Algorithms for protein comparative modelling and some evolutionary implications

Bruno Contreras-Moreira

Biomolecular Modelling Laboratory

London Research Institute

February 2004

CANCER RESEARCH UK

overview

1. Acknowledgements

2. Introduction: what is protein comparative modelling (5 slides)

3. Comparison of alignment techniques: defining domains and selecting templates (7 slides)

4. Recombination of protein models: in-house and CASP5 benchmarks (17 slides)

5. A relation between exonic structure of genes and protein structure: recombination of protein domains (5 slides)

6. Conclusions

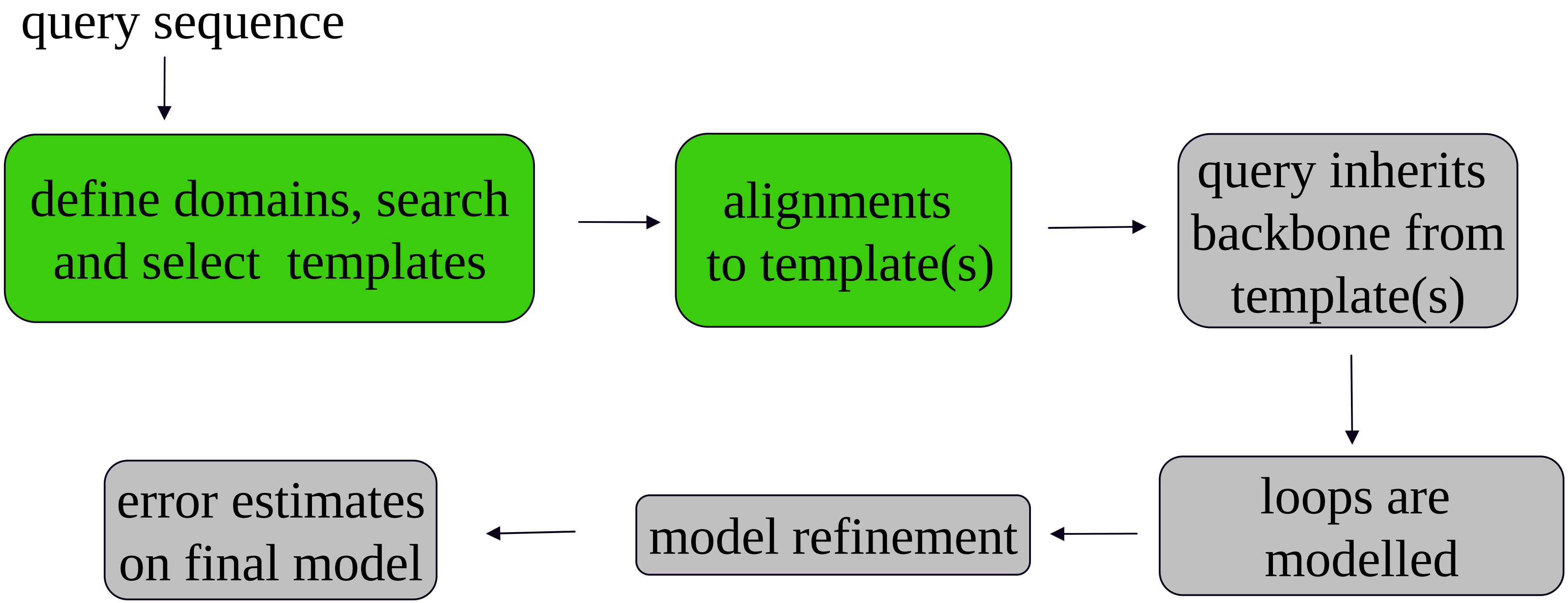
1. Acknowledgements

I would like to thank...

- the Biomolecular Modelling Lab:
  - Paul Bates
  - Paul Fitzjohn
  - Graham Smith
  - Raphäel Chaleil
  - Páll Jónsson
  - Chris Page
  - José Jiménez
- Neil McDonald, John Sgouros, Nancy Hogg, Helen McNeill, Giampietro Schiavo and David Perkins
- María, Joana, Aphrodite, Nikolakis and Óscar
- Cancer Research UK

2. Comparative modelling

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



query sequence

define domains, search and select templates

alignments to template(s)

query inherits backbone from template(s)

loops are modelled

model refinement

error estimates on final model

Template: experimentally determined protein structure stored in the Protein Data Bank.

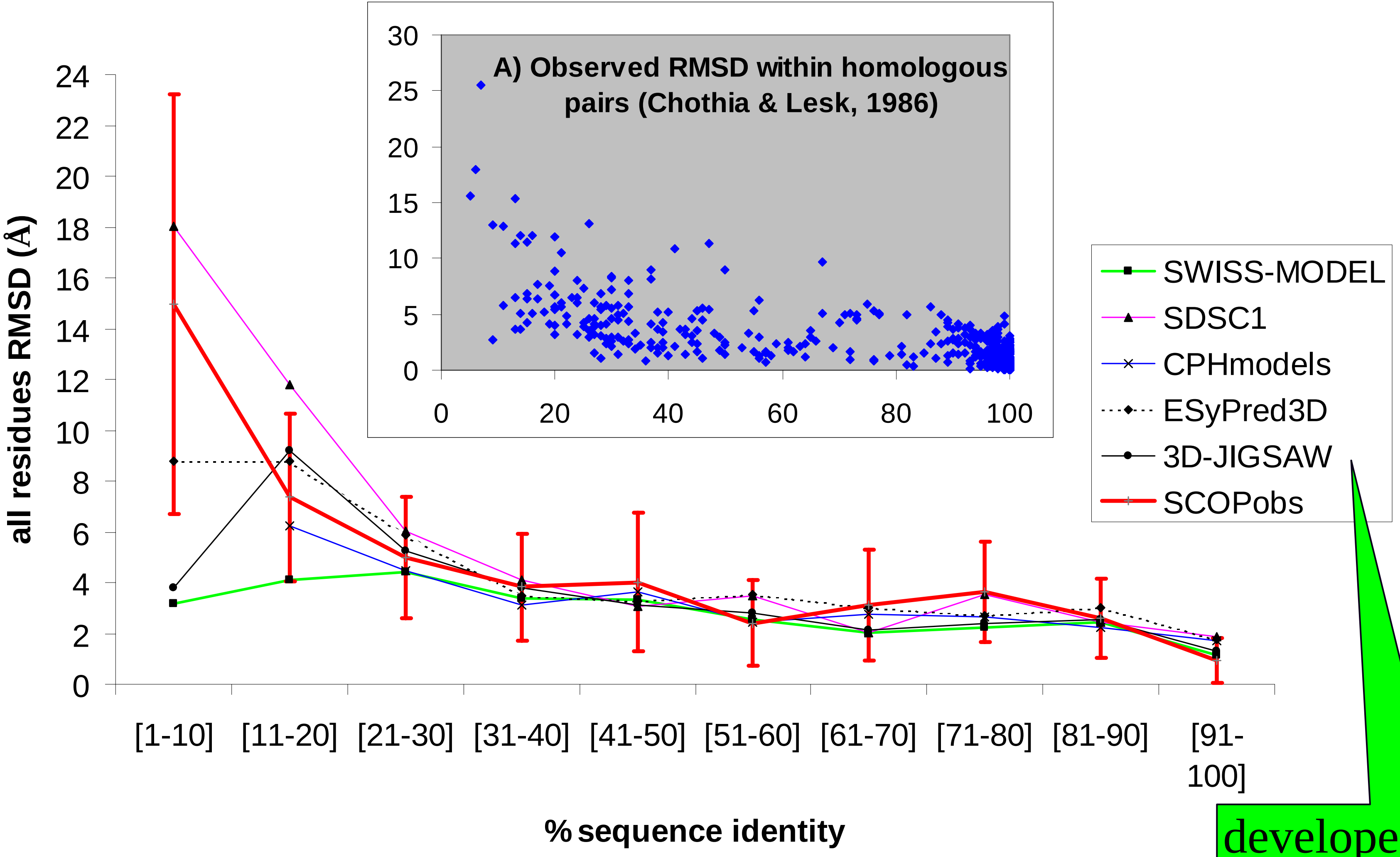
structural significance of sequence alignments

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

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

structural agreement = f(sequence similarity)

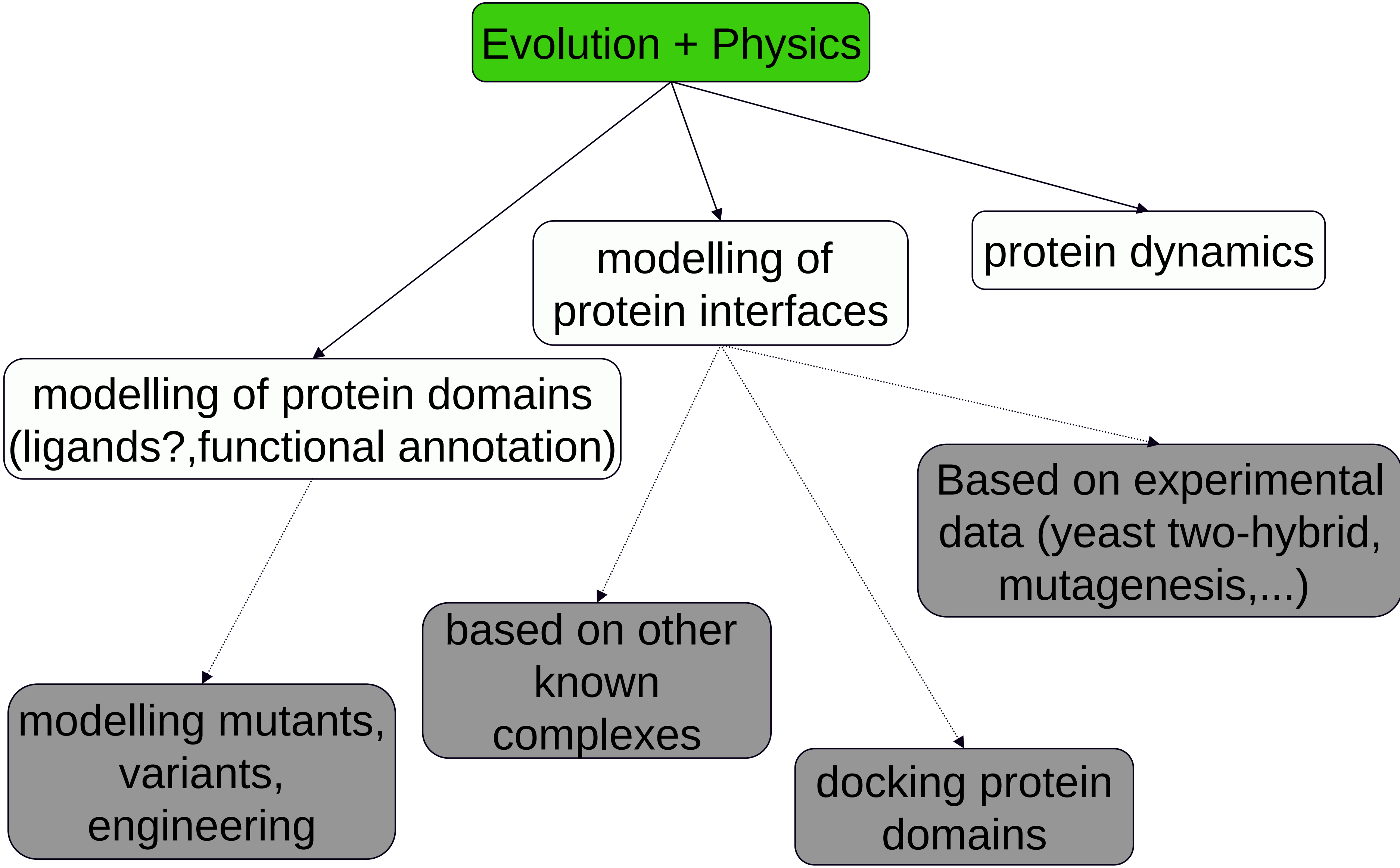
empirical foundations of comparative modelling

B) RMSD between protein models and their experimental structures in the PDB (from EVA project)



developed in our group

applications of protein comparative modelling (1)



Evolution + Physics

modelling of protein domains (ligands?,functional annotation)

modelling of protein interfaces

protein dynamics

modelling mutants, variants, engineering

based on other known complexes

docking protein domains

Based on experimental data (yeast two-hybrid, mutagenesis,...)

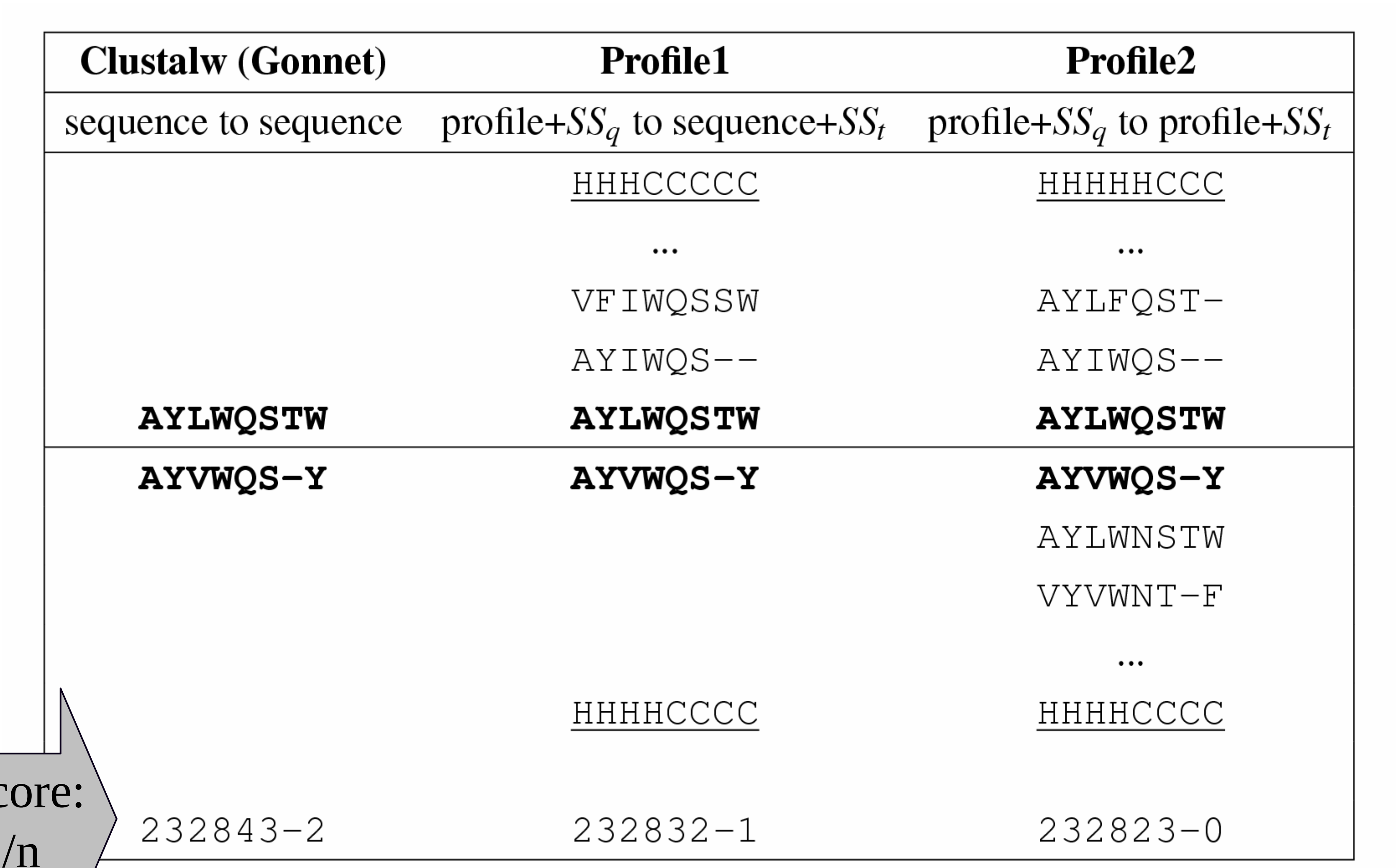
applications of protein comparative modelling (2)

