Towards the prediction of residues involved in the folding nucleus of proteins
Dimacs, May 2006

Jacques CHOMILIER, Mathieu LONQUETY
IMPMC, Paris
Nikolaos PAPANDREOU, AUA, Athens
Igor BEREZOVSKY, Harvard

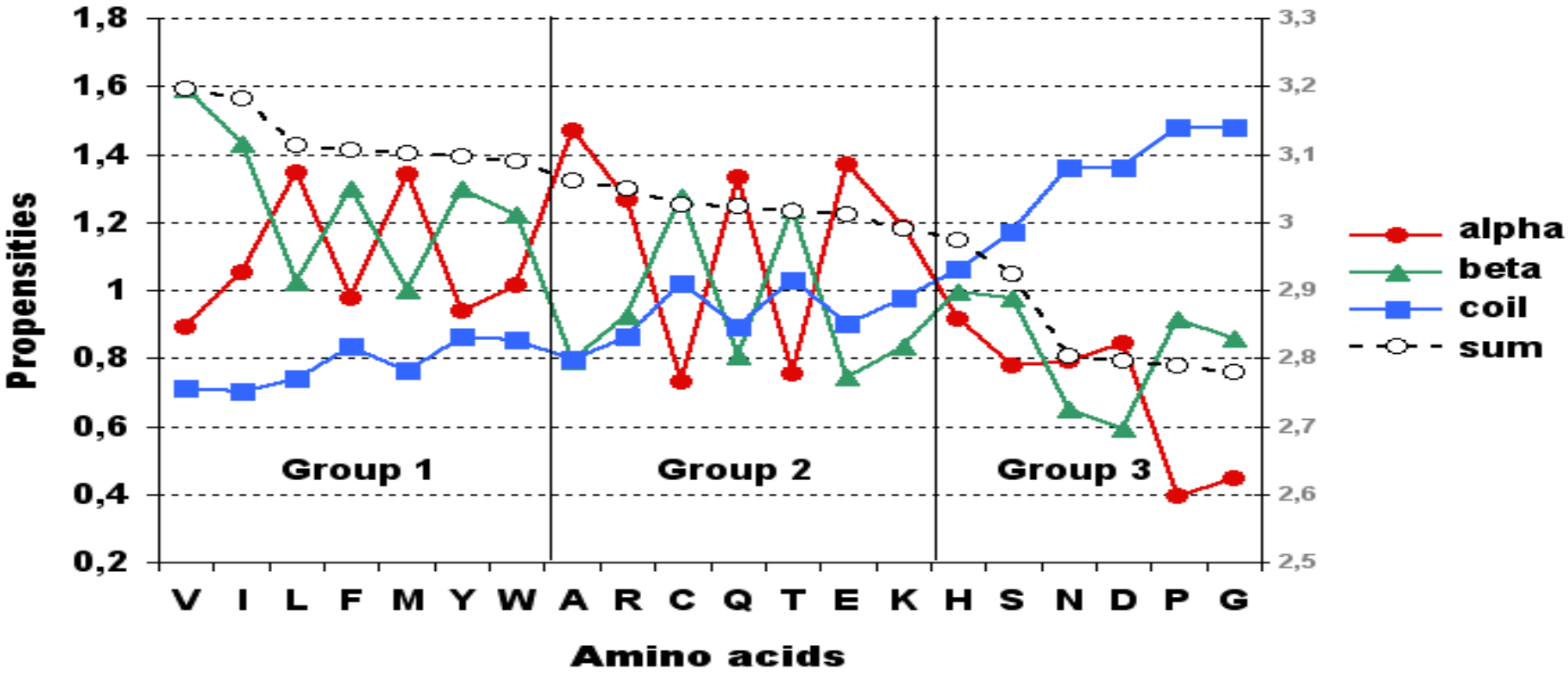
Topohydrophobic positions

- Bressler & Talmud (1944) : a globular protein is made of a hydrophobic core (1/3 of the AA)

- Analysis of the core from the structures
  - Families of structures. Sequence identity $\leq$ 25%
  - Superposition of structures
  - Derived multiple alignment
  - Positions with only hydrophobic residues (VILMFYW) are called Topohydrophobic positions

Ref: Poupon & Mornon. Proteins. 1998 33:329-42

Amino acid groups



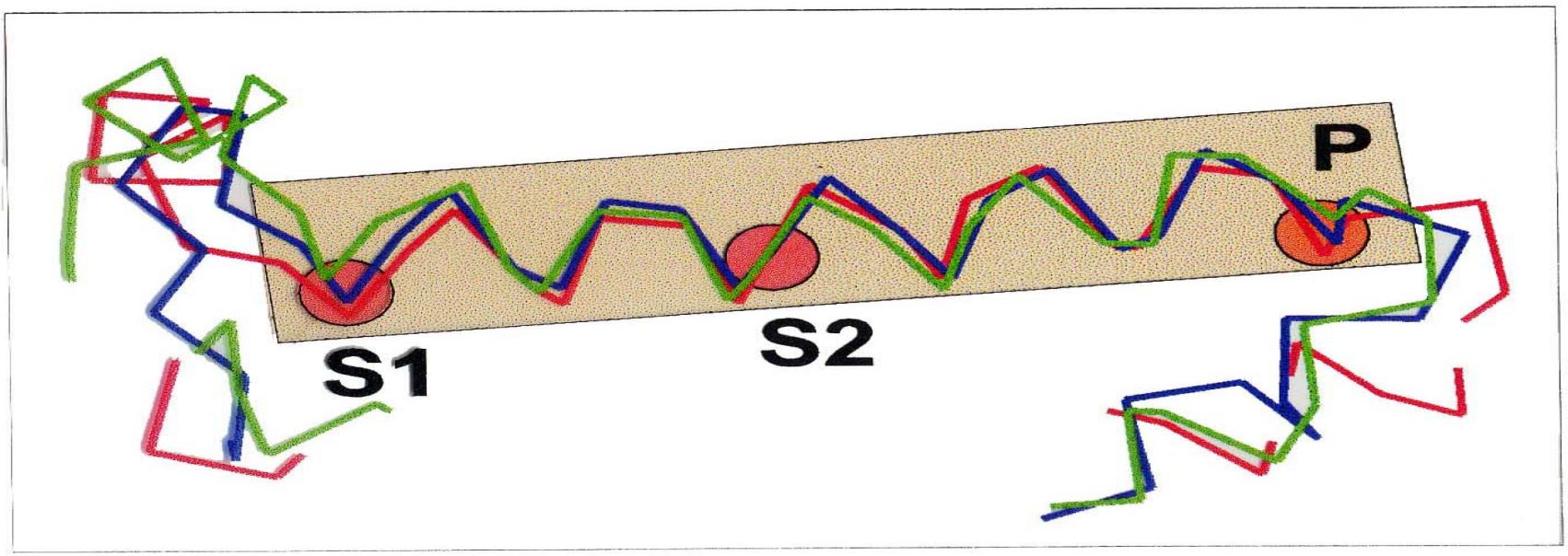
Strict = group 1 = VILFMYW

Extended = no group 3, 75% group 1 at least

Topohydrophobic positions

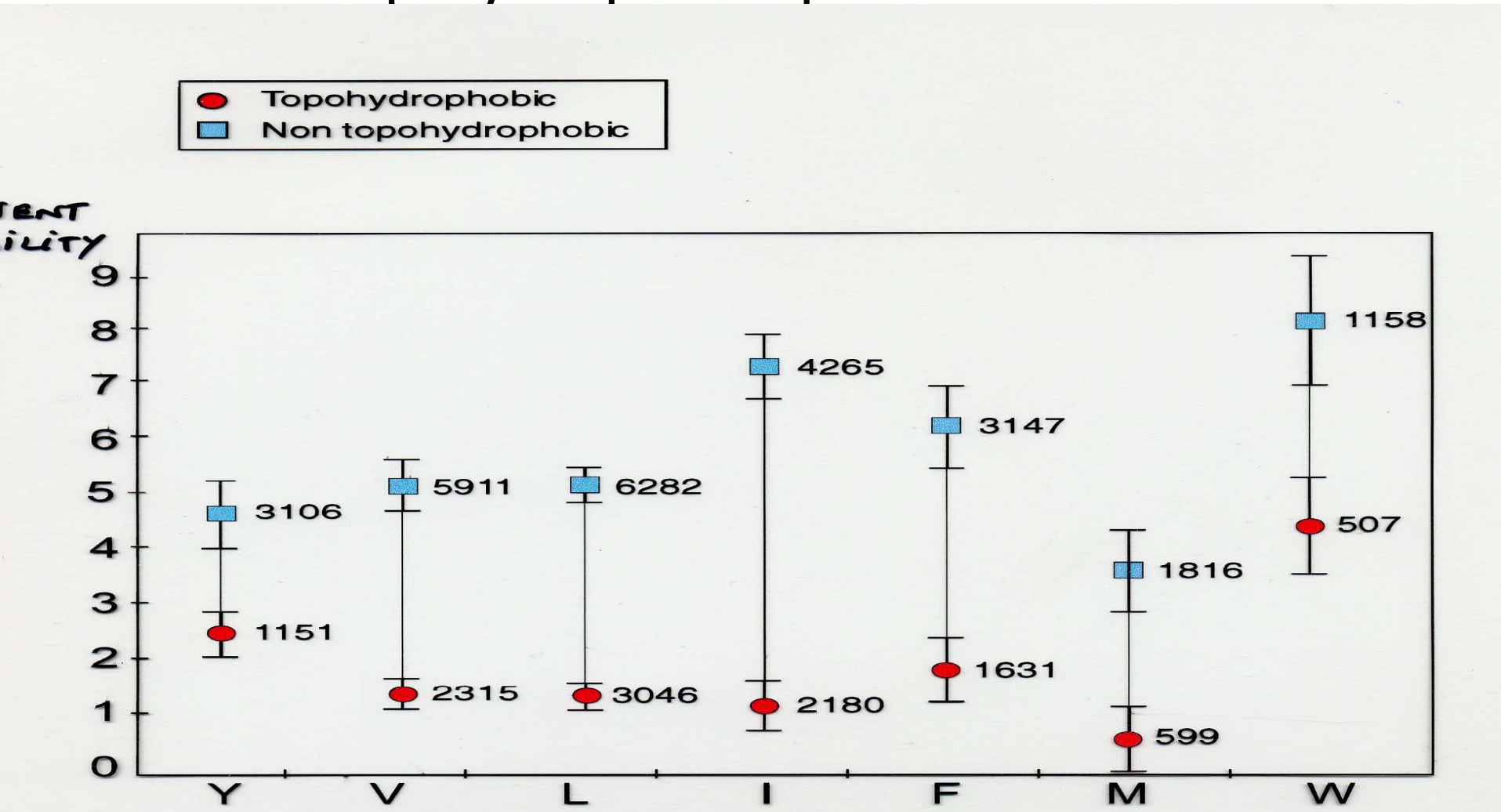
Globin family

```
                 S1          S2          P
1dxtb    . . PKVKAHGK  KVLGAFSDGL  AH
1hlbx  SSRQMQAHAI  RVSSIMSEYV  . .
1ecax    . . APFETHAN  RIVGFFSKII  GE
1flpx    . . PEMAAQAQ  SFKGLVSNWV  DN
1ashx  NDPFFAKQGQ  KILLACHVLC  AT
```

