Predicting structural and functional features starting at residue level

Laboratory of Bioinformatics I
Module 2

Emidio Capriotti
http://biofold.org/

Department of Pharmacy and
Biotechnology (FaBiT)
University of Bologna

Biomolecules
Folding and
Disease

From Sequence to Structure

```
>TargetSequence
MNPNQKIITIGSVCMTIGMANLILQIGNIISIWISHSIQLGNQNQIETCNQSV
ITYENNTWVNQTYVNISNTNFAAGQSVVSVKLAGNSSLCPVSGWAIYSKDNSV
RIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLM
SCPIGEVPSPYNSRFESVAWSASACHDGINWLTIGISGPDNGAVAVLKYNGII
TDTIKSWRNNILRTQESECACVNGSCFTVMTDGPSNGQASYKIFRIEKGKIVK
SVEMNAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQIGYIC
SGIFGDNPRPNDKTGSCGPVSSNGANGVKGFSFKYGNGVWIGRTKSISSRNGF
EMIWDPNGWTGTDNNFSIKQDIVGINEWSGYSGSFVQHPELTGLDCIRPCFWV
ELIRGRPKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTID
```

Computational Approach

Tertiary Predictions:

1. Comparative/Homology Modeling
2. Fold Recognition
3. De Novo Protein Structure Prediction

Template search

→ Comparative/Homology modelling requires:

1) the availability of a template
2) high sequence identity between target and template

→ Multiple sequence alignment and HMM are able to extend the applicability domain of comparative modelling (remote homology)

→ Example from the practicum: starting from the seed you adopted for modelling the Kunitz domain, how many similar domain can you recognize in SwissProt with simple sequence search? How many with your (or the PFAM) HMM?

A step further

→ What if similarity methods (simple or profile-based) fail (i.e. no suitable template can be detected in the PDB) ?

→ What are the possible scenarios?

1) Suitable templates DO NOT EXIST in the PDB

→ Ab Initio Methods are required

2) There are possible templates in the PDB, but they CANNOT BE RECOGNIZED.

→ Fold recognition/Threading methods can be adopted

Ab Initio predictions

Difficult because search space is huge. Much larger conformational space

Goal: Predict Structure only given its amino acid sequence
In theory: Lowest Energy Conformation

Difficult for sequences larger that 150aa

Rosetta (David Baker lab) one of best (CASP evaluation)

MD Force Field

$$U(\vec{R}) = \underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{U_{angle}} + \underbrace{\sum_{dihedrals} k_i^{dihe}[1 + \cos(n_i\phi_i + \delta_i)]}_{U_{dihedral}} + \underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{nonbond}}$$

$$\boldsymbol{r}(t + \delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\delta t$$

$$\boldsymbol{v}(t + \delta t) = \boldsymbol{v}(t) + \boldsymbol{a}(t)\delta t$$

$$\boldsymbol{a}(t) = \boldsymbol{F}(t)/m$$

$$\boldsymbol{F} = -\frac{d}{d\boldsymbol{r}}U(\boldsymbol{r})$$

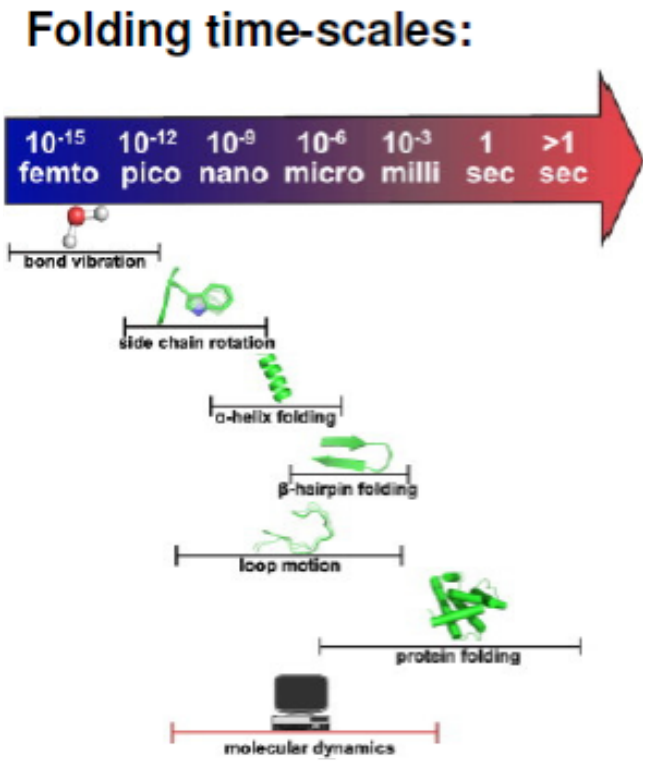
U_{bond} = oscillations about the equilibrium bond length
U_{angle} = oscillations of 3 atoms about an equilibrium bond angle
$U_{dihedral}$ = torsional rotation of 4 atoms about a central bond
$U_{nonbond}$ = non-bonded energy terms (electrostatics and Lenard-Jones)

