Computer Matchmaking in the Protein Sequence/Structure Universe

Thomas Huber

Supercomputer Facility
Australian National University
Canberra

email: Thomas.Huber@anu.edu.au

The ANU Supercomputer Facility

- A facility available to all members of the ANU
- Mission: support computational science through provision of HPC infrastructure and expertise
- Fujitsu collaboration at ANU
  - System software development
  - Mathematical subroutine library
  - Computational chemistry project
    - 5-6 persons
    - porting and tuning of basic chemistry code to Fujitsu supercomputer platforms
    - current code of interest
      - Gaussian98, Gamess-US, ADF
      - Mopac2000, MNDO94
      - Amber, GROMOS96

Resources

- Fujitsu VPP300 (vector processor)
  - 13 processors, 142 MHz (2.2 Gflop)
  - Distributed memory, 8*512MB, 5*2GB
  - crossbar interconnect, 570 MB/s
- SUN E3500
  - 8 processors, 400 MHz Ultra2 (800 Mflop)
  - 8 GB shared memory
- SGI PowerChallenge
  - 20 processors, 195 MHz R10k (390MFlop)
  - 2 GB shared memory
- alpha Beowulf cluster
  - 12+1 processors, 533Mhz alpha (1GFlop)
  - 256 MB memory per node
  - Fast ethernet connection, 12.5 Mb/s

Resources (cont.)

- Fujitsu AP3000 (“workstation cluster”)
  - 12 processors, 167 MHz Ultra2 (330Mflop)
  - 128 MB memory per node
  - Fast AP-Net (2D Torus), 200MB/s

- Future:
- ANU is host of APAC
  - ≈1 Tflop system
  - 300-500 processors

Protein Structure Prediction

- Basic choices in molecular modelling
- Why is fold recognition so attractive
- Basics of fold recognition
  - Representation
  - Searching
  - Scoring
- Special purpose sequence/structure fitness function
- How successful are we?
- How to do better

sequence

magic box

structure

Three basic choices in molecular modelling

- Representation
  - Which degrees of freedom are treated explicitly
- Scoring
  - Which scoring function (force field)
- Searching
  - Which method to search or sample conformational space

Why is fold recognition attractive?

- Conformational search problem notorious difficult
- searching in a library of known protein folds:
  - finding the optimum solution is guaranteed

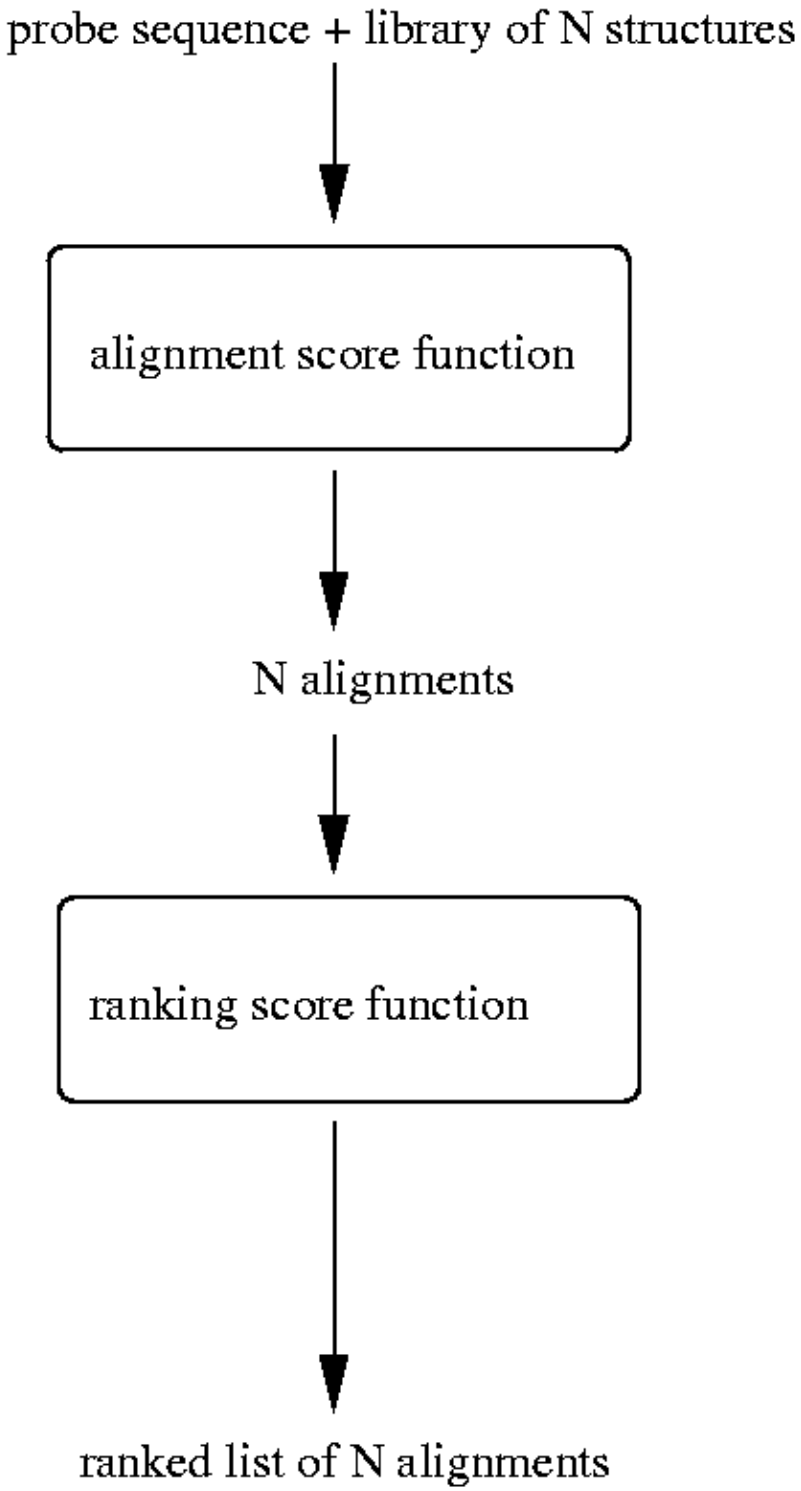
Is fold recognition useful?

- In how many ways do protein fold?
  - $\approx 10^4$ protein structures determined
  - $\approx 10^3$ protein folds

