# Protein Structure Prediction: On the Cusp between Futility and Necessity?

Thomas Huber

Supercomputer Facility
Australian National University
Canberra

email: Thomas.Huber@anu.edu.au

# The ANU Supercomputer Facility

- Mission: support computational science through provision of HPC infrastructure and expertise
- ANU is host of APAC
  - >1 Tflop (300-500 processors by 2002)
  - first machines now up and running
- Fujitsu collaboration at ANU
  - System software development
  - 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

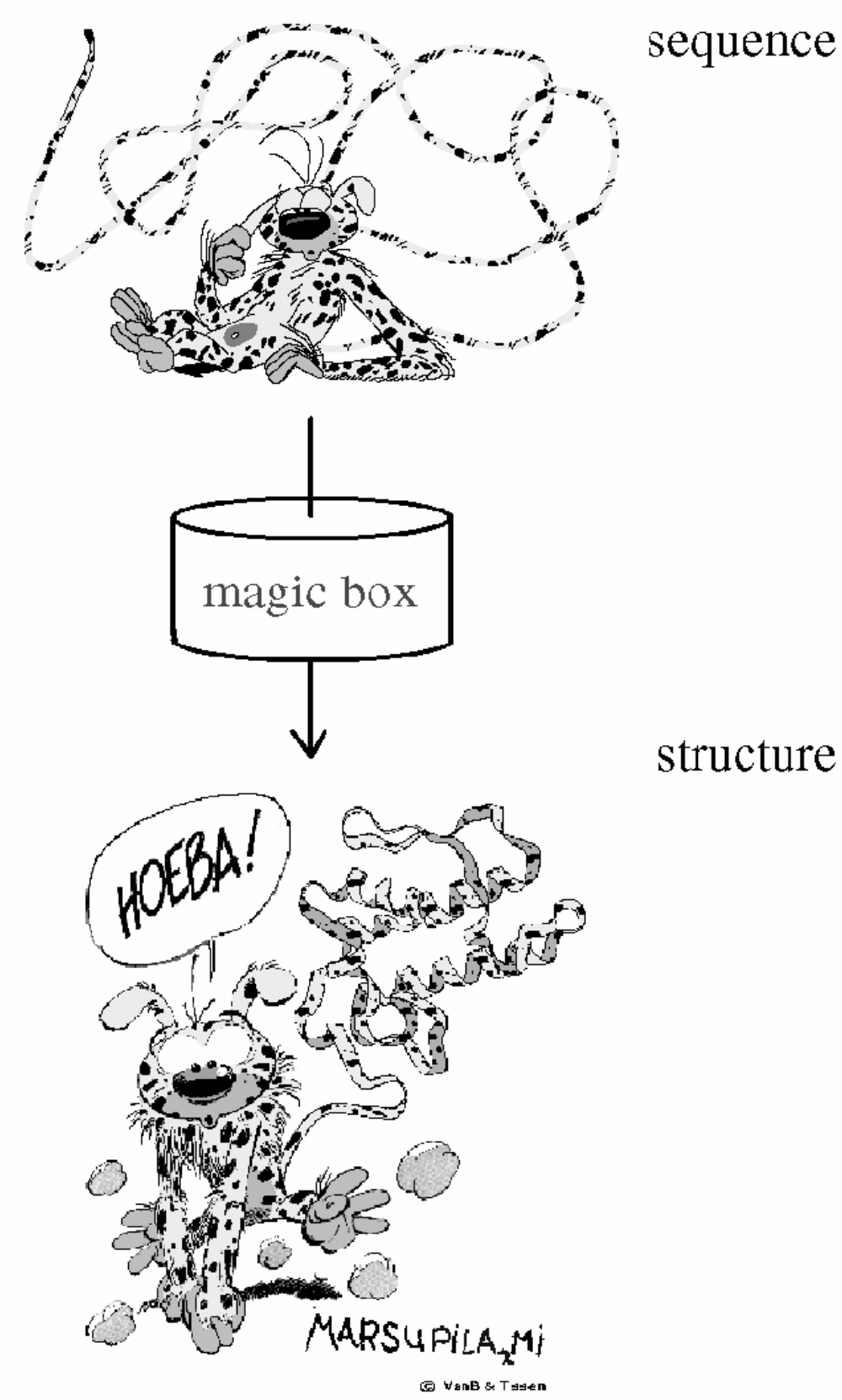
# My work

- Fujitsu collaboration
  - Responsible for MD software
    - porting and tuning to Fujitsu Supercomputer platforms
  - Collaboration with The Institute for Physical and Chemical Research (Riken), Japan.
    - Riken designed purpose specific hardware for MD simulation
      - MD-machine >1Tflop sustained performance (20 Gflop per chip)
      - Gorden Bell prize finalist (best performance for money)
    - We wrote biomolecular simulation software
- Research
  - Protein structure prediction

# Today's talk

- Something old
  - Protein structure prediction
  - Basics of protein fold recognition
  - How to build a low resolution force field
- Something new
  - How to improve fold recognition
  - Performance assessment
- Something for the future
  - Where is fold recognition useful
  - Perverting the concept of fold recognition
- Something new (for future work)
  - Model calculations

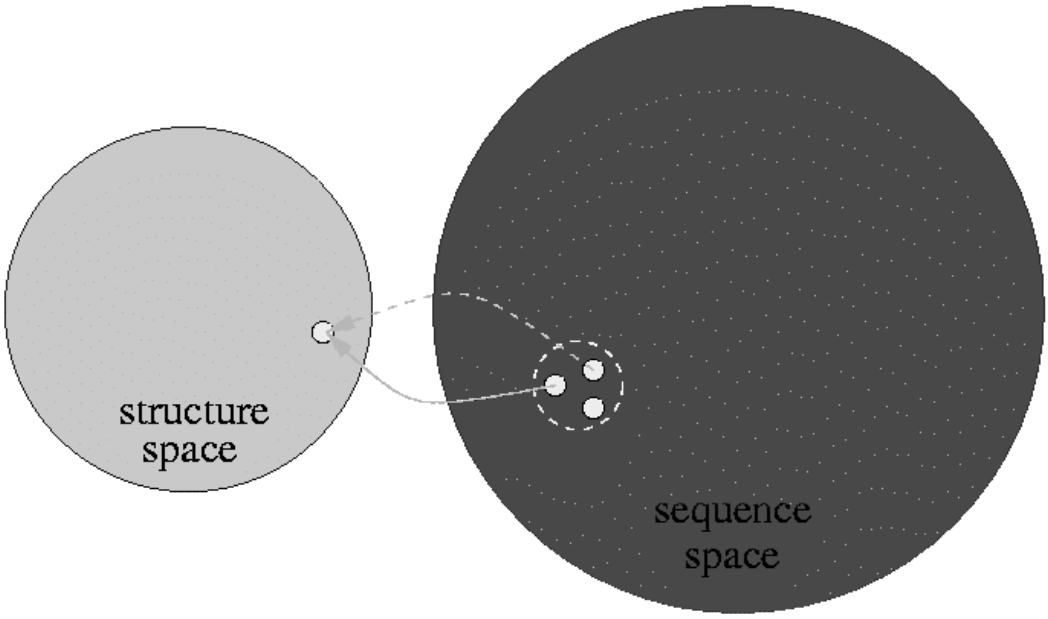
# Protein Structure Prediction



# Two Approaches

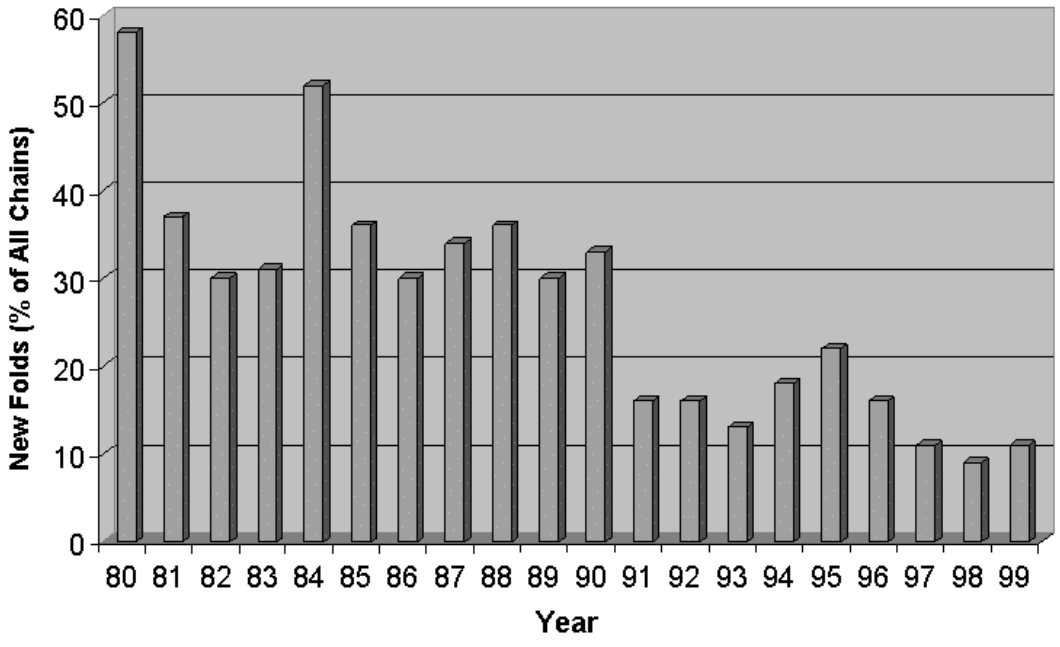
- Direct (ab initio) prediction
  - Thermodynamics: Structures with low energy are more likely

