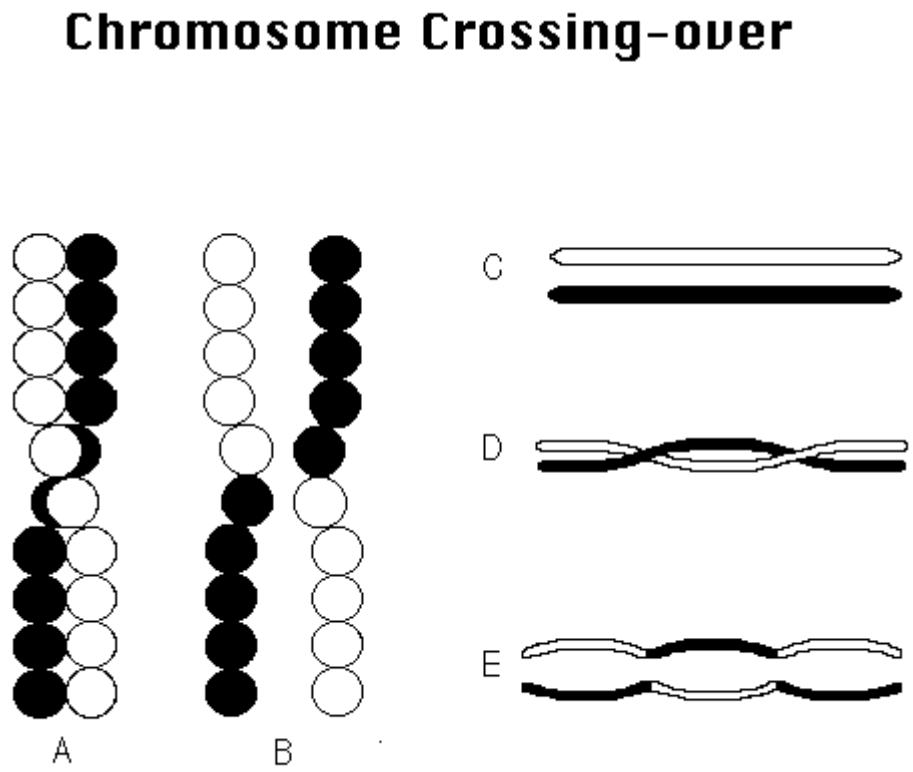
in silico protein recombination applied to Comparative Modelling

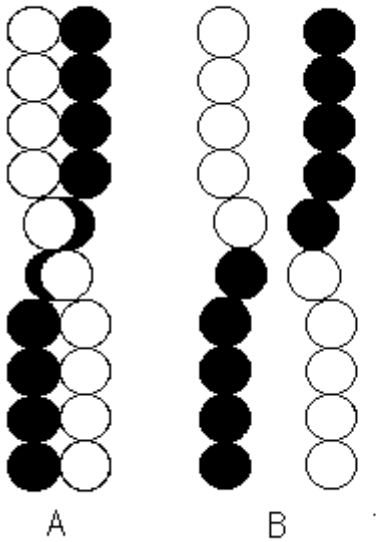
Bruno Contreras-Moreira,
Paul W. Fitzjohn and Paul A.Bates
Biomolecular Modelling Laboratory
London Research Institute
SAC-CASP5, December 2002

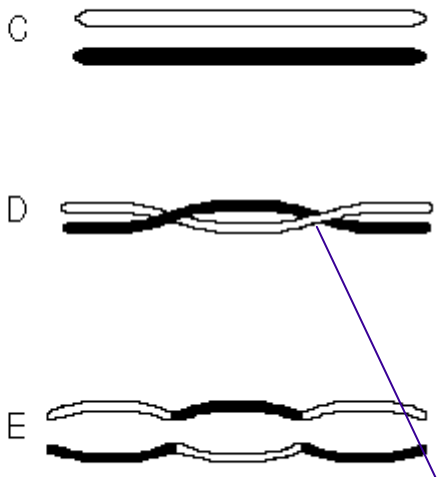
CANCER RESEARCH UK

the biological inspiration



the computational analogy: genetic algorithms





encoded solution

1111000111111111001

0010010000100111001

1111000000100111001

0010010111111111001

crossover point

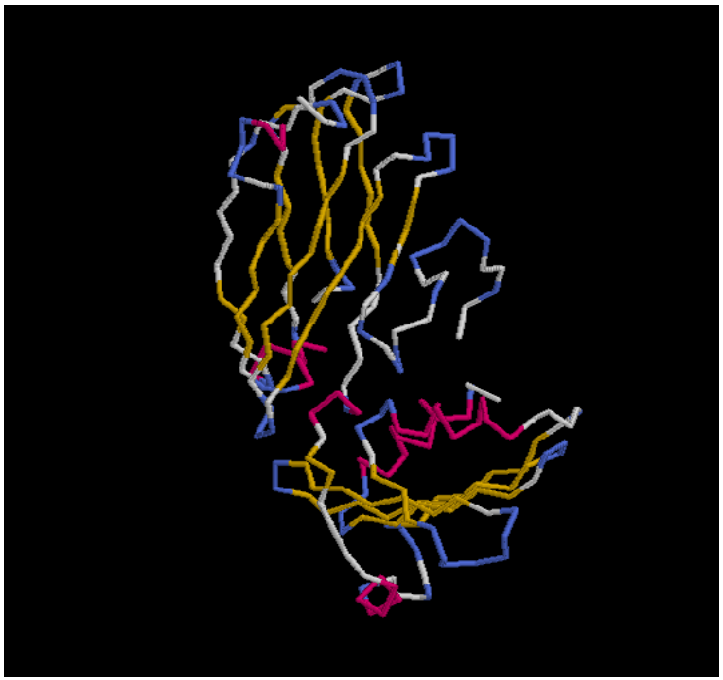
a genetic algorithm applied to Comparative Modelling

- how are solutions coded?
- genetic operators
- definition of fitness
- design of the algorithm

