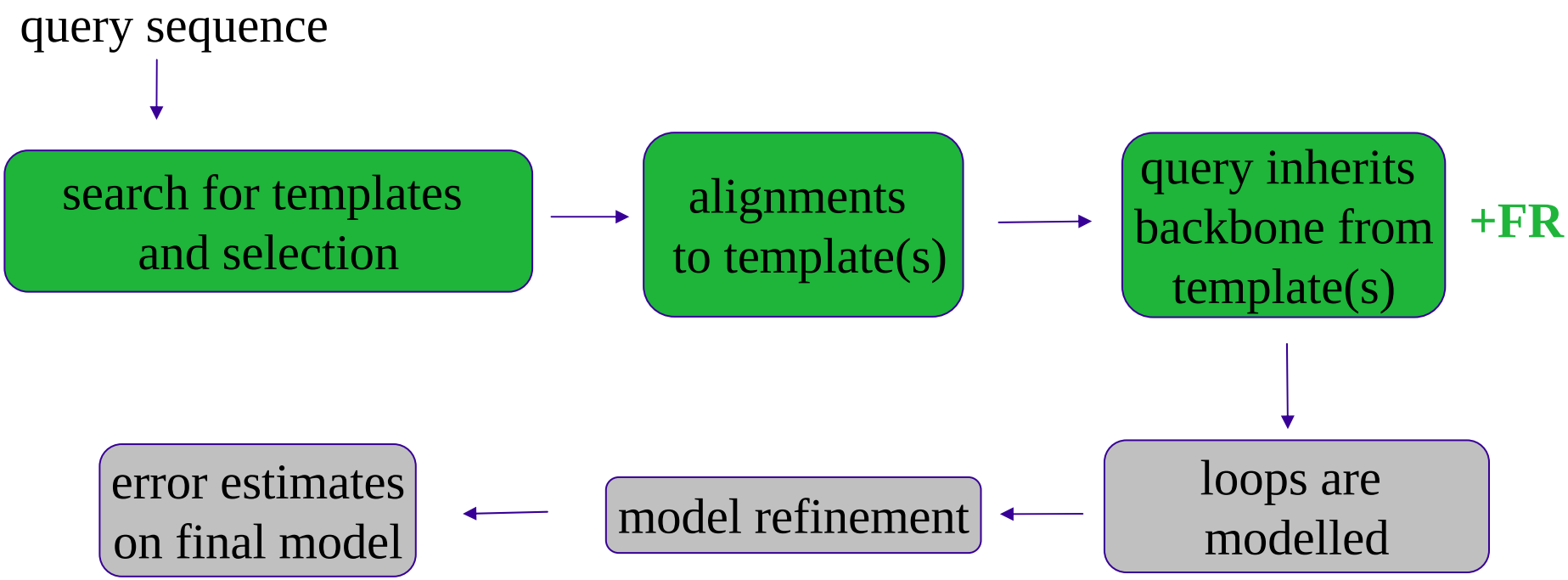
in silico protein recombination: a genetic algorithm applied to template and alignment selection in comparative modelling

Bruno Contreras-Moreira
Biomolecular Modelling Laboratory
London Research Institute
September 2003

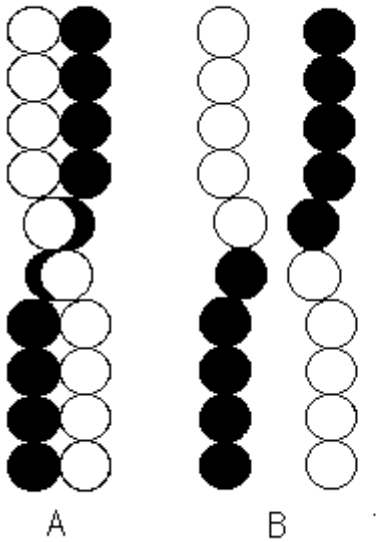
CANCER RESEARCH UK

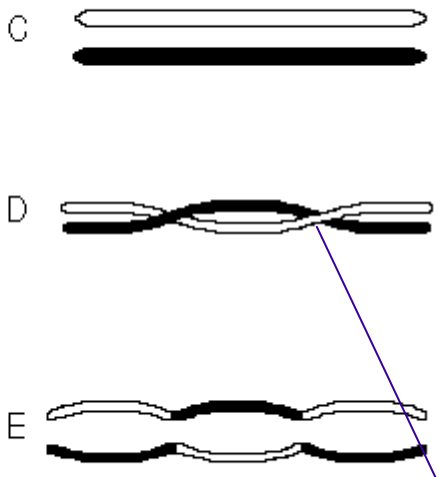
comparative modelling

Predictive technique to build a molecular model for a sequence based on homologous proteins whose structure is known.



chromosome evolution & computational analogy: genetic algorithms





encoded solution

1111000111111111001

0010010000100111001

1111000000100111001

0010010111111111001

crossover point

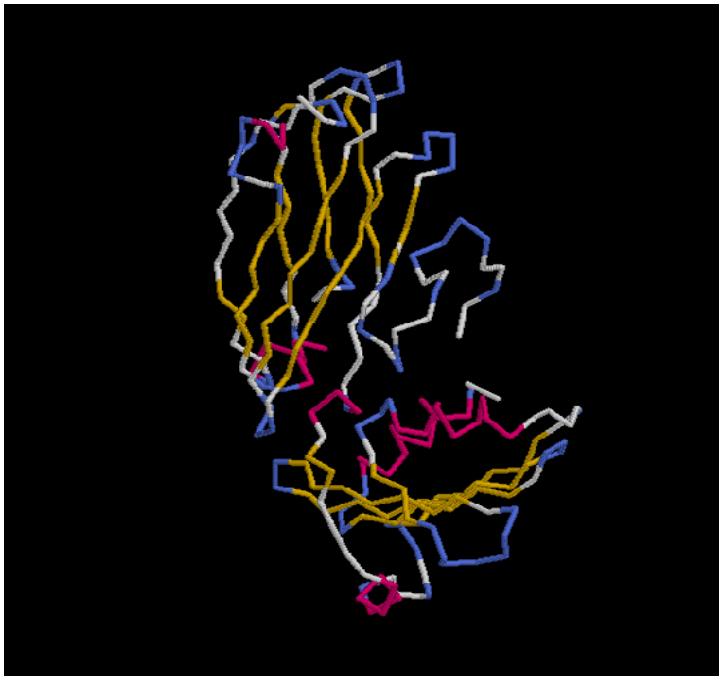
a genetic algorithm applied to Comparative Modelling

- 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

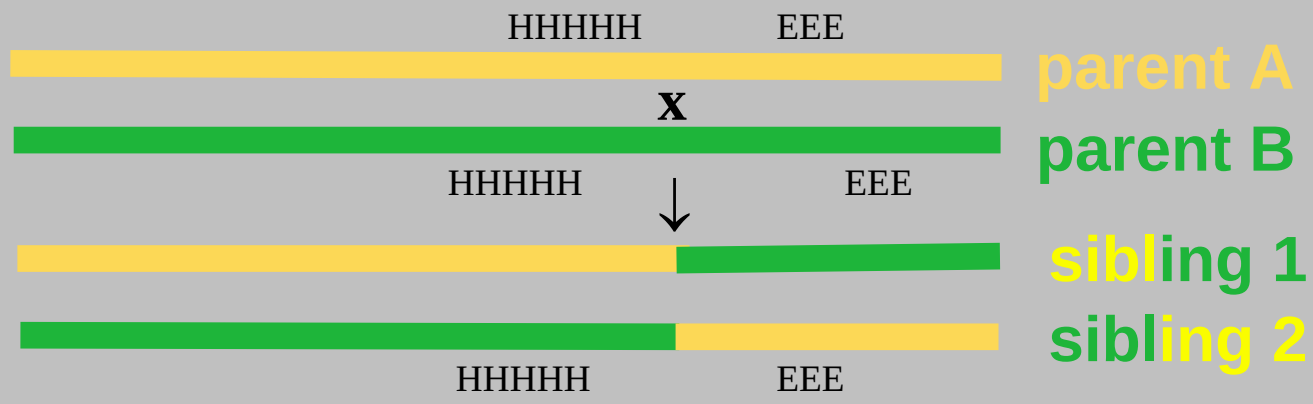
```
T0134   GEP-VQNGAPEEE--QLPPESSYSLLAENSYVKMTCDIRGSLQEDSQVTVAIVLENRSS
1qts_A GSPGIRLGSSEDNFARFVCKNNGVLF-ENQLLQI--GLKSEFRQNLG-RMFIFYGNKTS
SS      CCCCCCCCCCCCCCHHHHCCCCCEEEE-ECCCEEE--EEEEEEECCEE-EEEEEEEECCC
```



