Protein Sequence Design

- Why conventional force fields will not work ?

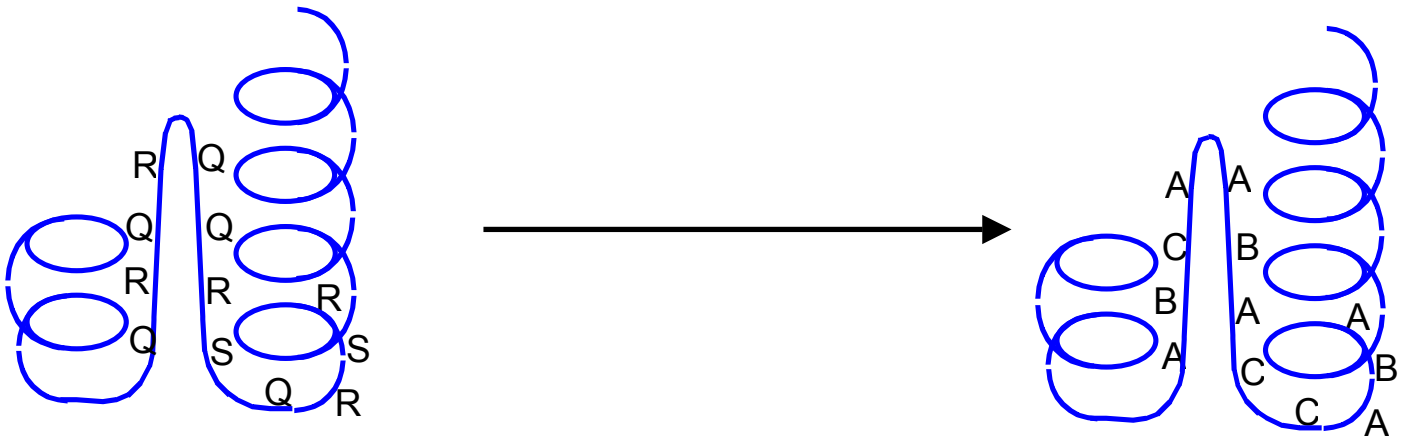
  +

  - different kinds of score functions / force fields
  - search problem

  - outrageous claims
  - remarkable successes

- Definitions…

Basic idea



native protein

"improved" protein

Rule

- structure should not change

Method

- the sequence should be predicted…

What might be useful ?

You have an important protein

- favourite enzyme
- binding protein (transport, receptor, ..)

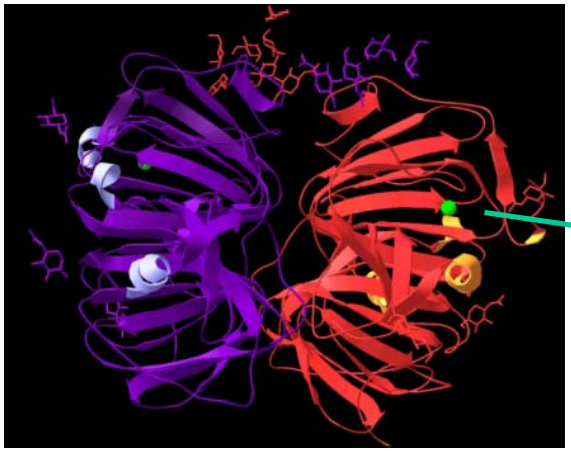
Two reasonable aims

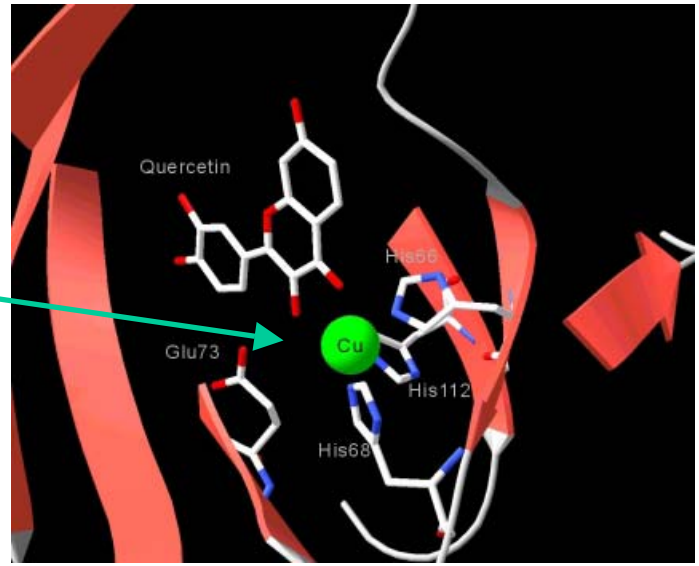
- change / improve activity specificity / binding
- change overall protein stability / solubility

Activity / Specificity

- how hard ?

Changing Activity / Specificity





To change activity

- know where every atom is to Å accuracy
- change residues and still know
- understand the chemistry / reactions / binding intimately
  - reactions are not a classical phenomenon
- predict substrate / product affinities….

What usually happens ? What really works ?