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

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

3. Comparing alignment techniques



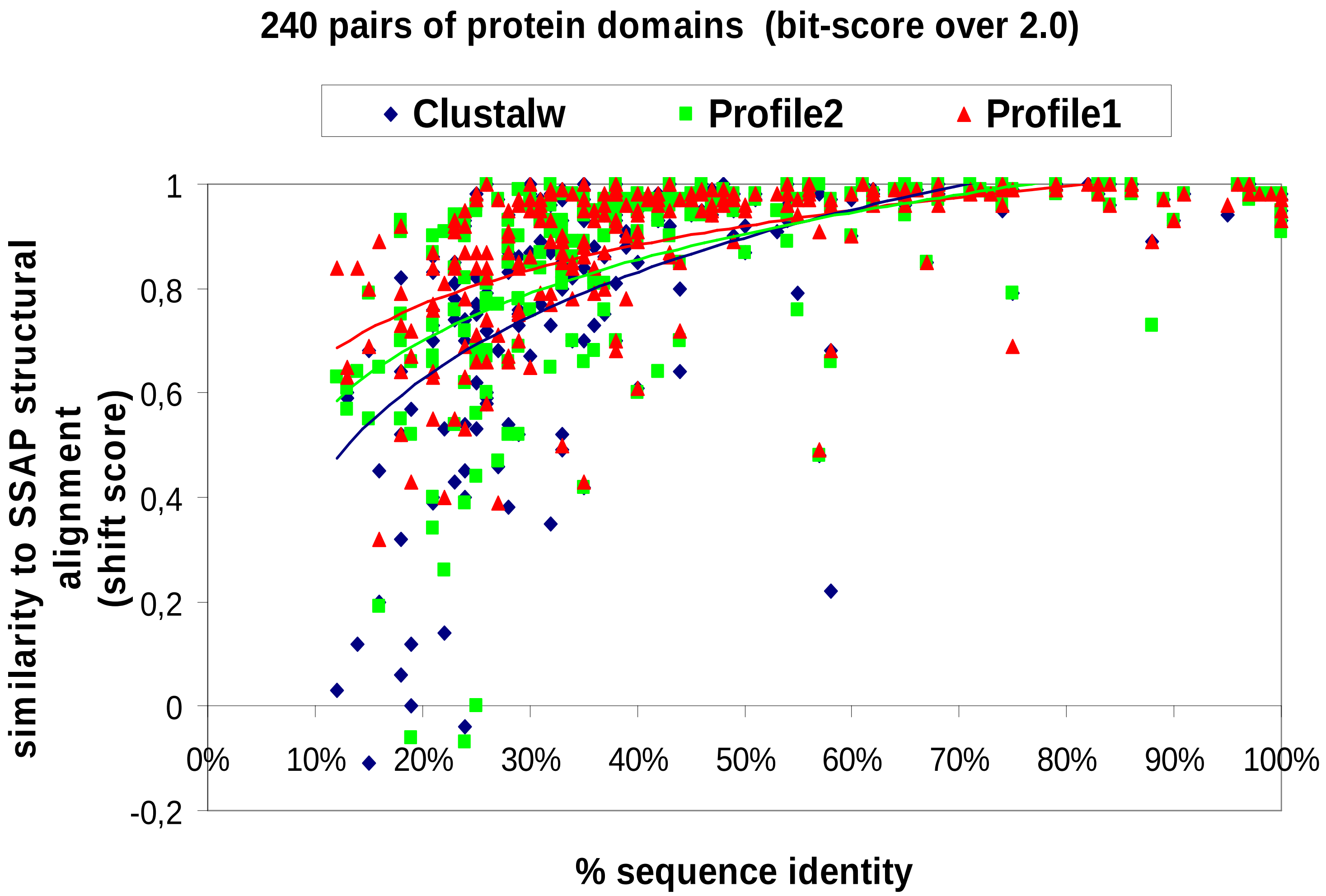
Clustalw (Gonnet)	Profile1	Profile2
sequence to sequence	profile+SS_q to sequence+SS_t	profile+SS_q to profile+SS_t
	HHHCCCCC	HHHHHCCC
	...	...
	VFIWQSSW	AYLFQST-
	AYIWQS--	AYIWQS--
AYLWQSTW	**AYLWQSTW**	**AYLWQSTW**
AYVWQS-Y	**AYVWQS-Y**	**AYVWQS-Y**
		AYLWNSTW
		VYVWNT-F
		...
	HHHHCCCC	HHHHCCCC
232843-2	232832-1	232823-0

bit-score: $\Sigma s_i/n$

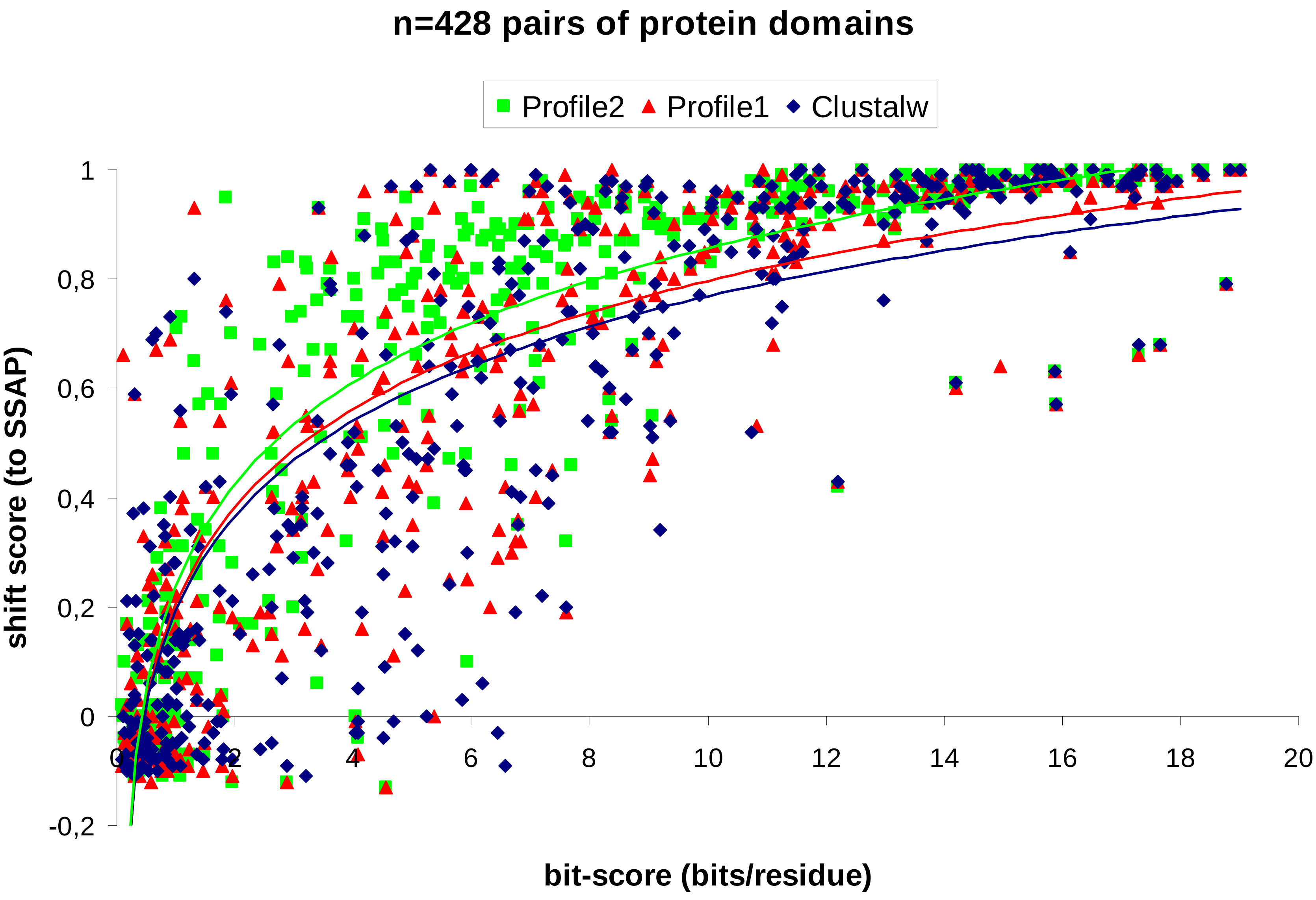
q = query, *t* = template, *SS* = secondary structure

alignment accuracy

A cut-off for the bit-score was found to evaluate alignments: 95% of alignments with shift-score > 0.5 have bit-scores > 2.0

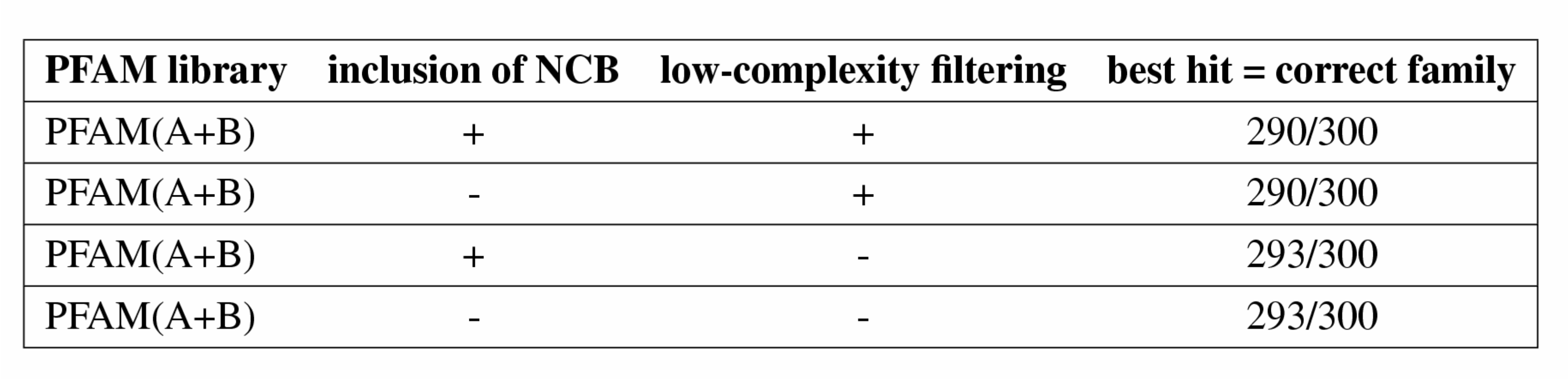


predictive value of bit-scores (R^2 ~ 0.7)



defining protein domains and finding templates

1) query sequence against profile library:
PFAM profiles + IMPALA: 290/300

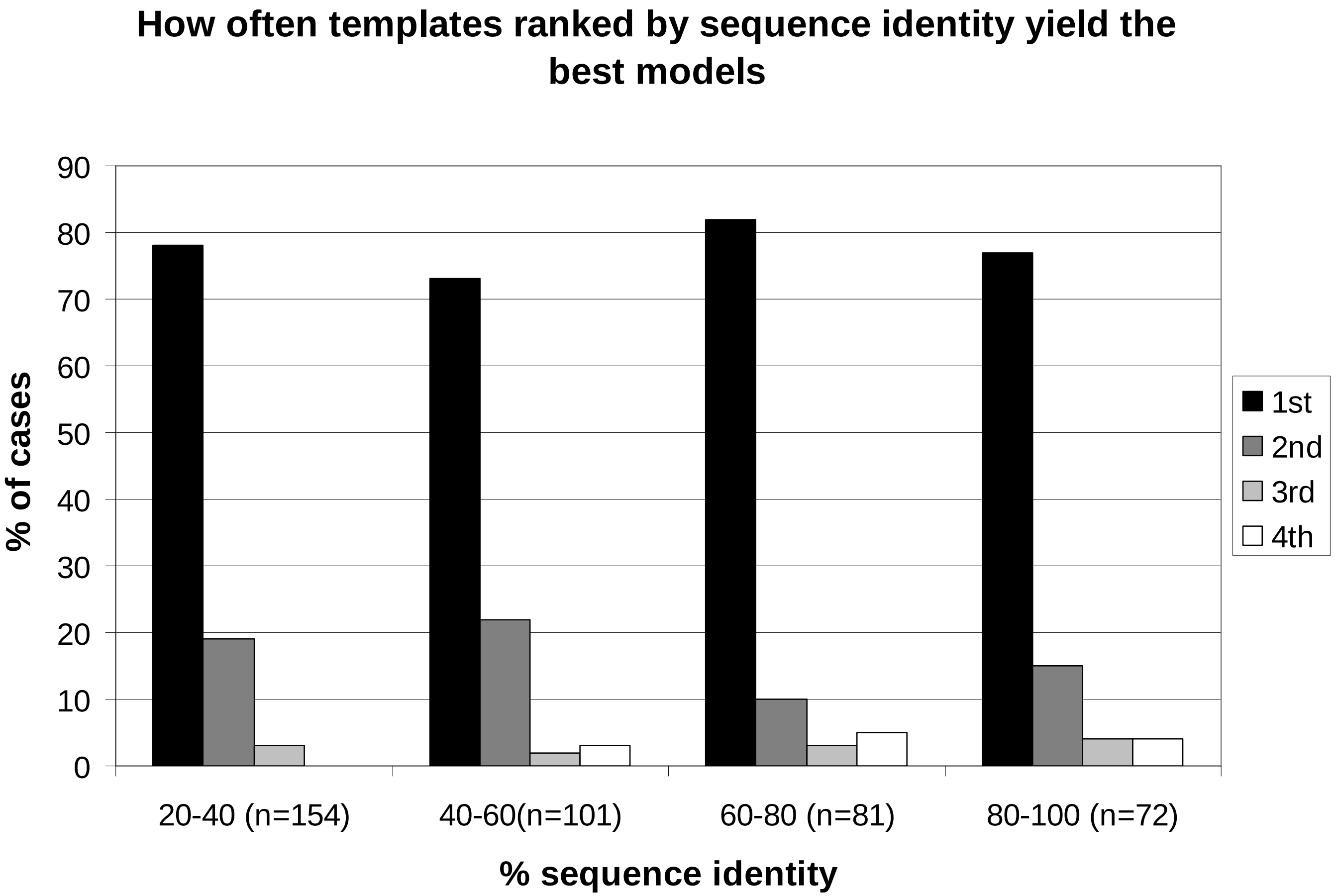


PFAM library	inclusion of NCB	low-complexity filtering	best hit = correct family
PFAM(A+B)	+	+	290/300
PFAM(A+B)	-	+	290/300
PFAM(A+B)	+	-	293/300
PFAM(A+B)	-	-	293/300

