BIOINFORMÁTICA Y PROTEÍNAS

EDGAR ANTONIO REYES MONTAÑO

NOVIEMBRE 25 DE 2011

Biology/Chemistry of Protein Structure

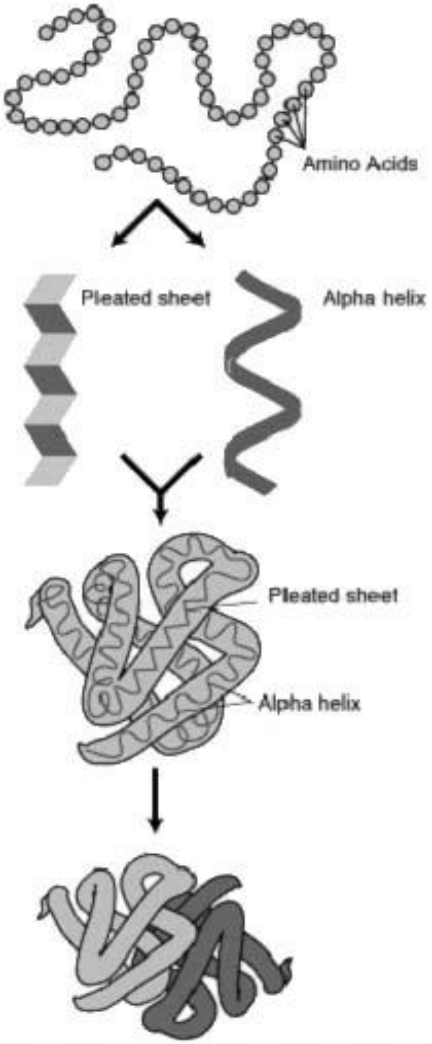
STRUCTURE

Primary

Secondary

Tertiary

Quaternary



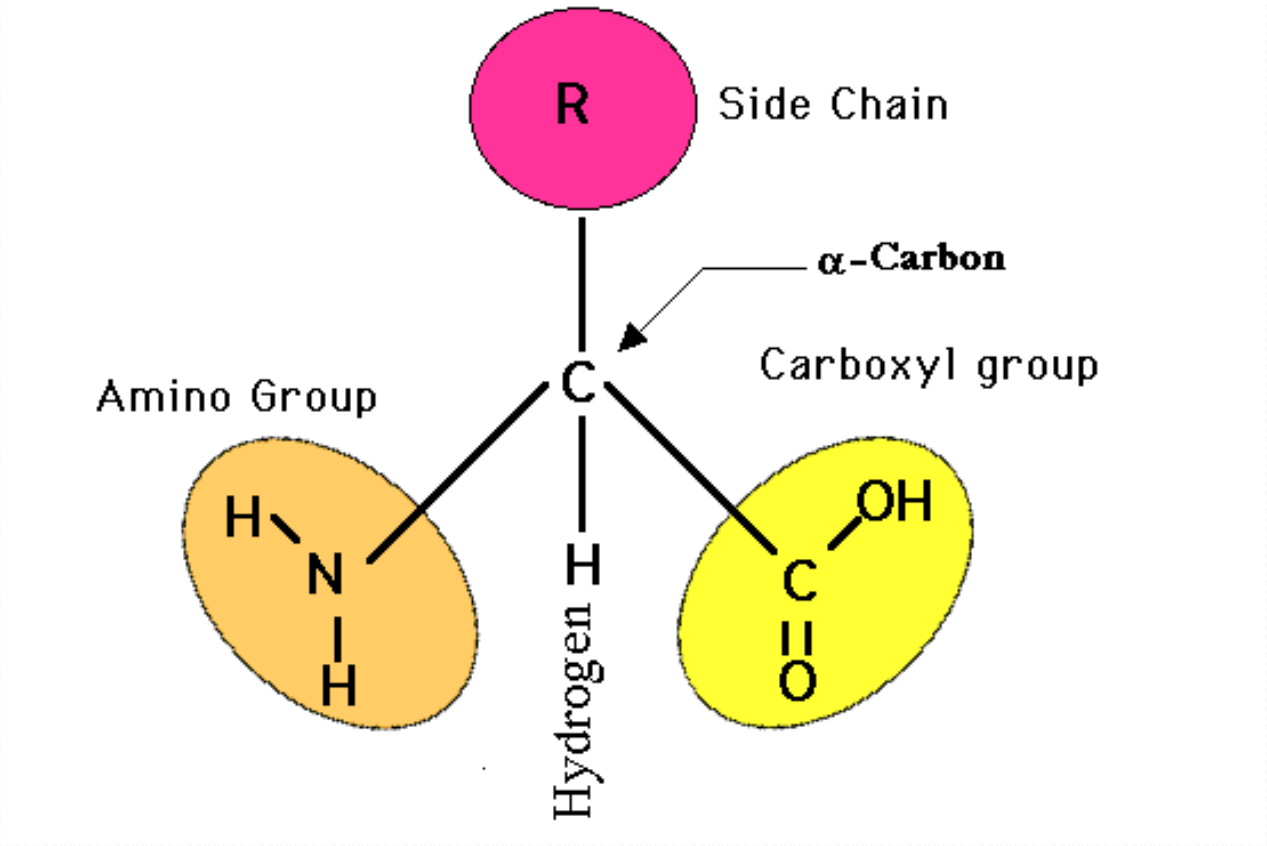
Assembly

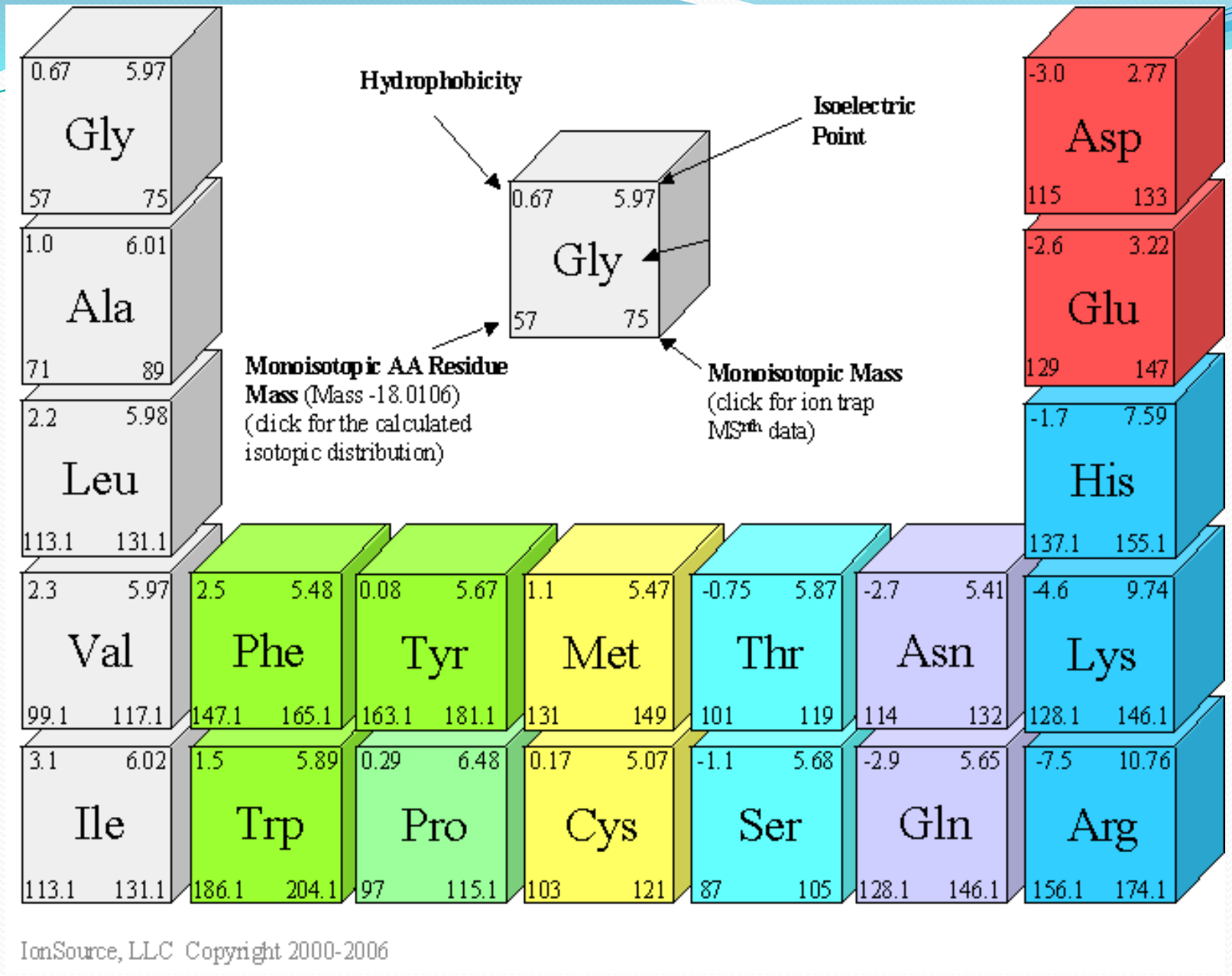
Folding

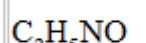
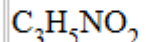
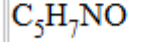
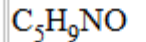
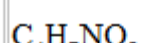
Packing