- Prediction by induction



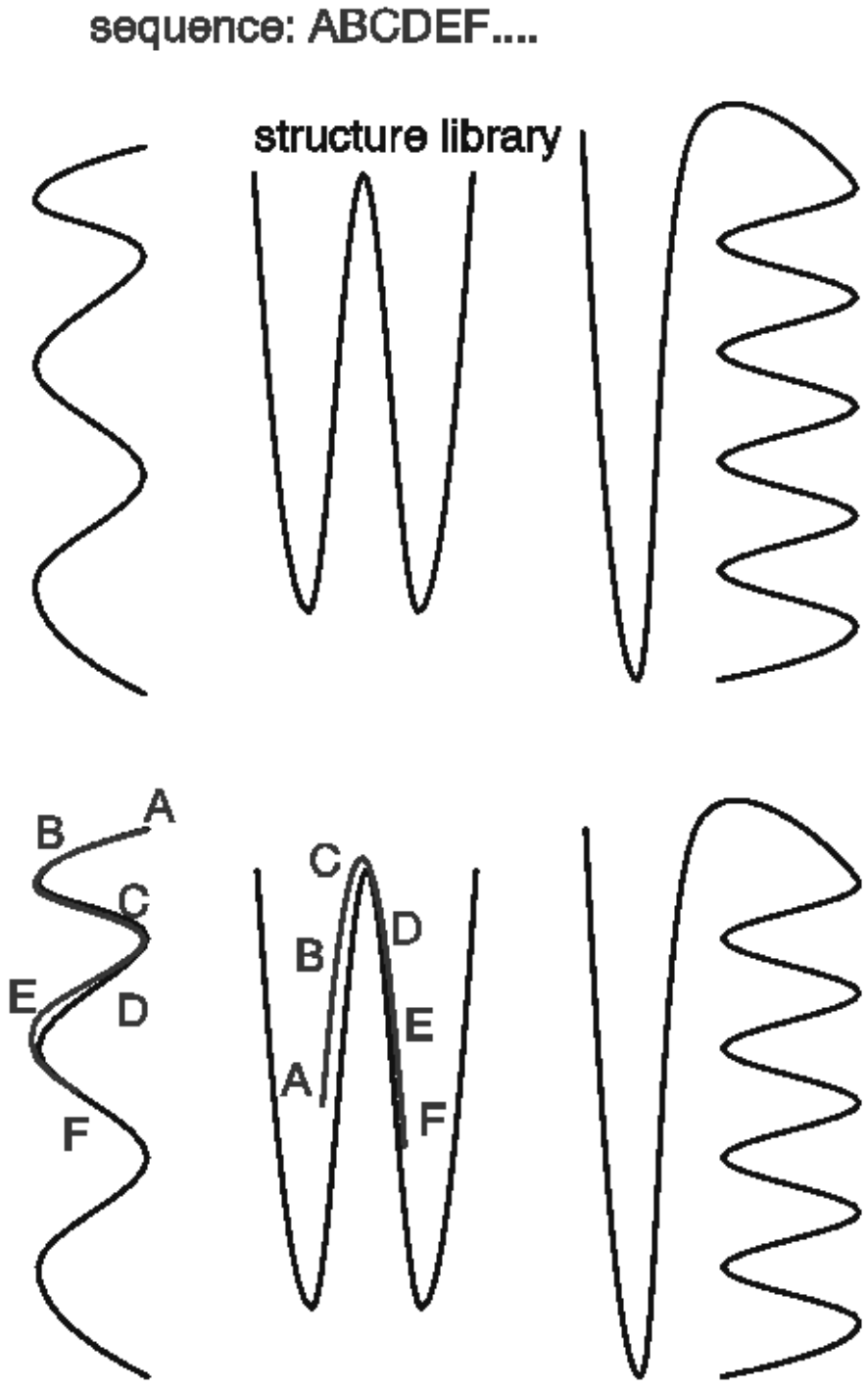
# Fold recognition

- More moderate goal:
  - Recognise if sequence matches a protein structure
- Why is fold recognition attractive?
  - Search problem notorious difficult
  - Searching in a library of known folds:
    - finding the optimum solution is guaranteed
- Is this useful?
  - $\approx 10^4$ protein structures determined
  - $< 10^3$ protein folds