proteins models are implicitly coded solutions

- **linear molecules**: arrays of residues connected by peptide bonds

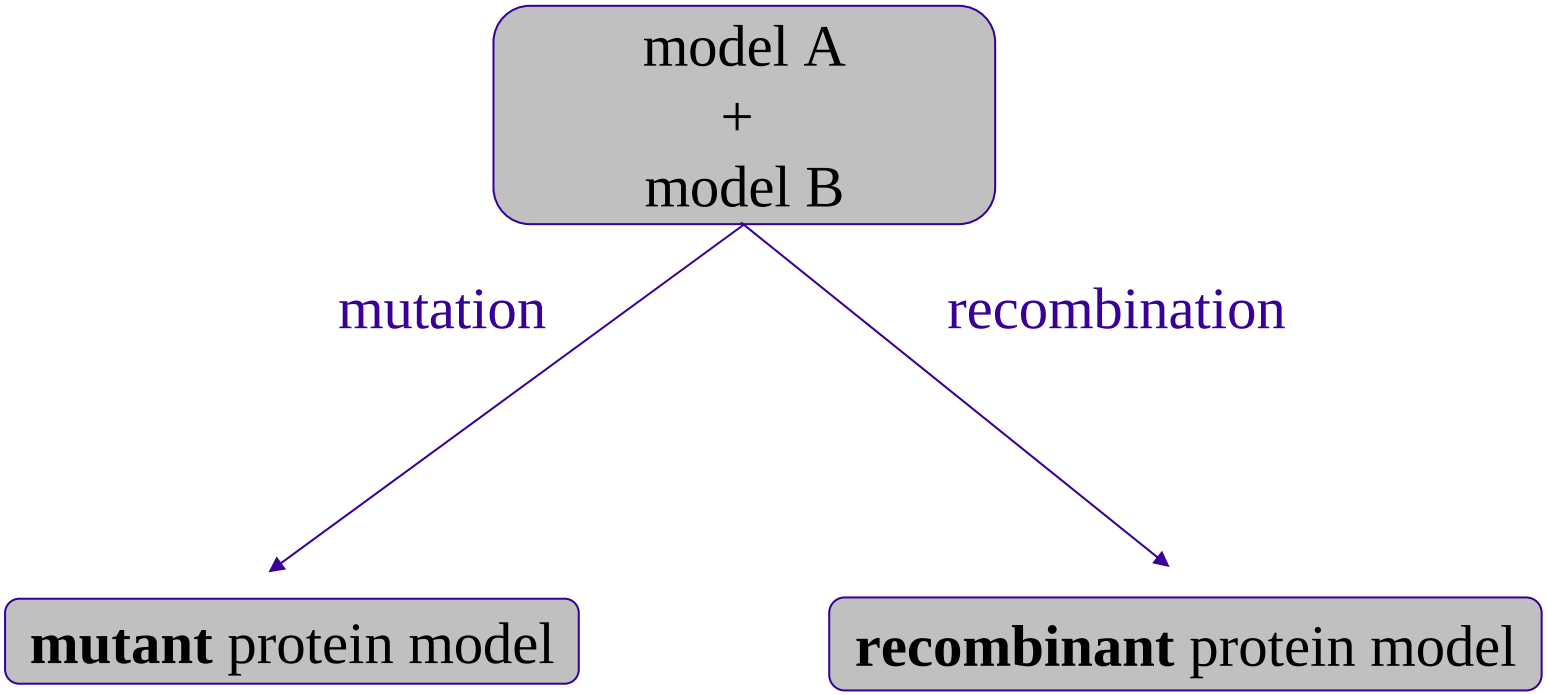
•**fitness** = likelihood of its fold



```
T0134   GEP-VQNGAPEEE--QLPPESSYSLLAENSYVKMTCDIRGSLQEDSQVTVAIVLENRSS
1qts_A GSPGIRLGSSEDNFARFVCKNNGVLF-ENQLLQI--GLKSEFRQNLG-RMFIFYGNKTS
SS      CCCCCCCCCCCCCCHHHHCCCCCEEEE-ECCCEEE--EEEEEEECCEE-EEEEEEEECCC
```

(1model = 1PDB template + 1alignment)

genetic operators



recombination

model A
+
model B

•sequence alignment
•superimpose on Cβ of equivalent residues
•refine fit on close equivalent residues (2 x Cα-Cβ)
•draw crossover point (!helix && !strand, after STICK)

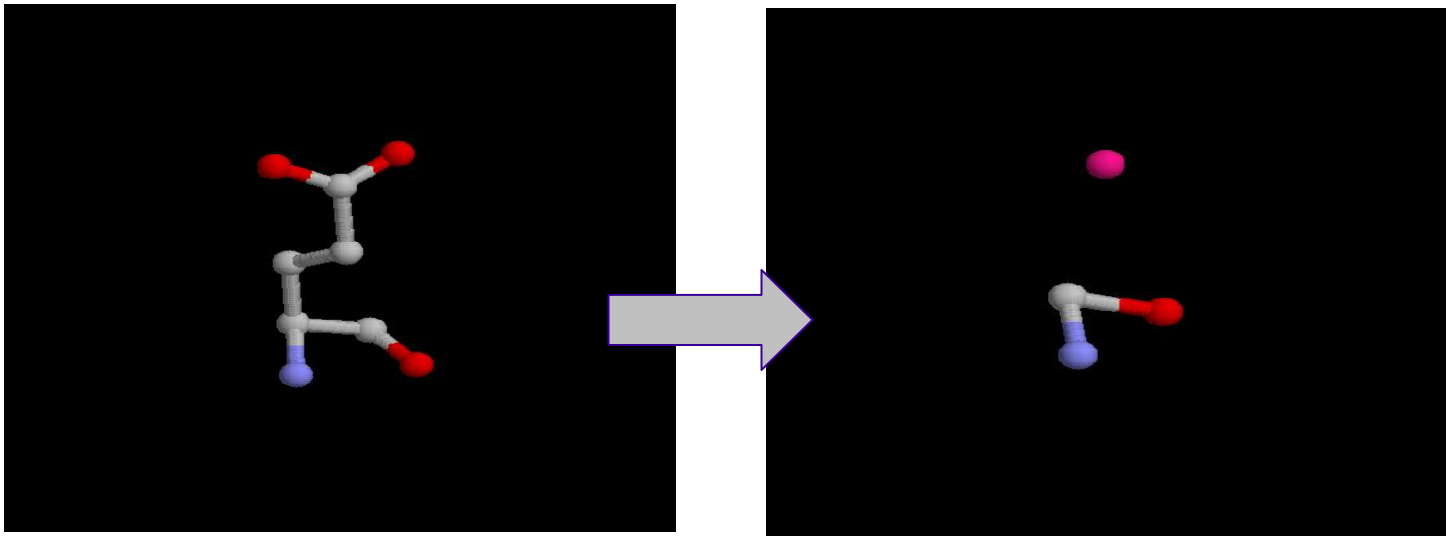
mutation

model A
+
model B

•sequence alignment
•superimpose on Cβ of equivalent residues
•all-atom Cartesian average (no checks)

protein fitness

fitness(p) = **internal_contacts(p)** + solvation(p)



$\sum_i \sum_j (A_{ij}/r_{ij}^{9}) - (B_{ij}/r_{ij}^{6})$ (in Kcal/mol)

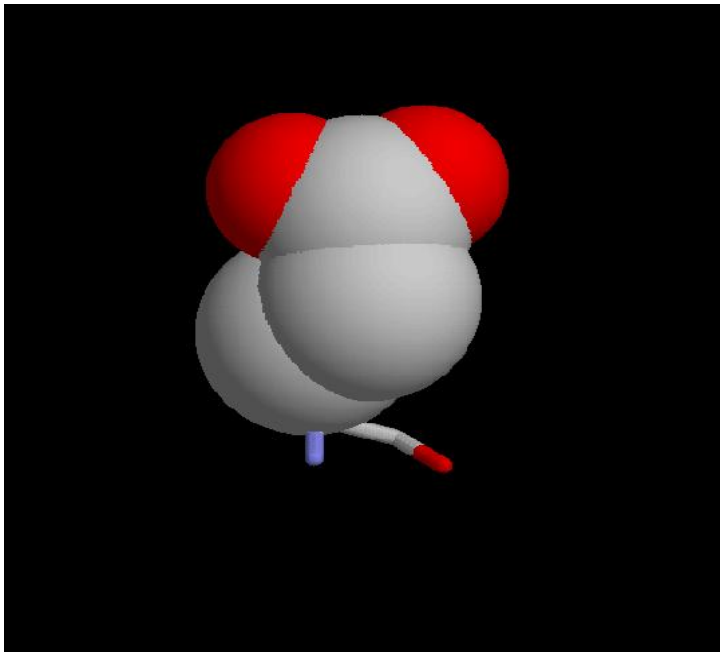
where i,j are pairs of pseudoatoms in protein p

