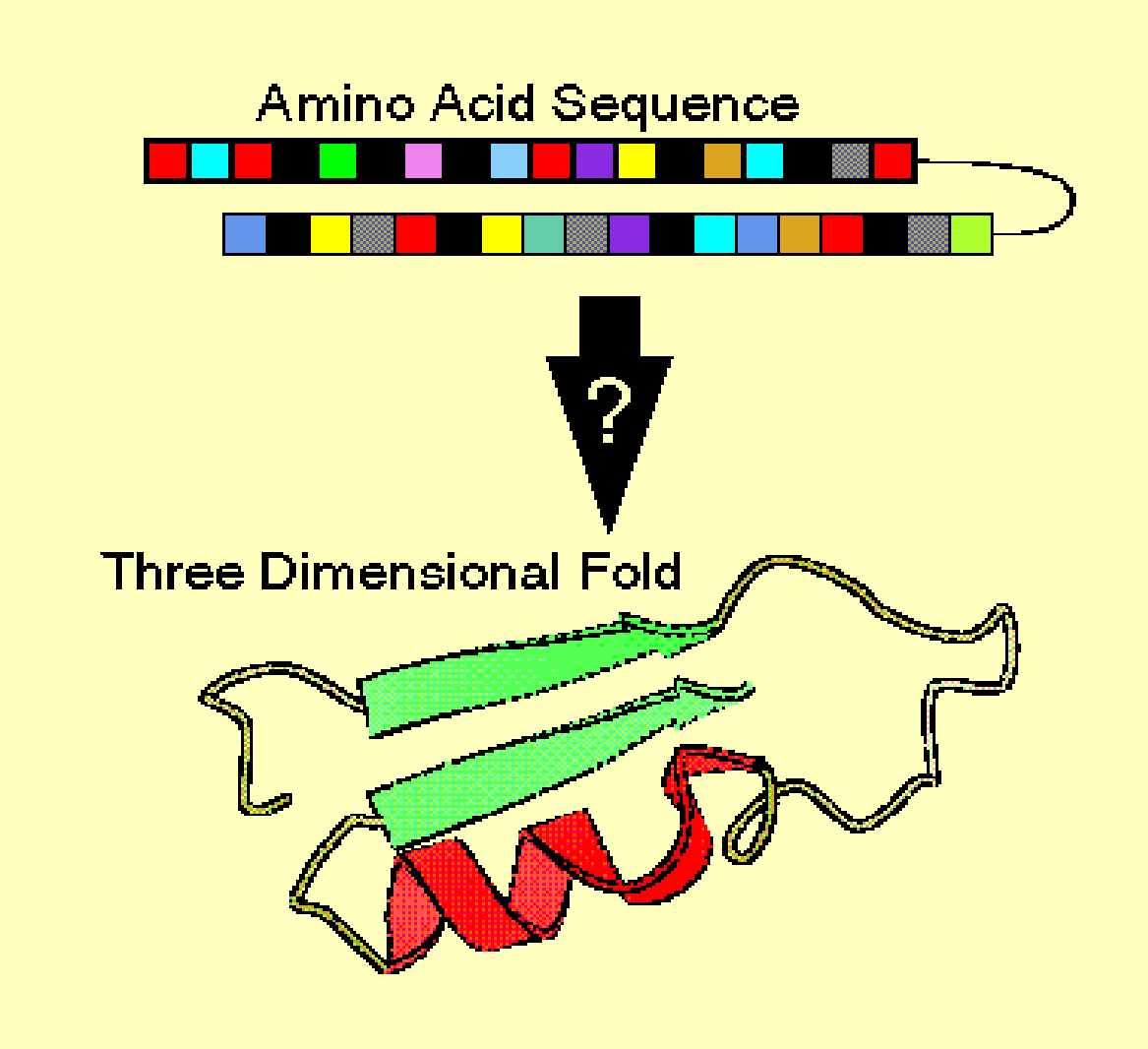
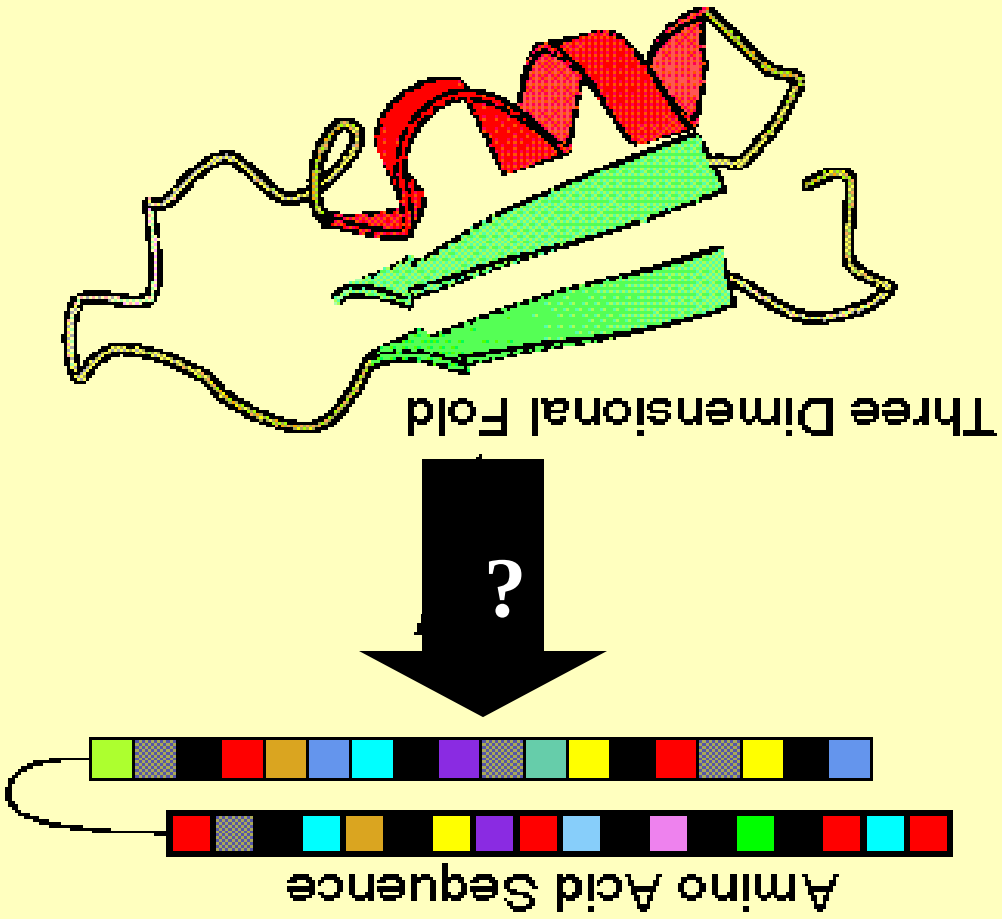
The Protein Folding Problem



- **The native state is uniquely determined by the sequence**
- **The native state is thermodynamically stable and reachable from different starting conditions.**
- **Only few sequences are proteins**
- **Only few conformations are native states**
- **The folding time is very rapid (0.01-100 sec)**

The Inverse Folding Problem



- Given a desired structure to find an aminoacid sequence that folds on it
- Protein functionality is controlled by its native conformation
- Powerful DNA-recombination techniques allow to modify the sequence of amino-acids
- To solve the problem would allow to design new proteins with new functionality (drug design)

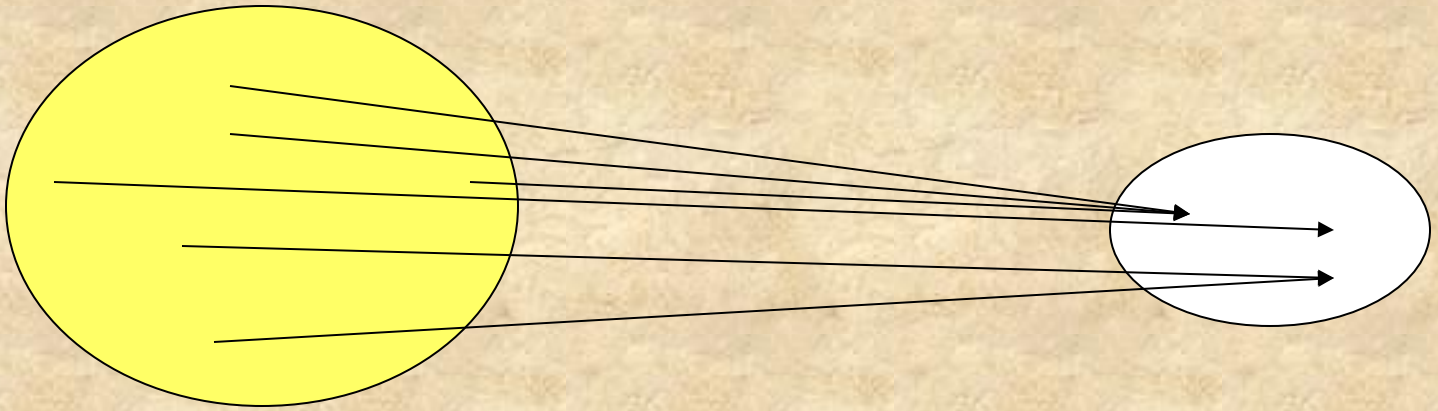
Protein folding is complex

- 20 type of amino acids with distinct side chains
- huge number of degrees of freedom
- polymer chain constraint (length 50-1000)
- steric constraints (excluded volume)
- role of the aqueous solvent
- quantum chemistry

Very interesting fact

~ 100.000 sequences exist and only
~ 1.000 folds

→ mapping many to one



Protein sequences have undergone evolution but folds have not…. they seem immutable

Compactness-Hydrophobicity

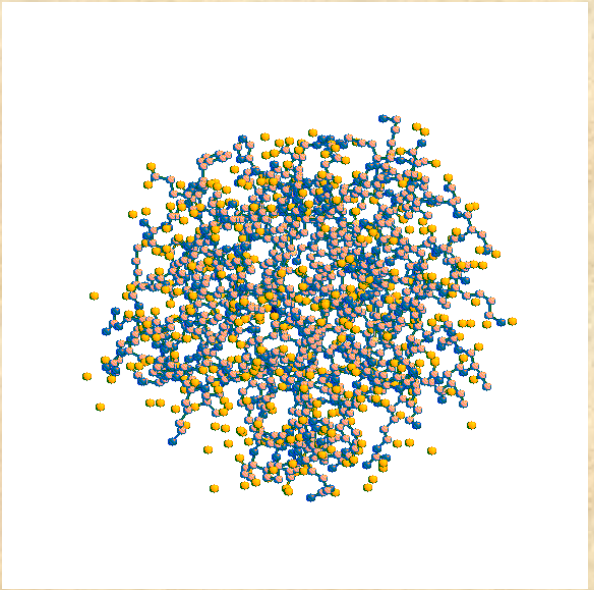
COARSE-GRAINED MODELS

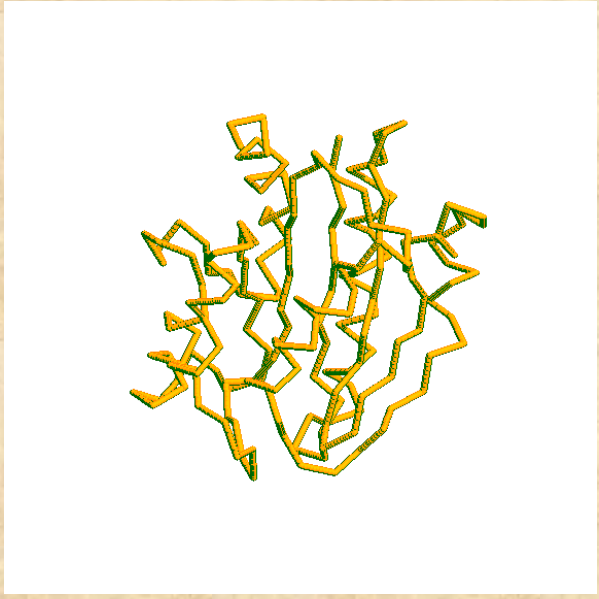
Coarse grained representation

Too many details can obscure ,
rather than illuminate physical principles

To consider just few degrees of freedom for each amino-acid (e.g.C_α)

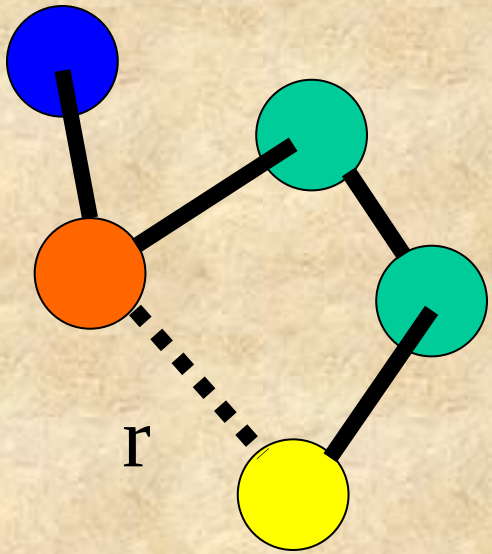
Effective interactions between these “amino-acids” are postulated to arise on integrating out the other degrees of freedom

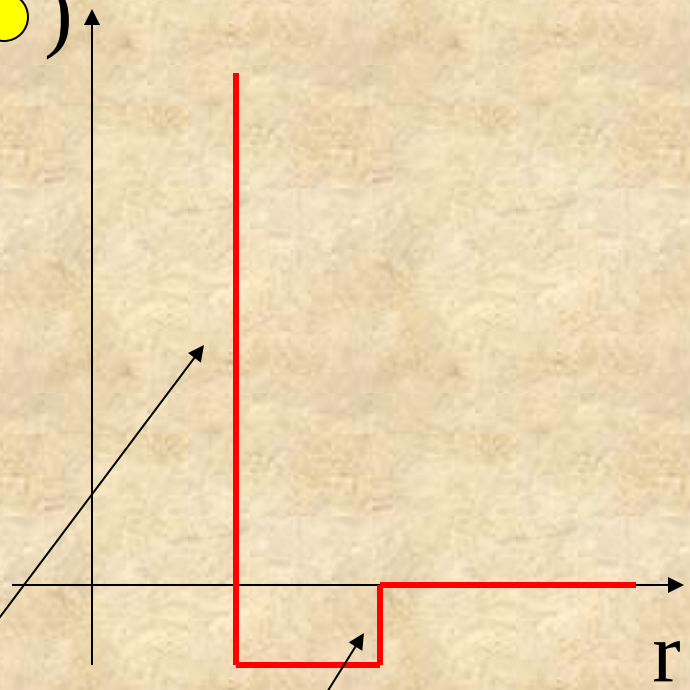




How to get it???