(1model = 1PDB template + 1alignment)

recombination

```
model recombination( model A , model B)
{
    do sequence_alignment( A , B );
    do sequence_superimposition_Cβ( A , B );
    do refine_superimposition_close_Cβ( A , B );
    do draw_crossover_point( A , B ); /* out of SS? */
    return create_model(A , B , crosspoint );
}
```



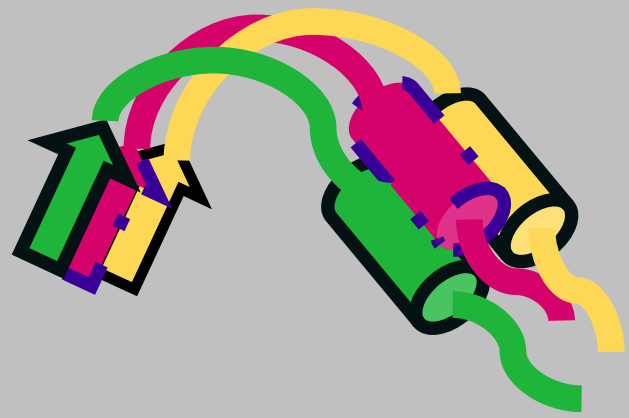
mutation

```
model mutation( model A , model B)
{
    do sequence_alignment( A , B );
    do sequence_superimposition_Cβ( A , B );
    return create_Cartesian_average_model(A , B);
    /* quality checks, minimization? */
}
```

parent A

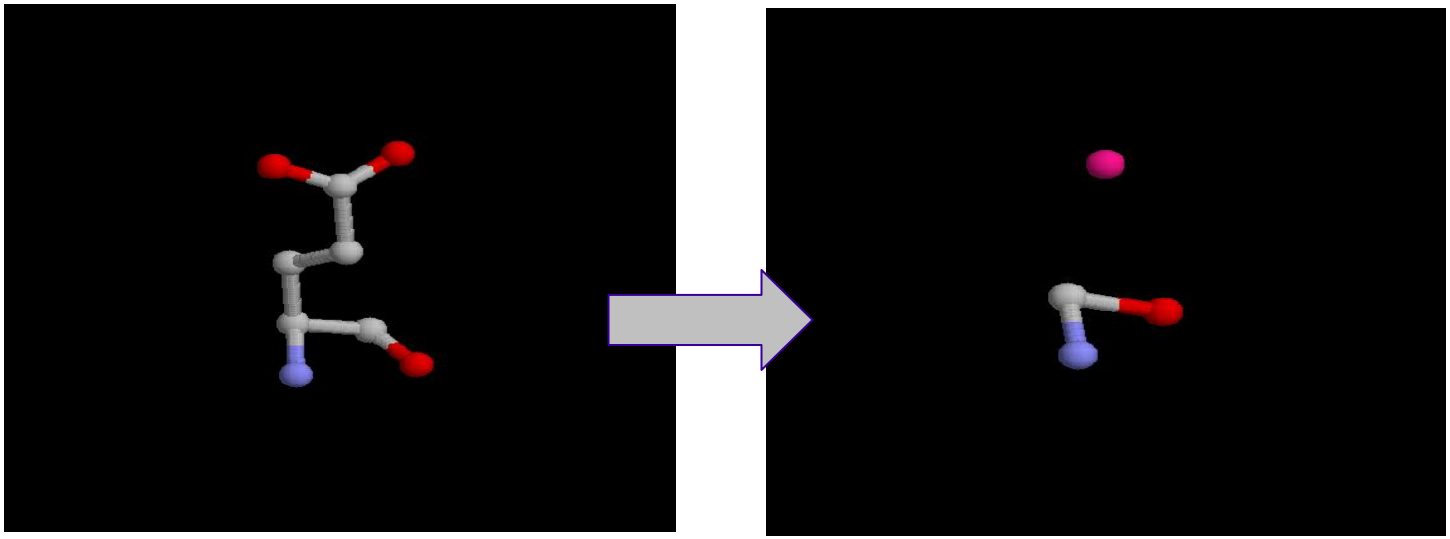
parent B

sibling



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)

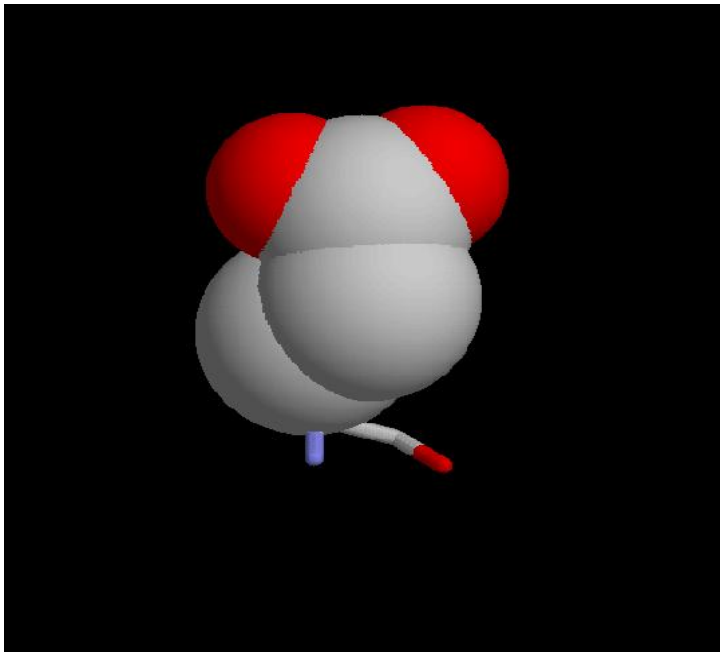
where i,j are pairs of pseudoatoms in protein p

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)**



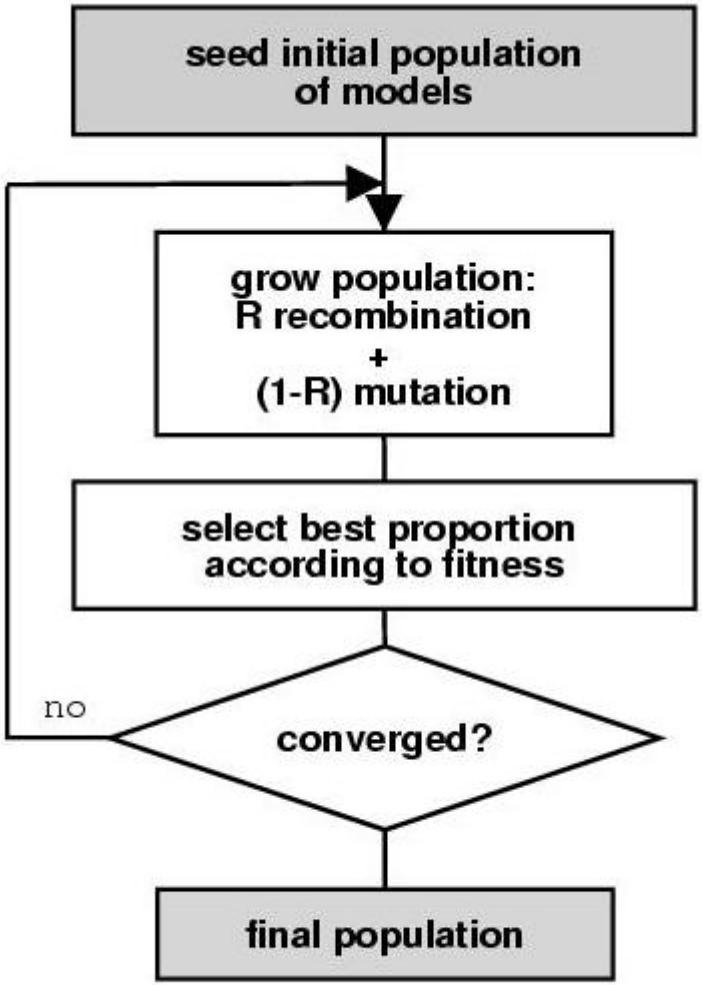
$\sum_i (SA_i \,.\, \Delta Gsolv_i)$ (in Kcal/mol)