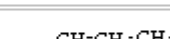
Interaction

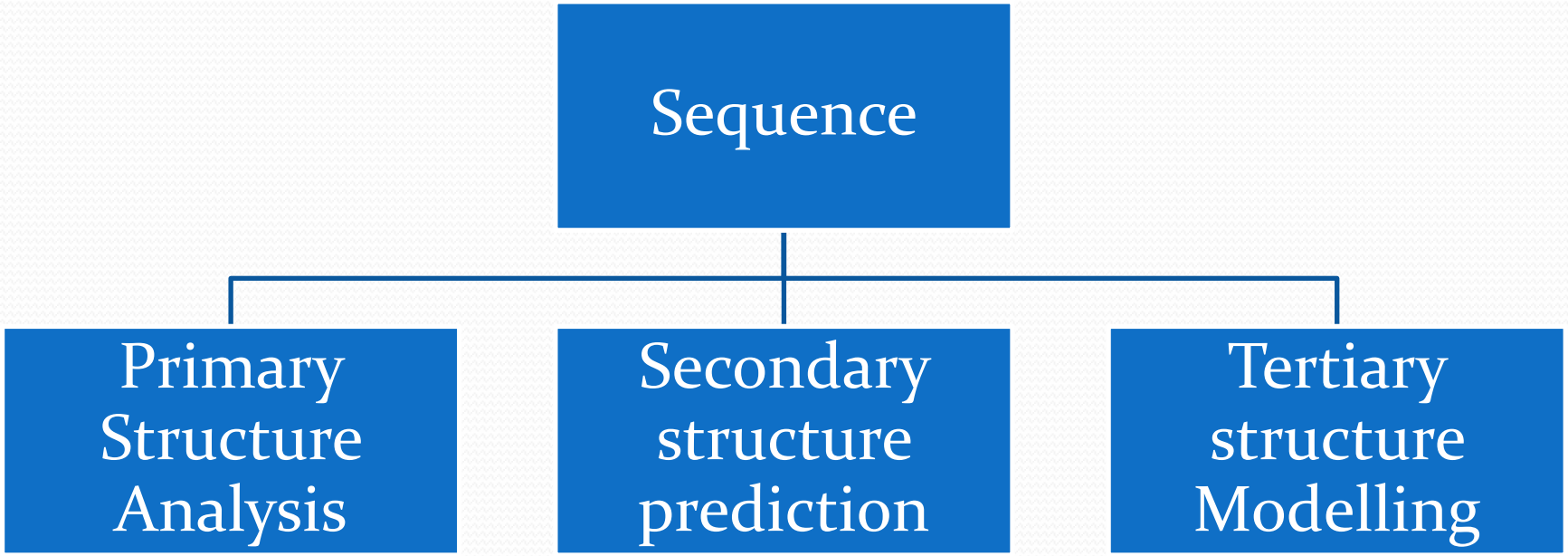
PROCESS


R
Side Chain
α-Carbon
Carboxyl group
Amino Group
C
H
N
H
H
Hydrogen
C
OH
O

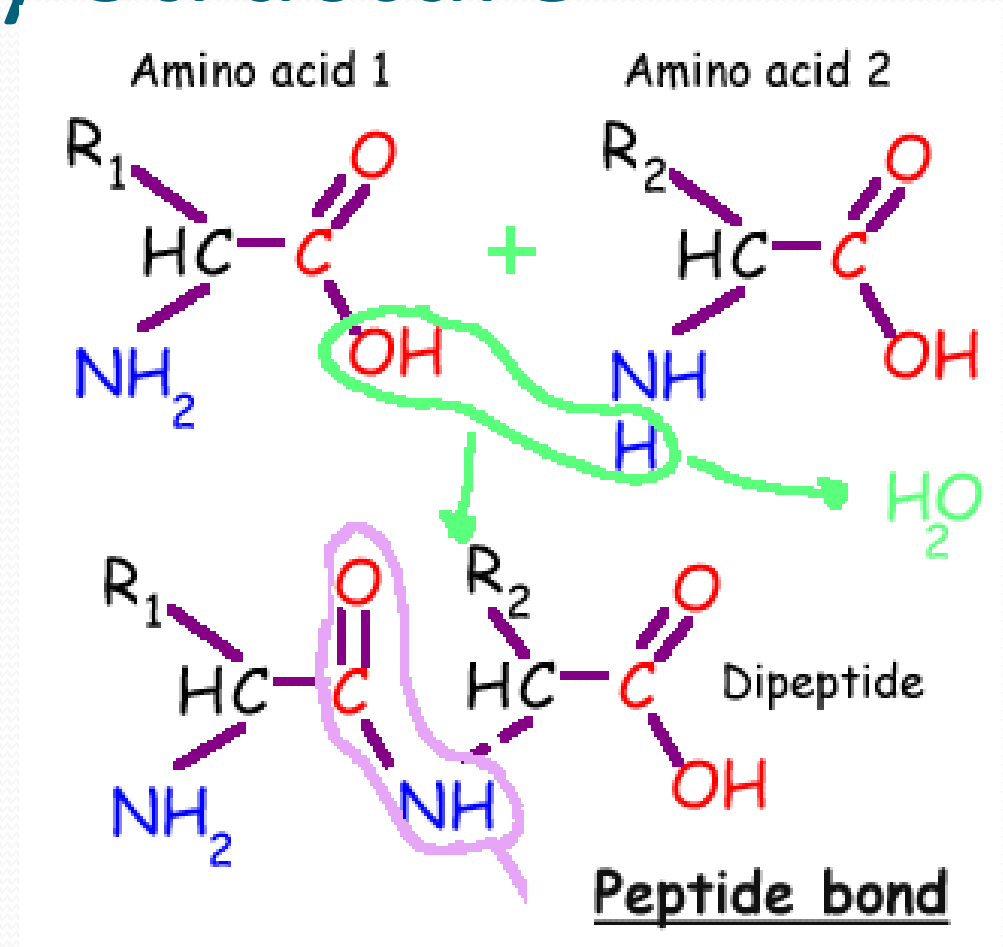

Hydrophobicity
Isoelectric Point
0.67 5.97
Gly
57 75
Monoisotopic AA Residue Mass (Mass -18.0106) (click for the calculated isotopic distribution)
Monoisotopic Mass (click for ion trap MS^nth data)
0.67 5.97 Gly 57 75
1.0 6.01 Ala 71 89
2.2 5.98 Leu 113.1 131.1
2.3 5.97 Val 99.1 117.1
2.5 5.48 Phe 147.1 165.1
0.08 5.67 Tyr 163.1 181.1
1.1 5.47 Met 131 149
-0.75 5.87 Thr 101 119
-2.7 5.41 Asn 114 132
-4.6 9.74 Lys 128.1 146.1
3.1 6.02 Ile 113.1 131.1
1.5 5.89 Trp 186.1 204.1
0.29 6.48 Pro 97 115.1
0.17 5.07 Cys 103 121
-1.1 5.68 Ser 87 105
-2.9 5.65 Gln 128.1 146.1
-7.5 10.76 Arg 156.1 174.1
-3.0 2.77 Asp 115 133
-2.6 3.22 Glu 129 147
-1.7 7.59 His 137.1 155.1
IonSource, LLC Copyright 2000-2006

AA Codes		AA Residue Composition	Mono.	Avg.	Structure
Gly	G	C_2H_3NO	57.021464	57.05	
Ala	A	C_3H_5NO	71.037114	71.08	
Ser	S	$C_3H_5NO_2$	87.032029	87.08	
Pro	P	C_5H_7NO	97.052764	97.12	
Val	V	C_5H_9NO	99.068414	99.07	
Thr	T	$C_4H_7NO_2$	101.04768	101.1	
Cys	C	C_3H_5NOS	103.00919	103.1	
Leu	L	$C_6H_{11}NO$	113.08406	113.2	
Ile	I	$C_6H_{11}NO$	113.08406	113.2	
Asn	N	$C_4H_6N_2O_2$	114.04293	114.1	

AA Codes		AA Residue Composition	Mono.	Avg.	Structure
Asp	D	$C_4H_5NO_3$	115.02694	115.1	
Gln	Q	$C_5H_8N_2O_2$	128.05858	128.1	
Lys	K	$C_6H_{12}N_2O$	128.09496	128.2	
Glu	E	$C_5H_7NO_3$	129.04259	129.1	
Met	M	C_5H_9NOS	131.04048	131.2	
His	H	$C_6H_7N_3O$	137.05891	137.1	
Phe	F	C_9H_9NO	147.06841	147.2	
Arg	R	$C_6H_{12}N_4O$	156.10111	156.2	
Tyr	Y	$C_9H_9NO_2$	163.06333	163.2	
Trp	W	$C_{11}H_{10}N_2O$	186.07931	186.2	


Sequence
Primary Structure Analysis
Secondary structure prediction
Tertiary structure Modelling

Primary Structure



Primary structure analysis

1
- Signal Peptide
- Hydrophobicity analysis

2
- Transmembrane domains
- Subcellular location

3
- Motif, prints, blocks
- Phylogeny

Protocol…

- Sequence
  - Databases: ncbi, Swissprot, Pir, etc
  - Experimental
- Signal peptide
- Molecular weight, isoelectric point
- Other Parameters
- potential cleavage sites

- Hydrophobicity
- TM
- TM2
- TM3
- TM4
- Subcellular location
- SL2

Protein substitution matrices

- Protein substitution matrices are significantly more complex than DNA scoring matrices.
- Proteins are composed of twenty amino acids, and physico-chemical properties of individual amino acids vary considerably.
- A protein substitution matrix can be based on any property of amino acids: size, polarity, charge, hydrophobicity.
- In practice the most important are **evolutionary substitution matrices**.