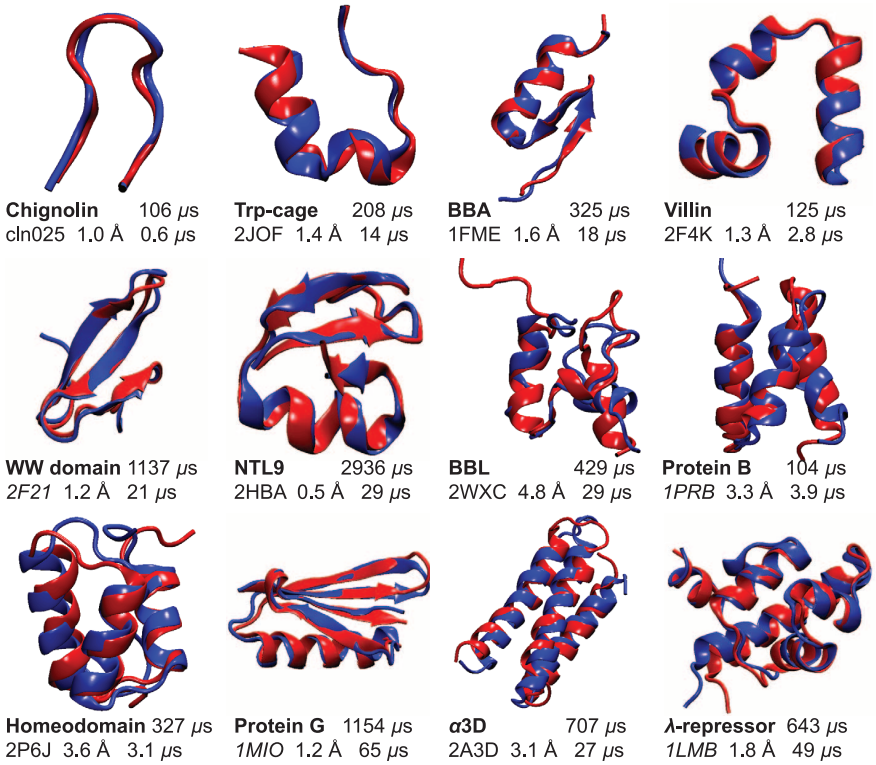
One of the most popular forcefield is CHARMM
(Chemistry at HARvard Macromolecular Mechanics)

https://www.charmmtutorial.org

MD Limitations

- Requires powerful hardware or computing time
- Limited to small simple proteins
- Can not take in to account chaperone activity
- Criteria for success??