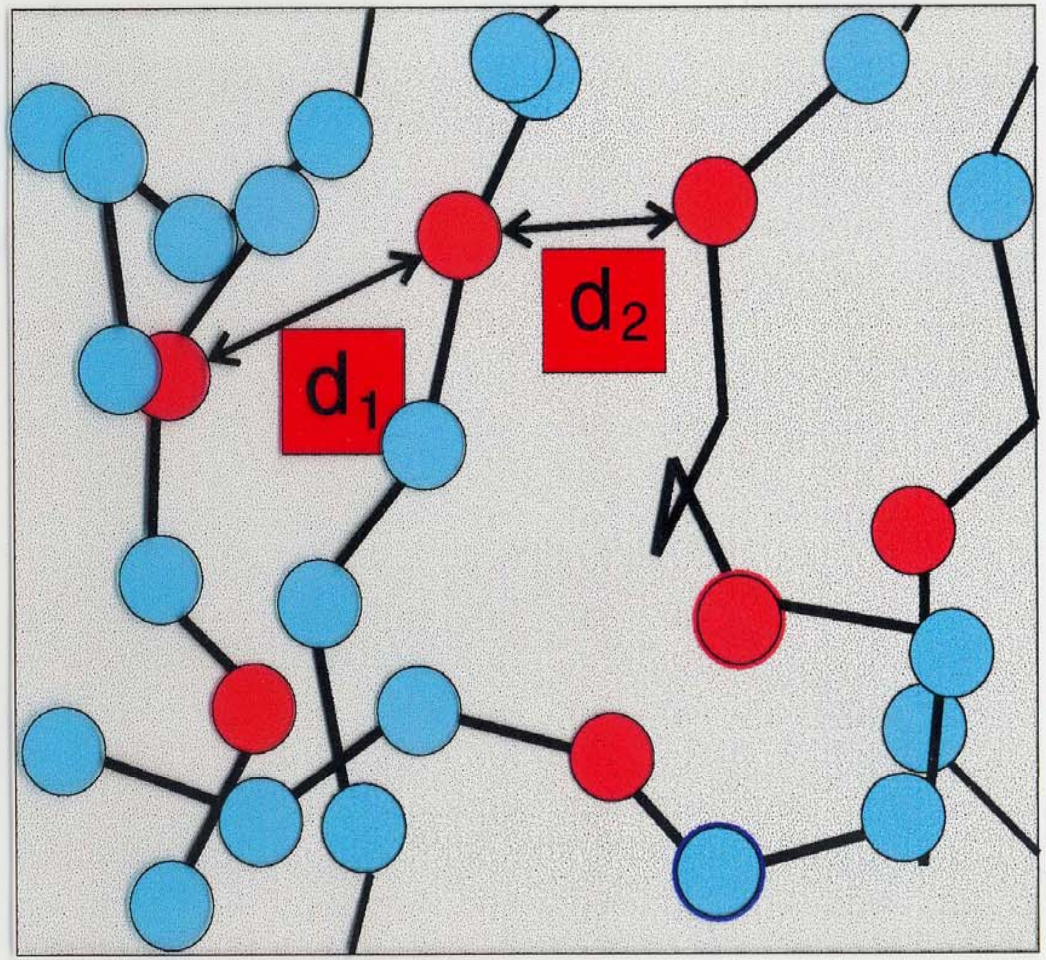


Solvent accessibility

Hydrophobic AA more buried at topohydrophobic positions



Topohydrophobic positions network



M = 5,85 Å
σ = 1,25 Å

Topohydrophobic positions form a network within the internal core

The core of the core

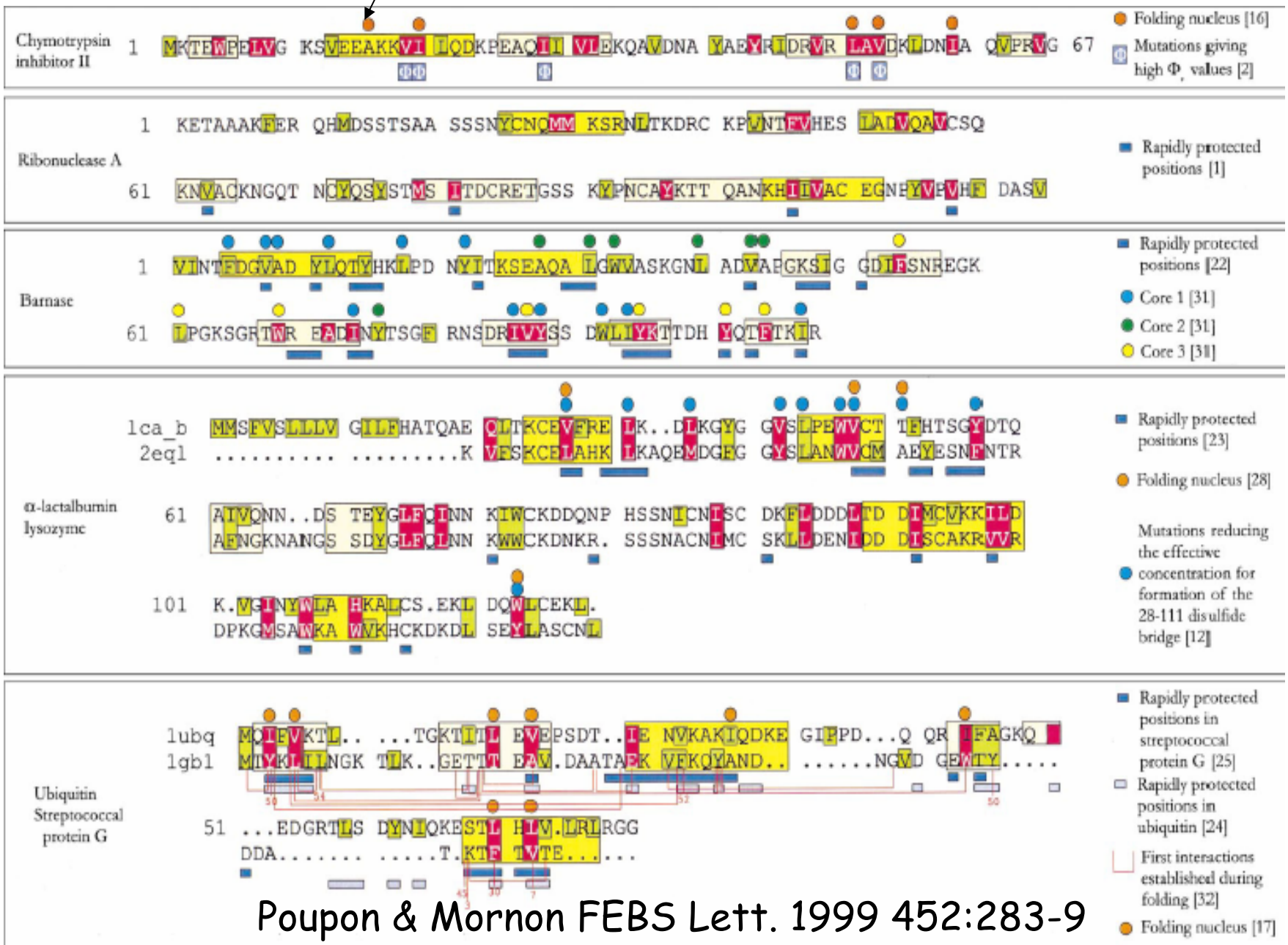
- Mean number of Topohydrophobic positions in:
  - Helices = 2.25
  - Strands = 1.67
  - Loops = 0.54

- Residues occupying TH positions are related by a set of distances smaller than other unconserved hydrophobic positions

- One third of Hydrophobic are TH

- Statiscally correspond to the folding nucleus

The folding nucleus



Poupon & Mornon FEBS Lett. 1999 452:283-9

Limits or difficulties

- Both ways possible to determine Topohydrophobic positions : Structure or Sequence
- Structural family of high divergence <25% ID: Algorithms do not give same results
- Multiple alignment difficult for sequences <25% ID (Not automatic)

Automatic TH

⟹ Retrieve members of families from PDB bank with CE

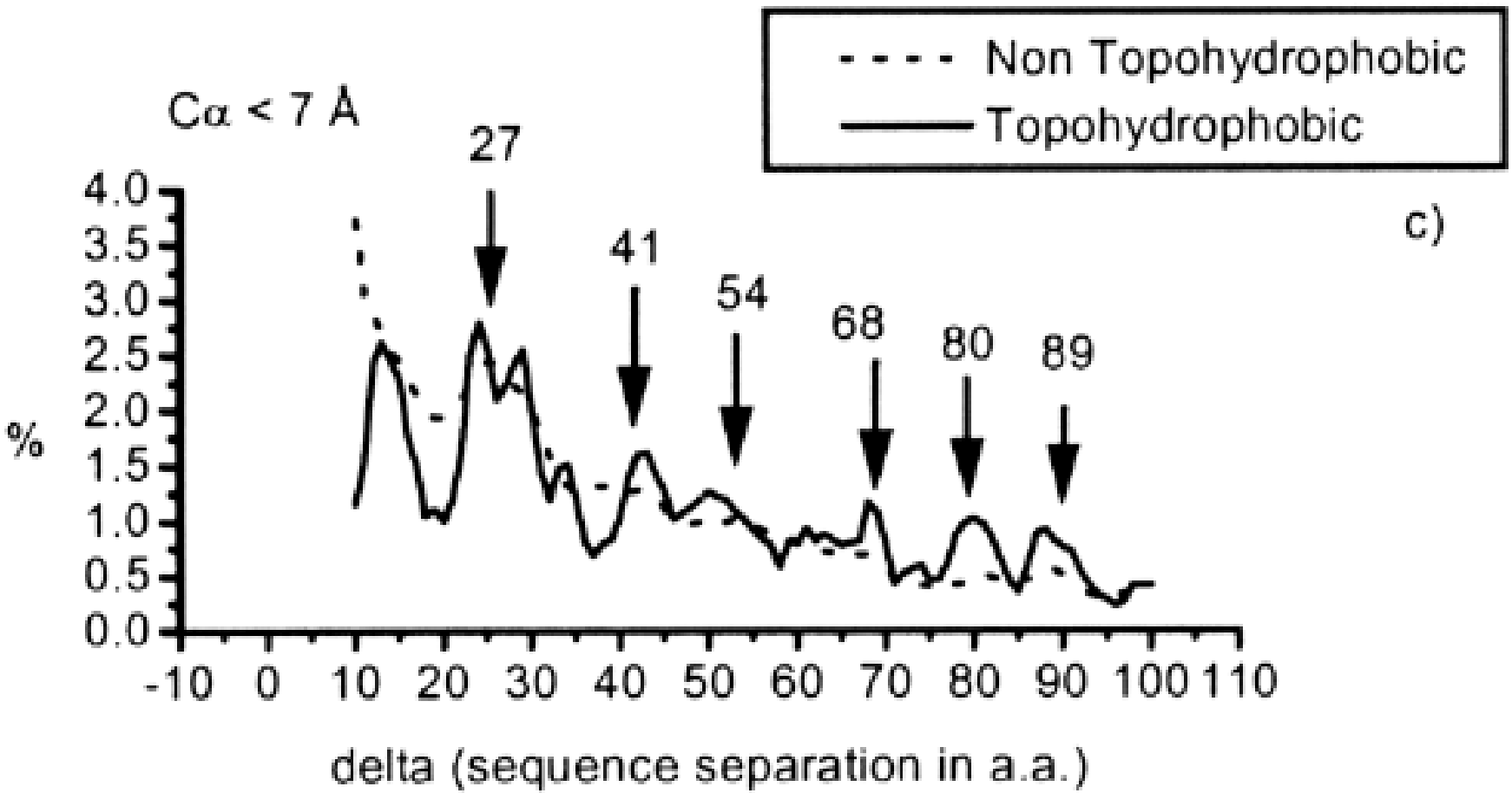
⟹ **3 servers of Multiple structural alignment**

- SSM (Secondary Structure Matching)
- CE (Combinatorial Extension)
- MATRAS

⟹ Choice of a consensus of the two programs which give consistent results

Topohydrophobic positions

Distance distribution (in sequence) among TH which are close in 3D space : frequency of separation



Comparative literature

Universally conserved positions in protein folds… Shakhnovich… JMB (1999) 291:177-196

Conserved Key Amino Acids Positions (CKAAPs)… P. Bourne… Proteins (2001) 42:148-163. /ckaaps.sdsc.edu/

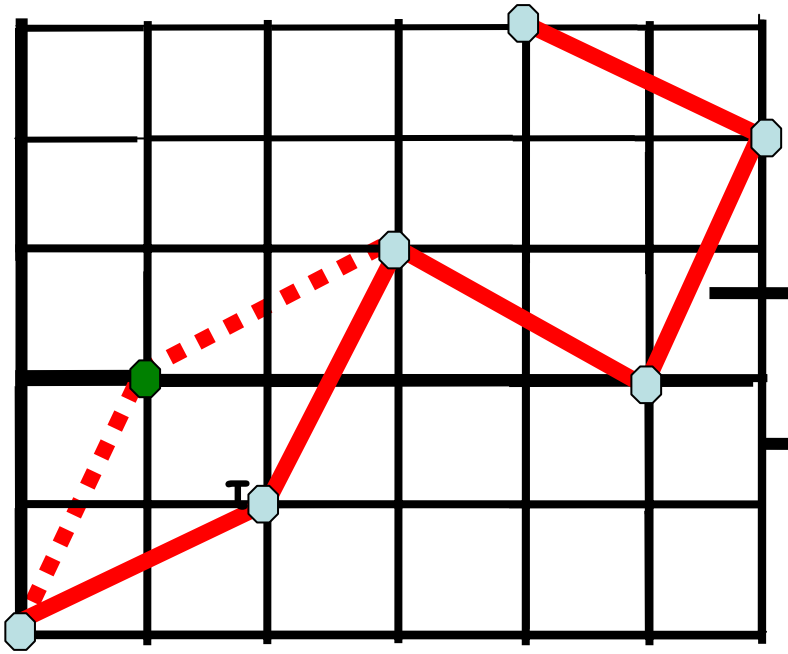
Non functional conserved residues in globins and their possible role as a folding nucleus. Ptitsyn… JMB (1999) 291:671-682

Protein structural alignments and functional genomics. Lesk… Proteins (2001) 42:378-382

How to predict the folding nucleus?