B(,)





Hard core

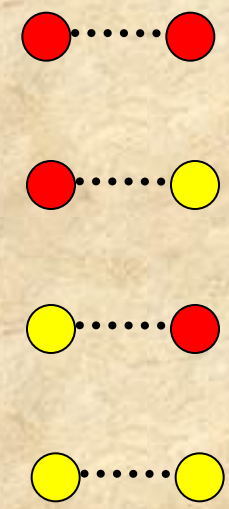
Threshold

HP Model

Only two kinds of aminoacids: H Hydrophobic

P Polar

$$\frac{E}{k_B T}$$

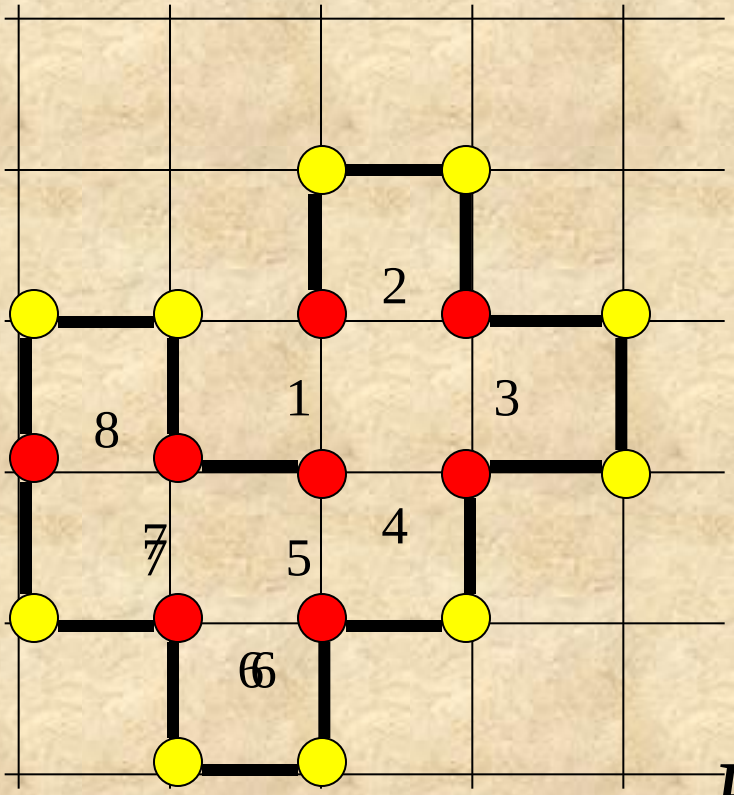


-1

0

0

0



$E = -8$

How to extract the potentials

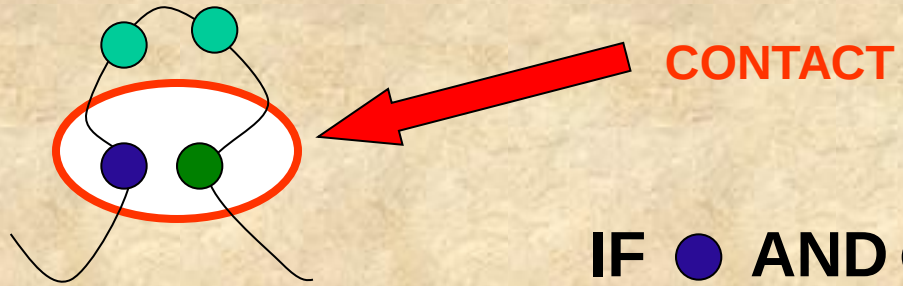
- Broadly speaking, scoring functions can be divided into the following classes:
  - Physical effective energy functions
    - Derived from a fundamental analysis of the forces between the particles ,based on terms from molecular mechanics forcefields (GoldScore, DOCK, AutoDock)
    - Greater computational cost
  - Knowledge-based potentials
    - Derived by a statistical analysis of known protein structures (PMF, DrugScore, ASP)
    - They are more robust and easier to compute

Knowledge-based potentials

Features to which an energy can be assigned

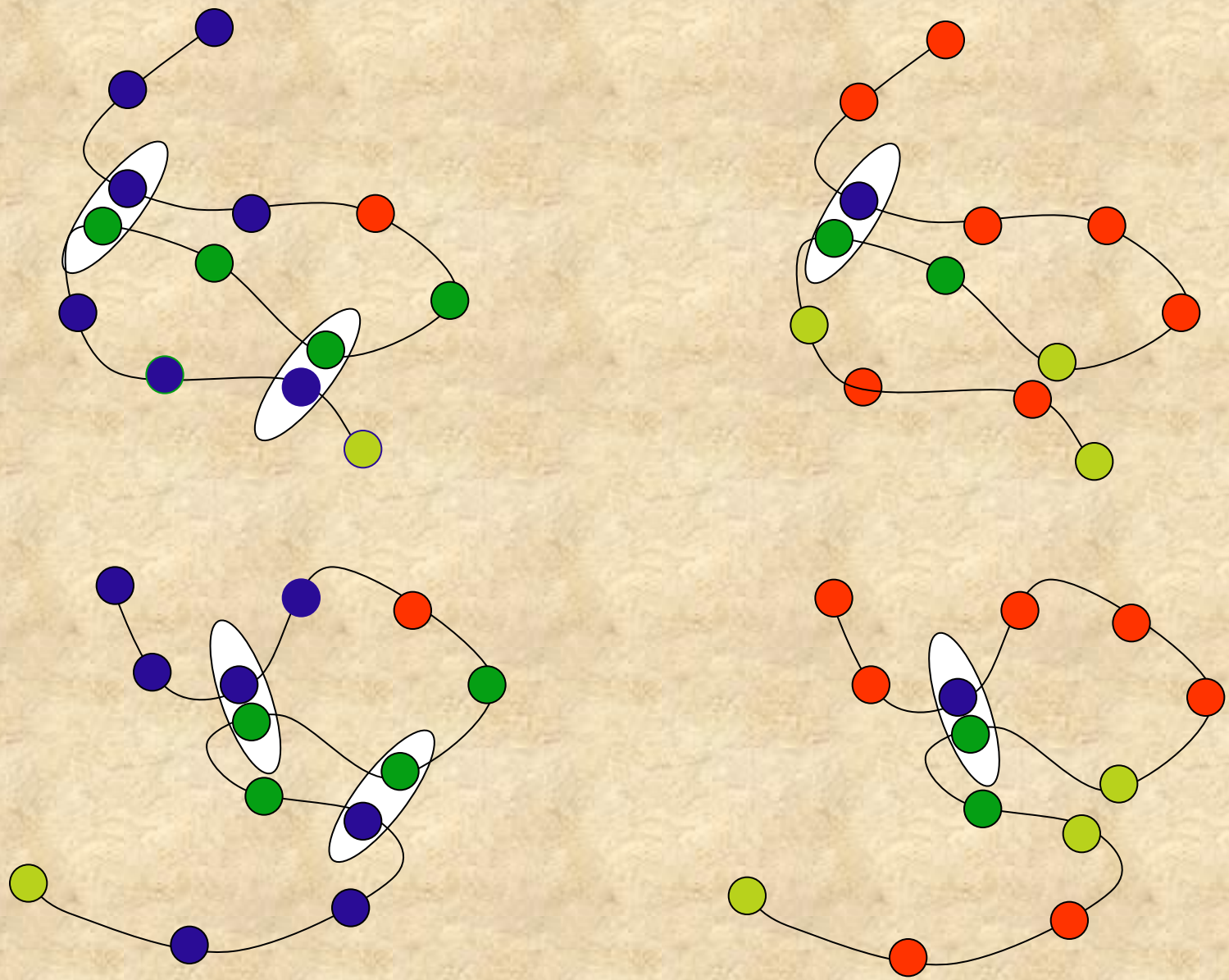
Parameters are then compiled from probabilities observed in a database of experimentally determined proteins

SIMPLEST EXAMPLE: PRESENCE OF A CONTACT BETWEEN TWO AMINOACIDS



IF ● AND ● LIKE EACH OTHER, THIS CONFORMATION IS FAVOURED

IF ● AND ● LIKE EACH OTHER THEY SHOULD BE FOUND VERY OFTEN IN CONTACT



$$score = -kT \log\left(\frac{p(r)_{observed}}{p(r)_{reference}}\right) \quad \begin{matrix} < 0 \\ > 0 \end{matrix}$$

Second Key point: to select the right features

And how to define them!!!

1) **Presence of a contact between two aminoacids**
2) **Solvent accesibility**
3) **Torsional angles**
4) **Presence of secondary structures elements**
5) **….**
6) **….**

Potenziali knowledge-based

Approccio statistico probabilistico

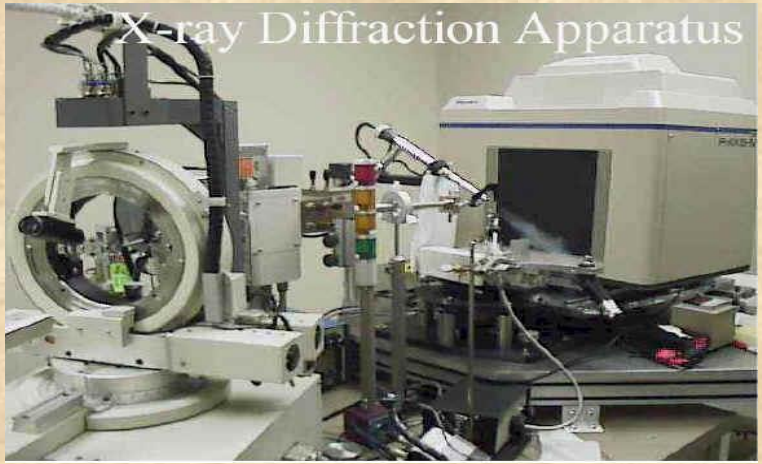
$$E_{ab} = -\log\left(\frac{\dfrac{n_{ab}^{c}}{n_{ab}}}{\dfrac{\sum_{ab} n_{ab}^{c}}{\sum_{ab} n_{ab}}}\right)$$

$$n_{ab}^{c} = \#\ \text{of contacts between}\ a\ \text{and}\ b$$

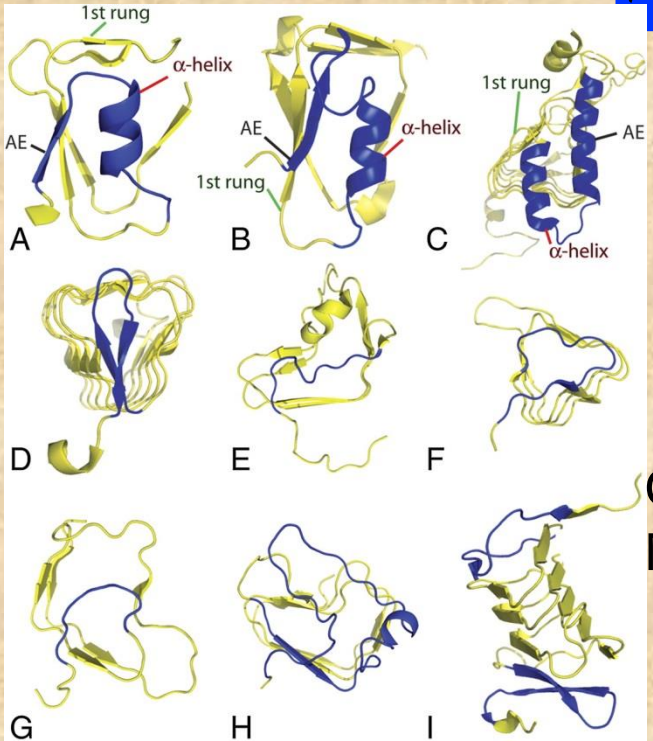
$$n_{ab} = \#\ \text{of pairs}\ a\ \text{and}\ b$$

Critical Assessment of *Protein Structure Prediction (CASP)*

PROTEIN SEQUENCE



EXPERIMENTAL TARGET



COMPUTATIONAL MODELS

PROTEIN AGGREGATION

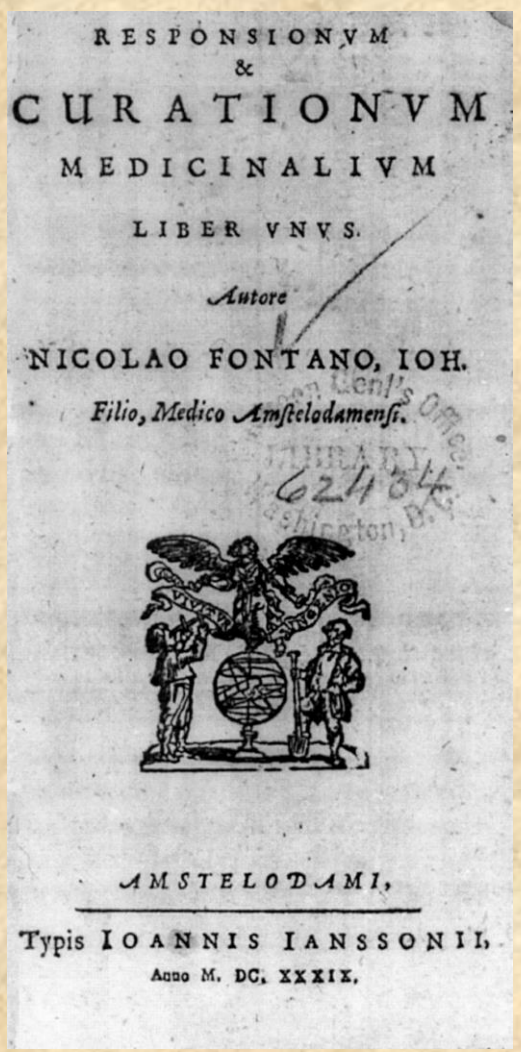
Protein-misfolding diseases

A broad range of human diseases arises from the failure of specific peptide or protein to adopt, or to remain in, its native functional conformational state

Disease	Associated proteins	Affected tissues
Amyloidosis—systemic		
Primary systemic amyloidosis	Ig light chain	Most tissues
Ig heavy-chain-associated amyloidosis	Ig heavy chain	Most tissues
Secondary (reactive) systemic amyloidosis	SAA	Most tissues
Senile systemic amyloidosis	Transthyretin	Microvasculature
Hemodialysis-related amyloidosis	β_2-Microglobulin	Osteoarticular tissues
Hereditary systemic ApoAI amyloidosis	ApoA-I	Liver, kidney, heart
Hereditary systemic ApoAII amyloidosis	ApoA-II	Kidney, heart
Finnish hereditary amyloidosis	Gelsolin	Most tissues
Hereditary lysozyme amyloidosis	Lysozyme	Kidney, liver
Hereditary cystatin C amyloid angiopathy	Cystatin C	Most tissues
Amyloidosis—localized		
Injection-localized amyloidosis	Insulin	Skin, muscles
Hereditary renal amyloidosis	Fibrinogen	Kidney
Senile seminal vesicle amyloid	Lactoferrin, seminogelin	Seminal vesicles
Familial subepithelial corneal amyloidosis	Lactoferrin	Cornea
Cataract	Crystallin	Eye
Medullary thyroid carcinoma	Calcitonin	Thyroid tissues
Neurodegenerative diseases		
Alzheimer's disease	Amyloid-β, tau	Brain
Parkinson's disease	α-Synuclein	Brain
Lewy-body dementia	α-Synuclein	Brain
Huntington's disease	Huntington	Brain
Spongiform encephalopathies	Prion	Brain, peripheral nervous system
Hereditary cerebral hemorrhage with amyloidosis	Cystatin C	Cerebral vasculature
Amyotrophic lateral sclerosis	Superoxide dismutase 1	Brain
Familial British dementia	Abri	Brain
Familial Danish dementia	ADan, amyloid-β	Brain
Familial amyloidotic polyneuropathy	Transthyretin	Peripheral nervous system
Frontotemporal dementias	Tau	Brain
Other diseases		
Diabetes mellitus	IAPP, amylin	Pancreas (islet)
Atherosclerosis	Modified LDL	Arteries
Sickle cell anemia	Hemoglobin	Erythrocytes

Amyloid from amylum (latin) = starch

Matthias Schleiden, German botanist, 1838



RESPONSIONVM
&
CURATIONVM
MEDICINALIVM
LIBER VNVS.