and A and B are statistical potentials

(taken from Robson and Osguthorpe (1979) *J.Mol.Biol.***132(1)**:19-51, code by Paul Fitzjohn)

protein fitness

fitness(p) = internal_contacts(p) + **solvation(p)**



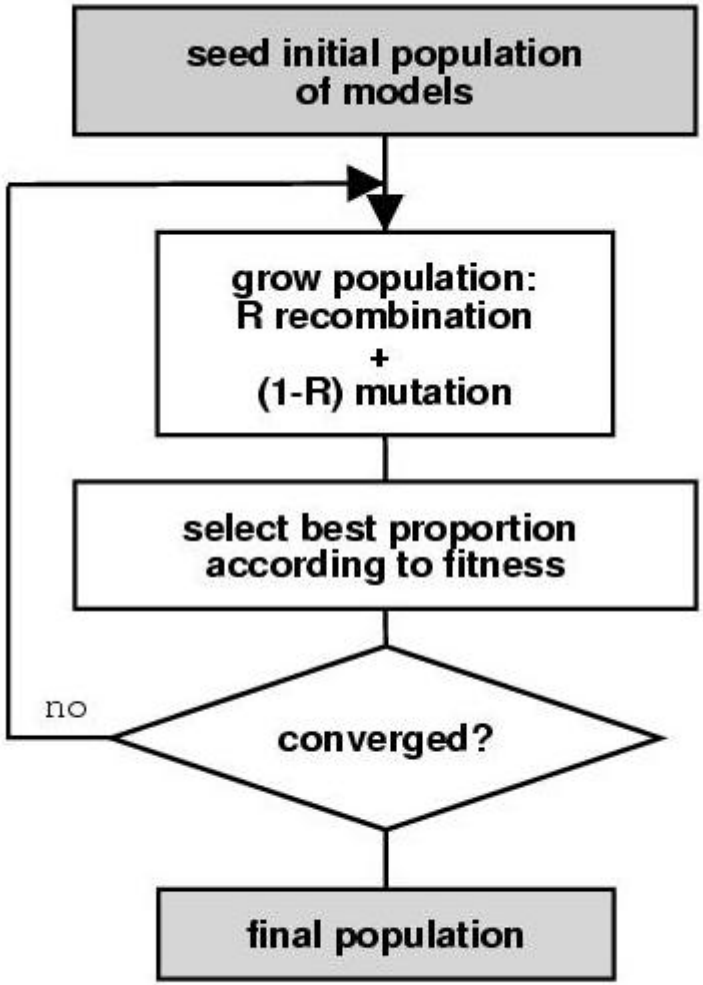
$\sum_i (SA_i \,.\, \Delta Gsolv_i)$ (in Kcal/mol)

where i is a residue in protein p, SA is the side-chain solvent accessible area calculated by NACCESS* and ΔGsolv¶ is the experimental solvation free energy change for each residue type

* NACCESS (Hubbard and Thornton see http://wolf.bms.umist.ac.uk/naccess

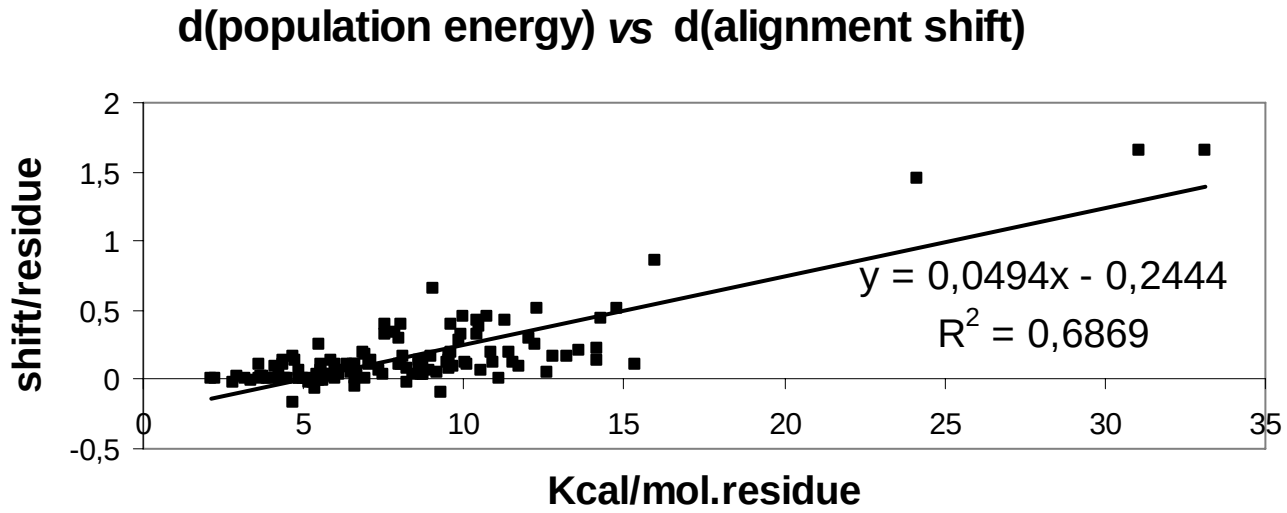
¶ Eisenberg and MacLachlan (1986) *Nature*, **319**: 199-203.

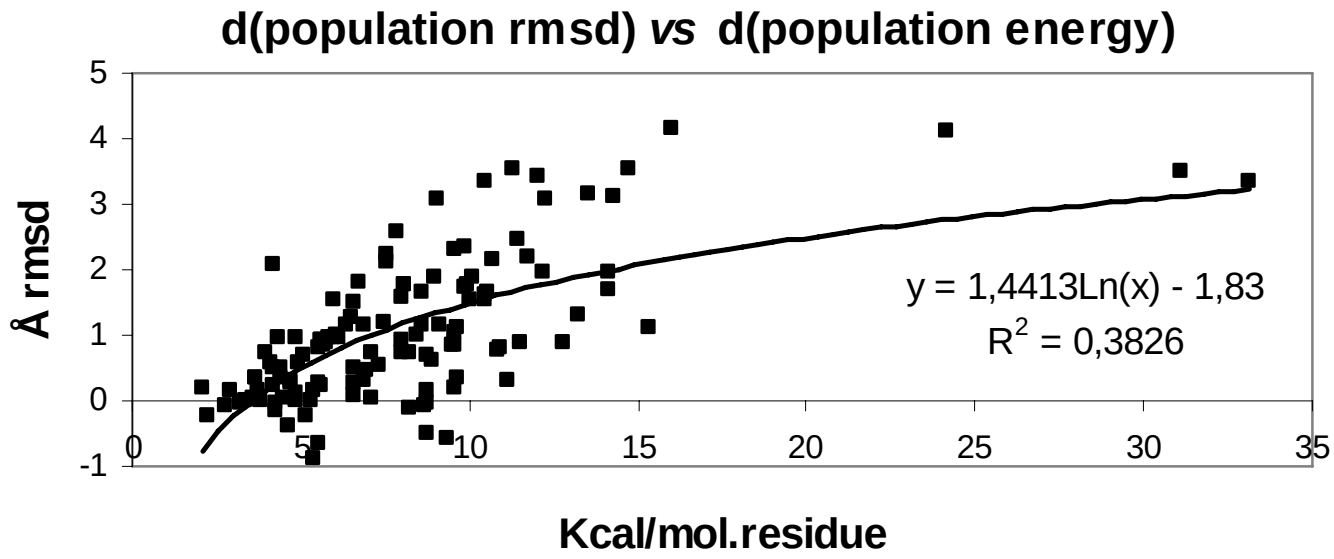
in silico protein recombination algorithm