- Prediction of topohydrophobic positions
- Lattice simulation
- Monte Carlo procedure

Folding simulation

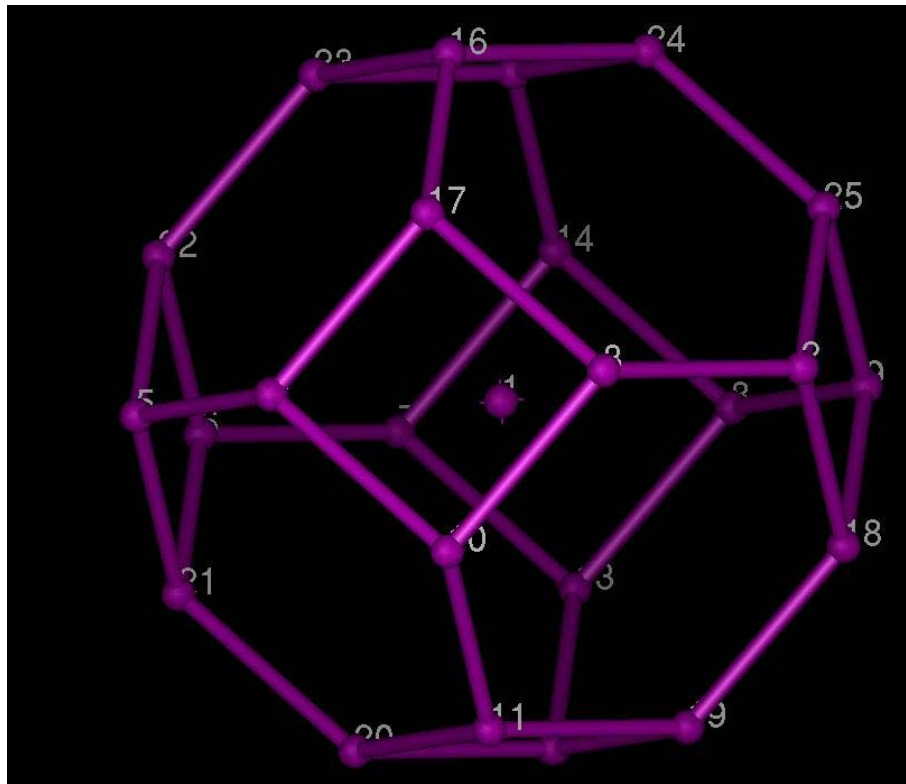


7 values for τ : 64° to 143°

3.8 Å

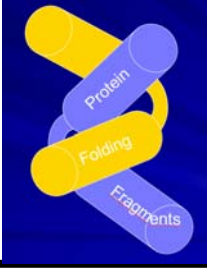
1.7 Å

24 first neighbours



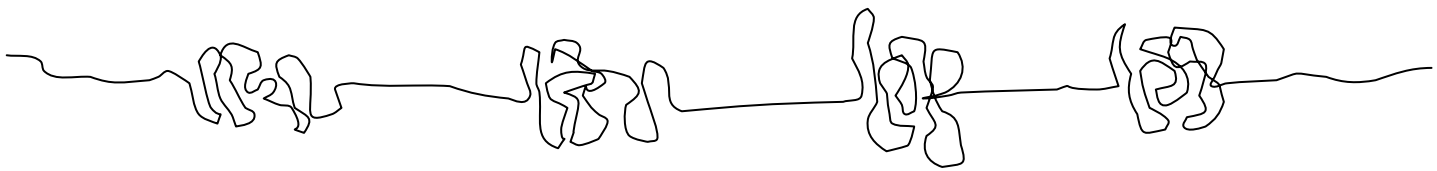
Lattice (2,1,0)
Skolnick, Kolinski *J. Mol. Biol. 221:499 (1991)*

Lattice simulation



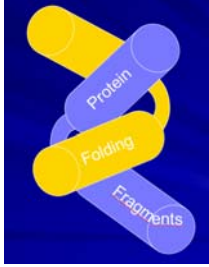
Initial state: unfolded chain; 100 initial states

Observation of compact fragments at the beginning of the simulation (10^6 MC steps)



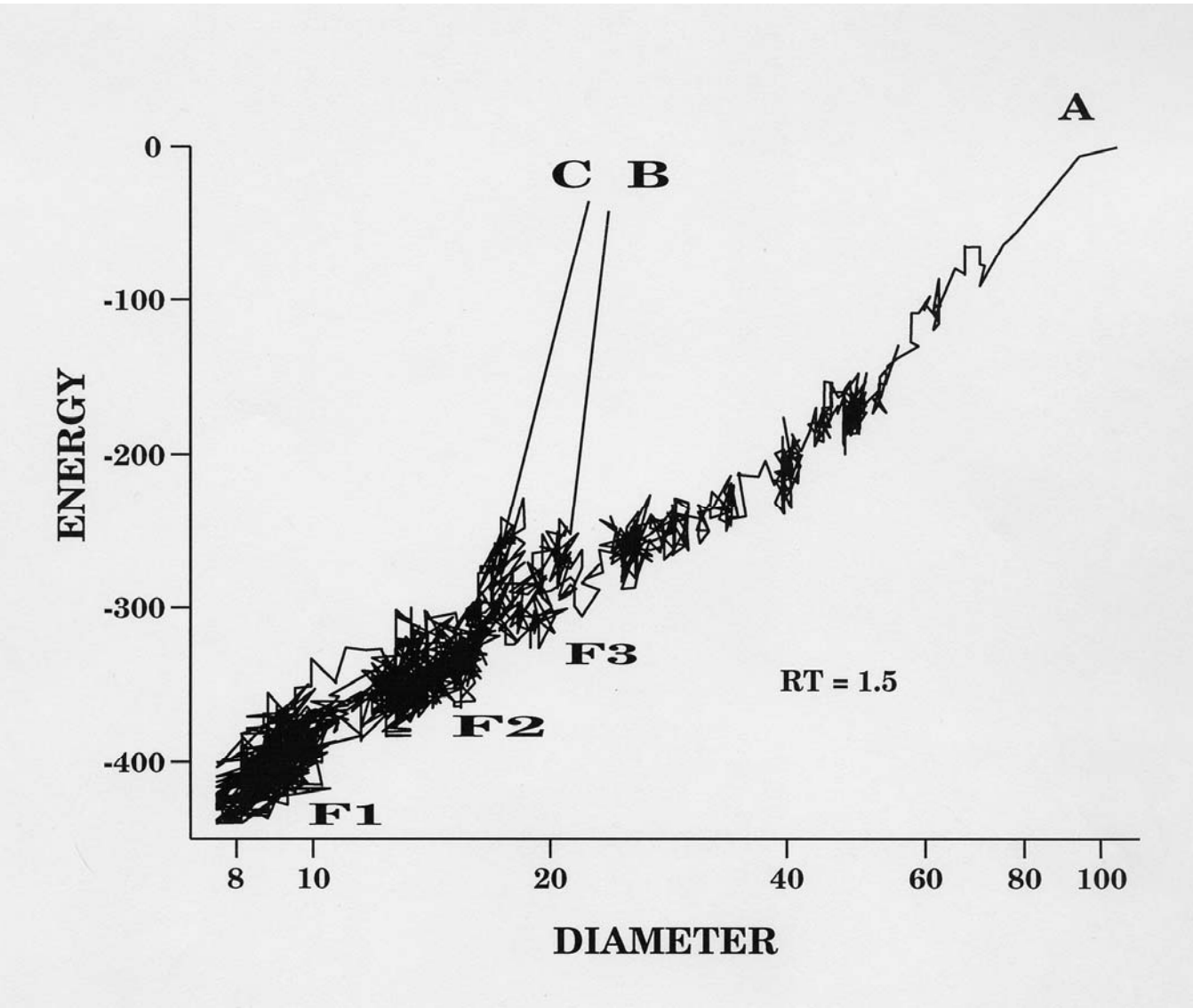
Fragments are stable in sequence
Inter fragment regions = loops

Time of simulation



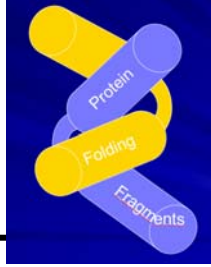
- t_{min} = INT(10^5L/50)
  - L length of the sequence
- t_{max} = 10 t_{min}
- Typical 10^5-10^6 MC steps

First steps of simulation (~10^6 MC)