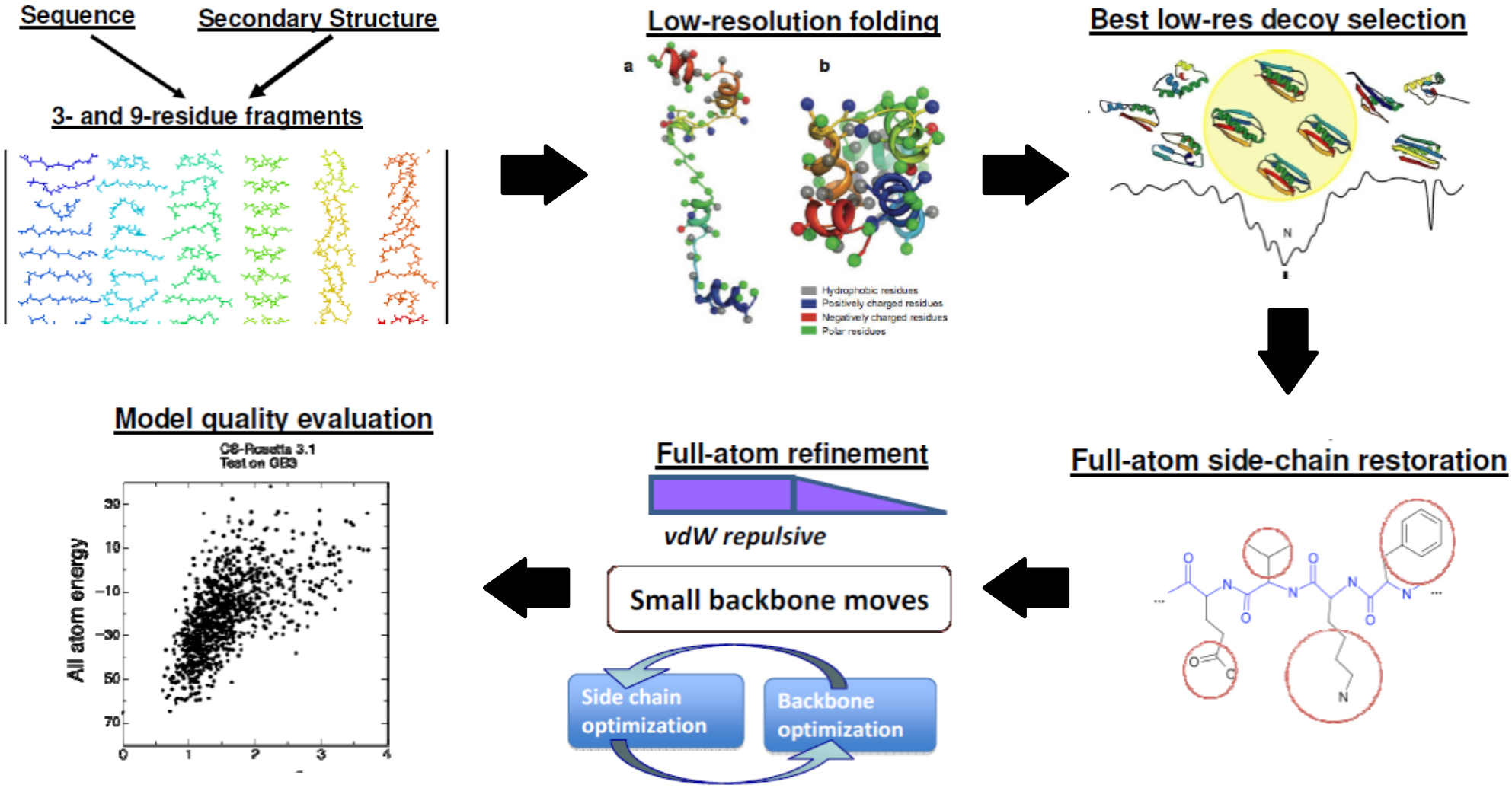




Lindorff-Larsen et al. Science 2011.

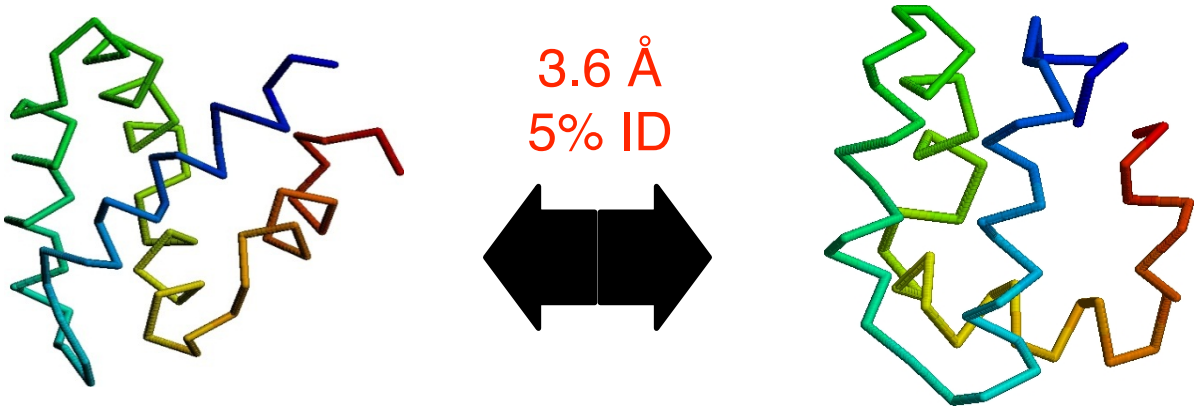
Fragment-based predictions

Rosetta is one of the most accurate fragment-based prediction methods.



Fold Recognition

• Proteins that do not have similar sequences sometimes have similar three-dimensional structures (such as B-barrel TIM fold)



NK-lysin (1nkl)

Bacteriocin T102/as48 (1e68)

• A sequence whose structure is not known is fitted directly (or “threaded”) onto a known structure and the “goodness of fit” is evaluated using a discriminatory function

