βαλβεσ

A Molecular Replacement Pipeline

Fei Long

MRC-LMC Cambridge
UK

CCP4

MRC | Laboratory of Molecular Biology

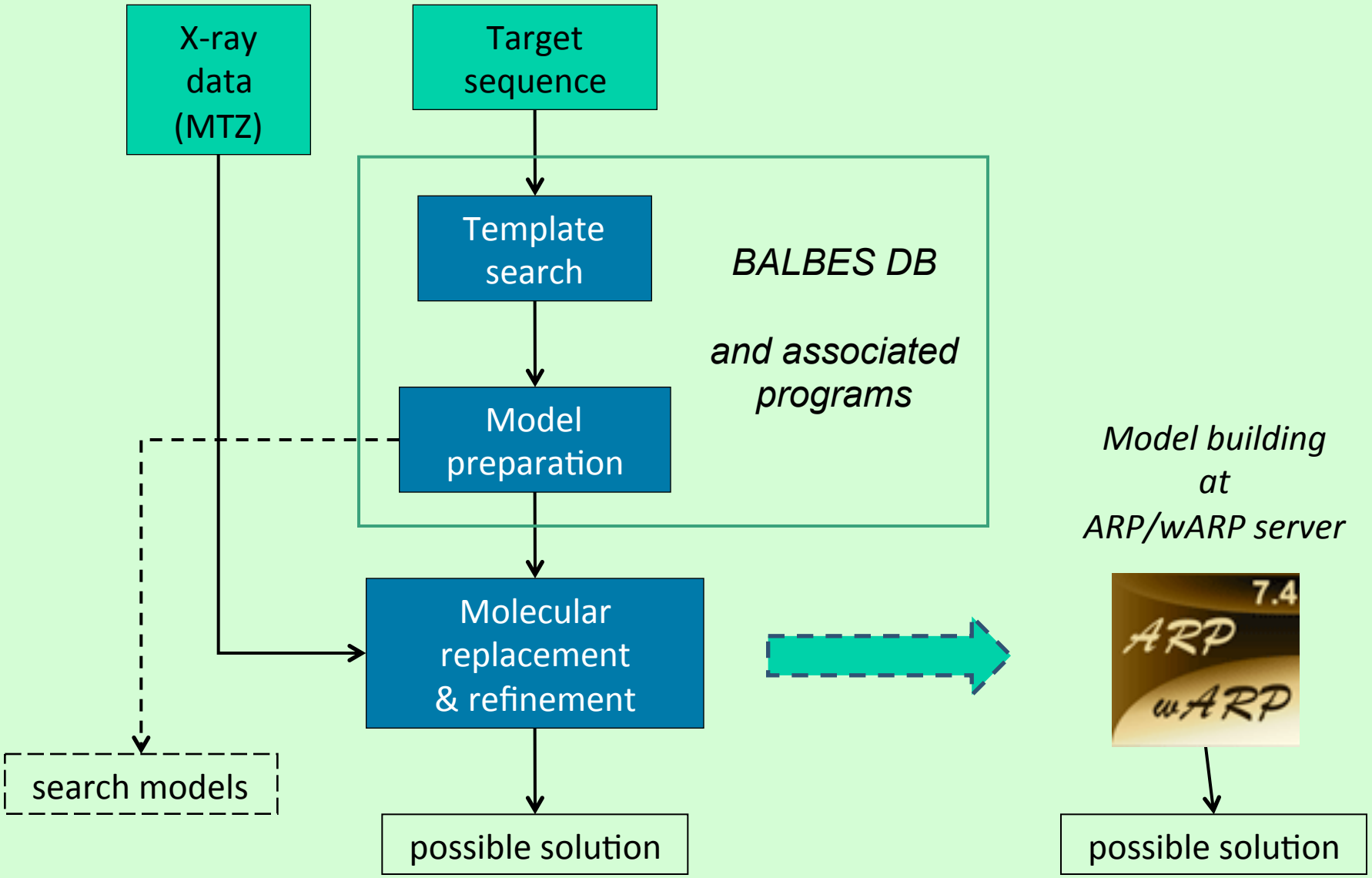
What is BALBES

BALBES: An automated system integrating all components required to solve protein structures via molecular replacement methods(MR).

- An internal database specifically designed for MR.
- Decision making protocols and algorithms for tasks at different stage during the structure-solving processes.
- The state of art engine programs such as MOLREP, REFMAC and others.