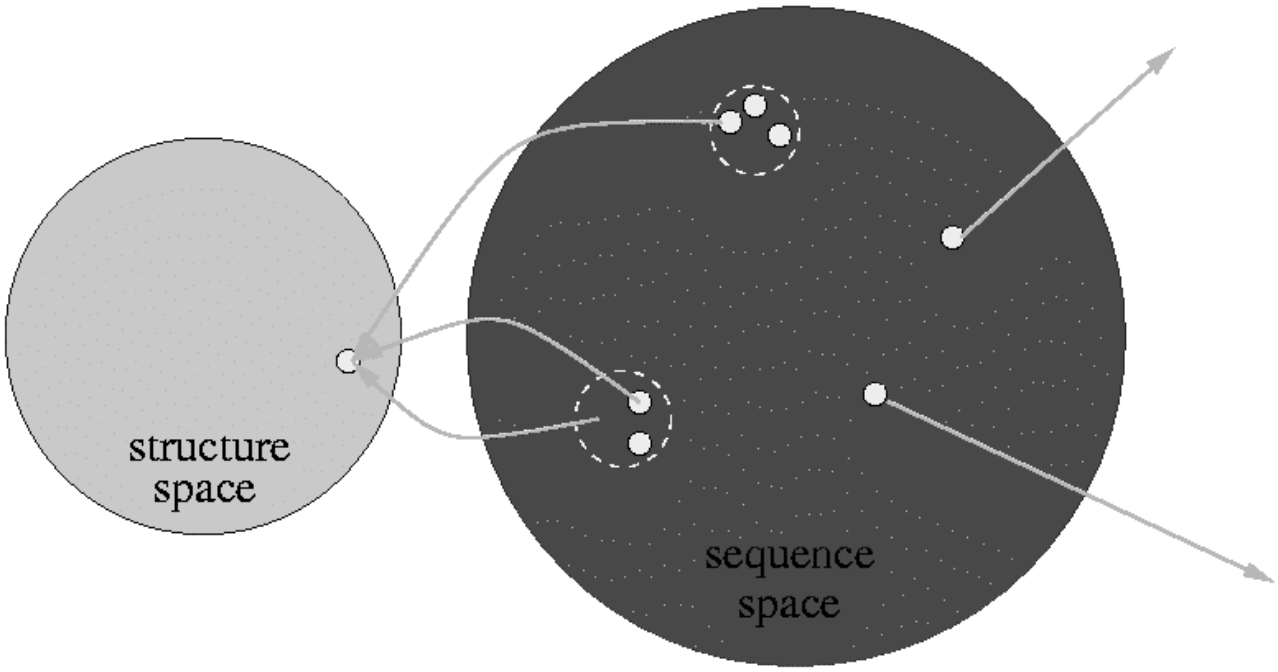


# Fold Recognition = Computer Matchmaking

- Structure Disco

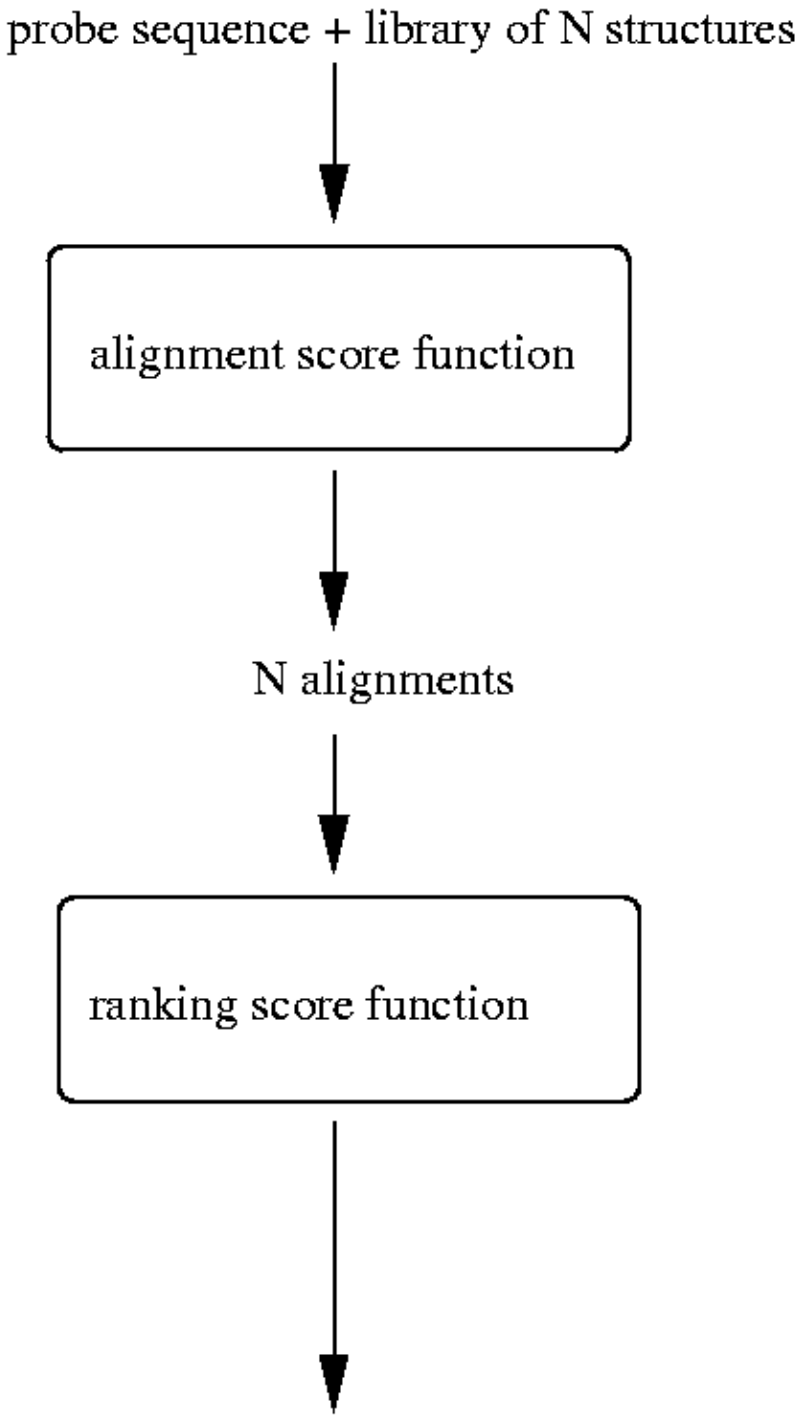


# Why is Fold Recognition better than Sequence Comparison?




- Comparison is done in structure space not in sequence space

# Sausage: 2 step strategy

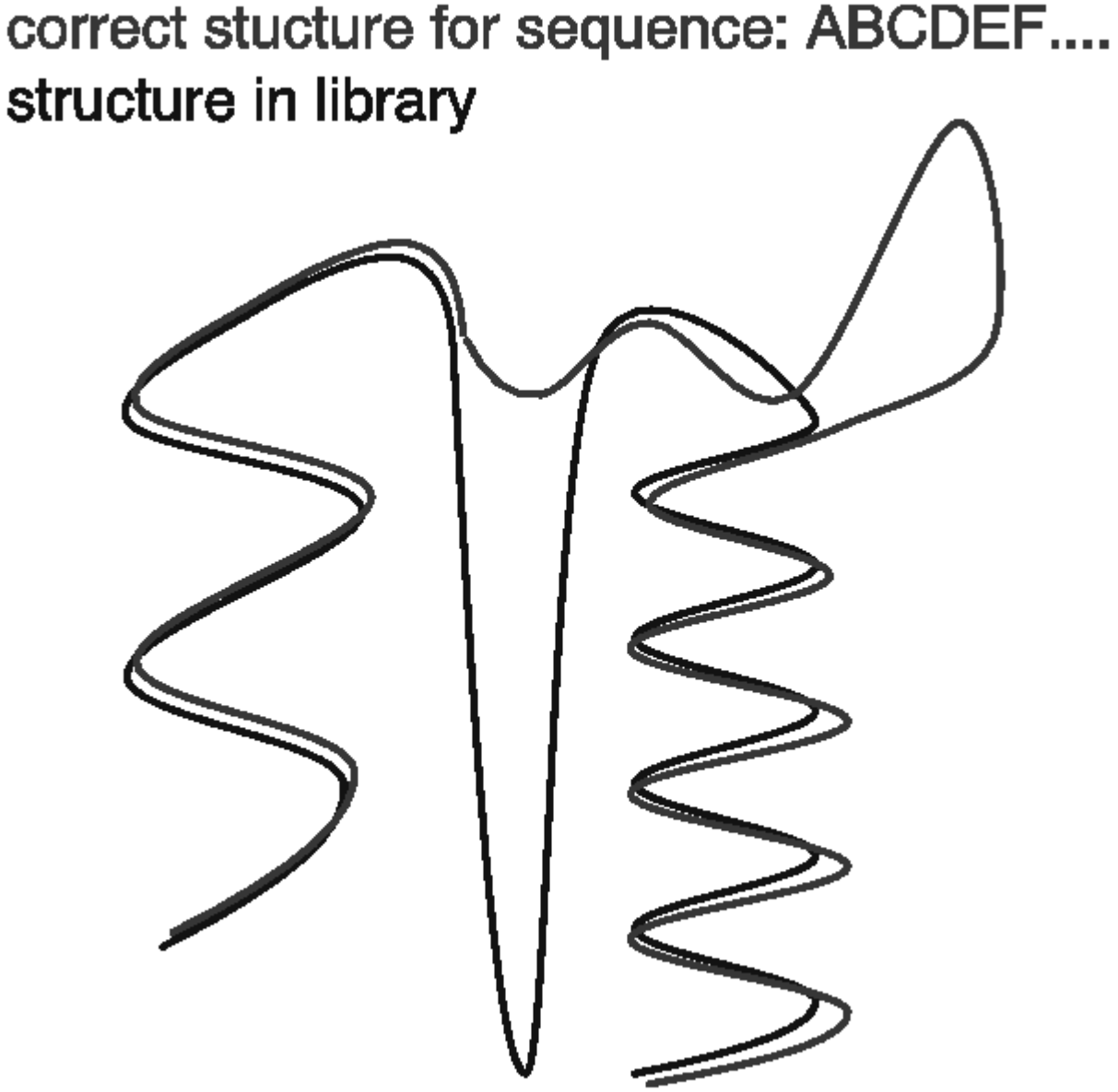


# Three basic choices in molecular modelling

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

# Sequence-Structure Matching
# The search problem

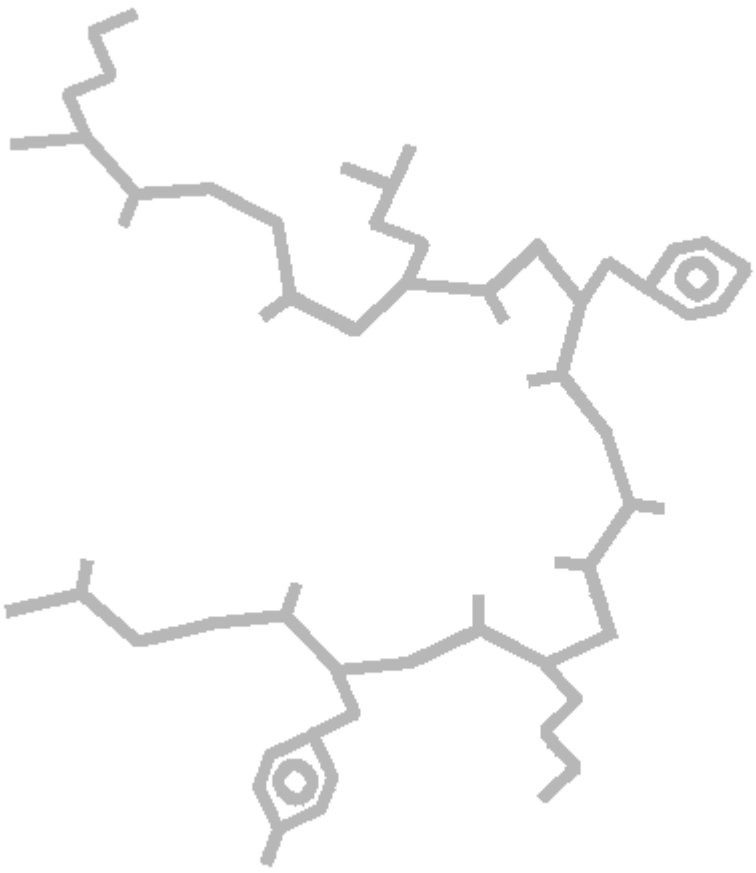
- Gapped alignment = combinatorial nightmare



# Model Representation

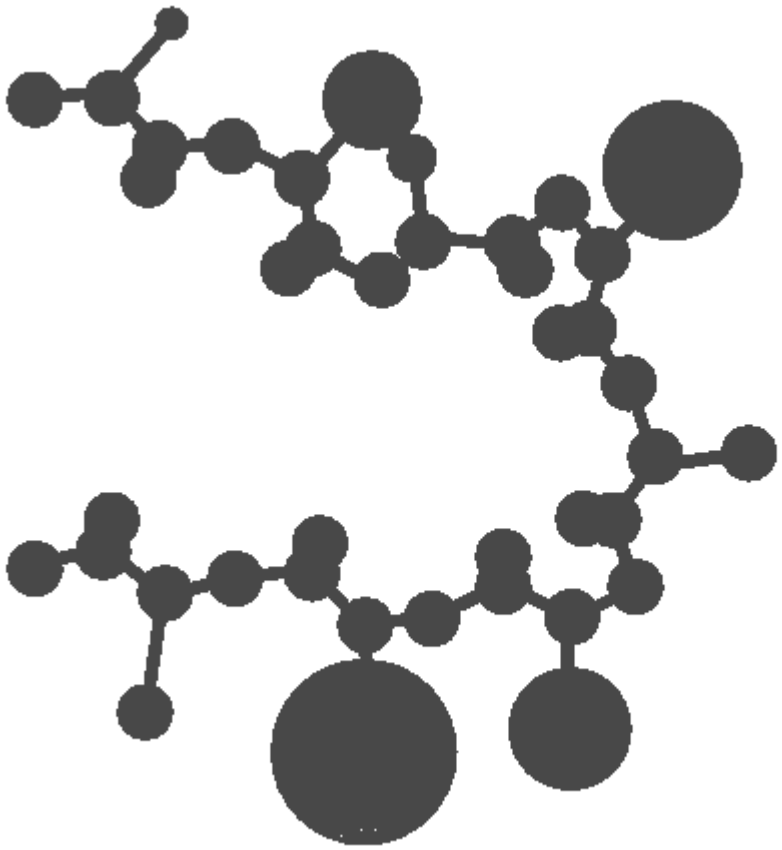
1. Conventional MM

   (structure refinement)



heavy atoms

# 4. Low resolution
(structure prediction)



backbone +
1 sidechain

# Scoring

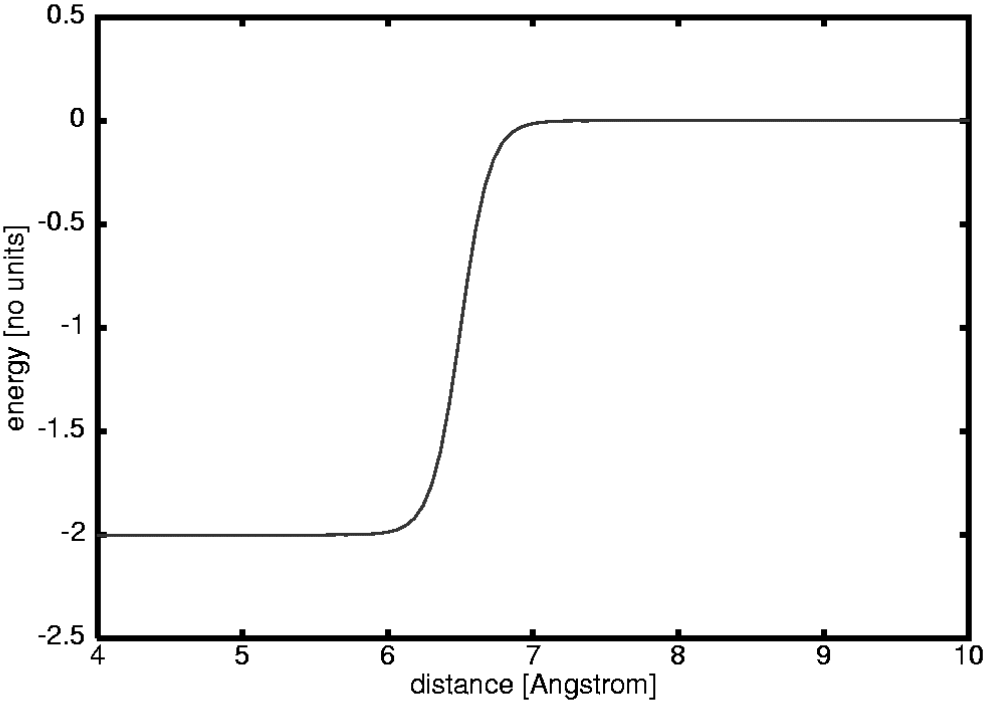
- Quality of prediction is given by

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

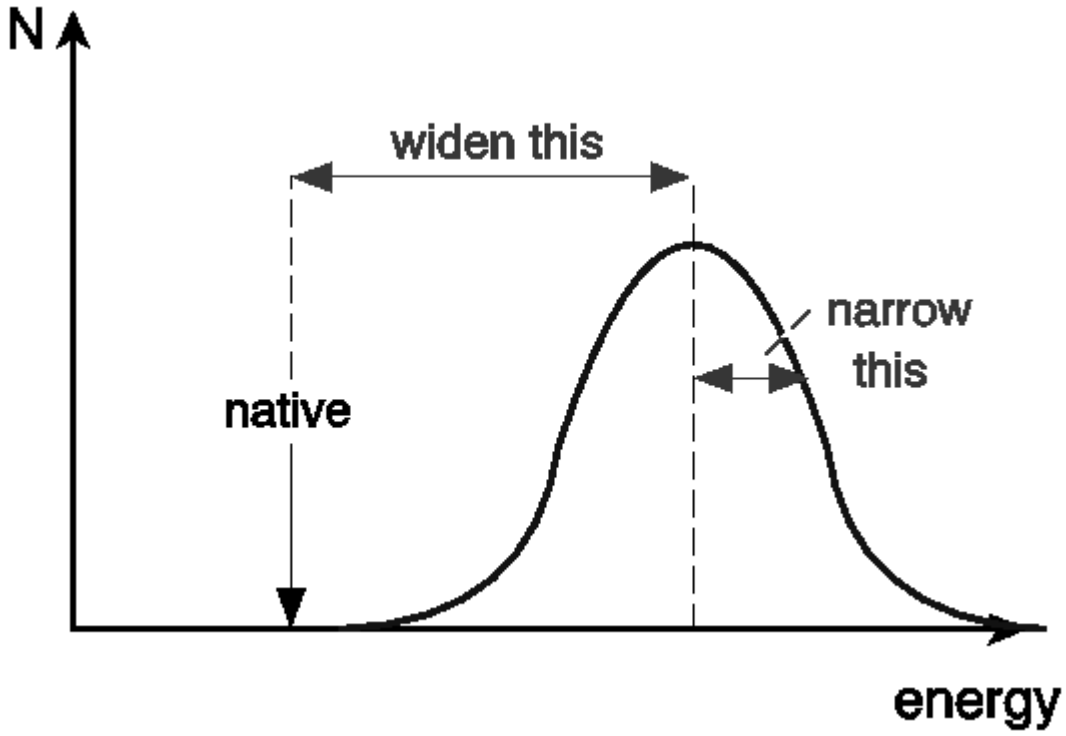
- Functional form of interactions
  - simple
  - continuous in function and derivative
  - discriminate two states