Threading & Fold Recognition

Generalization of comparative modeling method

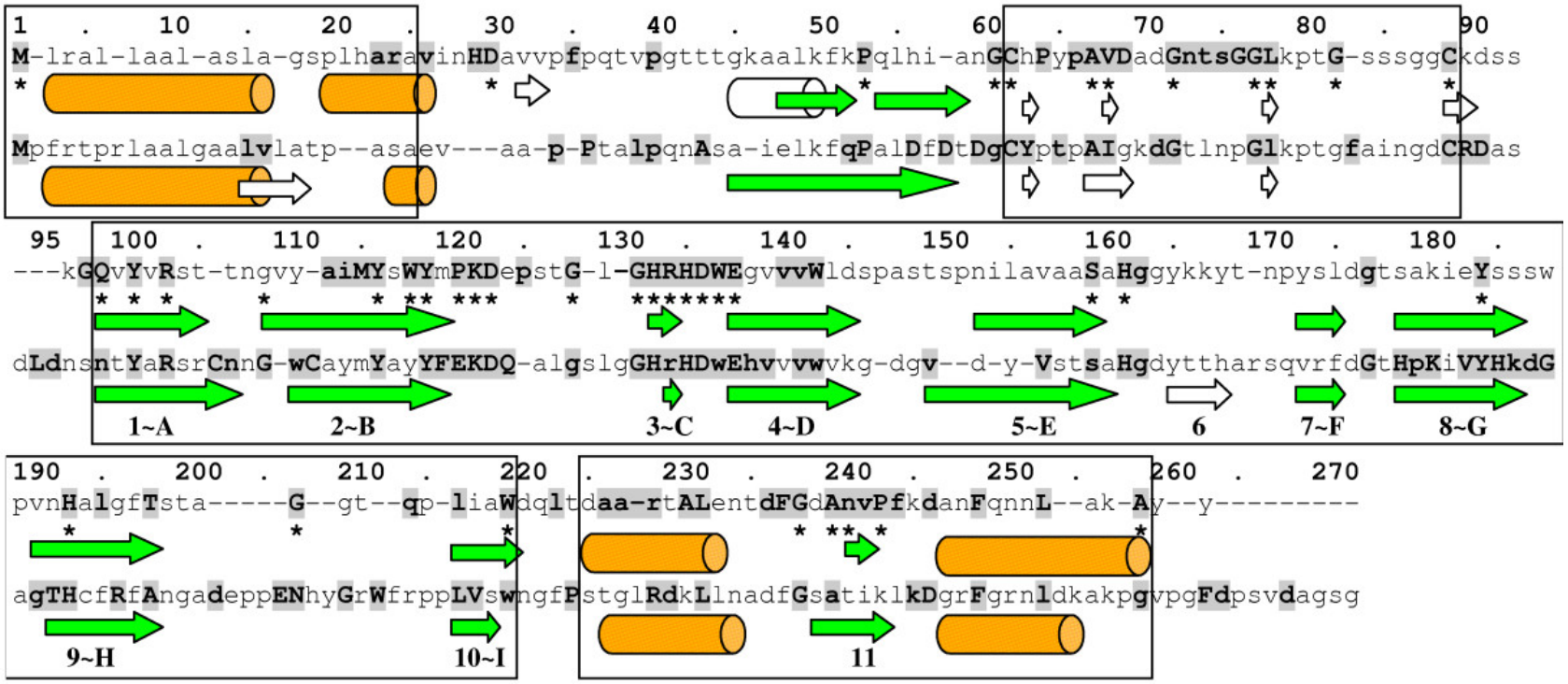
- Homology Modeling: Align sequence to sequence

- Threading: Align sequence to structure (templates)
For each alignment, the probability that that each amino acid residue would occur in such an environment is calculated based on observed preferences in determined structures.

Rationale:
- Limited number of basic folds found in nature
- Amino acid preferences for different structural environments provides sufficient information to choose the best-fitting protein fold (structure)

Fold Recognition approach

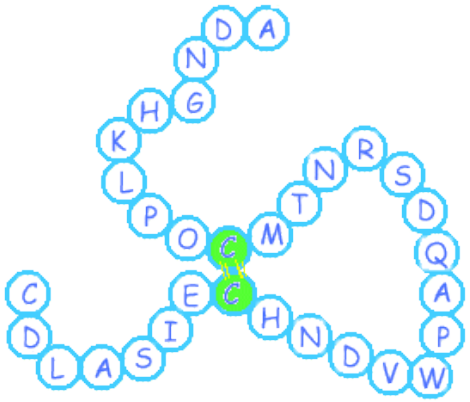
Even if the sequence loses any detectable similarity, secondary structure (and other features such as solvent accessibility profile, disulfide bonds…) should be more conserved