2) query sequence against database of sequences:
PFAM + PDB sequences + PSI-Blast: 300/300
plus: domain splitting

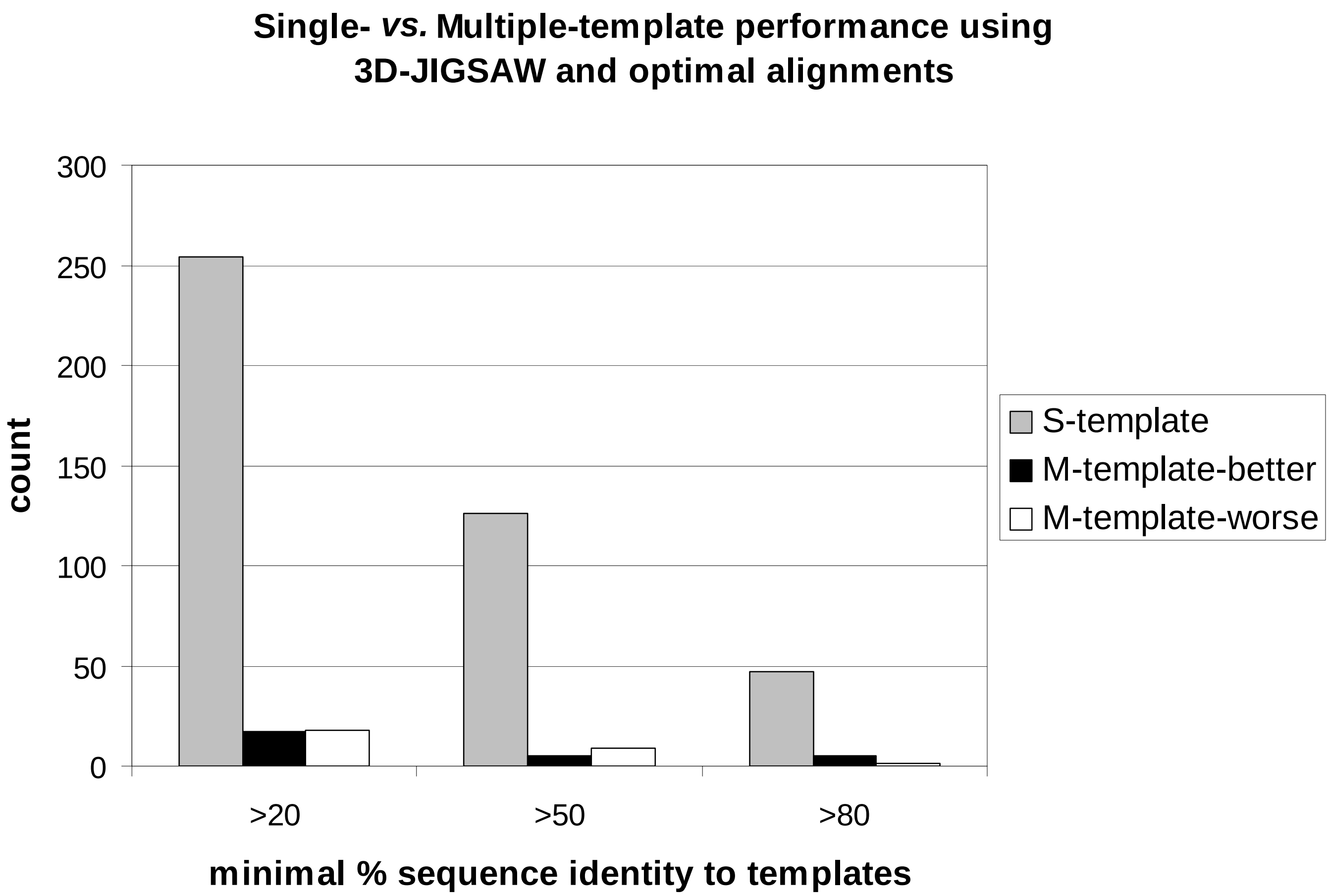
NCB = non-conserved blocks

selecting templates (1)



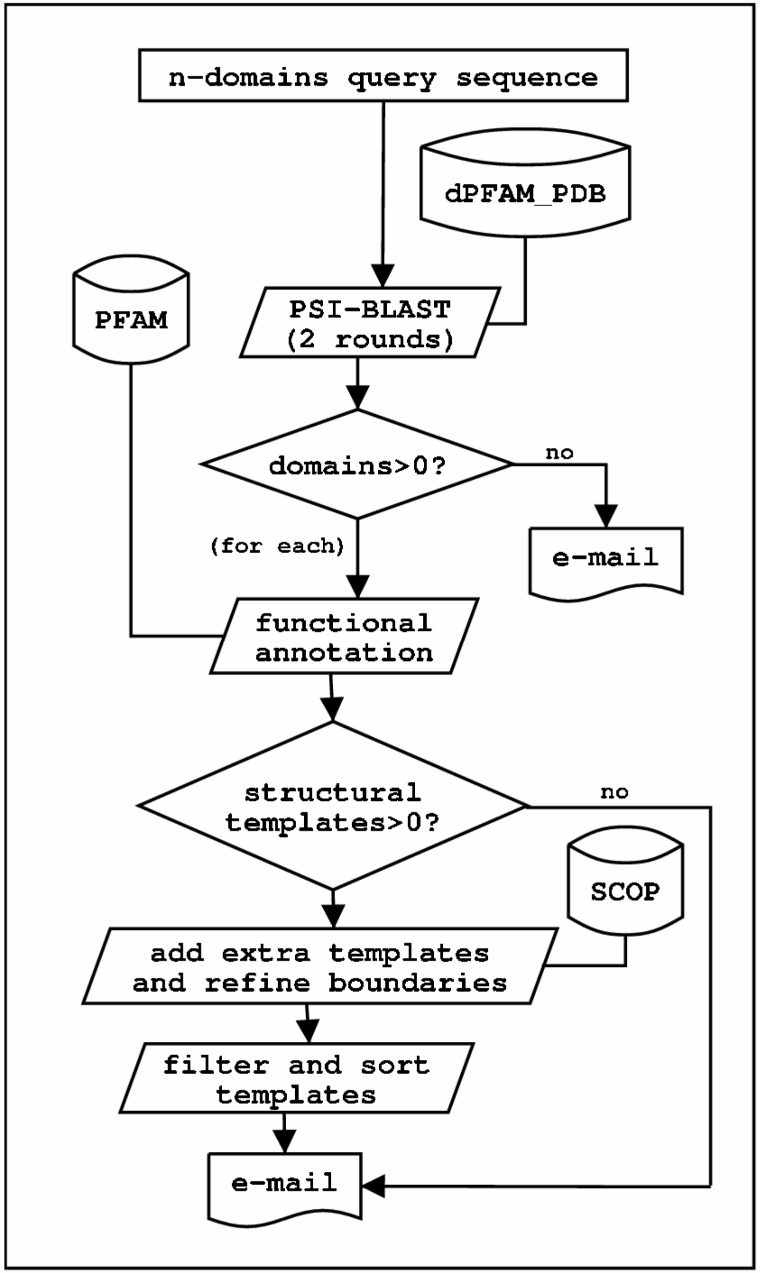
Using our comparative modelling program 3D-Jigsaw (Bates & Sternberg (1999) *Proteins*, Suppl.3:47-54).

selecting templates (2)



DomainFishing

Contreras-Moreira & Bates (2002)
Bioinformatics,18:1141-1142.



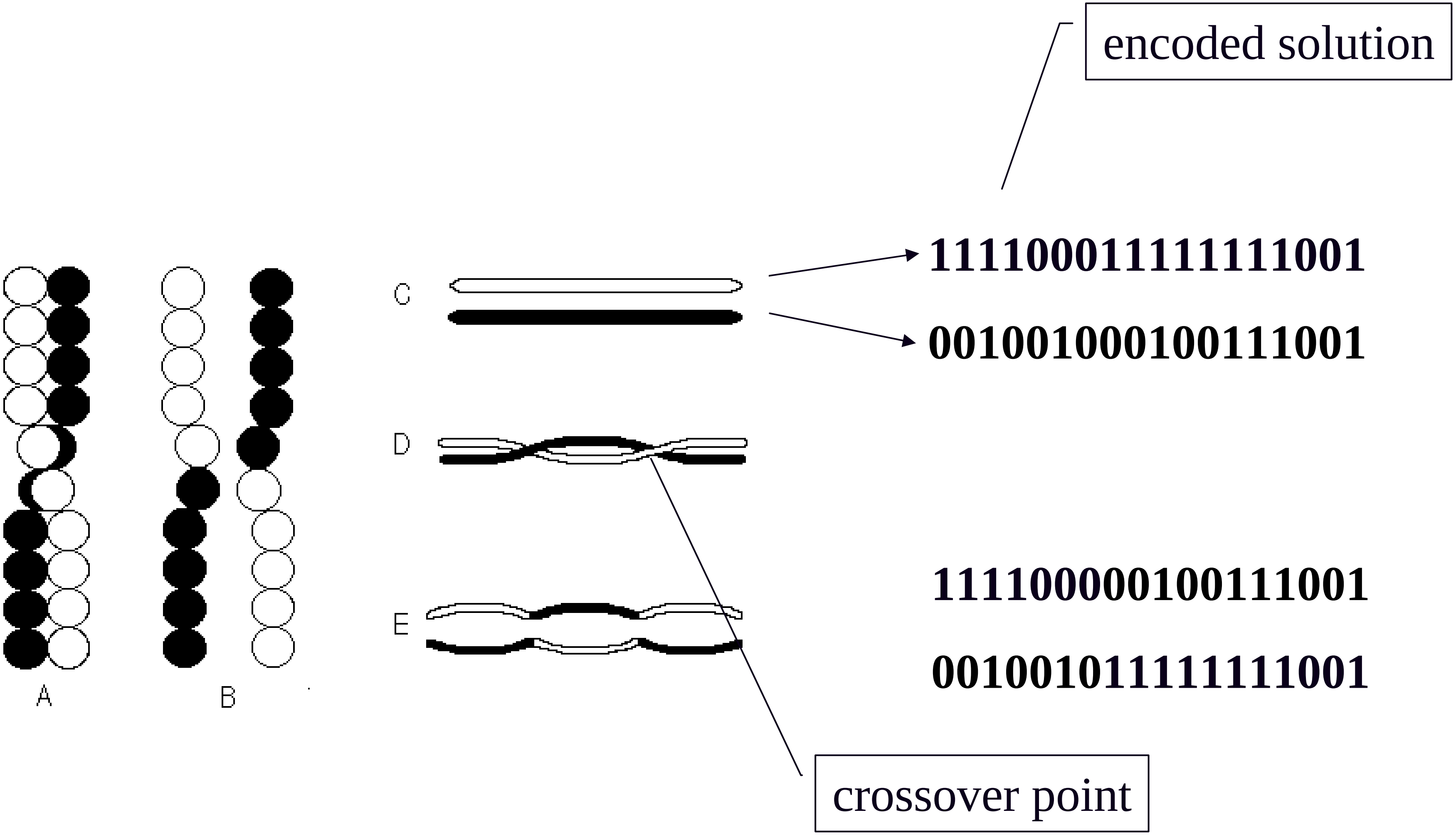
4. Recombining protein models

So far we have learnt:

- Although some alignment techniques are on average better than others, none is perfect and often “worse” procedures produce better alignments.
- Sequence-based evaluators (such as bit-scores) can aid in the task of ranking alignments, but they can’t resolve very similar alignments.
- Selecting templates is not trivial and therefore using only one template is not a good idea.