in silico protein recombination: performance

(benchmark on 130 SCOP families)





in silico protein recombination: evaluation

PROBLEMS

• models in the last population have sometimes **broken loops**

•models need often to be **minimized** after the simulation

•longer **computing time** than traditional methods

•current **mutation** implementation does not help much

ADVANTAGES

•converges close to the best initial model

•is able to recover alignment errors

•often last population contains different conformations

in silico protein recombination: example T0134

INITIAL POPULATION

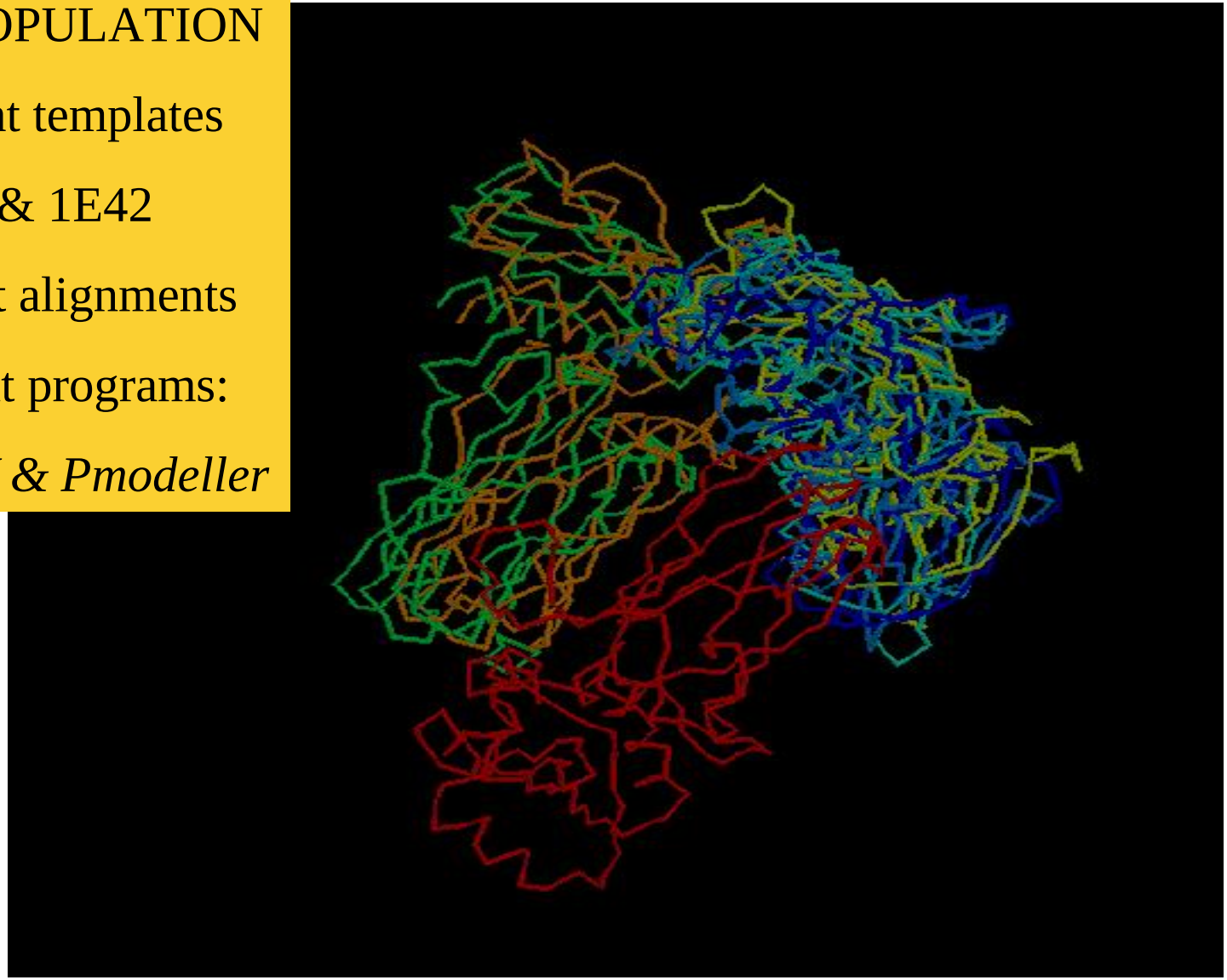
•2 different templates

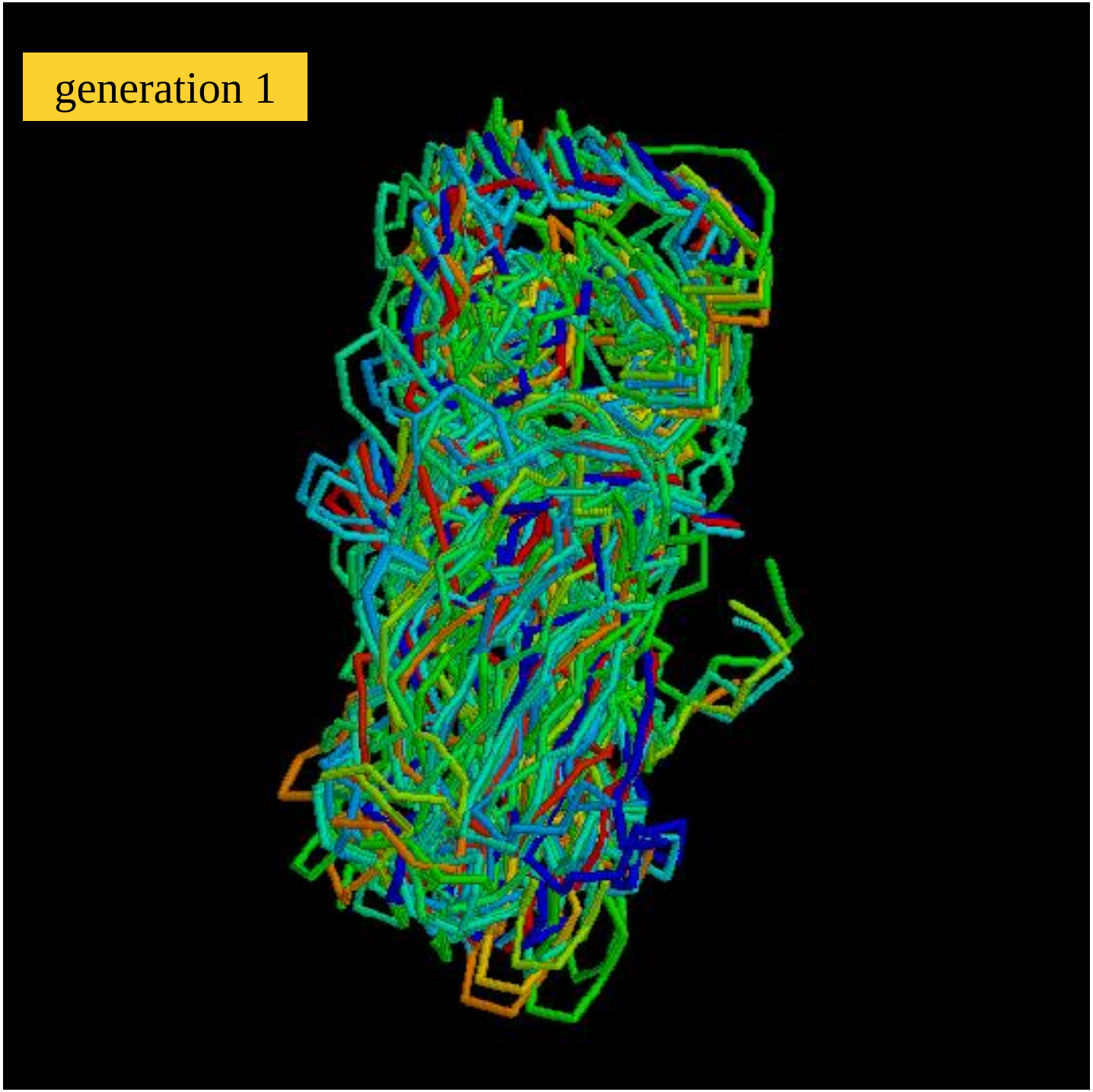
1QTS & 1E42

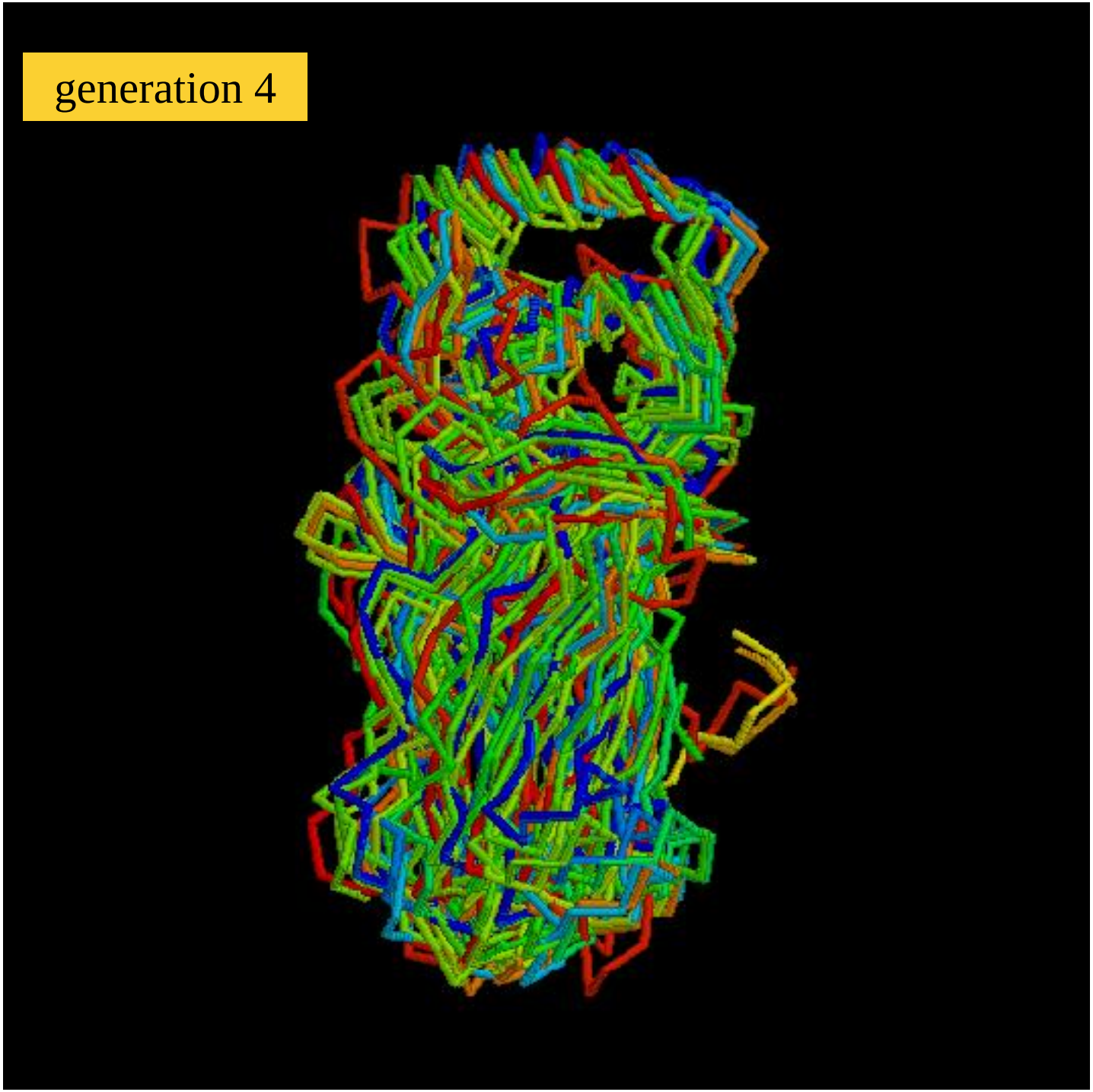
•8 different alignments

•2 different programs:

3D-JIGSAW & Pmodeller

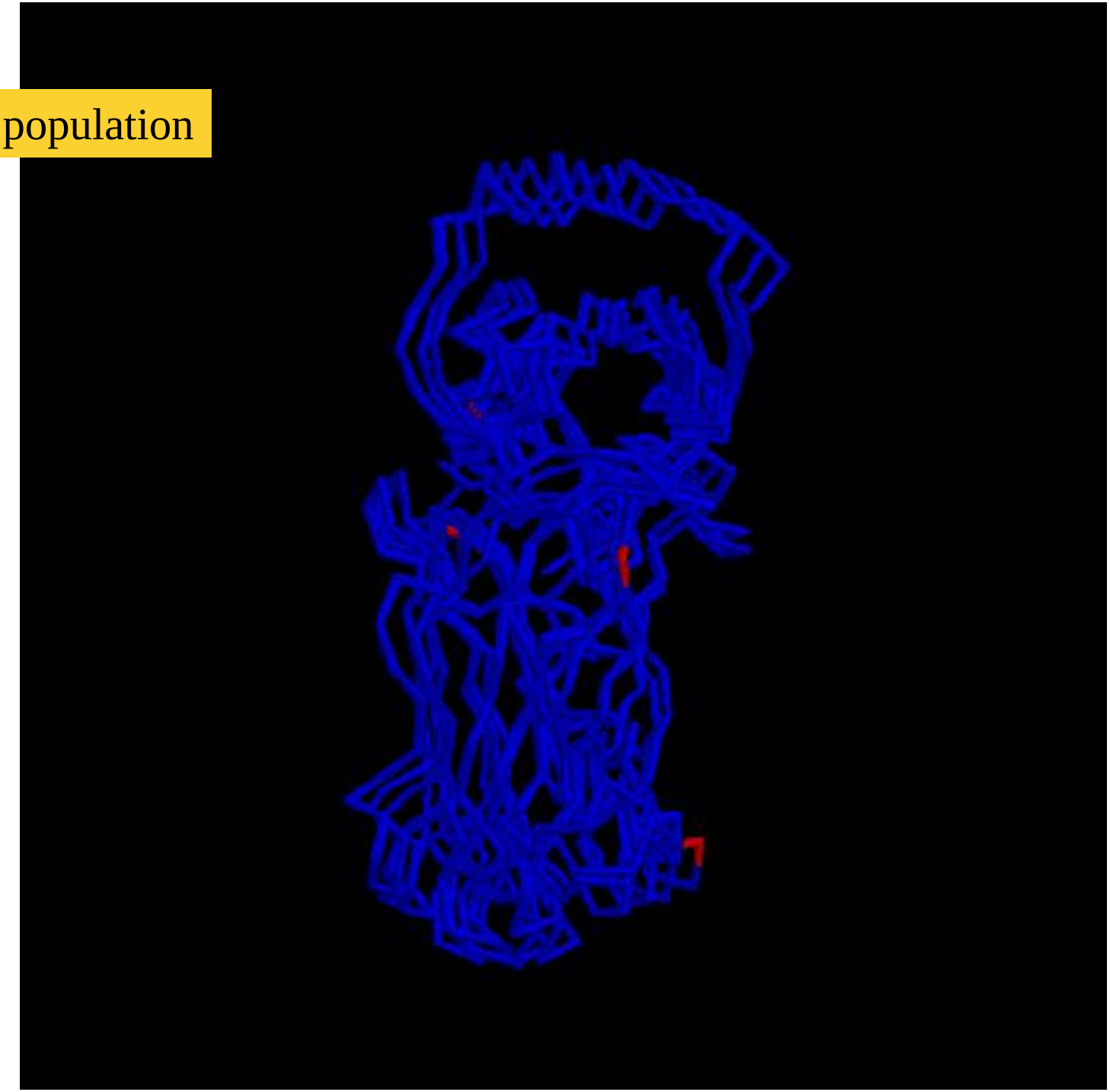


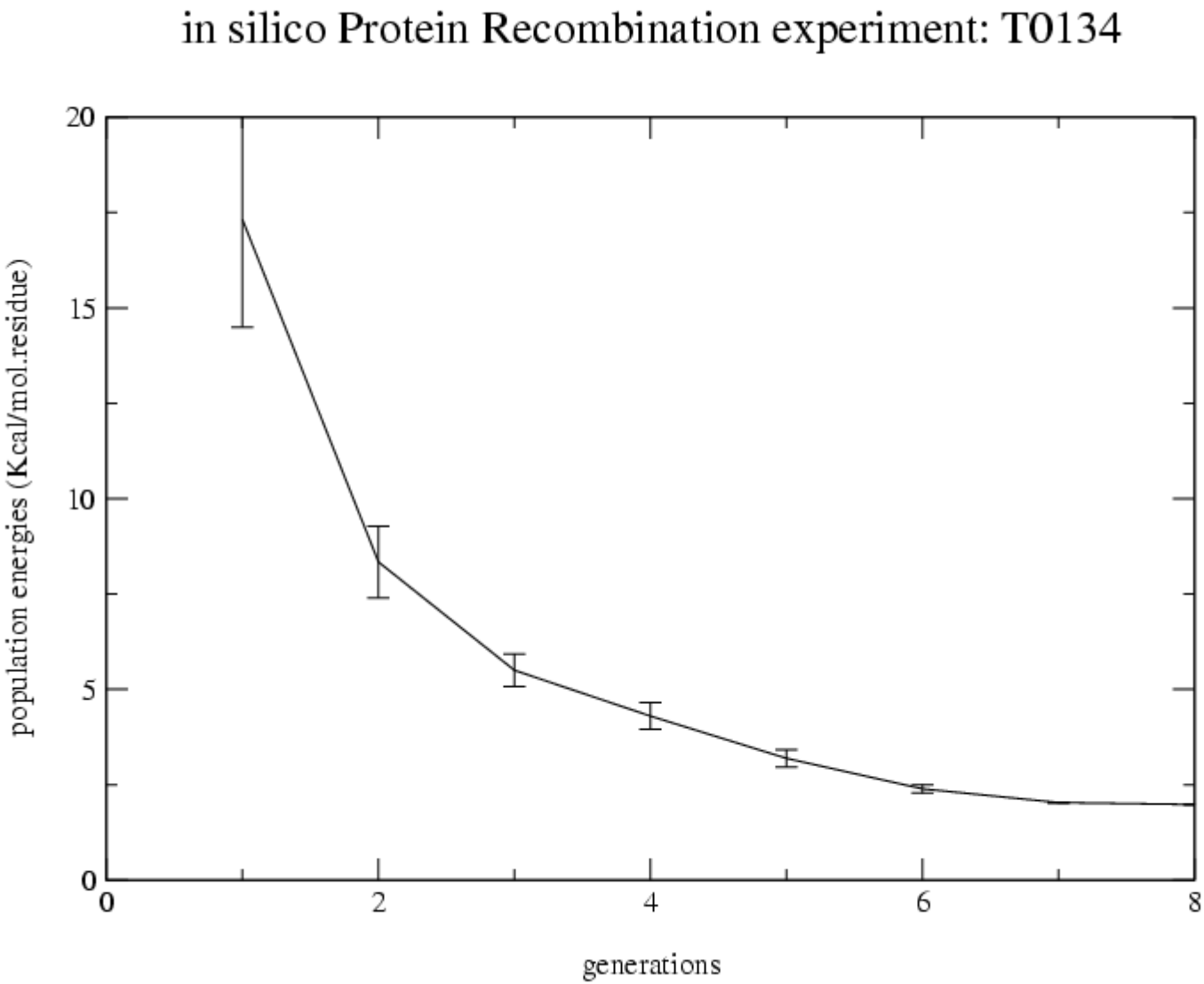

generation 1


generation 4

generation 7

last population




in silico Protein Recombination experiment: T0134
population energies (Kcal/mol.residue)
generations
0
5
10
15
20
0
2
4
6
8

in silico protein recombination (test version)

http://www.bmm.icnet.uk/3djigsaw/recomb

Description

This program performs artificial selection (through recombination + mutation) over a population of protein atomic models seeded by the user, with the aim of obtaining a ore accurate and energetically favourable atomic conformation than any starting model but based on all. So please make sure that your input file contains only models for the same protein. Enjoy yourself.

job identifier: your_job
your e-mail address:

Please specify a file (PDB format, TER-minated chains):

Browse...

or

... paste your PDB coordinates here:

pop size 30 selected prop 0.75 forced improvement 0.001 last gen 50 mutation rate 20