Picture of dioxygenase (Dijkstra) from http://www.xray.chem.rug.nl/Gallery/Qd.htm

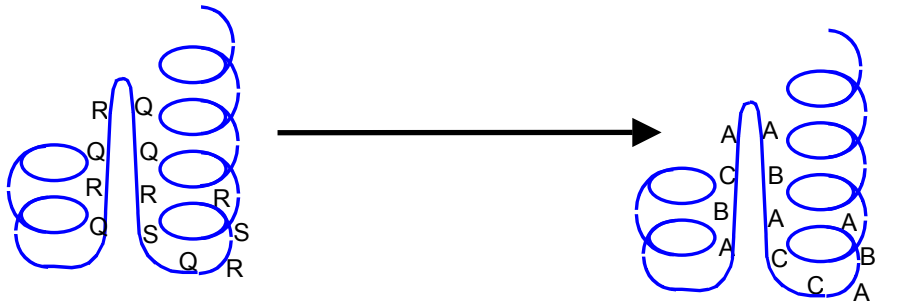
Really changing activity of a protein

Randomise + selection

- randomised genes in bugs
- phage display
- in vitro "evolution"
- …

- Reconsider the sequence design problem

limited version of problem



Rule

- structure should not change

Method

- the sequence should be predicted (not found by experiment)

Limitations

- do not worry about activity
- just make a better structure

Implication

- we should be able to fix residues

Applications

Make a protein more thermostable

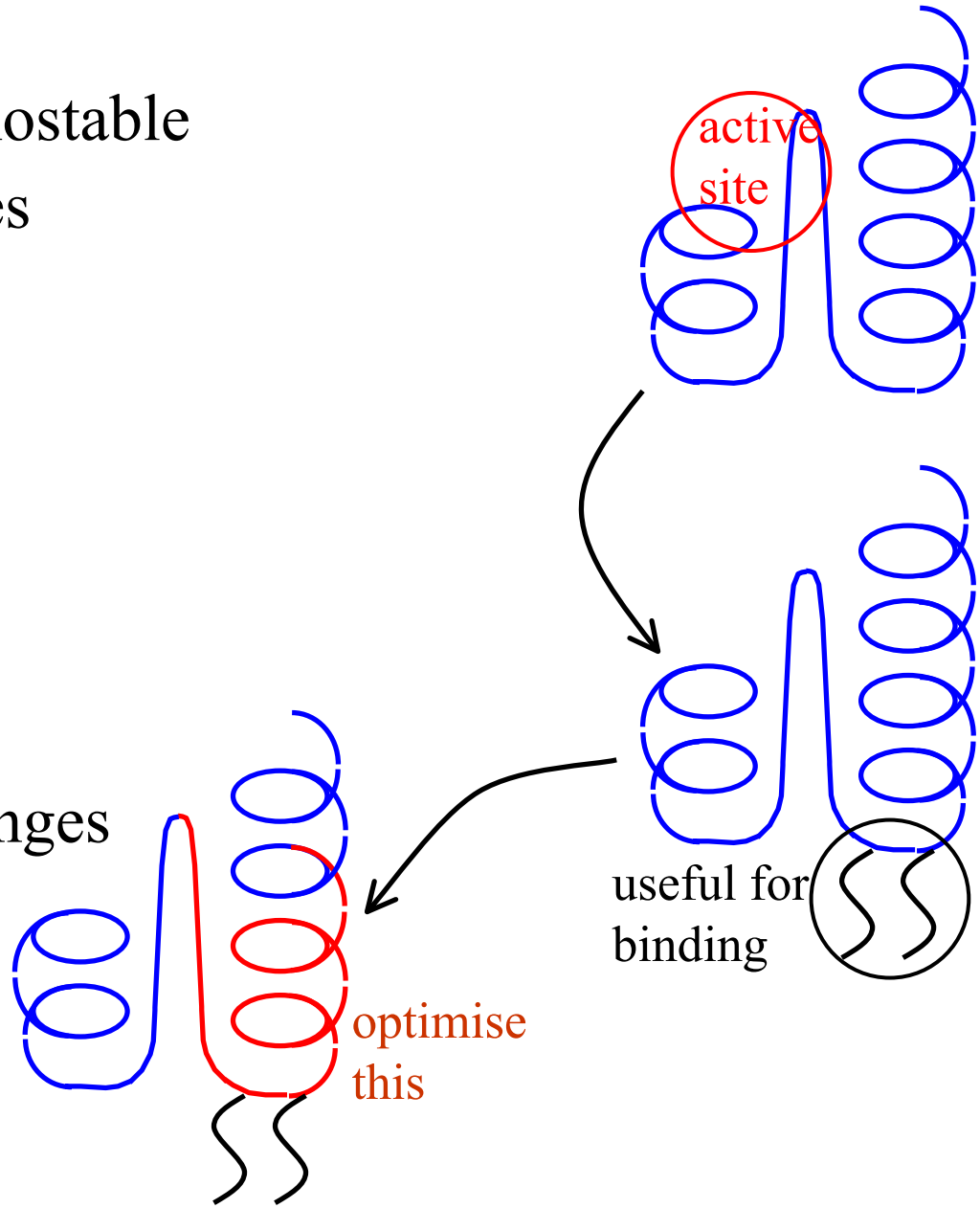
- washing powder enzymes
- industrial catalysts

Stable to other changes

- pH
- solvents
- ionic denaturing

Tolerant of engineered changes

- special residues
- minimisation



Realistic

Our goal ?