We concluded that we needed a way of combining different alignments and templates. This was called *in silico protein recombination* and implemented as a genetic algorithm.

chromosome evolution & computational analogy: genetic algorithms



encoded solution

111100011111111001

001001000100111001

111100000100111001

001001011111111001

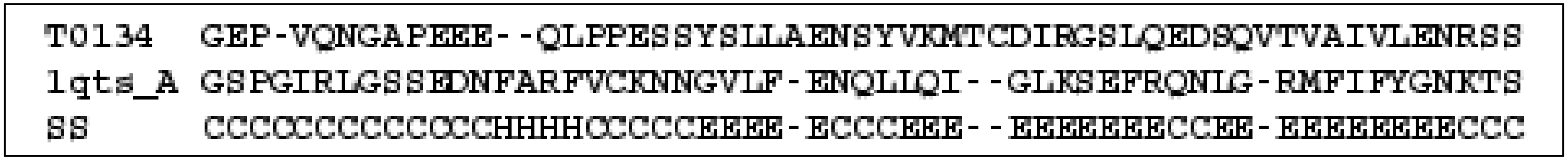
crossover point

a genetic algorithm applied to Comparative Modelling

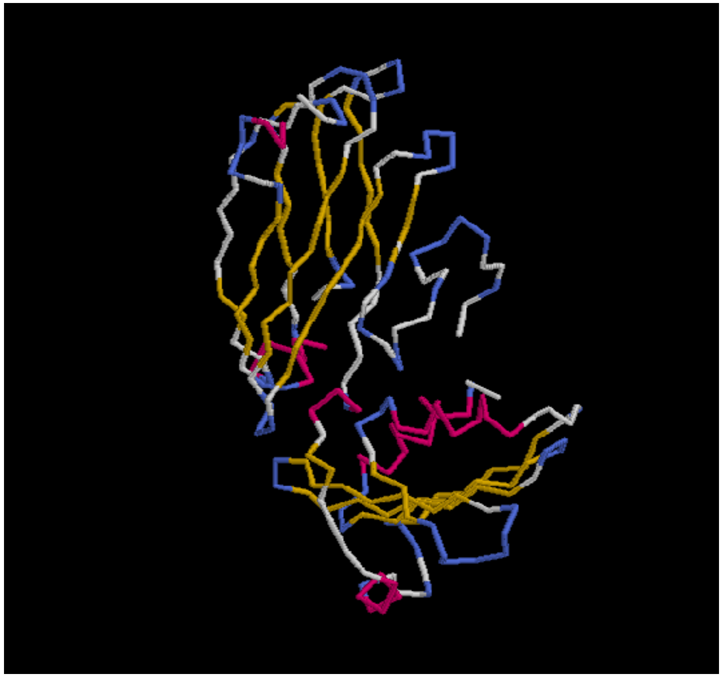
- how are solutions encoded?
- genetic operators
- definition of fitness
- design of the algorithm

proteins models are implicitly coded solutions

- **linear molecules**: strings of residues connected by peptide bonds
- **fitness** = likelihood of its fold



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

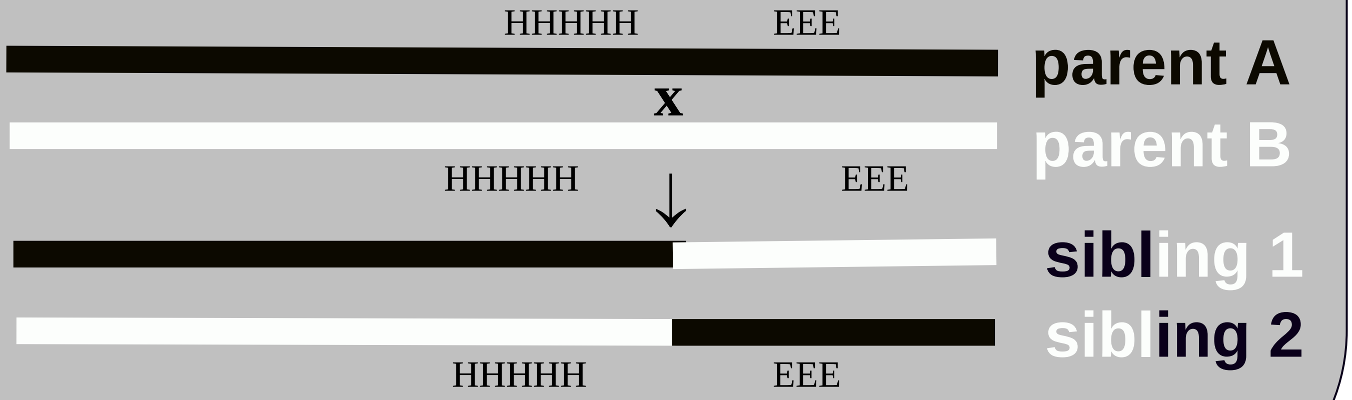


$\text{potential\_solution}_i = \text{model}_i =$

$f(\text{ PDBtemplate}_j \text{ , alignment}_k \text{ })$

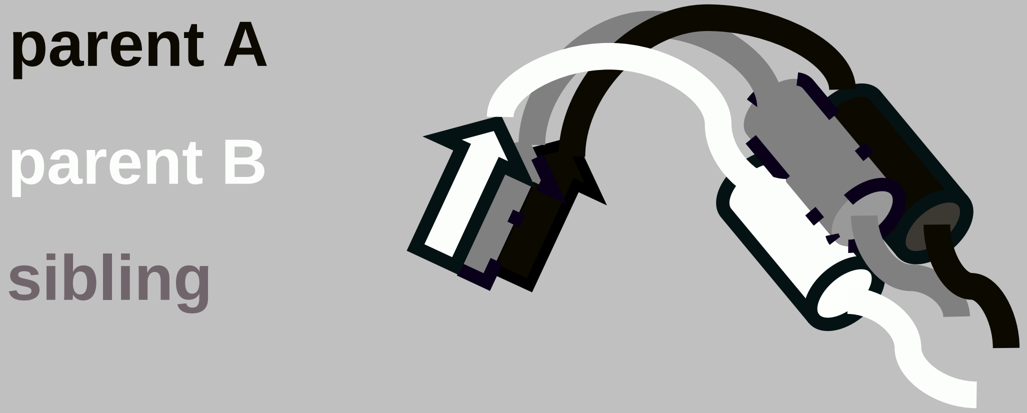
recombination

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



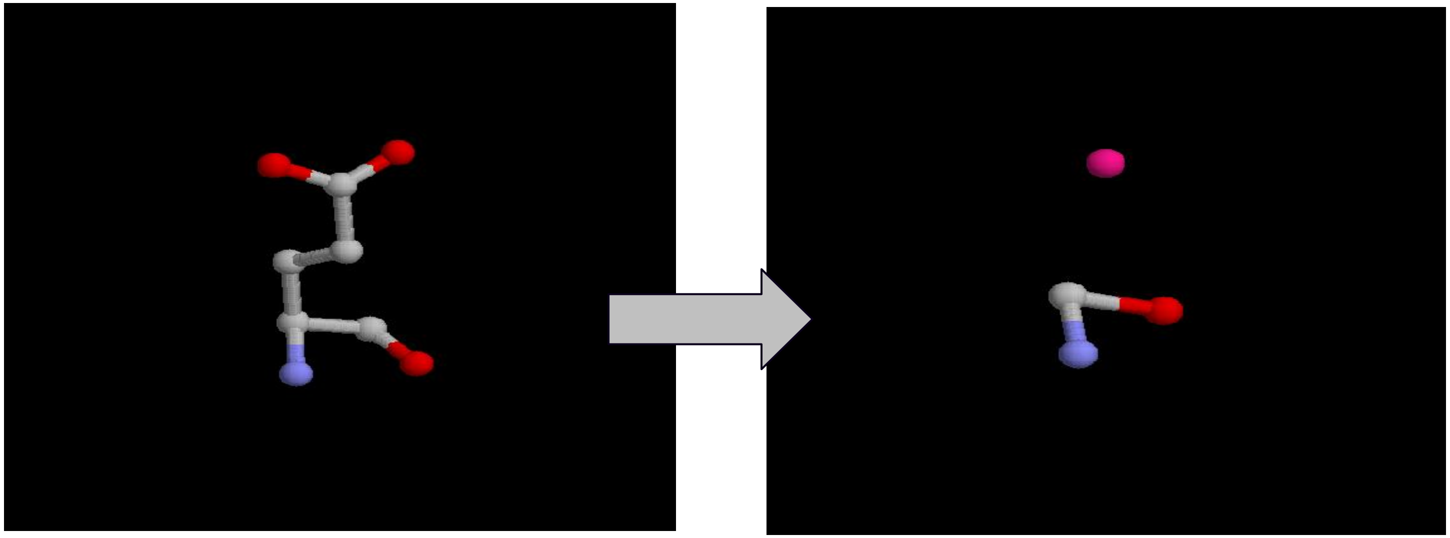
mutation

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



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)

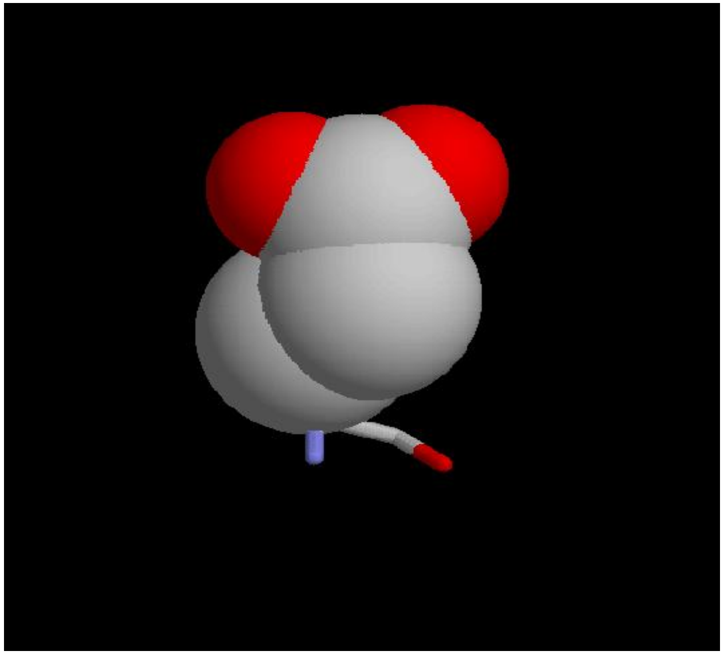
where *i,j* are pairs of pseudoatoms in protein *p*

and *A* and *B* are statistical potentials

(Robson & Osguthorpe (1979) *J.Mol.Biol.*,132:19-51, coded by Paul Fitzjohn)

protein fitness

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



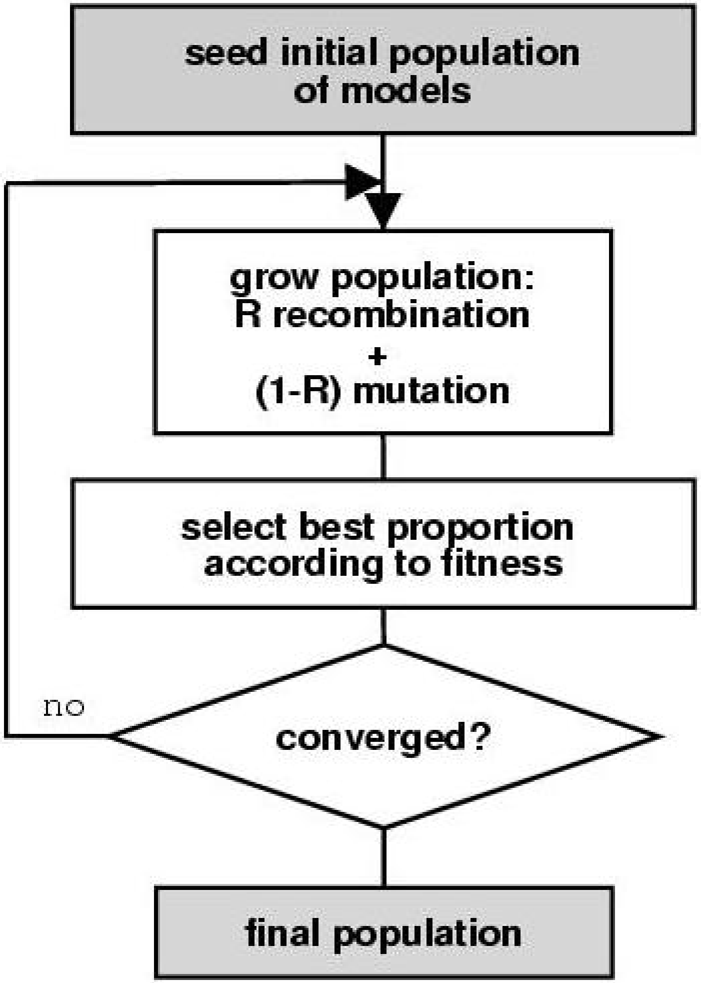
$\sum_i (SA_i \cdot \Delta Gsolv_i)$ (in kcal/mol)

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

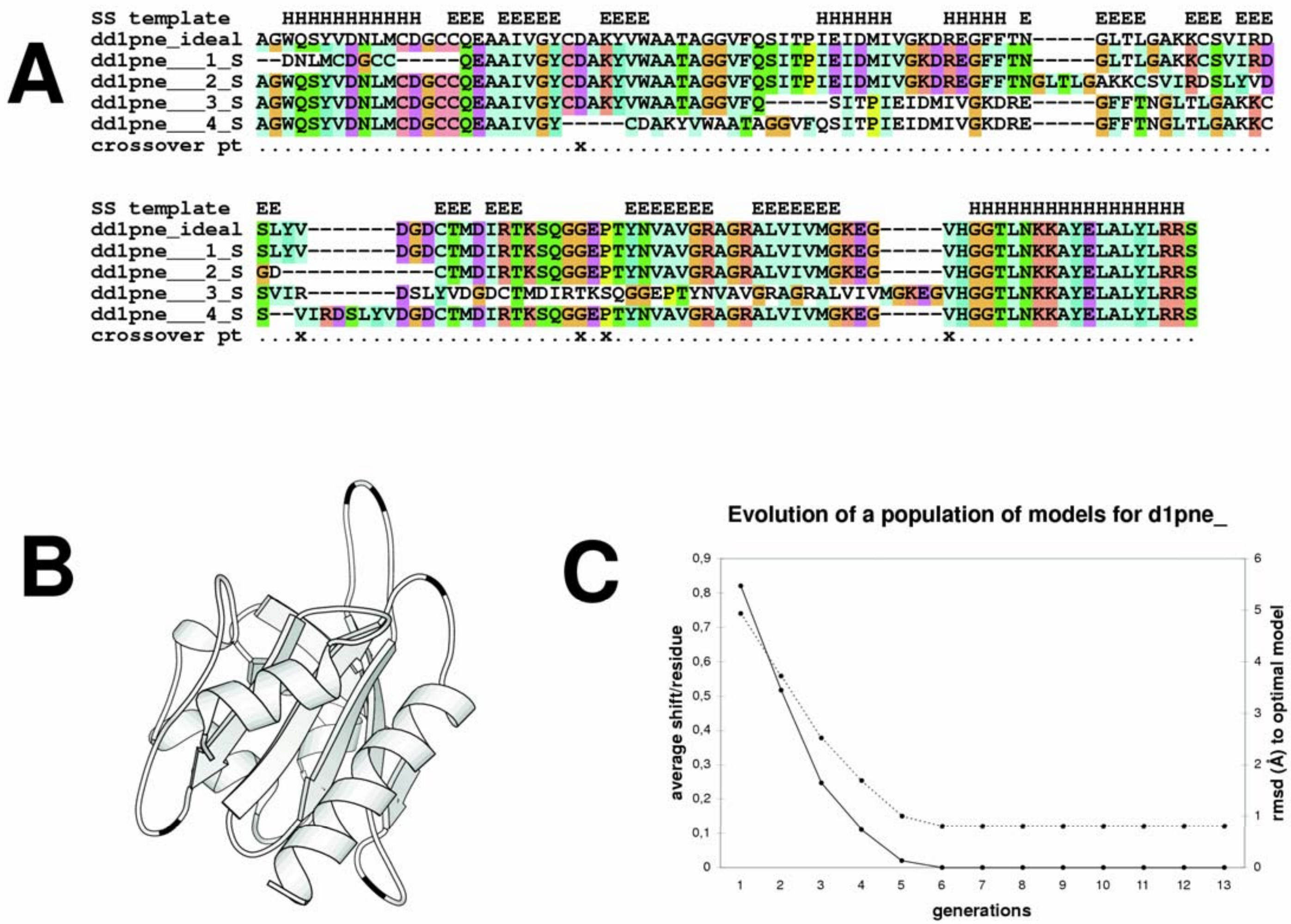
¶ Eisenberg and MacLachlan (1986) *Nature*, **319**: 199-203.