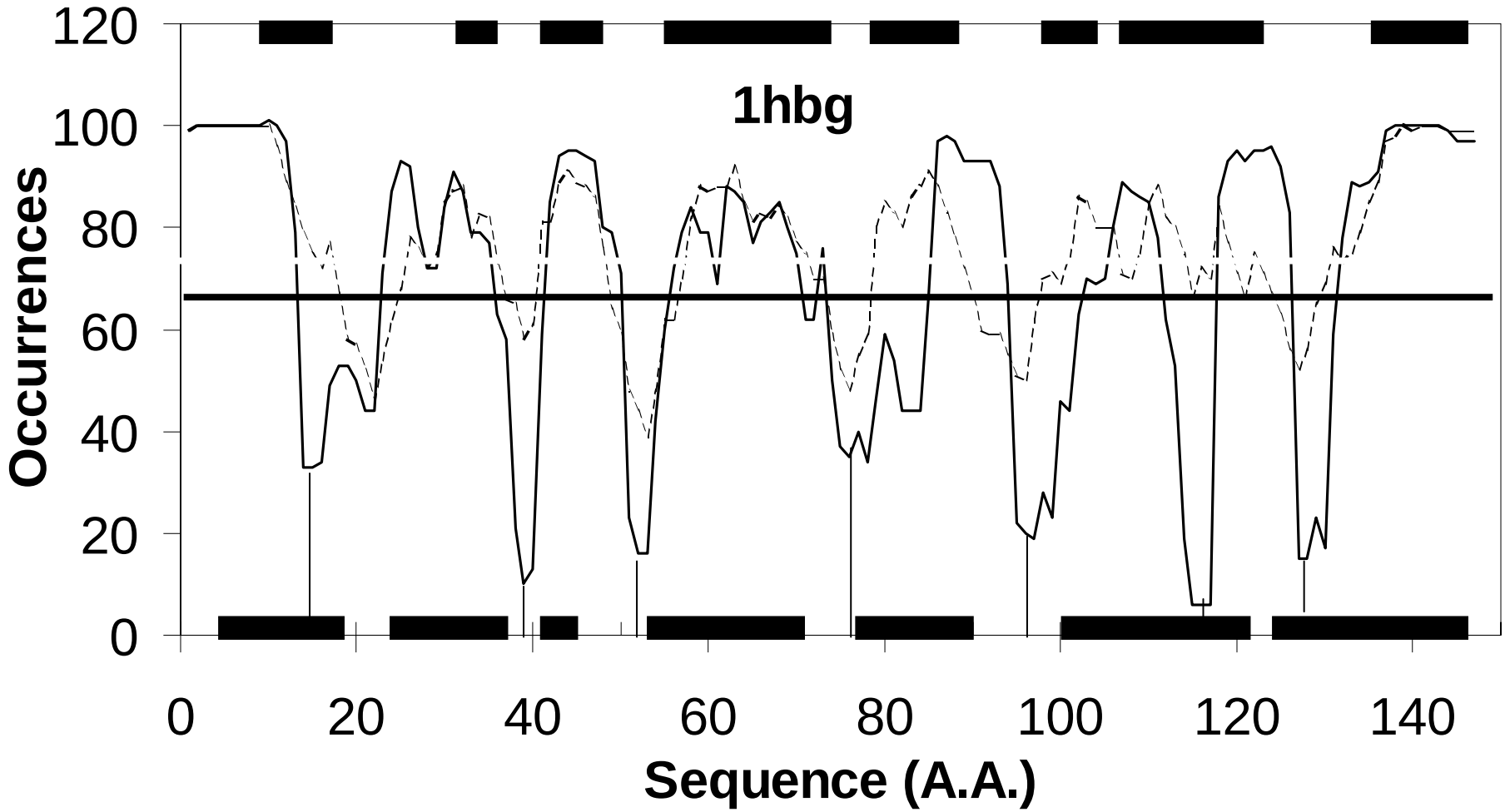


- FKBP
- 3 inital conformations *A, B, C*
- States of 3, 2 and 1 fragment

Fragments in the first MC steps

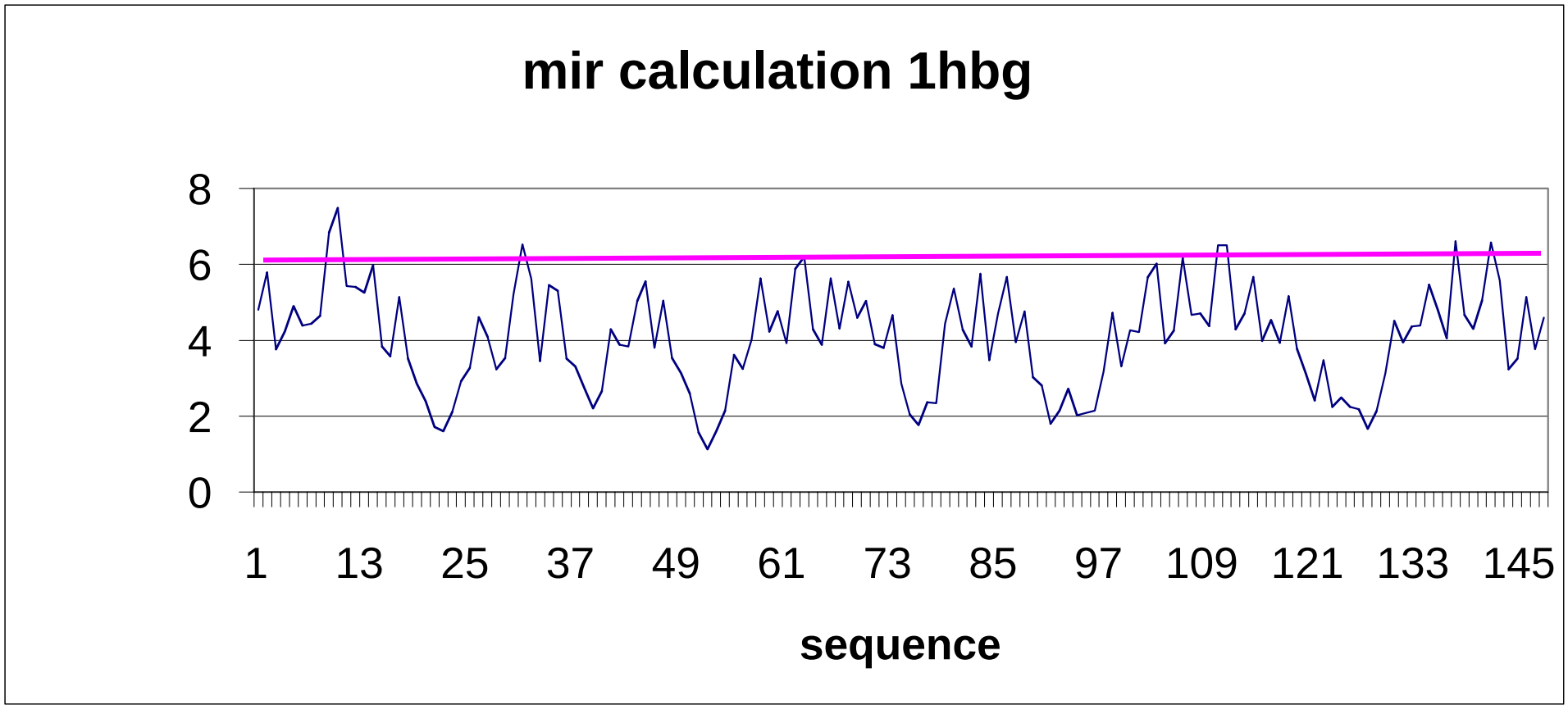


Bottom : secondary structures

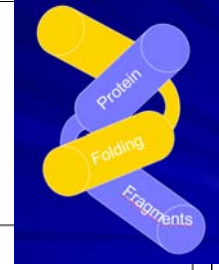


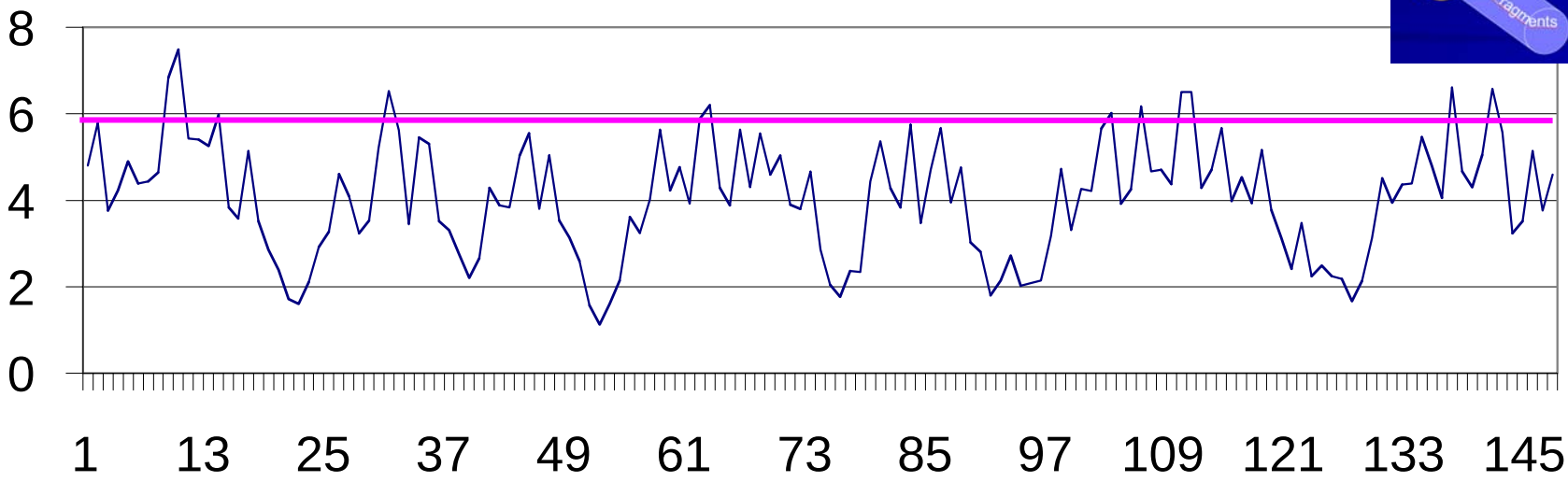
Mean Number of contacts during simulation

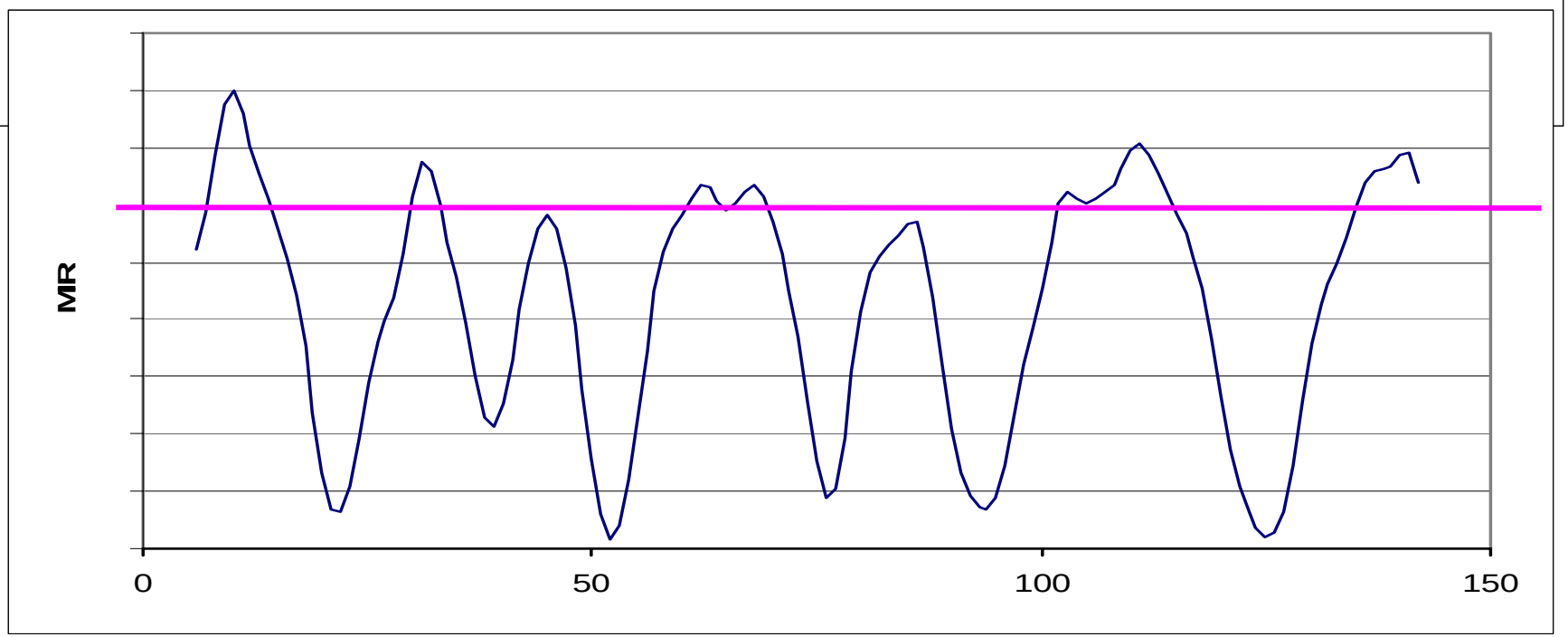
For each residue, number of non-covalent neighbours (NCN)
MIR=(NCN ≥ 6), Most Interacting Residues

