Protein design

[Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L. and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science, 302, 1364-1368]

What about folds that are **not** seen in SCOP or CATH?

Some are:

- physically impossible;
- not yet sampled by evolution;
- not observed by a structural biologist.

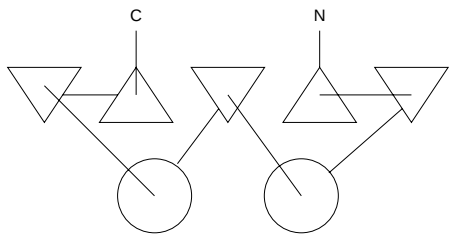
Goal was to achieve a highly stable protein with a new fold.

Approach to designing Top7 sequence

```
for i = 1 to 172 {
    generate starting structure;
    for j = 1 to 5 {
        for k = 1 to 15 {
            optimise sequence for fixed backbone;
            optimise backbone coordinates for fixed sequence;
        }
    }
}
```

Starting models are generated using a *de novo* approach ("Rosetta").
Assemble fragments taken from known structures.
Scoring function includes distance constraints from 2-D diagram.

TOPS cartoon of Top7



Optimise sequence

polar amino acid at the 22 surface β-sheet positions
(=> 75 rotamers per position)

any amino acid (except Cys) at the other 71 positions
(=> 110 rotamers per position)

Find combination of rotamers (and hence the sequence) with the lowest energy, using Monte Carlo search.

Optimise structure (1)

Measure energy

(i) Perturb backbone

a) choose between 1 and 5 residues at random and make small random adjustments to their main-chain torsion angles (ϕ,ψ),

or

b) replace the backbone of 1, 2 or 3 consecutive residues with a randomly selected fragment from the PDB, and adjust torsions of neighbouring residues to minimise the displacement of the downstream part of the chain.

Optimise structure (2)

(ii) Optimise side-chain structure

for those positions with higher energy after (i), replace current side-chain conformation with lowest energy rotamer.

(iii) Optimise backbone structure

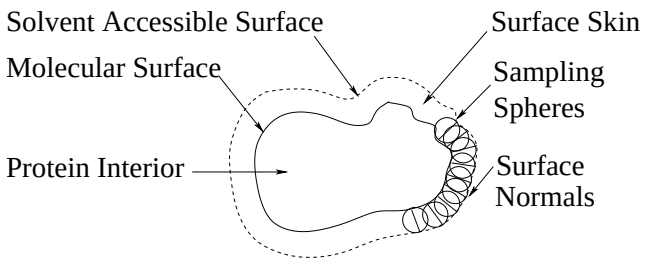
optimise ϕ and ψ again in a 10-residue window around the perturbation site.

Measure energy again, and use Metropolis criterion to decide whether to accept or reject.

Steps (i), (ii) and (iii) are repeated several thousand times.

After every 20 such moves, a full combinatorial optimisation of side-chain rotamer conformations was carried out.

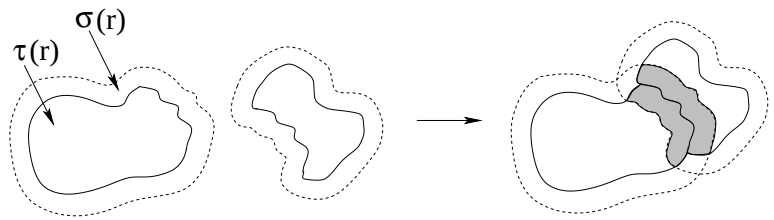
3D protein shape density representation in Hex



$$\sigma(\underline{r}) = \begin{cases} 1; & \underline{r} \in \text{surface skin} \\ 0; & \text{otherwise} \end{cases} \qquad \tau(\underline{r}) = \begin{cases} 1; & \underline{r} \in \text{protein atom} \\ 0; & \text{otherwise} \end{cases}$$

[Ritchie & Kemp (2000) Proteins **39:178–194**]