in silico protein recombination algorithm


seed initial population of models
grow population: R recombination + (1-R) mutation
select best proportion according to fitness
converged?
no
final population

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

Protein recombination example: bovine profilin

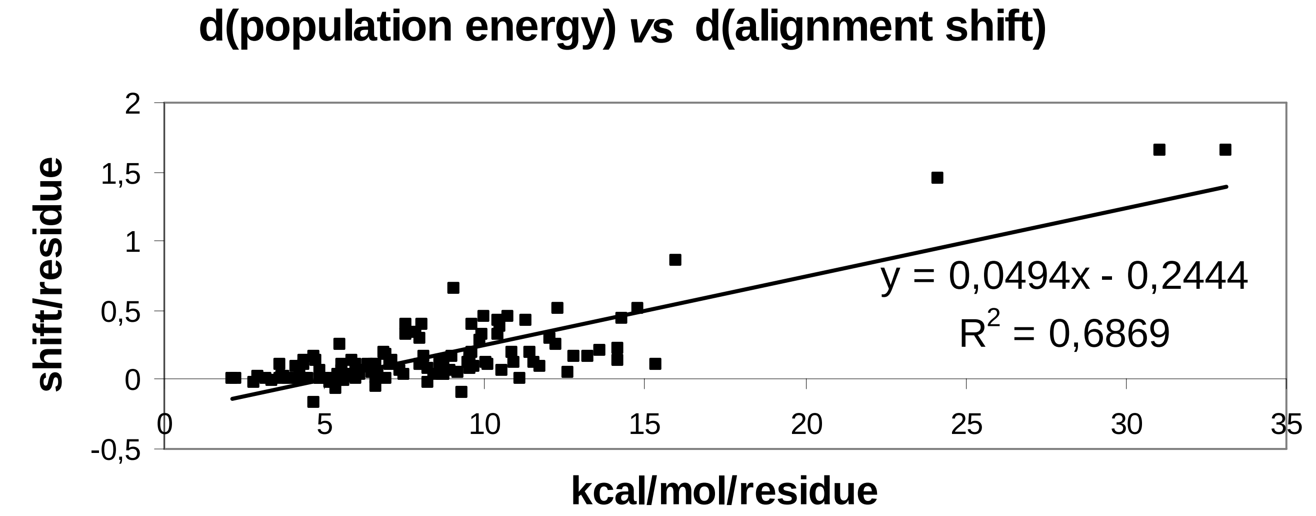


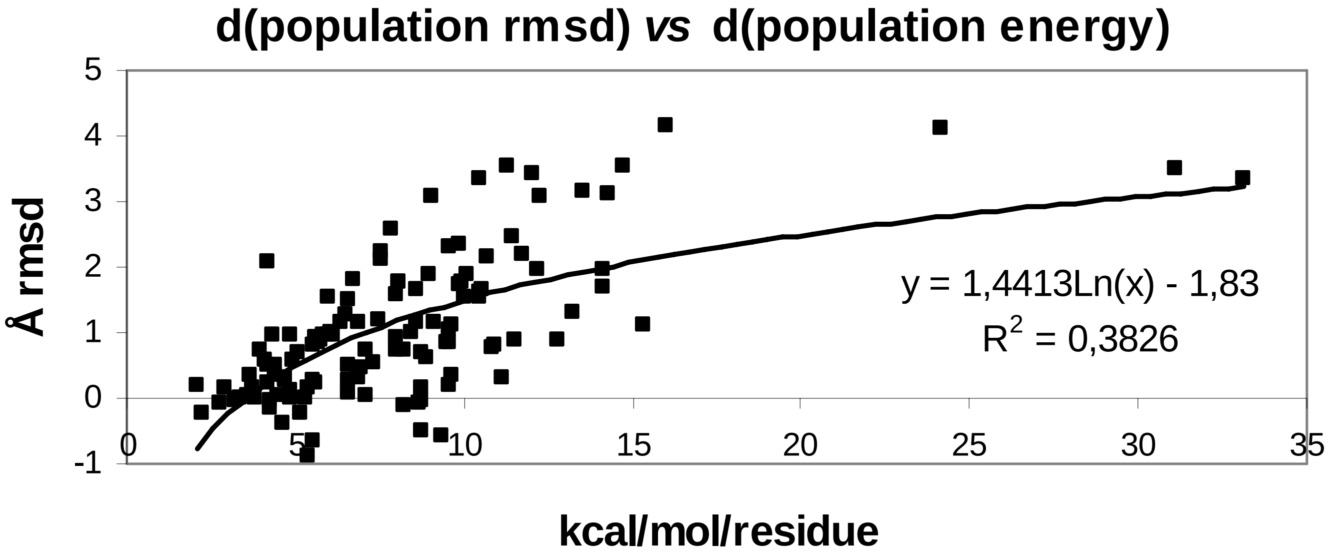
```
SS template    HHHHHHHHHHH  EEE EEEEE   EEEE             HHHHHH    HHHHH E     EEEE   EEE EEE
dd1pne_ideal AGWQSYVDNLMCDGCCQEAAIVGYCDAKYVWAATAGGVFQSITPIEIDMIVGKDREGFFTN-----GLTLGAKKCSVIRD
dd1pne___1_S --DNLMCDGCC-----QEAAIVGYCDAKYVWAATAGGVFQSITPIEIDMIVGKDREGFFTN-----GLTLGAKKCSVIRD
dd1pne___2_S AGWQSYVDNLMCDGCCQEAAIVGYCDAKYVWAATAGGVFQSITPIEIDMIVGKDREGFFTNGLTLGAKKCSVIRDSLYVD
dd1pne___3_S AGWQSYVDNLMCDGCCQEAAIVGYCDAKYVWAATAGGVFQ-----SITPIEIDMIVGKDRE-----GFFTNGLTLGAKKC
dd1pne___4_S AGWQSYVDNLMCDGCCQEAAIVGY-----CDAKYVWAATAGGVFQSITPIEIDMIVGKDRE-----GFFTNGLTLGAKKC
crossover pt .........................x......................................................

SS template  EE            EEE EEE        EEEEEEE   EEEEEEE          HHHHHHHHHHHHHHHHH
dd1pne_ideal SLYV-------DGDCTMDIRTKSQGGEPTYNVAVGRAGRALVIVMGKEG-----VHGGTLNKKAYELALYLRRS
dd1pne___1_S SLYV-------DGDCTMDIRTKSQGGEPTYNVAVGRAGRALVIVMGKEG-----VHGGTLNKKAYELALYLRRS
dd1pne___2_S GD------------CTMDIRTKSQGGEPTYNVAVGRAGRALVIVMGKEG-----VHGGTLNKKAYELALYLRRS
dd1pne___3_S SVIR-------DSLYVDGDCTMDIRTKSQGGEPTYNVAVGRAGRALVIVMGKEGVHGGTLNKKAYELALYLRRS
dd1pne___4_S S--VIRDSLYVDGDCTMDIRTKSQGGEPTYNVAVGRAGRALVIVMGKEG-----VHGGTLNKKAYELALYLRRS
crossover pt ...x.....................x.x..........................x...................
```

1pne, Cedergen-Zeppezauer et al. (1994) *J.Mol.Biol.*,240:459-475.

protein recombination: performance

(in-house benchmark on 130 protein families)



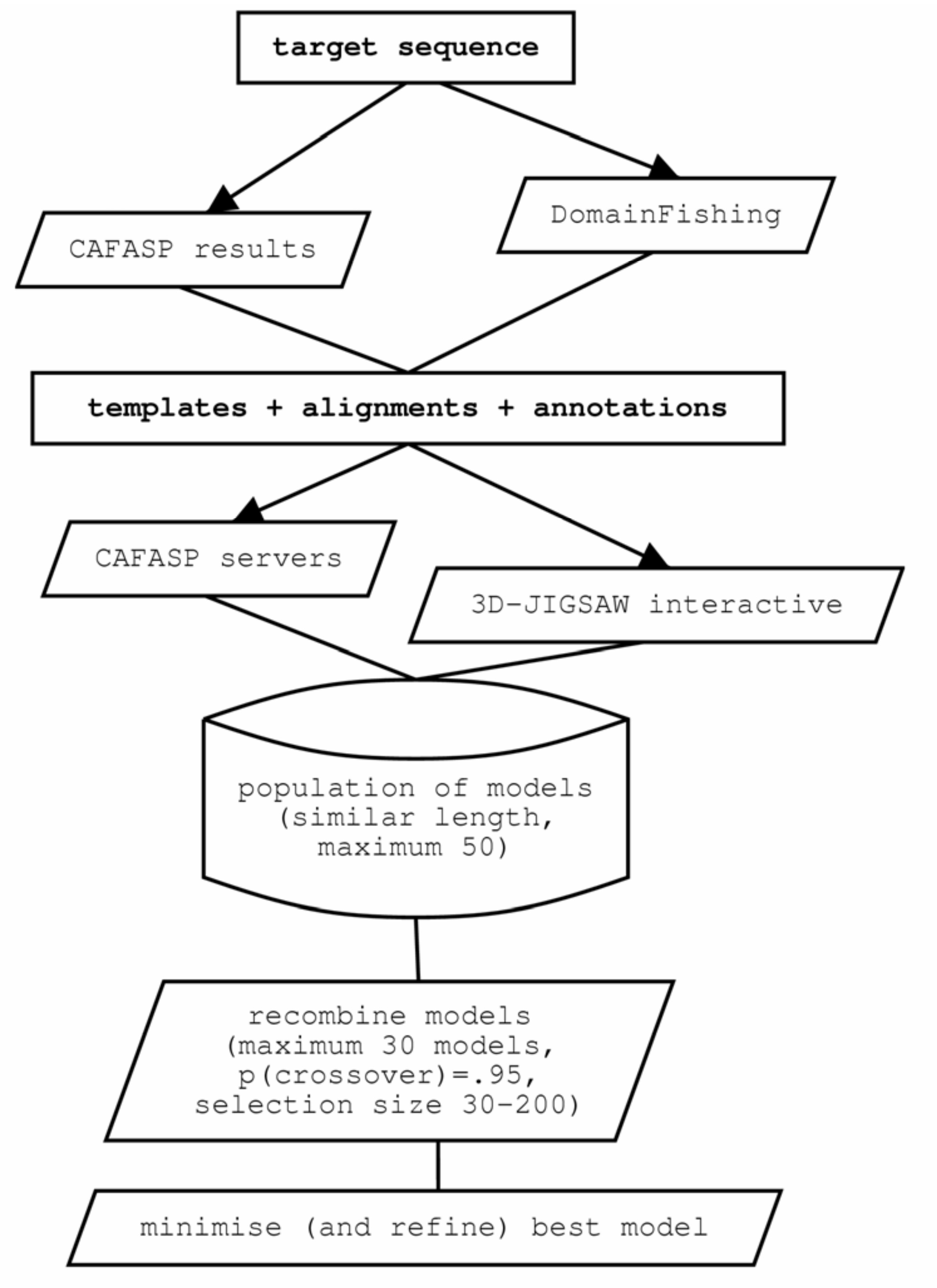


protein recombination: CASP5 benchmark

CASP5: 5th Critical Assessment of techniques for protein Structure Prediction (67 proteins).

Contreras-Moreira,Fitzjohn, Offman, Smith & Bates (2003) *Proteins*, 53:424-429.

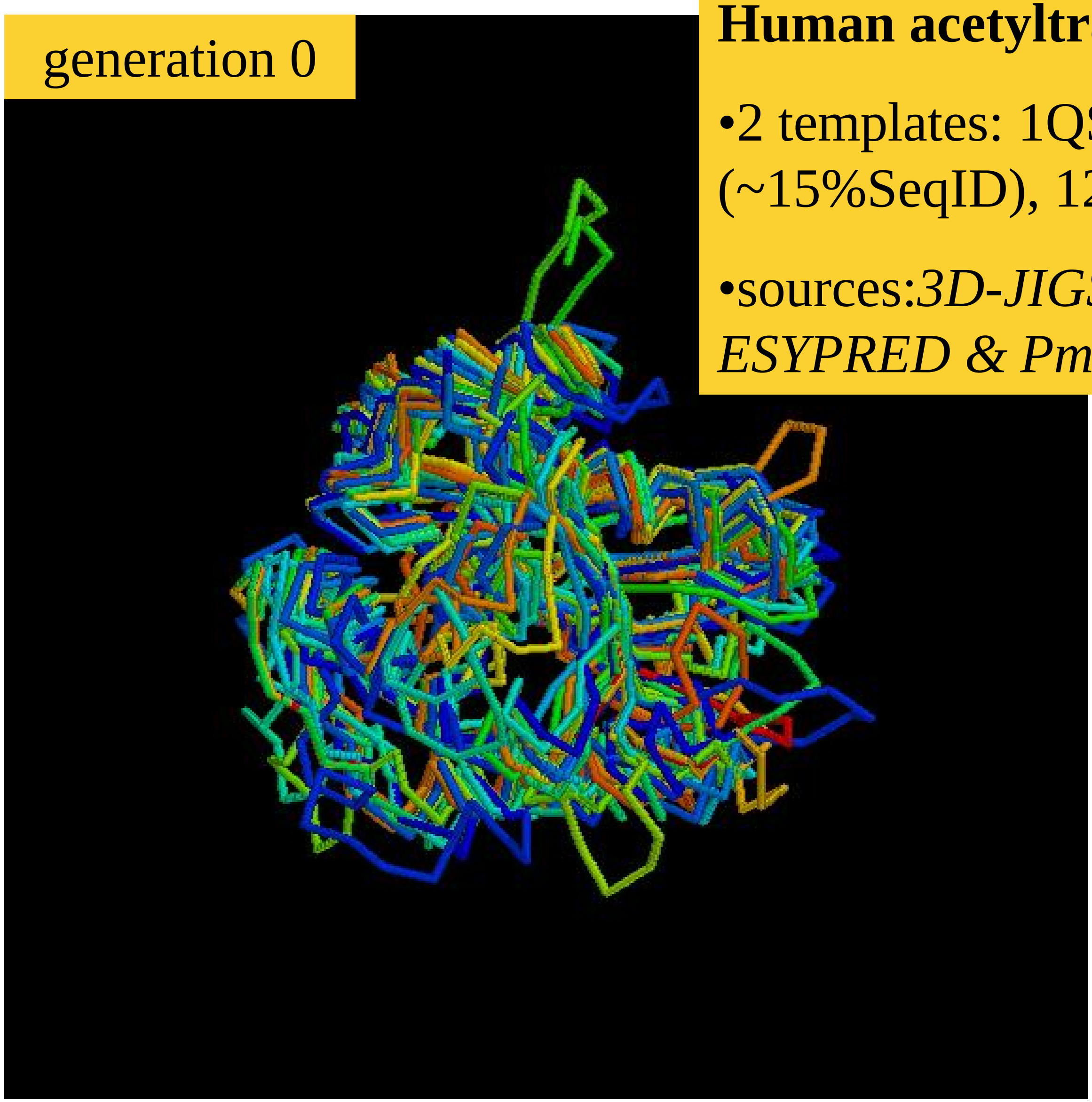
CAFASP is a web server that collects automatic predictions from servers around the world.