Fold Recognition = Computer Matchmaking

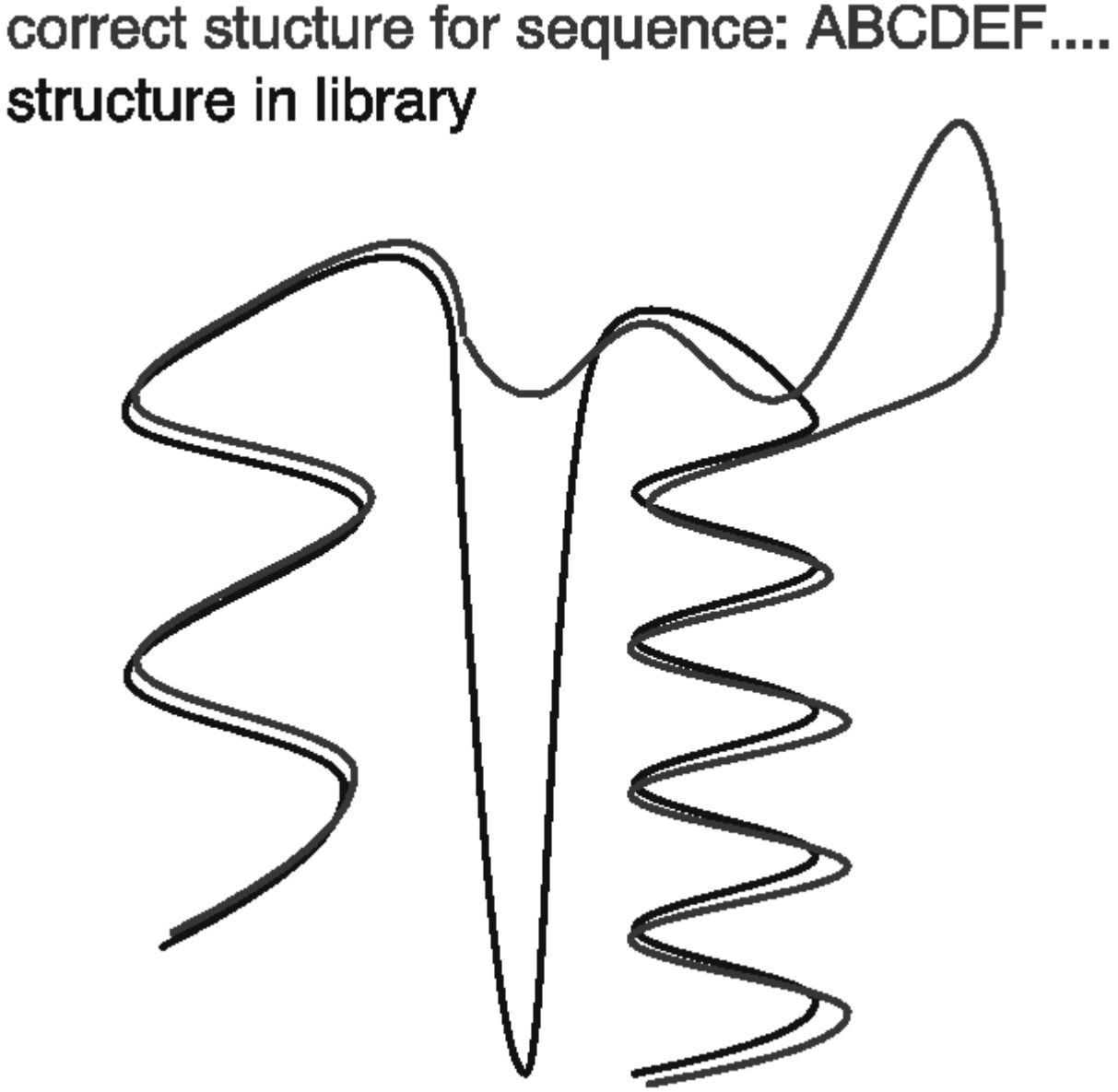
- Structure Disco

Sausage: 2 step strategy

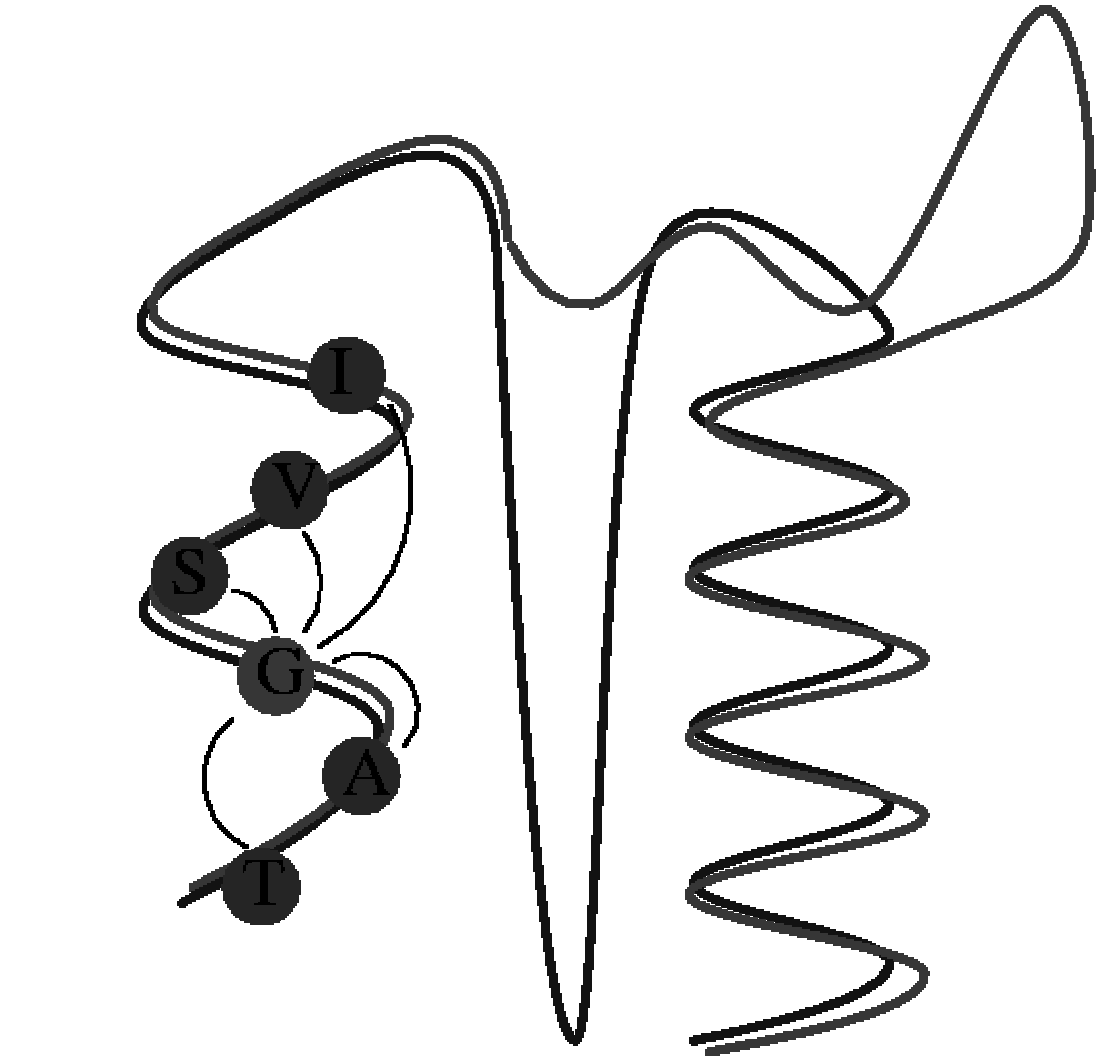


Sequence-Structure Matching
The search problem

- Gapped alignment = combinatorial nightmare



1. Double Dynamic Programming

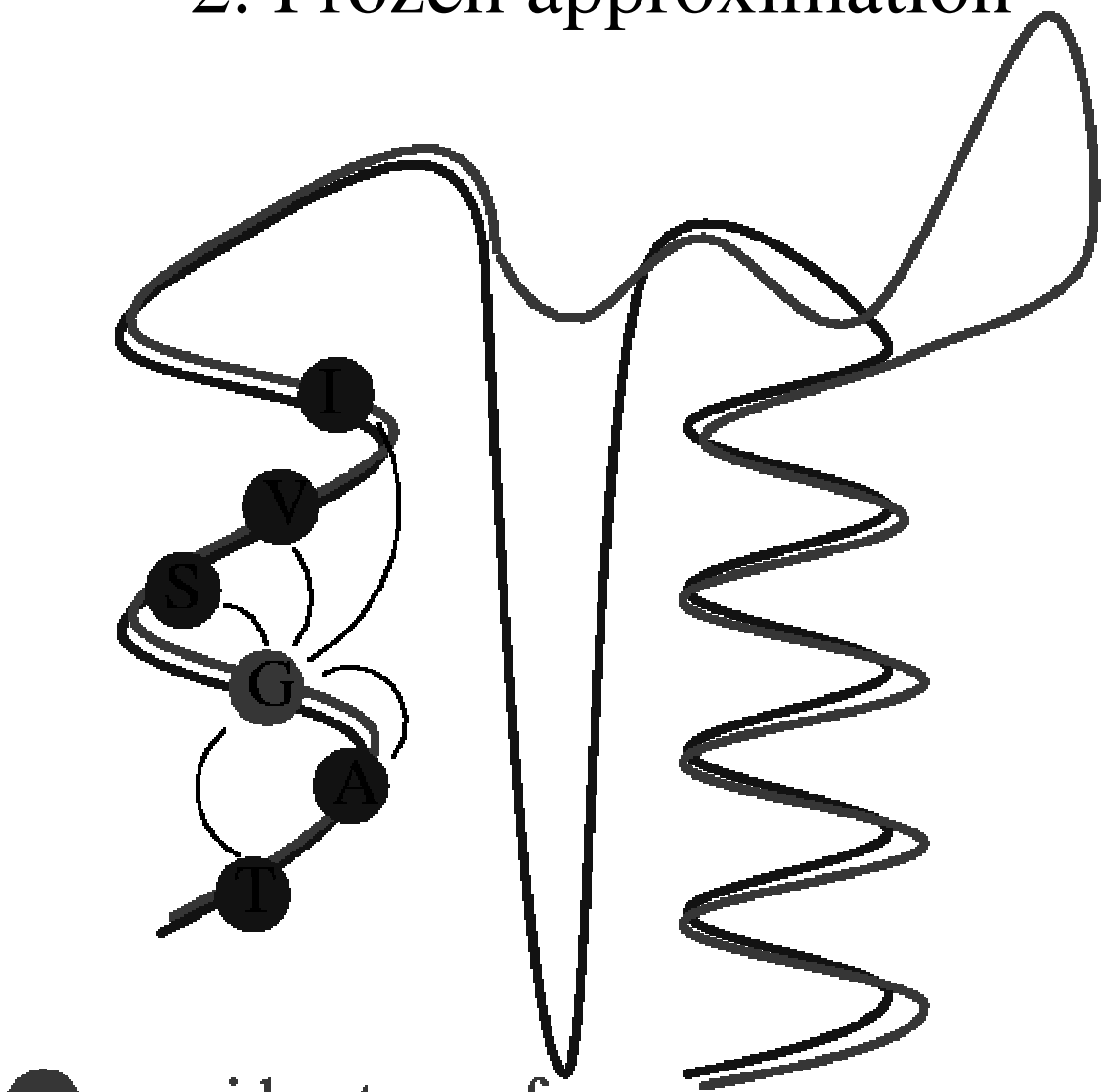


residue type of sequence

residue type of best neighbour

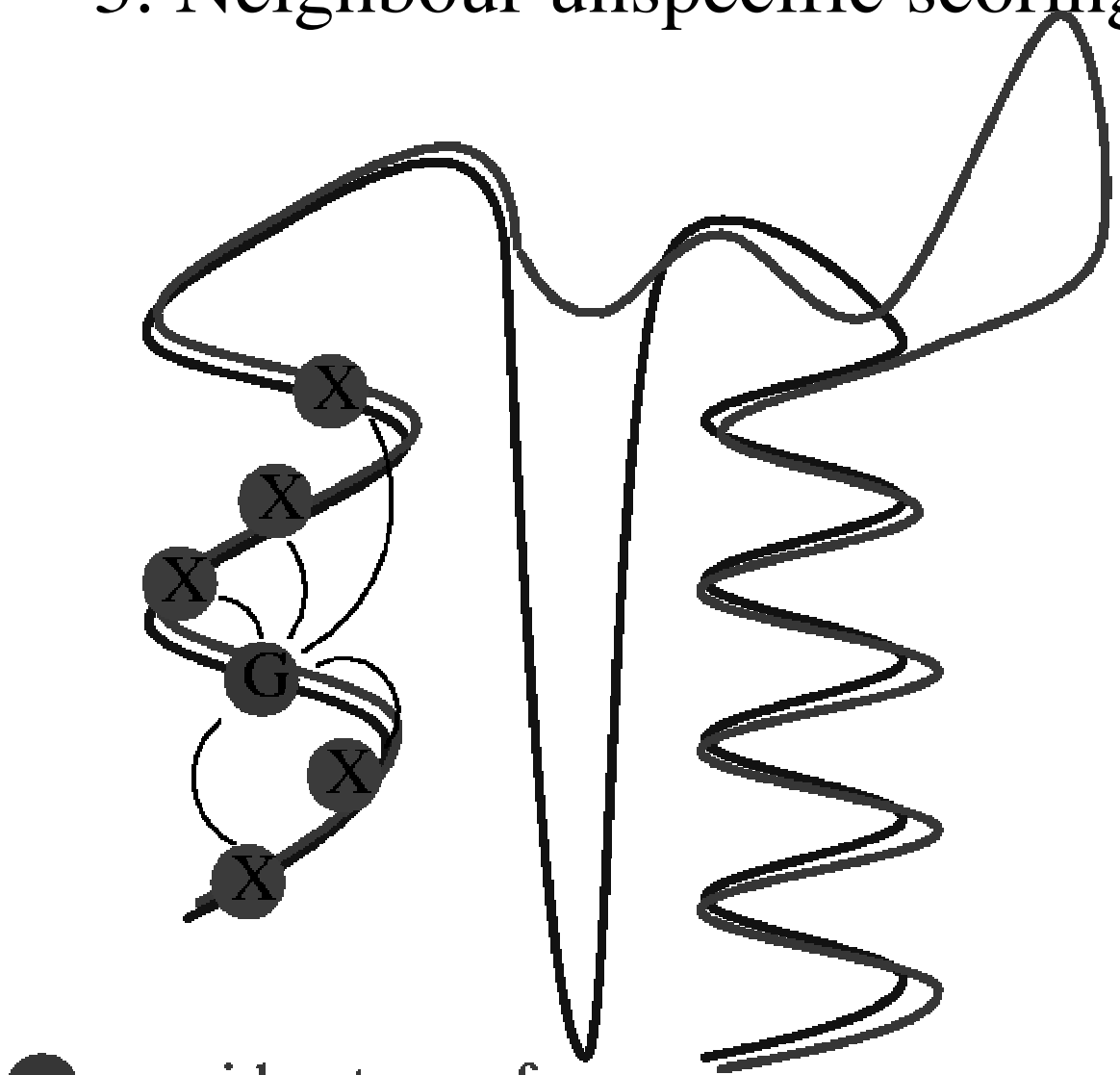
- **Advantage:** pair specific scoring