- optimise for thermostability or ΔG (folding)

Two aspects

- score function (energy / stability / happiness)
- search…

Sequence Search Problem

20 amino acids

- at each position 20 ×20 ×20 …
- 20^N possibilities / exponential growth

Some quick hacks

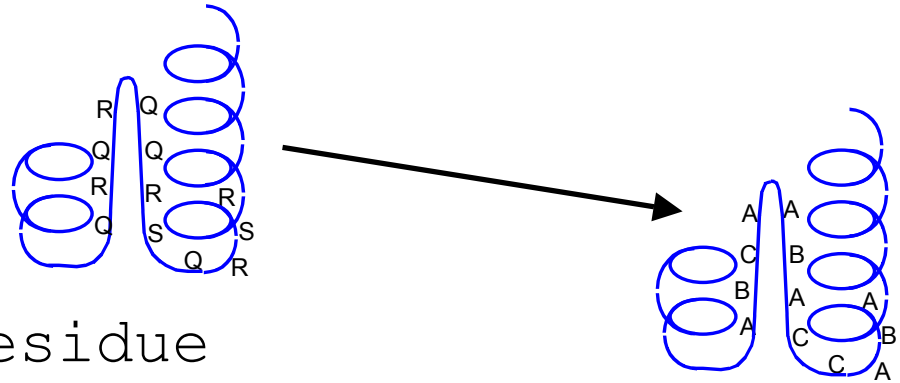
- polar / charged residues at surface / hydrophobic in core
- still exponential (consider $100^2 \approx 10^{30}$)

Real methods

- branch and bound / pruning
- self consistent mean field
- MC
  - should not really work
  - sometimes does

Assume searching is easy

Searching with energy



Simple method

```
for each site {
    best := native residue
    for amino acids 1 .. 20 {
        insert residue
        calculate energy
        if new better than best
            best := new residue
    }
}
```

- what will happen ?
- what do I mean by energy ?

Atomistic Energy

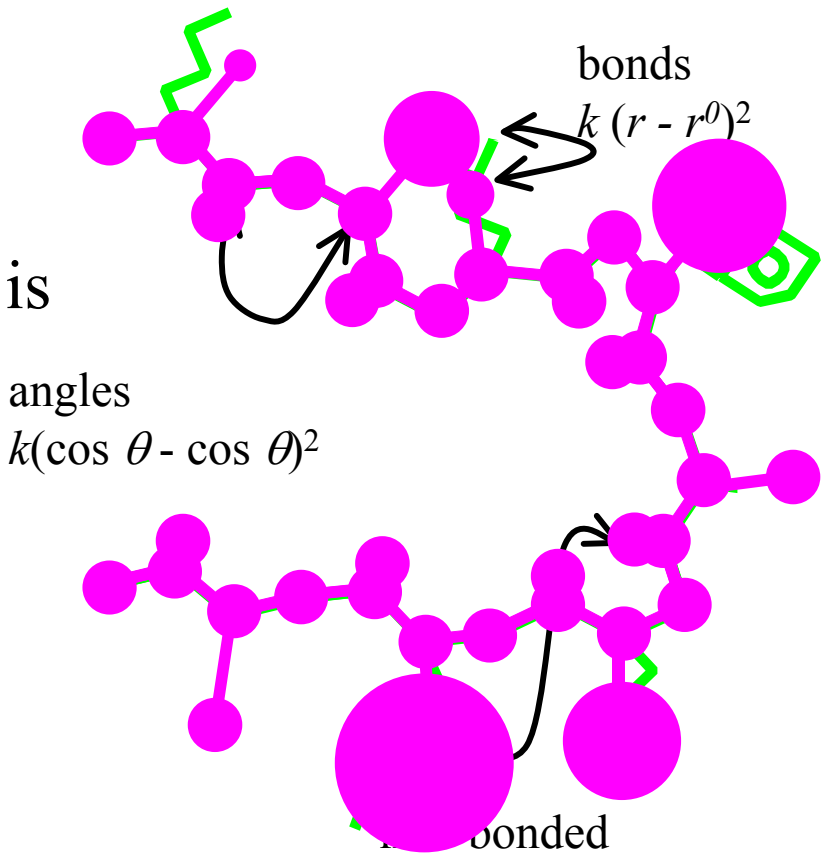
Good

- potentially best energies

Bad

- need to know where every atom is
- change one residue
  - perturbs others
  - moves backbone