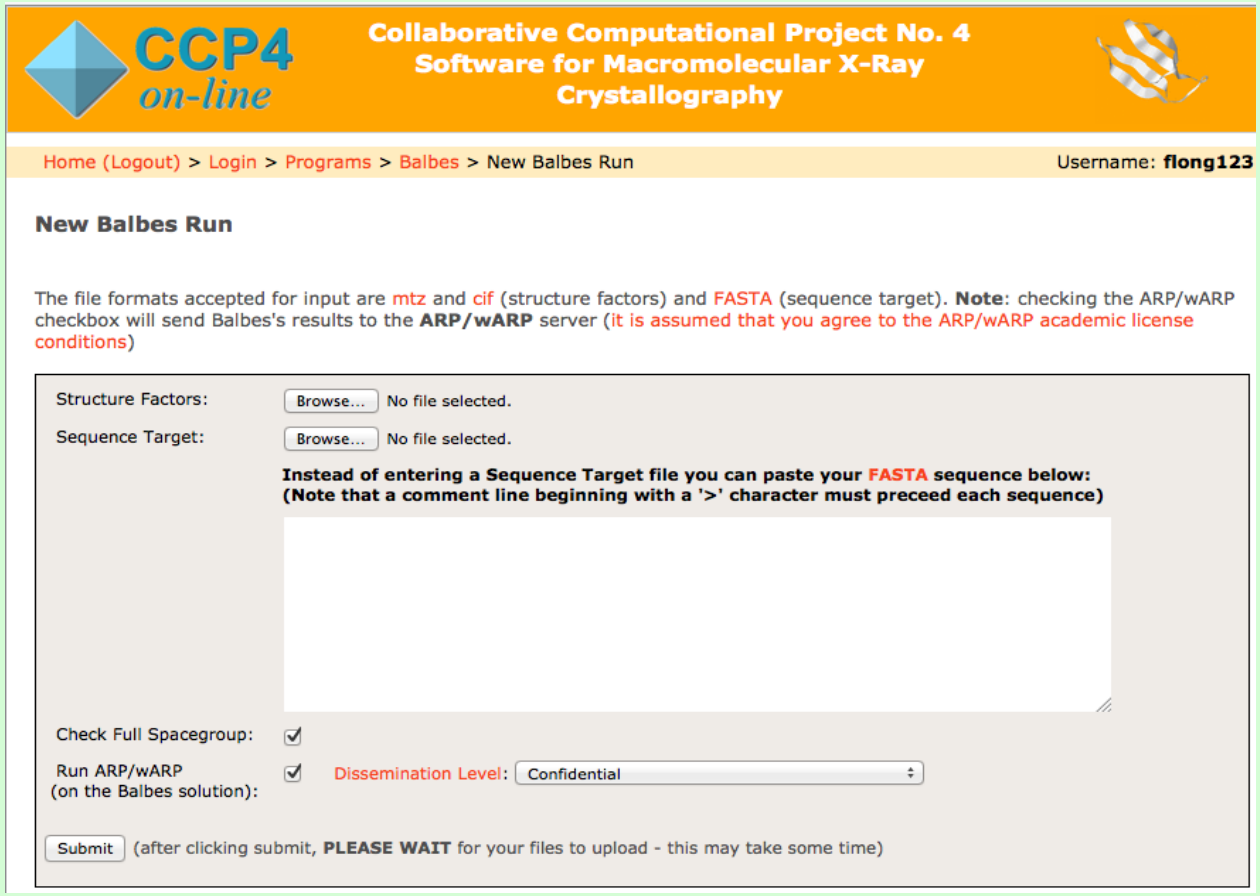
As an automated pipeline, BALBES tries *to minimise users' intervention. The only thing a user needs to do is to provide two input files (a structure factor file and a sequence file)*