- **Disadvantage:** $O(N^5)$

2. Frozen approximation



- **Advantage:** pair specific scoring
- **Disadvantage:** Sequence memory from template

3. Neighbour unspecific scoring



residue type of sequence

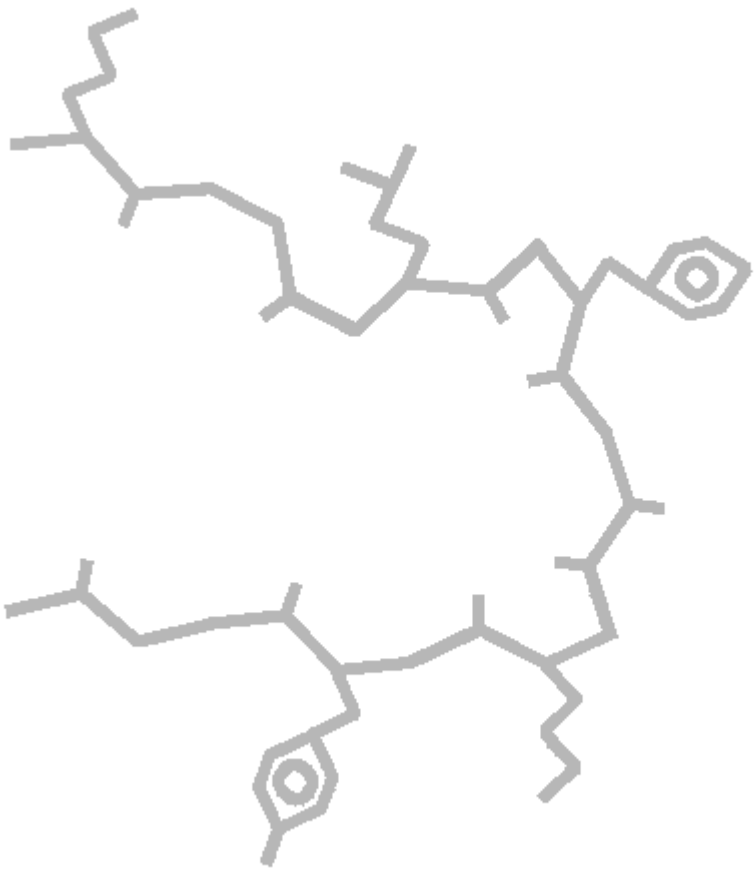
X residue with average properties (no identity)

- **Advantage:** no sequence memory from template

Model Representation

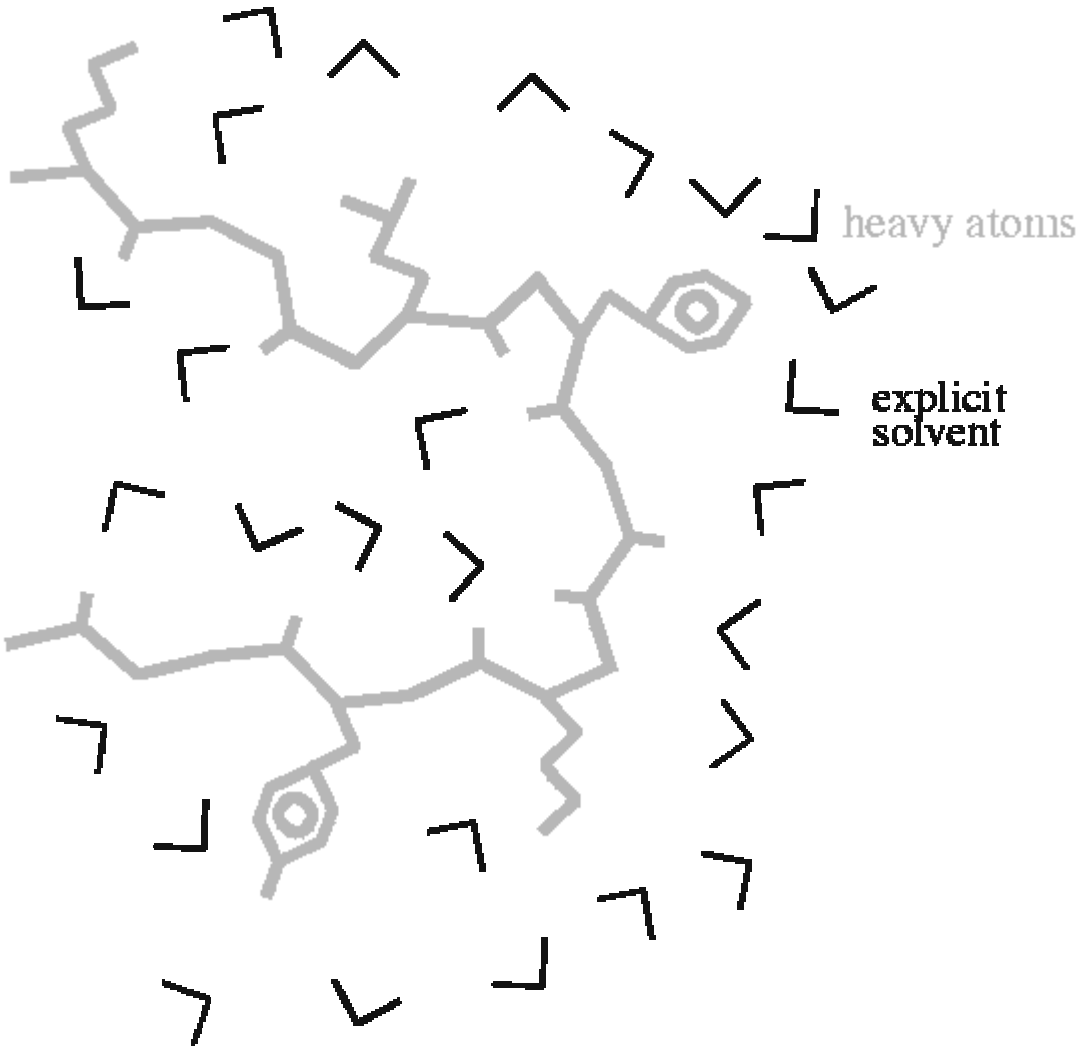
1. Conventional MM

   (structure refinement)