Matrices

- PAM (Point accepted mutation)
  - PAM120, PAM250 (Number of substitutions/100 residues)
- BLOSUM (Blocks substitution matrix)
  - Blosum 62, Blosum50 (identity)
- Proteins related
- BLOSUM newest

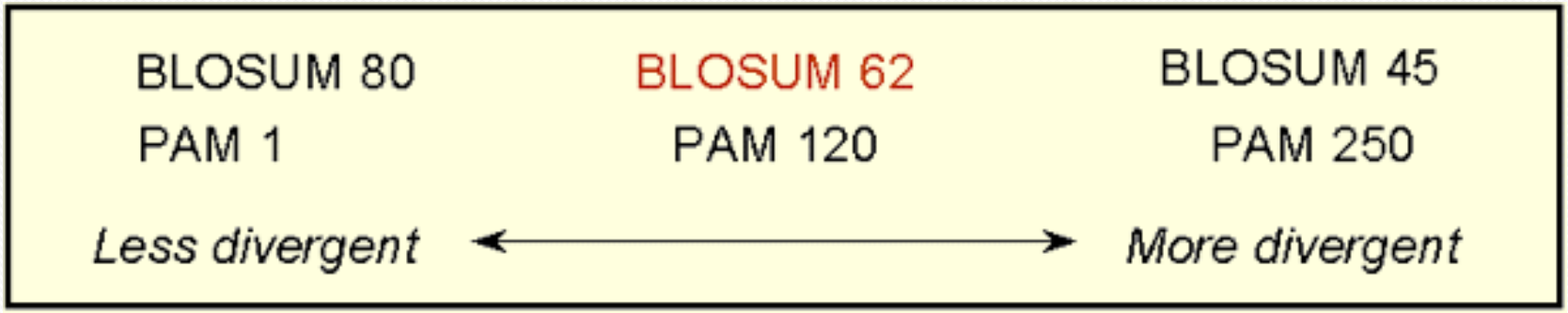
BLOSUM62 substitution matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

PAM 250

AMINOÁCIDO ORIGINAL

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A (Ala)	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R (Arg)	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N (Asn)	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D (Asp)	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C (Cys)	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q (Gln)	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E (Glu)	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G (Gly)	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H (His)	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I (Ile)	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L (Leu)	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K (Lys)	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M (Met)	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F (Phe)	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P (Pro)	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S (Ser)	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T (Thr)	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W (Trp)	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y (Tyr)	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V (Val)	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	2	4	17


BLOSUM 80
PAM 1
BLOSUM 62
PAM 120
BLOSUM 45
PAM 250
Less divergent
More divergent

Motifs??

Some family proteins conserve a short sequence and it´s related to function

Not 100% similarity

Phylogenetic analysis

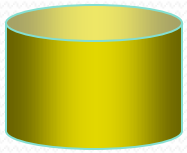
Conservation

Related to function

Conserved

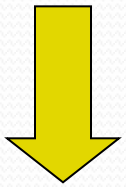
Prediction
Remote
Homologous

Motif databases



- Profiles
- Blocks
- Prints

ALRDF**ATHD**DF
SMTAE**ATHD**SI
ECDQA**ATHE**AS



A–T–H–[DE]

[AC]–x–V–x(4)–{E,D}.

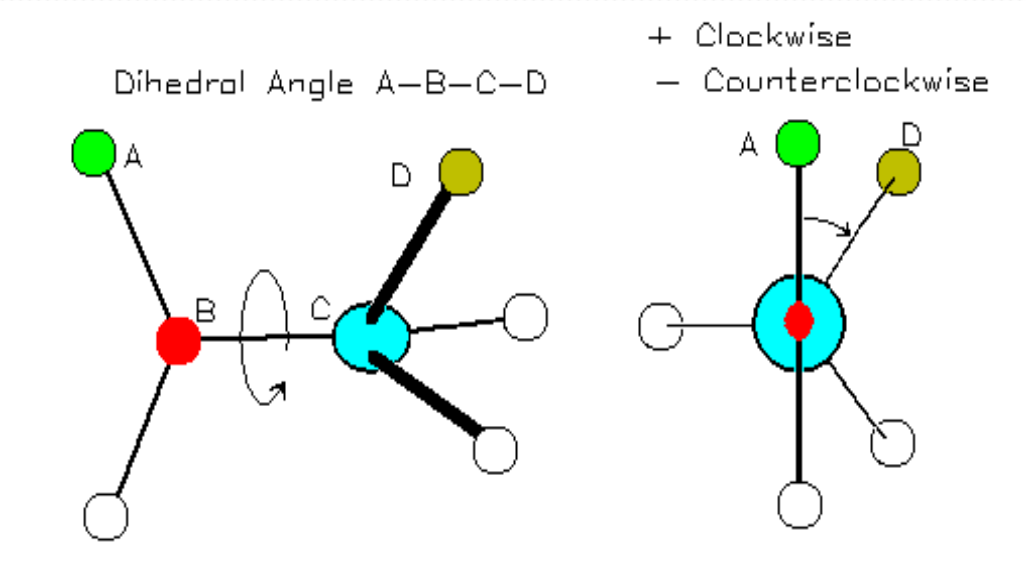
[Ala or Cys]–any–Val–any–any–any–any–{any but Glu or Asp}

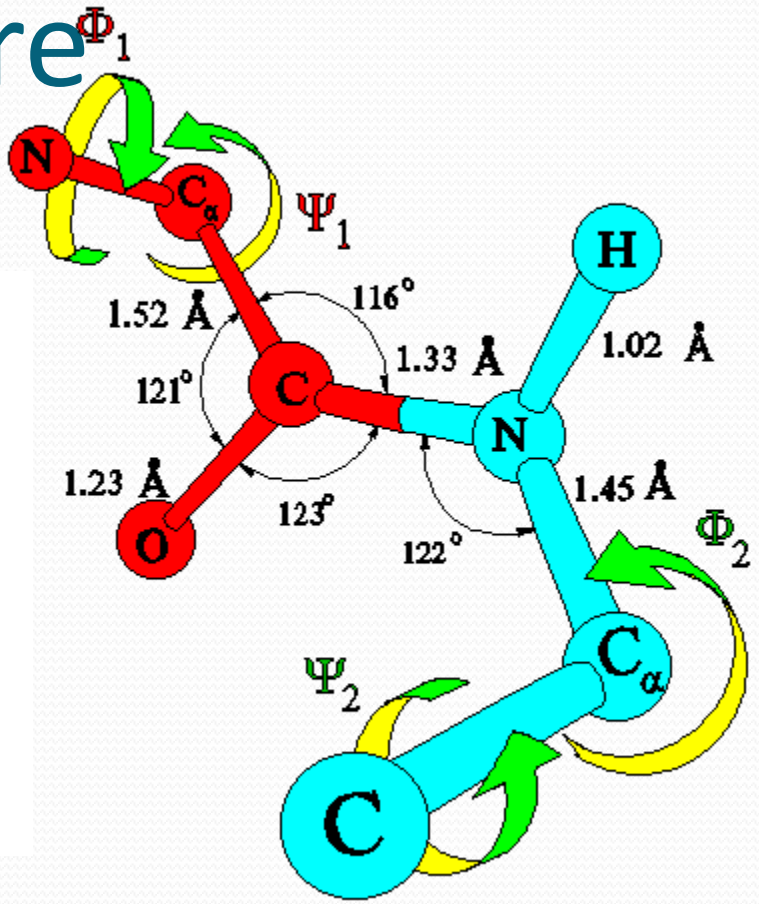
http://ca.expasy.org/tools/scanprosite/

http://myhits.isb-sib.ch/cgi-bin/motif_scan

http://smart.embl-heidelberg.de/index2.cgi

Secondary structure





Protein structure - bonding

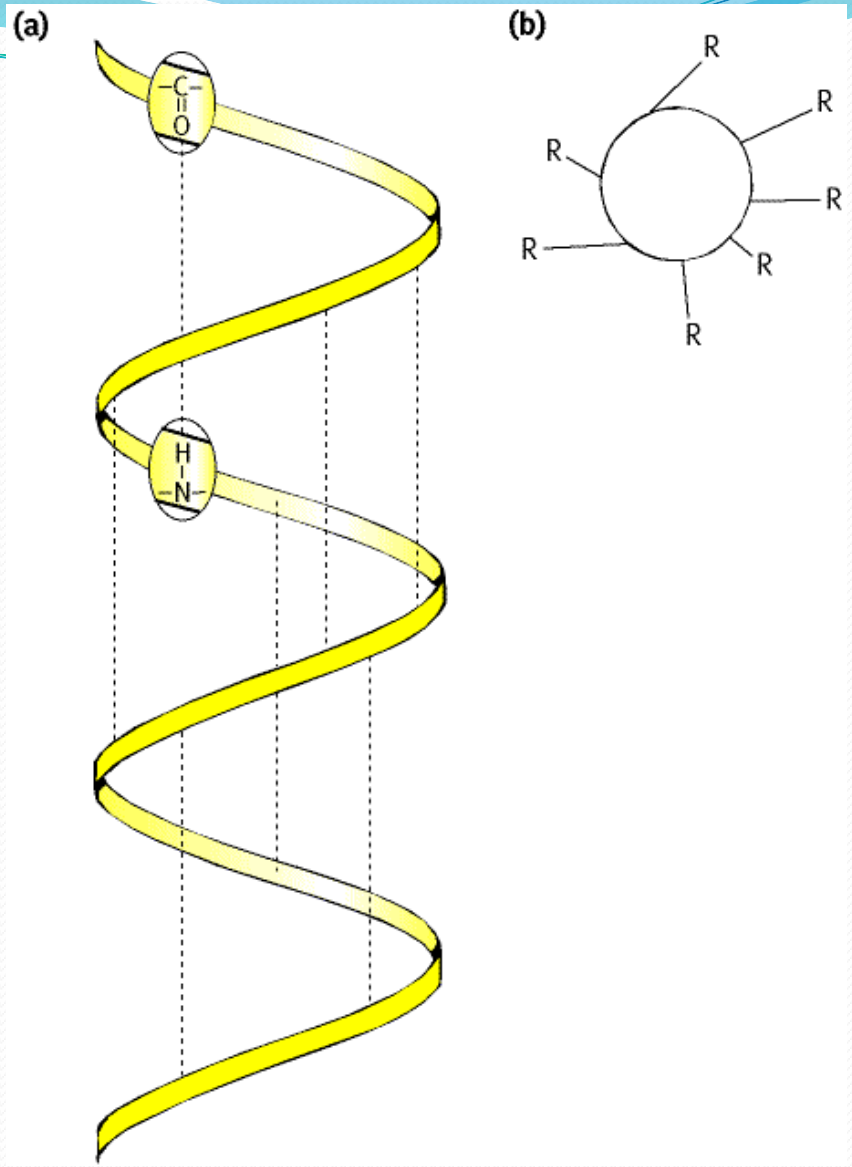
- 5 bonds or forces determine structure
  - Peptide bond
  - Hydrogen bond
  - Disulfide bond
  - Ionic bond
  - Hydrophobic force

