422	0.885	0.373	3.11
536	66	0.344	0.983	0.338	3.09
633	64	0.549	0.713	0.391	3.00
686	57	0.323	0.964	0.312	2.81
472	61	0.390	0.891	0.348	2.62
667	59	0.326	0.903	0.295	2.20
673	59	0.459	0.743	0.341	2.15
19	44	0.244	0.987	0.240	1.81
674	59	0.178	0.980	0.175	1.15
679	55	0.163	0.995	0.162	1.00
545	64	0.406	0.691	0.280	0.80
245	60	0.060	0.942	0.057	-0.55

–

Wootton, J.C. and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Meth in Enzym*, **266**, 554-571

Ward, J. J., McGuffin, L. J., Bryson K., Buxton, B. F. & Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20:**2138-2139.

Other 1D characteristics

ExPASy Proteomics tools **http://www.expasy.ch/tools**

COIL – Coiled-coil regions.
PSORT - prediction of signal proteins and localisation sites
SignalP - prediction of signal peptides

ChloroP - prediction of chloroplast peptides
NetOGlyc - prediction of O-glycosilation sites in mammalian proteins
Big-PI - prediction of glycosil -phosphatidyl inositol modification sites
DGPI - prediction of anchor and breakage sites for GPI

NetPhos - prediction of phosphorylation sites (Ser, Thr, Tyr) in eukaryotes
NetPicoRNA - prediction of cleavage sites for proteases in the picornavirus
NMT - prediction of N-miristoilation of N-terminals
Sulfinator - predicts sulphattation sites in tyrosines