  $\Rightarrow$ hyperbolic tangent function

$$E_{ij} = k_{ij}[1 - \tanh(d_{ij} - d^0)]$$



# Parametrisation 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

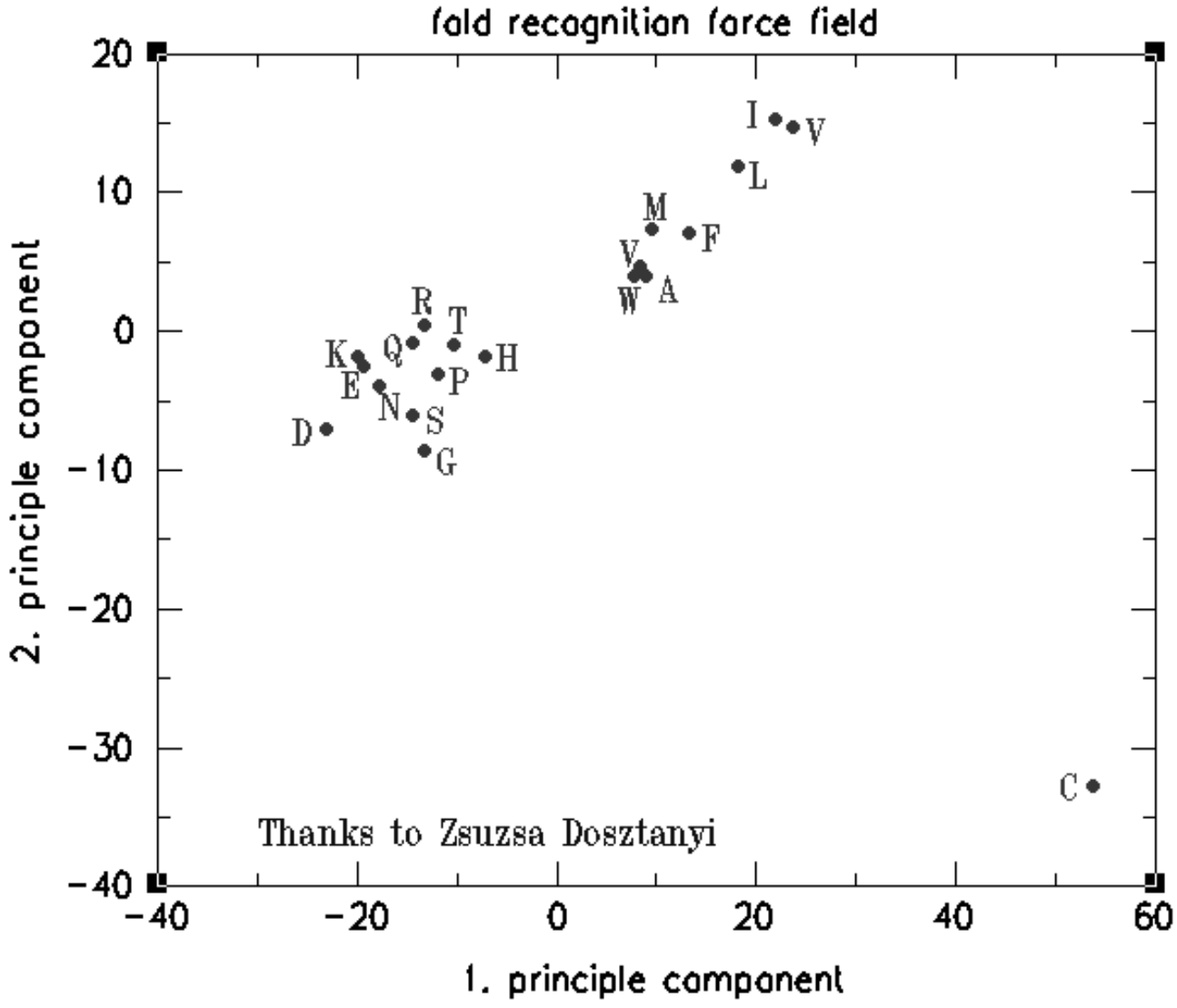
- 893 non-homologous proteins
  - Representative subset of PDB
  - < 25% sequence identity
  - 30-1070 amino acids
- $>10^7$ mis-folded structures

$\Rightarrow$ 2 force fields

- Neighbour unspecific (alignment)
  - 336 parameters
- Neighbour specific (ranking alignments)
  - 996 parameter

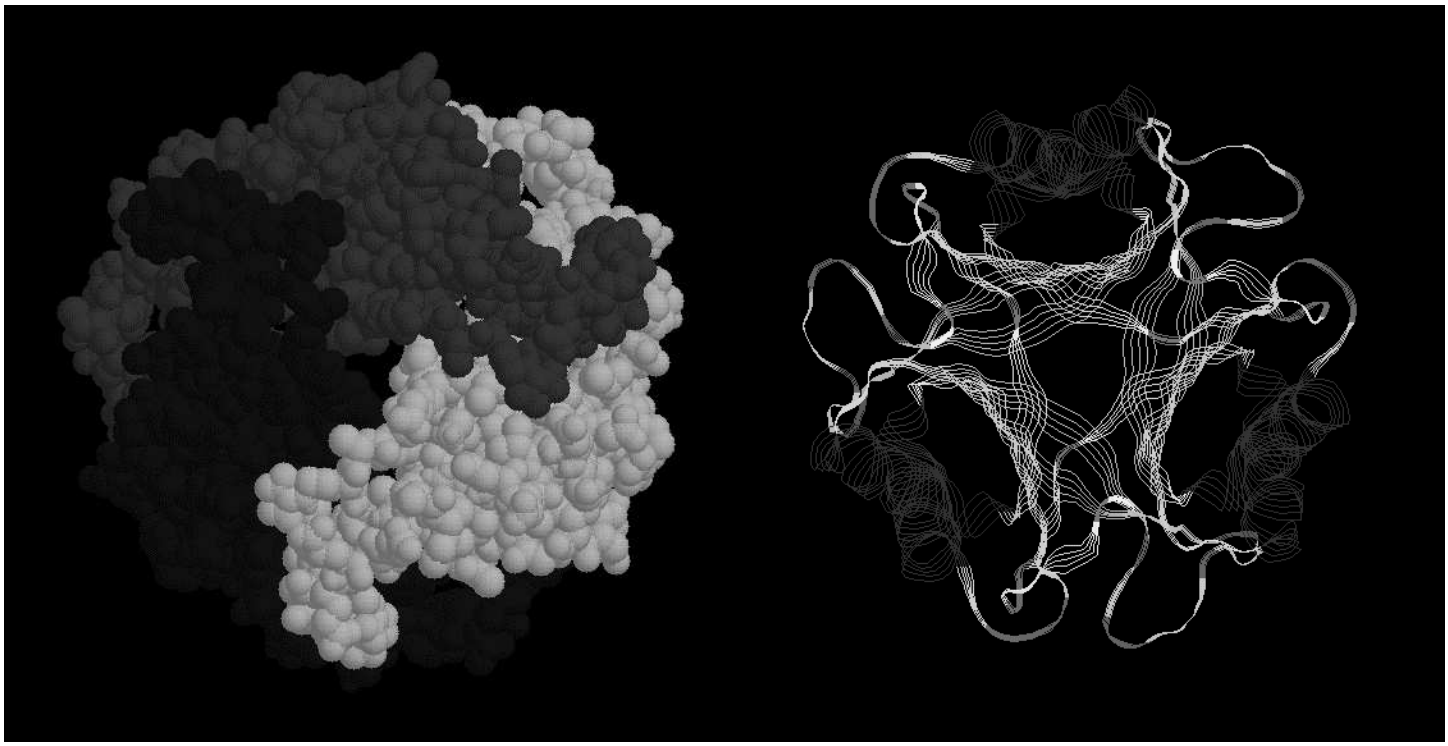
! Parameters well determined !

# Is Our Scoring Function Totally Artificial?



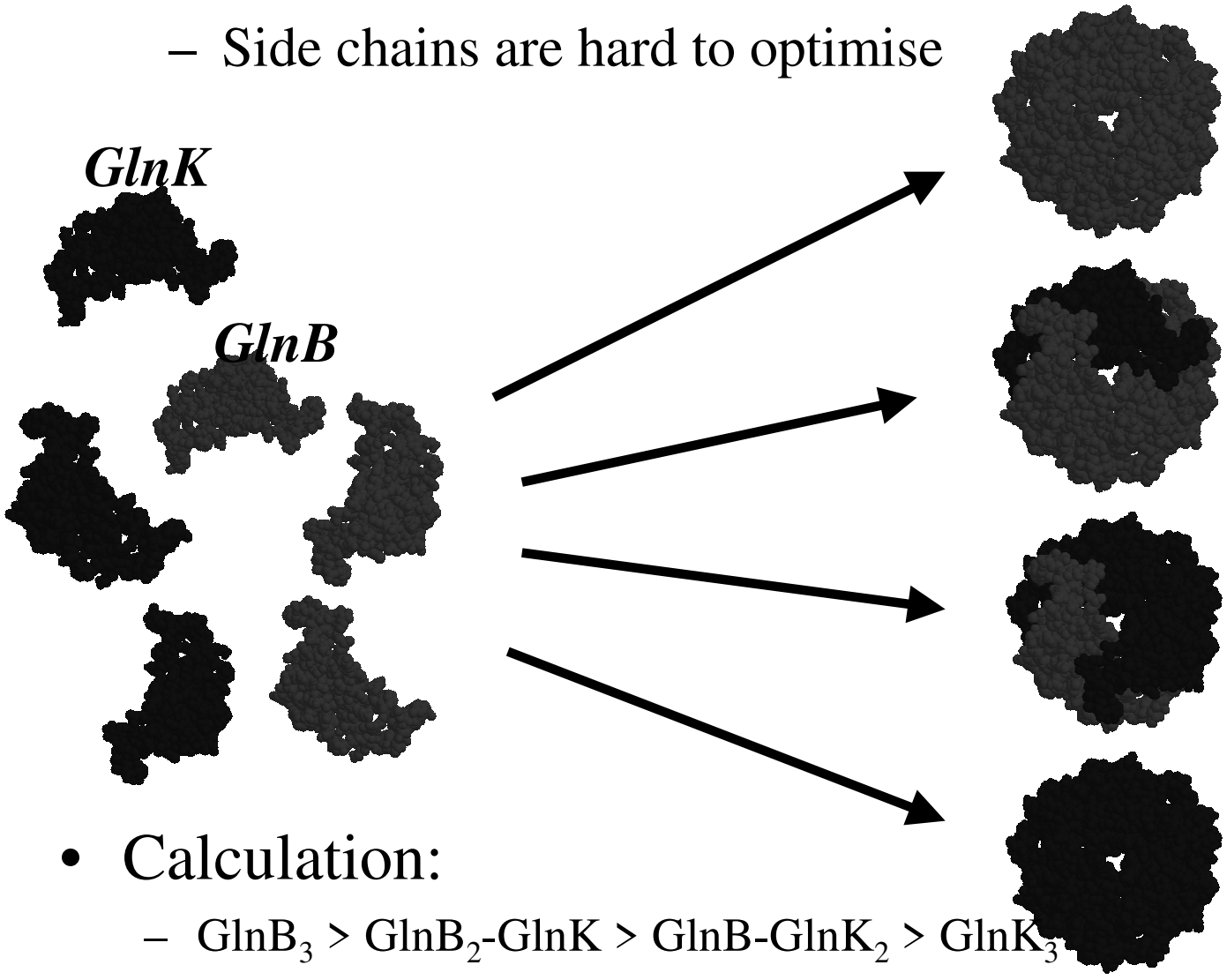
# Trimer Stability

- Nitrogen regulation proteins
  - 2 protein (PII (GlnB) and GlnK)
  - 112 residues
  - sequence: 67% identities, 82% positives
  - structure: 0.7Å RMSD
  - trimeric
  - Dr S. Vasudevan: hetero-trimers