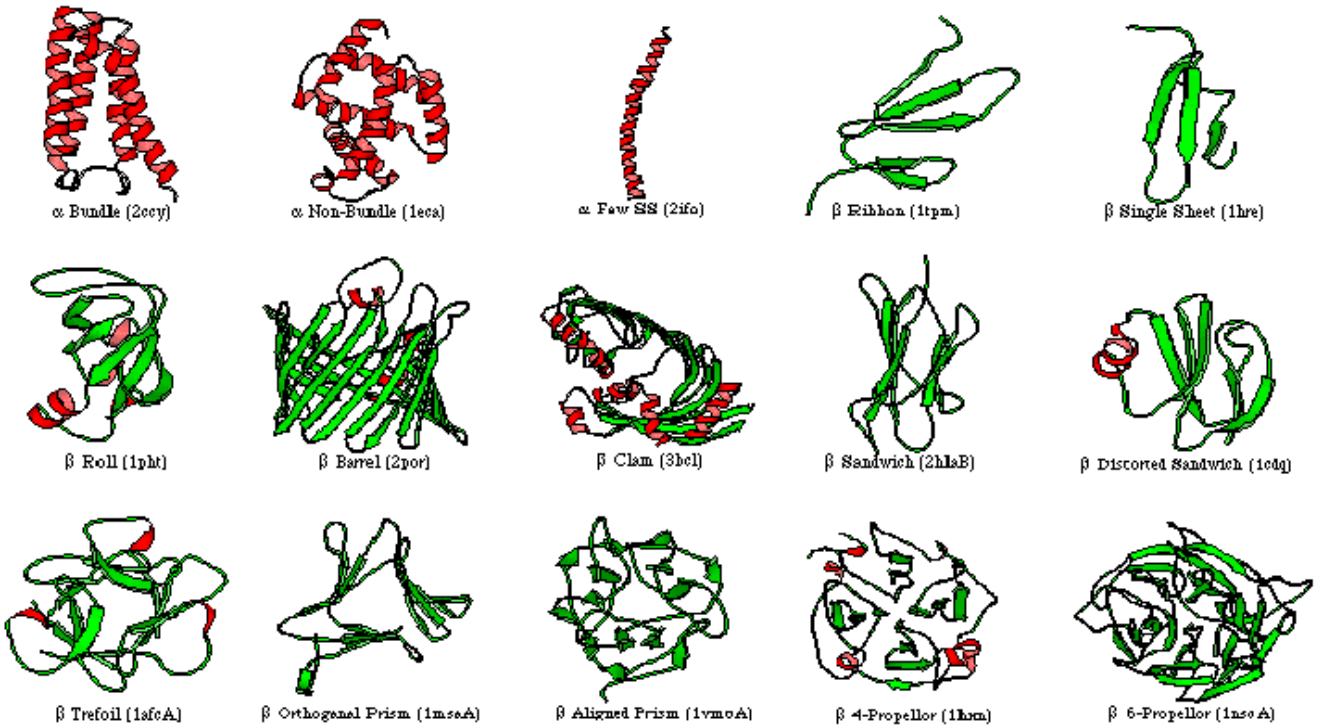


Threading

Does the sequence “fit” on any of a library of known 3D structures?

```
>C562_RHOSH
TQEPGYTRLQITLHWAIAGL…
```





Orengo et al. Structure 1997

Mapping Problem (I)