Secondary protein structure

- Peptide chains fold into secondary structures:
  - α - helix
  - β - pleated sheet
  - Random coil

α - helix

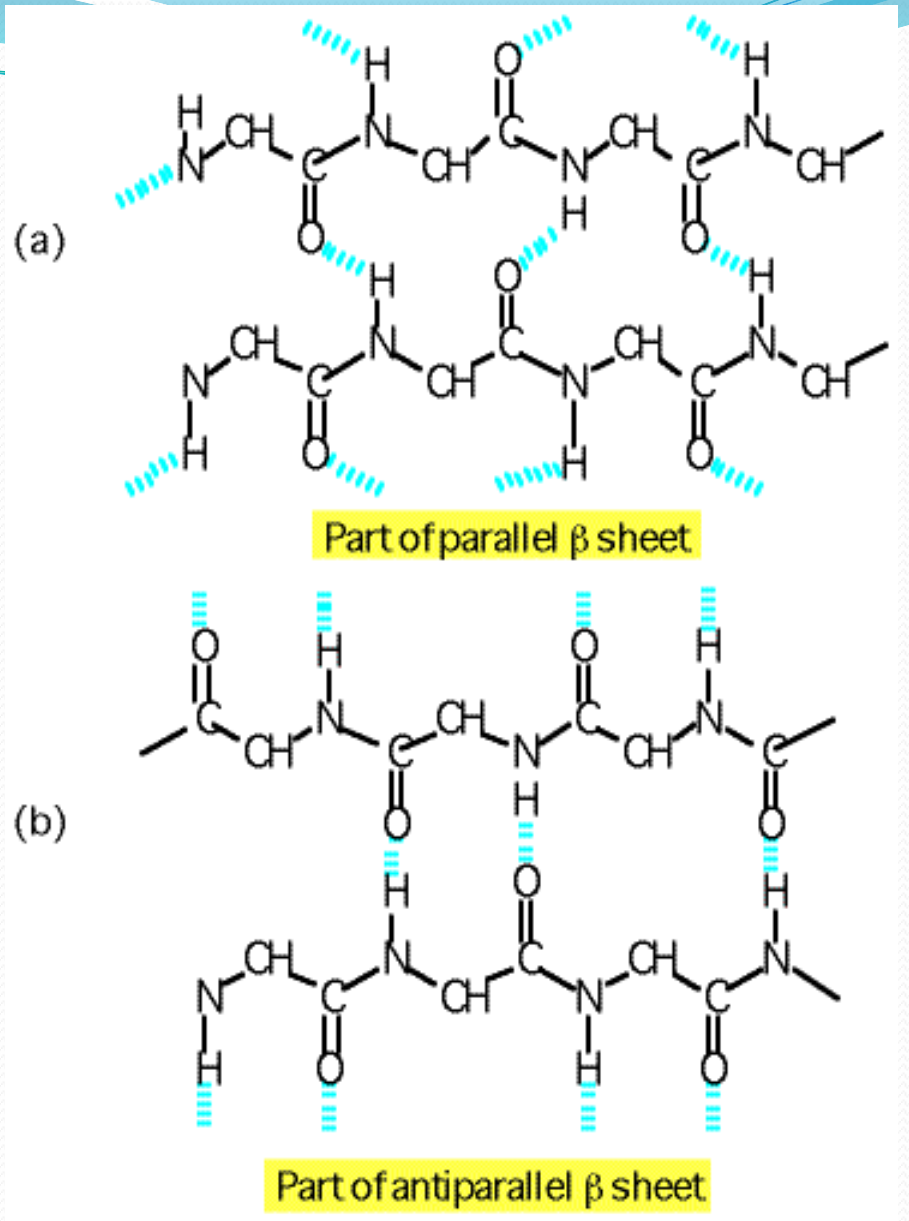
- Shape maintained by hydrogen bonds between C=O and N-H groups in backbone
- R groups directed outward from coil



From: Elliott, WH. Elliott, DC. (1997) Biochemistry and Molecular Biology. Oxford: Oxford University Press. p28

β - pleated sheet

- Structure maintained by hydrogen bonds between C=O and N-H groups in backbone
- R groups directed above and below backbone



From: Elliott, WH. Elliott, DC. (1997) Biochemistry and Molecular Biology. Oxford: Oxford University Press. p29

Random coil

From: Elliott, WH. Elliott, DC. (1997) Biochemistry and Molecular Biology. Oxford: Oxford University Press. p27

- Not really random structure, just non-repeating
  - 'Random' coil has fixed structure within a given protein
  - Commonly called 'connecting loop region'
  - Structure determined by bonding of side chains (i.e. not necessarily hydrogen bonds)

Secondary Structure Prediction

- *Given a protein sequence $a_1a_2...a_N$, secondary structure prediction aims at defining the state of each amino acid ai as being either H (helix), E (extended=strand), or O (other) (Some methods have 4 states: H, E, T for turns, and O for other).*
- *The quality of secondary structure prediction is measured with a "3-state accuracy" score, or Q_3. Q_3 is the percent of residues that match "reality" (X-ray structure).*

Quality of Secondary Structure Prediction

Determine Secondary Structure positions in known protein structures using DSSP or STRIDE:

1. Kabsch and Sander. Dictionary of Secondary Structure in Proteins: pattern recognition of hydrogen-bonded and geometrical features. Biopolymer 22: 2571-2637 (1983) (DSSP)
2. Frischman and Argos. Knowledge-based secondary structure assignments. Proteins, 23:566-571 (1995) (STRIDE)

Limitations of Q_3

ALHEASGPSVILFGSDVTVPPASNAEQAK Amino acid sequence

hhhhhooooeeeeoooeeeooooohhhhh Actual Secondary Structure

ohhhooooeeeeooooooeeeoooohhhhhh **Q3=22/29=76%**

(useful prediction)

hhhhhooooohhhhoooohhhooooohhhhh **Q3=22/29=76%**

(terrible prediction)

- Q3 for random prediction is 33%
- Secondary structure assignment in real proteins is uncertain to about 10%; Therefore, a “perfect” prediction would have Q3=90%.

Early methods for Secondary Structure Prediction

- *Chou and Fasman*

  (Chou and Fasman. Prediction of protein conformation. Biochemistry, 13: 211-245, 1974)

- *GOR*

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

Chou and Fasman

- *Start by computing amino acids propensities to belong to a given type of secondary structure:*

$$\frac{P(i/Helix)}{P(i)} \qquad \frac{P(i/Beta)}{P(i)} \qquad \frac{P(i/Turn)}{P(i)}$$

Propensities > 1 mean that the residue type I is likely to be found in the Corresponding secondary structure type.

Chou and Fasman

Amino Acid	α-Helix	β-Sheet	Turn
Ala	1.29	0.90	0.78
Cys	1.11	0.74	0.80
Leu	1.30	1.02	0.59
Met	1.47	0.97	0.39
Glu	1.44	0.75	1.00
Gln	1.27	0.80	0.97
His	1.22	1.08	0.69
Lys	1.23	0.77	0.96
Val	0.91	1.49	0.47
Ile	0.97	1.45	0.51
Phe	1.07	1.32	0.58
Tyr	0.72	1.25	1.05
Trp	0.99	1.14	0.75
Thr	0.82	1.21	1.03
Gly	0.56	0.92	1.64
Ser	0.82	0.95	1.33
Asp	1.04	0.72	1.41
Asn	0.90	0.76	1.23
Pro	0.52	0.64	1.91
Arg	0.96	0.99	0.88

Favors α-Helix

Favors β-strand

Favors turn

Chou and Fasman

Predicting helices:

- find nucleation site: 4 out of 6 contiguous residues with P(α)>1
- extension: extend helix in both directions until a set of 4 contiguous residues has an average P(α) < 1 (breaker)
- if average P(α) over whole region is >1, it is predicted to be helical

Predicting strands:

- find nucleation site: 3 out of 5 contiguous residues with P(β)>1
- extension: extend strand in both directions until a set of 4 contiguous residues has an average P(β) < 1 (breaker)
- if average P(β) over whole region is >1, it is predicted to be a strand

Chou and Fasman

Position-specific parameters for turn:

Each position has distinct amino acid preferences.