Upload and Recombine! Reset help

contrera@cancer.org.uk BMM disclaimer

Thanks to
the Biomolecular Modelling Laboratory

Paul Bates

Paul Fitzjohn

Pall Jonsson

Graham Smith

Chris Page

Marc Offman

www.bmm.icnet.uk

Thanks to Cancer Research UK and
to SAC-CASP5 organizers

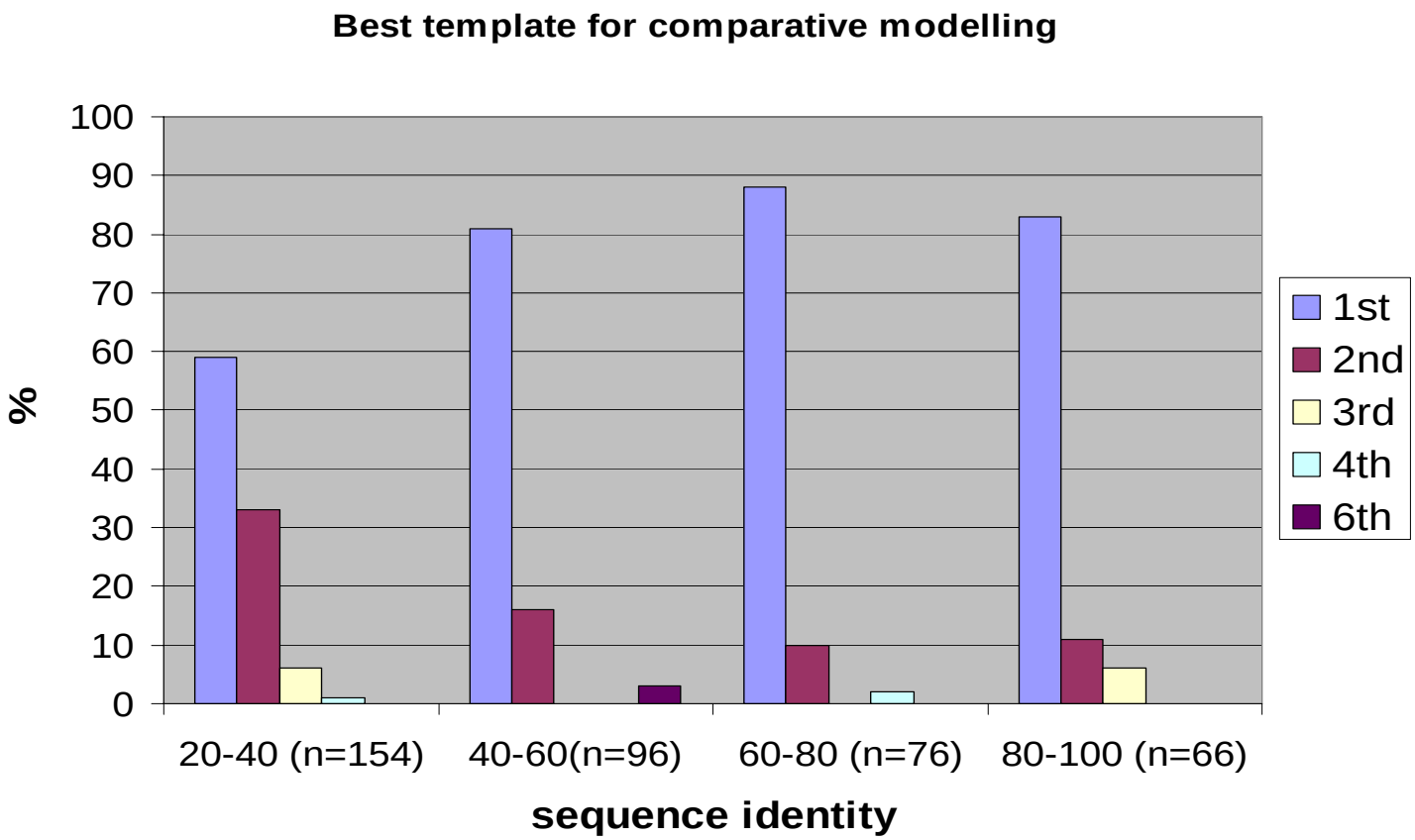
Possible applications of comparative modelling*

Depending on the sequence identity between query and template:

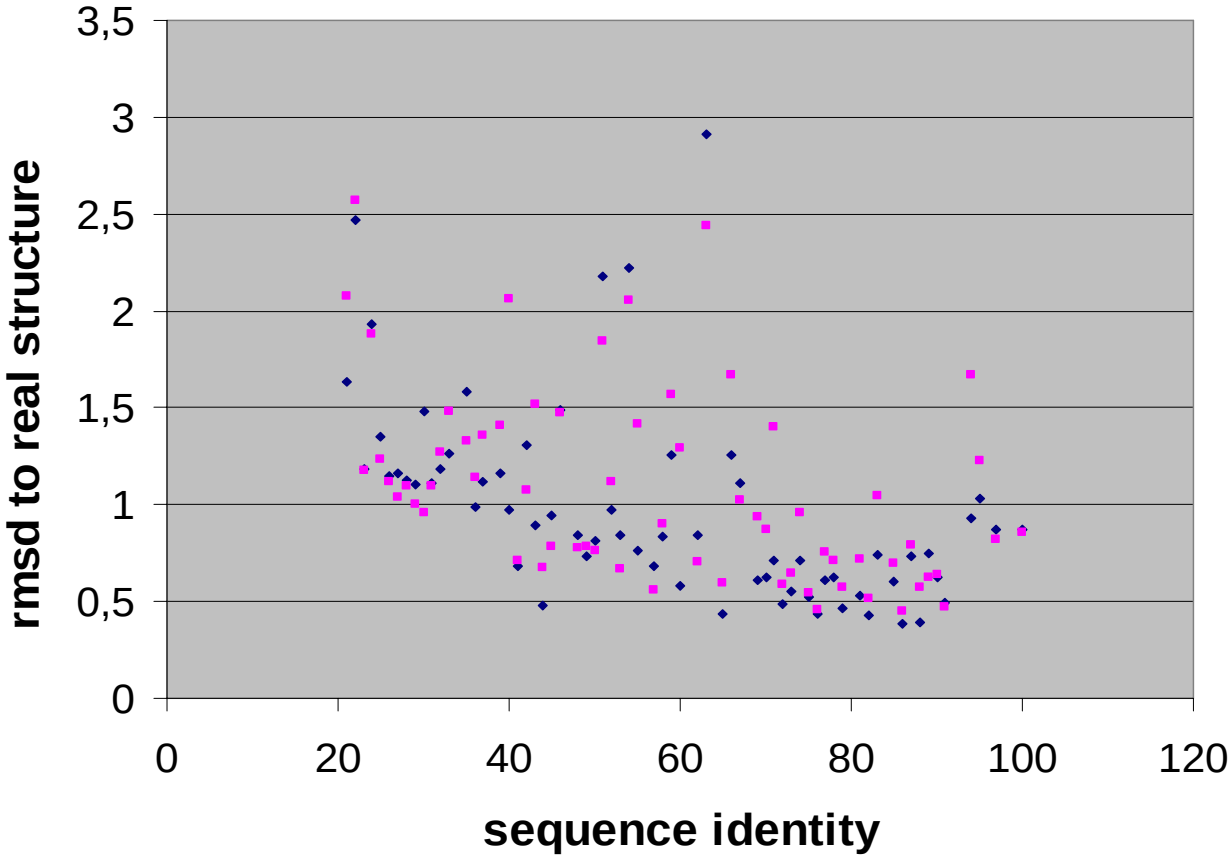
- >**90%** virtual ligand screening
- >**40%** defining antibody epitopes
- >**40%** molecular replacement in X-ray crystallography
- >20% support site directed mutagenesis
- >20% fitting into low resolution electron density maps

