Protein Structure Prediction
¼ century of disaster

- Claims
- Frauds
- Hopes
- Progress

Protein structure who cares ?

- 2 ½ to 3 decades
- many kilograms of literature
- many prizes awarded (last year Max-Planck)
- listed as a "grand challenge" problem
- IBM's big blue
- competitions / comparisons (like chess / Go)

What is prediction ?

- protein sequence -> coordinates like x-ray
- similarity to known structure ?
  - boring

Justification for prediction

Absolute basis for

- drug design

Useful

- understanding enzymes
- rational protein engineering
- industrial use (sensors, catalysts)

Justification for this talk 1970's

- 1975 Levitt M Computer simulation of protein folding
  50 residues, "such an approach .. understanding and simulation of … biological assembly processes"

- 1978 Sternberg MJ, Thornton JM.
  "In principle, it is possible to predict theoretically the three-dimensional structure of a protein from its amino acid sequence"

Levitt and Warshel, Nature 253, 694-698, (1975)
Sternberg and Thornton, Nature27115-20 (1978)

Recent

- 1998 Duan and Kollman
  36 residues, 1000 ns
  256 processors, 2 months
  do not find native structure

- 1999 Chipot…
  11 residues 100's ns
  do not find expected structure

- 2003 Klepeis and Floudas
  "ASTRO-FOLD, a novel and complete approach for the ab initio prediction of protein structures given only the amino acid sequences of the proteins"

Duan Y, Kollman PA, Science 282, 740-744 (1998)
Chipot … Pohorille, Proteins 36, 383-399 (1999)
Klepeis & Floudas, Biophys. J. 85, 2119-2146 (2003)

Long history

- 1969 Scheraga, Calculation of polypeptide conformation. Harvey Lect. 63:99-138
- 1970 Gibson and Scheraga, Minimization of polypeptide energy. IX. A procedure for seeking the global minimum of functions with many minima., Comput Biomed Res. 1970, 375-384
- 1974 Chou and Fasman, Prediction of protein conformation, Biochemistry;13:222-245
- 1975 Levitt and Warshal, Computer simulation of protein folding, Nature 253, 694
- 1977 McCammon.. Karplus, Dynamics of folded proteins, Nature, 267, 585

Patents

- 1985, Levinthal and Fine, "… energy and pairwise central forces of particle interactions" includes "a means for calculating energy and force values for each *ij* pair… force on each particle"

- 1993, Skolnick and Kolinksi, "A computer system and method are disclosed for determining a protein's tertiary structure from a primary sequence of amino acid residues"

- 1995, Eisenberg, "Method to identify protein sequences that fold into a known three-dimensional structure"

- 1997, Rose and Srinivasan "A computer-assisted method for predicting the three-dimensional structure of a protein fragment from its amino acid sequence"

Books

- "Prediction of Protein Structure …" Fasman 1989
- "Protein Folding Problem & Tertiary Structure Prediction", Merz and le Grand, 1994
- "Protein structure prediction: A practical approach" Sternberg, 1996
- "Protein structure prediction: Methods…", Webster 2000
- "Protein Folding Problem & Tertiary Structure Prediction", Friesner, 2001
- "Protein structure prediction: A bioinformatic ..." Tsigelny, 2002
- …

Long History

- 1969 Scheraga, Calculation of polypeptide conformation. Harvey Lect. 63:99-138
- ….
- 1999, Liwo, … Scheraga, Protein structure prediction by global optimization…, PNAS, 96, 5482-5485

- 2001, Proteins editorial – simulations are not enough for publication

Grounds for optimism

Anfinsen

- take a protein
- denature
- remove denaturing
- it refolds

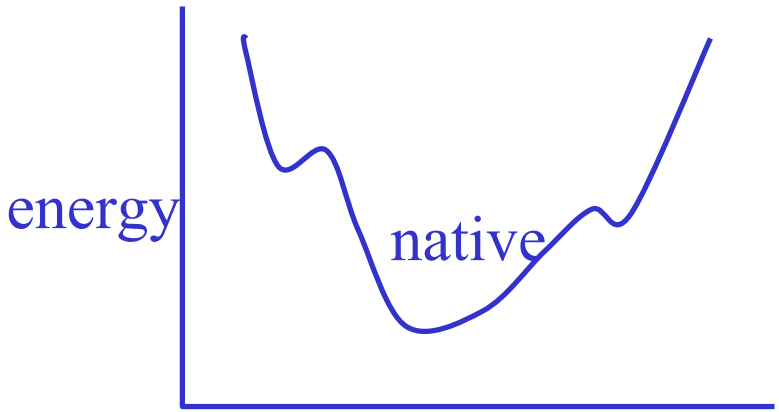
Two consequences

- the information for the structure is in the sequence
- protein finds its free energy minimum

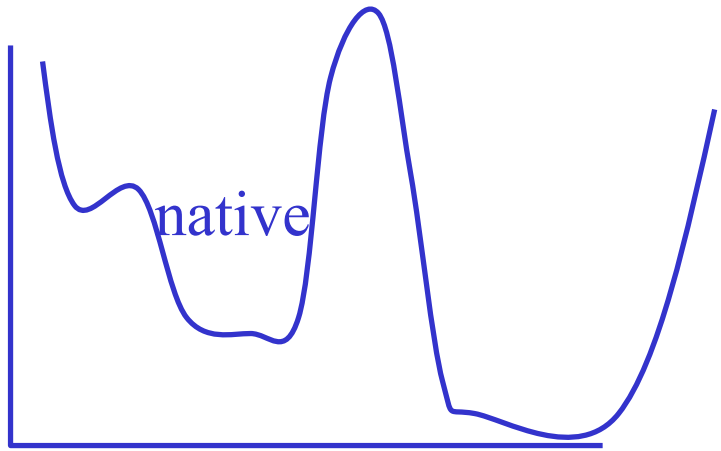
Anfinsen or bust

How true ?