Examples:

-At position 2, Pro is highly preferred; Trp is disfavored

-At position 3, Asp, Asn and Gly are preferred

-At position 4, Trp, Gly and Cys preferred

	f(i)	f(i+1)	f(i+2)	f(i+3)
Ala	0.060	0.076	0.035	0.058
Arg	0.070	0.106	0.099	0.085
Asp	0.147	0.110	0.179	0.081
Asn	0.161	0.083	0.191	0.091
Cys	0.149	0.050	0.117	0.128
Glu	0.056	0.060	0.077	0.064
Gln	0.074	0.098	0.037	0.098
Gly	0.102	0.085	0.190	0.152
His	0.140	0.047	0.093	0.054
Ile	0.043	0.034	0.013	0.056
Leu	0.061	0.025	0.036	0.070
Lys	0.055	0.115	0.072	0.095
Met	0.068	0.082	0.014	0.055
Phe	0.059	0.041	0.065	0.065
Pro	0.102	0.301	0.034	0.068
Ser	0.120	0.139	0.125	0.106
Thr	0.086	0.108	0.065	0.079
Trp	0.077	0.013	0.064	0.167
Tyr	0.082	0.065	0.114	0.125
Val	0.062	0.048	0.028	0.053

Chou and Fasman

Predicting turns:

- for each tetrapeptide starting at residue i, compute:
  - P_{Turn} (average propensity over all 4 residues)
  - F = f(i)*f(i+1)*f(i+2)*f(i+3)

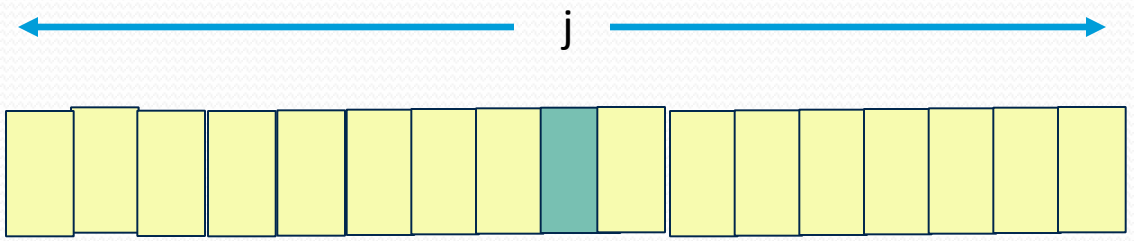
- if $P_{Turn} > P\alpha$ and $P_{Turn} > P\beta$ and $P_{Turn} > 1$ and F>0.000075 tetrapeptide is considered a turn.

Chou and Fasman prediction:

http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

The GOR method

Position-dependent propensities for helix, sheet or turn is calculated for each amino acid. For each position j in the sequence, eight residues on either side are considered.



A helix propensity table contains information about propensity for residues at 17 positions when the conformation of residue j is helical. The helix propensity tables have 20 x 17 entries.
Build similar tables for strands and turns.

GOR simplification:
The predicted state of AAj is calculated as the sum of the position-dependent propensities of all residues around AAj.

GOR can be used at : http://abs.cit.nih.gov/gor/ (current version is GOR IV)

Accuracy

- Both Chou and Fasman and GOR have been assessed and their accuracy is estimated to be Q3=60-65%.

(initially, higher scores were reported, but the experiments set to measure Q3 were flawed, as the test cases included proteins used to derive the propensities!)

Neural networks

The most successful methods for predicting secondary structure are based on neural networks. The overall idea is that neural networks can be trained to recognize amino acid patterns in known secondary structure units, and to use these patterns to distinguish between the different types of secondary structure.

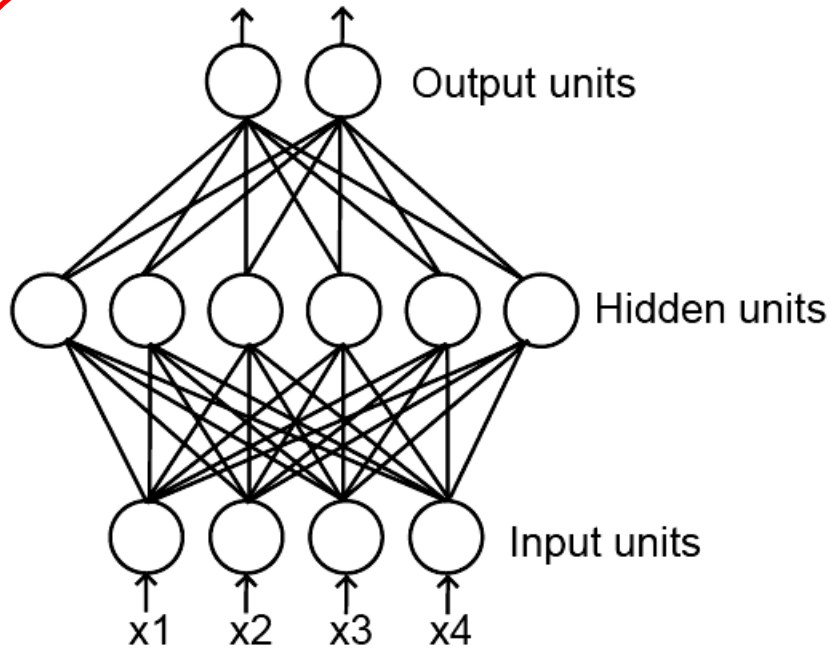
Neural networks classify "input vectors" or "examples" into categories (2 or more).
They are loosely based on biological neurons.

Neural networks

The most successful methods for predicting secondary structure are based on neural networks. The overall idea is that neural networks can be trained to recognize amino acid patterns in known secondary structure units, and to use these patterns to distinguish between the different types of secondary structure.

Neural networks classify "input vectors" or "examples" into categories (2 or more).
They are loosely based on biological neurons.

Neural Network



A complete neural network is a set of perceptrons interconnected such that the outputs of some units becomes the inputs of other units. Many topologies are possible!

Neural networks are trained just like perceptron, by minimizing an error function:

$$E = \sum_{i=1}^{Ndata} \left(NN(X^i) - t(X^i) \right)^2$$

Neural networks and Secondary Structure prediction

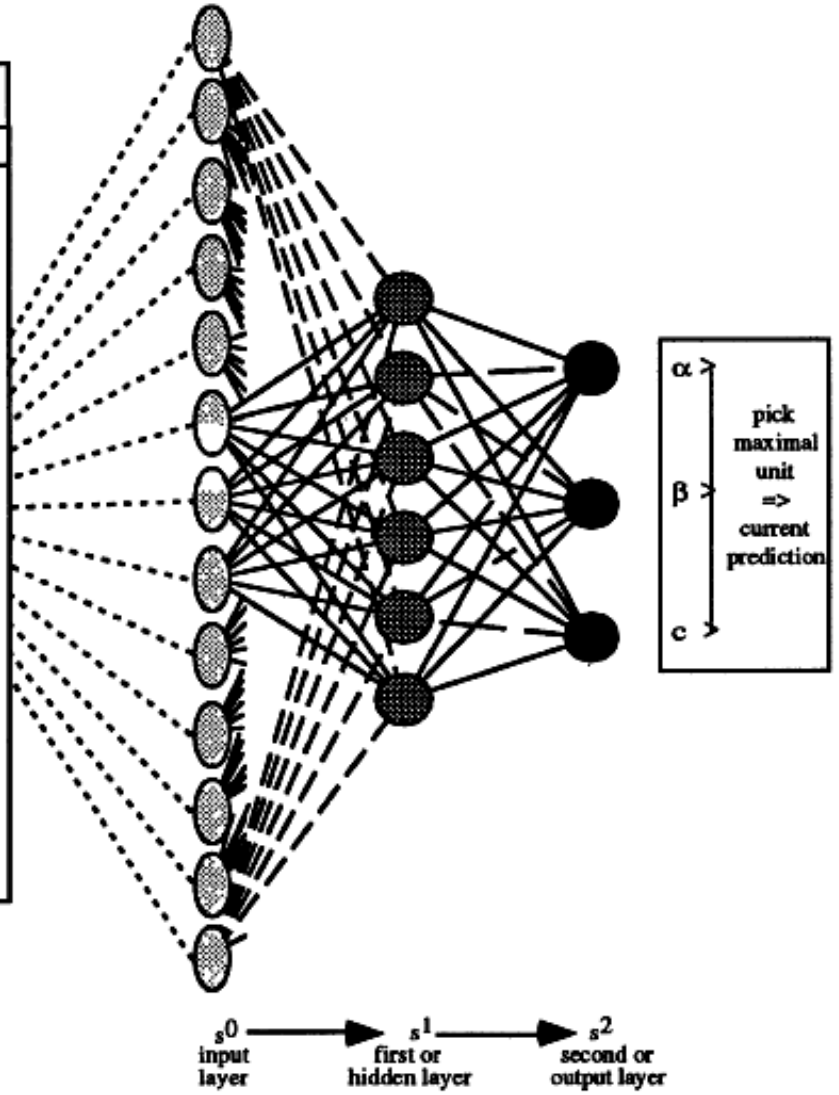
Experience from Chou and Fasman and GOR has shown that:

- In predicting the conformation of a residue, it is important to consider a window around it.
- Helices and strands occur in stretches
- It is important to consider multiple sequences

PHD: Secondary structure prediction using NN

Biophysics: Rost and Sander — *Proc. Natl. Acad. Sci. USA 90 (1993)* — 7559

Protein	Alignments	profile table
		GSA PD NT EKQ C VH IR LM YF W
:	: : : :	
G	G G G G	5. . . . .
Y	Y Y Y Y	. . . . . 5 . .
I	I I E E	. . . . . 2 . . . . 3 . . . . .
Y	Y Y Y Y	. . . . . 5 . .
D	D D D D	. . . . 5 . . . . .
P	P P P P	. . . 5 . . . . .
E	A E A A	. . 3 . . . . 2 . . . . .
D	V V E E	. . . . 1 . . 2 . . . 2 . . . . .
G	G G G G	5 . . . . .
D	D D D D	. . . . 5 . . . . .
P	P P P P	. . . 5 . . . . .
D	D T D D	. . . . 4 . 1 . . . . .
D	N Q N N	. . . . 1 3 . . . 1 . . . . .
G	G N G G	4 . . . . 1 . . . . .
V	V I V V	. . . . . 4 . 1 . . . . .
N	E P K K	. . . 1 . 1 . 1 2 . . . . .
P	P P P P	. . . 5 . . . . .
G	G G G G	5 . . . . .
T	T T T T	. . . . . 5 . . . . .
D	E K S A	. 1 1 . 1 . . 1 1 . . . . .
F	F F F F	. . . . . 5 .
:	: : : :	