What is BALBES



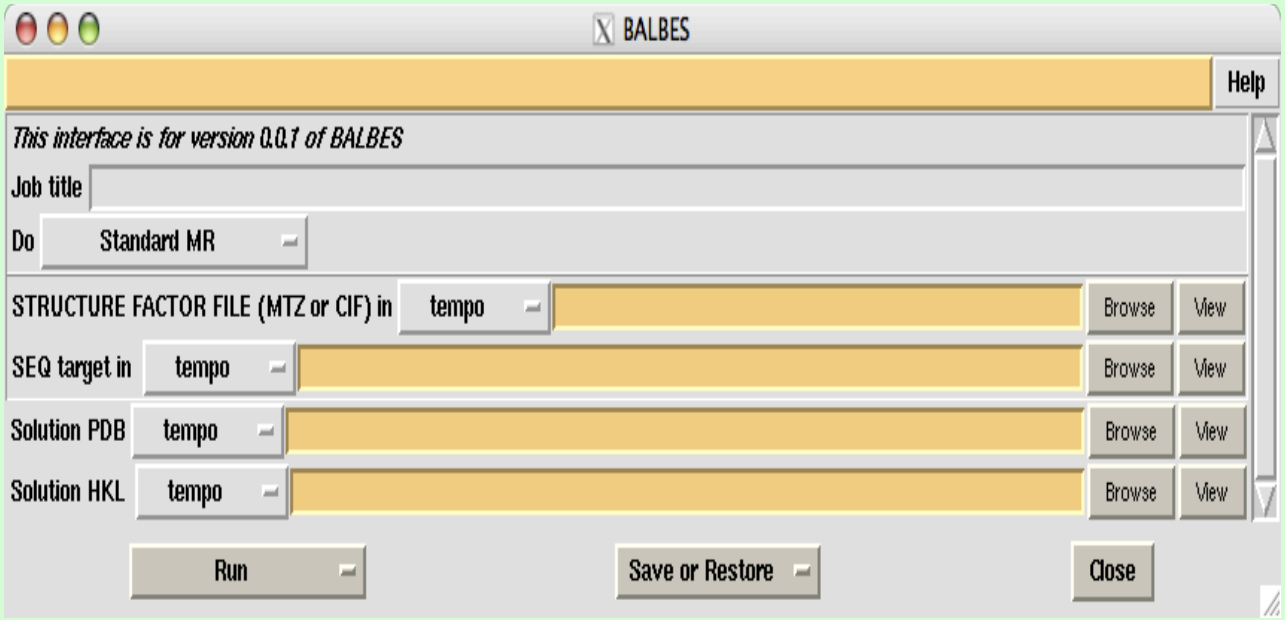
How to Run BALBES

1. Run BALBES from CCP4 on-line web server



How to Run BALBES

2. Run BALBES using CCP4i if you have installed CCP4 suite at your local computers



How to Run BALBES

3. Run BALBES from command line if you have installed CCP4 suite at your local computers (for advanced users)

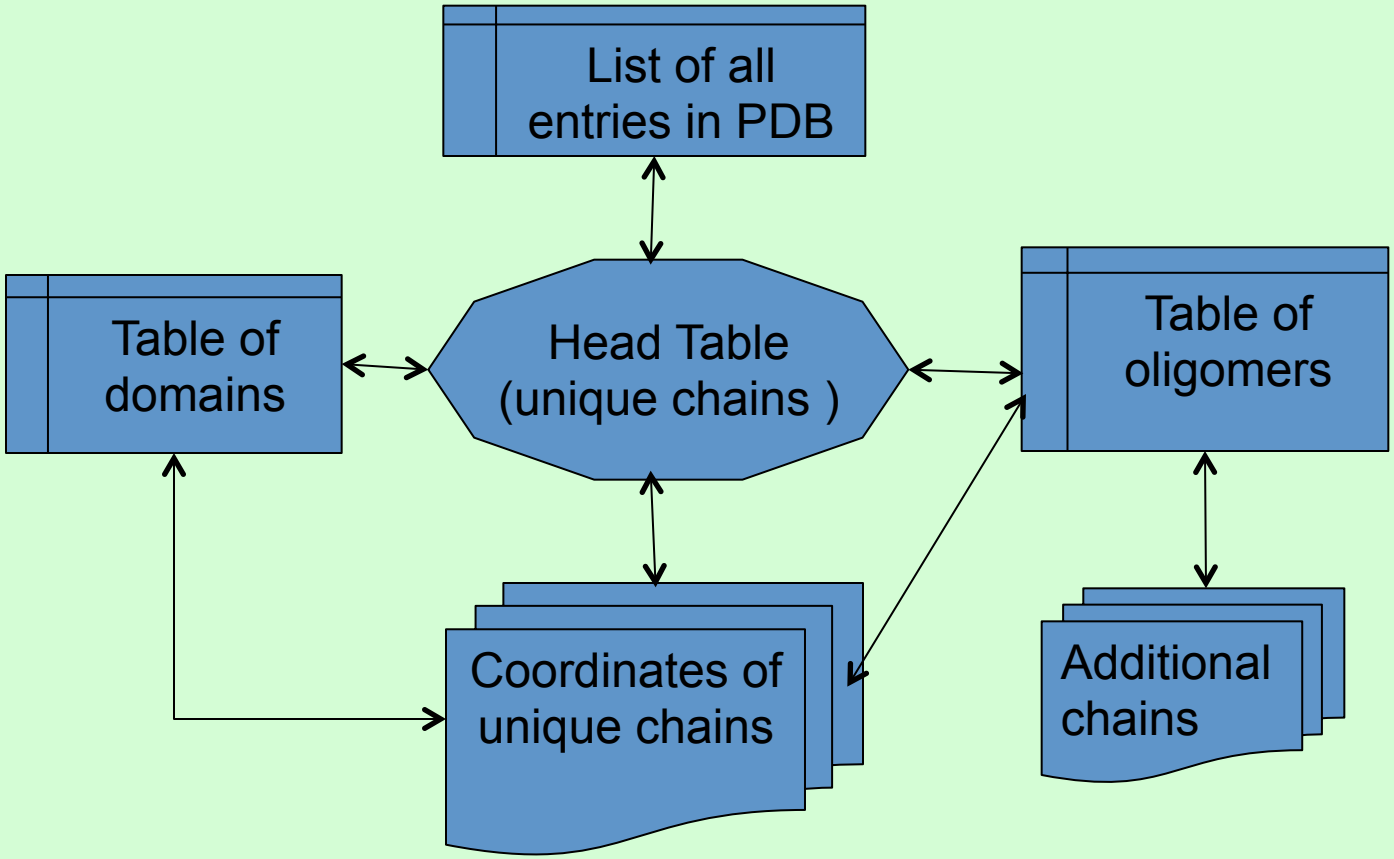
- More options
- Embedding BALBES as a component in other programs and pipelines.