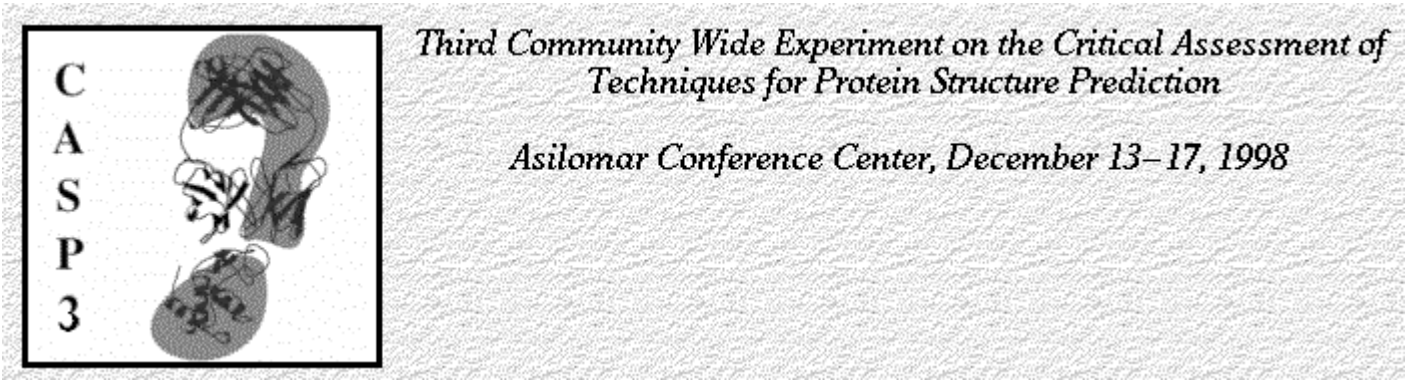
# Hetero-trimer Stability

- What is the most/least stable trimer
- Why use a low resolution force field?
  - Structures differ (0.7Å RMSD)
  - Side chains are hard to optimise




- Calculation:
  - $GlnB_3$ > $GlnB_2$-GlnK > GlnB-$GlnK_2$ > $GlnK_3$
- Experiment:
  - $GlnB_3$ > $GlnB_2$-GlnK > GlnB-$GlnK_2$ > $GlnK_3$

# Does it work with Fold Recognition?




- 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 (our methodology) 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

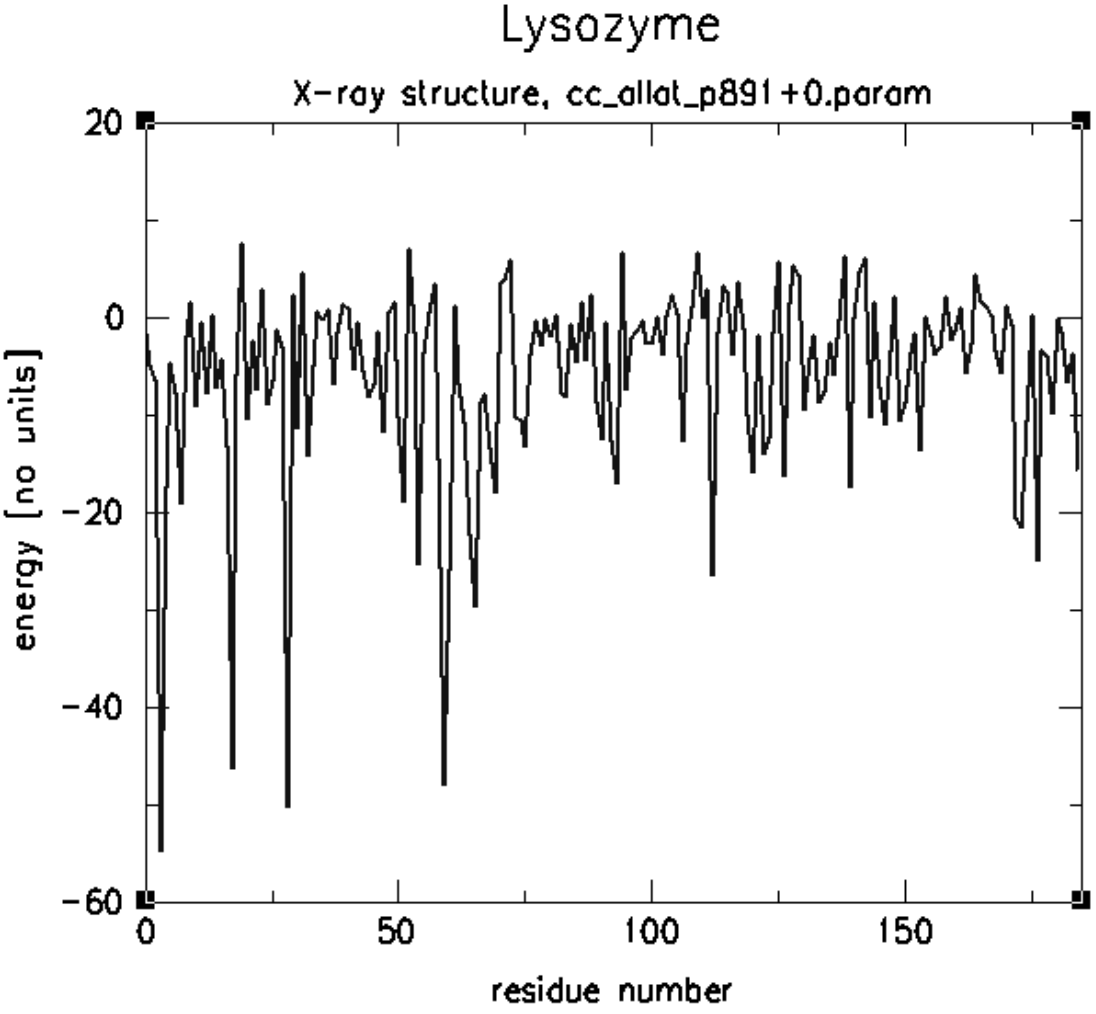
## (computationally by Arne Elofsson)

- 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 |

# Improvements to Fold Recognition

- Noise vs signal




- Average profiles
- Geometry optimised structures

# Structure Optimisation

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

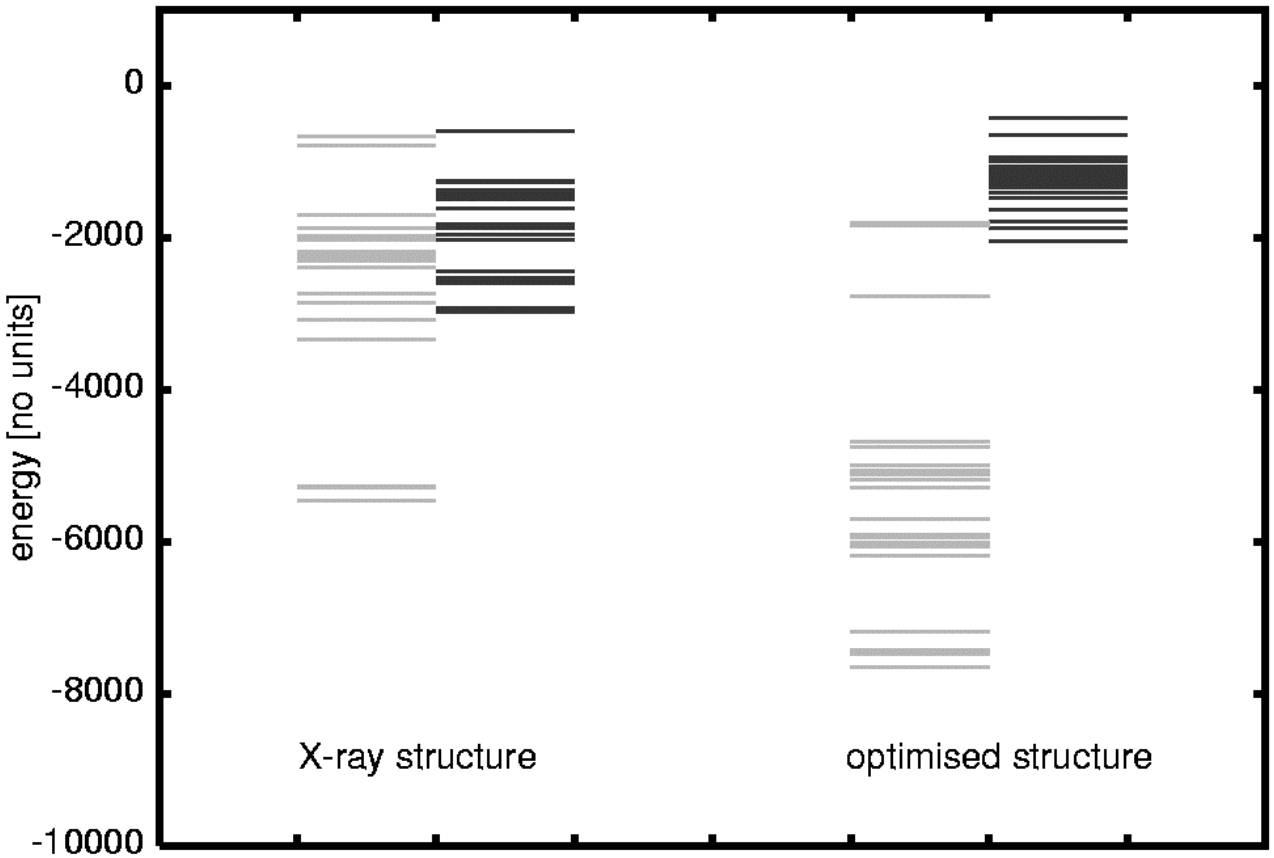
$\Rightarrow$Optimise structure (coordinates) for fold recognition