- dogma
- does not work for all proteins
- some are kinetically trapped



thermodynamic



kinetically trapped

- consequences…

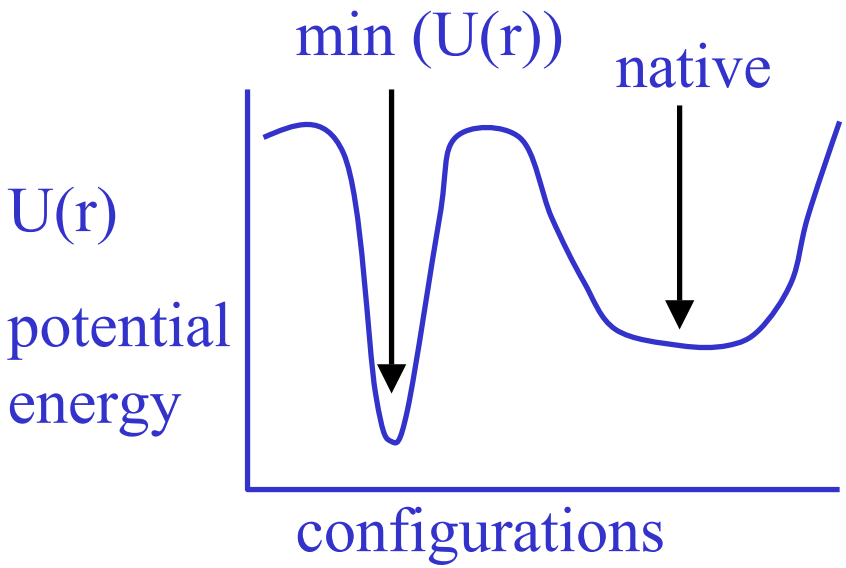
Is energy enough ?

We have a perfect model for potential energy

- do we want the minimum ?

Can we find the native ?

- requires
  - molecular dynamics, MC other simulation method



Are problems real ?

- β -fibril proteins
- poorly defined structures
- the world will turn into ^{56}Fe

What we want

Potential energy function

- the most visited structures

Free energy function

- the best scoring structure
- free energy is not a function of a structure

Truth !

- perfect model for physics + infinite time
  - no problem

Reality

- any method which predicts x-ray structures

Simple models

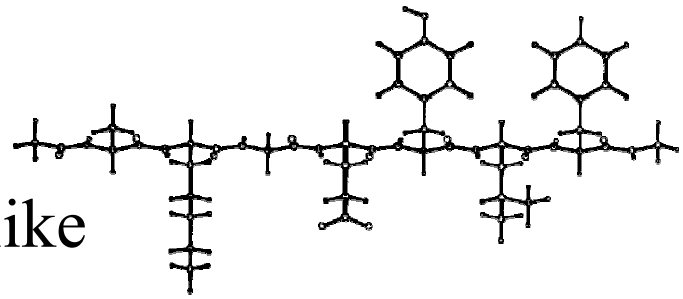
Model physics

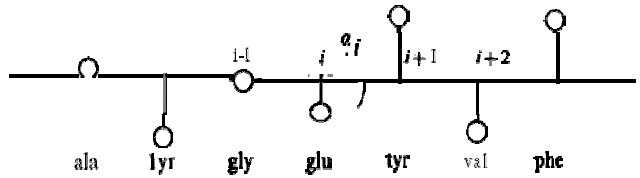
- what level ?
  - QM ?
  - atomistic ? (later)
  - simplified physical

- atomistic ?
  - too many degrees of freedom
  - search problem too big
  - too many local minima

simplified models 1975

- "under certain conditions, the method succeeds in 'renaturing' BPTI"
- model
  - sidechain-sidechain Lennard-Jones like
  - number neighbour estimate solvation
  - near neighbour special treatment
  - backbone H-bonds
- calculations
  - BPTI
- calculated structures "have features in common with x-ray"
- "an exciting application .. prediction of the conformation of an unknown protein"





Levitt, M, J. Mol. Biol, 1976, 104, 59-107, ".. rapid simulation of protein folding"

simple models post 1975

2001 (Sessions et al)

- one pt / sidechain + one for backbone
- simple interaction forms
- fixed backbone geometry

2001 (Hassinen and Peräkylä)

- one pt / sidechain
- backbone dipoles
- accurate backbone geometry
- lots of dynamics terms (flexible bonds, angles, …)

Some elegant fitting methods (Gerber mid 90's)

Any better ?