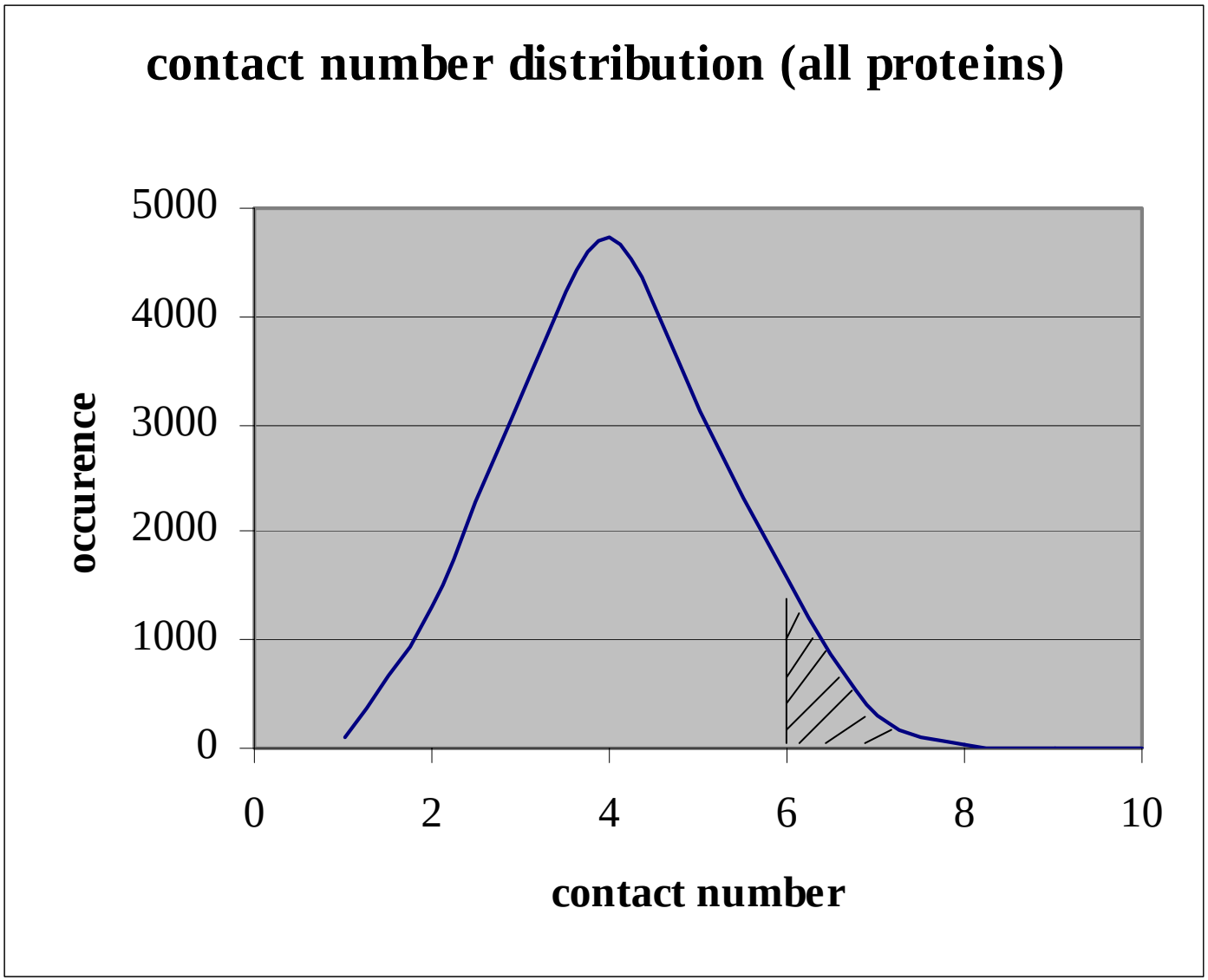


mir calculation 1hbg





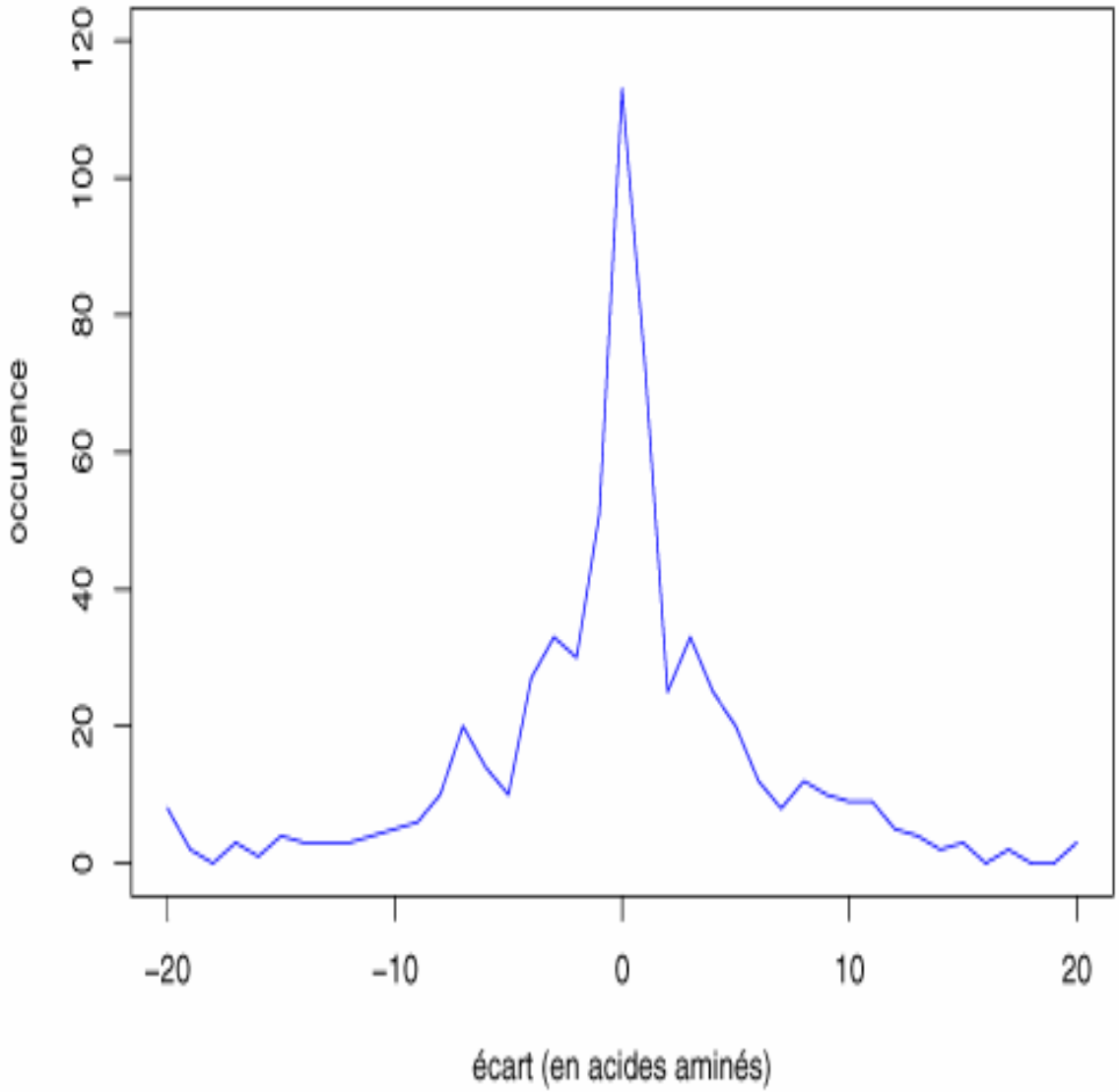




13% of residues have NCN ≥ 6

92% of MIR are hydrophobic (VIMWYLF)

Most Interacting Residues (MIR)



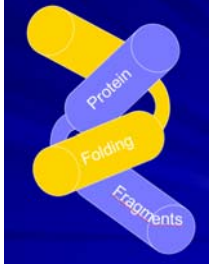
92% of MIR are Hydrophobic

MIR are in compact fragments ⇒ Core

65 % MIR: topohydrophobes ±3AA

Multiple alignment:90%

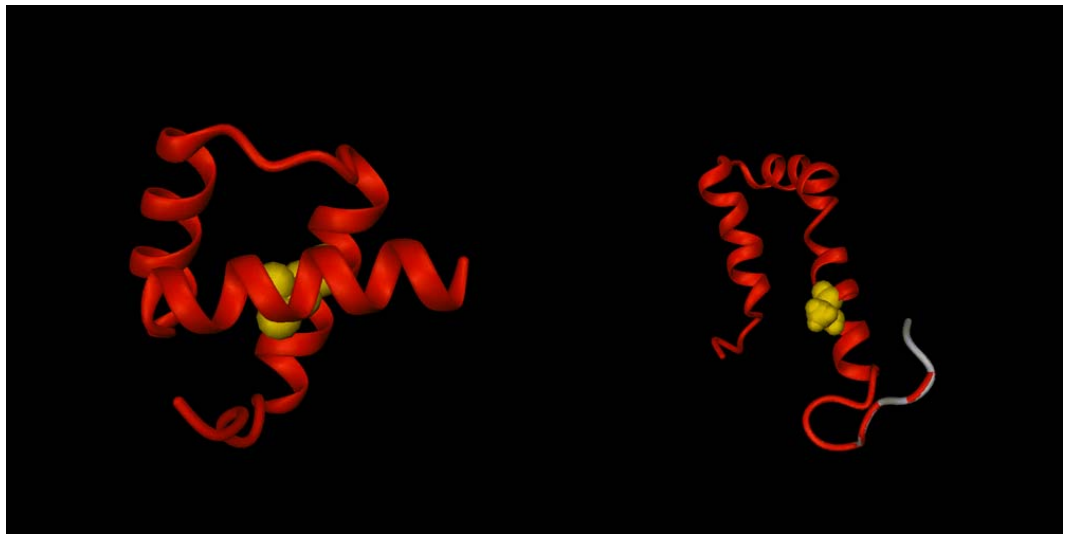
MIR & nucleus



- Prediction of the folding nucleus :
  - MIR = Prediction of topohydrophobic positions from a sequence or a multiple alignment
  - Residues involved in the folding nucleus do correspond to TH

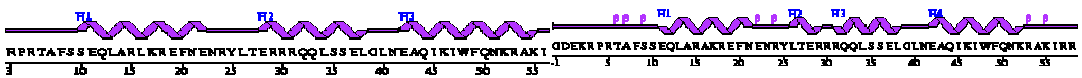
1enh
Homeodomain

ASA=4000Å²



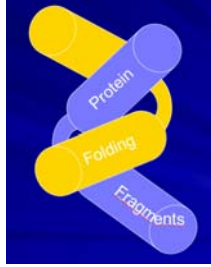
1ztr
L16A

ASA6500Å²



- Function is concerned since mutation of some nucleus residues destroys compacity of the globule

MIR & nucleus

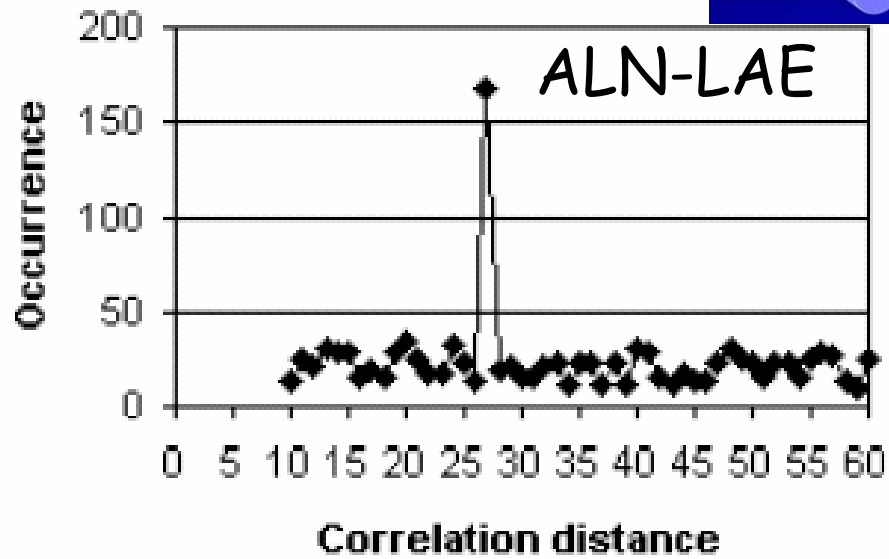


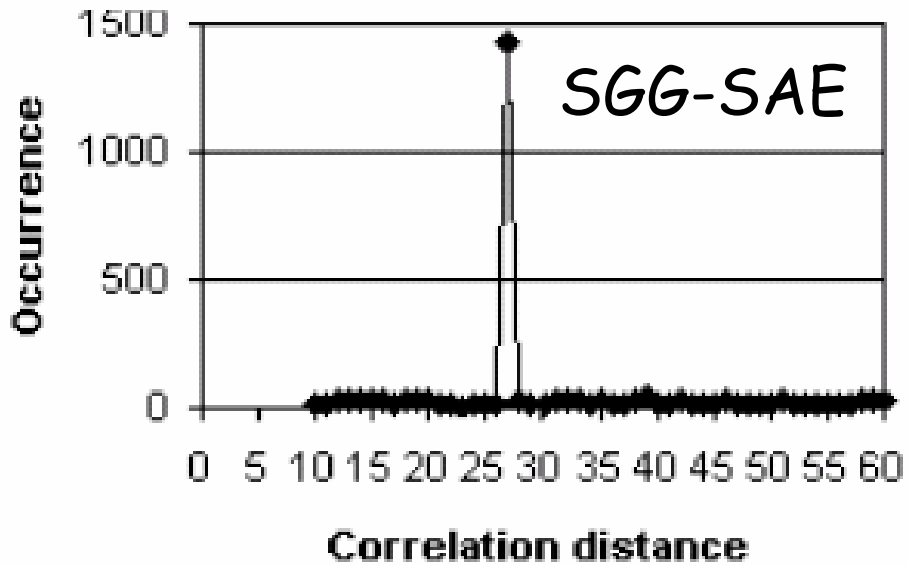
- Prediction of the folding nucleus : overprediction with the MIR?
- Some do not fall into the core
- How to avoid them?
  - Multiple prediction with several distantly related sequences
  - Other approaches

MIR & tripeptides

Different approaches to separate both classes of MIR: (Barrowed from Ed Trifonov & E. Aharonovsky, JBSD 2005 22:545)

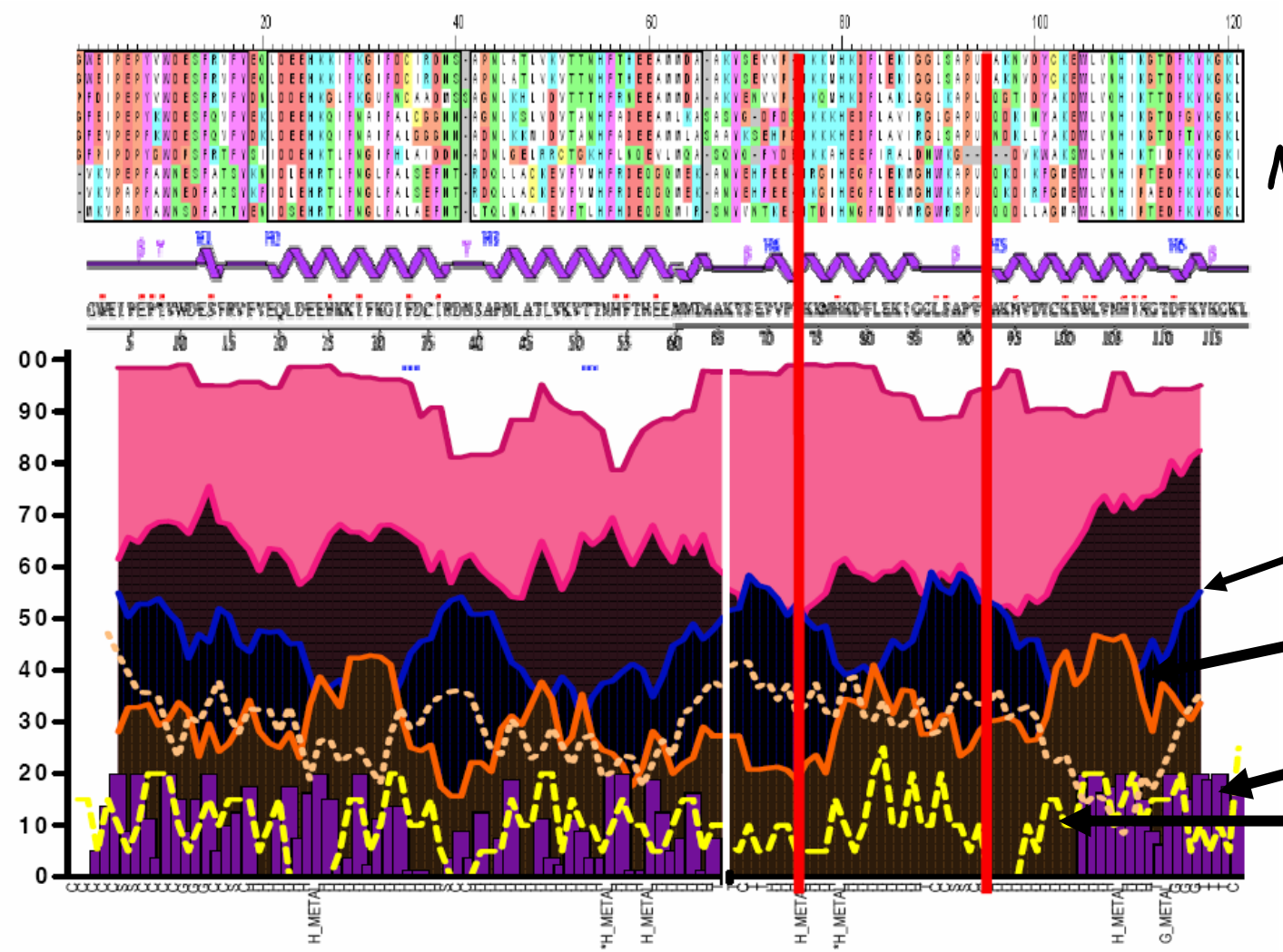
Some tripeptides are anchor points close to MIR