balbes –f *structure_factors_file*
-s *sequence_file*
–o *output_directory*

-f required
-s required
-o optional

Internal Database in BALBES

Organization of BALBES database



Internal Database in BALBES

Some details of BALBES database

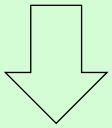
- About 30000 Non-redundant chains selected from protein structures the of high resolutions in PDB.

- More than 50000 domain definitions
  - ✓ defined by their three-dimensional compactness and separability from other parts
  - ✓ flexible parts remove.
  - ✓ hierarchically organized according to 3D similarity

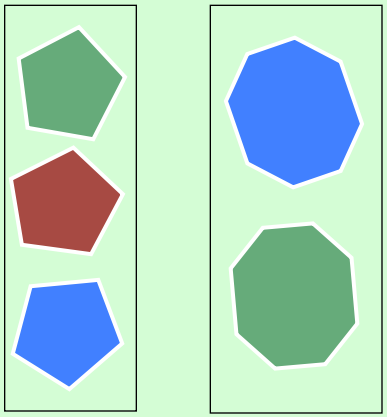
- Multimers
  - generated using PISA definition

Model preparation using BALBES DB

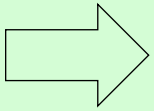
MARPVPEETVATERVKE

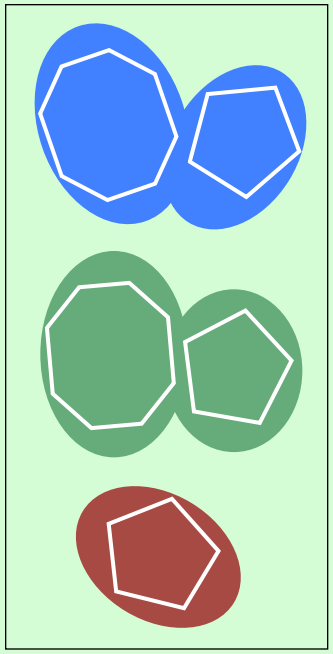


MARPVPEETVATERVKE

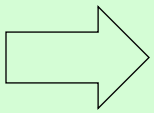


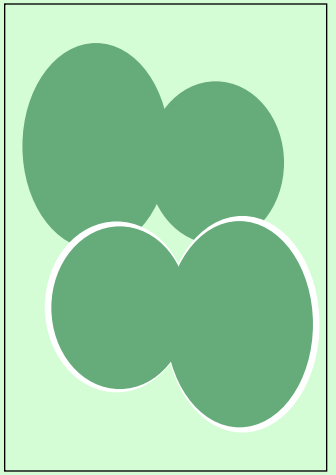
Domain models





Chain models





Multimeric models

- All models are corrected
  - by sequence alignment
  - by accessible surface area
- Several input sequences mean protein-protein complex
  - hetero-multimeric models will be generated if possible