Protein shape complementarity



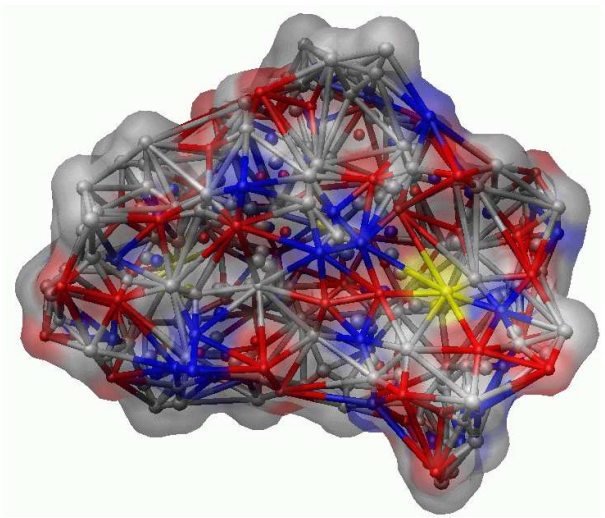
Favourable: $\int(\sigma_A(\underline{r}_A)\tau_B(\underline{r}_B) + \tau_A(\underline{r}_A)\sigma_B(\underline{r}_B))\mathrm{d}V$

Unfavourable: $\int \tau_A(\underline{r}_A)\tau_B(\underline{r}_B)\mathrm{d}V$

Score: $S_{AB} = \int(\sigma_A\tau_B + \tau_A\sigma_B - Q\tau_A\tau_B)\mathrm{d}V$

Penalty Factor: $Q = 11$

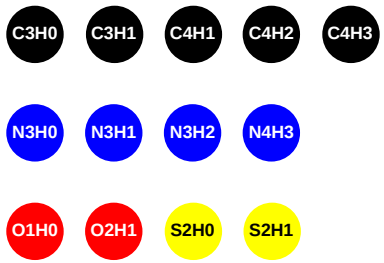
Surface representation



Atomic group in proteins

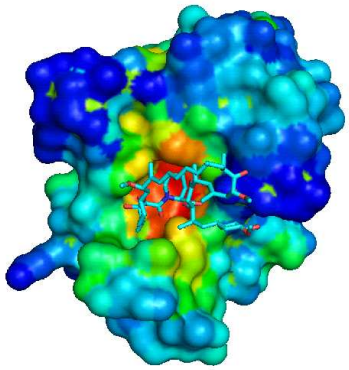
Classifcation proposed by Tsai et al. (J. Mol. Biol., 1999, 290:253-266), based on:

- heavy-atom types,
- the number of covalently attached hydrogen atoms, and
- the number of all covalently attached atoms.