* Baker & Sali (2001) Science 294: 93-96

Selecting templates



Single *vs* Multiple template modelling

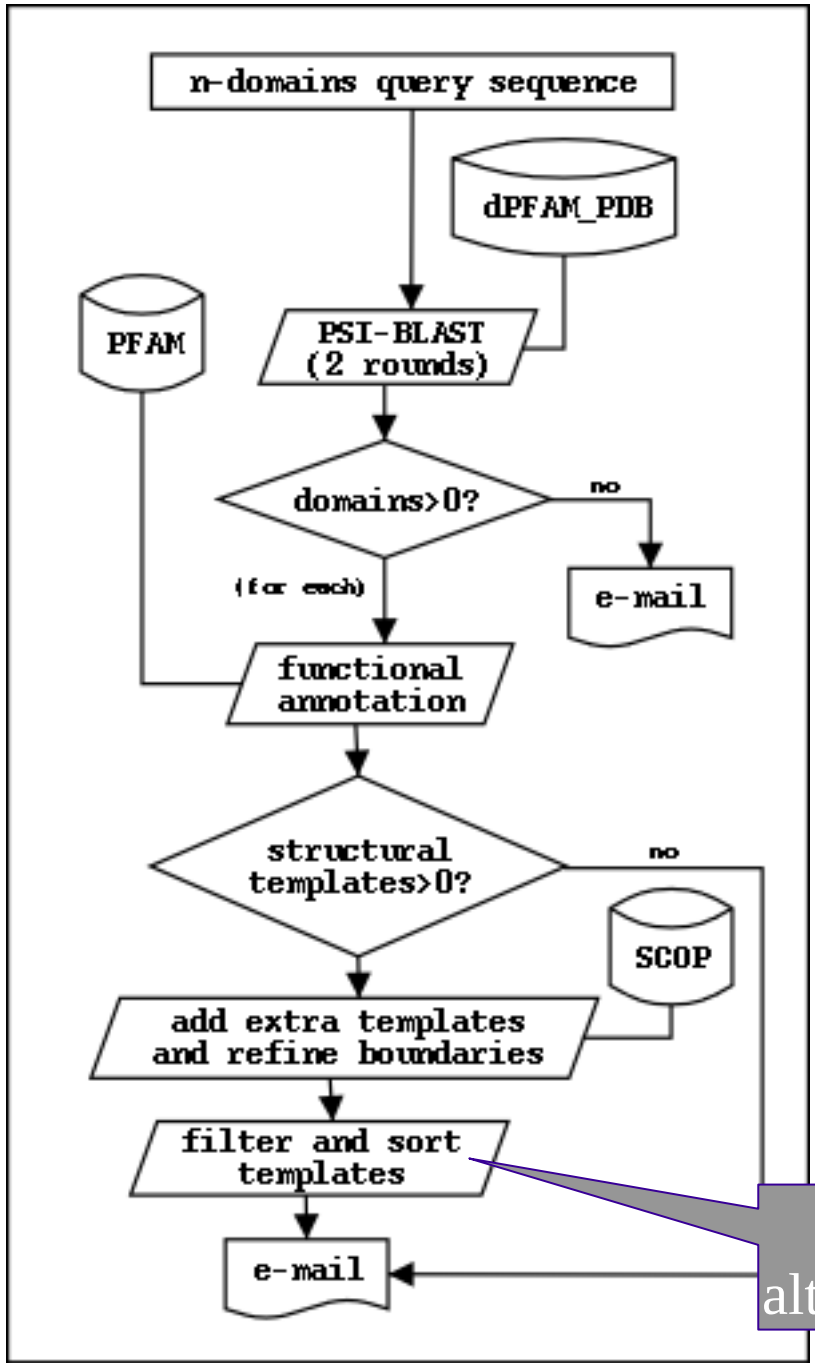


Overall:

54/97 S

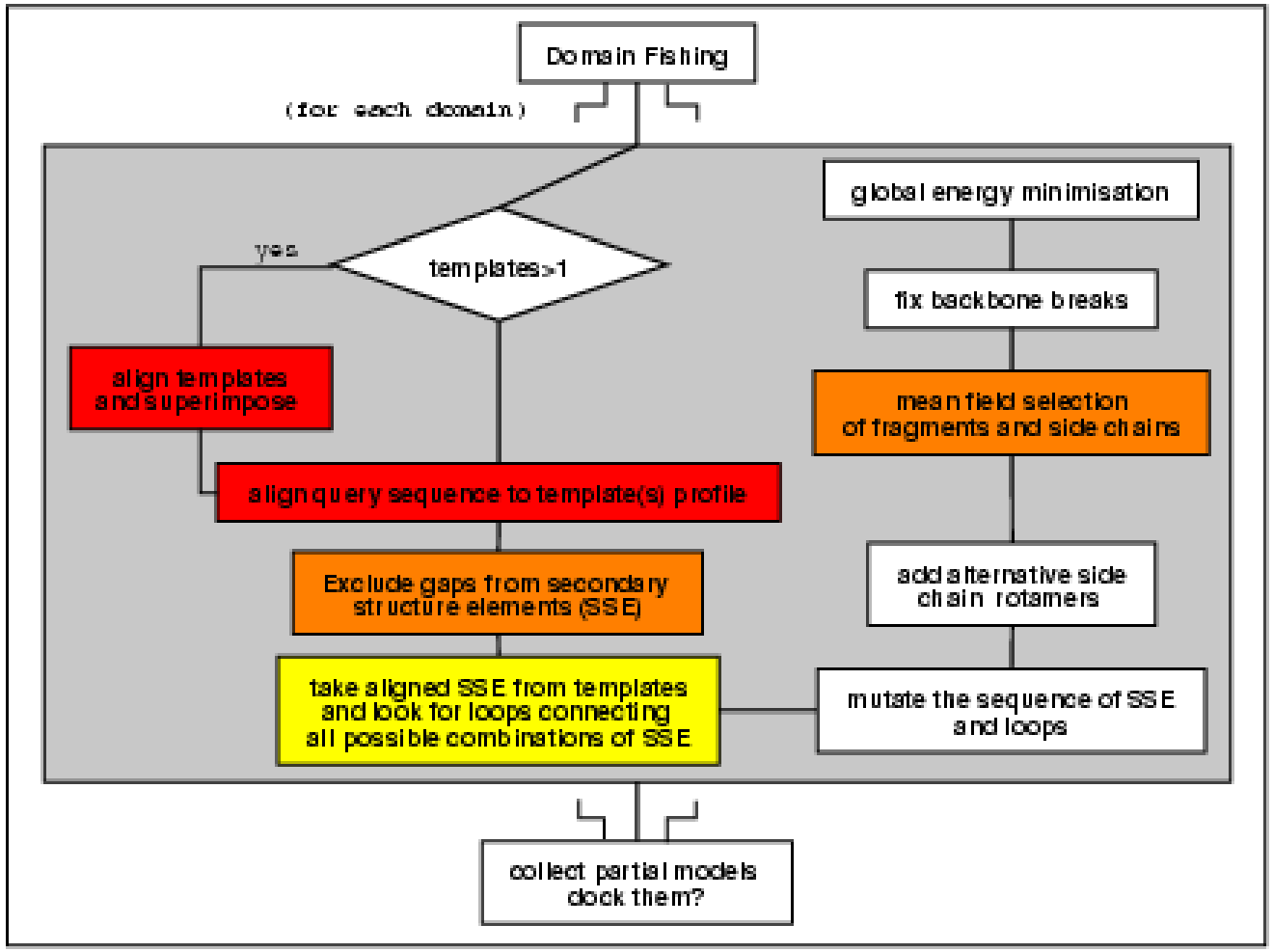
43/97 M

Domain Fishing



3D-JIGSAW

Example



Conclusion

We have done:

- automatic domain identification
- improved alignments
- multidomain modelling

We want to do next:

- better template selection (energies)
- connecting domains
- different multi-template strategies