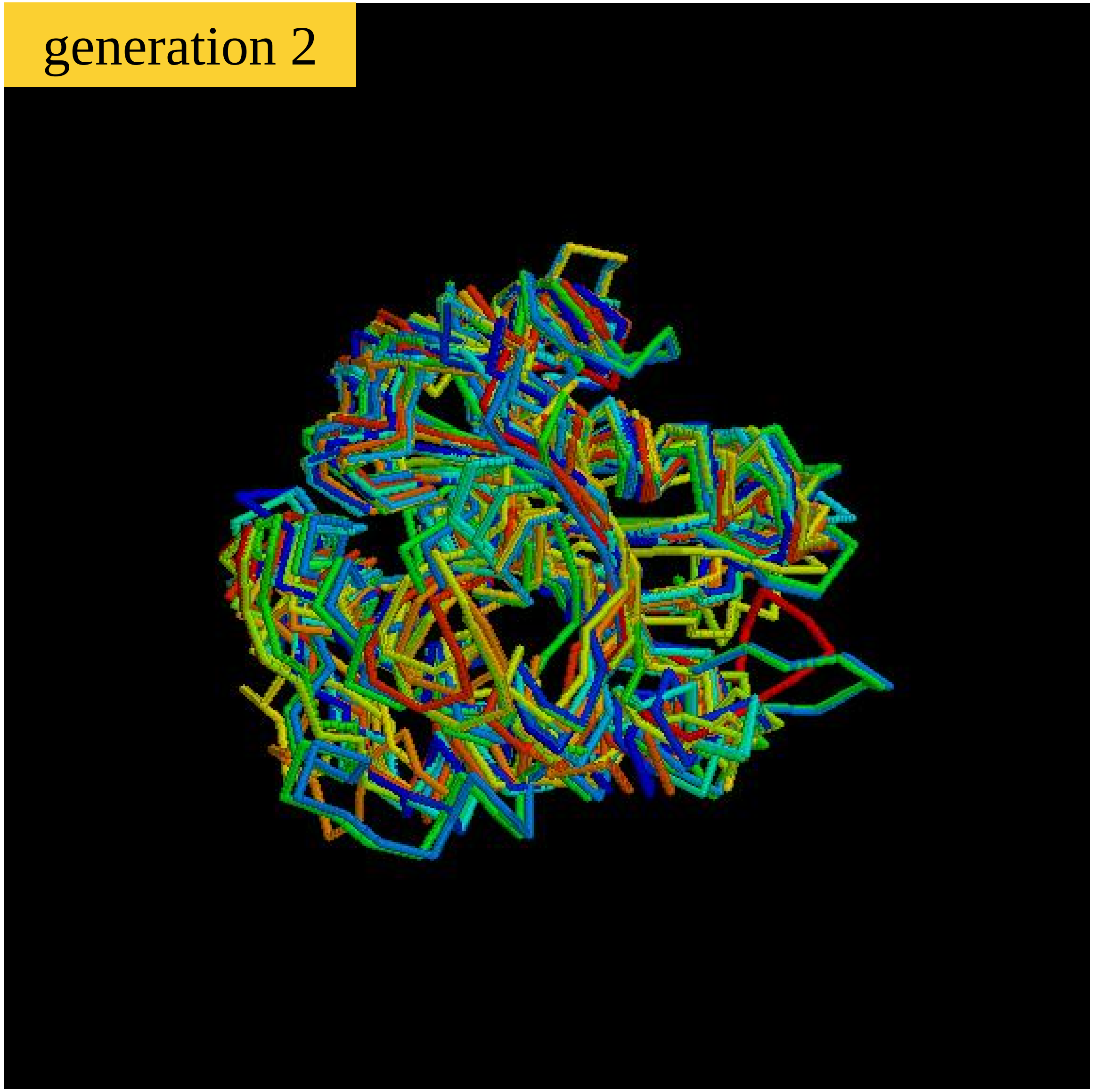

target sequence
CAFASP results
DomainFishing
templates + alignments + annotations
CAFASP servers
3D-JIGSAW interactive
population of models (similar length, maximum 50)
recombine models (maximum 30 models, p(crossover)=.95, selection size 30-200)
minimise (and refine) best model

CASP5 example: T0192

Human acetyltransferase

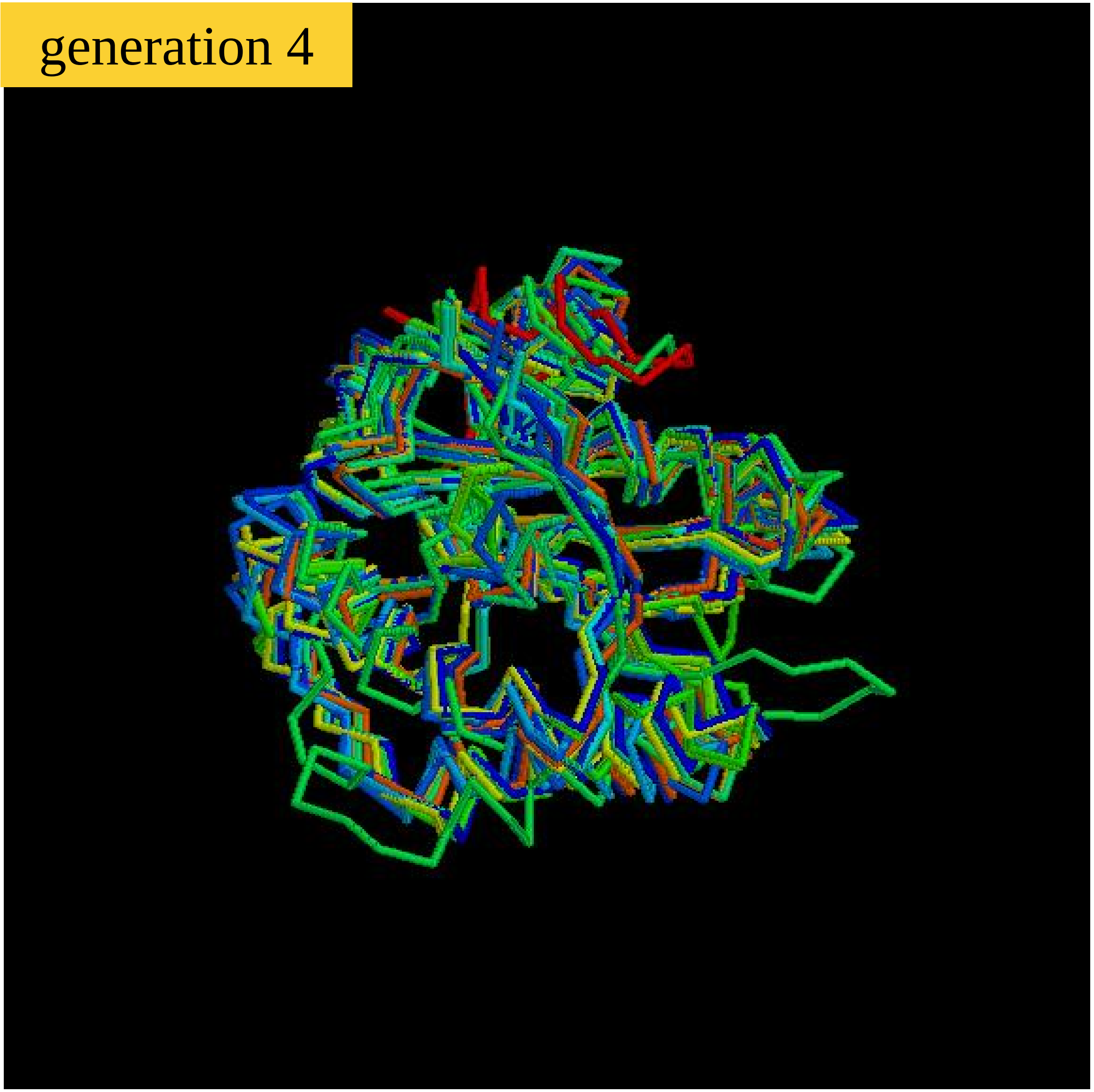
- 2 templates: 1QSM & 1QSO (~15%SeqID), 12 alignments
- sources:*3D-JIGSAW,FAMS, ESYPRED & Pmodeller*