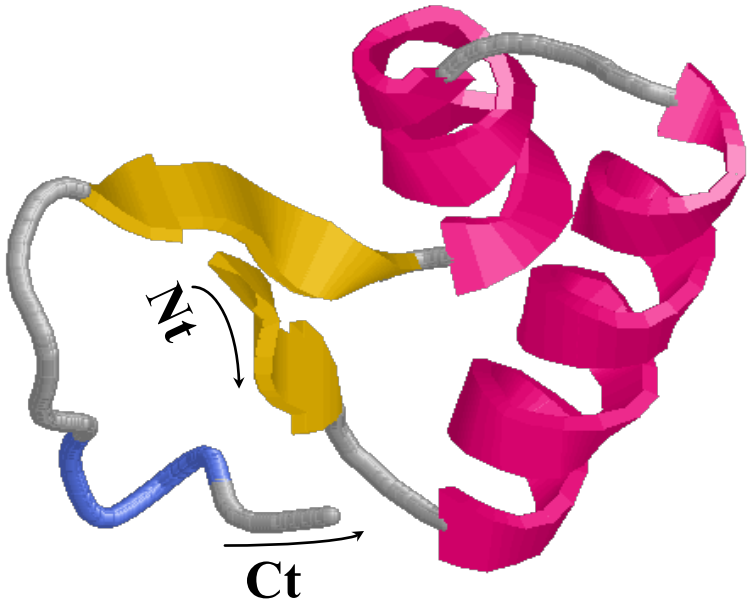
Covalent structure

TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN

Secondary structure

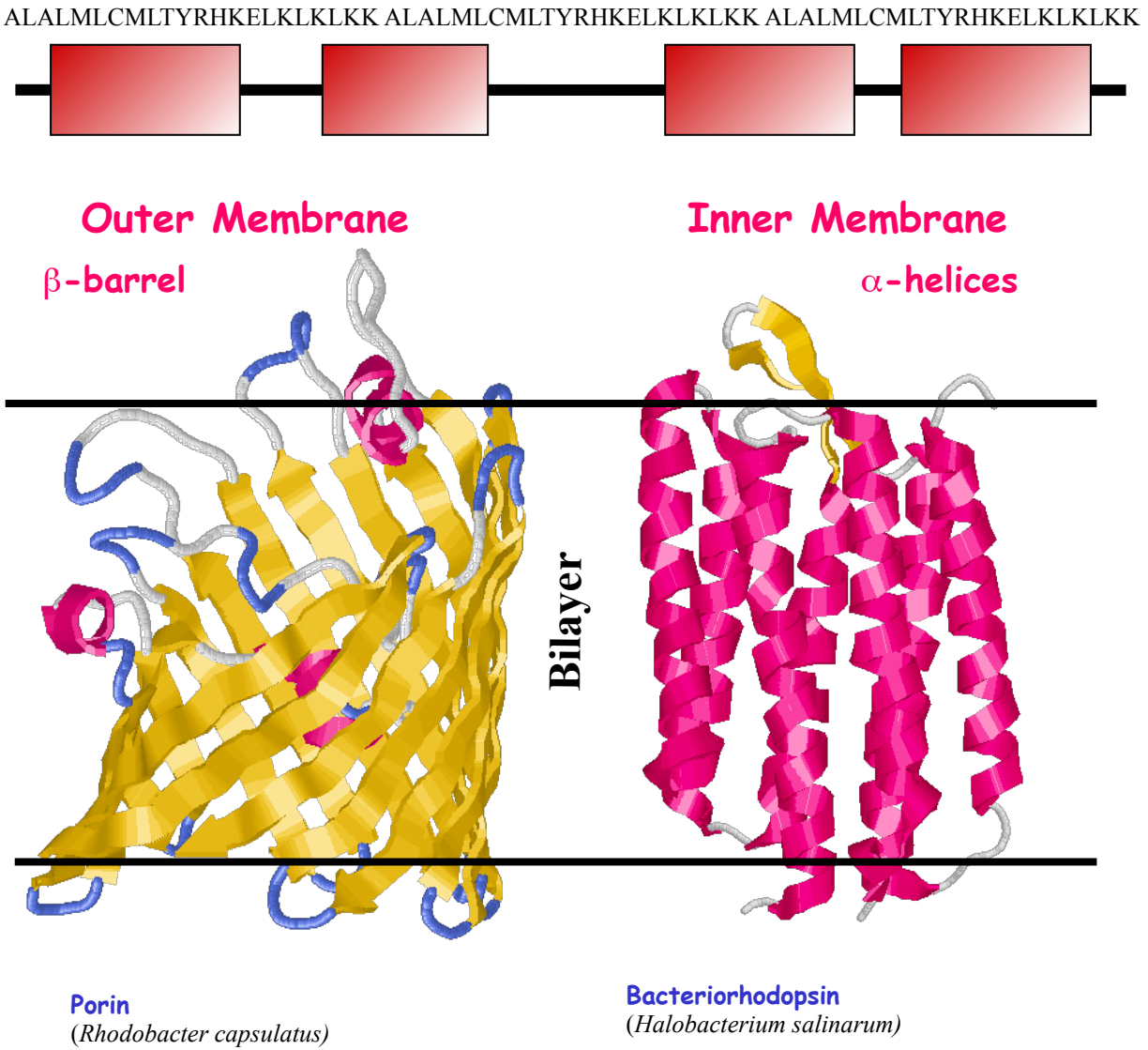
EEEE..HHHHHHHHHHHH....HHHHHHHH.EEEE..........

3D structure



Mapping Problem (II)

Topography: position of Trans Membrane Segments along the sequence



A simple approach

Propensity scales

For each residue

- The association between each residue and the different features is statistically evaluated

- Physical and chemical features of residues

A propensity value for any structure can be associated to any residue

HOW?

Chou-Fasman (I)

Given a set of known structures we can count how many times a residue is associated to a structure.

Example:

AL**A**KSL**A**KPSDTL**A**KSDFREKWEWLKLLK**A**L**A**CCKLS**AA**L
hh**h**hhhh**h**hcccccc**c**cccccchhhhhhhhh**h**h**h**hhhhhh**hh**h

N(A,h) = 7, N(A,c) =1, N= 40

P(A,h) = 7/40, P(A,c) = 1/40

Is that enough for estimating a propensity?

Chou-Fasman (II)

We need to estimate how much independent the residue-to-structure association is.

$$P(h) = 27/40,\ P(c) = 13/40,\ P(A) = 8/40$$

If the structure is independent of the residue:

$$P(A,h) = P(A)P(h)$$

The ratio $P(A,h)/P(A)P(h)$ is the propensity

The prediction method

The Chou-Fasman method was published in 1974 and the propensity scales were calculated on a set of 19 proteins.

Helical Residues[b]	P_α		β-Sheet Residues[c]	P_β	
$Glu^{(-)}$	1.53		Met	1.67	
Ala	1.45	H_α	Val	1.65	H_β
Leu	1.34		Ile	1.60	
$His^{(+)}$	1.24		Cys	1.30	
Met	1.20		Tyr	1.29	
Gln	1.17		Phe	1.28	
Trp	1.14	h_α	Gln	1.23	h_β
Val	1.14		Leu	1.22	
Phe	1.12		Thr	1.20	
$Lys^{(+)}$	1.07		Trp	1.19	
Ile	1.00	I_α	Ala	0.97	I_β
$Asp^{(-)}$	0.98		$Arg^{(+)}$	0.90	
Thr	0.82		Gly	0.81	i_β
Ser	0.79	i_α	$Asp^{(-)}$	0.80	
$Arg^{(+)}$	0.79		$Lys^{(+)}$	0.74	
Cys	0.77		Ser	0.72	
Asn	0.73		$His^{(+)}$	0.71	b_β
Tyr	0.61	b_α	Asn	0.65	
Pro	0.59		Pro	0.62	
Gly	0.53	B_α	$Glu^{(-)}$	0.26	B_β

[a] Chou and Fasman (1974). [b] Helical assignments: H_α, strong α former; h_α, α former; I_α, weak α former; i_α, α indifferent; b_α, α breaker; B_α, strong α breaker. I_α assignments are also given to Pro and Asp (near the N-terminal helix) as well as Arg (near the C-terminal helix). [c] β-sheet assignments: H_β, strong β former; h_β, β former; I_β, weak β former; i_β, β indifferent; b_β, β breaker; B_β, strong β breaker. b_β assignment is also given to Trp (near the C-terminal β region).