Secondary Structure Prediction

-Available servers:

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

- PHD: http://cubic.bioc.columbia.edu/predictprotein/

- PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred/

- NNPREDICT: http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html

- Chou and Fassman: http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

-Interesting paper:

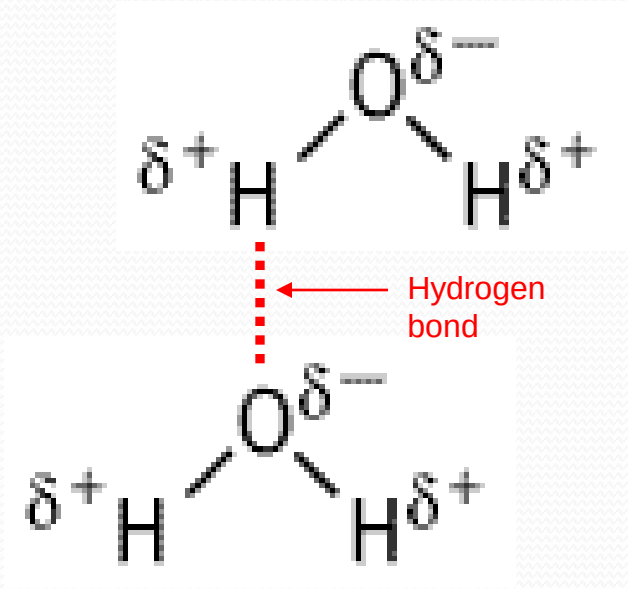
- Rost and Eyrich. EVA: Large-scale analysis of secondary structure prediction. Proteins 5:192-199 (2001)

Performances (monitored at CASP)

CASP	YEAR	# of Targets	<Q3>	Group
CASP1	1994	6	63	Rost and Sander
CASP2	1996	24	70	Rost
CASP3	1998	18	75	Jones
CASP4	2000	28	80	Jones

Tertiary protein structure

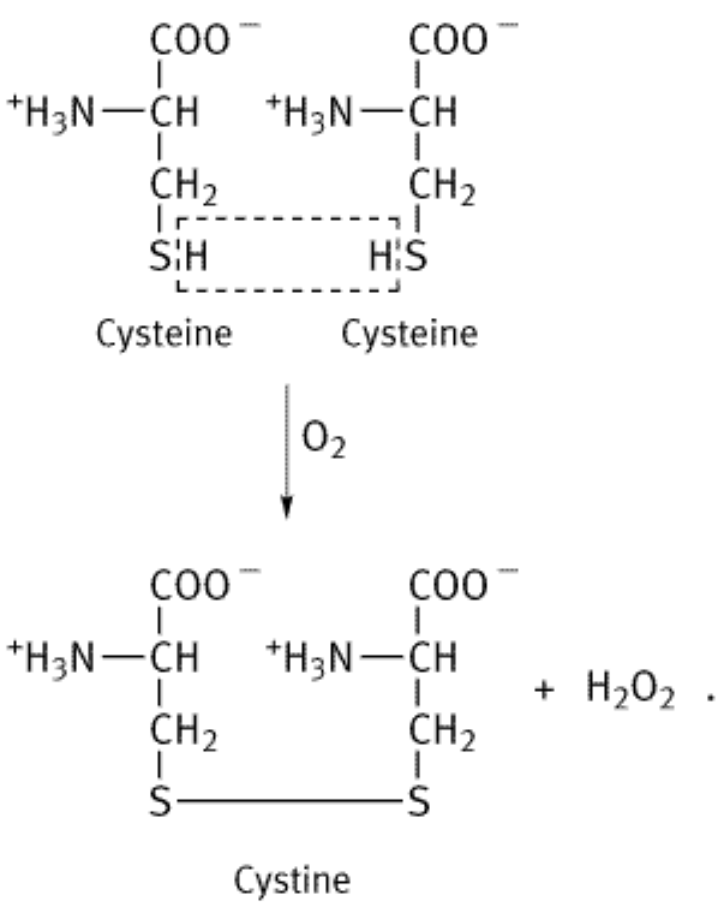
- Secondary structures fold and pack together to form tertiary structure
  - Usually globular shape
- Tertiary structure stabilised by bonds between R groups (i.e. sidechains)



- H bonds weak allowing to be broken and reformed easily
  - Allows structural change
    - produces 'functional' molecules

Tertiary structure - H bond

- Covalent bond between sulfur atoms on two cysteine amino acids

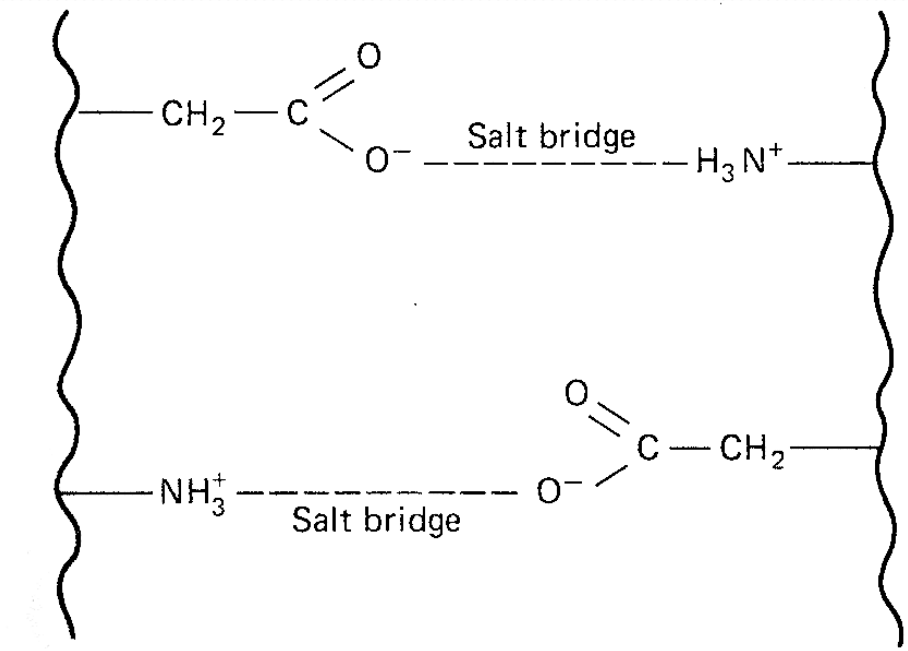


Tertiary structure - disulfide bond

From: Elliott, WH. Elliott, DC. (1997) Biochemistry and Molecular Biology. Oxford: Oxford University Press. p32

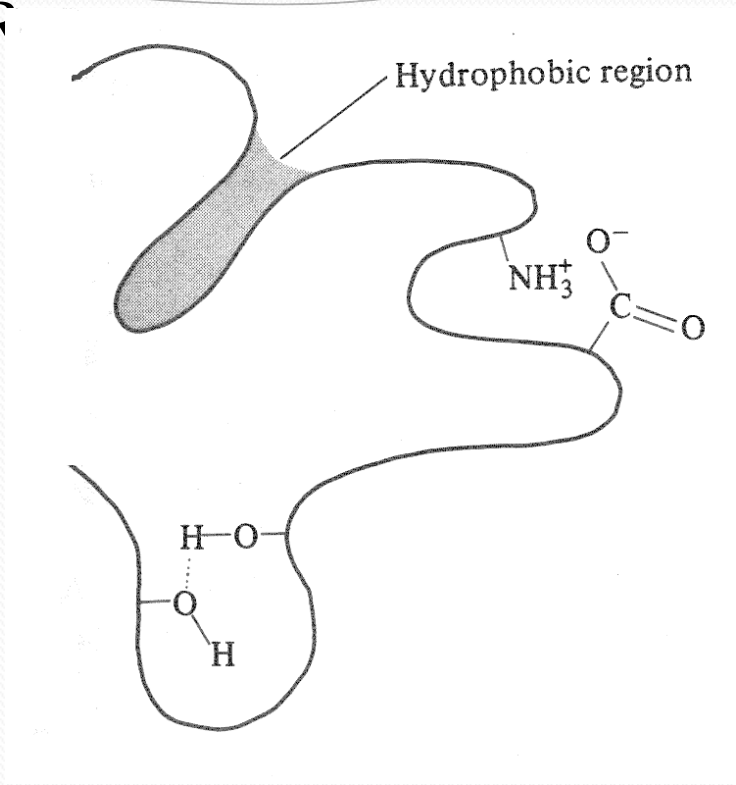
Tertiary structure - ionic bond

- Ions on R groups form salt bridges through ionic bonds



From: Summerlin, LR. (1981) Chemistry for the Life Sciences. New York: Random House, p459

- Close attraction of non-polar R groups through dispersion forces
- Very weak but collective interactions over large area stabilise structure
- Repel polar and charged molecules/particles



Tertiary structure - hydrophobic forces

Bettelheim & March (1990) Introduction to Organic & Biochemistry (International Edition) Philadelphia: Saunders College Publishing, p302

Methods for structure prediction

- Analysis of Primary Structure
- Secondary structure prediction
- Homology modelling
  - Building a 3D model on the basis of similar sequences
- Threading
  - Threading the sequence on all known protein structures, and testing the consistency
- *ab initio* prediction of tertiary structure
  - For proteins of normal size, it is almost impossible to predict structures *ab initio*.
  - Some results have been obtained in the prediction of oligopeptide structures.