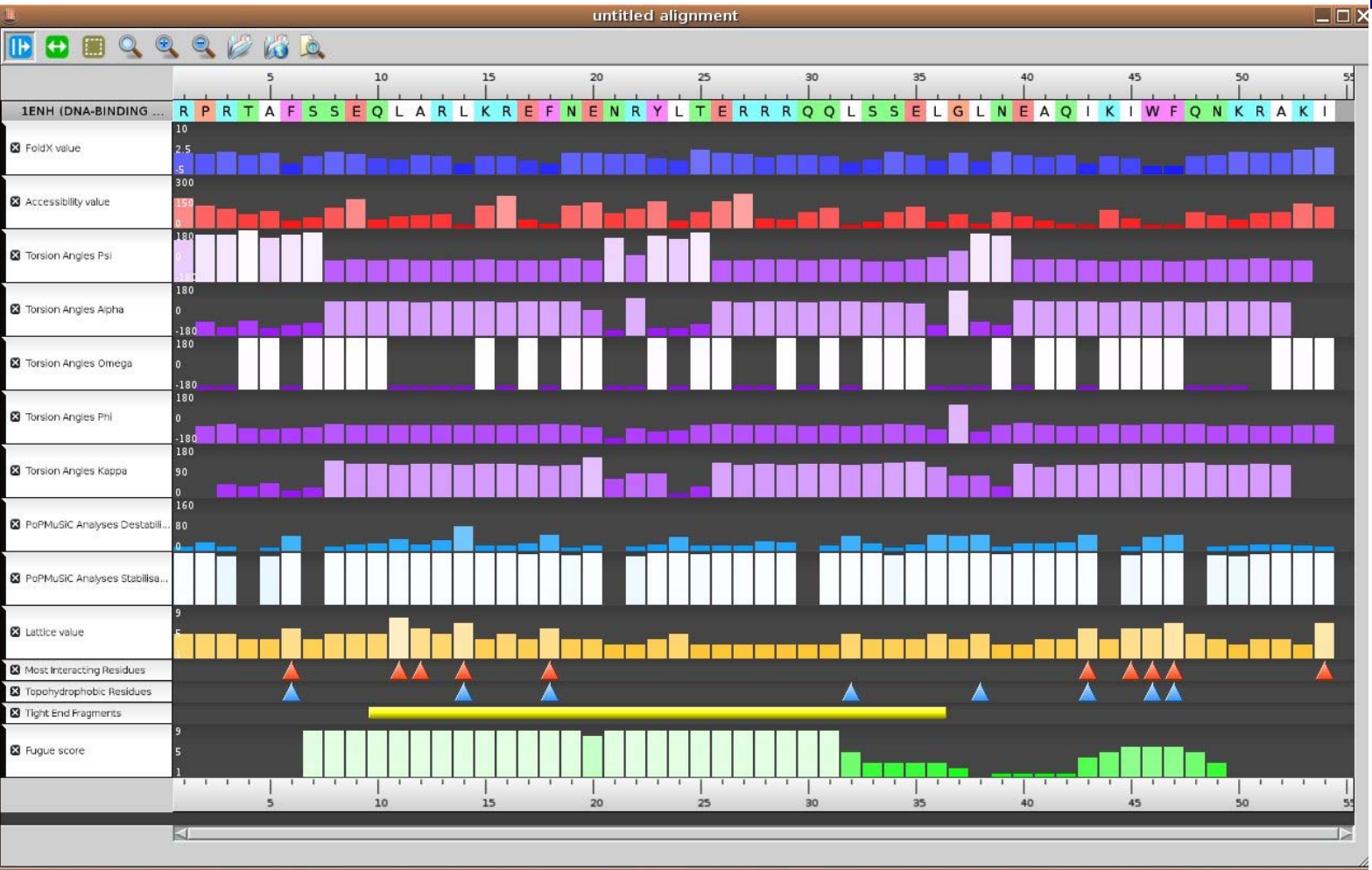
Protein Folding Fragments

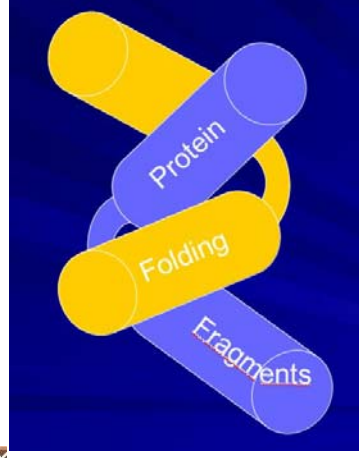
- MIR compared to foldons (M. Rooman), prints (T. Attwood… (this picture is a courtesy of M. Corpas)

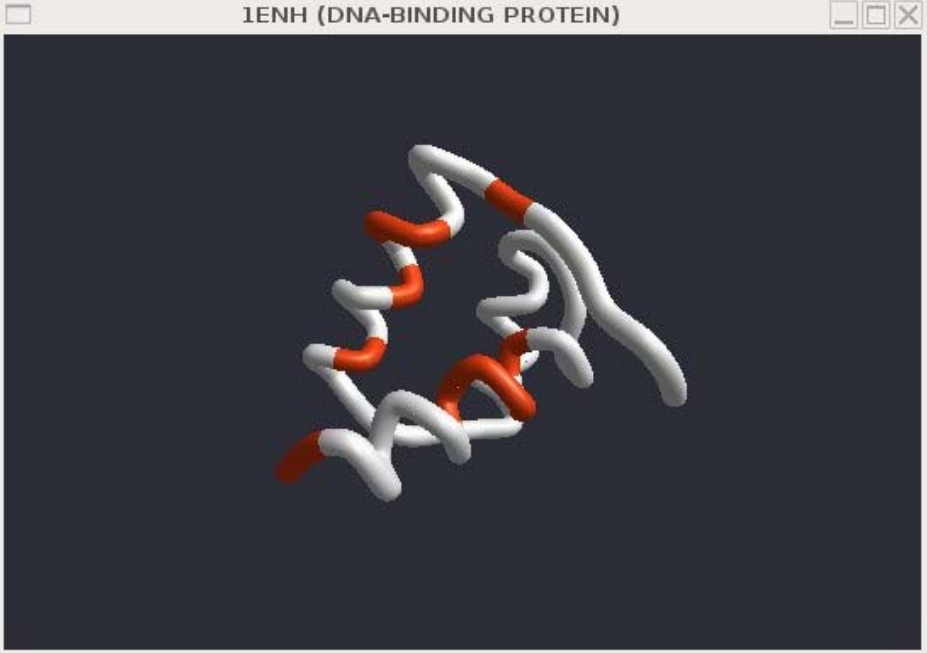


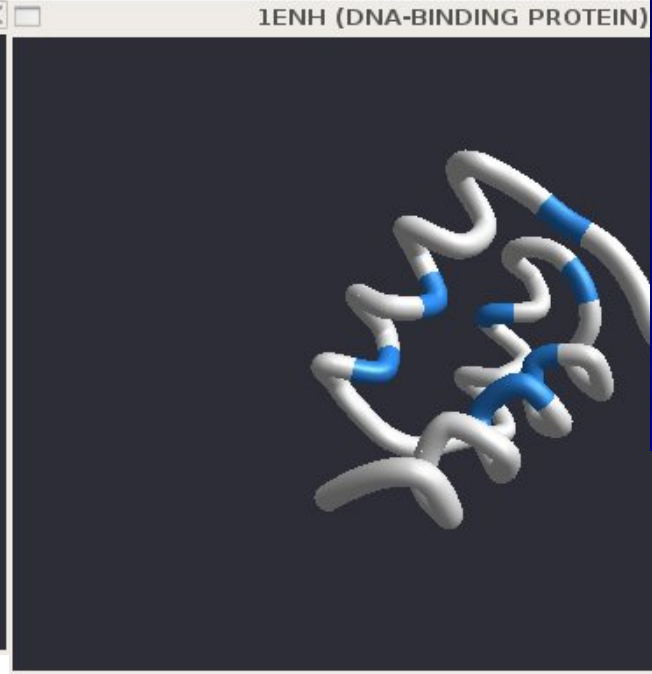
Cinema & Ambrosia

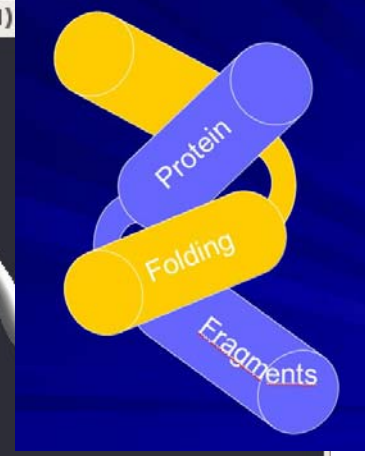
Xml structural database maintained in Manchester (Terri Attwood & Steve Pettifer):
Functional annotation in the future

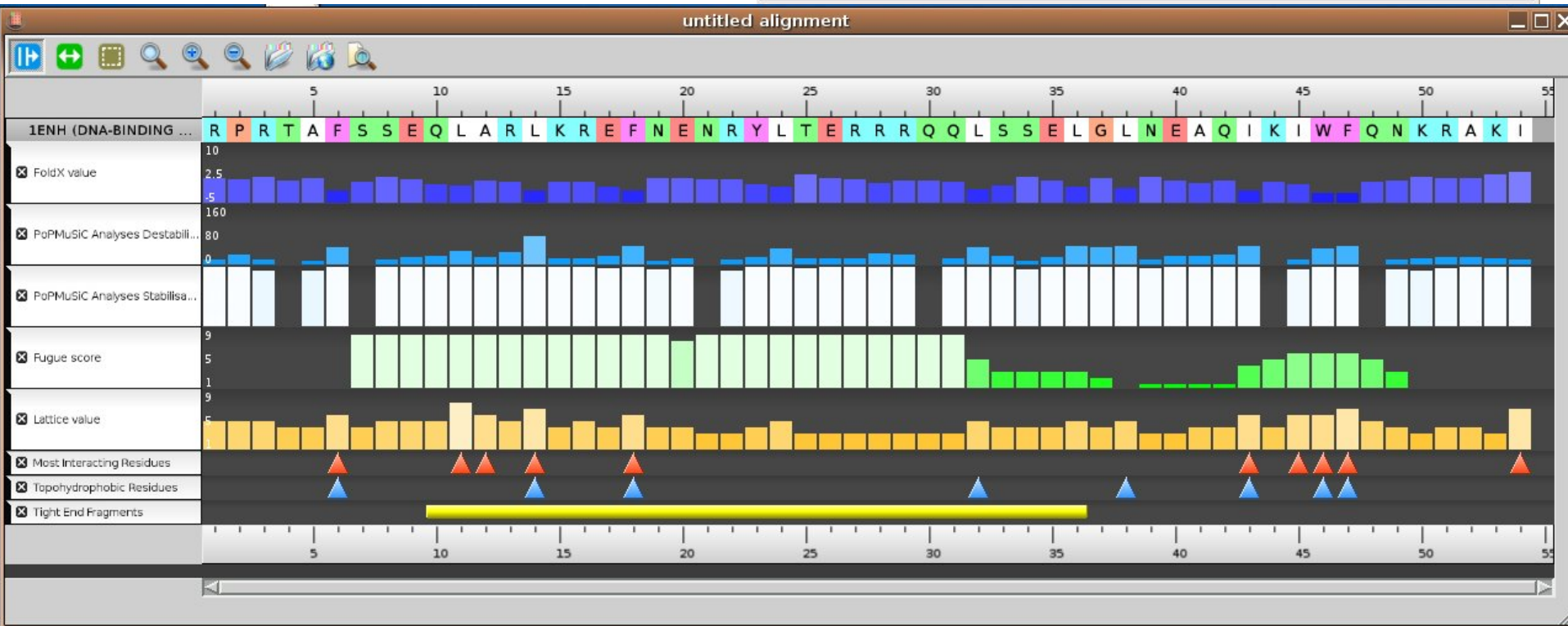




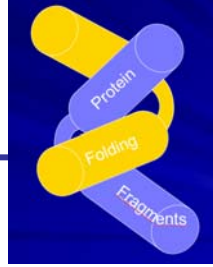

1ENH (DNA-BINDING PROTEIN)


1ENH (DNA-BINDING PROTEIN)


Protein
Folding
Fragments


untitled alignment
1ENH (DNA-BINDING ...
R P R T A F S S E Q L A R L K R E F N E N R Y L T E R R R Q Q L S S E L G L N E A Q I K I W F Q N K R A K I
FoldX value
PoPMuSiC Analyses Destabili...
PoPMuSiC Analyses Stabilisa...
Fugue score
Lattice value
Most Interacting Residues
Topohydrophobic Residues
Tight End Fragments