Organization for search models

Homologue Models from one input sequence

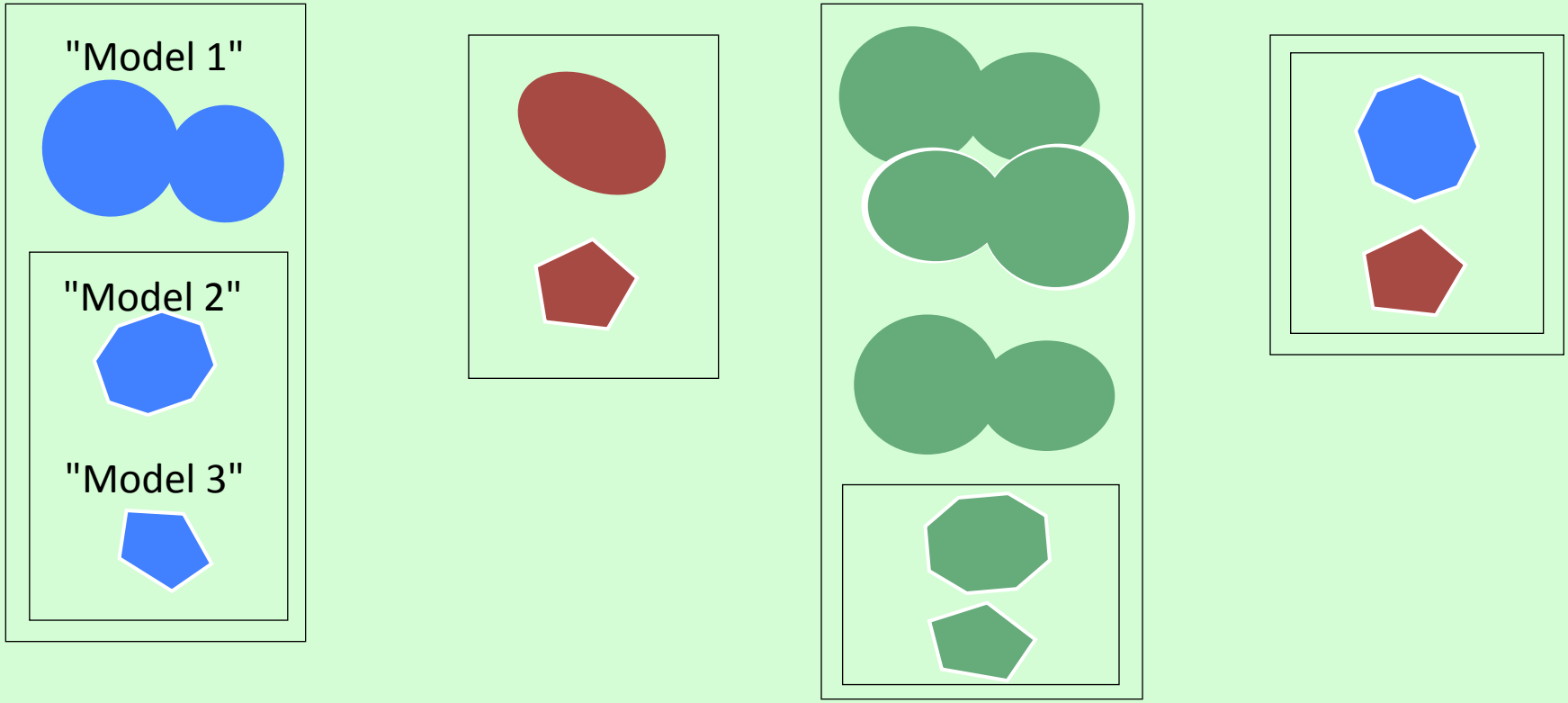
Sequence identity decreases

"Structure 1"

"Structure 2"

"Structure 3"

"Multidomain model"



Organization for search models

Homologue Models from one input sequence

Homologues

dimeric

monomeric

"multi-domain"