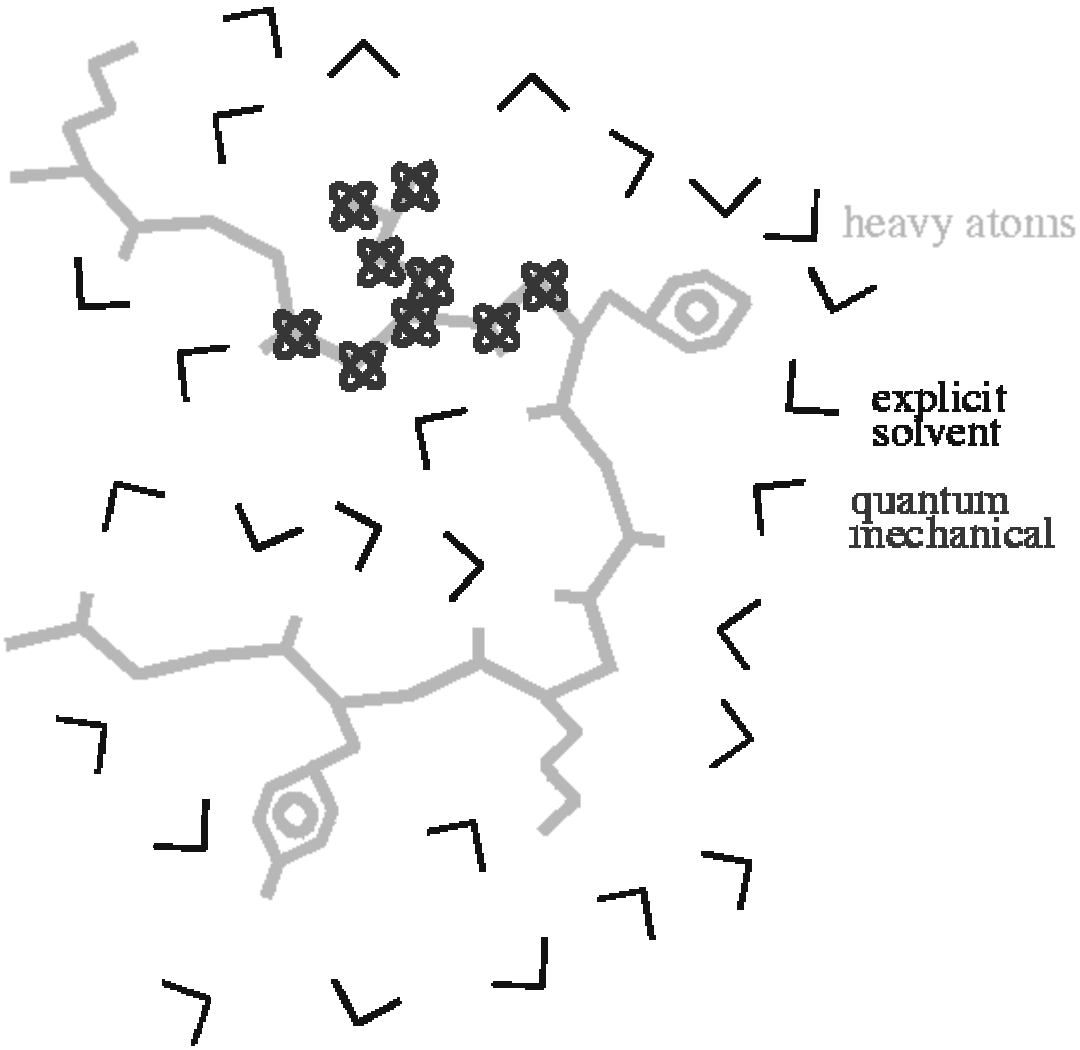


heavy atoms

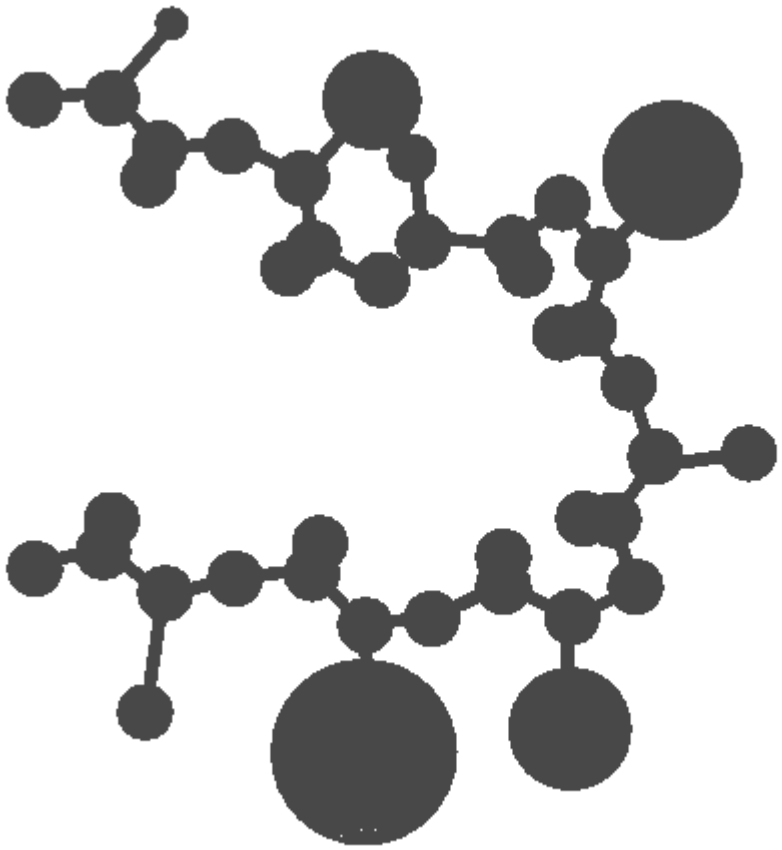
2. MM with solvation
(local dynamics)