Mutations



MIR calculations are sensible to point mutation

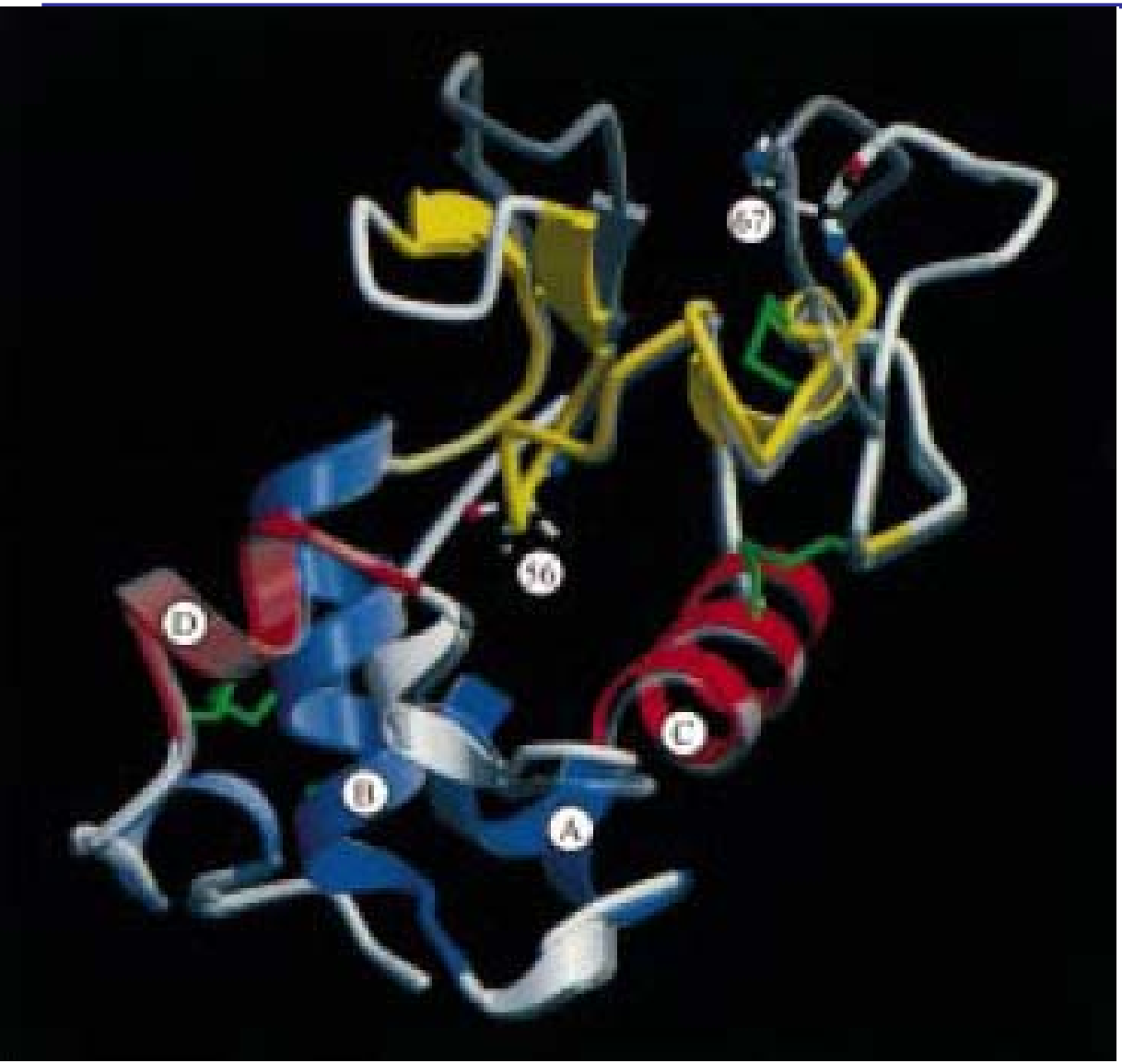
On a limited test set, mutations giving rise to amyoid behavior are located at MIR positions

Lysozyme: Two mutations give rise to amyloid

I56T

D67H

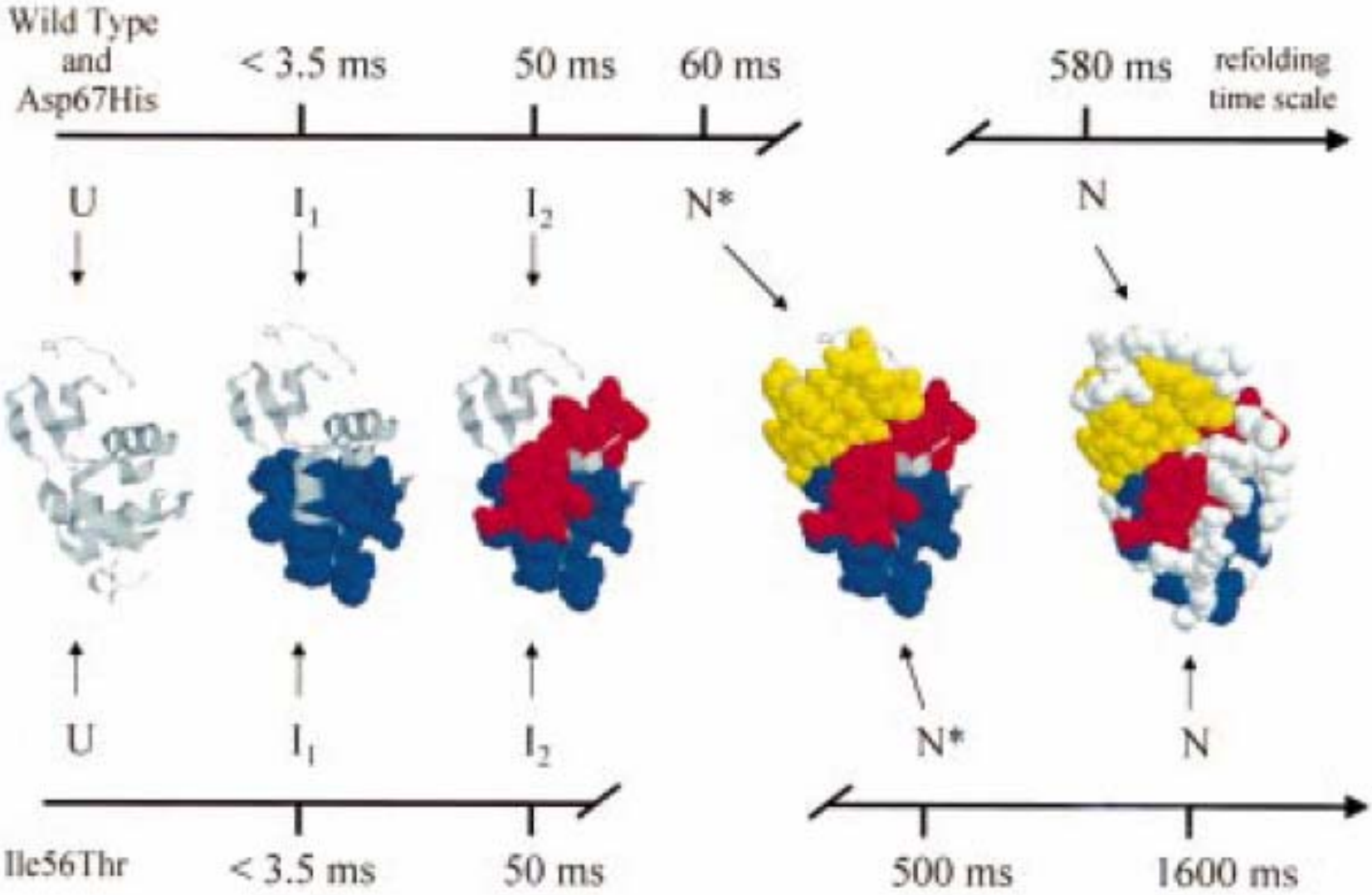
Lysozyme



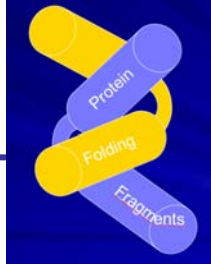
D67, in a loop, β domain

I56 is at the interface between both domains

Lysozyme folding rate



Lysozyme



Lactalbumin (1f6re) and lysozymes (1iiz, 1ix0, 1jwr)

	1f6rE	1ix0	1iizA	1jwrA
1f6rE	100.000	33.913	30.435	36.522
1ix0		100.000	33.913	97.391
1iizA			100.000	36.522
1jwrA				100.000

Strong MIR are conserved

Mutations : I56T and D67H. I56 is a MIR D67 is not

EQ**L**TKCE**VF**RE**L**K--DLKGYGGVS**L**PE**WV**CTT**F**HTSGYDTQAIVQNN--DSTEYG**LF**Q**I**NNK**I**WCKD
KR**F**TRCG**LV**NE**L**RKQGFDE--NL-MRD**WV**C**L**VENESARYTDKIANVNKNGSRDYG**LF**Q**I**NDKYWCSK
KVFERCE**L**ART**L**KRLGMDGYRGIS**L**AN**WMCL**AK**W**ESGYNTRATNYNAGDRSTD**Y**G**IF**QANSRYWCND
KV**F**ERCE**L**ART**L**KRLGMDGYRGIS**L**AN**W**MC**L**AK**W**ESGYNTRATNYNAGDRSTD**Y**G**IF**Q**I**NSRYWCND