Gibbs… Sessions, Proteins, 43, 186-202 (2001), Ab initio protein structure prediction …simplified off-lattice model
Hassinen & Peräkylä, J. Comput Chem (2001), New energy terms for reduced protein models…

2001 physical model results

	1975	2001
proteins for parameterising	1	100's
proteins for testing	1	few or 100's
reproduce native structures	a bit	some to 4 Å / some not
dynamics	EM + randomisation	MD, MC

Progress

- better testing (a bit)
- much more computer time
- local structure good
- better parameterisation / residue specific effects

Questions…

Questions from physical models

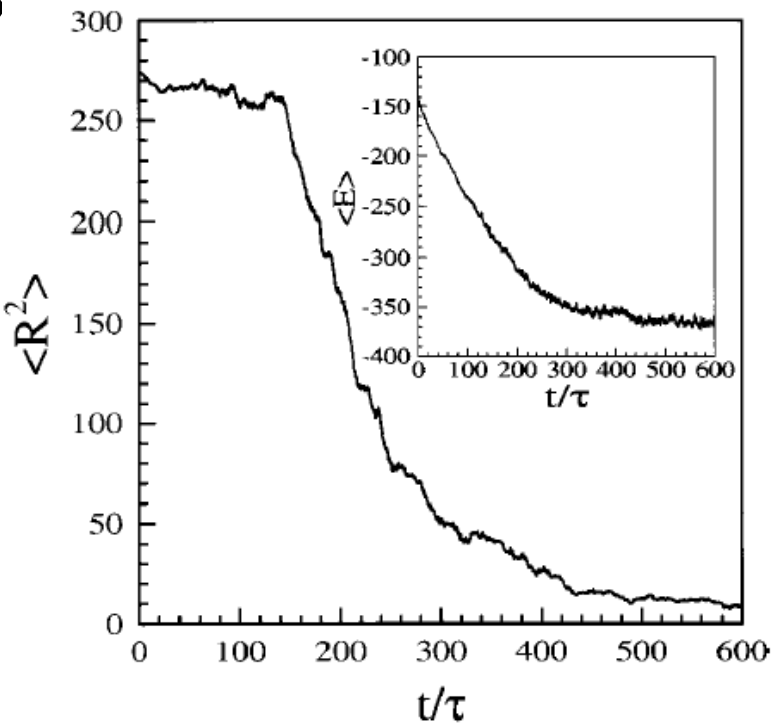
If we do better

most important:

- do we know which terms helped ?

Can we represent some of the physics ?

- folding kinetics
- heat capacity
- collapse ?
- what system ?
  - dumb homopolymer



Srinivas, Yethiraj, Bagchi, (2001), J Phys Chem B, 105, 2475-2478

Atomistic MD

- 1977 50 residues < 10 ps (10^{-12}) in vacuo
- 1980's 100 residues many 10's ps often in water

- 1998 Duan and Kollman, water
  36 residues, 1000 ns
  256 processors, 2 months
  do not find native structure

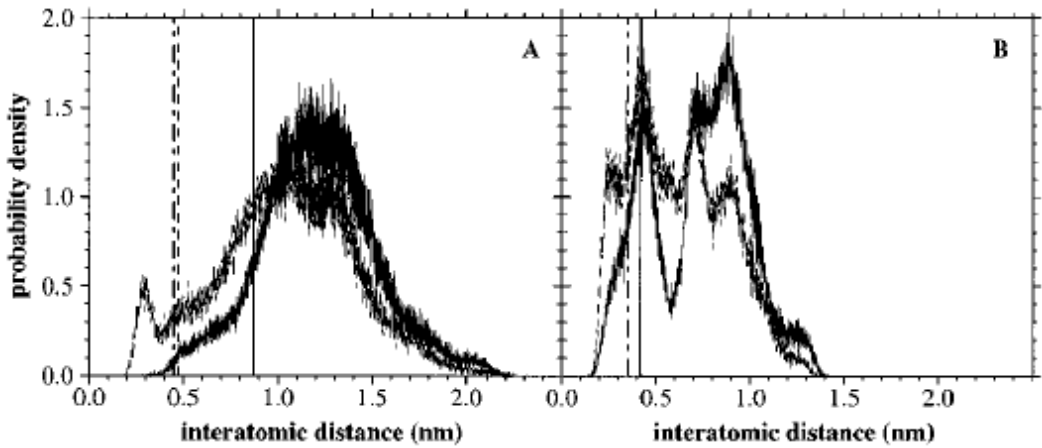
- 2001 Daura et al, water
  6 residues, several x 100ns

- 2003 Folding at home
  some lower bounds on some folding events

Duan Y, Kollman PA, Science 282, 740-744 (1998)
Daura, … Seebach, van Gunsteren, J. Am Chem. Soc, 123, 2393-2404 (2001)