Autore

NICOLAO FONTANO, IOH.

Filio, Medico Amstelodamensi.

AMSTELODAMI,

Typis IOANNIS IANSSONII,
Anno M. DC. XXXIX.

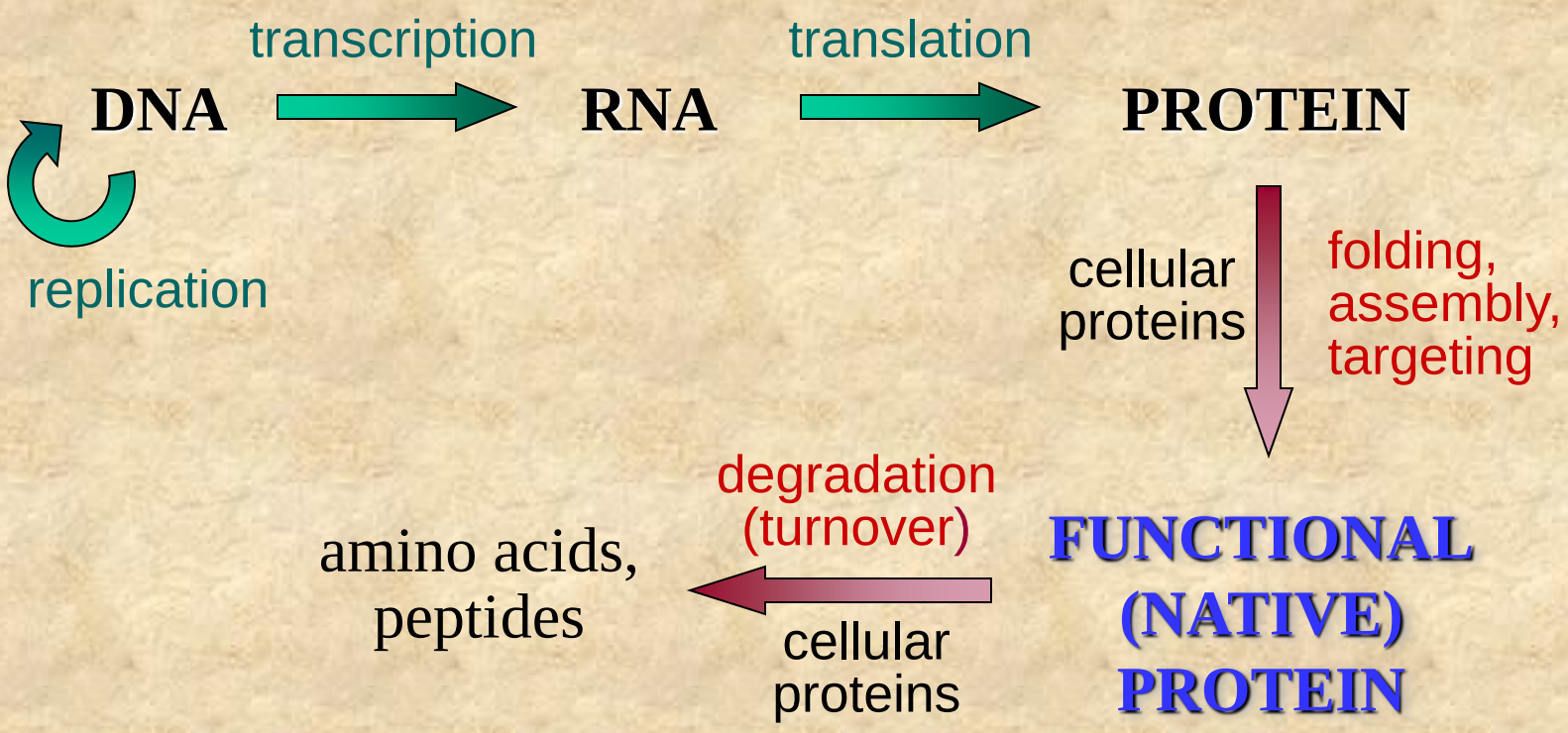
Samuel Wilks (1824-1911)

52 year old man with lardaceous viscera.

No starch -> albuminoid nature

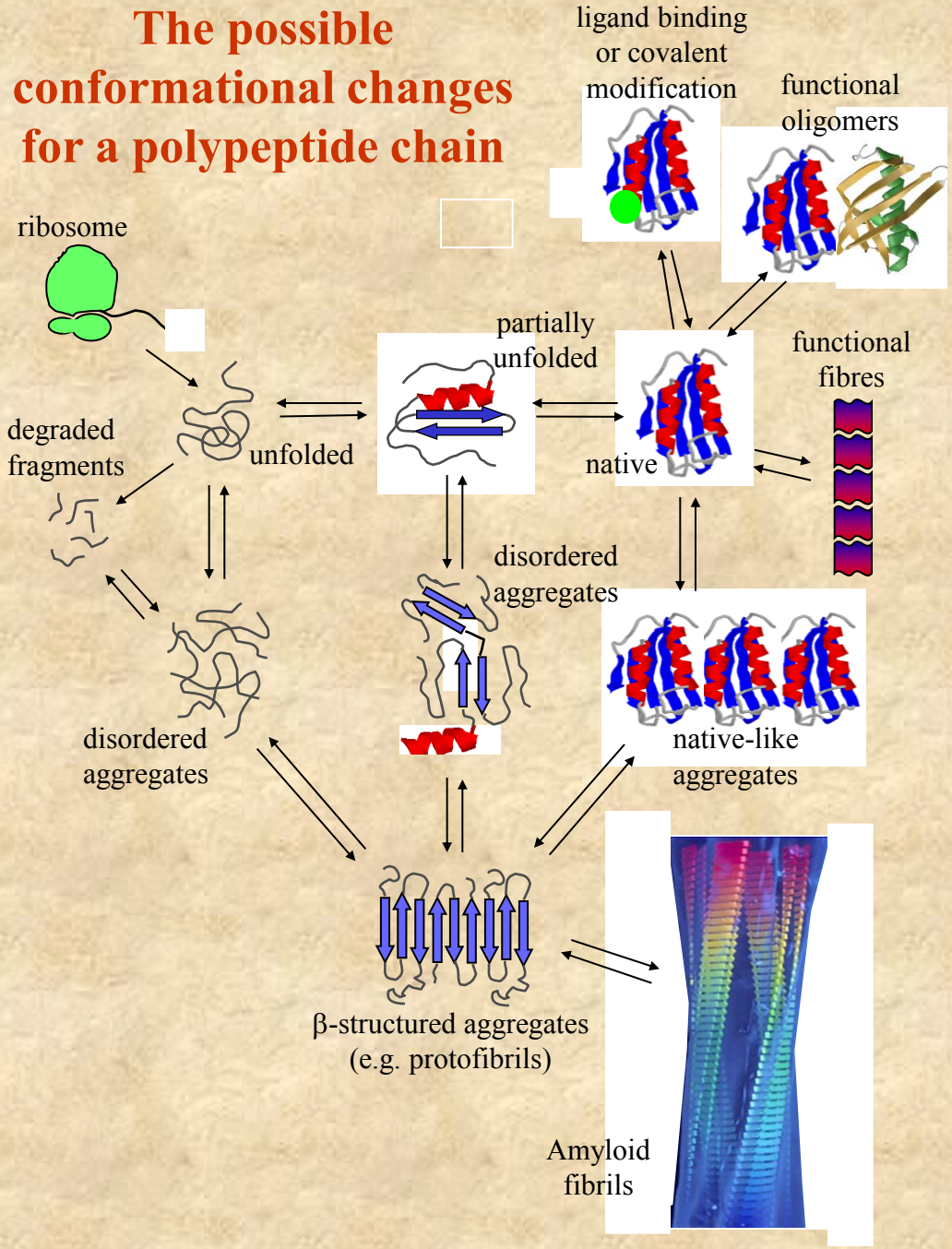
Maybe the first observed case:large spleen filled with white stones (1639)

Central dogma of biology



PROTEIN QUALITY CONTROL SYSTEM

The possible conformational changes for a polypeptide chain

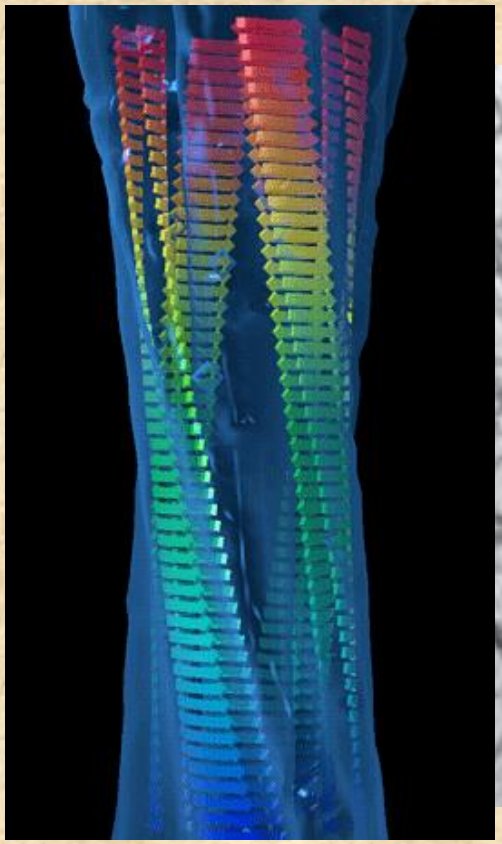


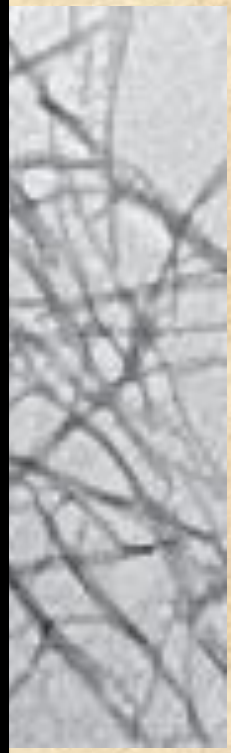
Chiti & Dobson, 2006
Ann. Rev. Biochem. 75, 333-66