# How are Structures Optimised?

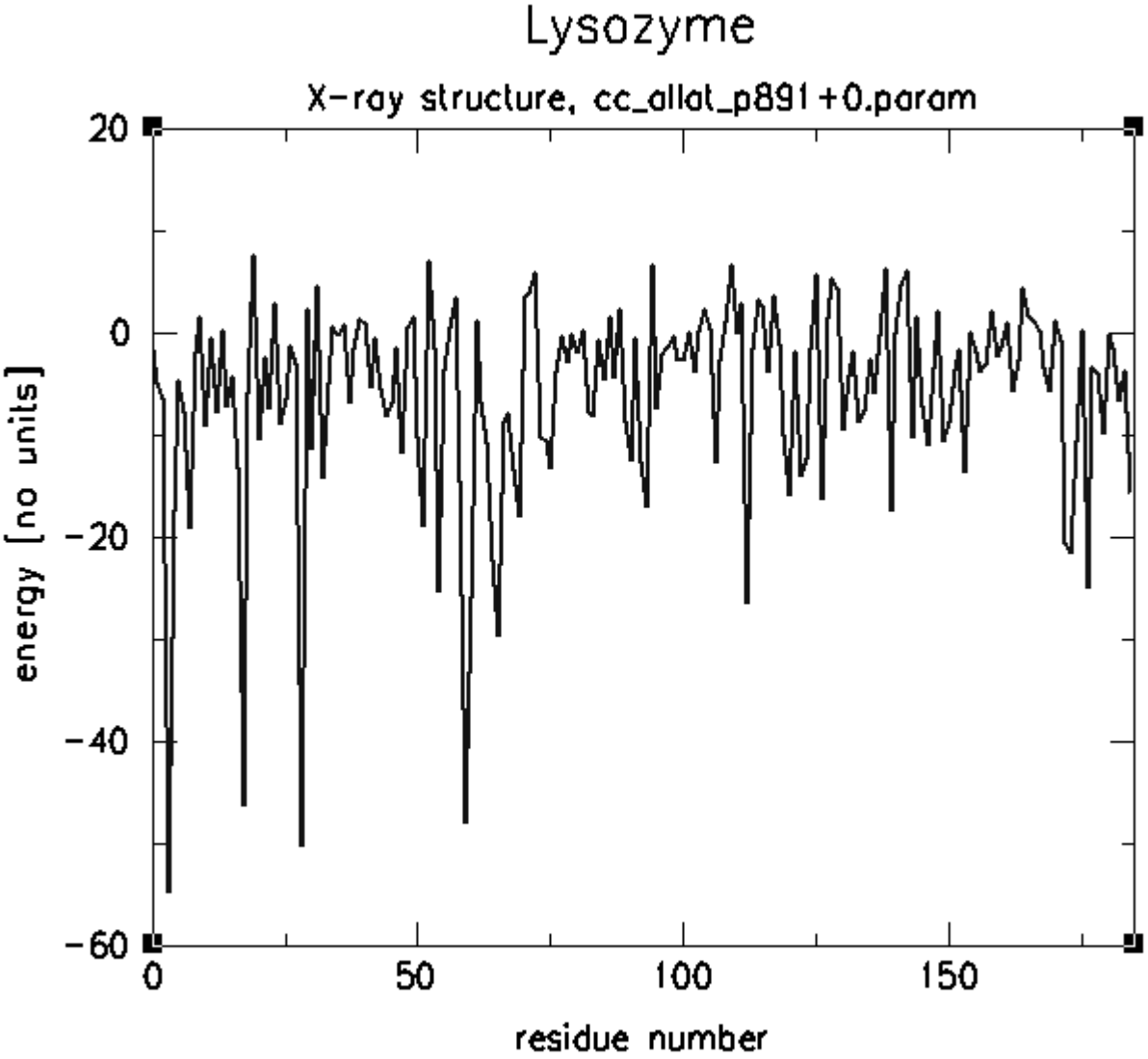
- Goal:
  - NOT to minimise energy of structure
  - BUT increase energy gap between correctly and incorrectly aligned sequences
- Deed:
  - 20 homologous sequences (<95%)
  - 20 best scoring alignments from (893) “wrong” sequences
  - change coordinates to maximise energy gap between “right” and “wrong”
    - restraint to X-ray structure (change <1Å rmsd)
    - 100 steps energy minimisation
    - 500 steps molecular dynamics
- Hope:
  - important structural features are (energetically) emphasised

# Effect of Structure Optimisation

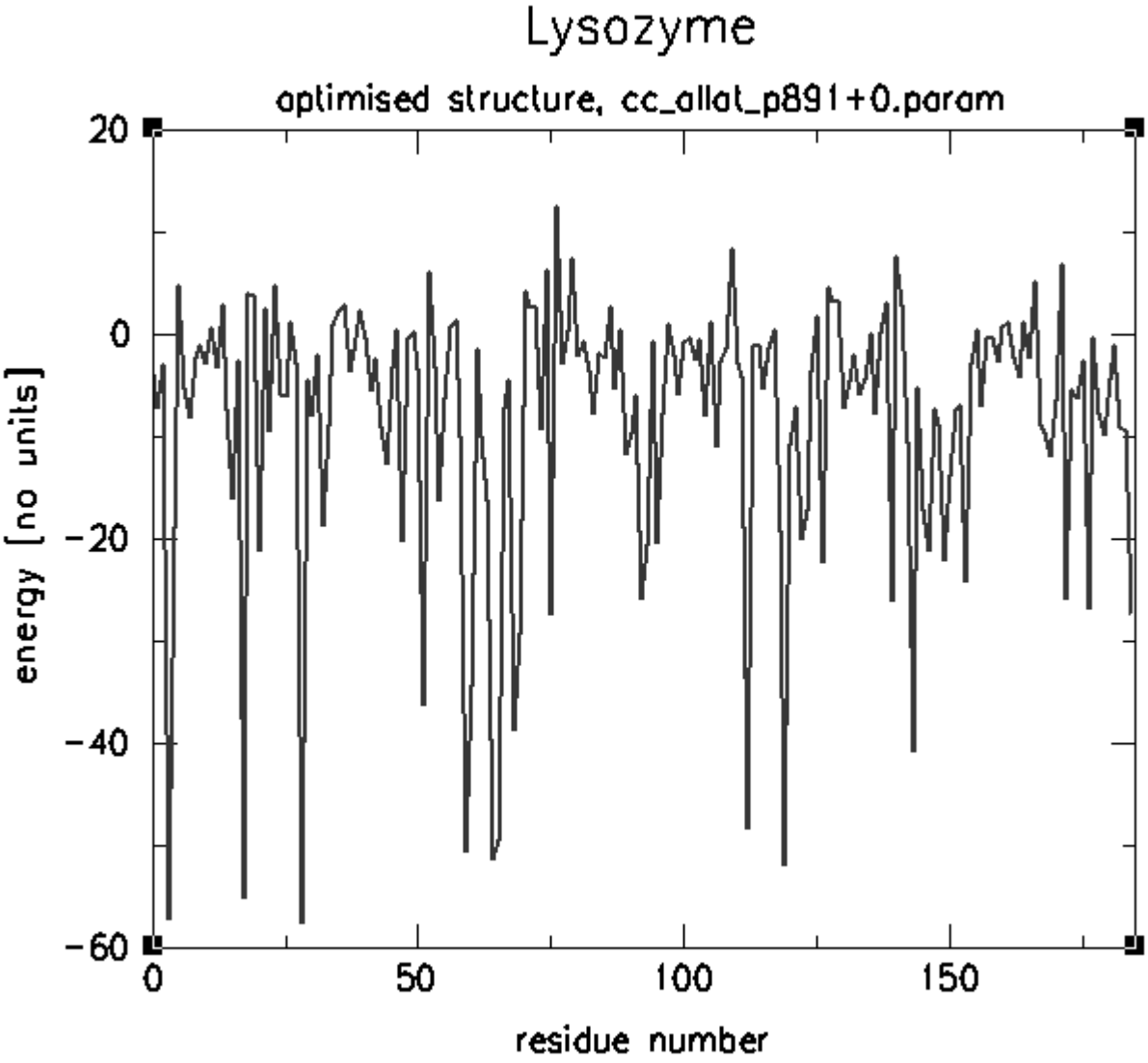
- Lyzosyme (153l_)



# 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 with force field term
- Correlated mutations in sequence

```
Query: 1     RTDCYGNVNRIDTTGASCKTAKPEGLSYCGVSASKKI
Sbjct: 1     RTGCYGDVNRVDTTGASCKSAKPEKLNYCGVAASRKI
Sbjct: 28    -TGCYGSVSRIDTTGASCRTAKPEGLSYCGVRASRTI
```

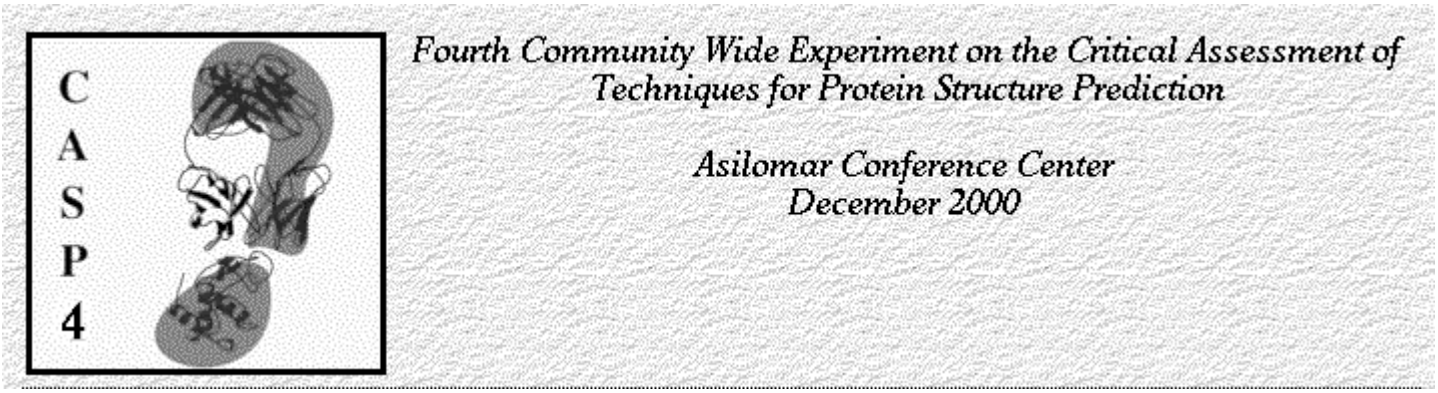
$$d_{ij} \propto c_{ij} = \frac{\left\langle \left(s_i - \langle s_i \rangle\right)\left(s_j - \langle s_j \rangle\right) \right\rangle}{\sigma_i \sigma_j}$$

  - can reflect distance information
  - yet untested (by us)

# Where are we now?

- Cassandra package
  - fast O(N) alignment
  - structural optimised library
  - side chain modelling
  - fully automatic predictions
- Extensive testing with big test sets
  - Mock prediction for 595 test sequences
  - Homologous structure with < 25% sequence identity in library
  - ≈25%, homologous structure ranks #1
  - ≈ 45% correct hit in top 10
  - average shift error of alignment ≈ 4
- Confidence of prediction
  - Predicting new folds

# Structure Prediction Olympics 2000




- CASP4 experiment
  - held April - September 2000
  - 43 target sequences
    - ≈30 no sequence homology detectable with sequence-sequence alignment techniques
  - 154 prediction groups
  - Cassandra predictions
    - top 5 predictions for all targets are submitted
    - no human intervention (why?)
- Leap frog or being frogged?
  - Results to be published in December

# CASP4: T111

- Protein Name: enolase
- Organism: E. coli
- # amino acids: 436
- Homologous sequence of known structure: YES!
- Structure solved by molecular replacement.