atomistic MD progress

- small protein
  - possible intermediate (wrong answer)
- 6 residue peptide



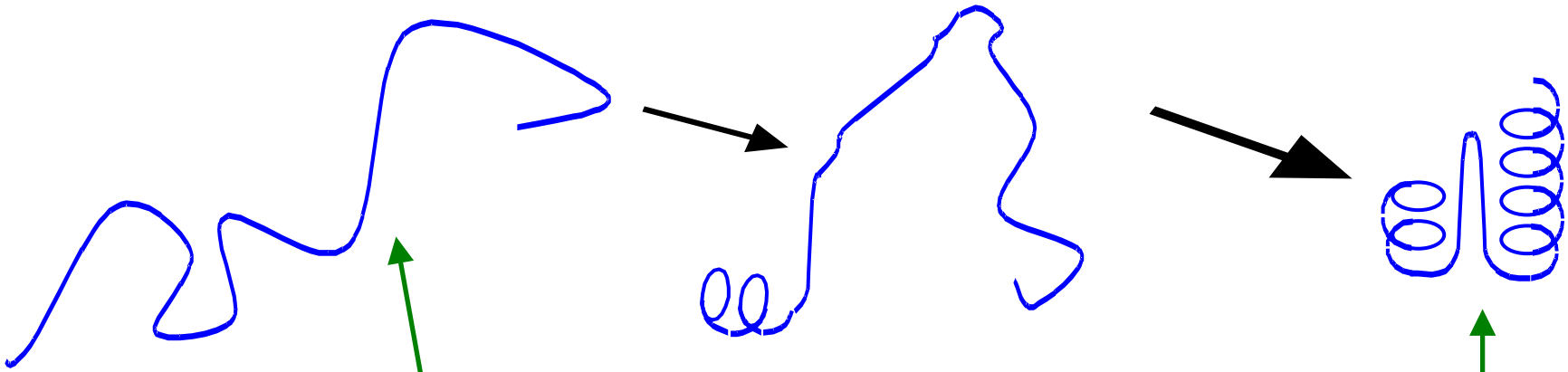
good	bad
complete folding / unfolding	6 residues
agrees with experimental data	one peptide
	testing ?
	clustering of structures (later)

Forgetting physics

- Kuntz … Kimelman 1976
  - "we will *not* attempt to determine the correct forces"
  - "nor are we concerned with simulation of the protein folding process"
- why is this useful ?

Kuntz, Crippen, Kollman, Kimelman, J. Mol. Biol, 106, 983-994 (1976), Calculation of protein tertiary structure

Modelling folding



Physics + time

- folding easy

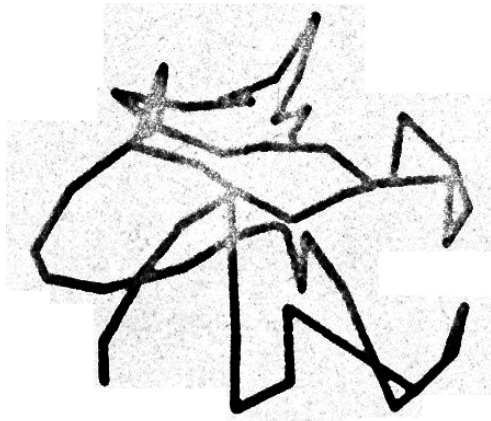
Folding is too hard

long range electrostatics ?
H bonds ?
everything is solvated

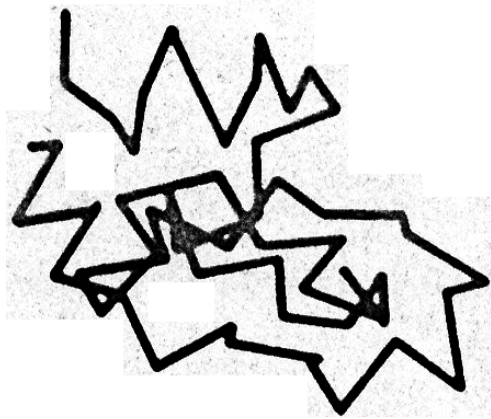
hydrophobic core
dense packing
different H bonds

Forgetting physics 1976

- one point / residue
- chain connectivity
- hydrophobic partitioning (from centre of mass)
- simple interaction matrix



- limitations
  - no parameterisation data
- Today ?

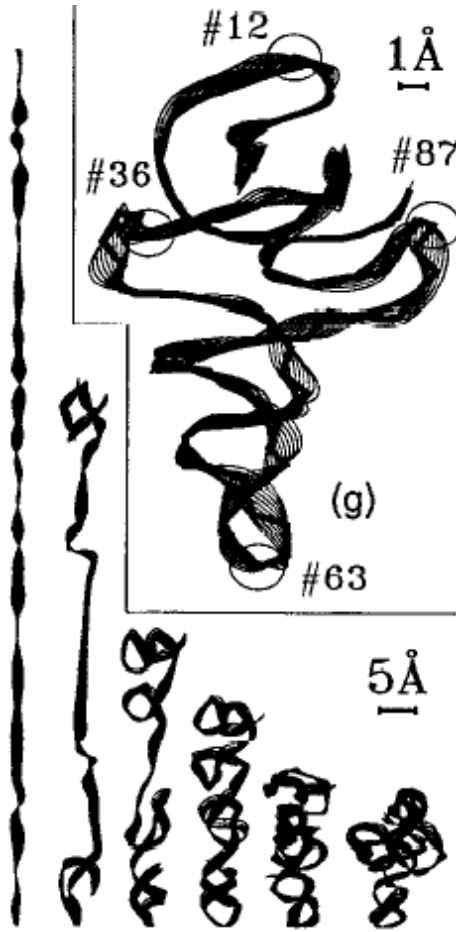


Kuntz, Crippen, Kollman, Kimelman, J. Mol. Biol, 106, 983-994 (1976), Calculation of protein tertiary structure

non-physical today

- lots of parameterisation data
- complete, automatic adjustment of parameters
  - discrimination function approach

- scary result
  - allow all possible pairwise interactions
  - for some cases, proteins cannot be folded



Vendruscolo & Domany, J. Chem. Phys, 109, 11101-11108 (1998), Pairwise contact potentials are unsuitable for protein folding
Grubmüller and Tavan, J. Chem. Phys, 101, 5047-5058 (1994), Molecular dynamics … simplified protein model

A hierarchical approach ?