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

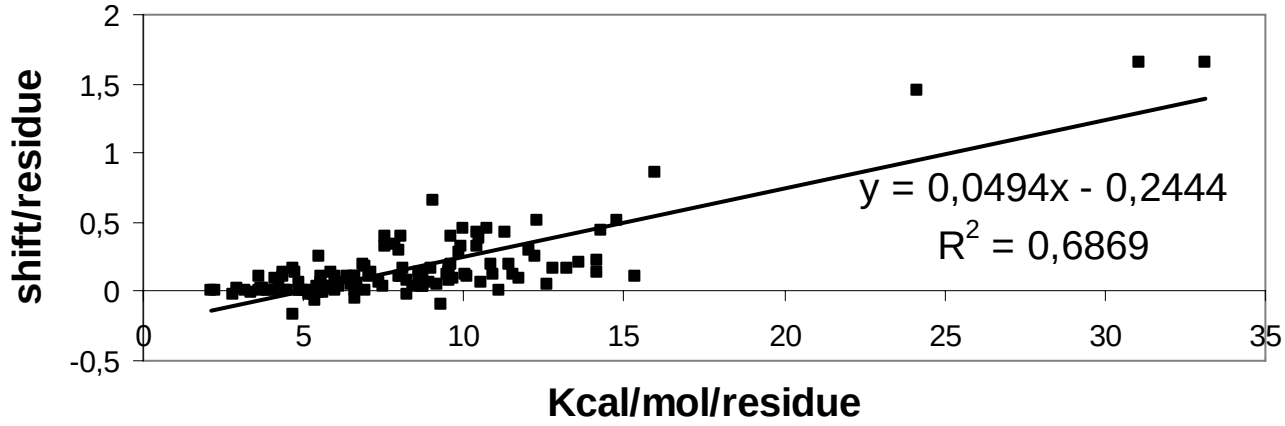


Contreras-Moreira, Fitzjohn and Bates (2003) *J Mol Biol*, **328**: 593-608.

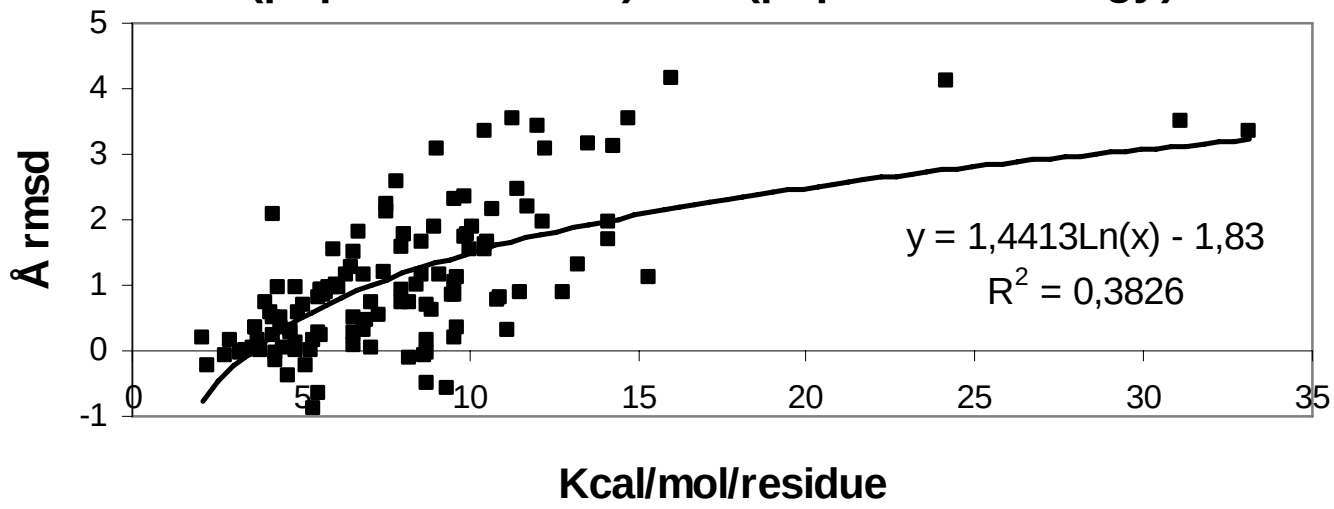
in silico protein recombination: performance

(benchmark on 130 SCOP families)

d(population energy) *vs* d(alignment shift)



d(population rmsd) *vs* d(population energy)