- Alternative…



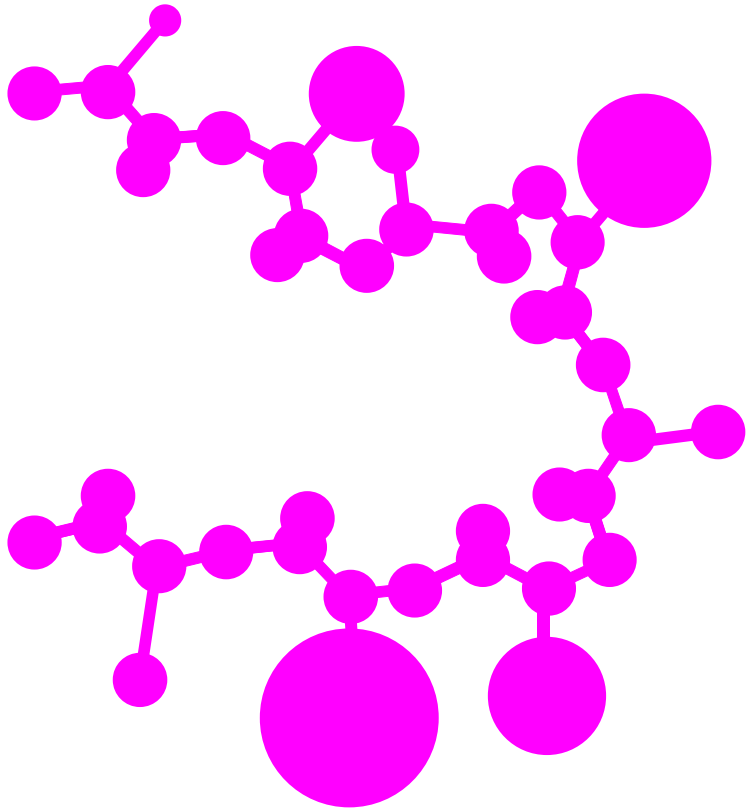
Coarse-grain energy

Good

- fewer interactions
- not sensitive to exact geometry
- can encode important properties

Bad

- No good at sidechain packing
- potentially less accurate



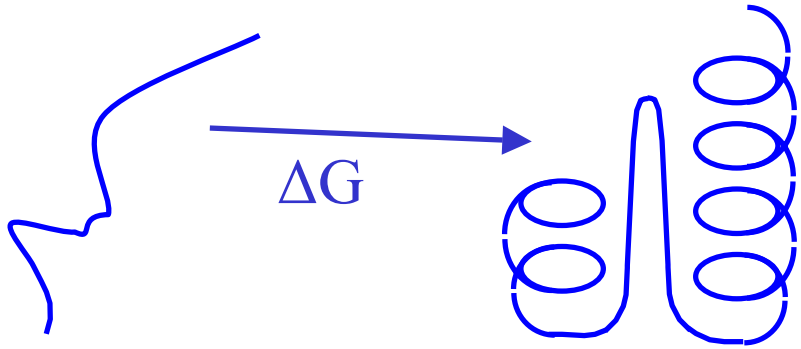
Why are we chasing energy ?

- god likes free energy
- what calculation would you want ?

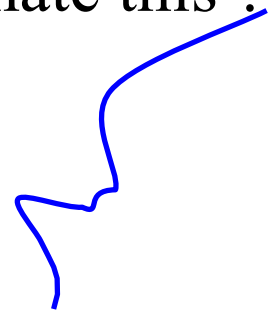
Serious free energies

Stability

(forget kinetics, barriers…)

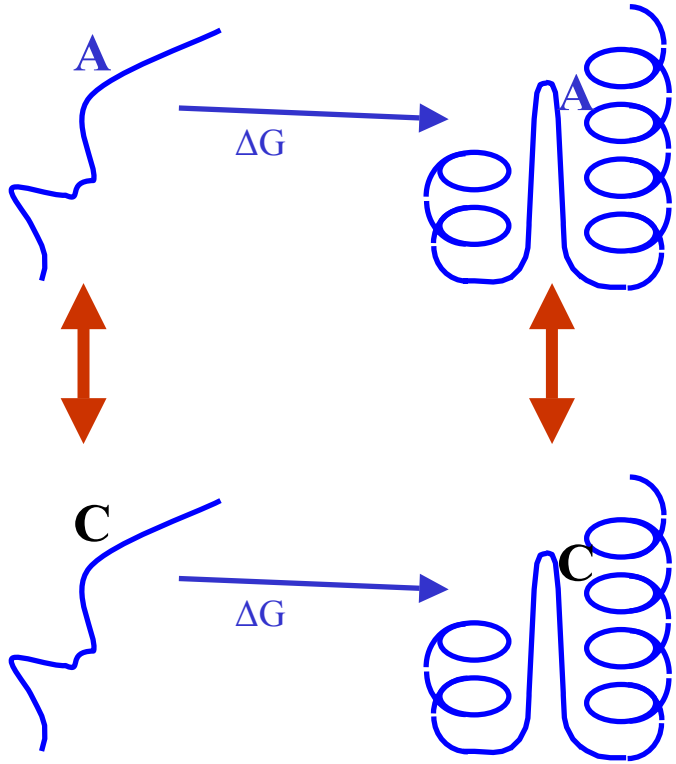


- can we estimate this ?
- what is ?



- To calculate the effect of a residue change…

native residue



modified residue

direct change too hard

- implies a free energy cycle

serious free energy

- implies a knowledge of unfolded states

in practice ? ….