Fix local structure and assemble big pieces

- reason for the field of secondary structure prediction

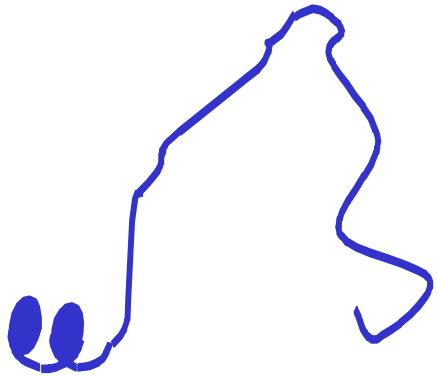
History

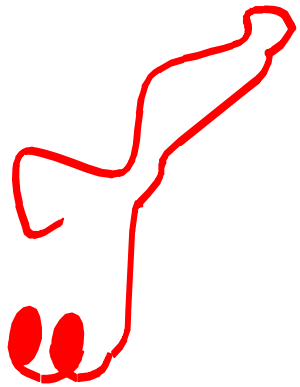
- 1974 Chou and Fasman, … GOR
  - single residue methods
  - 60 % or less
- 1990's
  - 66 % ? (machine learning)
- today
  - 76 % (sequence profiles)
    (every combination possible, nets of nets)
- is this enough ?

Rost, B. J. Struct Biol. 134, 204,-218 (2001) .. Protein secondary structure continues to rise

¾ secondary structure

Is this enough ?

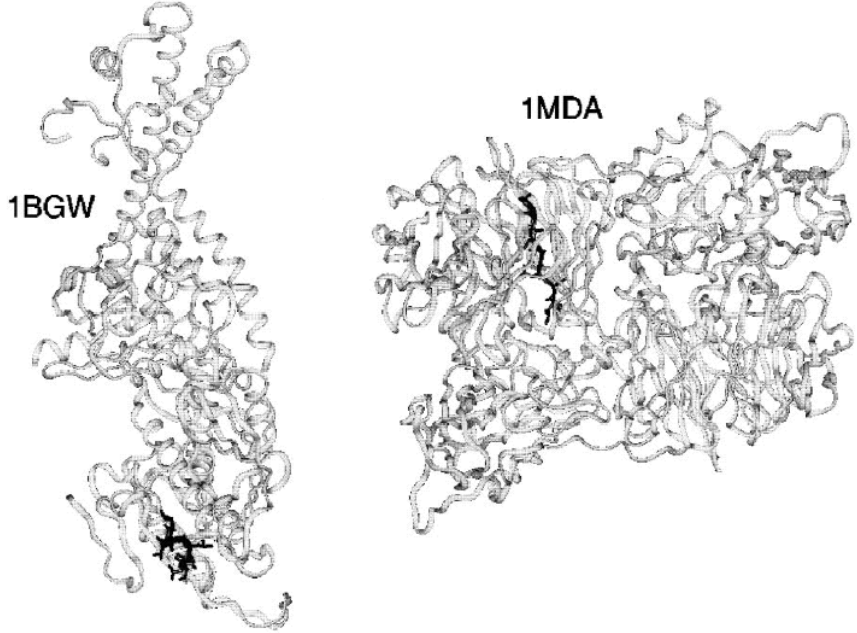




- a pure building block approach will fail

Fundamental problem

- secondary structure is not a local property



7-mer pair, 1amp and 1gky

Can one use this information ?

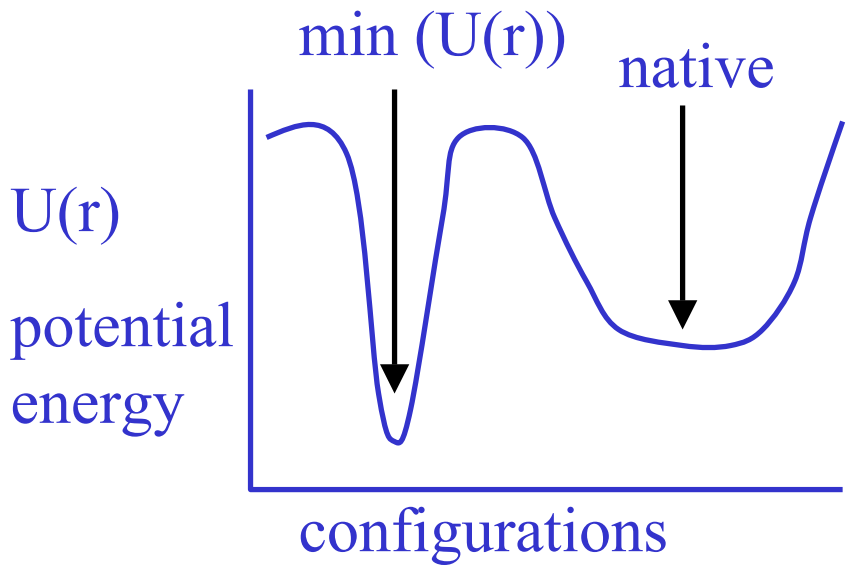
- combine with others
- are dihedrals always in their best position ?