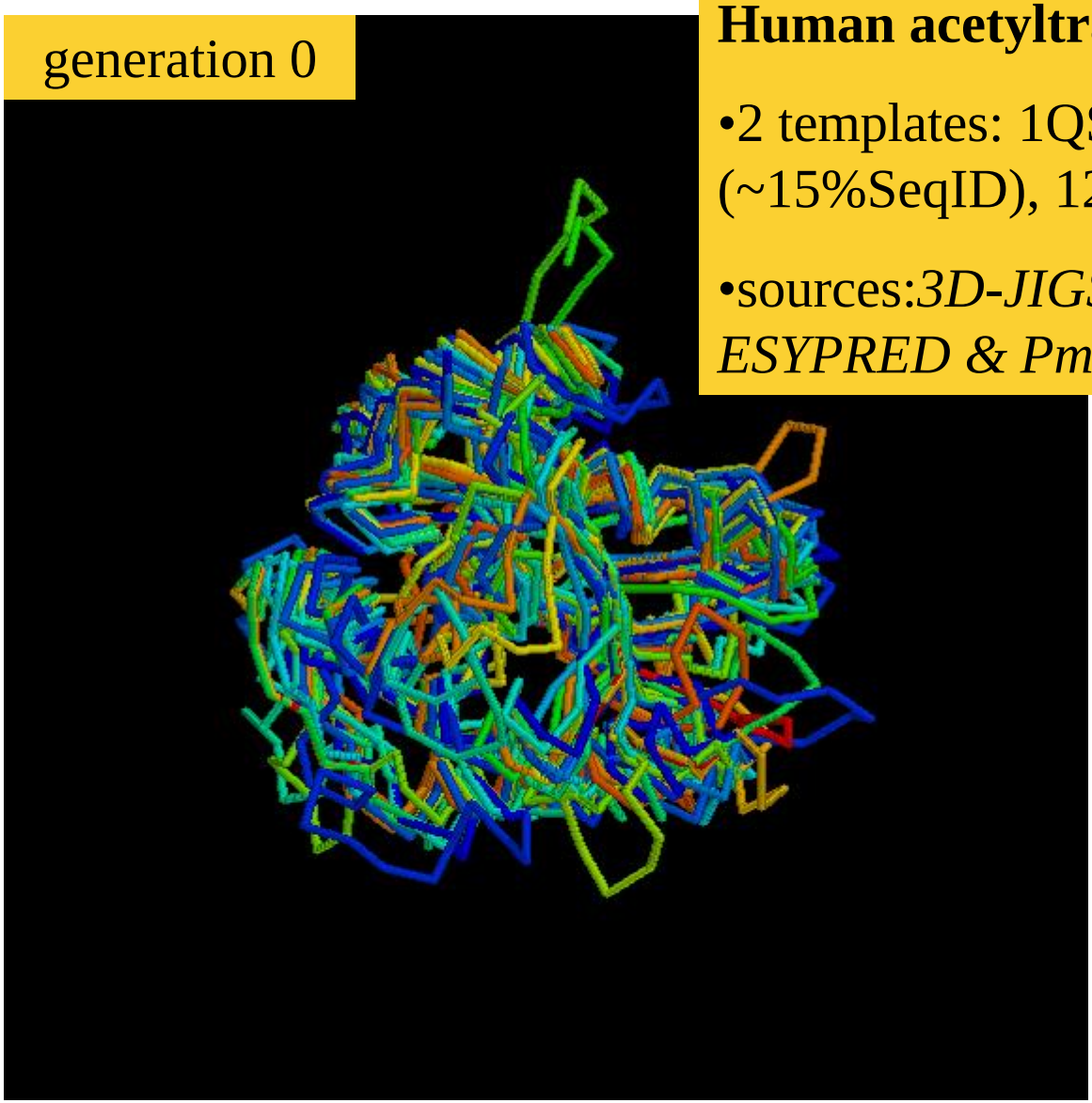


CASP5 example: T0192

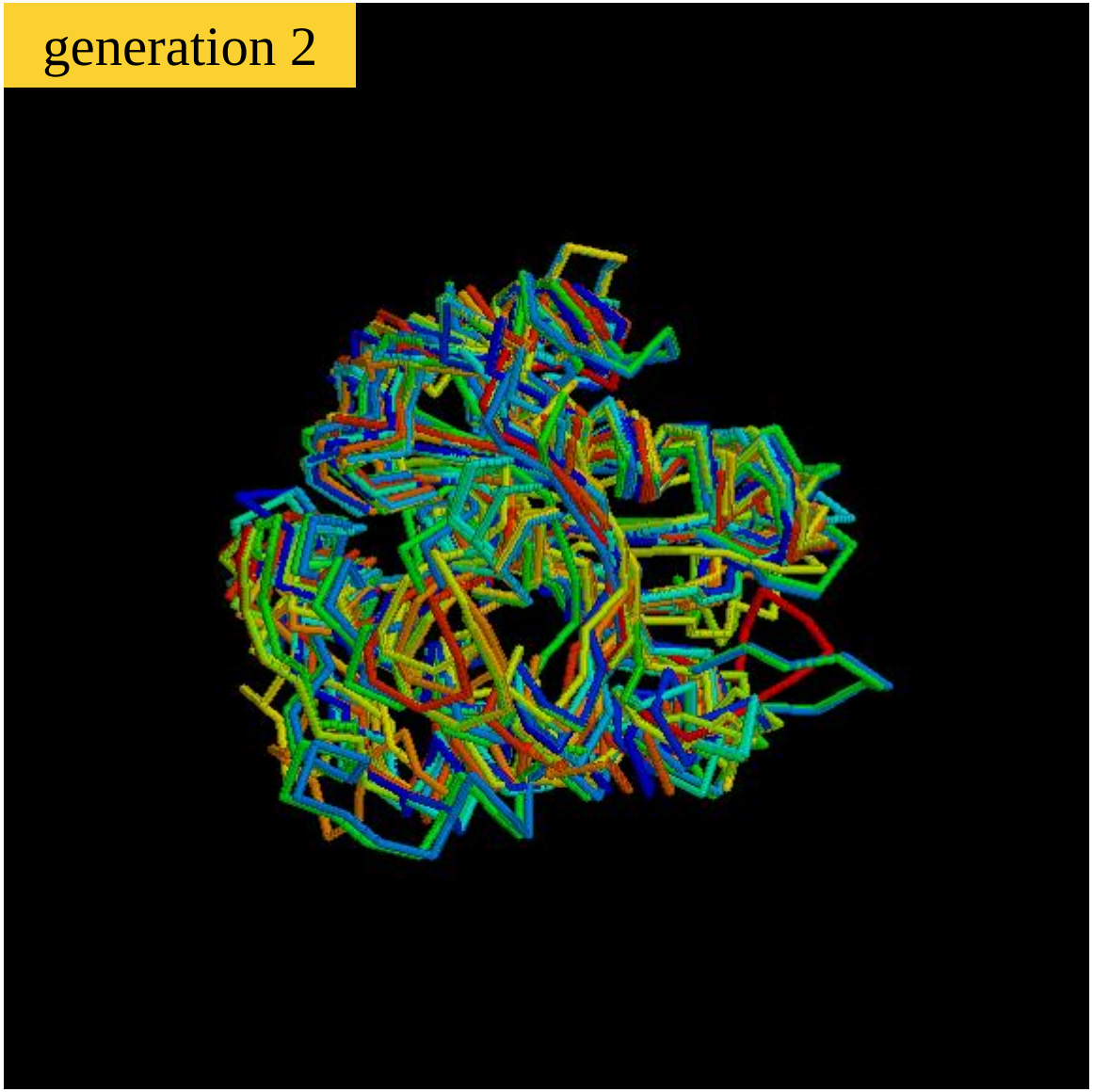
Human acetyltransferase

•2 templates: 1QSM & 1QSO (~15%SeqID), 12 alignments

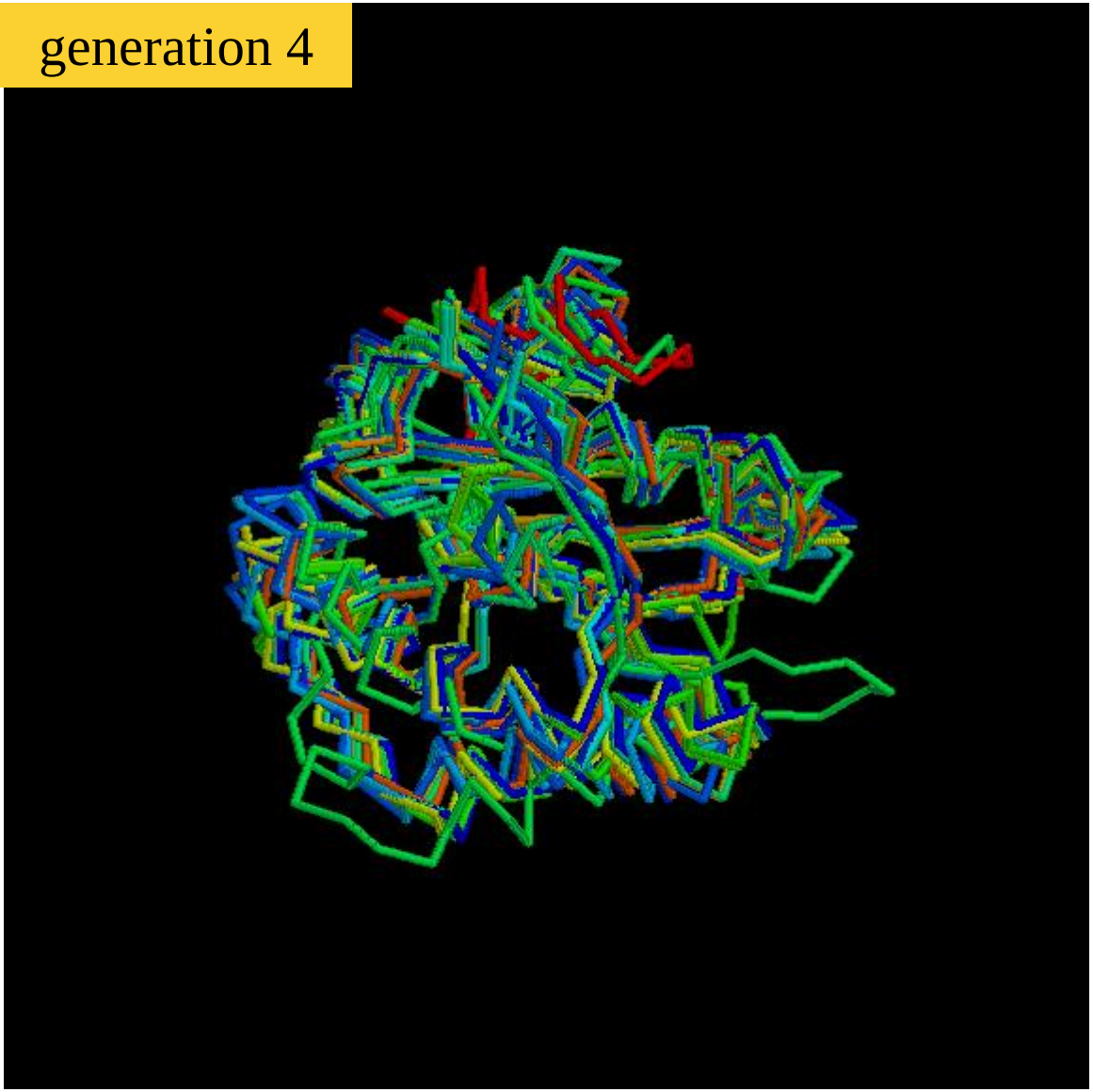
•sources:*3D-JIGSAW*,*FAMS*, *ESYPRED & Pmodeller*