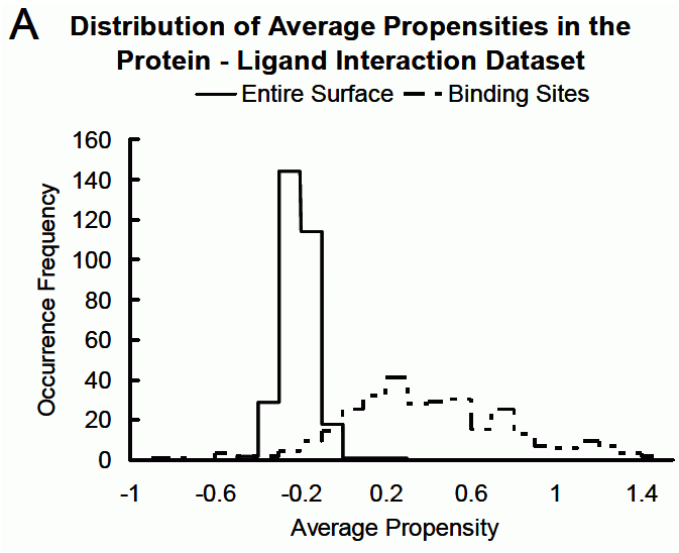


Surface Triplet Propensities

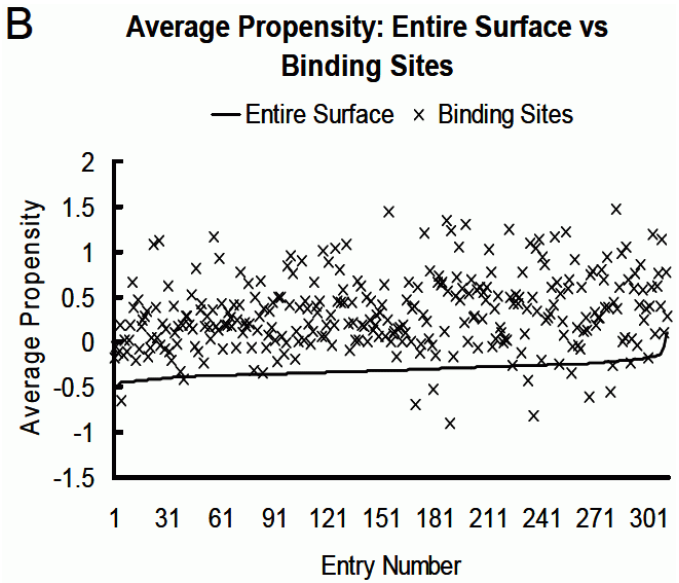
Mehio, W., Kemp, G.J.L., Taylor P. and Walkinshaw, M.D. (2010) Identification of Protein Binding Surfaces using Surface Triplet Propensities.
Bioinformatics