3. QM with solvation
(enzyme reactions)



4. Low resolution
(structure prediction)



backbone +
1 sidechain

Scoring

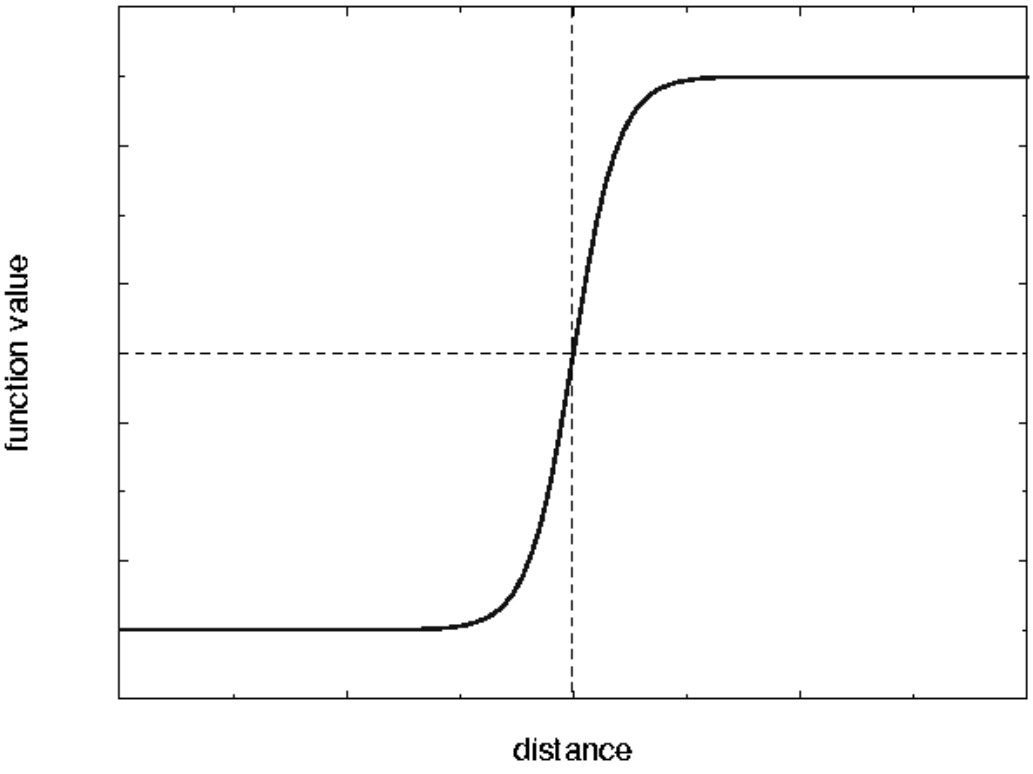
- Quality of prediction is given by

$$E = \sum_{ij} E_{ij}$$

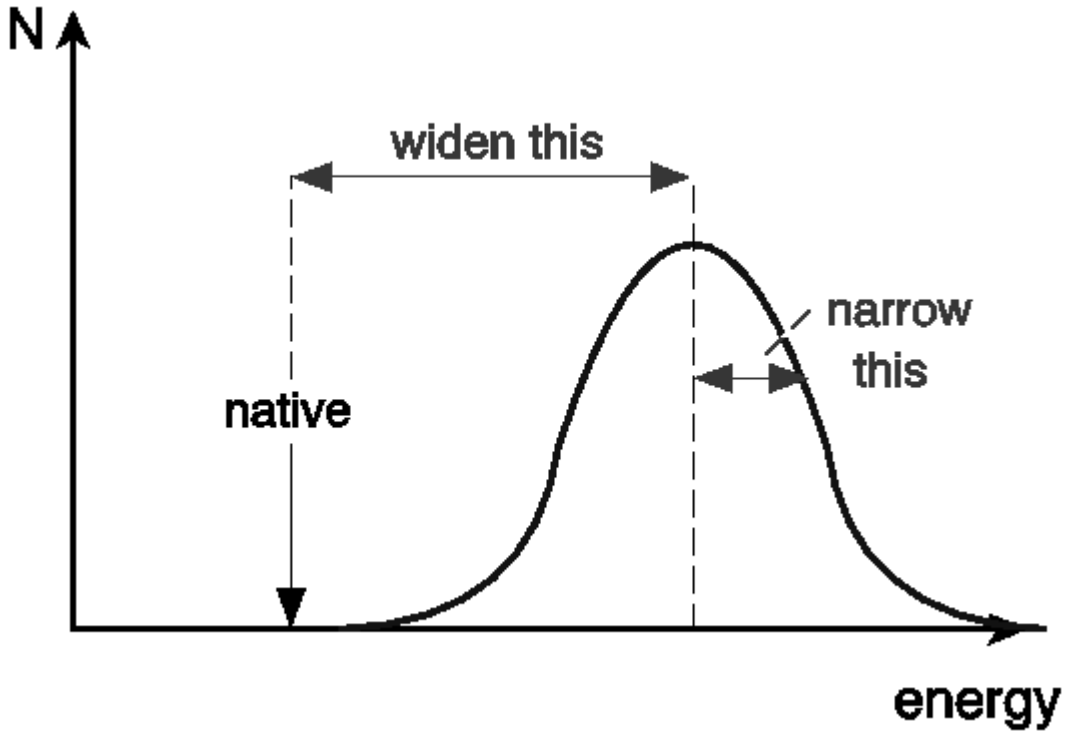
- Functional form of interaction
  - simple
  - continuous in function and derivative
  - discriminate two states

  $\Rightarrow$ hyperbolic tangent function

$$E_{ij} = k_{ij}[1 + \tanh(d - d_0)]$$



Parameterisation of Discrimination Function



- Gaussian distribution

$$N(E) \propto \exp\left(- \frac{(E - \langle E \rangle)^2}{2\sigma^2} \right)$$

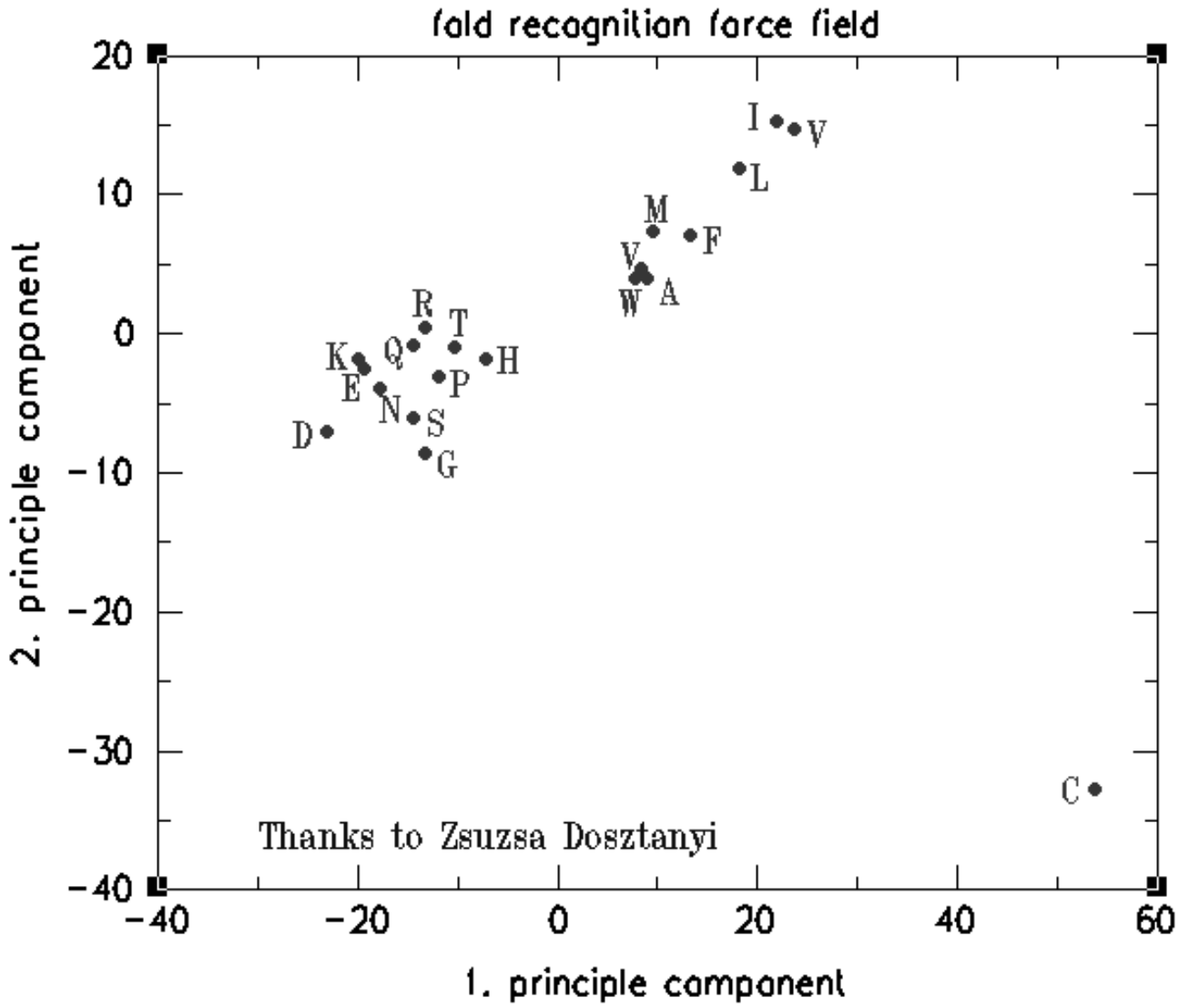
$$\text{z-score} = \frac{E - \langle E \rangle}{\sigma}$$

$\Rightarrow$ Minimisation of z-score with respect to parameters

Size of Data Set

- 893 non-homologous proteins
  - < 25% sequence identity
  - 30-1070 amino acids
- $>10^7$ mis-folded structures
- 996 force field parameters
  - parameters well determined

Is Our Scoring Function Totally Artificial?



Does it work?

- Blind test of methods (and people)
  - methods always work better when one knows answer
- ≈30 proteins to predict
- ≈90 groups (≈40 fold recognition)
  - Torda group one of them
  - All results published in Proteins, Suppl. **3** (1999).

Fold Recognition
Official Results
(Alexin Murzin)

Grp	Name	Score
005	Jones (1)	5| 17
212	Bryant (2)	5| 17
166	SB-Fold (3)	3| 12
217	Sippl (4)	3| 11
176	Unagi (5)	3| 09
066	Fischer (6)	2| 11
040	Olszewski (7)	2| 11
201	Tatsuya (8)	2| 07
009	Xu-Ying (9)	2| 07
033	Elofsson (10)	2| 05
156	Benner-Cohan (11)	2| 03
085	Park (12)	1| 06
035	Baker (13)	1| 06
003	Hubbard (14)	1| 06
142	Taylor (15)	1| 05
061	Yang (16)	1| 04
074	Sternberg (17)	1| 04
028	Godzik (18)	1| 04
045	Torda (19)	1| 04
147	BMERC (20)	1| 02

Fold Recognition Predictions Re-evaluated

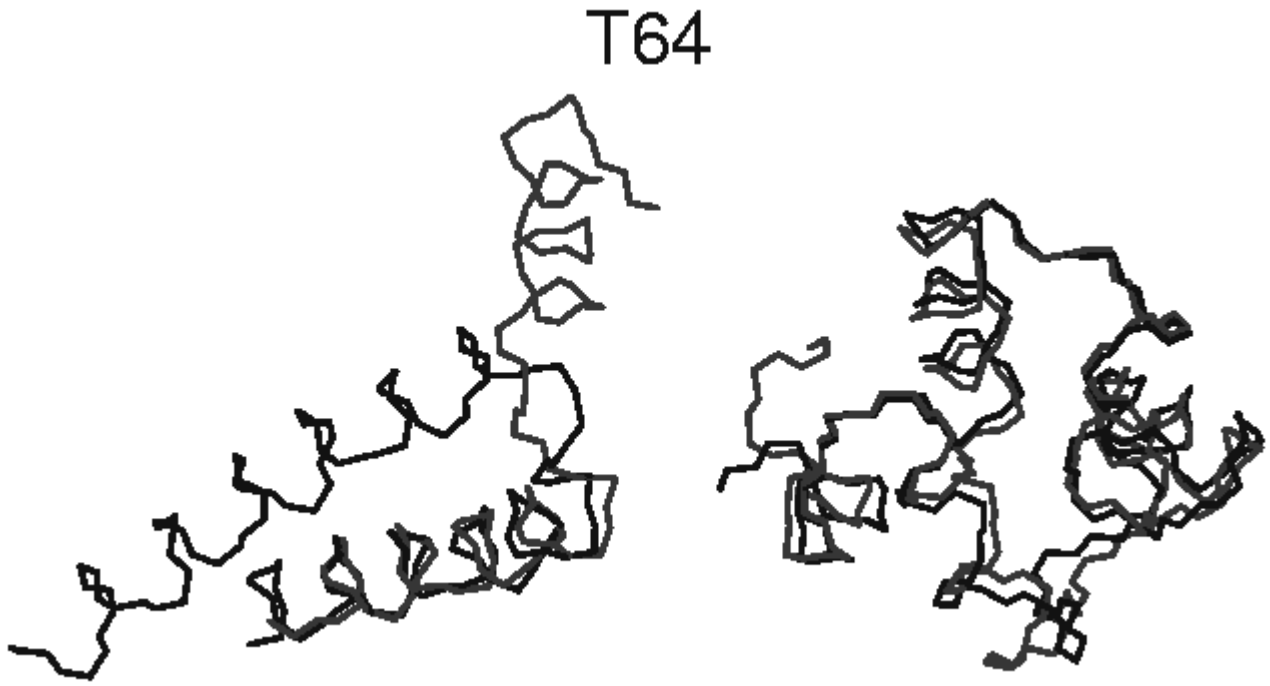
(computationally by Arne Elofsson)