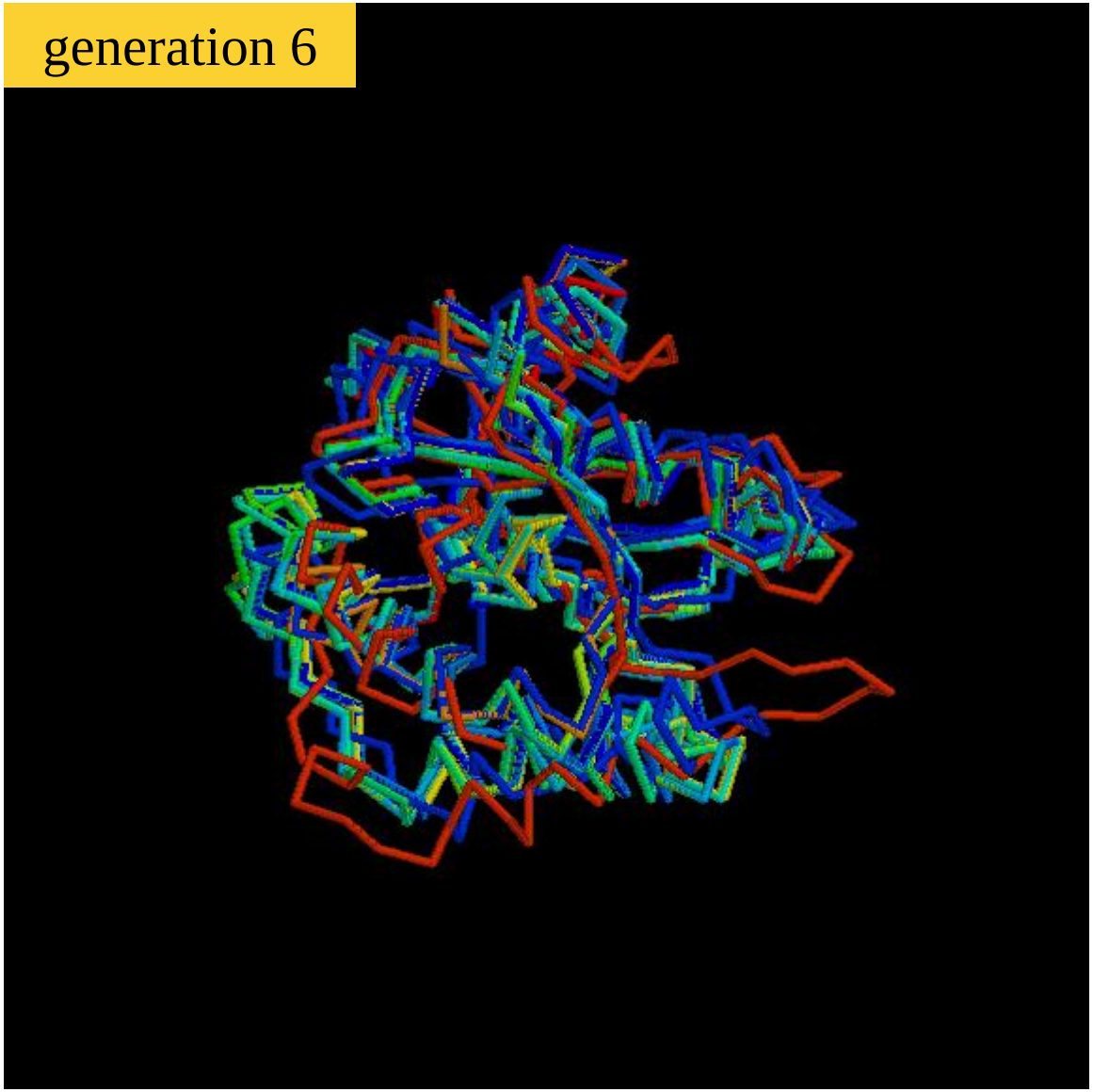
generation 2



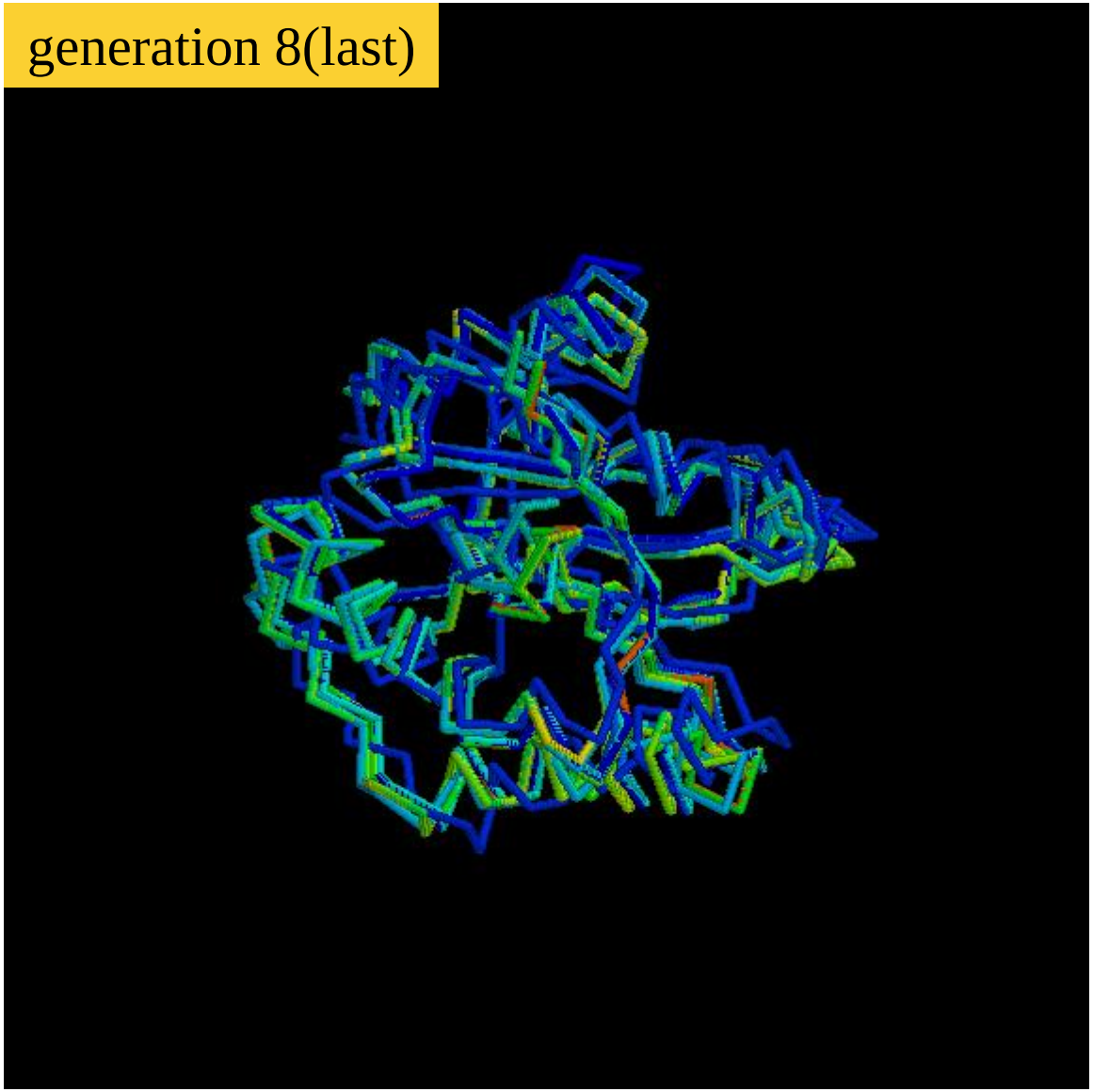
generation 4

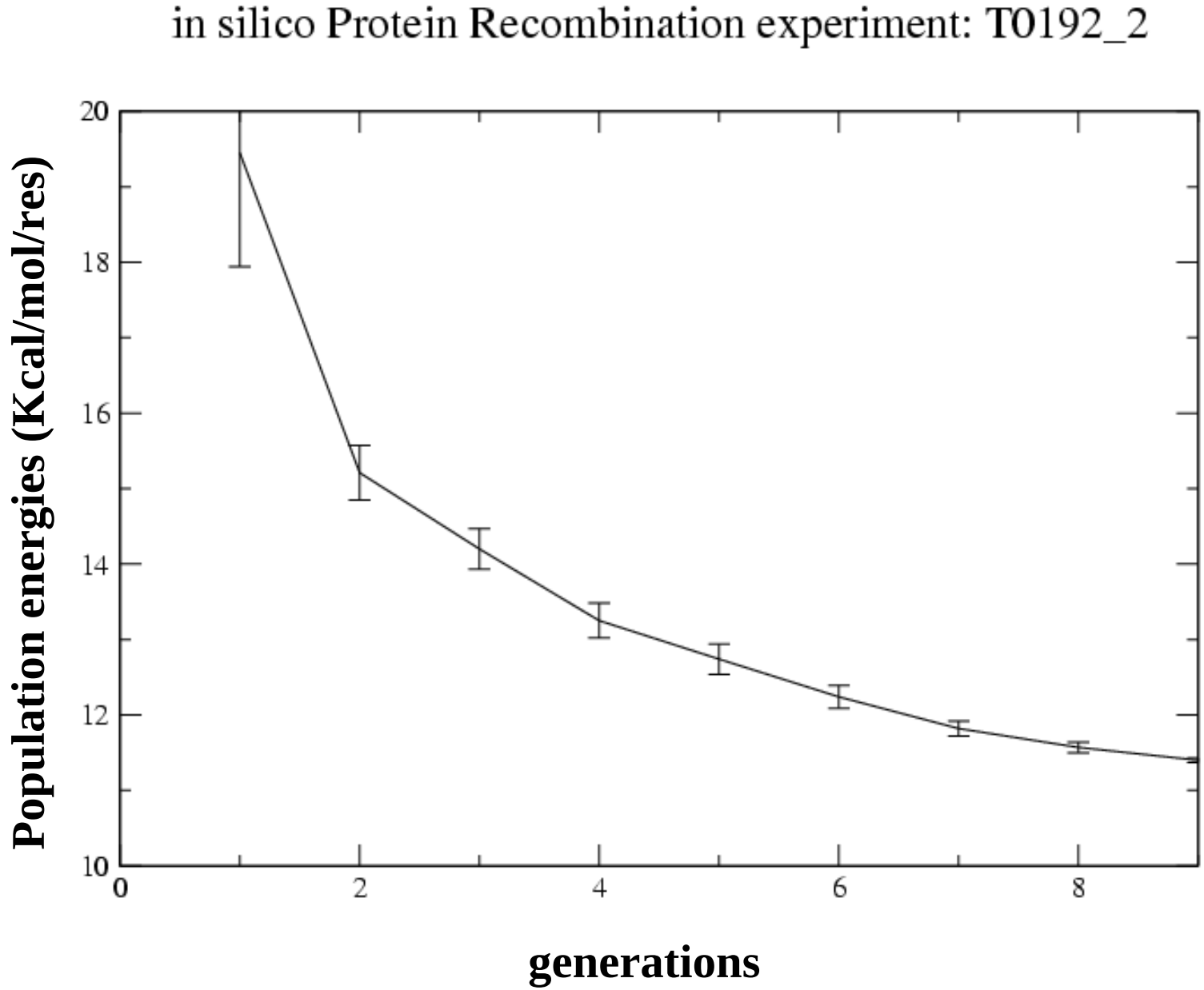


generation 6



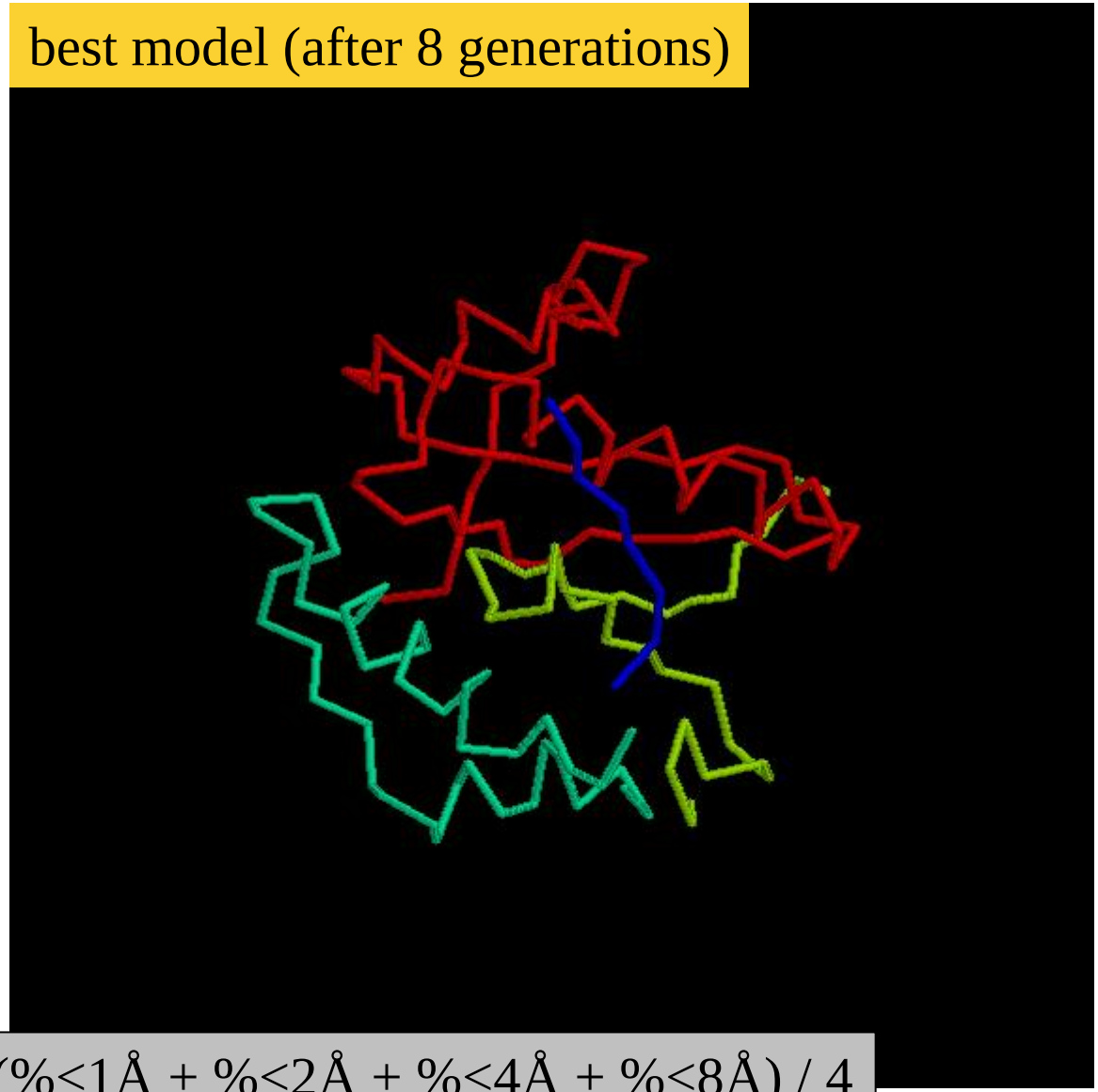
generation 8(last)




in silico Protein Recombination experiment: T0192_2
Population energies (Kcal/mol/res)
20
18
16
14
12
10
0
2
4
6
8
generations

model	GDT_TS	AL_4
mod1	45	61
mod2	63	81
mod3	57	72
mod4	54	64
mod5	54	64
mod6	61	80
mod7	61	76
mod8	61	80
mod9	62	78
mod10	65	77
mod11	62	78
mod12	60	71
average	*58*	*74*
rec_8gen	61	81
bestCASP5	*66*	*85*

best model (after 8 generations)



GDT = (%<1Å + %<2Å + %<4Å + %<8Å) / 4

AL_4 = %(<4Å AND shift±4)