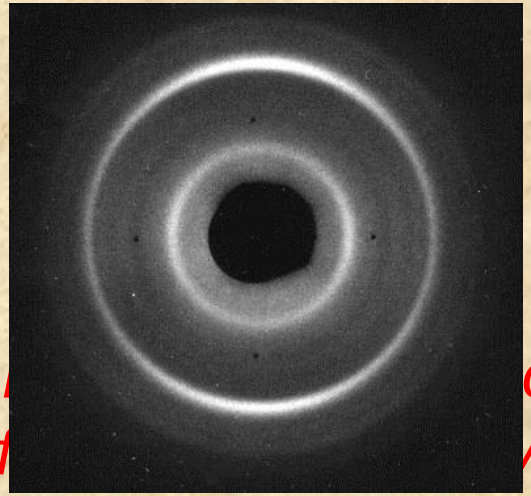
Amyloid fibrils

Insoluble fibrillar aggregates

Highly organized macrostructure a few nm in diameter



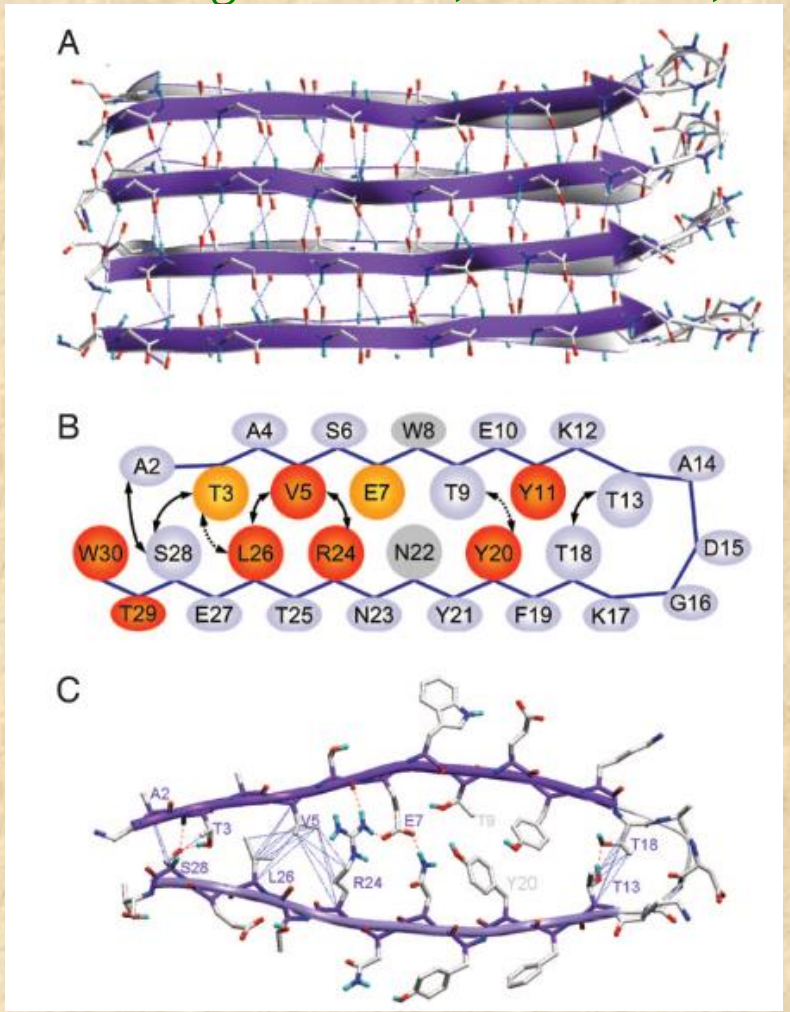




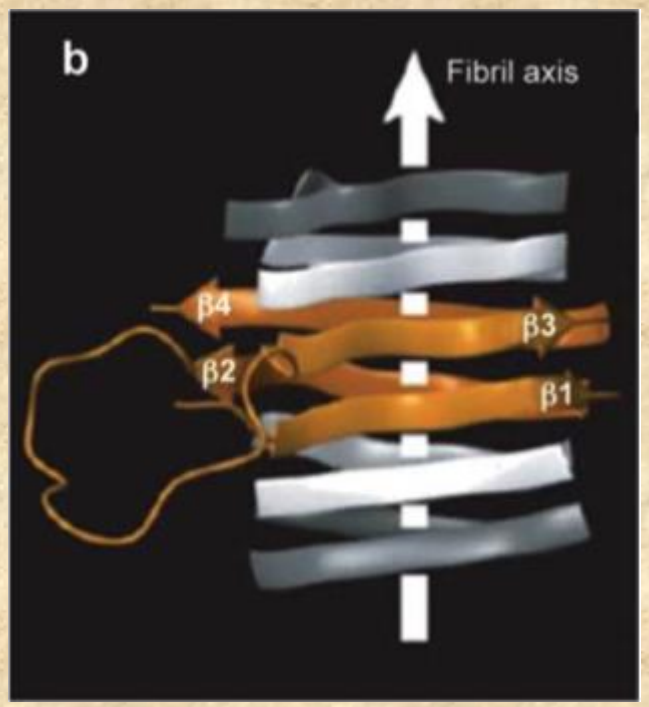
A ope image
of vitro

Diffraction pattern: signature of cross β structure with β-strands orthogonal to the fibril axis

Ferguson et al., PNAS **103**, 16248 (2006)

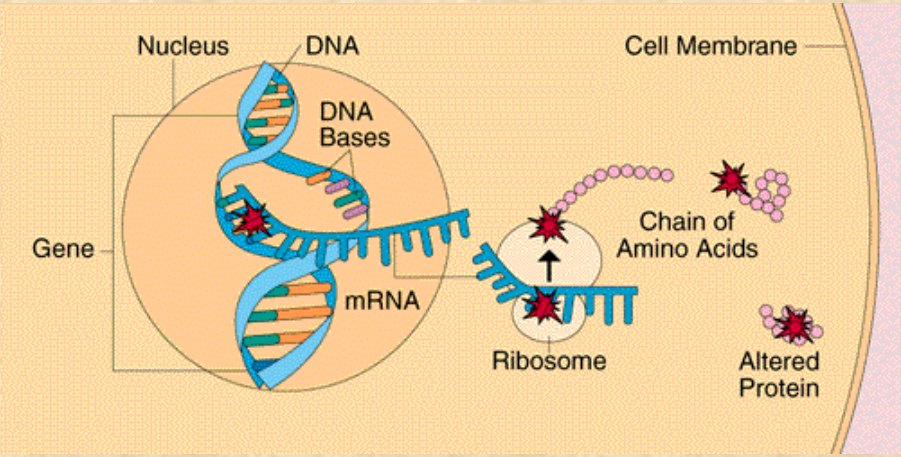


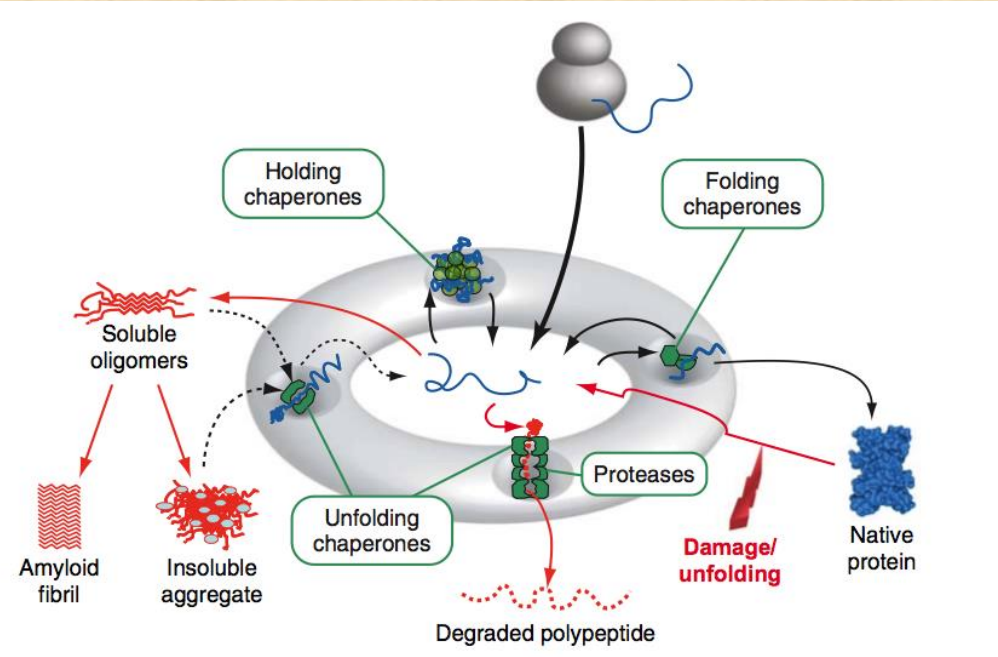
Solid state NMR atomic level structure of amiloyd fibrils of WW domain in human CA150 (a transcriptional activator involved in Huntington's disease)



Protein aggregation can occur due to a variety of causes:

•Individuals may have mutations that encode for proteins that are particularly sensitive to misfolding and aggregation.





•Troubles in Protein Quality Control System:

As many of the diseases increase in frequency with age, it seems that cells lose the ability to clear misfolded proteins and aggregates over time.

• Infection

Open Problems:

- **STRUCTURE AT ATOMIC DETAIL**
- **PROCESS OF FORMATION**

very little is yet known about the structure of the amyloid protofibrils and unstructured aggregates that precede their formation

- **TOXICITY**

the precise origin of pathogenic nature of the amyloid deposits and their precursors remains elusive in each pathological condition associated with the formation of this species

THE RATIONAL DESIGN OF SUCCESSFUL THERAPEUTIC STRATEGIES REQUIRES FURTHER CHARACTERIZATION OF THE PROCESS OF AMYLOID FORMATION

PHYSICAL APPROACHES: UNIVERSALITY

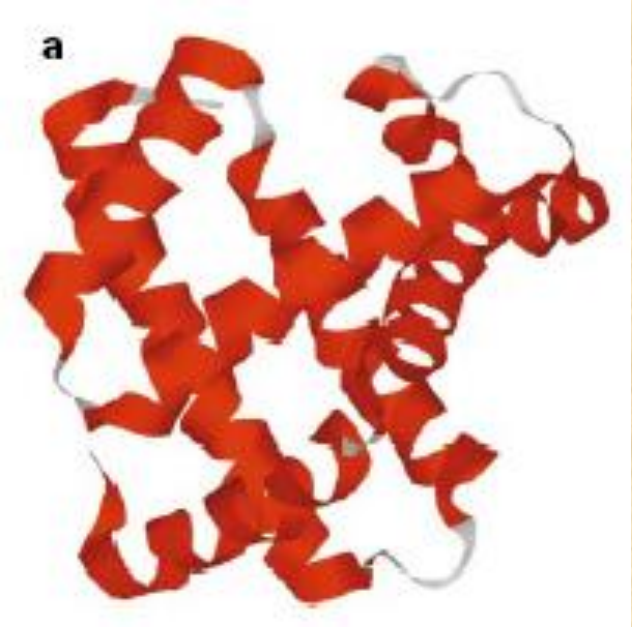
AMYLOID FORMATION IS NOT LIMITED TO THE FEW PROTEINS ASSOCIATED WITH DISEASES….

Fandrich, Fletcher, and Dobson, *Nature* **410**, 165-166 (2001)

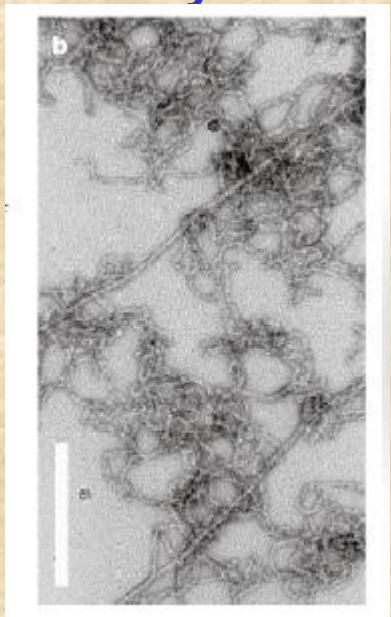

brief communications

Amyloid fibrils from muscle myoglobin

Even an ordinary globular protein can assume a rogue guise if conditions are right.

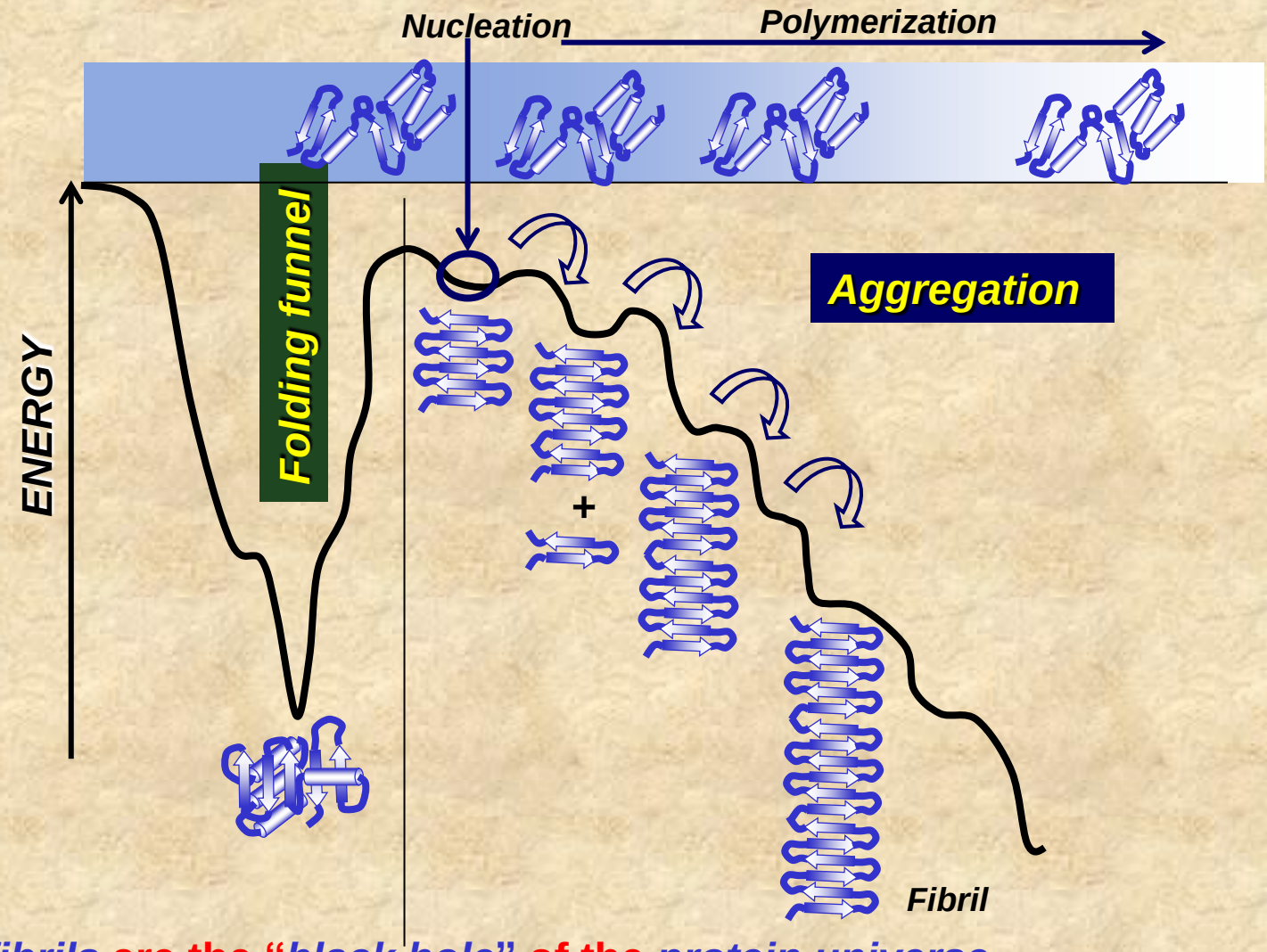


Myglobin is a compact and highly soluble protein without any native state properties to suggest that it has a predisposition to form amyloid fibrils.



pH 9.0 at T=65 C

PHYSICAL APPROACHES: ENERGY LANDSCAPE



Amyloid fibrils are the "*black hole*" of the *protein universe.*

The amyloid structure is the most stable in the free energy landscape of a protein conformation, even more stable than the native state…and it has the ability to attract new protein molecules