- 8-mer pair, 1pht and 1wbc

Mezei, M, Prot. Eng, 11, 411-414 (1998) Chameleon sequences in the PDB
from Sudarsanam, S, Proteins 30, 228-231 (1998), "… identical octapeptides can have different conformations"

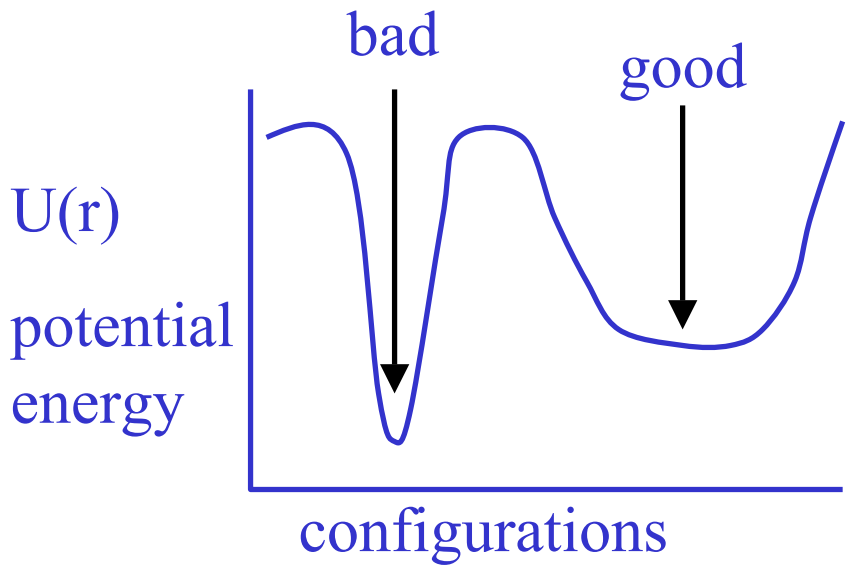
Maybe force field approach is wrong

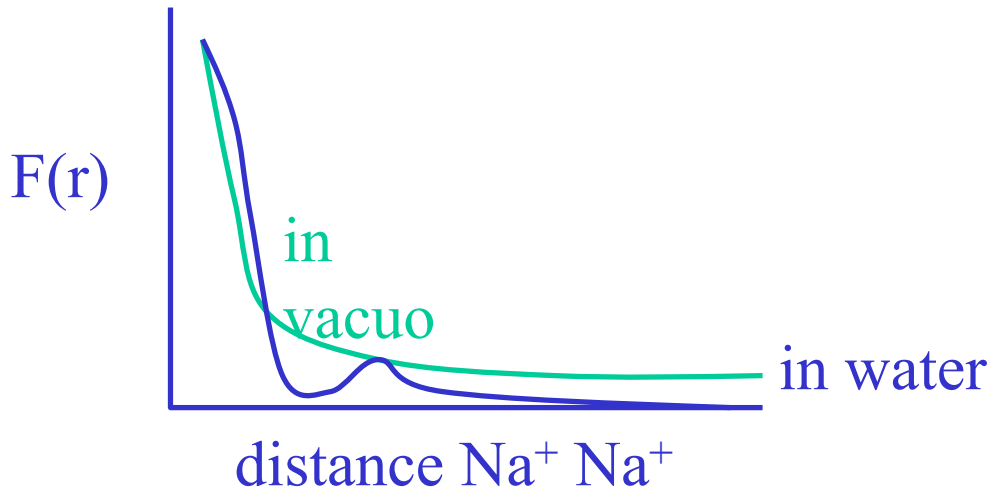
- Potential energy is too difficult
  - requires sampling
- Approximate free energies
- Potentials of mean force



Potential of mean force

- estimate the happiness of a minimum from a single point
- when is it good ?
  - simple systems
- when is it bad ?
- apply to proteins





Protein potentials of mean force

- 1990 Sippl, 1992 Jones et al
- 1985 Miyazawa & Jernigan
- 1978 Warme and Morgan
- 1976 Tanaka and Scheraga
- 1971 Pohl

Pohl, Nature, 234, 277 (1971)
Warme and Morgan, J. Mol. Biol 118, 273-287
Tanaka and Scheraga, Macromolecules, 9, 142-159 (1976)
Miyazawa & Jernigan, Macromolecules, 18, 534 (1985)
Sippl, J. Mol. Biol. 213, 819 (1990)

Innovations in potentials of mean force

Progress ?

- parameterisation
  - early – residue – residue contacts
  - distance dependent
  - atomic detail
  - parameterisation in terms of angles, …

Data

- 1978 not much
- 2003 more

Subject of hate..

Maybe it is really a searching problem…

Ben-Naim, J. Chem. Phys. 107, 3698, "Statistical potentials… are these meaningful ?" (1997)
Thomas and Dill, J. Mol. Biol., 257 (1996) "Statistical… How accurate are they ?" (1996)

Do not mention genetic algorithms

A mass graveyard

1970 Gibson and Scheraga, Minimization of polypeptide energy. IX. A procedure for seeking the global minimum of functions with many minima.

- molecular dynamics
  - SD, Langevin,
  - high dimensional
- potential energy contouring
- simulated annealing
- biased MC
- replica exchange
- self consistent mean field
- genetic algorithms
- branch and bound
- pole dropping
- minimum tunnelling
- chain growth
- function modification
  - deformation, diffusion eqn

State of the art ?

- 1975 looked promising
- 1988 just a bit more cpu time
- 1995 Rose, Srinivasan LINUS
- 2003 Klepeis and Floudas
  "ASTRO-FOLD, a novel and complete approach for the ab initio prediction of protein structures given only the amino acid sequences of the proteins"
- what was said in 1995 ?

Srinivasan, R. & Rose, G. D. (1995) Proteins Struct. Funct. Genet. 22, 81–99.