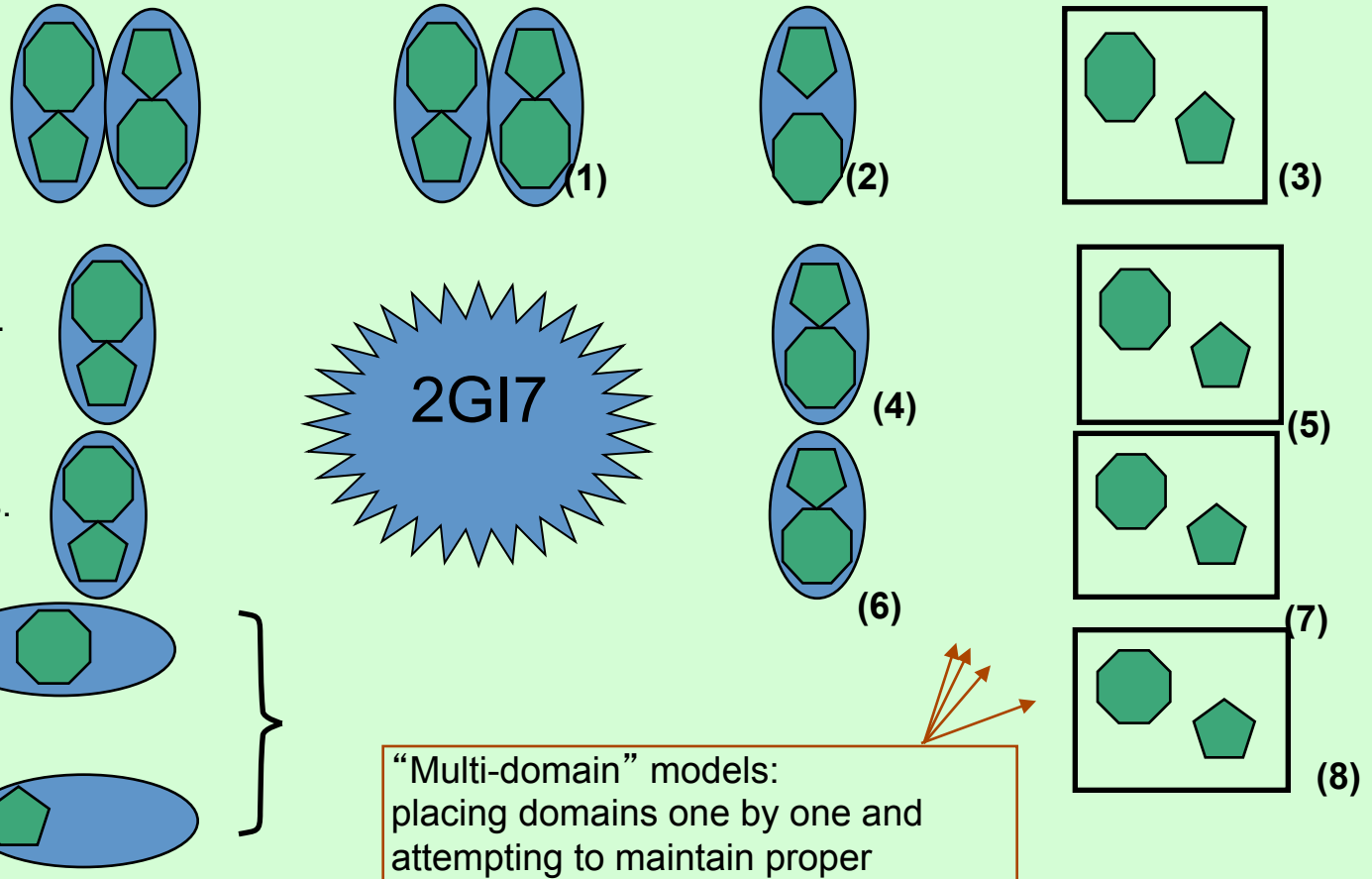
1p7q: homo-dimer; each monomers is formed by two domains.
Identity = 45%

1ufu: monomer formed by two domains.
Identity = 45%

2d3v: monomer formed by two domains.
Identity = 46%

xxxx: contains domain 1
Identity = 42%

yyyy: contains domain 2
Identity = 56%



"Multi-domain" models:
placing domains one by one and attempting to maintain proper composition of the asymmetric unit

Multi-domain model (test case)