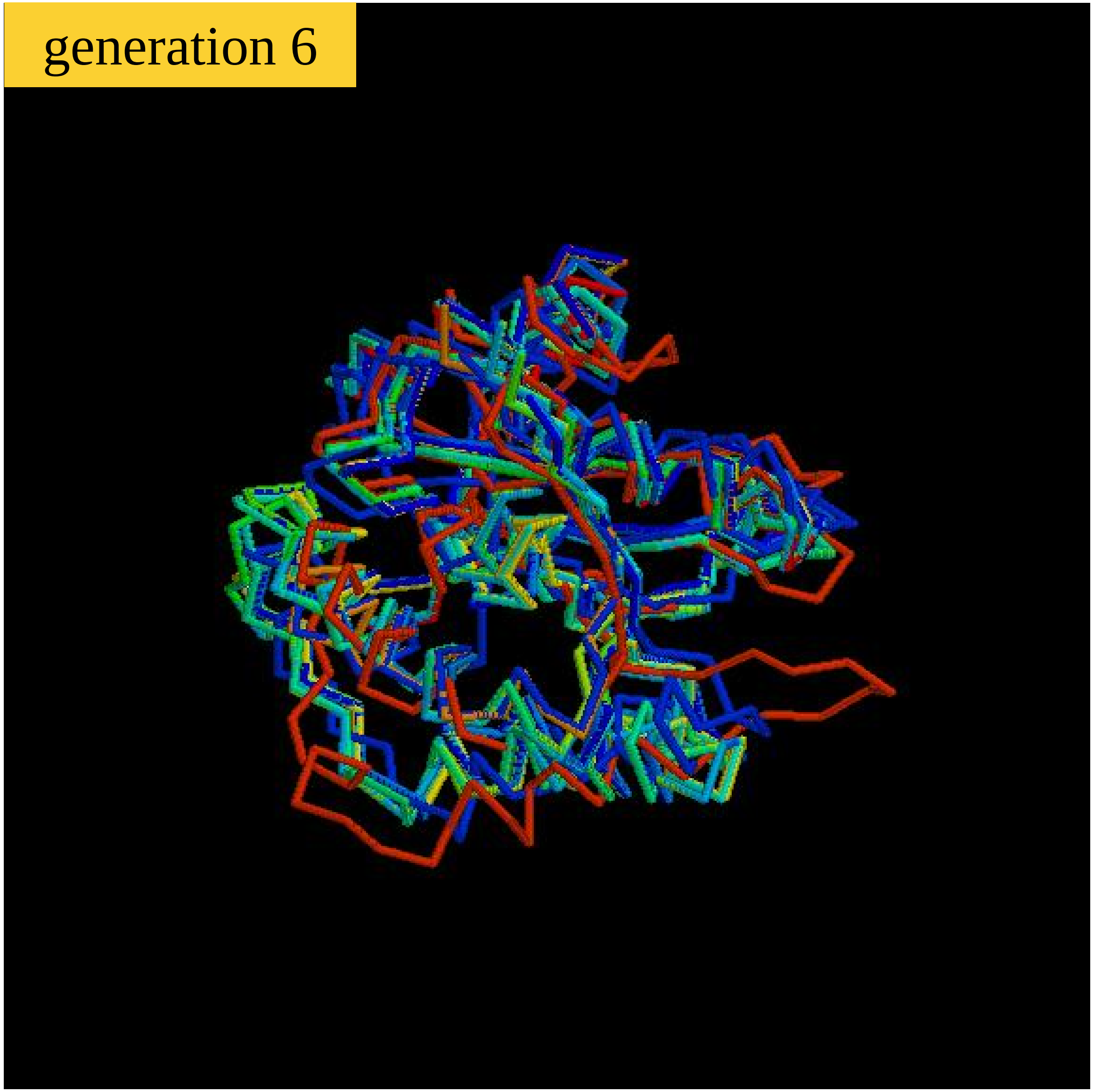




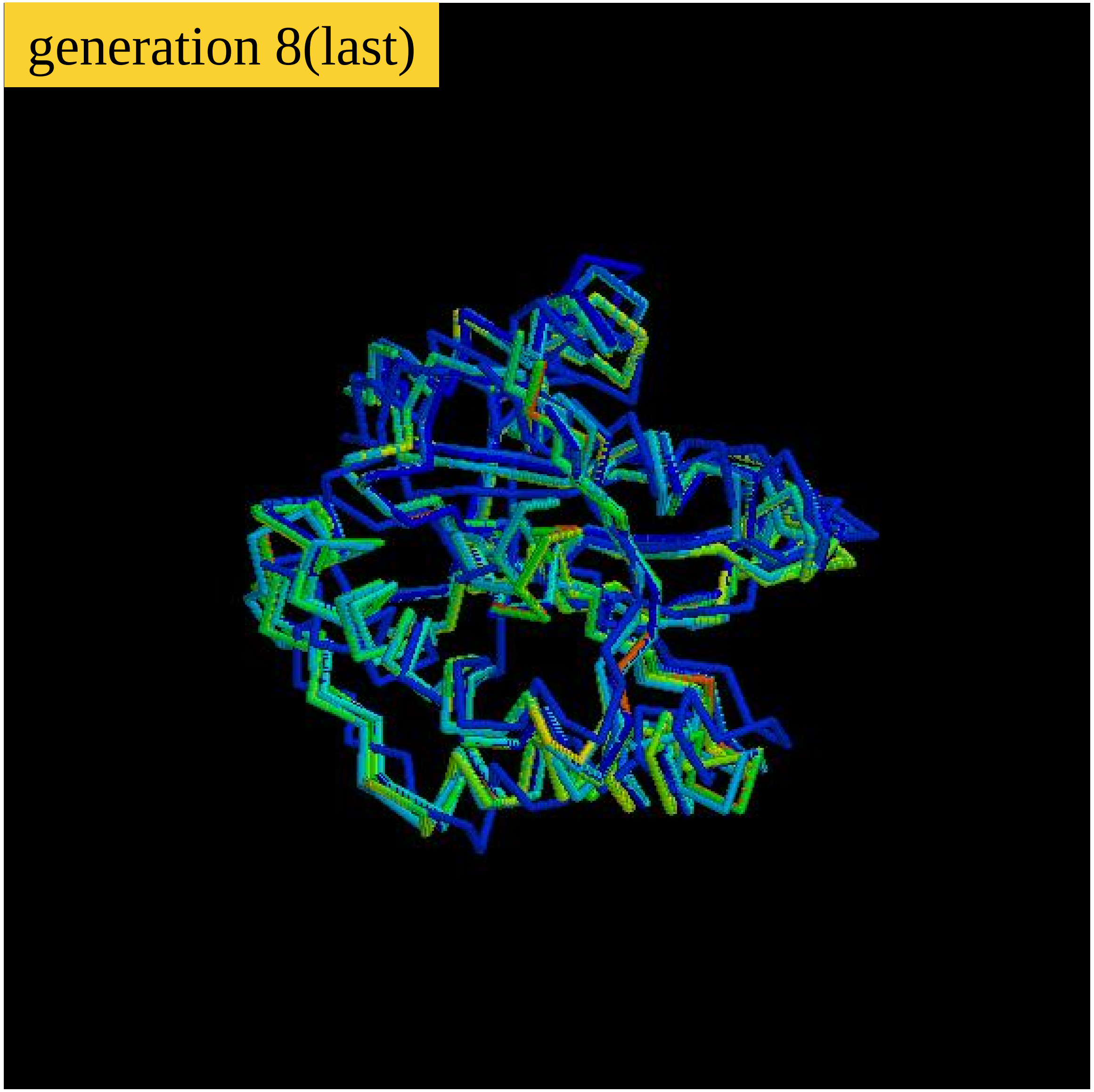
generation 2

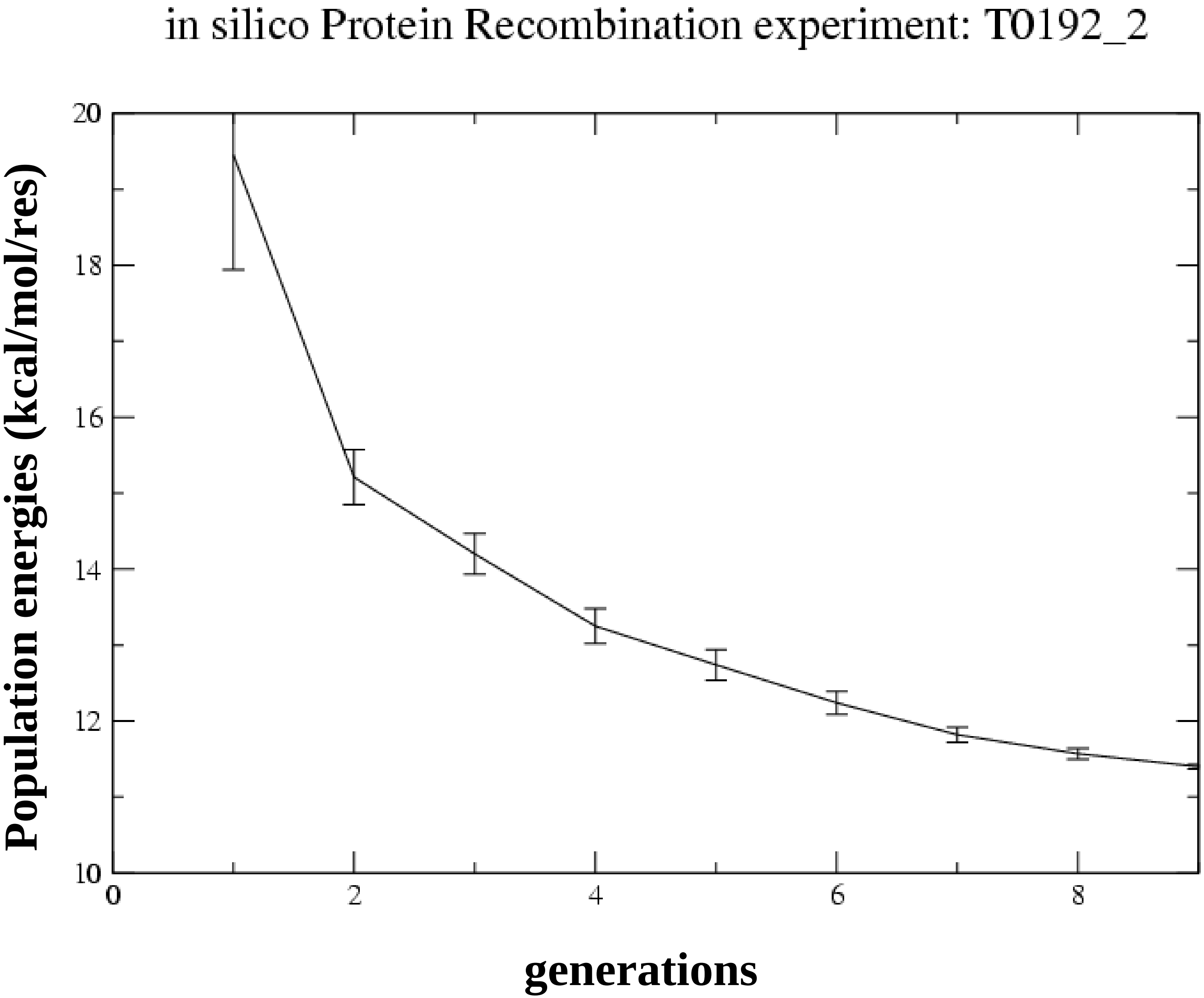
generation 4




generation 6

generation 8(last)

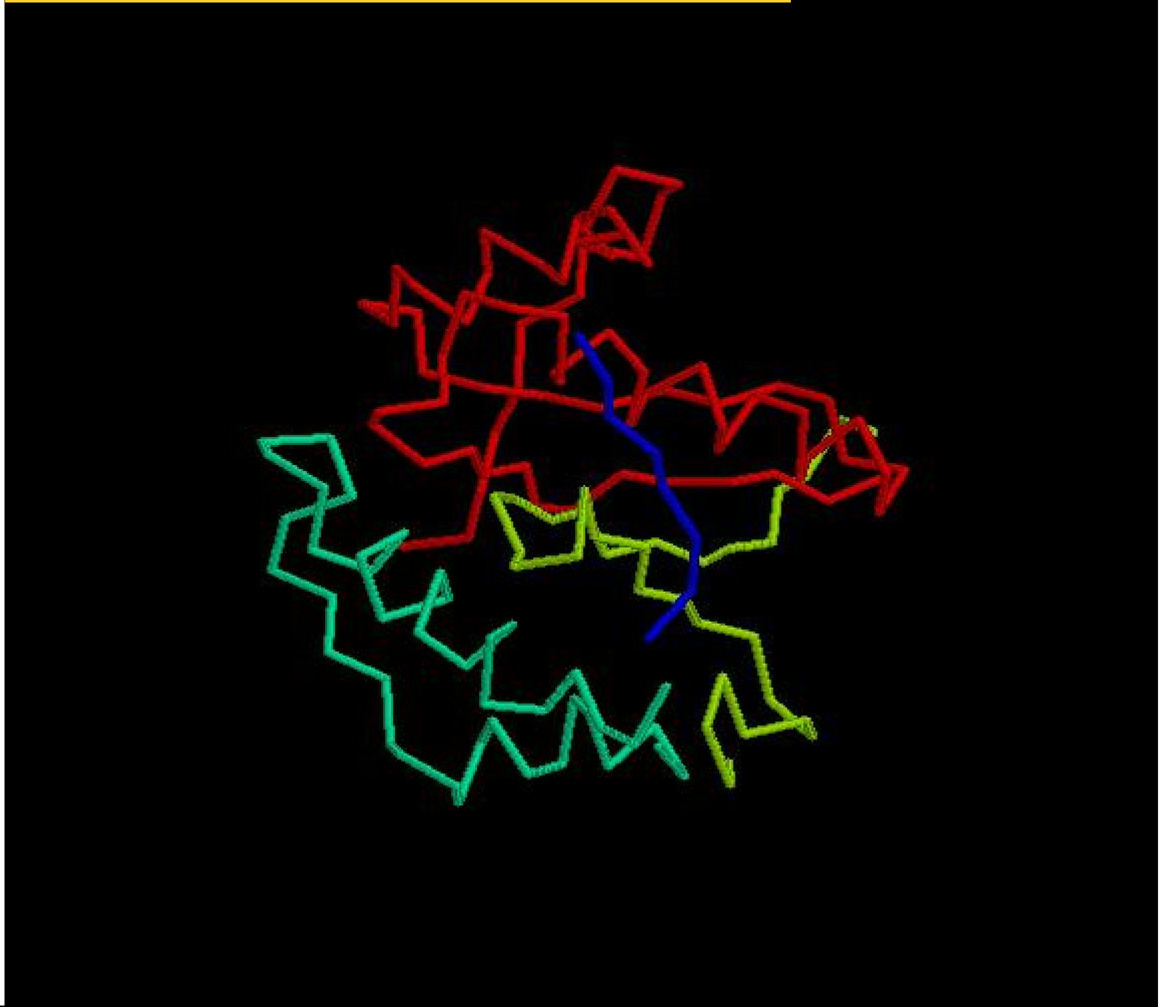




in silico Protein Recombination experiment: T0192_2

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: CASP5 summary

- Targets with an obvious fold, assessed by Anna Tramontano (La Sapienza, Roma):
  - protein recombination is among the 10 top methods (out of ~200) in terms of alignment quality, but is worse in atomic deviation terms (RMSD).

- Fold recognition targets, evaluated by Nick Grishin (Howard Hughes Medical Institute,Dallas):
  - protein recombination is among the top 10 methods in both alignment and RMSD terms.

in silico protein recombination: evaluation

ADVANTAGES

- converges close to the best initial model in a population
- it is able to recover some alignment errors
- often last population contains alternative conformations (?)

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

5. A relation between exonic structure of genes and protein structure (in collaboration with Páll Jónsson)

Protein set: 684 human and mouse experimental structures from the PDB (100< size <300 res) with their intron-exon boundaries mapped by aligning their amino acid sequence back to their genomic DNA sequence.

Contreras-Moreira, Jónsson & Bates (2003) *J.Mol.Biol.*,333:1057-1071.

Intron-exon boundaries in the context of 2^{ary} structure

Secondary structure, 3-state structure	$\mathbf{f_{obs\ introns}}$	$\mathbf{f_{exp\ introns}}$	**Difference**
C - Not in a secondary structure element (loops)	**776 (32%)**	**544 (22%)**	**+43%**
C - Residue in isolated β-bridge	29 (1%)	31 (1%)	-6%
C – Hydrogen-bonded turn	308 (13%)	288 (12%)	+7%
C – Bend	260 (11%)	265 (11%)	-2%
E - Extended β-strand	**430 (18%)**	**537 (22%)**	**-20%**
H - α-helix	**570 (23%)**	**702 (29%)**	**-19%**
H - 3_{10} helix	73 (3%)	80 (3%)	-9%
H – 5-helix	1 (0%)	0 (0%)	-

Subset of intron-exon boundaries	$\mathbf{end_{obs}}$	$\mathbf{end_{exp}}$	$\mathbf{mid_{obs}}$	$\mathbf{mid_{exp}}$
all β-strands	**184 (41%)**	**45 (10%)**	**266 (59%)**	**405 (90%)**
conserved β-strands	**13 (21%)**	**6 (10%)**	**49 (79%)**	**56 (90%)**
all α-helices	**114 (20%)**	**58 (10%)**	**465 (80%)**	**521 (90%)**
conserved α-helices	**15 (25%)**	**6 (10%)**	**45 (75%)**	**54 (90%)**

Intron-exon boundaries & protein function

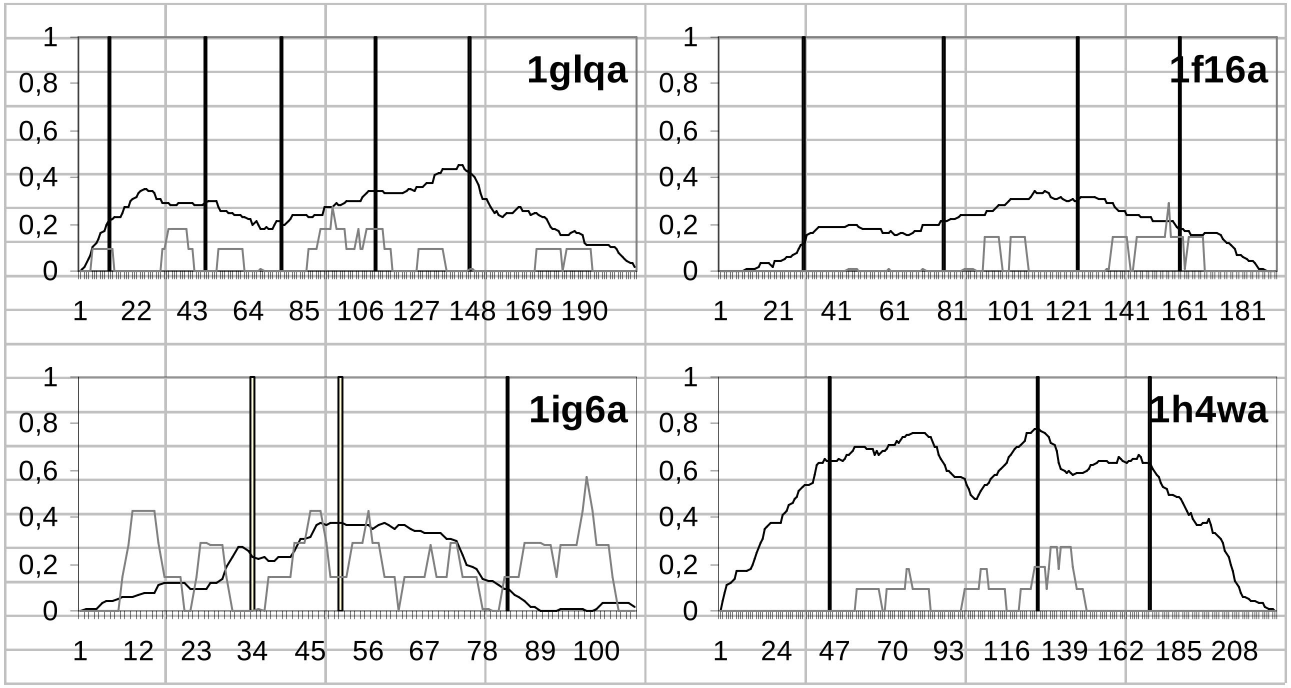
test	**Obs**	**Exp**
Intron-exon boundaries <7Å functional sites	**55/308 (18%)**	**51/308 (17%)**
Intron-exon boundaries separate functional residues	**106/308 (34%)**	**100 (32%)**

Intron-exon boundaries & protein recombination

PDB chain	**annotation**	**Number of templates used for recombination and sequence identity range**	**Origin of homologous proteins (templates)**
1a66a	Rel homology domain, eukaryotic transcription factor.	11, 100%-23%	*H.sapiens, M.musculus, Anopheles gambiae*
1bv8a	Alpha-2-macroglobulin.	3, 100%-62%	*H.sapiens, Paracoccus denitrificans, R.norvegicus*
1b4qa	Glutaredoxin.	10, 100%-20%	*H.sapiens, phage T4, E.coli, S.scrufa*
1h4wa	Trypsin	14, 100%-38%	*R.rattus,S.scrufa,B.taurus, H.sapiens,E.coli, R.norvegicus*
…(22)			

NOTE: all cross-overs are allowed

Intron-exon boundaries & protein recombination



Over the 22 test cases there are 71 intron-exon boundaries, of which 56 (79%) have less than 5% of recombination frequency, compared to 65% expected by chance. The probability of this being a random deviation is p=0.01 for a χ^2_{1df}.