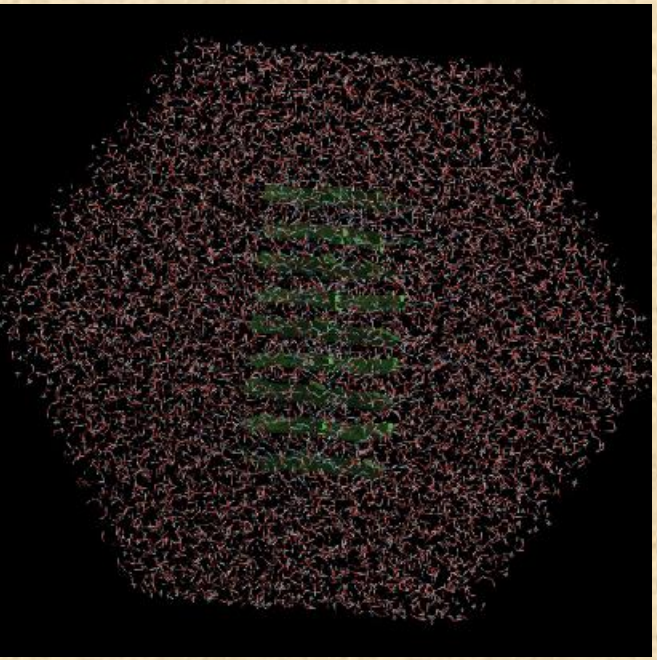
MULTISCALE APPROACH

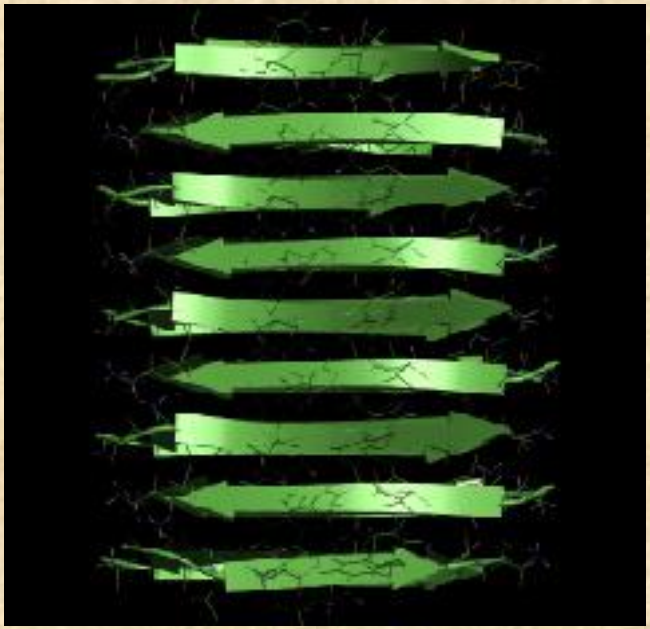
All-atom: Molecular Dynamics (MD) With EXPLICIT Solvent

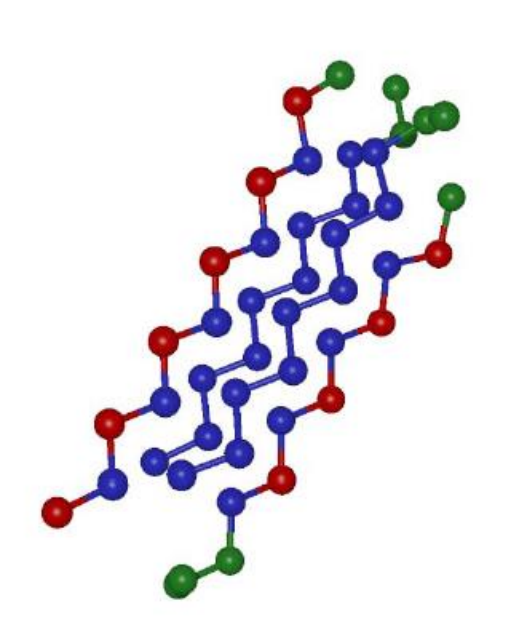
All-atom: Simulations with IMPLICITsolvent

Off-lattice minimalist models

MONTE CARLO SIMULATIONS







COARSE GRAINING

Energy function for aggregation propensity

SPECIFIC PAIRING OF TWO SEQUENCE STRETCHES OF THE SAME LENGTH

CHAIN 1

CHAIN 2

IS THERE A PART OF **CHAIN 1** WHICH PREFER TO FORM BETA-STRAND WITH ANOTHER PART OF **CHAIN 2**?

Do they like to form hydrogen bonds?

Energy function for aggregation propensity

Propensity of two residue types to be found paired in neighbouring strands within beta-sheets in globular proteins. (Samudrala and Moult, 1998)

$$E_{ab}^{p} = -\log\left(\frac{\dfrac{n_{ab}^{p}}{n_{ab}}}{\dfrac{\sum_{ab} n_{ab}^{p}}{\sum_{ab} n_{ab}}}\right) \qquad E_{ab}^{a} = -\log\left(\frac{\dfrac{n_{ab}^{a}}{n_{ab}}}{\dfrac{\sum_{ab} n_{ab}^{a}}{\sum_{ab} n_{ab}}}\right)$$

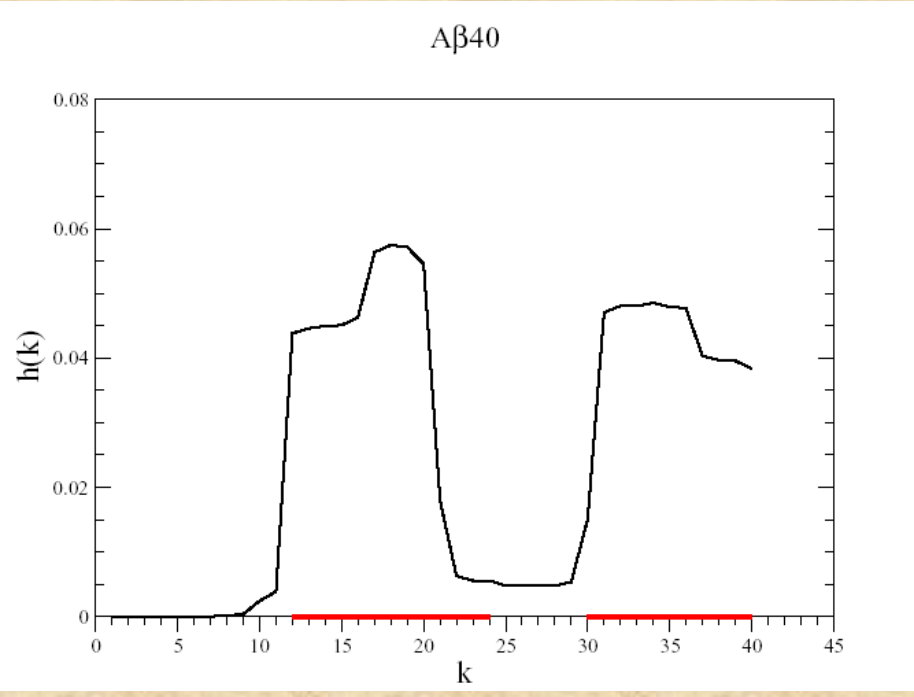
$$n_{ab}^{p(a)} = \#\ \text{of parallel (antip) contacts in strands between } a \text{ and } b$$

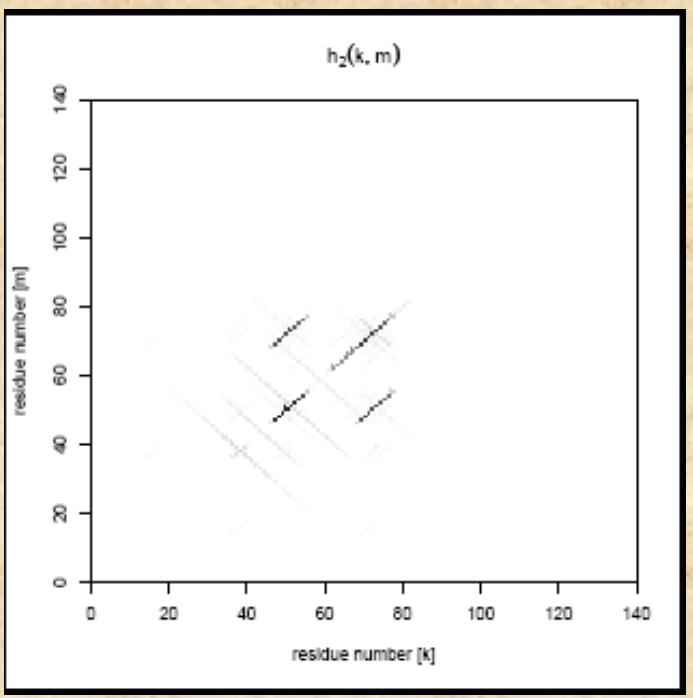
$$n_{ab} = \#\ \text{of pairs } a \text{ and } b$$

Prediction of specific pairings and sequence aggregation propensities

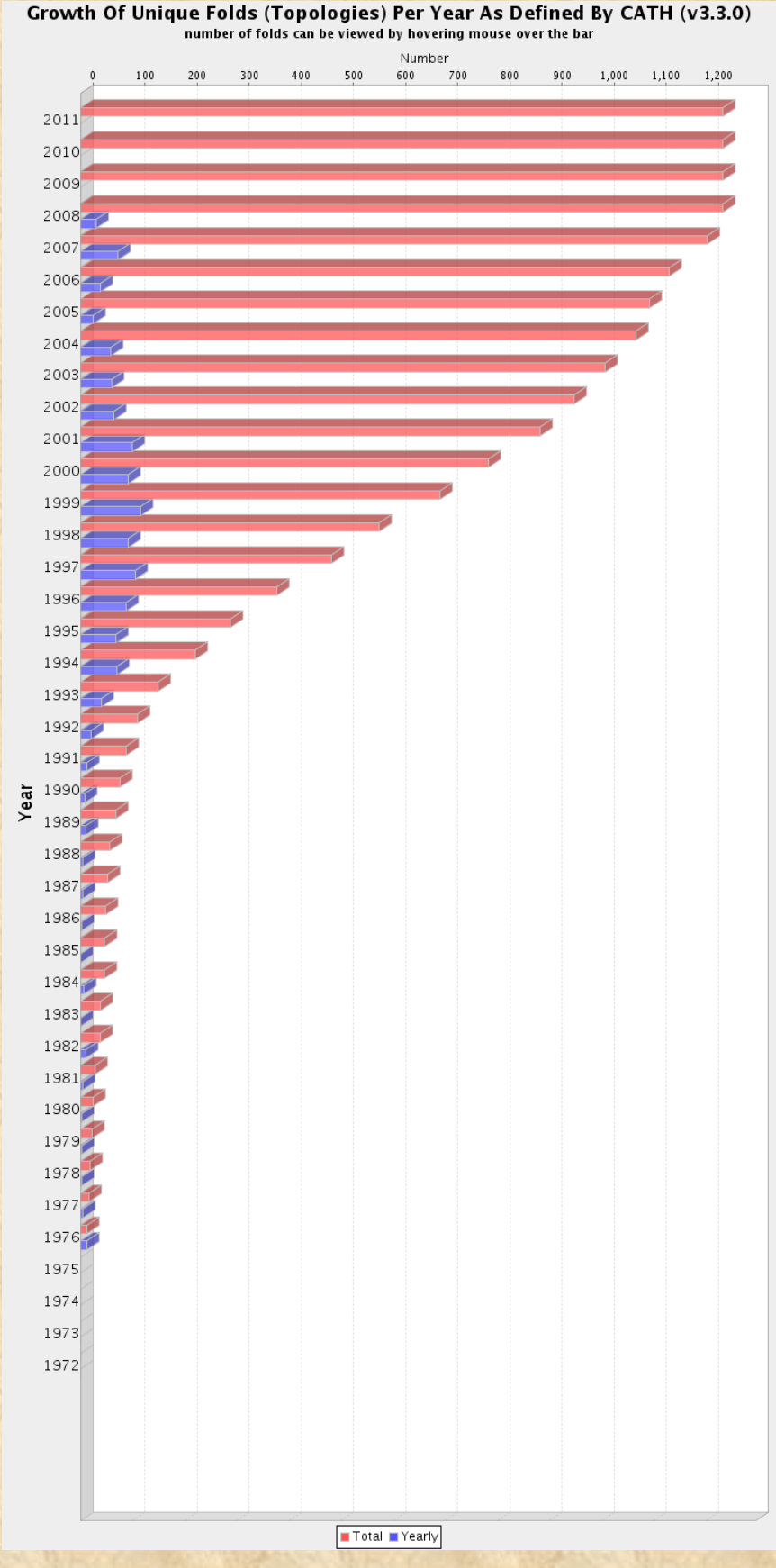
PROBABILITY THAT A GIVEN AMINO-ACID *K* BELONGS TO AN AGGREGATED SEGMENT OF LENGTH L (EITHER P OR AP)

PROBABILITY THAT A.A. *K* IN FIRST CHAIN FORMS AN HYDROGEN BOND WITH *J* IN THE SECOND CHAIN.

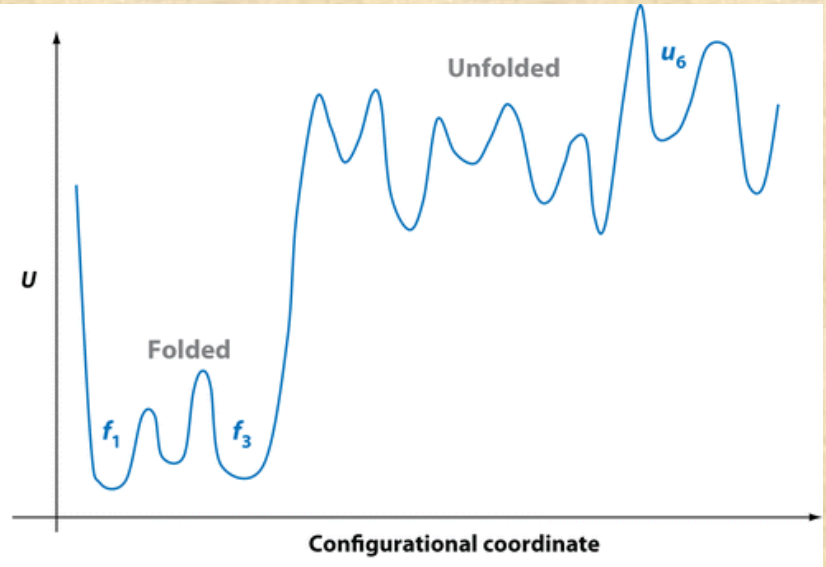




PRE-SCULPTED ENERGY


Growth Of Unique Folds (Topologies) Per Year As Defined By CATH (v3.3.0)
number of folds can be viewed by hovering mouse over the bar
Number
0
100
200
300
400
500
600
700
800
900
1,000
1,100
1,200
Year
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
1999
1998
1997
1996
1995
1994
1993
1992
1991
1990
1989
1988
1987
1986
1985
1984
1983
1982
1981
1980
1979
1978
1977
1976
1975
1974
1973
1972
Total
Yearly