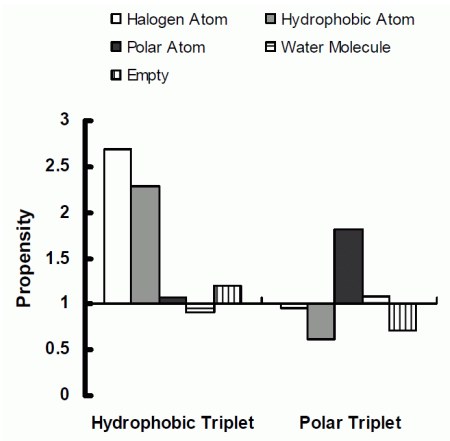
Surface Triplet Propensities



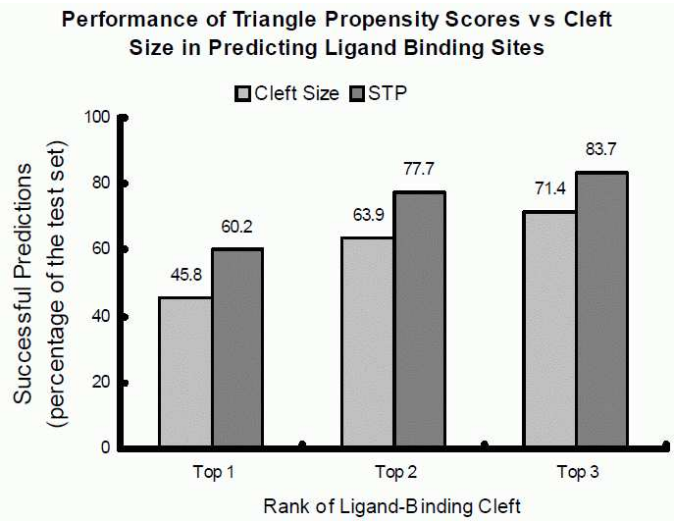
Surface Triplet Propensities



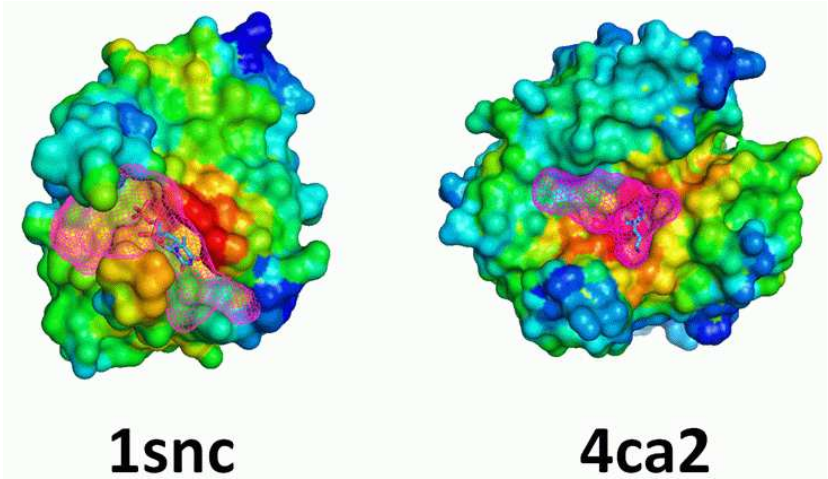
Surface Triplet Propensities



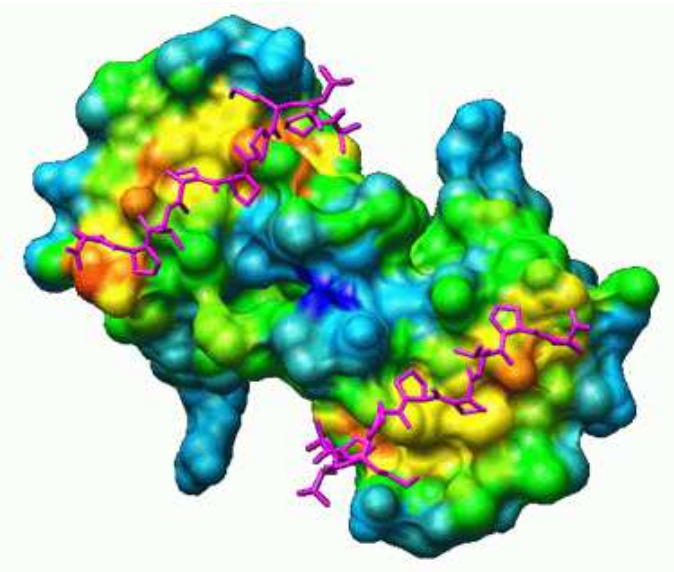
Surface Triplet Propensities



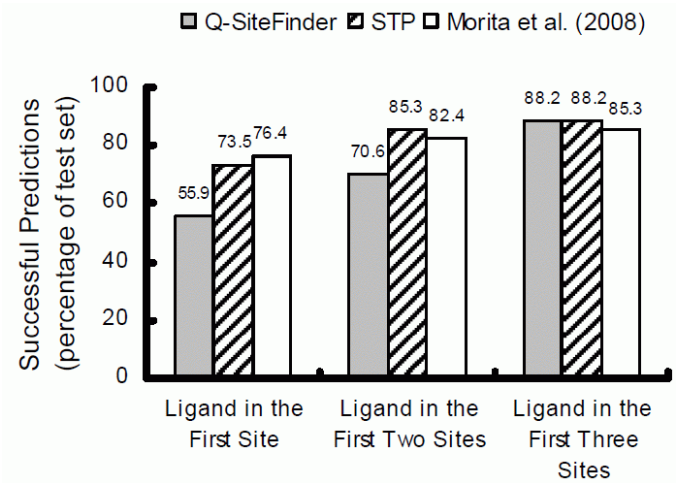
Surface Triplet Propensities



Surface Triplet Propensities



Surface Triplet Propensities



CombDock

[Inbar, Y., Benyamini H., Nussinov R. and Wolfson H.J. (2005) "Prediction of multimolecular assemblies by multiple docking". *J. Mol. Biol.*, **349**, 435-447]

- All pairs docking
  - — $N(N-1)/2$ pairs
  - — keep best K transformations for each pair
- Combinatorial assembly
  - — find best spanning tree representing a valid complex
  - — keep best D trees of size s starting at i
- Rescoring
  - — cluster (to avoid redundancy in solution set)
  - — geometric component
    - — large interface area and small steric overlap
  - — physico-chemical component
    - — count number of buried non-polar atoms

CombDock results (1)

Yeast RNA polymerase II elongation complex

10 protein chains

$K = 100$, 15 pairwise interactions predicted

10^{26} possible complexes

50188 complexes generated by combinatorial assembly

1113 complexes left after clustering

2nd ranked complex had RMSD of 1.37.sp 2 Human subunits modelled by homology
6th ranked complex had RMSD of 1.9.

How many spanning trees?

(i) If we have N vertices and 1 edge between each pair there are N^{N-2} spanning trees.

(ii) If we have K edges between each pair of vertices, then there are K^{N-1} graphs of type (i).

So there are $N^{N-2}K^{N-1}$ spanning trees.

Can't search the whole space!
So use a heuristic solution.

CombDock results (2)

Bovine arp2/3 complex

7 protein chains

$K = 100$

1.68×10^{17} possible complexes

5488 complexes generated by combinatorial assembly

145 complexes left after clustering

3nd ranked complex had RMSD of 1.2.sp 2 *Drosophila melanogaster* subunits modelled by homology
10th ranked complex had RMSD of 1.9.