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

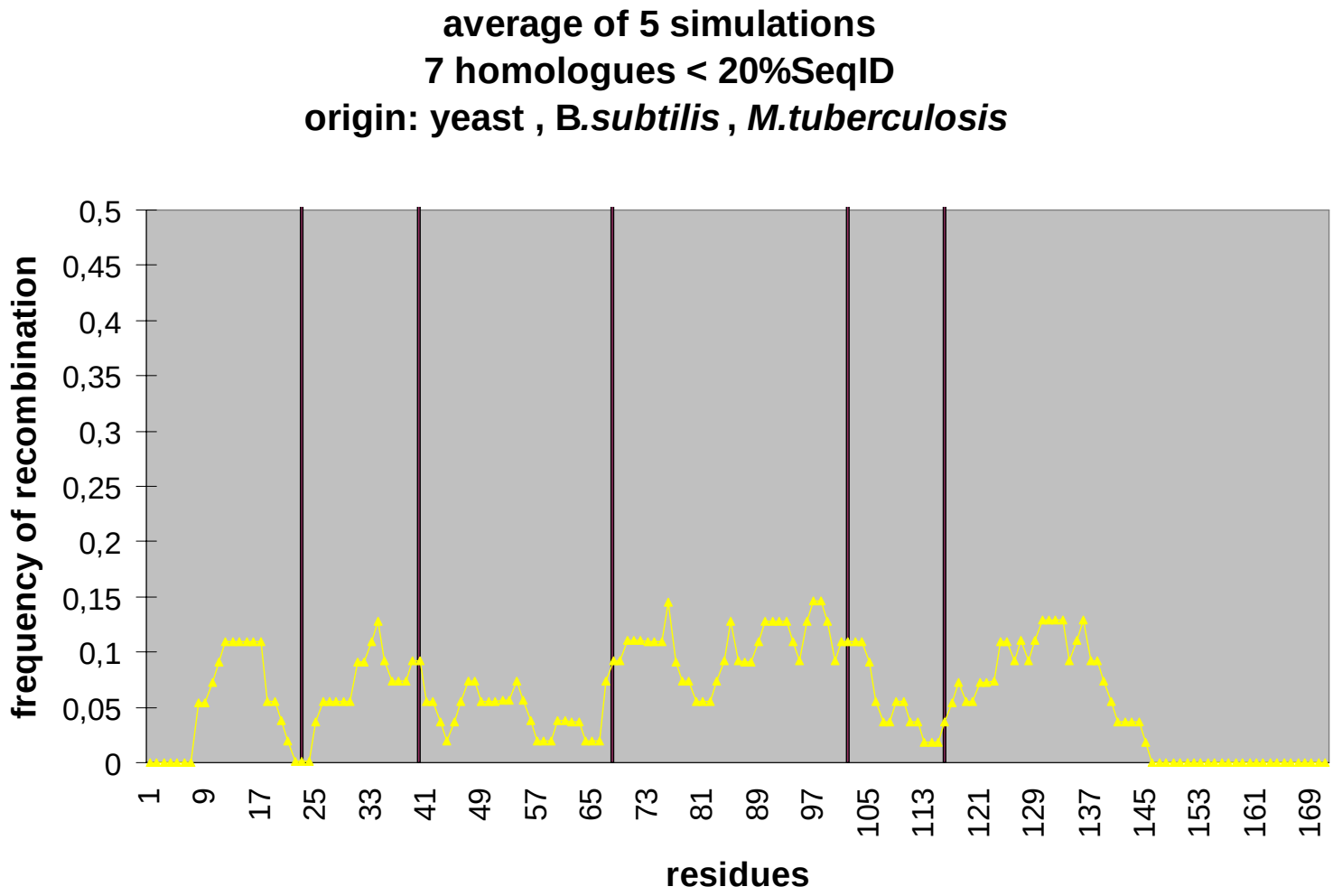
•is able to recover alignment errors

•often last population contains different conformations

EVOLUTIONARY INSIGHT

• high-frequency crossover spots seem to occur away from intron boundaries (work with Pall Jonsson)

Crossover points and introns boundaries: T0192



Biomolecular Modelling Laboratory

Paul Bates

Paul Fitzjohn

José Jimenez

Pall Jonsson

Chris Page

Graham Smith

&

Marc Offman

Fabien Birzele

(students from LMU München)

www.bmm.icnet.uk

etscape: 'in silico' Protein Recombination
Edit View Go Communicator

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:

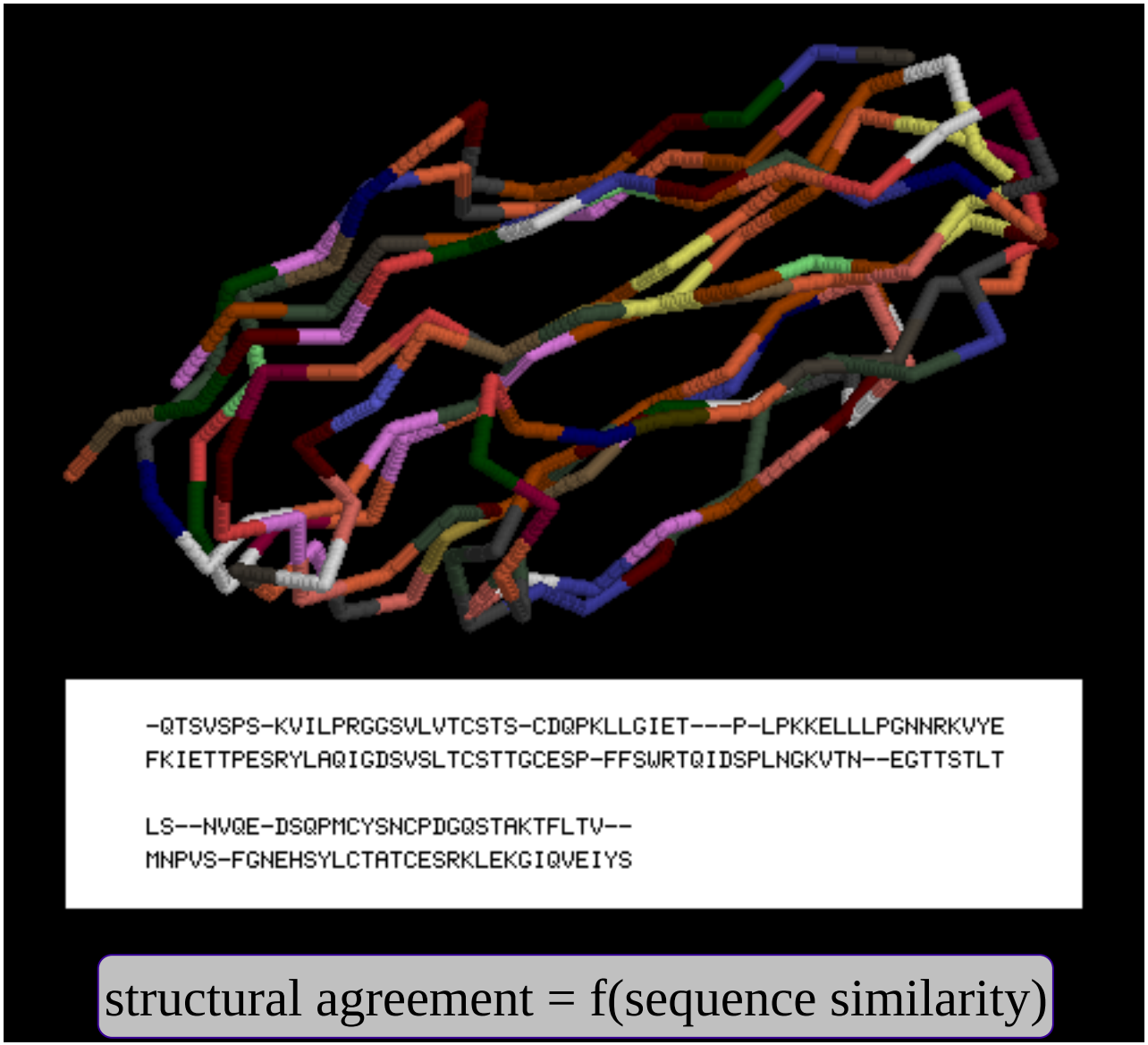
pop size 30 selected prop 0.75 forced improvement 0.001 last gen 50 mutation rate 20