SIMULATED ANNEALING



Zuckerman DM. 2011.
Annu. Rev. Biophys. 40:41–62

RANDOM EXPLORATION OF THE CONFORMATIONAL SPACE

E_i energy starting conformation

E_f energy final conformation

$$\Delta E = E_f - E_i < 0$$

if the new conformation has lower energy

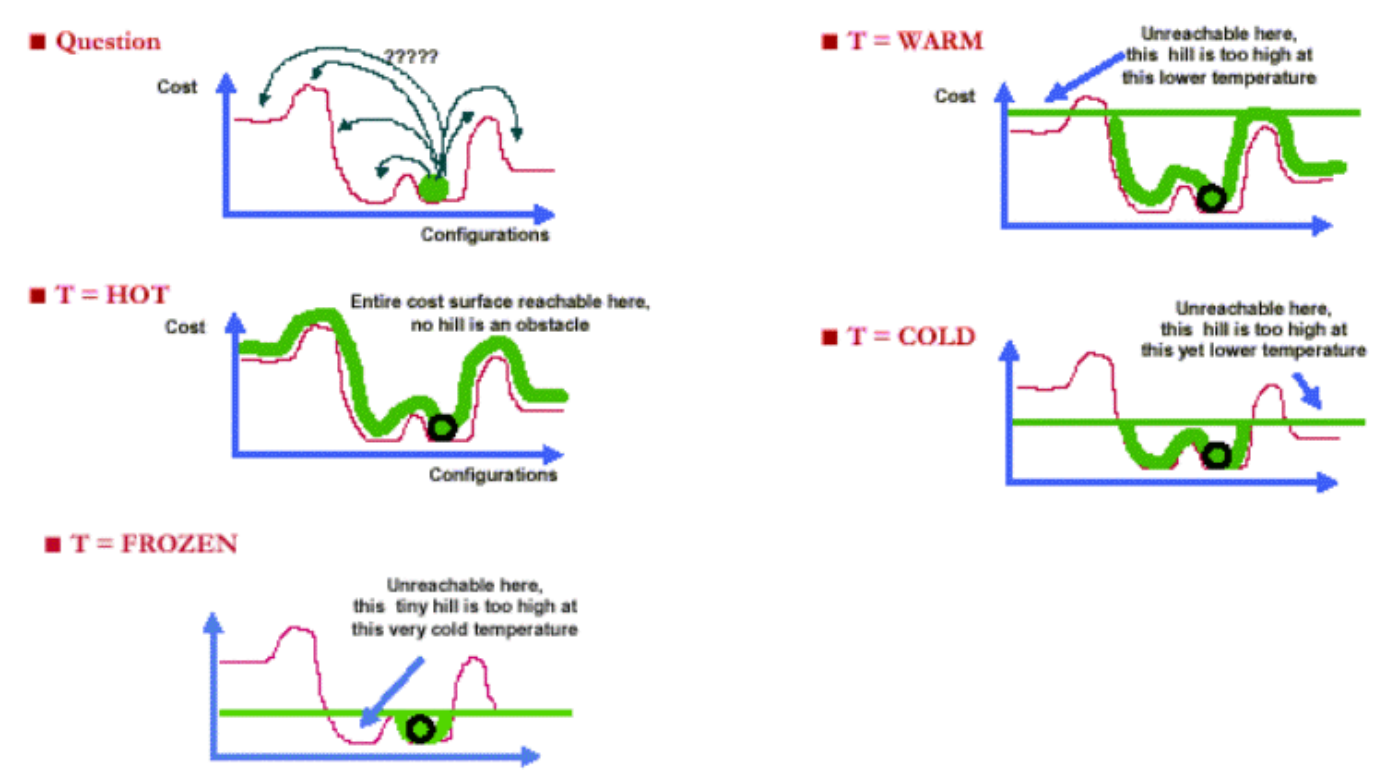
SIMULATED ANNEALING

$$X = \min \begin{cases} 1 \\ e^{-\frac{\Delta E}{T}} \end{cases}$$

MOVE IS ACCEPTED IF

$$Y < X$$

Y RANDOM NUMBER $[0,1]$



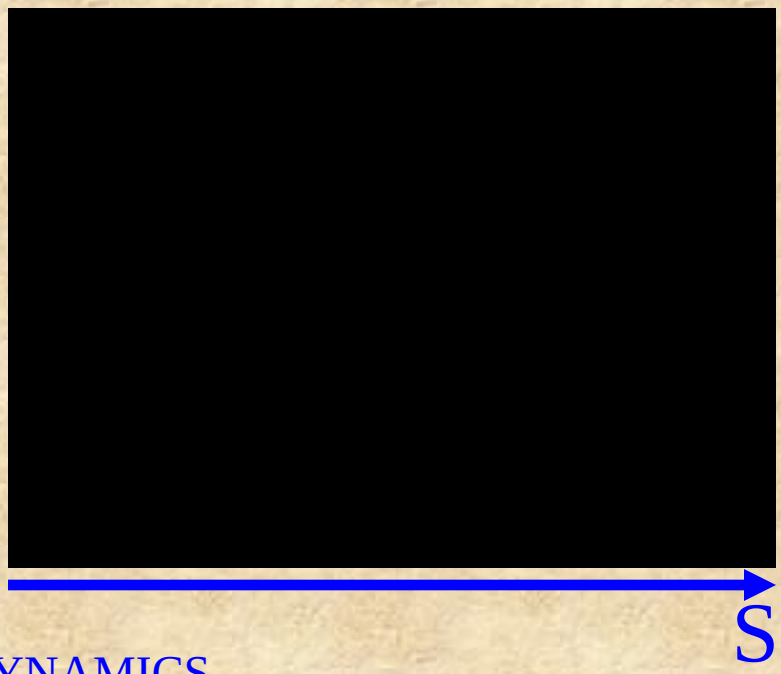
T IS SLOWLY DECREASED

METADYNAMICS

A Laio, M Parrinello,
Escaping free-energy minima
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF, **99**, 12562 (2002)

HOW TO FIND STABLE MINIMA WHICH ARE SEPARATED BY BARRIERS THAT CANNOT CLEARED IN THE AVAILABLE SIMULATION TIME

THE METHOD IS BASED ON AN ARTIFICIAL DYNAMICS (METADYNAMICS)



1) IDENTIFY COLLECTIVE VARIABLES S WHICH ARE ASSUMED TO PROVIDE A RELEVANT COARSE GRAINED DESCRIPTION OF THE SYSTEM

2) TO BIAS THE DYNAMICS ALONG THESE VARIABLES.

3) RUN IN PARALLEL SEVERAL MOLECULAR DYNAMICS EACH BIASED WITH A METADYNAMIC POTENTIAL

4) SWAPS OF THE CONFIGURATIONS

ATOMISTIC MODEL

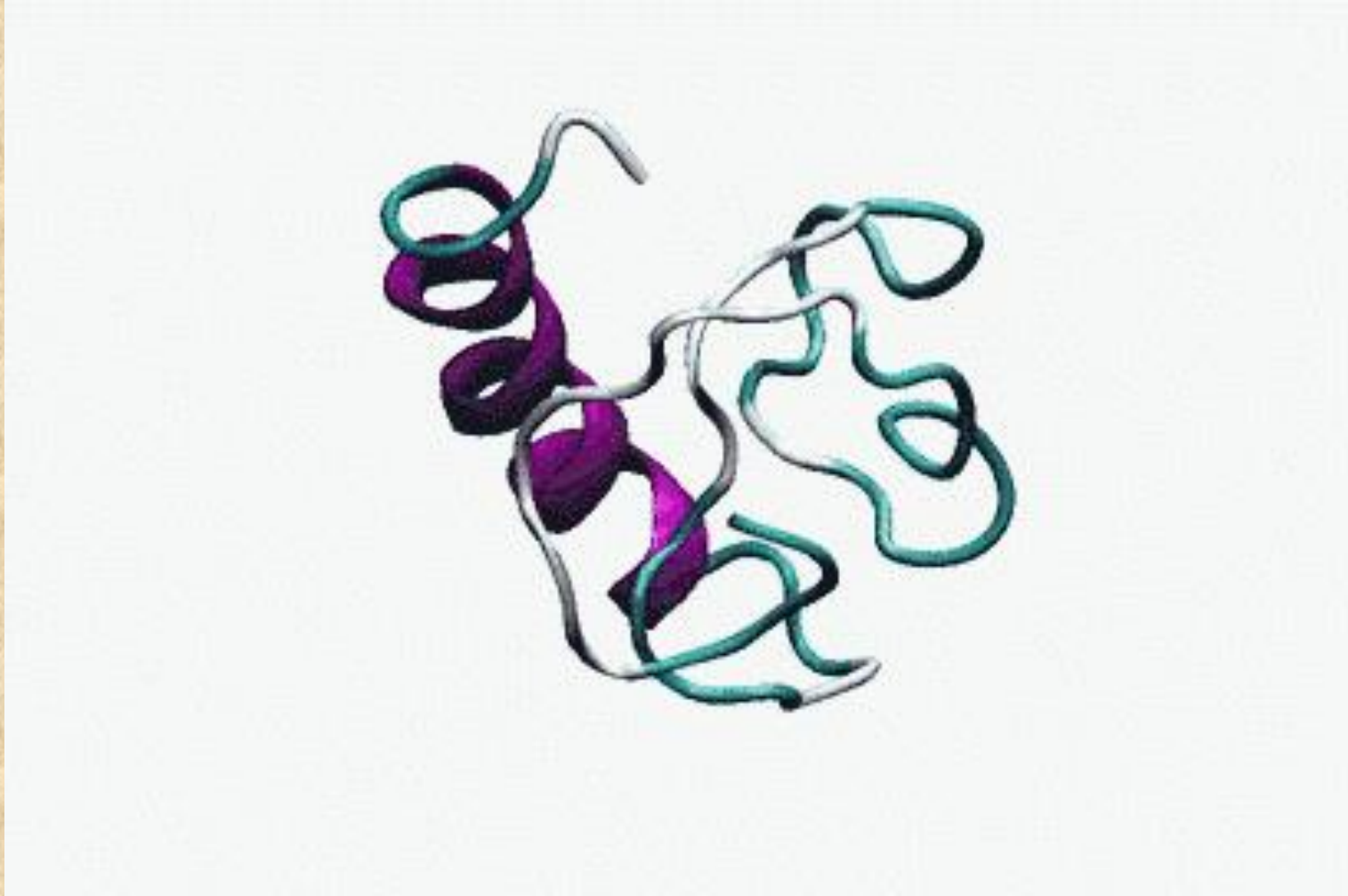
60 AMINO ACIDS POLYVALINE (VAL60)

- Why VAL? (is small but not too much)
- MD simulations with AMBER force field and package GROMACS
- Bias-exchange METADYNAMICS with 6 replicas
- Six collective variables linked to secondary structure elements

50 microseconds molecular dynamics simulation

We generate an ensemble of 30000 all-atom conformations

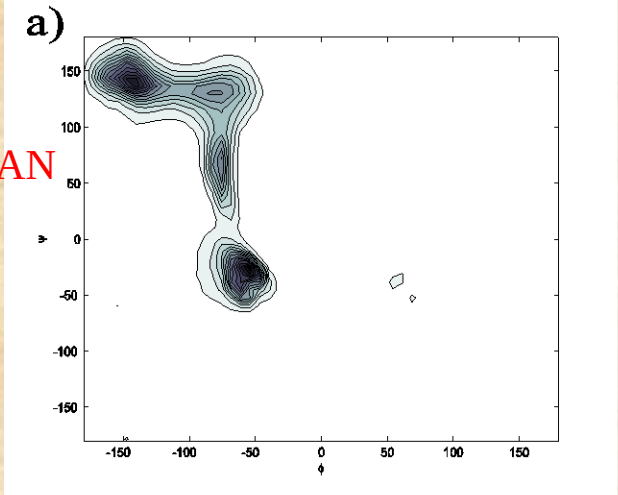
SIGNIFICANT SECONDARY STRUCTURE CONTENT AND SMALL RADIUS OF GYRATION

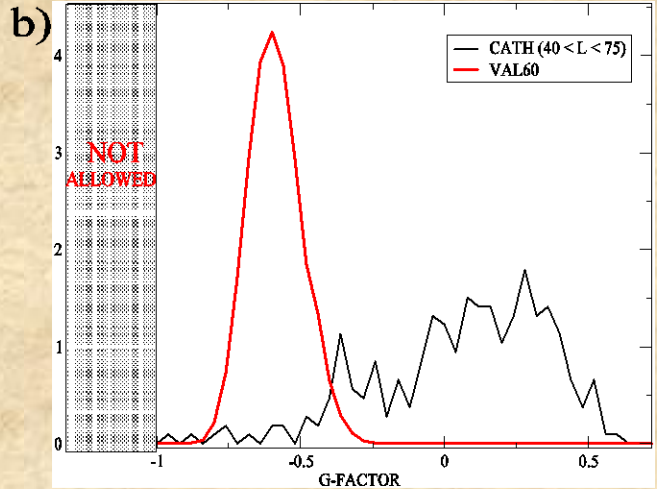


We verify they are local minima also for ALA-60

Structural quality resembles that of real protein

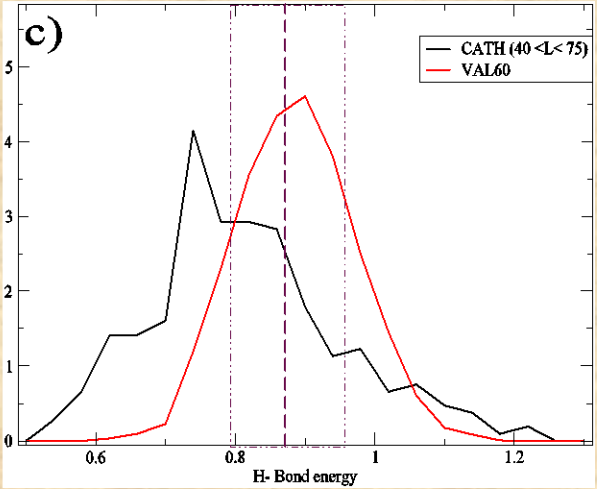
RAMACHANDRAN PLOT

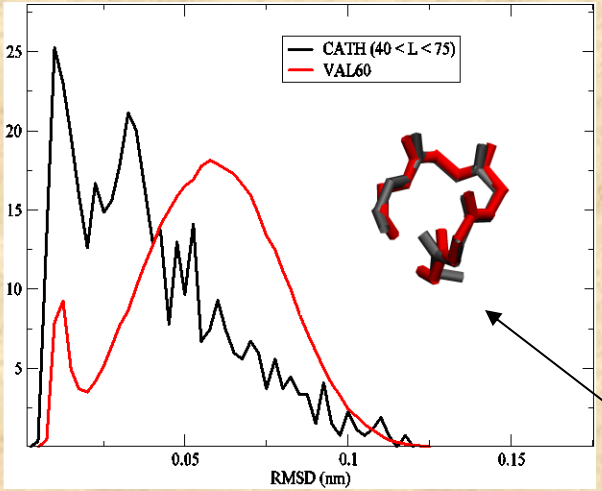




QUALITY MEASURE G-FACTOR

H-BOND ENERGY COMPUTED WITH PROCHECK





FRAGMENT DISTANCE < 0.6 A

0.7 A

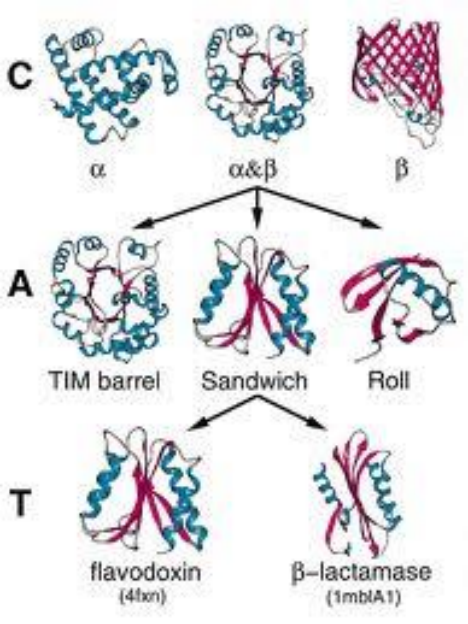
FIRST RESULT

FINDING BY MOLECULAR DYNAMICS AT AN ALL-ATOM LEVEL A LIBRARY OF 30000 PROTEIN LIKE STRUCTURES

http://datadryad.org/handle/10255/dryad.1922

RELATION BETWEEN VAL60 AND REAL PROTEINS

The *Class Architecture Topology and Homologous* superfamily protein structure classification (CATH) is one of the main databases providing hierarchical classification of protein domain structures.