First guess at "energy" function

Atomistic

- does not include anything like free energy

Coarse grain ?

- more ad hoc

Real functions

- some approximation
    - mysterious contributions
        - rotatable bonds / solvent entropy …

What happens ?

- 1992 ? Torda group gives up…

Early 1990's

- Optimise a sequence (Monte Carlo / genetic algorithm)
- 1993 – by swapping residues only
- 1994 - persuade sequence composition not to change too much

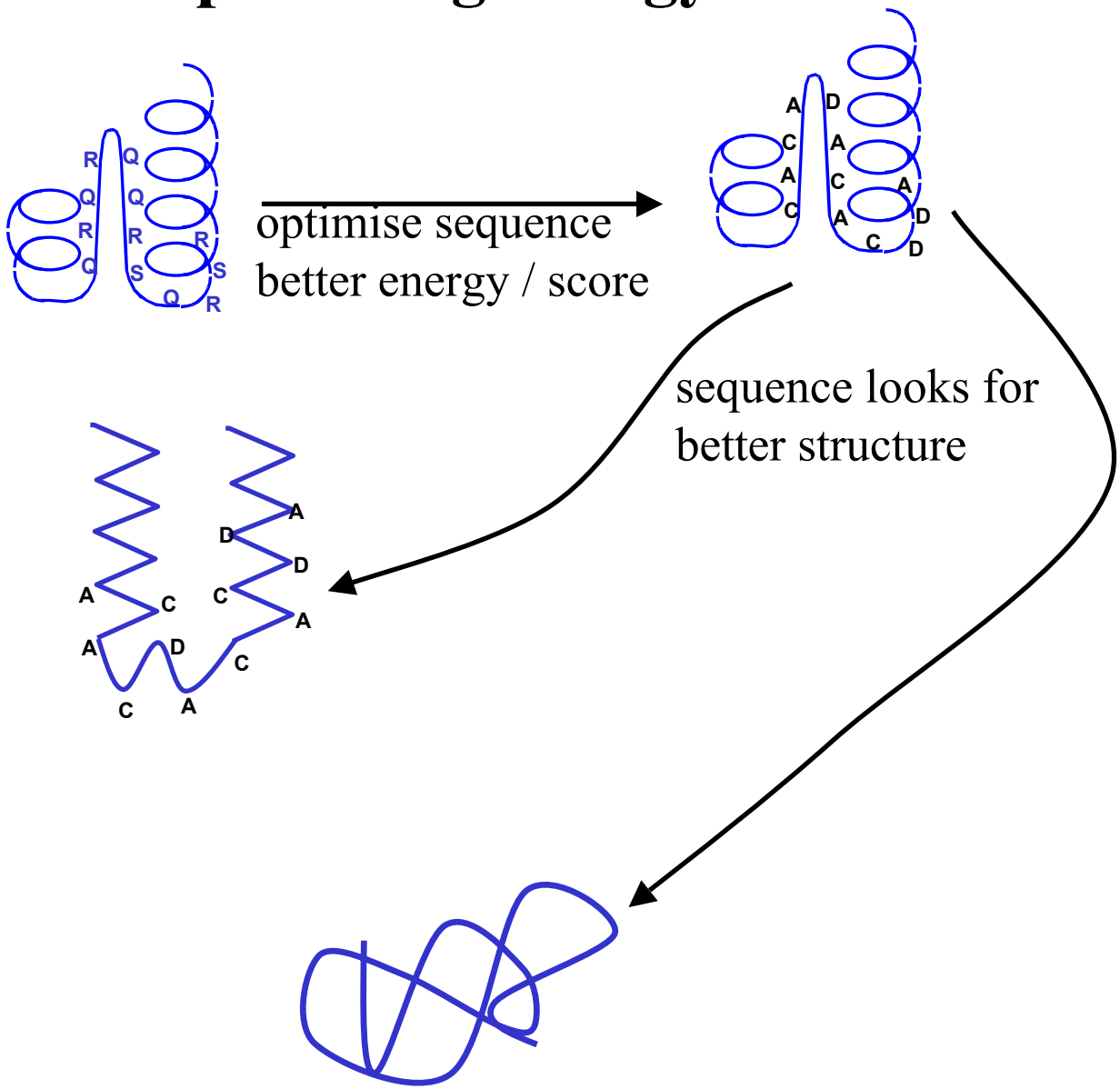
- Justifications ?
  - crazy Ising model analogy
  - composition / class tendency

- what really happens

Shakhnovich and Gutin, Protein Eng, 1993, 6, 793-800
Jones, D Protein Sci, 1994, 3, 567-574

result of optimising energy

- our intention
- maybe …
- more likely



Numerical explanation

My force field / score function / optimisation is tolerant

- of geometric errors

What is most dominant term in score function ?

- hydrophobic interactions
- disulfides
- some special terms
  - prolines at kinks, gly at exotic phi/psi…

Result ?

```
ACDFGAHKLMNPQRSTVW
        ↓
WWWWWDWWWWDWWWWWWW
```

Consequence

- different scoring function

Negative design

Conventional score function

- minimise energy / free energy
- happiness = - U (sequence | given a structure)
- makes better sequences
- sequences look for better structures

Negative design

- happiness = - [U(sequence | given a structure) –
  U (sequence | all other possible structures)]

Is this common in the literature ?