- Investigation of 5 computational (objective) evaluations
- Comparison with Murzin's ranking

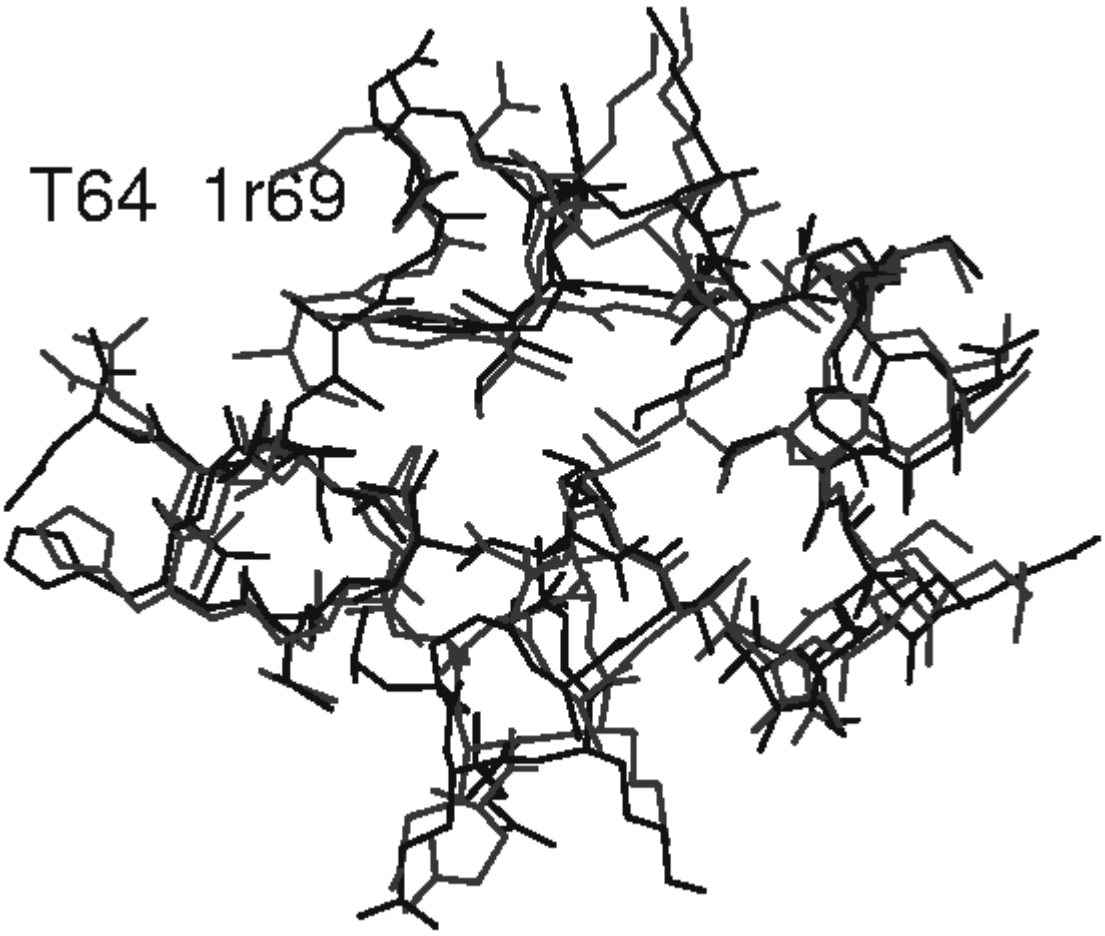
Group	Murzin	GDT	MAXsub	LG	FR	NoRes	Score
Sippl	4	1	2	1	1	3	3.75
Jones	1	2	1	3	3	1	3.16
Torda	19	9	5	2	4	10	1.11
Baker	13	3	9	5	5	5	0.92
Bryant	2	15	12	10	11	12	0.84
SB-fold	3	10	10	6	13	6	0.77
Valencia	23	25	20	8	2	9	0.75
Unagi	5	12	17	4	12	4	0.67
Godzik	18	6	3	30	15	29	0.65
Yang	16	5	4	13	18	7	0.64