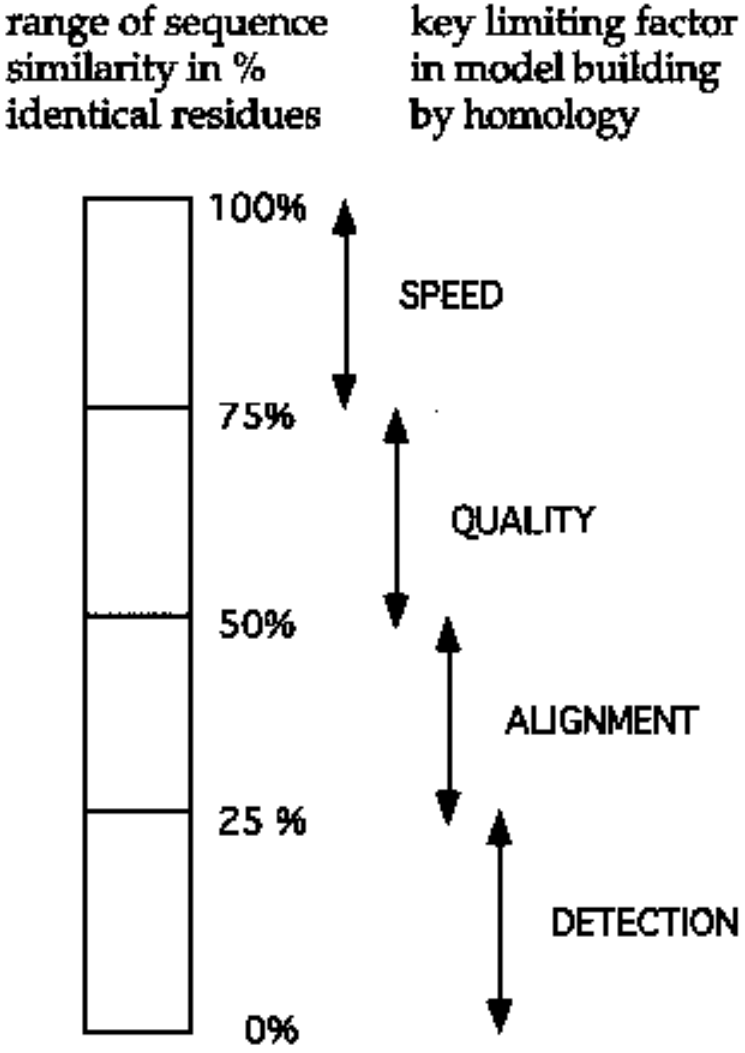
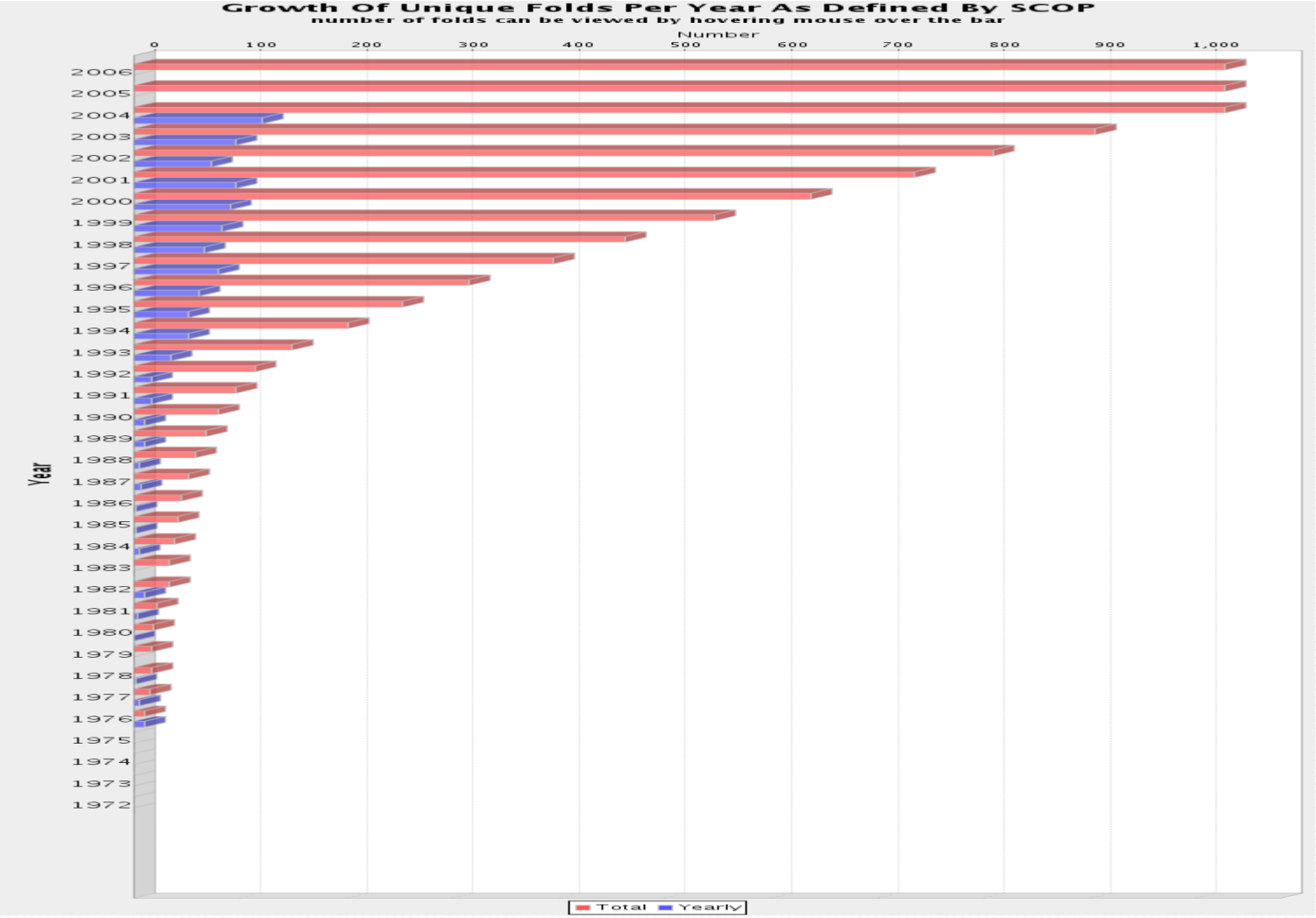


Figure 1. The main limiting steps for model building by homology as function of the percentage sequence identity between the structure and the model.

Protein folds



Basic concept

- In a given protein 3D structure is a more conserved characteristic than sequence
  - Some aminoacids are “equivalent” to each other
  - Evolutionary pressure allows only aminoacids substitutions that keep 3D structure largely unaltered
- Two proteins of “similar” sequences must have the “same” 3D structure

Possible scenarios

1. Homology can be recognized using sequence comparison tools or protein family databases (blast, clustal, pfam,…).

   Structural and functional predictions are feasible

2. Homology exist but cannot be recognized easily (psi-blast, threading)

   Low resolution fold predictions are possible. No functional information.

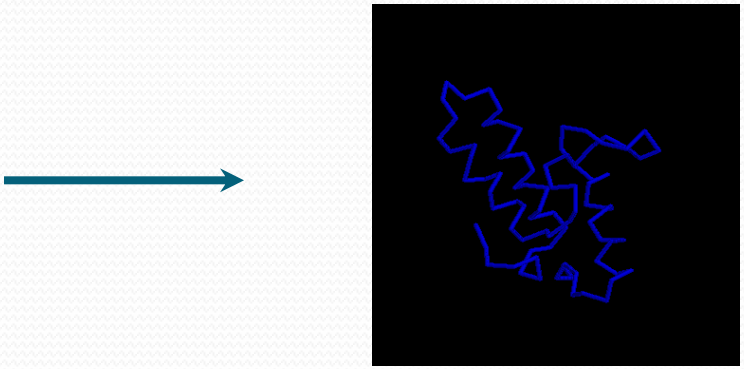
3. No homology

   1D predictions. Sequence motifs. Limited functional prediction. Ab-initio prediction

ab-initio prediction

- Prediction from sequence using first principles

AVVTW...GTTWVR



Ab-initio prediction

- "In theory", we should be able to build native structures from first principles using sequence information and molecular dynamics simulations: "Ab-initio prediction of structure"

  - Simulaciones de 1 μs de "folding" de una proteína modelo (Duan-Kollman: Science, 277, 1793, 1998).
  - Simulaciones de folding reversible de péptidos (20-200 ns) (Daura et al., Angew. Chem., 38, 236, 1999).
  - Simulaciones distribuidas de folding de Villin (36-residues) (Zagrovic et al., JMB, 323, 927, 2002).