- how have calculators responded

Ignoring negative design

- Original negative design paper 1995
- 1999, "de novo" design (no composition change)
- 2000 – Wodak's "DESIGNER"
  - never allowed more than a few residues to change
- 2001-2002 Serrano, "automatic design"
  - even fewer residues allowed to change
  - …

- Only 5 years after negative design noted
- Why is this so awful ? What would happen if properly searched ?
- Has everyone ignored negative design ?

Godzik, A, Protein Eng. 1995, 8, 409-416.
Koehl, P.; Levitt, M. J. Mol. Biol. 1999, 293, 1161-1181
Wernisch, L.; Hery, S.; Wodak, S.J.. J. Mol. Biol. 2000, 301, 713-736.
Ogata, K.; ….; Wodak, S.J. J. Biol. Chem. 2003, 278, 1281-1290.
Reina, J.; Lacroix, E.; … Serrano, L. Gonzalez, C. Nature Struct. Biol. 2002, 9, 621-627.
De La Paz, M.L.; ...; Serrano, L. J. Mol. Biol. 2001, 312, 229-246.
Fisinger, S.; Serrano, L.; Lacroix, E. Protein Sci. 2001, 10, 809-818.

Not ignoring negative design

Goal

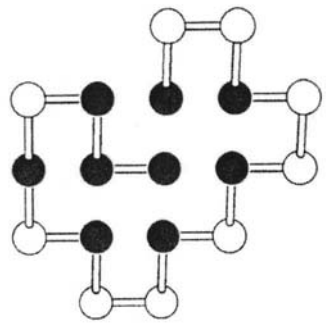
- happiness = - [U(sequence | given a structure) – U (sequence | all other possible structures)]

Requires

- visit every possible conformation
- make sure sequence is happier on native than alternatives

Number of possible conformations ?

- intractable
- can be done on toy systems (lattices)

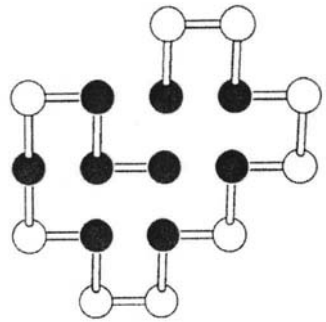


Try to

- identify important alternatives
- simultaneous optimise sequence and structure

TOO HARD

Cunning Goldstein Approach



Magic happiness function

- target structure + trial sequence
- return a number
  - includes effects of ensemble of alternative structures

Demonstration calculation

- lattice system + simple interaction function
  - statistical contact preferences
- vary sequence to minimise energy ?
  - makes lots of HHHH pairs (as discussed)
  - look for new function ..

Chiu, T.L., Goldstein, R.A. Protein Eng. 1998, 11, 749-752.

Magic interaction function

- For toy systems, we can search all structures
  - find lowest energy structure = native conformation

- Use minimisation to search for parameters where
  - preferred sequence scores well
  - native conformation scores better than alternatives

- Result:
  - a new set of interaction parameters

Chiu, T.L., Goldstein, R.A. Protein Eng. 1998, 11, 749-752.

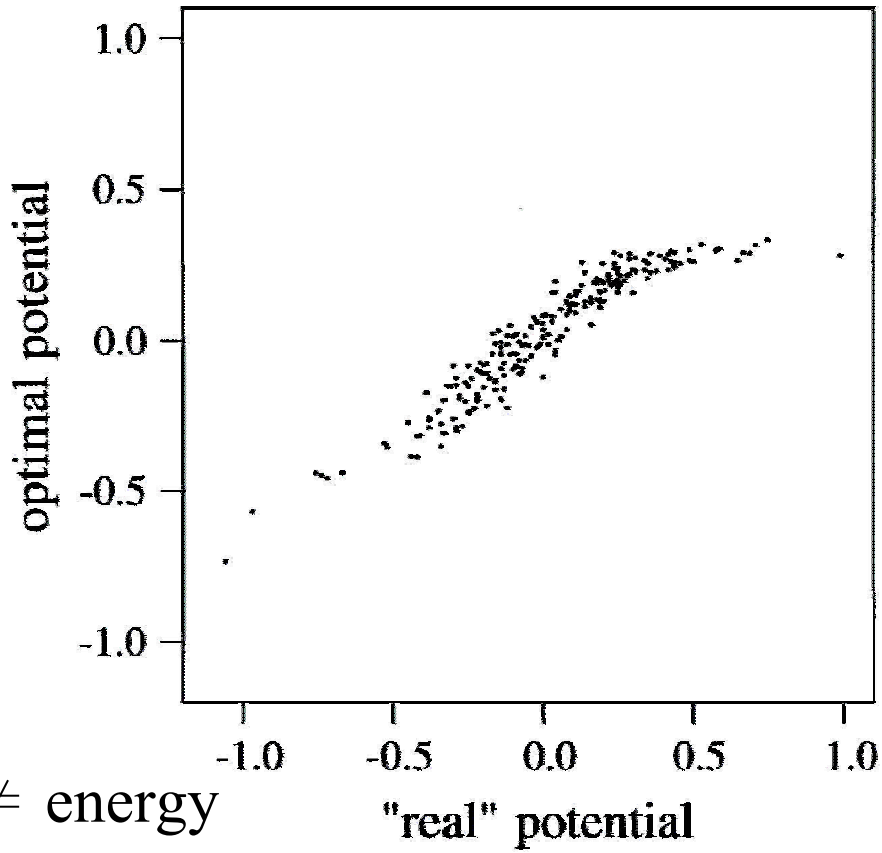
Properties of magic function

What is this function

- NOT a potential energy, free energy, …
- could not be used to predict structure
- a sequence optimisation function

What changes ?