L L MCL W Y □F I 56
F L L WMCL W IF I 67

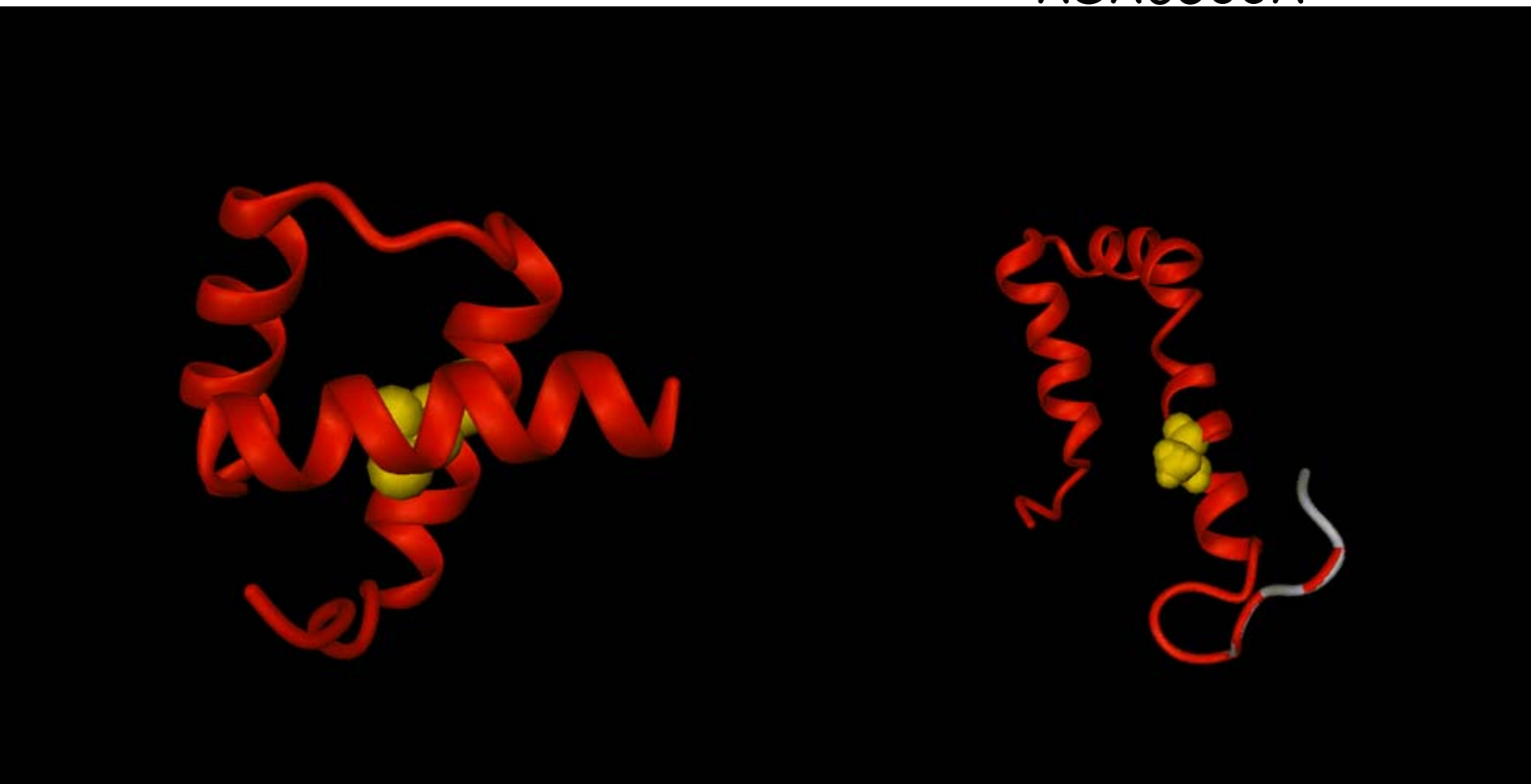
Effect of mutation on function

1enh
Homeodomain

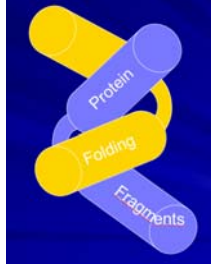
ASA=4000Å^2

1ztr
L16A

ASA6500Å^2



Amyloid fragments



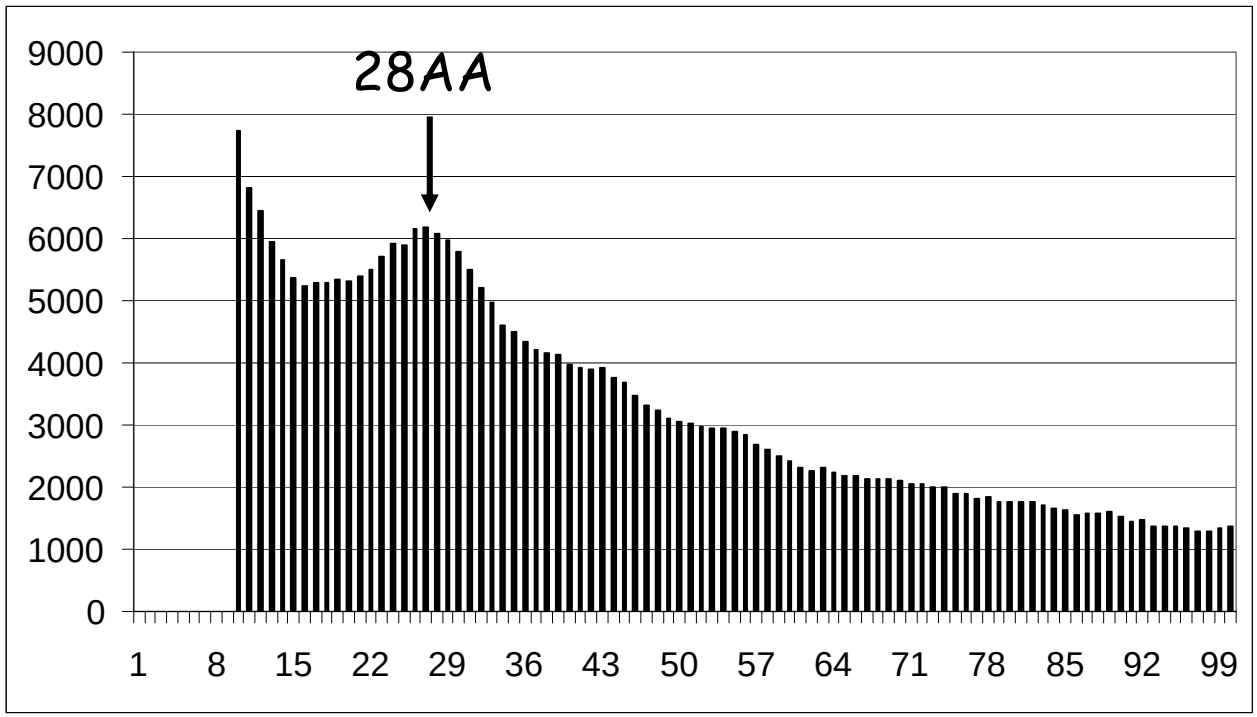
FUTURE :

Is there a correlation between fragments agregating ends and the presence of a MIR

MIR might delimitate fragments candidate for amyloid fibril formation

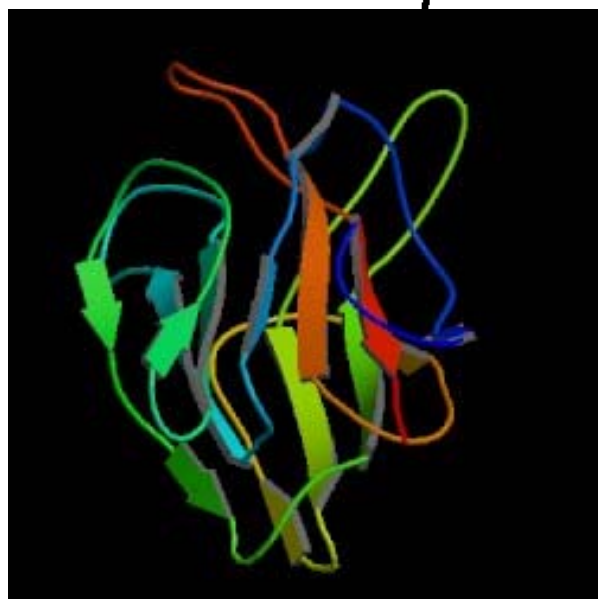
Protein Folding Fragments

Closed loop = protion of the backbone in between two contacts: Cα-Cα < 10 Å

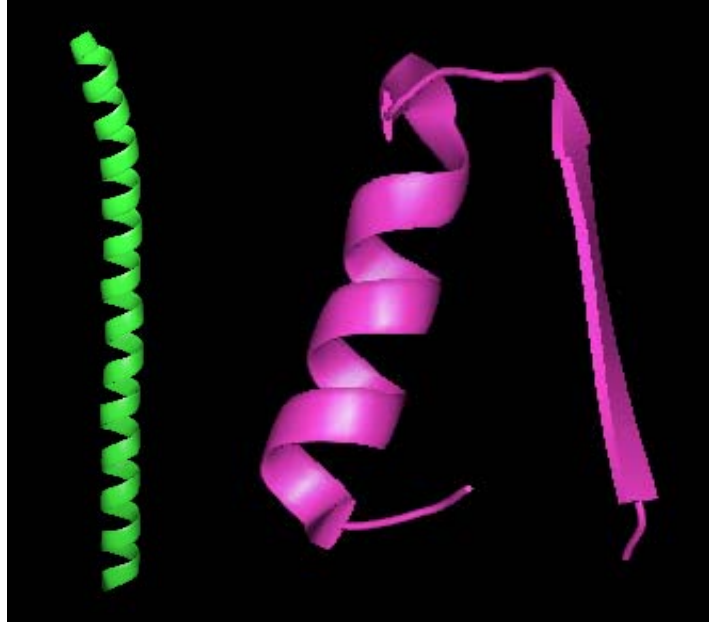


Sequence length between two neighbors

1VMO



TEF

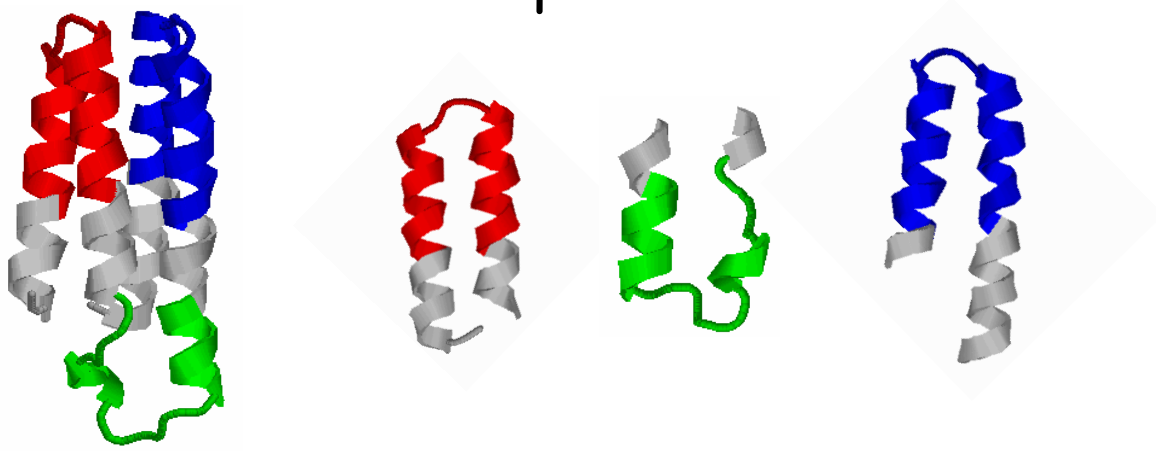


- Closed loops = 28 AA
  - ≈super SSR
  - mimimal length to fold

- Ends in the core
  - Topohydrophobic
  - Folding nucleus (Structuraly required)

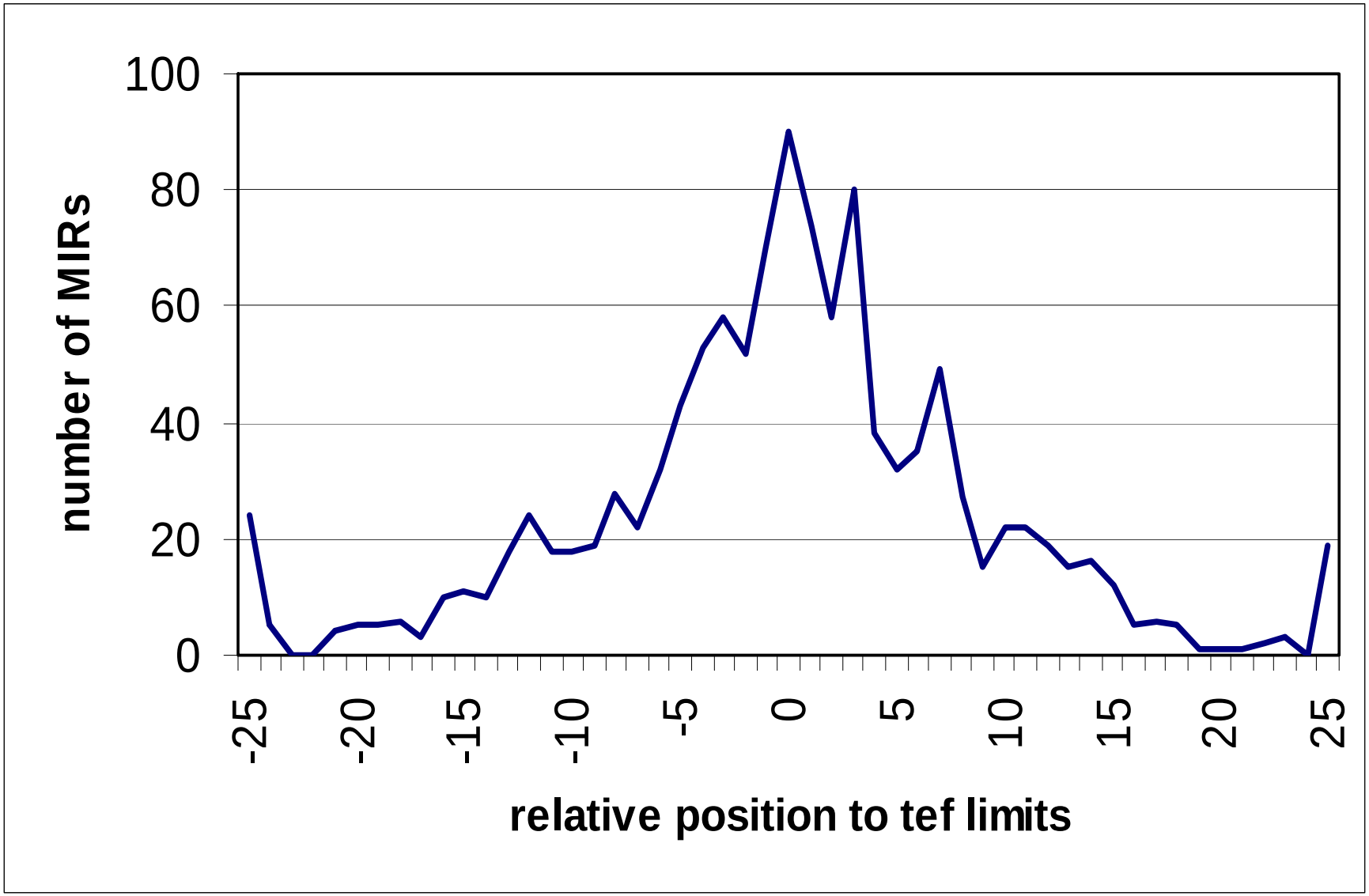
- Tightened End Fragments = Closed Loop + TH = TEF

Cytochrome b562

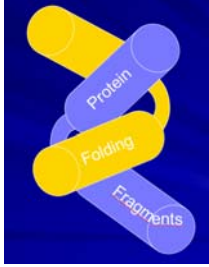


Comparison MIR & TEF

75% MIR in the TEF's ends are TH.



TEF & amyloid fragments



Prediction of MIR allows to predict TEF ends

Are TEF Autonomous Folding Units?

They must be compared to fragments involved in production of amyloid fibrils

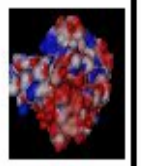
- http://bioserv.rpbs.jussieu.fr/

Séquences

Cet ensemble de liens et de services pour chercher, analyser, aligner des séquences protéiques

Structure

Cet ensemble de services permet la recherche et l'analyse des structures protéiques.

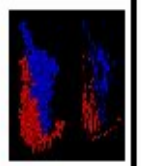


Modélisation des structures protéiques

Cet ensemble de services est relatif a la modelisation des structures proteiques

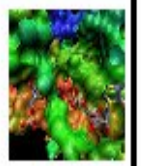
Interactions Protéiques

Cet ensemble de services est relatif aux interactions entre protéines



Petites Molécules

Cet ensemble de services est relatif aux interactions entre protéines et petites molécules (drogues)



Paris: Jean-Paul Mornon, Alain Soyer, Anne Lopes, David Perahia, Liliane Mouawad, Charles Robert

Athens: Elias Eliopoulos

Haifa: Edward Trifonov, Elik Aharonovsky

Heidelberg: Luis Serrano

Bruxelles: Marianne Rooman, Jean-Marc Kwasigroch, Dimitri Gillis

Manchester: Terry Attwood, Manuel Corpas, Steve Pettifer, Dave Thorne, James Sinnott