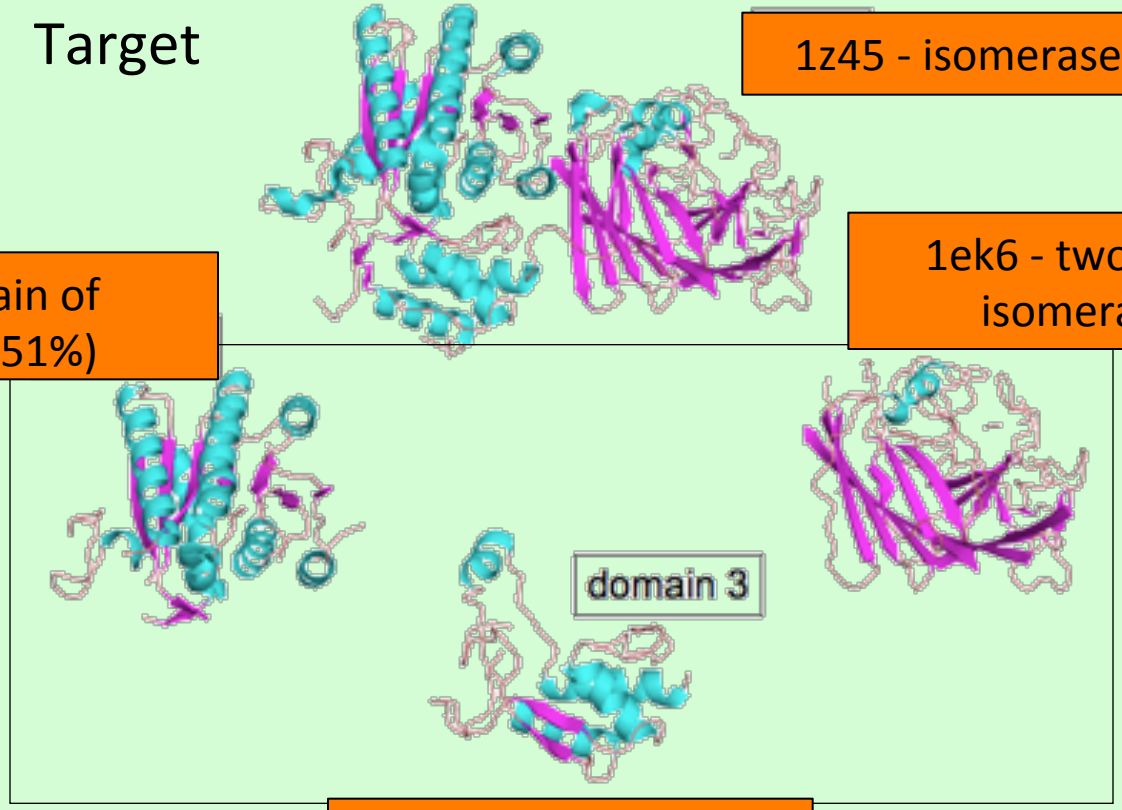
Target

1z45 - isomerase

1yga - domain of isomerase (51%)

1ek6 - two domains of isomerase (55%)

Multi-domain model



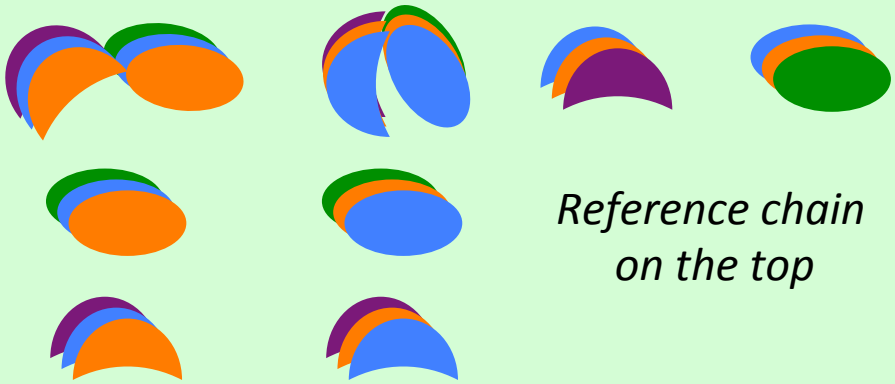
1udc - two domains of isomerase (49%)