Upload and Recombine! Reset help

contrera@cancer.org.uk BMM disclaimer

100% http://www.bmm.icnet.uk/servers/3djigsaw/recomb/html/help.html

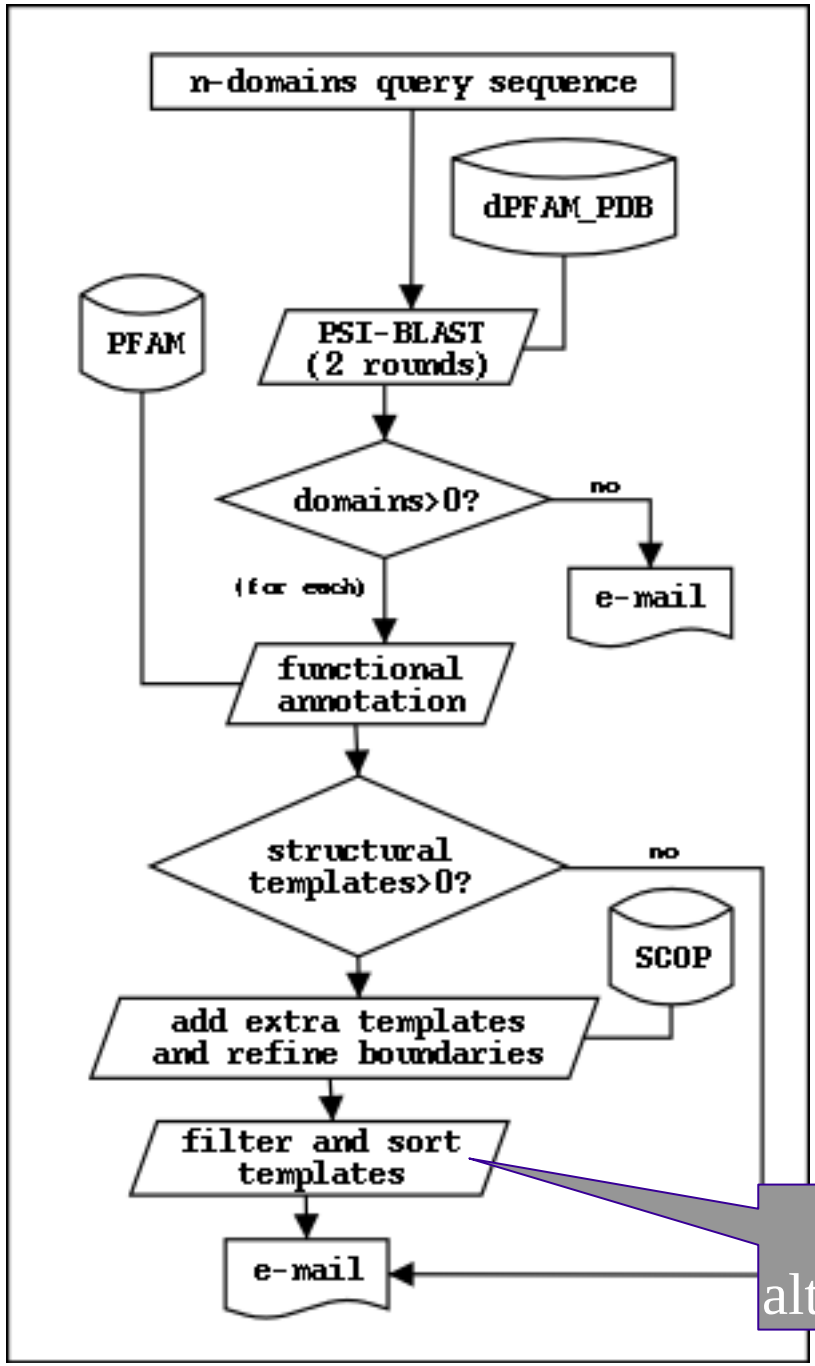
the basic principle



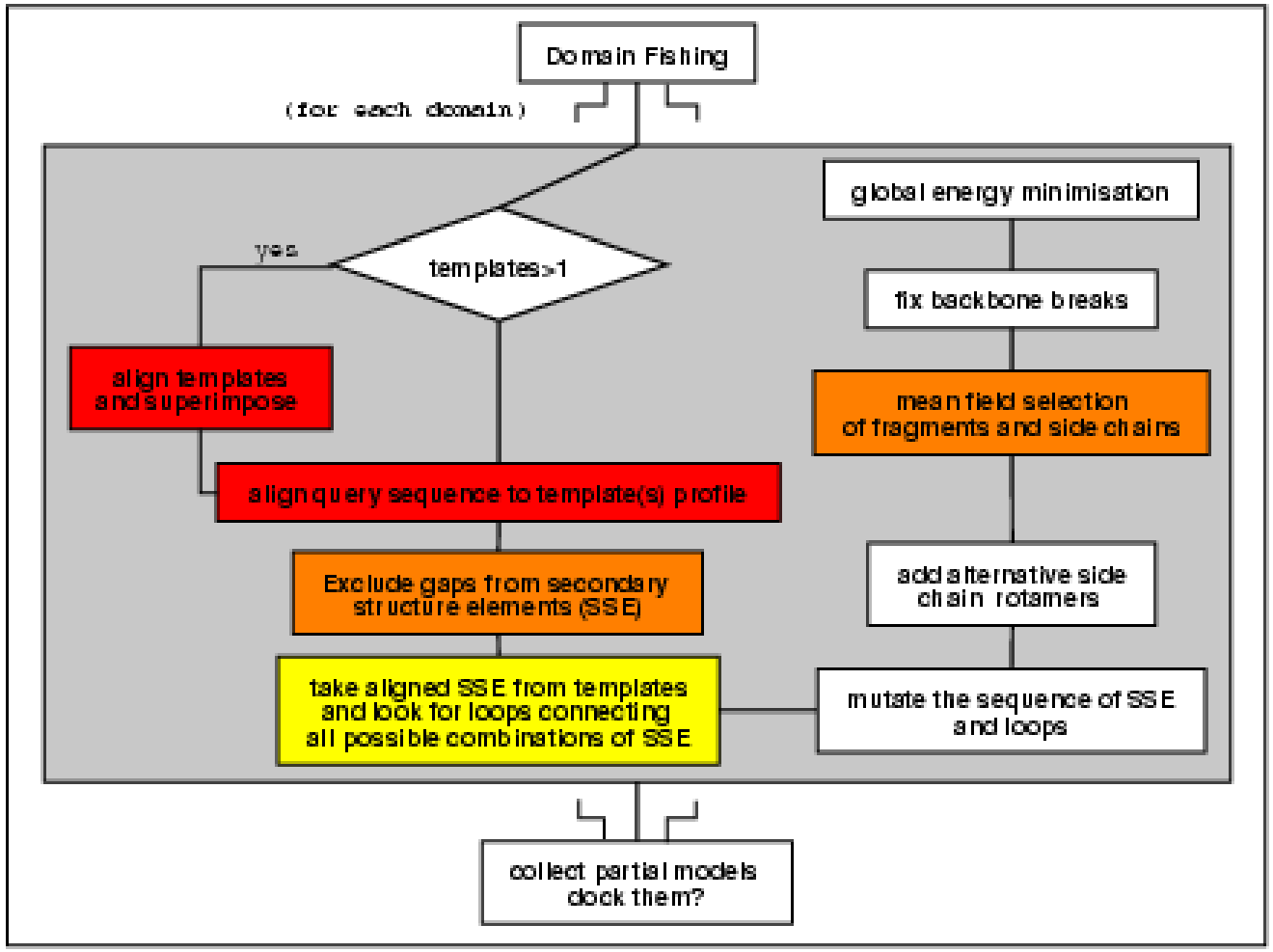
```
-QTSVSPS-KVILPRGGSVLVTCSTS-CDQPKLLGIET---P-LPKKELLLPGNNRKVYE
FKIETTPESRYLAQIGDSVSLTCSTTGCESP-FFSWRTQIDSPLNGKVTN--EGTTSTLT

LS--NVQE-DSQPMCYSNCPDGQSTAKTFLTV--
MNPVS-FGNEHSYLCTATCESRKLEKGIQVEIYS
```

structural agreement = f(sequence similarity)

Domain Fishing



3D-JIGSAW


Domain Fishing
(for each domain)
templates>1
yes
align templates and superimpose
align query sequence to template(s) profile
Exclude gaps from secondary structure elements (SSE)
take aligned SSE from templates and look for loops connecting all possible combinations of SSE
collect partial models dock them?
global energy minimisation
fix backbone breaks
mean field selection of fragments and side chains
add alternative side chain rotamers
mutate the sequence of SSE and loops

Netscape: Your results

File Edit View Go Communicator

Bookmarks Location: http://www.bmm.icnet.uk/~3djigsaw/dom_fish/display2.cgi?output/J79b3137 What's Related

Back Forward Reload Home Search Netscape Print Security Shop Stop

Interactive 3D-JIGSAW

Possible structural templates in PDB

name	from	to
1bza_# Model!?	28	287
1shv_A Model!?	26	292
1g56_A Model!?	26	292
1ck3_A Model!?	26	290
1jtd_A Model!?	27	288
1fqg_A Model!?	26	288
1btl_# Model!?	26	290
1bt5_A Model!?	26	290
1erq_A Model!?	26	288

truncated alignments? wrong templates? PDB code 1bza chain # first residue - last - align the query to this PDB!