… the bad news …

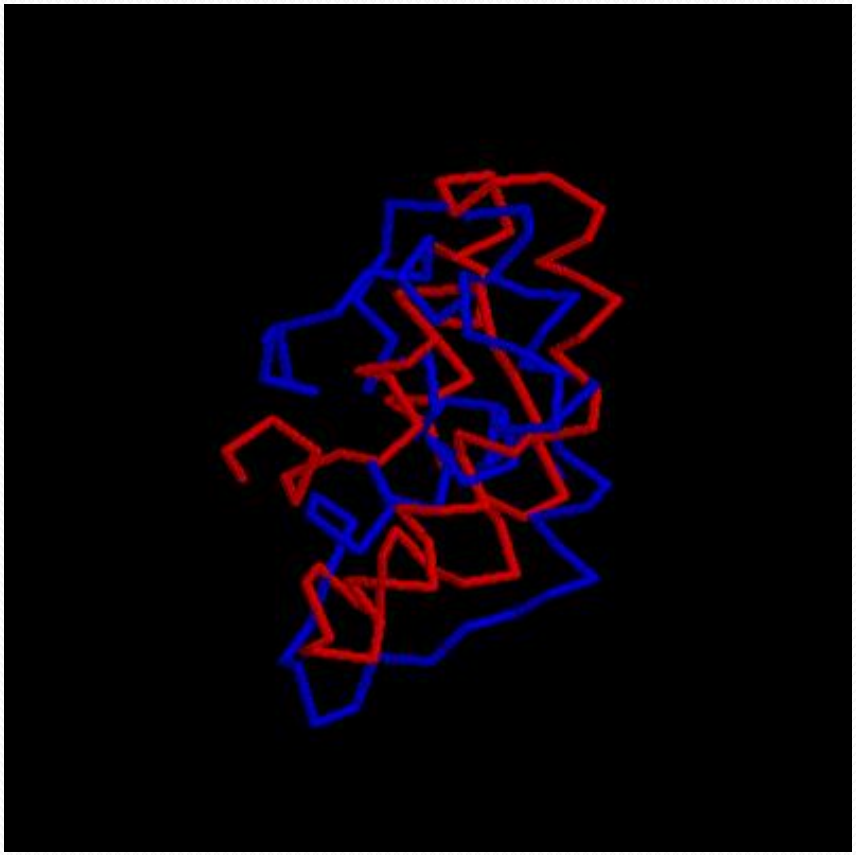
- It is not possible to span simulations to the “seconds” range
- Simulations are limited to small systems and fast folding/unfolding events in known structures
  - steered dynamics
  - biased molecular dynamics
- Simplified systems

typical shortcuts

- Reduce conformational space
  - 1,2 atoms per residue
  - fixed lattices
- Statistic force-fields obtained from known structures
  - Average distances between residues
  - Interactions
- Use building blocks: 3-9 residues from PDB structures

Results from ab-initio

- Average error 5 Å - 10 Å
- Function cannot be predicted
- Long simulations



Some protein from E.*coli*
predicted at 7.6 Å
(CASP3, H.Scheraga)

comparative modelling

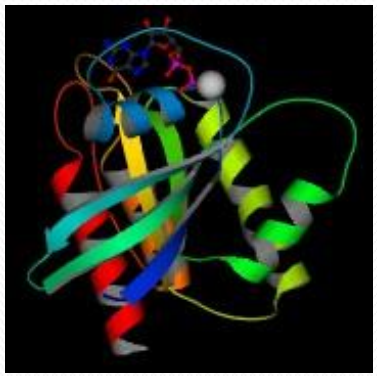
- The most efficient way to predict protein structure is to compare with known 3D structures

Threading

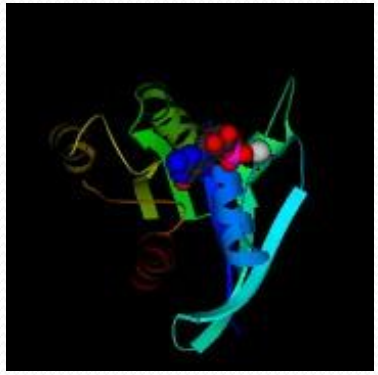
- Unknown sequence is “folded” in a number of known structures

- Scoring functions evaluate the fitting between sequence and structure according to statistical functions and sequence comparison

ATTWV....PRKSCT



.....

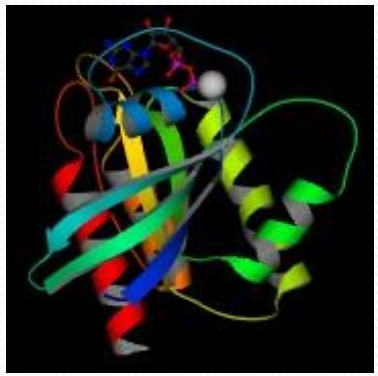


10.5 > 5.2

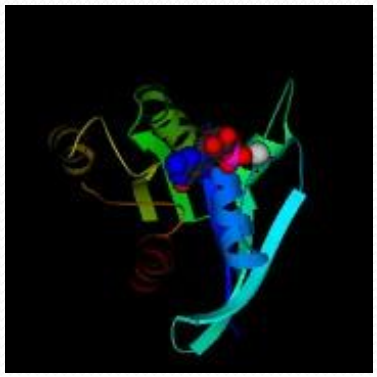
SELECTED HIT

ATTWV....PRKSCT
HHHHH....CCBBBB
eeebb....eeebeb

Sequence
Pred. Sec. Struc.
Pred. accesibility

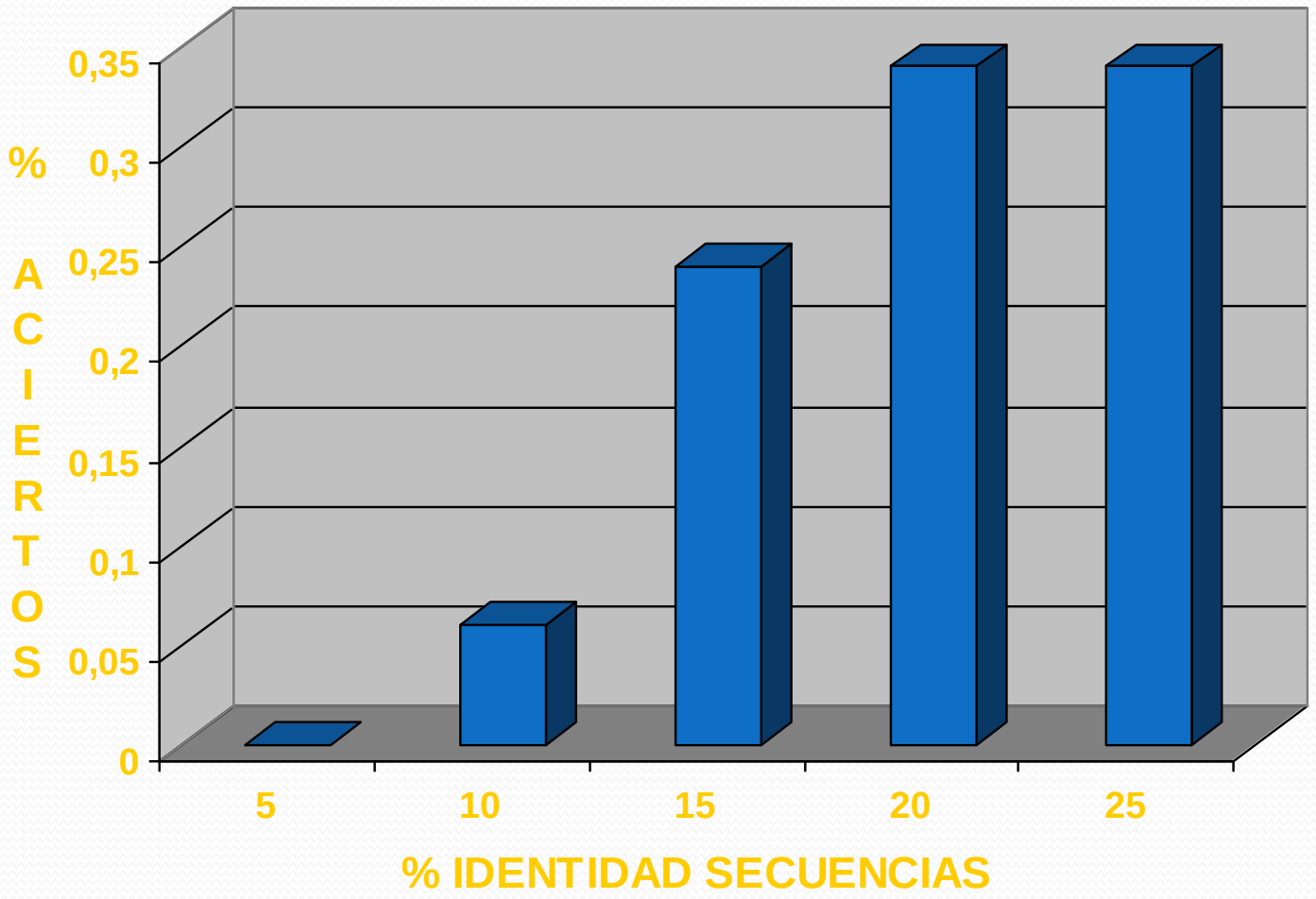


.....



Sequence	GGTV....ATTW	.....	ATTVL....FFRK
Obs SS	BBBB....CCHH	.....	HHHB.....CBCB
Obs Acc.	EEBE.....BBEB	.....	BBEBB....EBBE

Threading accurancy



http://toolkit.tuebingen.mpg.de/hhpred

Comparative modelling

- Good for homology >30%
- Accurancy is very high for homology > 60%
- Reminder
  - The model must be USEFUL
  - Only the “interesting” regions of the protein need to be modelled
- http://swissmodel.expasy.org/

Expected accurancy

- Strongly dependent on the quality of the sequence alignment
- Strongly dependent on the identity with “template” structures. Very good structures if identity > 60-70%.
- Quality of the model is better in the backbone than side chains
- Quality of the model is better in conserved regions

Quality test

- No energy differences between a correct or wrong model

- The structure must by “chemically correct” to use it in quantitative predictions

http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html