CASP3 Example

- 31% sequence identity

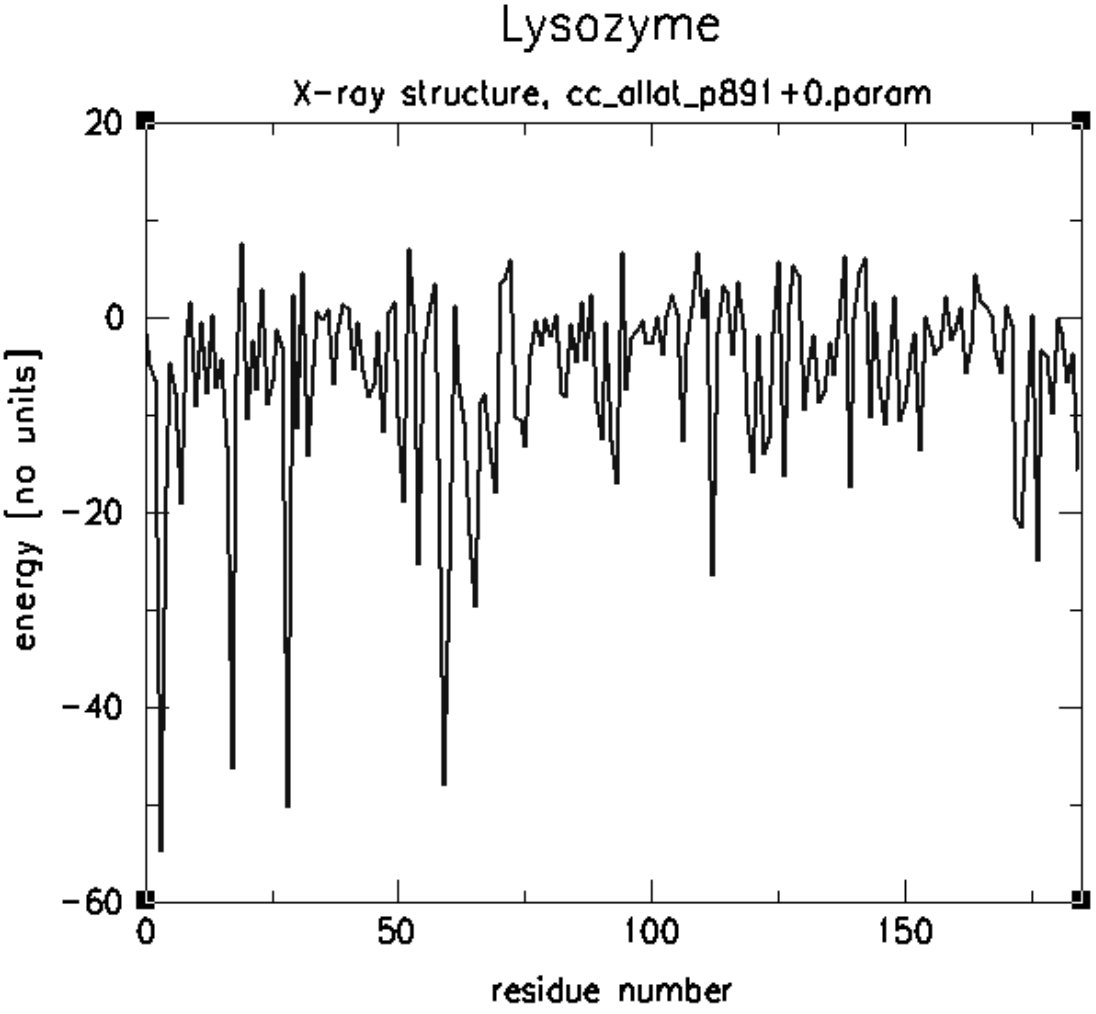


CASP3 Example



Improvements to Fold Recognition

- Noise vs signal



- Average profiles (Andrew Torda)
- Optimised Structures

Structure Optimisation

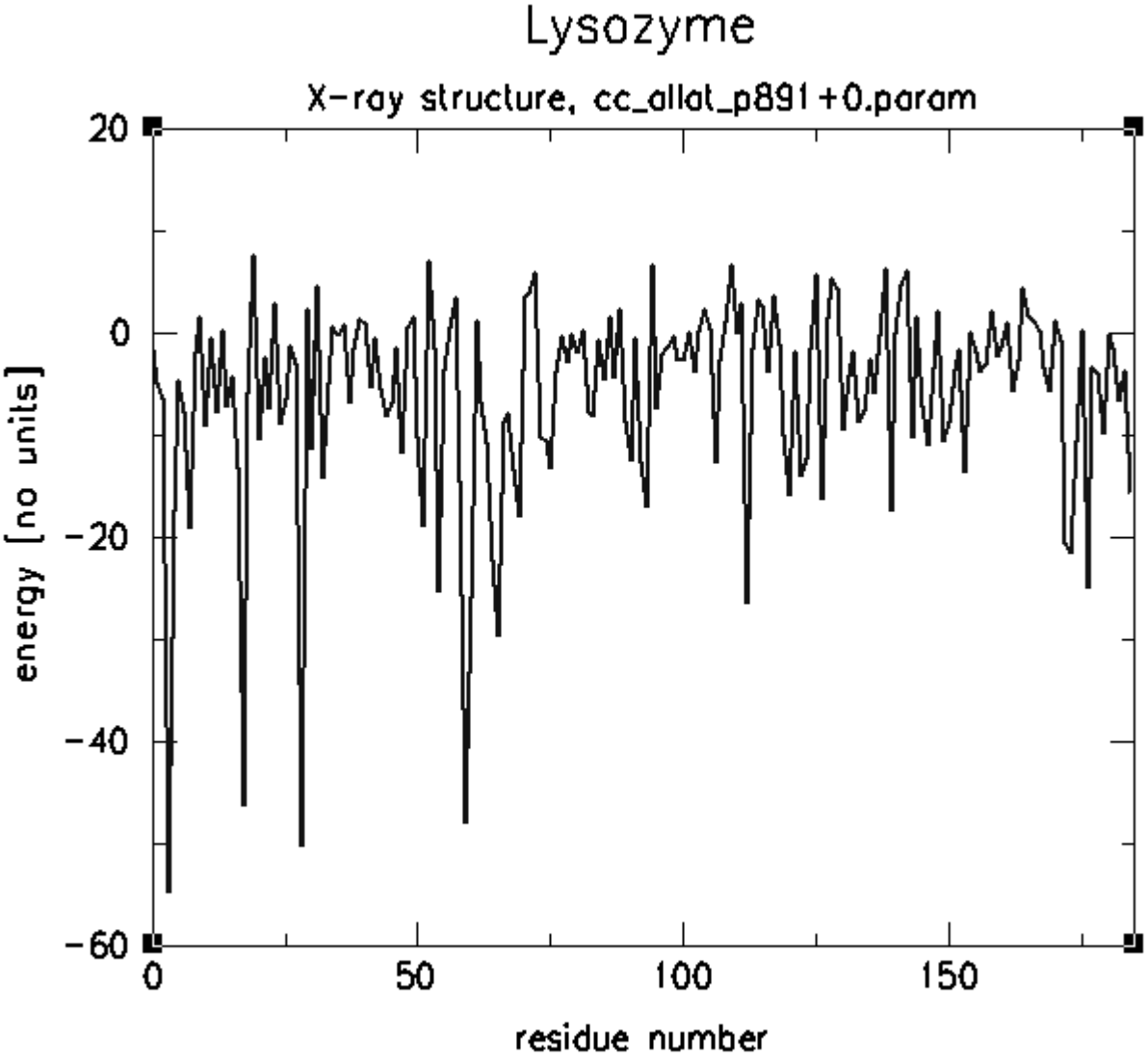
- X-ray structures
  - high (atomic) resolution, fit 1 sequence
- Structure for fold recognition
  - low resolution (fold level)
  - should fit many sequences

$\Rightarrow$Optimise structures for fold recognition

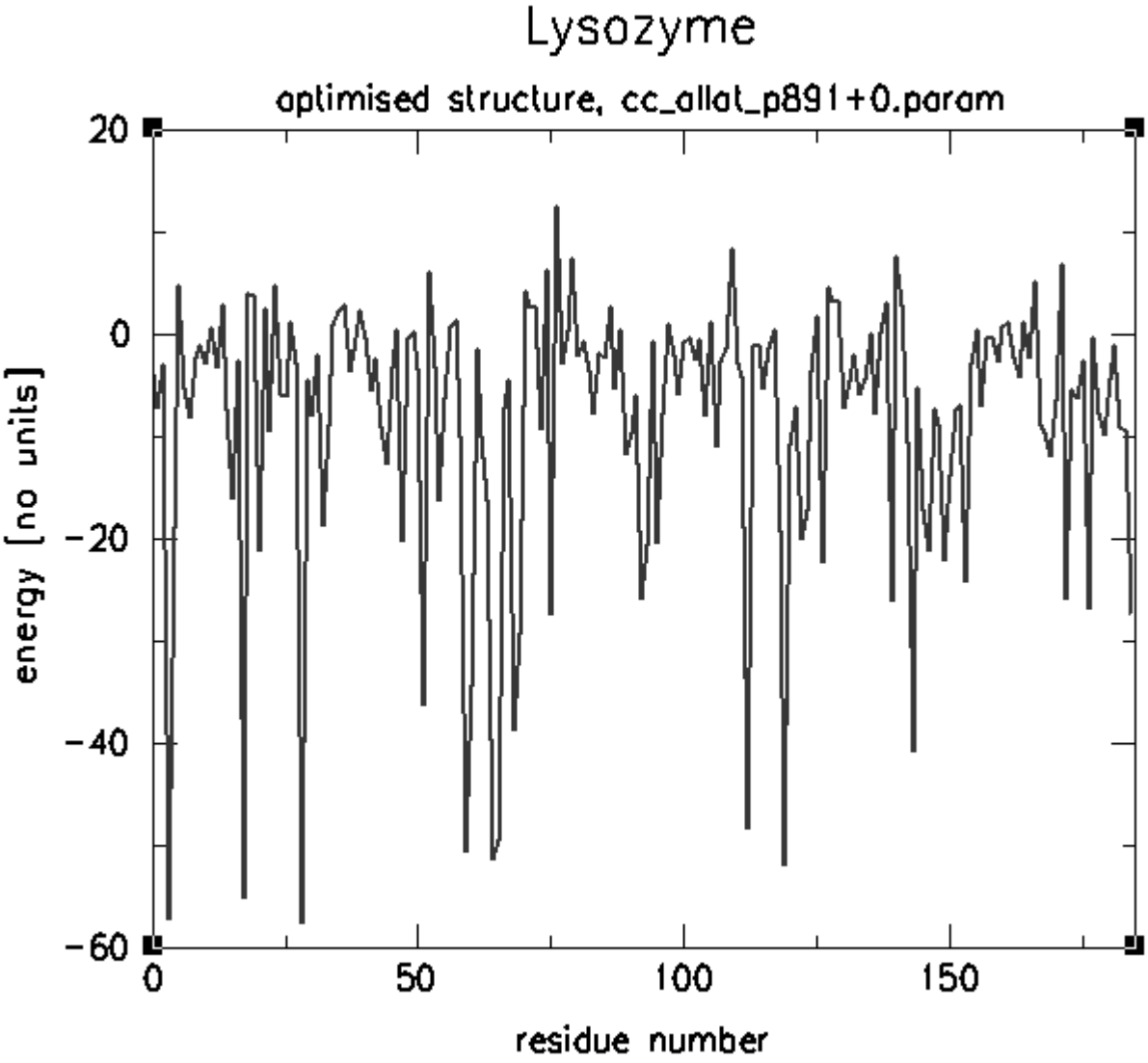
How are Structures Optimised?