Is this the answer ?



Demonstration of principle

- sequence optimisation function ≠ energy

Chiu, T.L., Goldstein, R.A. Protein Eng. 1998, 11, 749-752.

Real systems ?

- Mostly real laboratory engineering
- Calculations and demonstrations
  - two important examples

Mayo 1997

Mission

- small protein (27 residues) zinc finger
- find a new unrelated sequence which folds to same structure

Calculation

- allow (almost) all residues to change to (almost) anything
- branch and bound algorithm

Force field

- atomistic (slight modifications) + simple solvation

Results ?

Dahiyat, B.I.; Mayo, S.L Science 1997, 278, 82-87

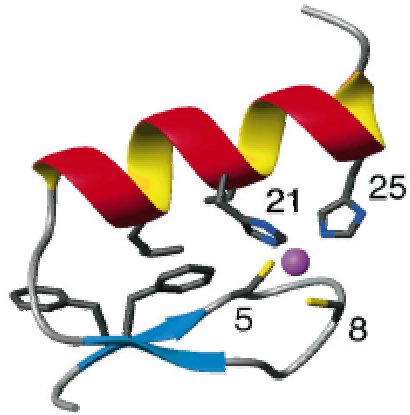
Mayo results

designed **QQYTAKIKGRTFRNEKELRDFIEKFKGR**

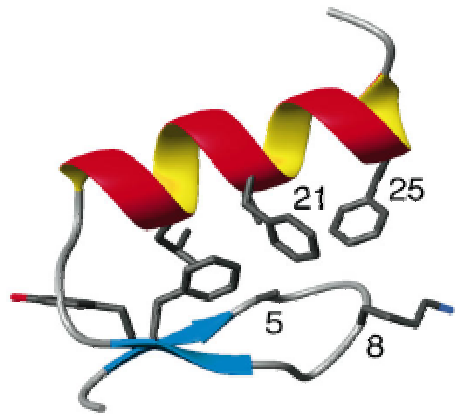
native **KPFQCRICMRNFSRSDHLTTHIRTHTGE**

New sequence

- about 20 % similar to start
- not related to any known protein (still)