SCIENCE/TECHNOLOGY

20 NOVEMBER 6, 1995 C&EN

Researchers Advance Ability To Predict Structures Of Folded Proteins

John Hopkins Journal

- "It used to take a world-class lab five years to solve a protein" "LINUS promises much quicker and cheaper structures, soon to be done in a day or an hour"
- "Over the very long term, the implications of LINUS are too big to see, because a discovery of this dimension changes the way one sees the world"

Trends and overview

Sack the atomistic simulators ?

atomistic MD

- physics has not changed
  - some methods can be implemented (PB/solvent/..)
- proteins are more stable (good ?)
- extrapolation gives good simulation of 50 residues in 2046
- current results do suggest feasibility
- one needs big blue ?
  - not a useful method
- optimistic school of thought
  - interesting conformational space is
    - very limited
    - can be found
- fundamental question from longest simulations
  - searching or score functions

Databases

- sequences and structures
- much less noise than 25 years ago
- still not enough for some areas

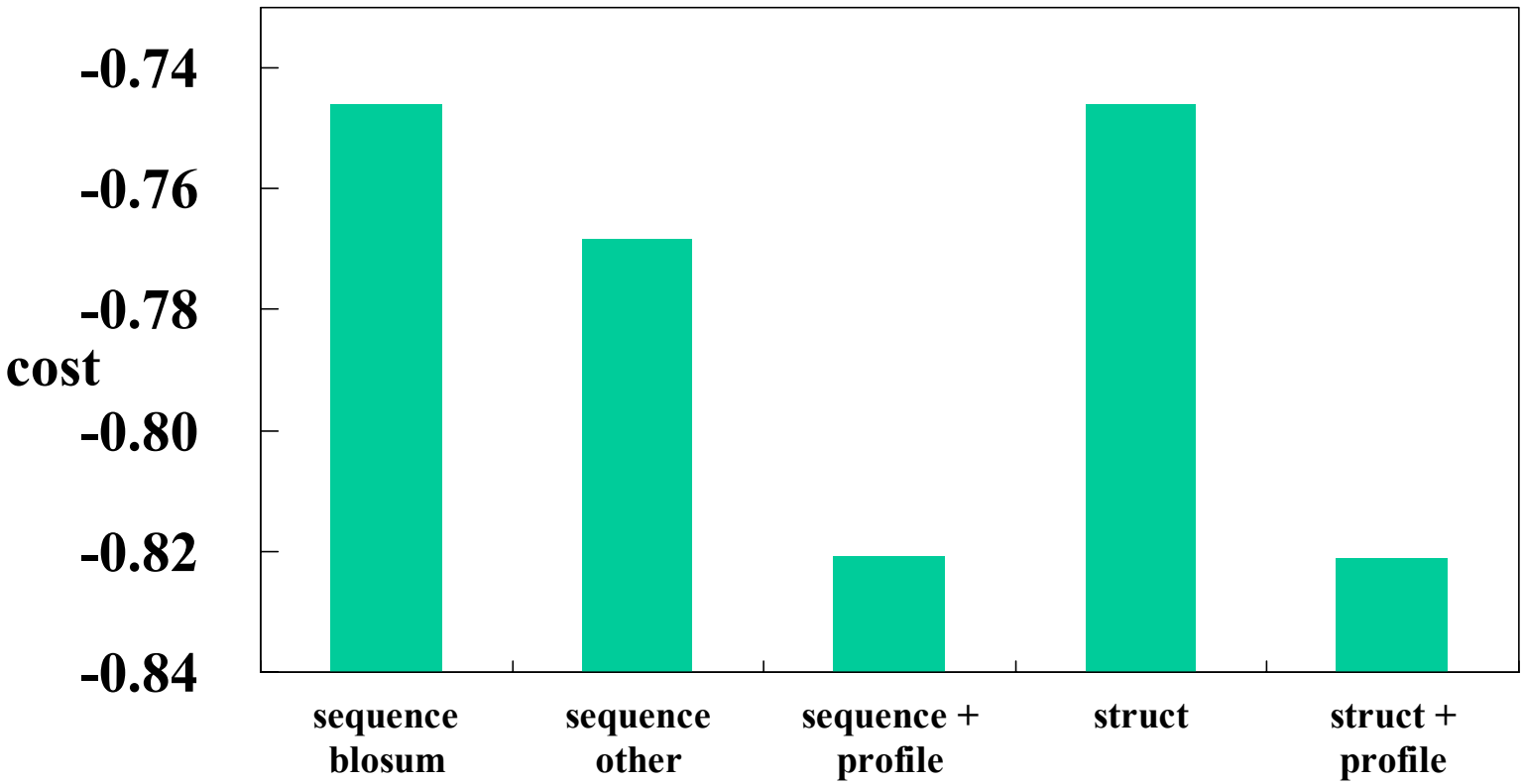
Fold recognition (sequence)

- biggest tangible usable advance
- textbooks
  - twilight zone of homology 20 – 25 %
- use of sequence profiles (psi-blast, HMMs)
  - routinely << 20 % reliable homologues
- example calculation