Chou and Fasman Biochemistry 1974

Updated Chou-Fasman

An update version of the Chou-Fasman propensity scales are available at the AAIndex database.

```
H CHOP780201
D Normalized frequency of alpha-helix (Chou-Fasman, 1978b)
R PMID:364941
A Chou, P.Y. and Fasman, G.D.
T Prediction of the secondary structure of proteins from their amino acid
  sequence
J Adv. Enzymol. 47, 45-148 (1978)
I     A/L     R/K     N/M     D/F     C/P     Q/S     E/T     G/W     H/Y     I/V
     1.42    0.98    0.67    1.01    0.70    1.11    1.51    0.57    1.00    1.08
     1.21    1.16    1.45    1.13    0.57    0.77    0.83    1.08    0.69    1.06
//

H CHOP780202
D Normalized frequency of beta-sheet (Chou-Fasman, 1978b)
R PMID:364941
A Chou, P.Y. and Fasman, G.D.
T Prediction of the secondary structure of proteins from their amino acid
  sequence
J Adv. Enzymol. 47, 45-148 (1978)
I     A/L     R/K     N/M     D/F     C/P     Q/S     E/T     G/W     H/Y     I/V
     0.83    0.93    0.89    0.54    1.19    1.10    0.37    0.75    0.87    1.60
     1.30    0.74    1.05    1.38    0.55    0.75    1.19    1.37    1.47    1.70
//
```

http://www.genome.jp/aaindex/

Secondary Structure

Given a new sequence a secondary structure prediction can be obtained by plotting the propensity values for each structure, residue by residue

	Y	S	P	Y	A	E	L	M	R	S	Y	G
P(H)	69	77	57	69	142	151	121	145	98	77	69	57
P(E)	147	75	55	147	83	37	130	105	93	75	147	75

Considering three secondary structures (H,E,C), the overall accuracy, as evaluated on an uncorrelated set of sequences with known structure, is very low

Accuracy = 50/60 %

Chou-Fasman Algorithm

Conformational parameter: P_a, P_β and P_t for each amino acid i

$$P_{i,x} = f_{i,x} / \langle f_x \rangle = (n_{i,x} / n_i) / (n_x / N)$$

Nucleation sites and extension

Clusters of four helical formers out of six propagated by four residues

if $$\langle P_a \rangle = \sum_{1}^{4} P_a / 4 \geq 1.00$$

Clusters of three β-formers out of five propagated by four residues

if $$\langle P_\beta \rangle = \sum_{1}^{4} P_\beta / 4 \geq 1.00$$

Clusters of four turn residues

if $$P_t = f_j \times f_{j+1} \times f_{j+2} \times f_{j+3} > 0.75 \times 10^{-4}$$

Specifics thresholds for $< P\alpha >$, $< P_\beta >$ and $< P_t >$ and their relatives values decide for the prediction

Kyte-Doolittle scale

It is computed taking into consideration the octanol-water partition coefficient, combined with the propensity of the residues to be found in known transmembrane helices

```
H KYTJ820101
D Hydropathy index (Kyte-Doolittle, 1982)
R PMID:7108955
A Kyte, J. and Doolittle, R.F.
T A simple method for displaying the hydropathic character of a protein
J J. Mol. Biol. 157, 105-132 (1982)
I     A/L     R/K     N/M     D/F     C/P     Q/S     E/T     G/W     H/Y     I/V
      1.8    -4.5    -3.5    -3.5     2.5    -3.5    -3.5    -0.4    -3.2     4.5
      3.8    -3.9     1.9     2.8    -1.6    -0.8    -0.7    -0.9    -1.3     4.2
//
```

Exercise

Develop your own alpha helix propensity scale based on the non redundant PDB structures with resolution below 2 Å and with more than 50 residues.

Compare your scale with the AAindex Chou-Fassman scale

Write a script that given a sequence and propensity scale calculates the smoothed score on a window sequence.

Second generation methods

The structure of a residue in a protein strongly depends on the sequence context

It is possible to estimate the influence of a residue in determining the structure of a residue close along the sequence. Usually windows from -8/8 to -13/13 are considered.

Coefficients P(A,s,i) estimate the contribution of the residue A in determining the structure s for a residue that is i positions apart along the sequence

GOR method

- Garnier, Osguthorpe & Robson
- Assumes amino acids up to 8 residues on each side influence the ss of the central residue.
- Frequency of amino acids at the central position in the window, and at -1, ... -8 and +1,....+8 is determined for a, b and turns (later other or coils) to give three 17 x 20 scoring matrices.
- Calculate the score that the central residue is one type of ss and not another.
- Correctly predicts ~64%.