- Structure solved by NMR

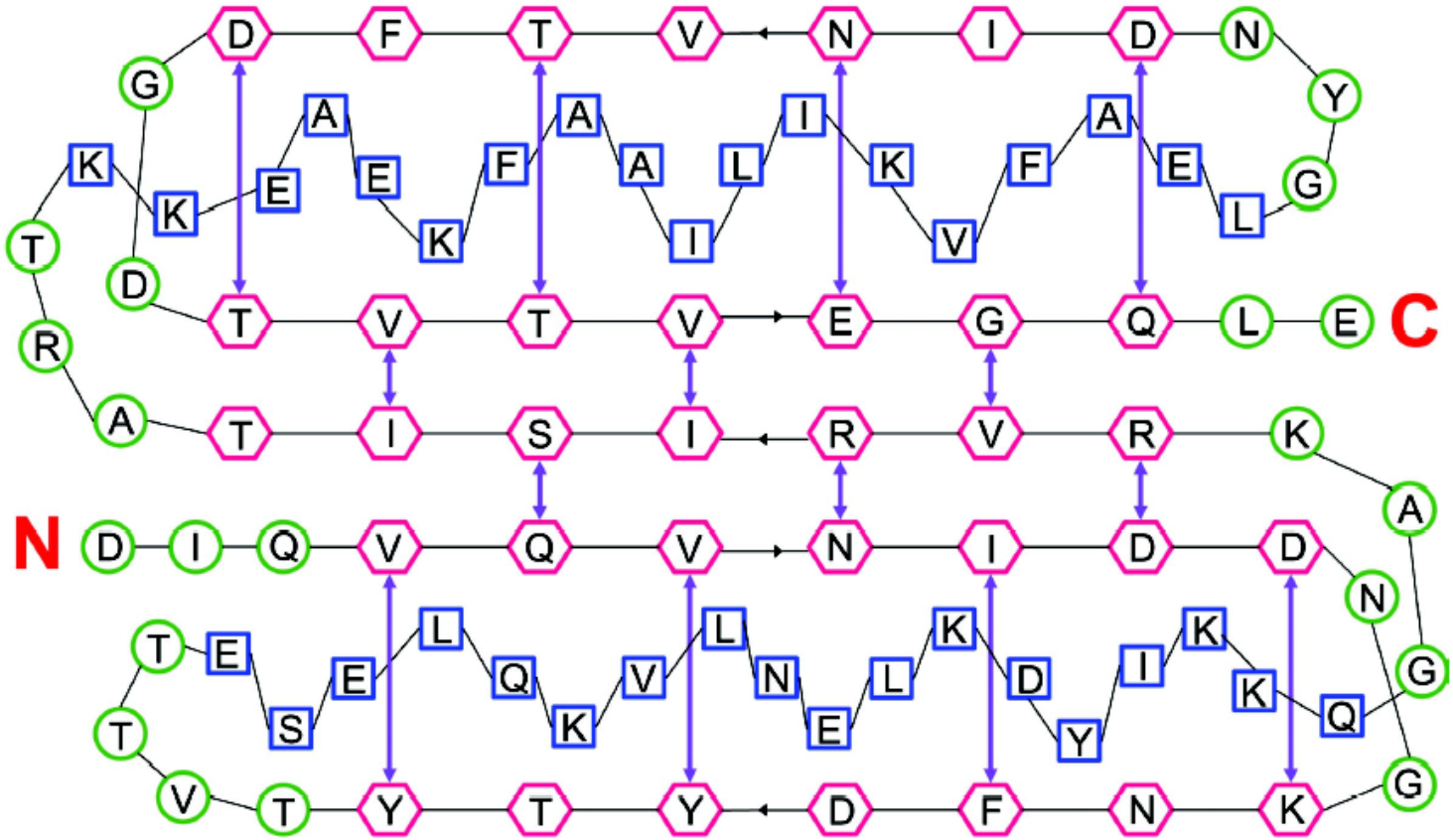


- Problem solved ?
- What was the secret
- More examples ?

Baker late 2003

Mission

- sketch a new protein topology
- build a sequence to fit it



Kuhlman, B.; Dantas, G.; Ireton, G.C.; Varani, G.; Stoddard, B.L.; Baker, D. Science 2003, 302, 1364-1368.

Methods

Generate coordinates from sketch

Simple Monte Carlo of sequence + some geometry

Force field

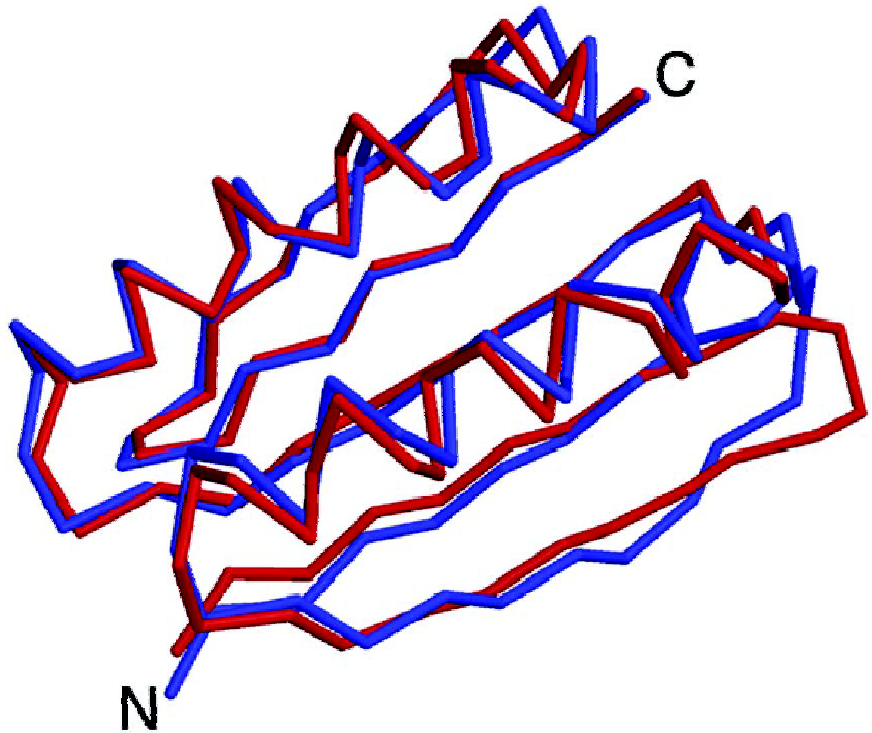
- atomistic
- more..

Results ?

Results

Find a sequence
- not like any known

Structure
- as predicted
- solved by X-ray
  - neat phasing trick !



100 % success ?
- not quite
- room for improvement (important)

Kuhlman, B.; Dantas, G.; Ireton, G.C.; Varani, G.; Stoddard, B.L.; Baker, D. Science 2003, 302, 1364-1368.

Implications

Score function

- modified Lennard-Jones
- rotamer preferences
- solvation approximation
- explicit Hbonds
- "statistical electrostatics"
- composition bias

- double counting
- potential energy, free energy, potentials of mean force …

What have we learnt ?

Bad

- Any conventional force field is a disaster
- negative design is essential
- fraudulent literature outweighs real results

Good

- functions do exist which sometimes work
- negative design can be implicit / accidental
- good results with an ugly score function

Future

- much room for improvement
- identification of important properties