300 STRUCTURES

40 < L<75

SIMILARITY: TM-SCORE (Zhang Scolnick 2005)

ALLIGNMENTS OF SECONDARY STRUCTURES ALLOWING INSERTIONS AND DELETIONS (COVERAGE)

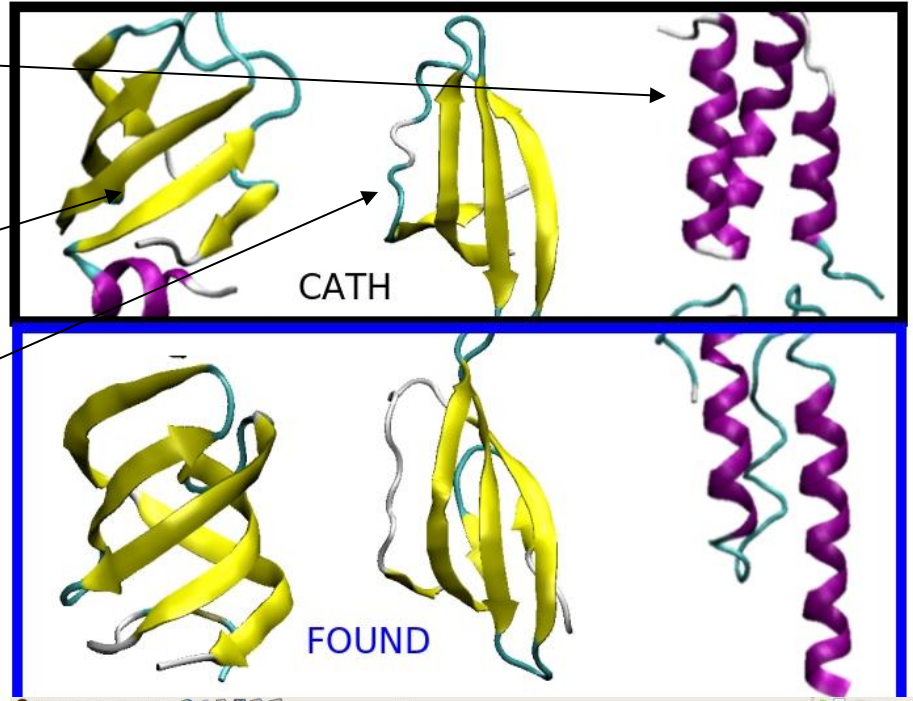
MINIMIZATION OF THE RELATIVE DISTANCE BETWEEN ALIGNED RESIDUES (RMSD)

TM=0.45

1x9b

1ib8

1g29



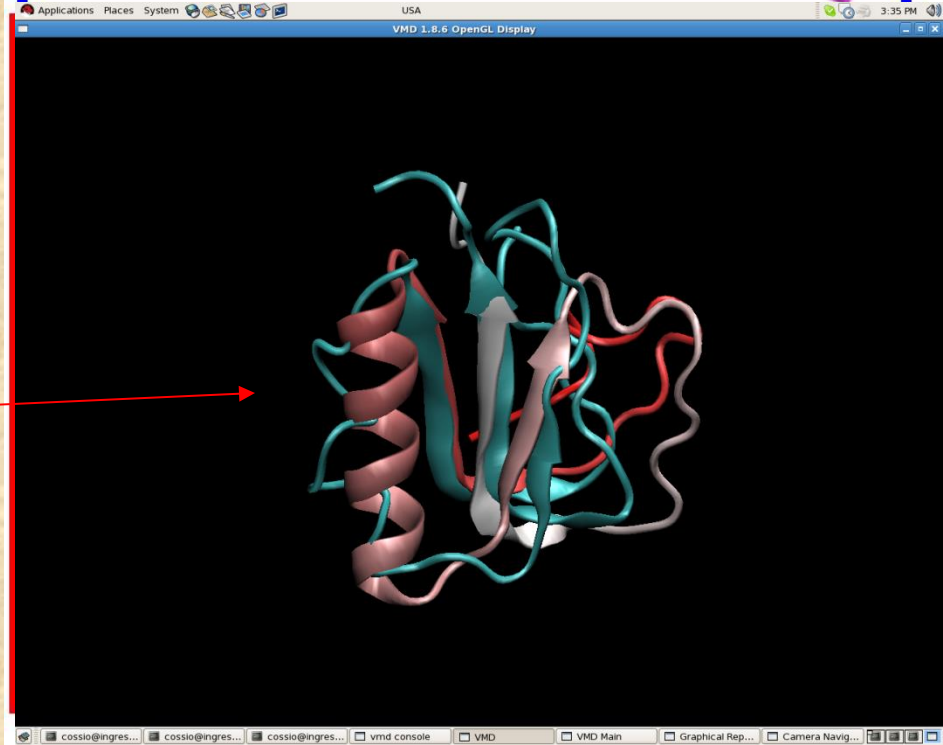
COMPARISON

VAL60 VS CATH

$40 < L < 75$

$\approx$ 300 FOLDS

1uxy



SECOND RESULT

THE COMPUTATIONAL SETUP USED IN THIS WORK ALLOW US TO EXPLORE THE MAJORITY OF THE FOLDS IN NATURE (AT LEAST FOR THESE LENGTHS)

COMPARISON

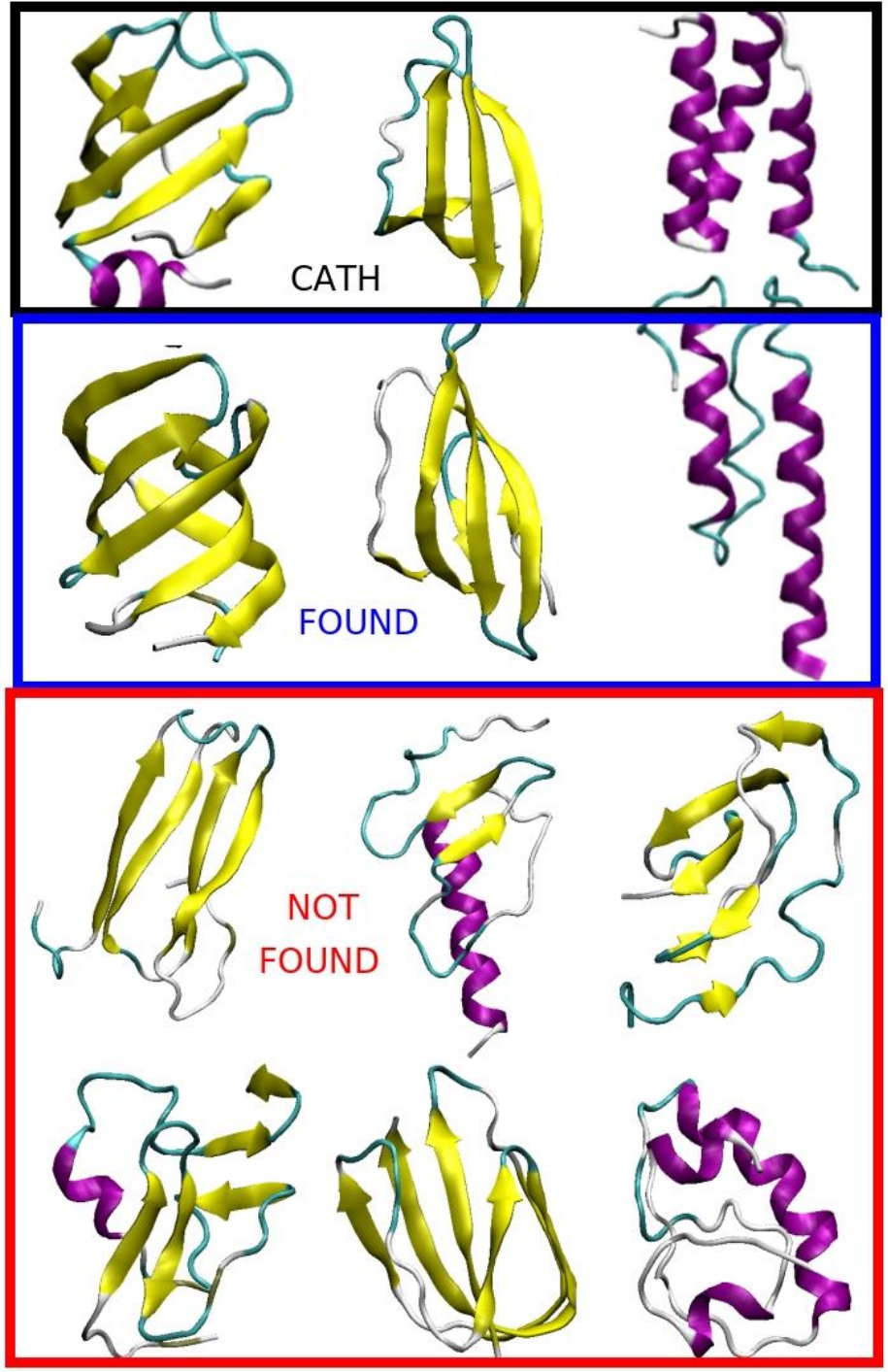
POLYVAL VS CATH

NOT ALL VAL60 ARE PRESENT IN CATH!!!!!!!

TM =0.45

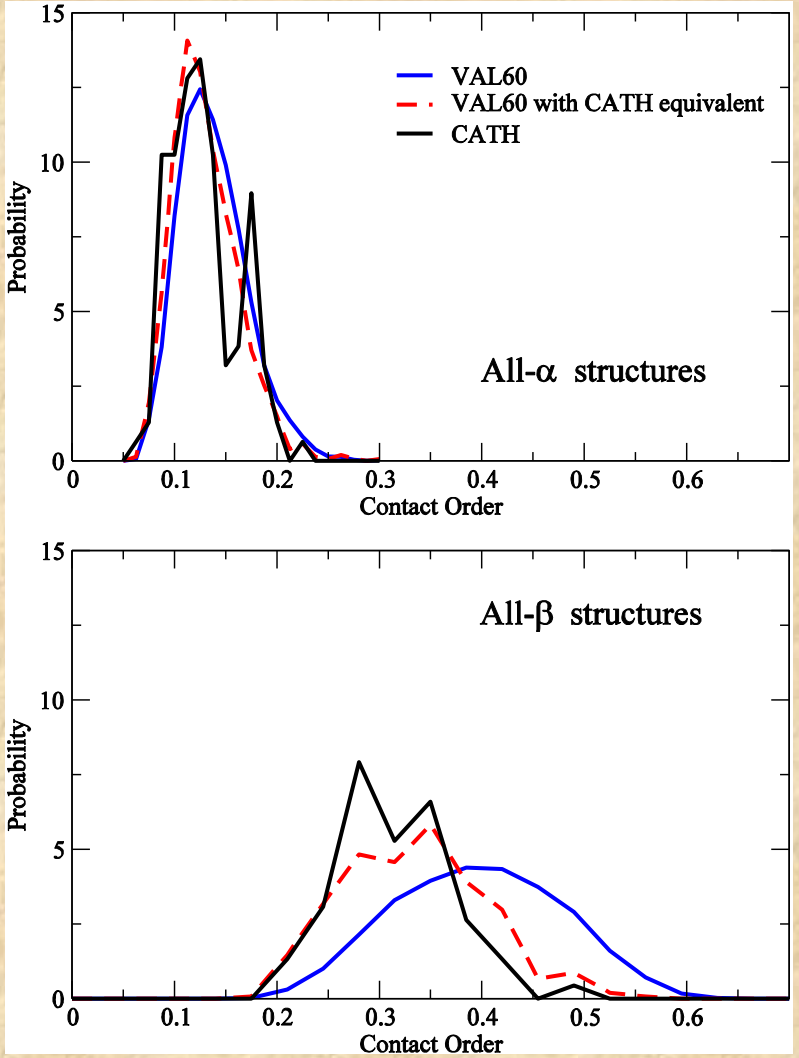
VAL60 → 7000

CATH → 300



THIS MIGHT JUST DEPEND ON THE CHOSEN SIMILARITY THRESHOLD

DO STRUCUTRAL DESCRIPTORS DISCRIMINATE BETWEEN CATH AND VAL60?