- Goal:
  - NOT to minimise energy of structure
  - BUT increase energy gap between correct alignments and incorrectly aligned sequence
- Deed:
  - 20 homologous sequences (<95%)
  - 20 best scoring alignments from (893) “wrong” sequences
  - change coordinates to maximise energy gap between “right” and “wrong”
    - 100 steps energy minimisation
    - 500 steps molecular dynamics
- Hope:
  - important structural features are (energetically) emphasised

Old Profile


Lysozyme
X-ray structure, cc_allat_p891+0.param
20
0
-20
-40
-60
energy [no units]
0
50
100
150
residue number

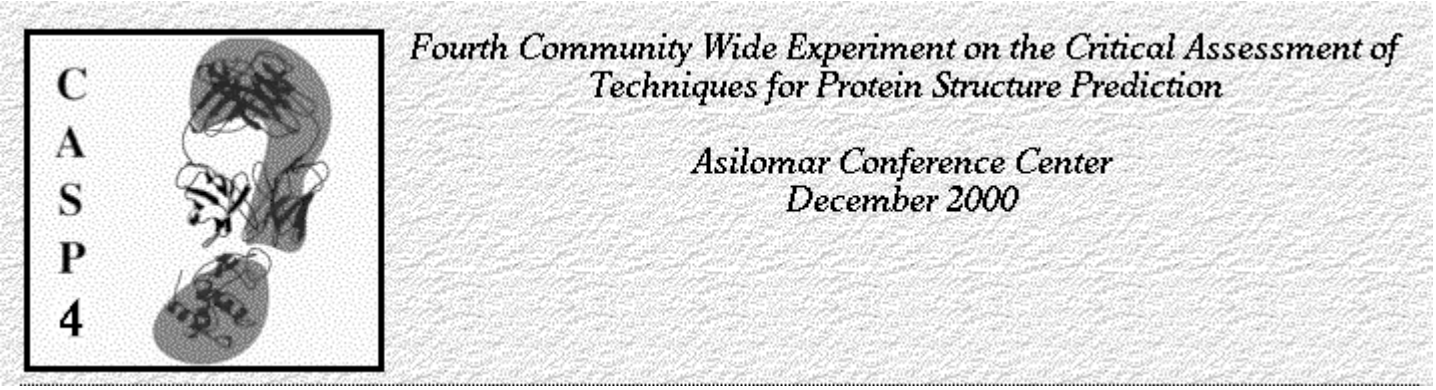
New Profile


Lysozyme
optimised structure, cc_allat_p891+0.param
energy [no units]
20
0
-20
-40
-60
0
50
100
150
residue number

More Information about Structure

- Predicted secondary structure
  - highly sophisticated methods
  - secondary structure terms not well reproduced by force field
  - easy to combine
- Sequence correlation
  - can reflect distance information
  - yet untested (by us)

What next?



- CASP4 (just announced)
  - Leap frog or being frogged?

- Stay tuned!

People

- At RSC
  - Andrew Torda
  - Dan Ayers
  - Zsuzsa Dostyani
- At ANUSF
  - Alistair Rendell

Want to try yourself?

- Sausage package freely available

  http://rsc.anu.edu.au/~torda

  or

  Thomas.Huber@anu.edu.au

Design of “better” proteins

- How to make more stable proteins?
  - Industrially very important
- How to design sequences which fold into a pre-defined structure?

Naïve Approach:

- Use physical force field
- Calculate energy difference of sequences

Why does this fail?

- Free energy all important measure

Why is it Hard to Calculate Free Energies?

- Free energy = ensemble weighted energy

$$F(N,V,T) = -k_B T \ln\left\langle \exp(+H / k_B T) \right\rangle$$

with ensemble average

$$\left\langle \exp(+H / k_B T) \right\rangle = \iint dp dr \exp(+H / k_B T)\pi(p,r)$$

$$\pi(p,r) \propto \exp(-H / k_B T)$$

$\Leftrightarrow$ delicate balance between contributions from high energy and low energy conformations

Model Calculations on a Simple Lattice

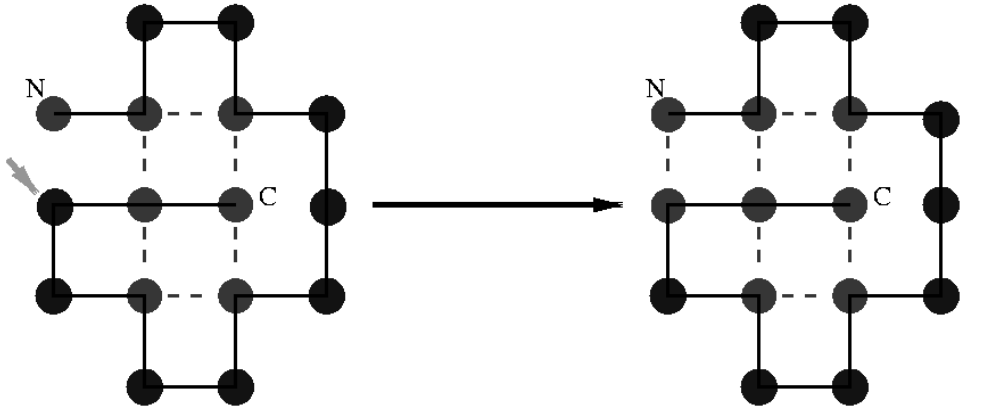
- Explore model "protein" universe
  - Square lattice
  - Simple hydrophobic/polar energy function (HH=1, HP=PP=0)
  - Chains up to 16-mers

  $\Rightarrow$ evaluation of all conformations (exact free energy)

  $\Rightarrow$ for all possible sequences

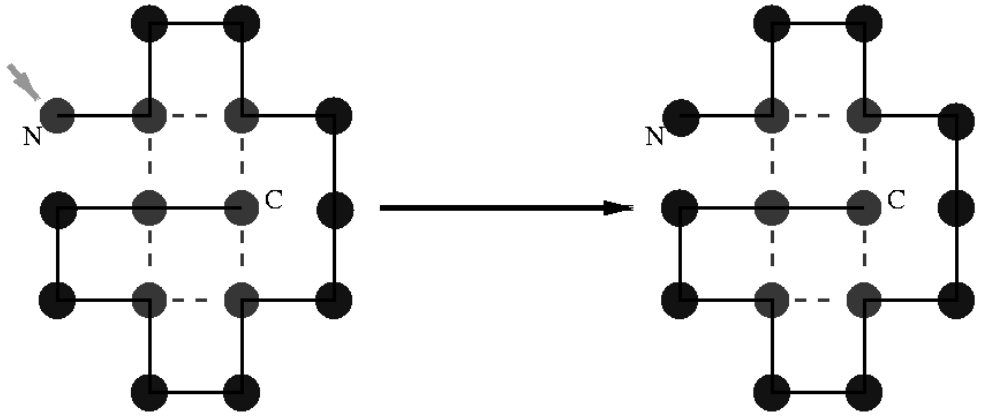
- "Our small universe"
  - 802074 self avoiding conformations
  - $2^{16} = 65536$ sequences
  - 1539 (2.3%) sequences fold to unique structure
  - 456 folds
  - 26 sequences adopt most common fold

Effect of sequence mutations



E = -6
F = -0.65981

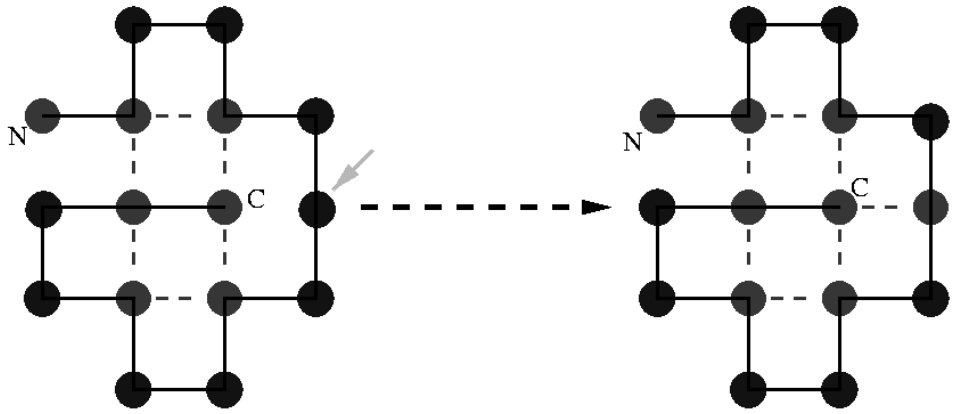
E = -7
F = -0.82074



E = -6
F = -0.65981

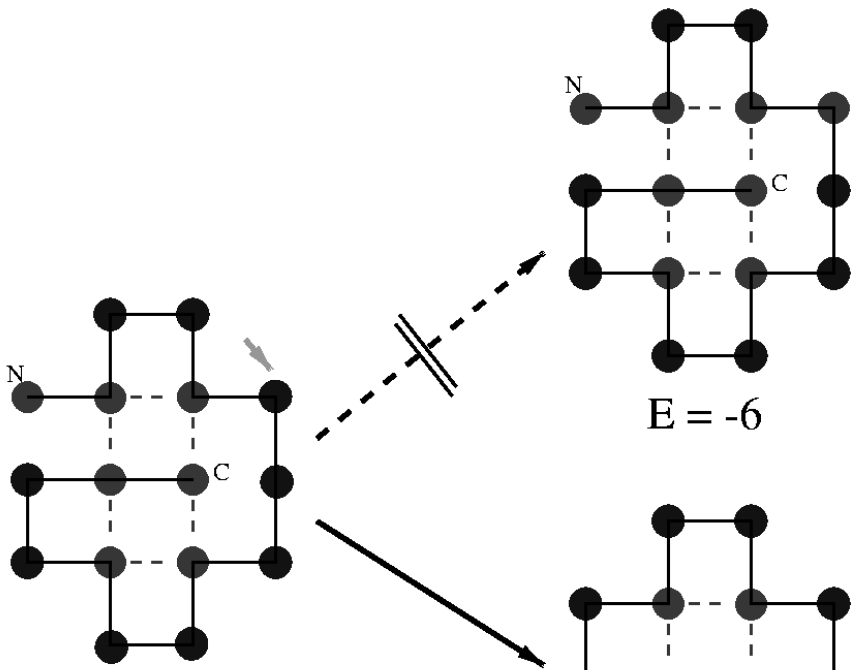
E = -6
F = -0.83885

Pitfalls



E = -6
F = -0.65981

E = -7
F = 0.00005



E = -6

E = -6
F = -0.65981

E = -7
F = -0.68215

Free energy approximation

- Question: Is there a simple function which approximates free energies