Altschul et al, Nucl. Acids Res. 25, 3389-3402, 1997

Sequence profiles

- 1500 x 2 alignments
  - quality of alignments



Fold recognition

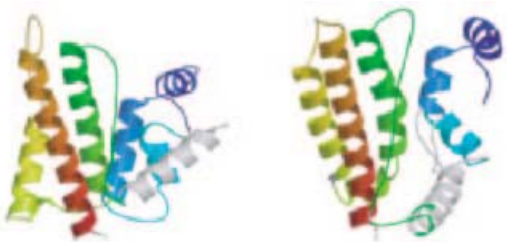
- belief
  - there is a finite number of protein folds
  - all we have to do is find the best for our sequence

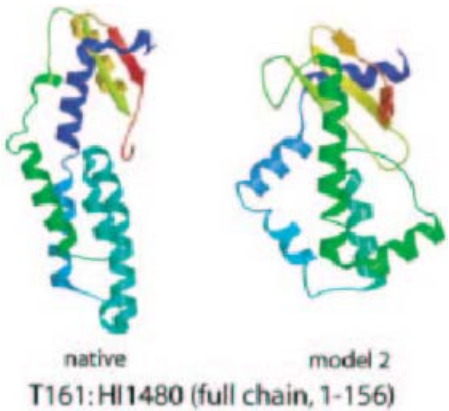
- limitations
  - breaks on unknown folds

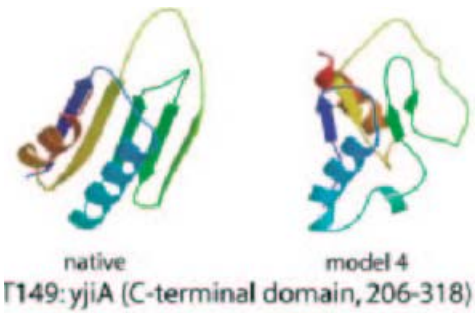
state of the art

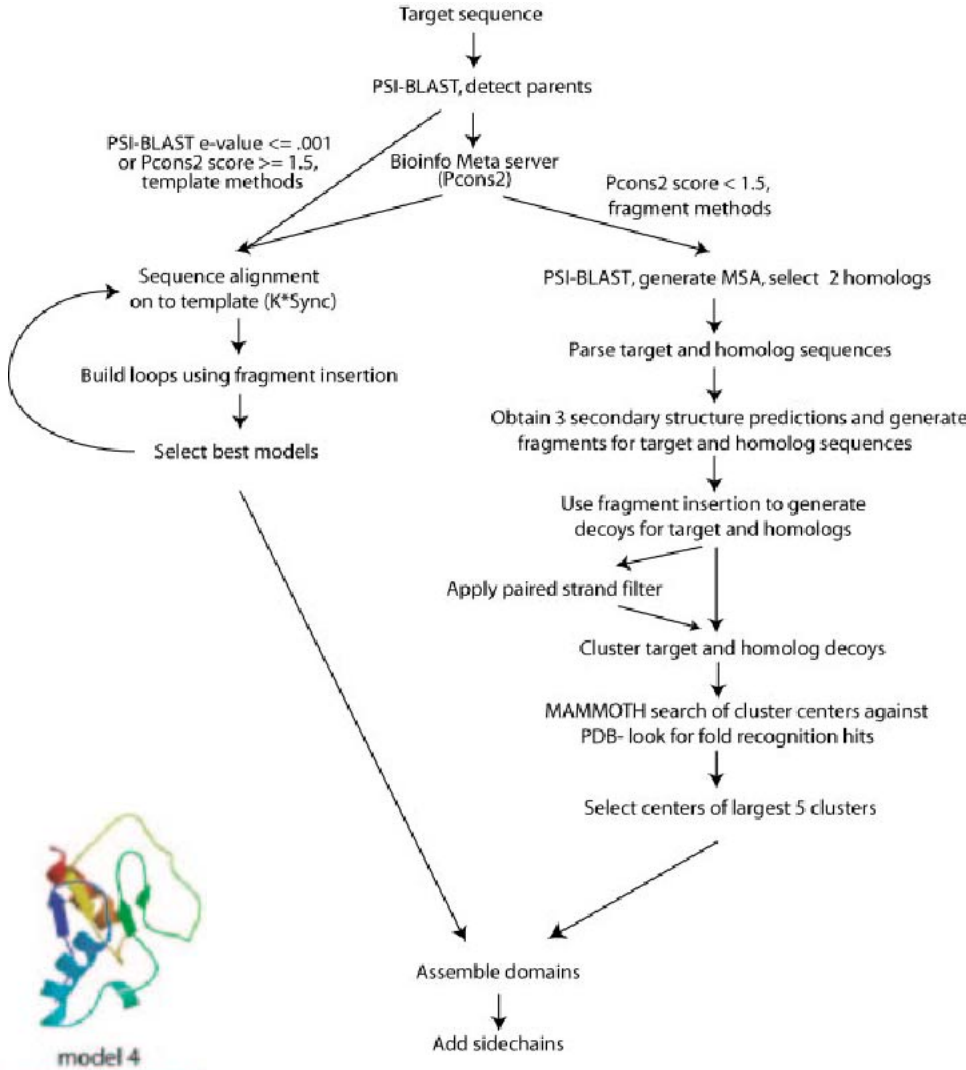
Best results ?

- Baker's fragment assembly
- automatic ?









Bradley, … Baker, Proteins, 53, 457-468 (2003)

Give up

Progress

- MD / bigger computers
- fold recognition
- occasional new structures
- parameterisation data
  - reliability and more complicated models

Signal to noise problem

- when are we happy ?

- noble goal
  - make calculators redundant