⇒✈-Blast search

- 4enl: Enolase
  - 431 residues aligned
  - 46% identities, 62% positives
  - Expect = $10^{-100}$

# Homologous structures to 4enl in fold library

- FSSP strucure-structure comparison

  $\Rightarrow$ 33 homologous structures

  < 13% sequence identity, > 3.6 Å RMSD, < 50% of full structure

| Name | Z | RMSD | nali | nstr | seqid | |
|---|---|---|---|---|---|---|
| 1a49A | 9.8 | 4.7 | 204 | 519 | 11 | Opt_bin/1a49A.bin |
| 1byb | 9.8 | 3.7 | 196 | 490 | 6 | Opt_bin/1byb_.bin |
| 1nar | 8.3 | 3.7 | 184 | 289 | 8 | Opt_bin/1nar_.bin |
| 1b5tA | 8.1 | 3.6 | 180 | 275 | 6 | Opt_bin/1b5tA.bin |
| 1aj2 | 8.1 | 3.8 | 175 | 282 | 8 | Opt_bin/1aj2_.bin |
| 1cnv | 7.8 | 3.6 | 177 | 283 | 9 | Opt_bin/1cnv_.bin |
| 1qba | 7.6 | 3.9 | 190 | 858 | 7 | Opt_bin/1qba_.bin |
| 1dhpA | 7.4 | 4.0 | 166 | 292 | 8 | Opt_bin/1dhpA.bin |
| 1onrA | 7.3 | 3.3 | 169 | 316 | 8 | Opt_bin/1onrA.bin |
| 4xis | 7.3 | 4.3 | 187 | 386 | 13 | Opt_bin/4xis_.bin |
| 1rpxA | 7.2 | 3.5 | 156 | 230 | 7 | Opt_bin/1rpxA.bin |
| 1pud | 7.2 | 4.2 | 191 | 372 | 5 | Opt_bin/1pud_.bin |
| 1smd | 7.1 | 3.8 | 180 | 496 | 6 | Opt_bin/1smd_.bin |
| 1eceA | 6.9 | 3.9 | 182 | 358 | 11 | Opt_bin/1eceA.bin |
| 1oyc | 6.5 | 4.3 | 183 | 399 | 9 | Opt_bin/1oyc_.bin |
| 1edg | 6.0 | 3.9 | 178 | 380 | 7 | Opt_bin/1edg_.bin |
| 1dosA | 5.9 | 3.8 | 161 | 343 | 11 | Opt_bin/1dosA.bin |
| 2dorA | 5.6 | 4.0 | 163 | 311 | 8 | Opt_bin/2dorA.bin |
| 1bd0A | 5.5 | 3.4 | 143 | 381 | 8 | Opt_bin/1bd0A.bin |
| 1a4mA | 5.4 | 3.7 | 153 | 349 | 9 | Opt_bin/1a4mA.bin |
| 2tpsA | 5.2 | 3.6 | 142 | 226 | 5 | Opt_bin/2tpsA.bin |
| 1uroA | 5.1 | 4.0 | 151 | 357 | 8 | Opt_bin/1uroA.bin |
| 1aq0A | 5.0 | 3.8 | 152 | 306 | 6 | Opt_bin/1aq0A.bin |
| 1tml | 4.9 | 4.0 | 146 | 286 | 10 | Opt_bin/1tml_.bin |
| 1uok | 4.9 | 4.1 | 175 | 558 | 8 | Opt_bin/1uok_.bin |
| 2plc | 3.6 | 4.6 | 149 | 274 | 7 | Opt_bin/2plc_.bin |
| 1nfp | 2.8 | 4.4 | 128 | 228 | 7 | Opt_bin/1nfp_.bin |
| 1wab | 2.6 | 3.8 | 108 | 212 | 7 | Opt_bin/1wab_.bin |
| 1auz | 2.4 | 3.3 | 83 | 116 | 1 | Opt_bin/1auz_.bin |
| 1mtyG | 2.4 | 3.9 | 61 | 162 | 4 | Opt_bin/1mtyG.bin |
| 8abp | 2.2 | 4.2 | 85 | 305 | 11 | Opt_bin/8abp_.bin |
| 1fuiA | 2.2 | 4.9 | 108 | 591 | 6 | Opt_bin/1fuiA.bin |
| 1be1 | 2.1 | 3.6 | 92 | 137 | 8 | Opt_bin/1be1_.bin |

# T111: Cassandra prediction

## Sorted by score:

| score | nali | name | |
|---|---|---|---|
| 7533.9 | 324 | "1a4mA" | adenosine deaminase |
| 7269.9 | 309 | "1onrA" | transaldolase |
| 7112.5 | 298 | "1rkd_" | ribokinase |
| 7016.9 | 359 | "1ch6A" | glutamate dehydrogenase |
| 7009.3 | 329 | "1dosA" | aldolase class ii |
| 6959.4 | 333 | "3pte_" | d-alanyl-d-alanine carboxypeptidase |
| 6866.3 | 323 | "1uroA" | uroporphyrinogen decarboxylase |
| 6810.6 | 303 | "1cipA" | guanine nucleotide-binding protein |
| 6788.4 | 352 | "1smd_" | amylase |
| 6785.8 | 277 | "1a4iB" | methylenetetrahydrofolate dehydrogenase |
| 6783.6 | 284 | "1dhpA" | dihydrodipicolinate synthase |
| 6771.2 | 364 | "1ajsA" | aspartate aminotransferase |

.
.

# T111: Cassandra prediction

**Sorted by score:**

| score | nali | name | |
|---|---|---|---|
| 7533.9 | 324 | "1a4mA" | adenosine deaminase |
| 7269.9 | 309 | "1onrA" | transaldolase |
| 7112.5 | 298 | "1rkd_" | ribokinase |
| 7016.9 | 359 | "1ch6A" | glutamate dehydrogenase |
| 7009.3 | 329 | "1dosA" | aldolase class ii |
| 6959.4 | 333 | "3pte_" | d-alanyl-d-alanine carboxypeptidase |
| 6866.3 | 323 | "1uroA" | uroporphyrinogen decarboxylase |
| 6810.6 | 303 | "1cipA" | guanine nucleotide-binding protein |
| 6788.4 | 352 | "1smd_" | amylase |
| 6785.8 | 277 | "1a4iB" | methylenetetrahydrofolate dehydrogenase |
| 6783.6 | 284 | "1dhpA" | dihydrodipicolinate synthase |
| 6771.2 | 364 | "1ajsA" | aspartate aminotransferase |

.
.

- Probability of this result by chance:

  $p = 1.36 \cdot 10^{-9}$

- BUT: Alignment is shifted!!!
  - ✈-Blast prediction is much better.

# Summary

- Urgency of Prediction
  - sequencing: fast & cheap
  - structure determination: hard & expensive
  - $\approx 10^4$ structures are determined
    - insignificant compared to all proteins
- Fold recognition
  - a feasible way to predict protein structure
  - is not perfect (9/10, 1/4)
  - requires special scoring functions
- Low resolution scoring functions
  - knowledge based
    - from database of known protein structures
    - only meaningful when database is big
    - data mining?
  - not necessarily physical
  - BUT capture important physical features

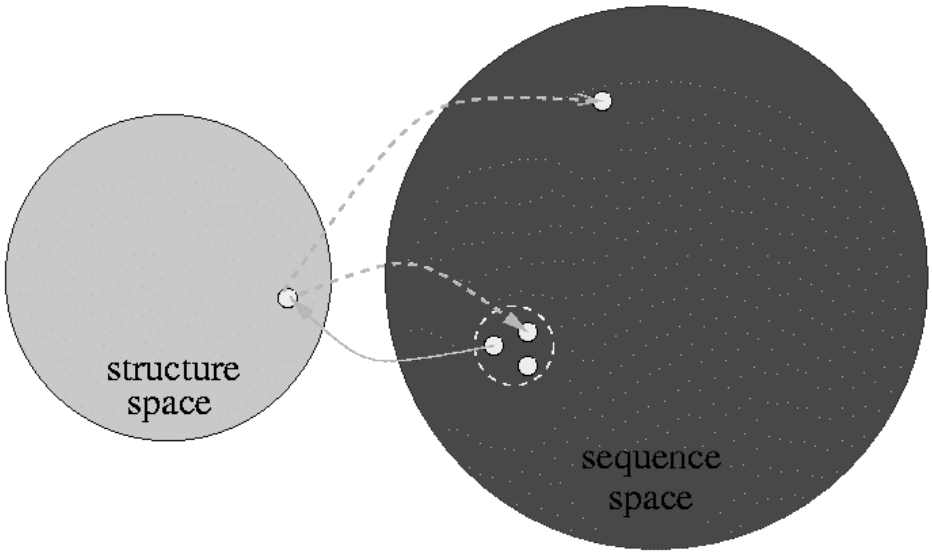
# Future work

- Large scale structure prediction
  - Fold recognition on genomic scale
    - 20% predicted protein >> what's in PDB
    - putative proteins
    - new folds
    - from structure to function (maybe too hard)

  $\Leftrightarrow$ why our CASP submissions are fully automatic
  - Experimentally assisted structure prediction
    - cross linking & MS
  - Prediction based structure determination
    - structure determination is much easier if a tentative model is already known
    - use experiment to confirm prediction

# What else?

- The inverse problem
  - Is there a sequence match for a structure?

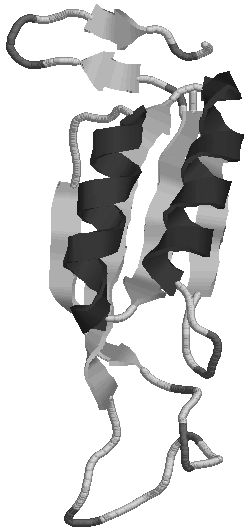



- Applications for the inverse problem
  - Fishing for putative sequences in genomic ponds
  - "Better" sequences for proteins

    What is "better"?
    - More stable
    - More soluble
    - Better to crystallise
    - Better function
    - etc.