6. Conclusions

1. Sequence alignment techniques are not perfect and, although it is possible to rank them, in certain situations “weaker” techniques can perform better than ”stronger” ones.

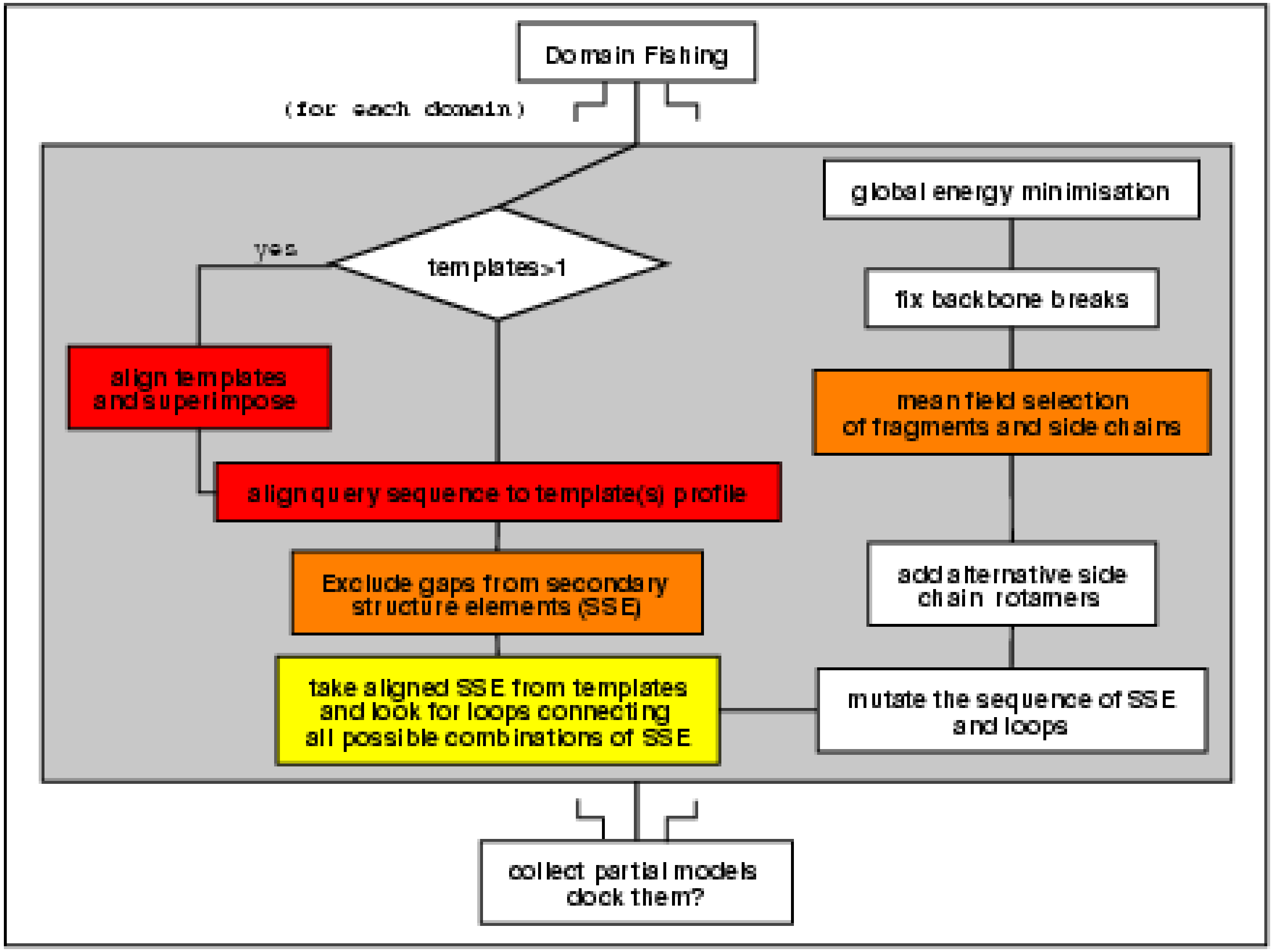
2. Protein recombination is able to construct protein models in a robust manner, with the ability to resolve at least some alignment conflicts and therefore correct errors. Our results (and others in CASP5) suggest this combinatorial approach can be equally useful for Fold Recognition purposes.

3. Introns do not populate randomly the genes in which they live, especially when protein secondary structure is considered. The observed preferences can be exploited for protein engineering purposes.

Biomolecular Modelling Laboratory

www.bmm.icnet.uk

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

mutate the sequence of SSE and loops

add alternative side chain rotamers

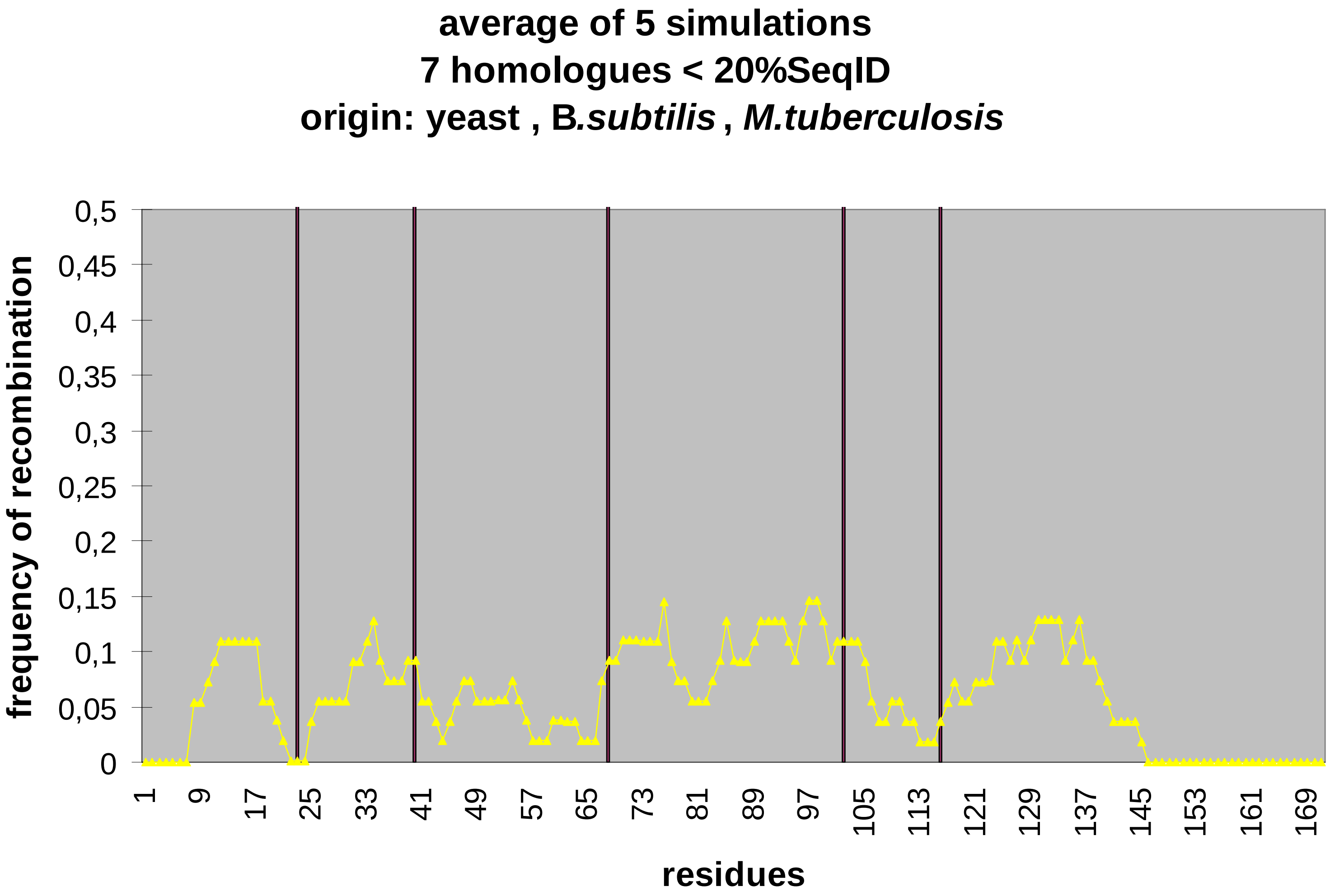
mean field selection of fragments and side chains

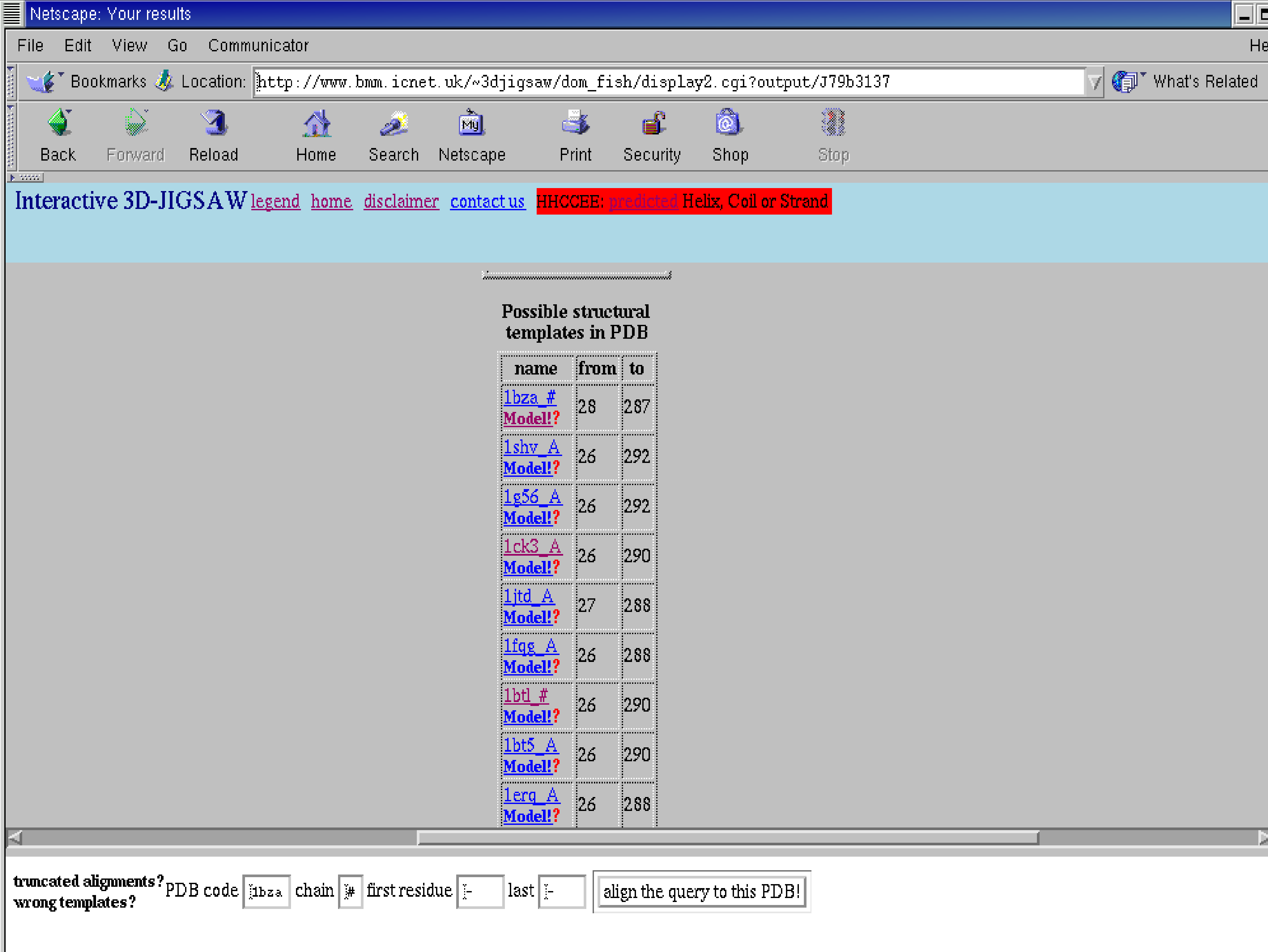
fix backbone breaks

global energy minimisation

collect partial models dock them?

Crossover points and introns boundaries: T0192





http://www.bmm.icnet.uk/~3djigsaw/dom_fish/display2.cgi?output/J79b3137

Interactive 3D-JIGSAW legend home disclaimer contact us HHCCEE: predicted Helix, Coil or Strand

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!