-CONTACT ORDER:
Average sequence separation between contacting residues
(related to folding rates Plaxco Simons Baker 1998)

-Real protein strucures were selected under a bias towards low CO

- protein structures are selected to be topologically less entangled

THIRD RESULT

THERE IS NO ONE-TO-ONE CORRESPONDENCE BETWEEN PDB LIBRARY AND THE ENSEMBLE OF COMPACT STRUCUTRES WITH SIGNIFICANT SECONDARY STRUCUTURE CONTENT (VAL60)

SUMMARY

- VAL60 SET IS REPRESENTATIVE OF REAL PROTEINS

  (PROTEINS FOLDS SELECTED BY GEOMETRY AND SIMMETRY AND NOT BY CHEMISTRY OF THE SEQUENCE)

- KNOWN FOLDS FORM ONLY A SMALL FRACTION OF THE FULL DATABASE
- NATURAL FOLDS ARE CHARACTERIZED BY SMALL CONTACT ORDER

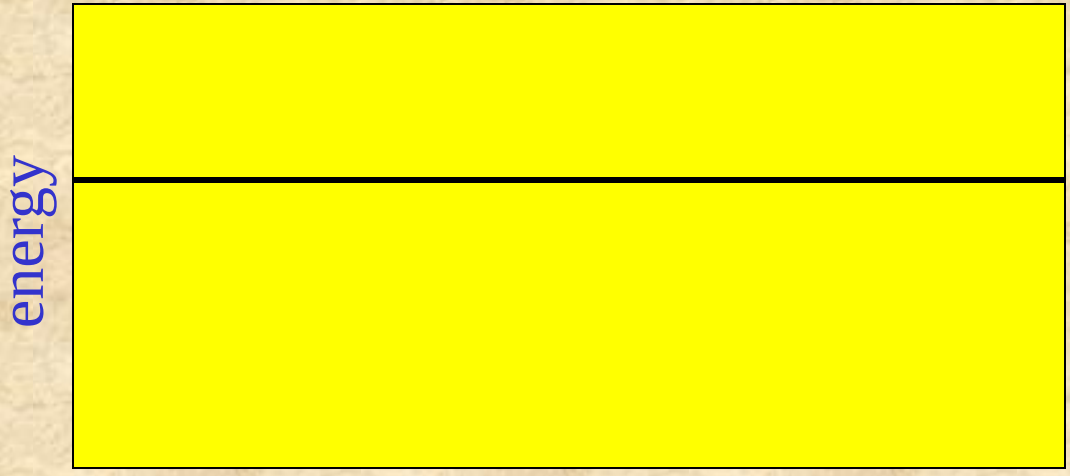
WHY

KINETIC ACCESSIBILITY

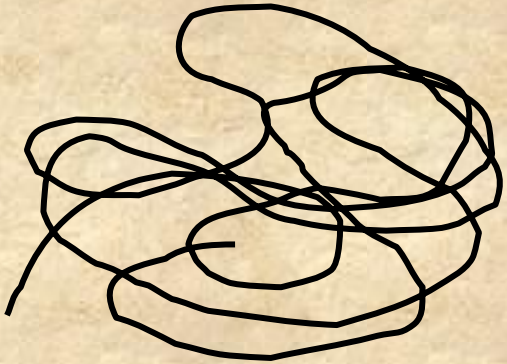
HIGHER CO ↔ HIGHER TENDENCY TO AGGREGATE?

Compact *versus* marginally compact phase

Homopolymer

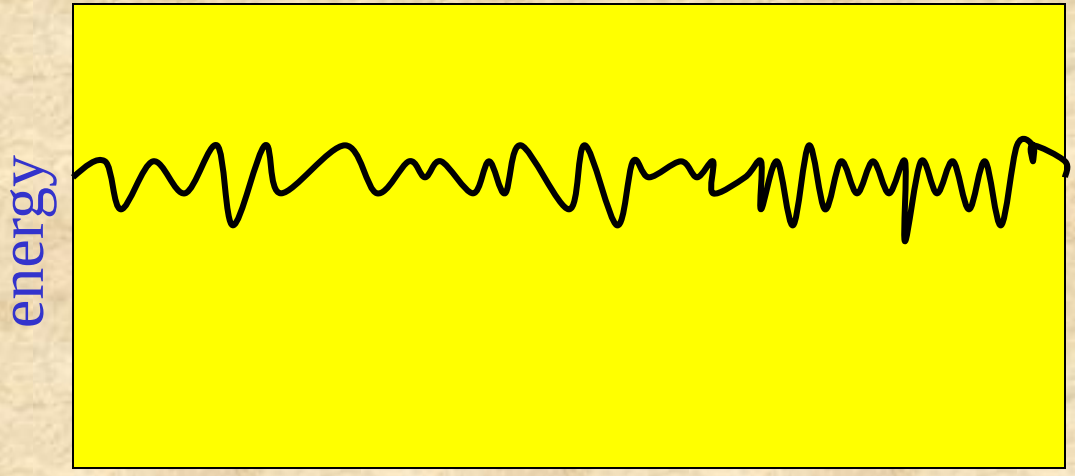


compact conformations

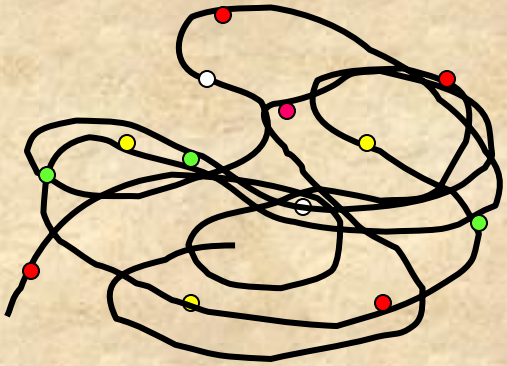


Compact *versus* marginally compact phase

Hetero-polymer

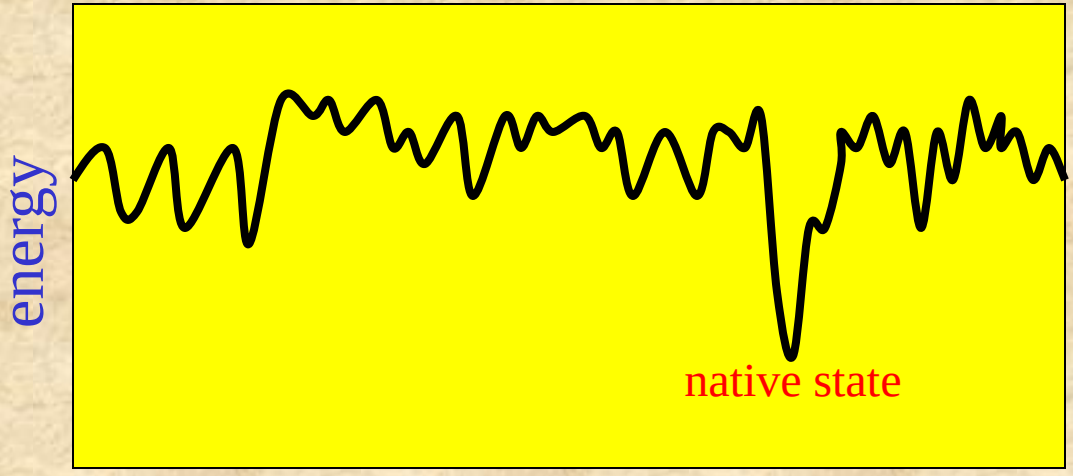


compact conformations

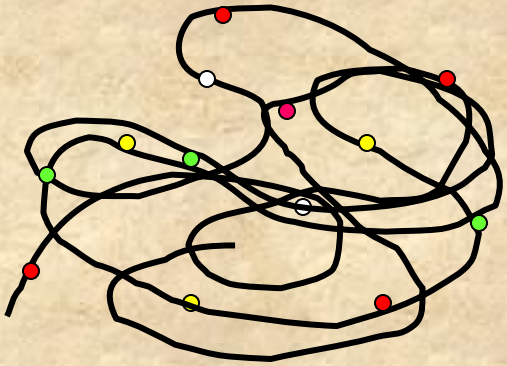


Compact *versus* marginally compact phase

Protein-like sequence

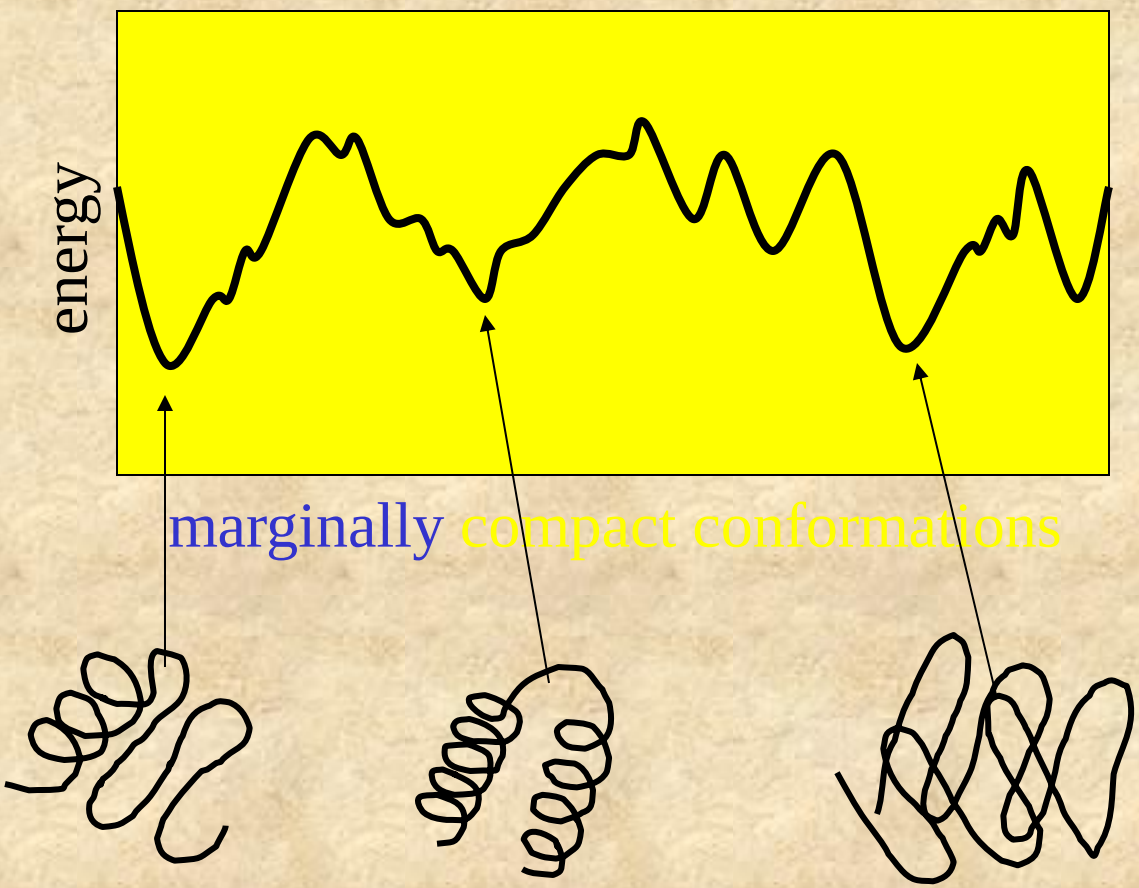


compact conformations



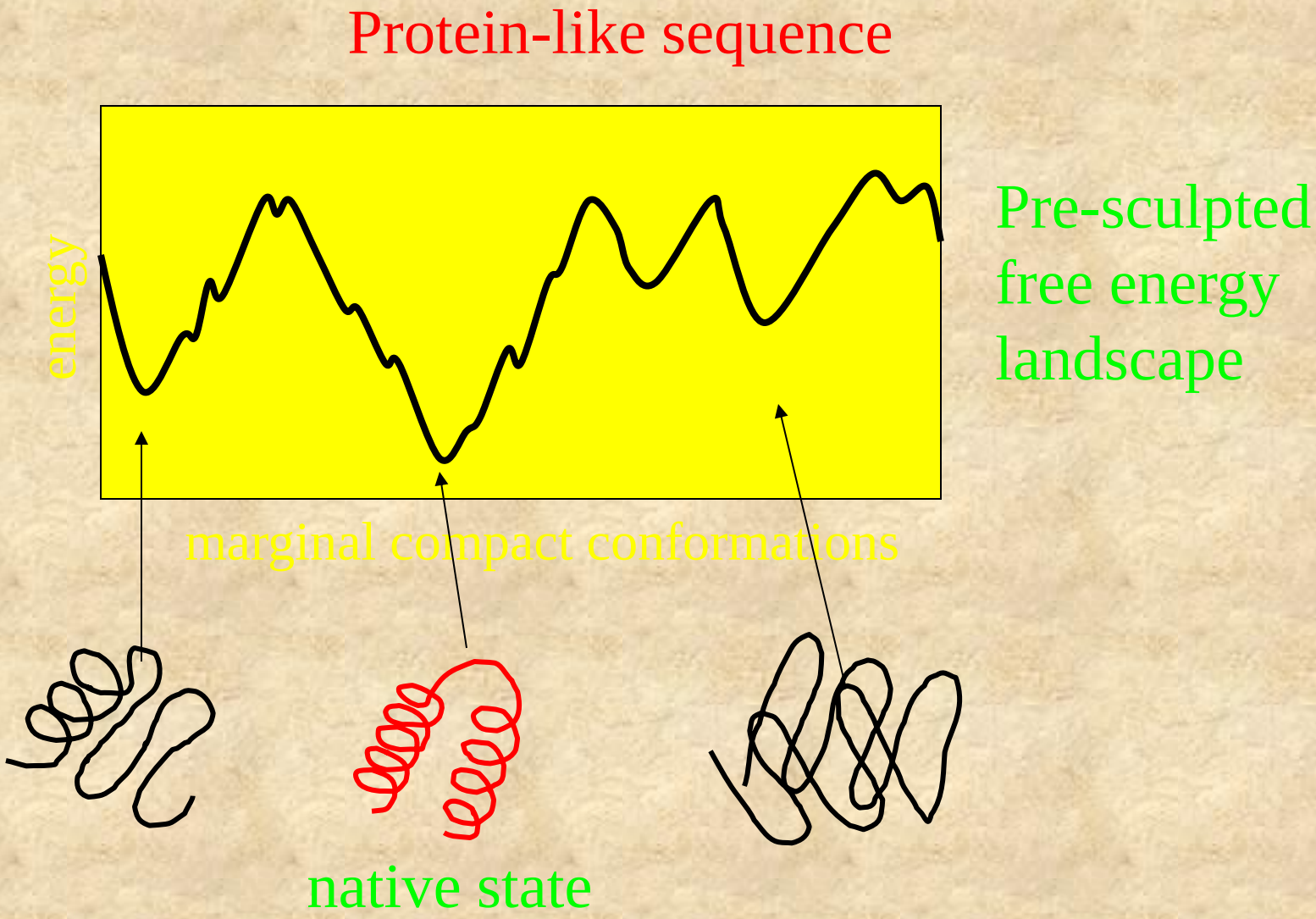
Compact *versus* marginally compact phase

Homopolymer



Pre-sculpted free energy landscape

Compact *versus* marginally compact phase



Others problems:

Unstructured proteins

Repeated proteins

Membrane proteins