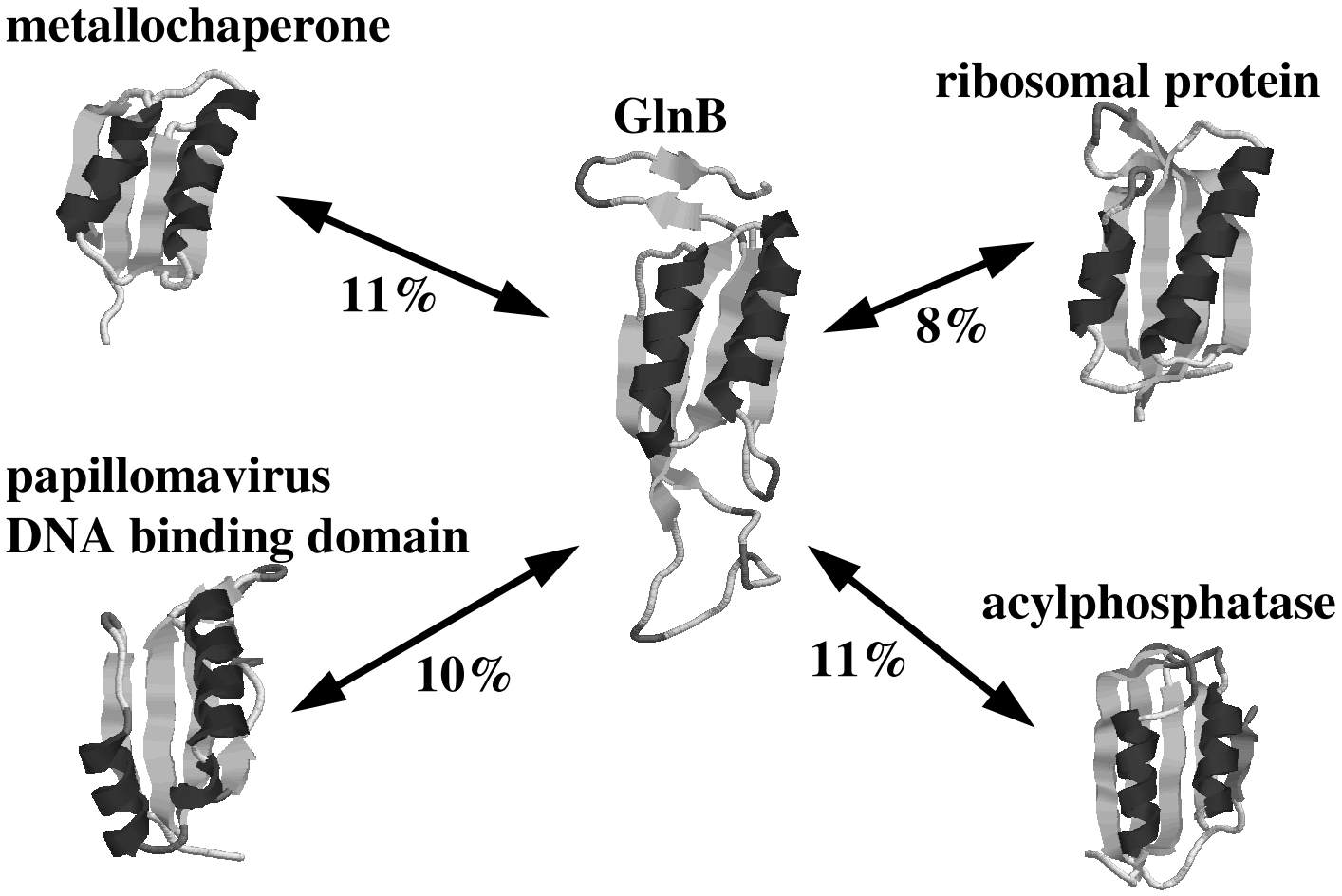
# Rational Protein Design

**GlnB**



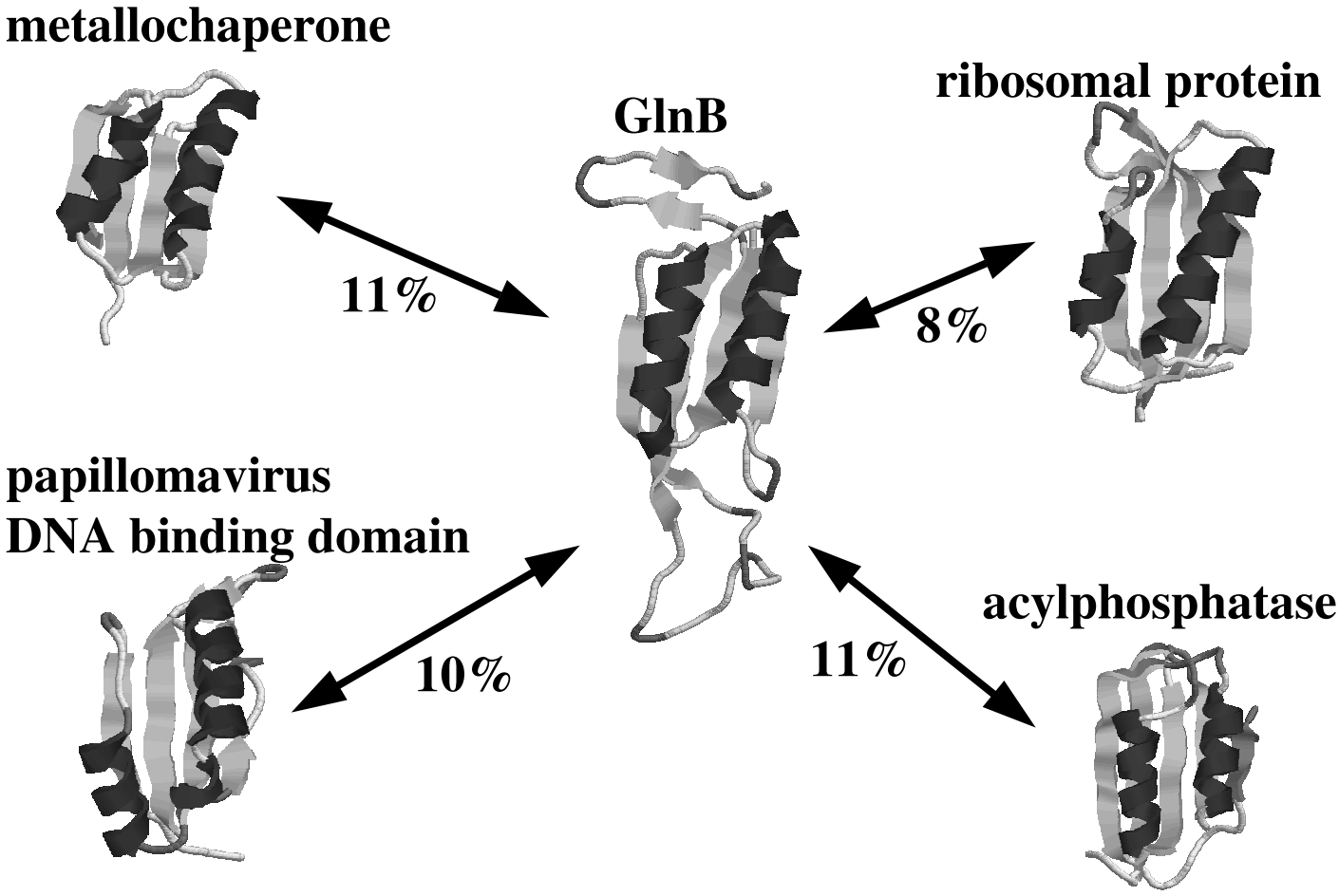
- Is there a “better” sequence for GlnB structure?

# Example GlnB




- Nature uses same fold motif for different functions

# Why important?




- Minimalistic proteins
- Many industrial applications
  - E.g. enzymes in washing powder
    - should be stable at high temperatures
    - work faster at low temperature
    - …

# Naïve Concoction

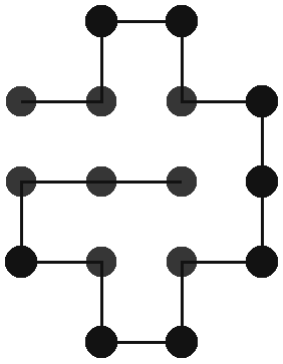
- Use energy score
  - e.g. score from low resolution force field
- Change sequence to lower energy

# Why naïve?

- Comparing energies of different sequences is like comparing apples with potatoes
- Free energy is all important measure
  - Is it possible to capture free energy in a simple function?

# 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



  ⇒ evaluation of all conformations (exact free energy)

  ⇒ for all possible sequences

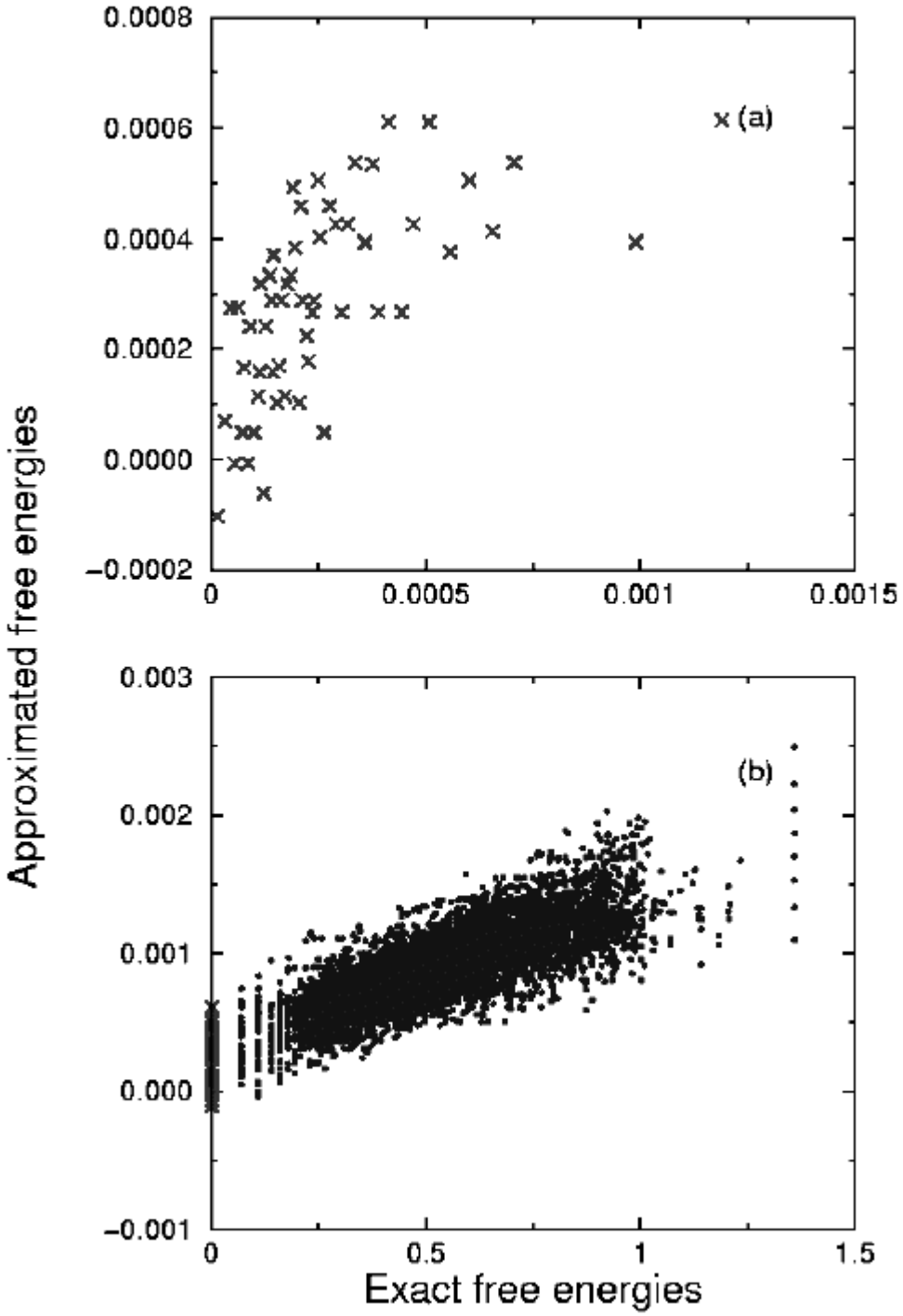
- "Our small universe"
  - 802074 self avoiding conformations
  - $2^{16} = 65536$ sequences
  - 1539 (2.3%) sequences fold to unique structure
  - 456 folds
  - 26 sequences adopt most common fold

# Free energy approximation

- Question: Is there a simple function which approximates free energy
  - Calculate free energies for all sequences
  - Select folding sequences and use them to fit new scoring function
  - correlate free energy and approximated free energy for all sequences

- Using simple 3 parameter HP matrix for fit does not work well
- BUT ...

# Extended Functional Form (5 parameters)



# People

- Sausage
  - Andrew Torda (RSC)
  - Dan Ayers (RSC)
  - Zsuzsa Dosztanyi (RSC)
  - Anthony Russell (RSC)
- GlnB/GlnK
  - Subhash Vasudevan (JCU)
  - David Ollis (RSC)
- At ANUSF
  - Alistair Rendell

# Want to try yourself?

- Sausage and Cassandra freely available

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

  Thomas.Huber@anu.edu.au