Scoring Matrix

$$S_{ss}^{ij} = \log \frac{P(ss_i \mid aa_{i+j})}{p(ss_i)}, j = -8, \ldots, 8$$

	.	A	Y	E	D	Y	R	H	F	S	.
	…	- 4	-3	-2	-1	0	1	2	3	4	…
A	..	..	..	..	..	..	..	..	..	..	
B	..	..	..	..	..	..	..	..	..	..	

Information Function

$$I(S_j; R_j) = \log \frac{P(S_j \mid R_j)}{p(S_j)}$$

Information function, $I(S_j;R_j)$:

S_j = one of three secondary structure (H, E,C) at position j

R_j = one of the 20 amino acids at position j

$p(S_j|R_j)$ = conditional probability for observing S_j having R_j

$p(S_j)$ = prior probability of having S_j

- Information that sequence R_j contains about structure S_j
  - I = 0 : no information
  - I > 0 : R_j favors S_j
  - I < 0 : R_j dislikes S_j

GOR approximation

- Secondary structure should depend on the whole sequence, R
- Simplification (1) : only local sequences (window size = 17) are considered

$$I = (S_j; \mathbf{R}) \approx I(S_i; R_{j-8}, \ldots, R_j, \ldots, R_{j+8})$$

- Simplification (2) : each residue position is statistically independent.
- For independent event, just add up the information

$$I(S_i; R_{j-8}, \ldots, R_j, \ldots, R_{j+8}) \simeq \sum_{m=-8}^{8} I(S_j; R_{j+m})$$

GOR Scores

$$I(S_i; R_{j-8}, \ldots, R_j, \ldots, R_{j+8}) \simeq \sum_{m=-8}^{8} I(S_j; R_{j+m})$$

Directional information measure for the α-helical conformation†

Amino acid residue	Residue position‡ (centinats) $j-8$		$j-6$		$j-4$		$j-2$		j	$j+2$		$j+4$		$j+6$		$j+8$
Gly	−5	−10	−15	−20	−30	−40	−50	−60	−86	−60	−50	−40	−30	−20	−15	−10	−5
Ala	5	10	15	20	30	40	50	60	65	60	50	40	30	20	15	10	5
Val	0	0	0	0	0	0	5	10	14	10	5	0	0	0	0	0	0
Leu	0	5	10	15	20	25	28	30	32	30	28	25	20	15	10	5	0
Ile	5	10	15	20	25	20	15	10	6	0	−10	−15	−20	−25	−20	−10	−5
Ser	0	−5	−10	−15	−20	−25	−30	−35	−39	−35	−30	−25	−20	−15	−10	−5	0
Thr	0	0	0	−5	−10	−15	−20	−25	−26	−25	−20	−15	−10	−5	0	0	0
Asp	0	−5	−10	−15	−20	−15	−10	0	5	10	15	20	20	20	15	10	5
Glu	0	0	0	0	10	20	60	70	78	78	78	78	78	70	60	40	20
Asn	0	0	0	0	−10	−20	−30	−40	−51	−40	−30	−20	−10	0	0	0	0
Gln	0	0	0	0	5	10	20	20	10	−10	−20	−20	−10	−5	0	0	0
Lys	20	40	50	55	60	60	50	30	23	10	5	0	0	0	0	0	0
His	10	20	30	40	50	50	50	30	12	−20	−10	0	0	0	0	0	0
Arg	0	0	0	0	0	0	0	0	−9	−15	−20	−30	−40	−50	−50	−30	−10
Phe	0	0	0	0	0	5	10	15	16	15	10	5	0	0	0	0	0
Tyr	−5	−10	−15	−20	−25	−30	−35	−40	−45	−40	−35	−30	−25	−20	−15	−10	−5
Trp	−10	−20	−40	−50	−50	−10	0	10	12	10	0	−10	−50	−50	−40	−20	−10
Cys	0	0	0	0	0	0	−5	−10	−13	−10	−5	0	0	0	0	0	0
Met	10	20	25	30	35	40	45	50	53	50	45	40	35	30	25	20	10
Pro	−10	−20	−40	−60	−80	−100	−120	−140	−77	−60	−30	−20	−10	0	0	0	0

† The data for Tables 1 to 4 are obtained from 25 proteins by Robson & Suzuki (1976), but the values quoted here are read from curves fitted through the directional plots. The coil values come from the same source but have not previously been quoted. Values are in centinats (nats × 100).

‡ For example, the information at position $j - 6$ is the information which the residue j carries about the conformation of any residue 6 away in the N-terminal direction and at position $j + 6$ about any residue 6 away in the C-terminal direction (see Robson & Suzuki, 1976). At position j, it is the information carried by the residue itself to be in the given conformation (single-residue information).

GOR performance

Information scores obtained on a set of 25 proteins.

Accuracy = 60-65 % (Considering three secondary structures (H,E,C), and evaluating the overall accuracy on an uncorrelated set of sequences with known structure)

The contribution of each position in the window is independent of the other ones. No correlation among the positions in the window is taken in to account.

Garnier et al. J Mol Biol. 1978.