Ensemble models

Ensemble models are generated if possible

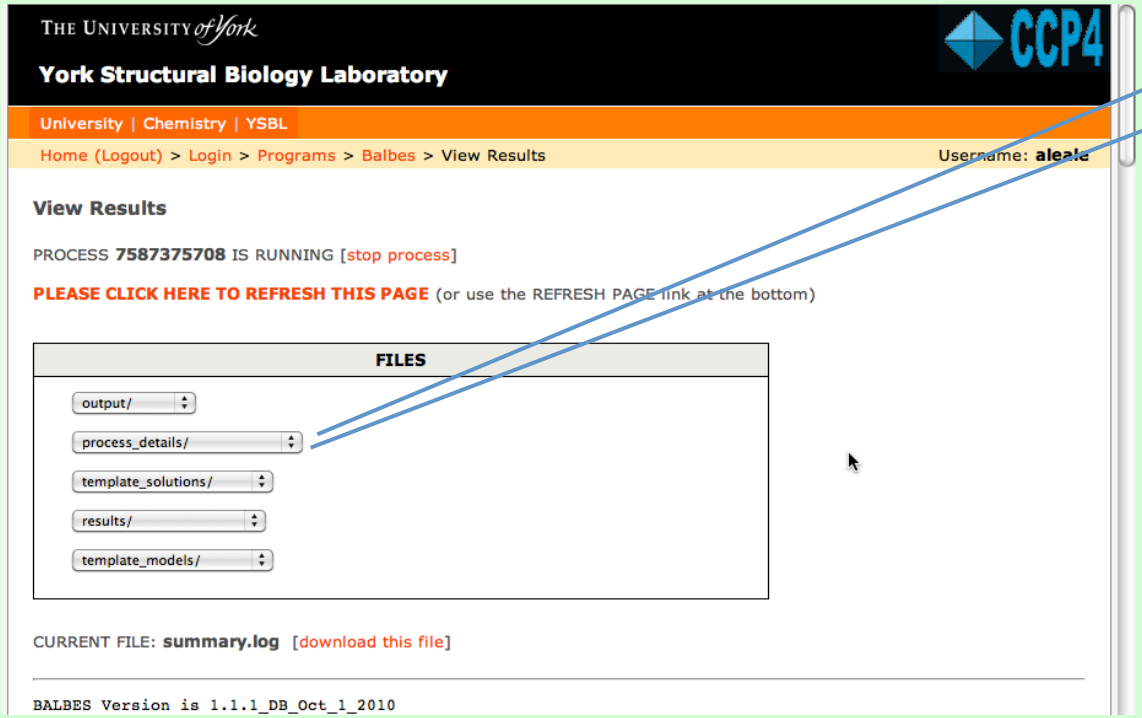
Homologues from the Balbes database:

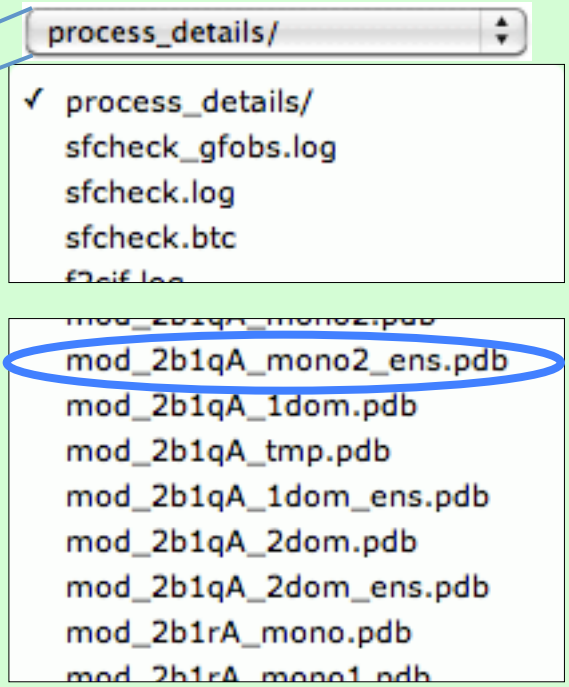
Ensemble search models:

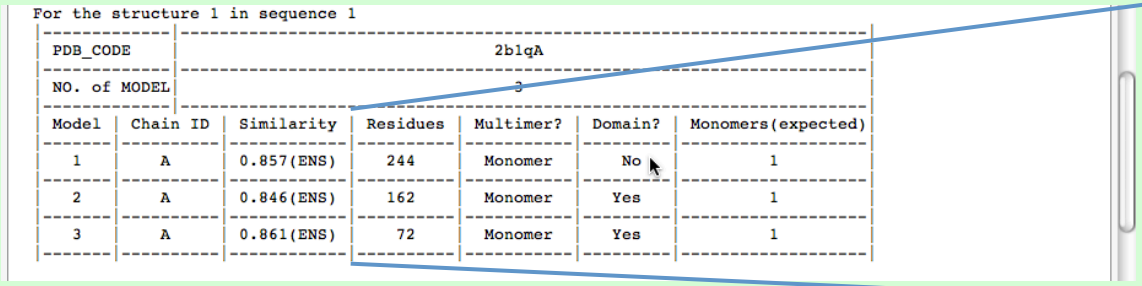


Reference chain on the top

Ensemble models from Balbes

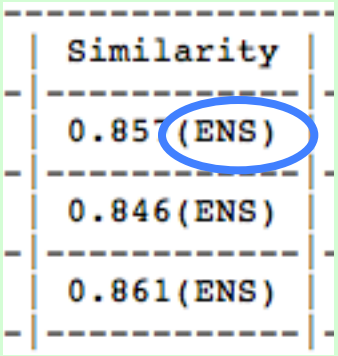






For the structure 1 in sequence 1

PDB_CODE	2b1qA					
NO. of MODEL	3					
Model	Chain ID	Similarity	Residues	Multimer?	Domain?	Monomers(expected)
1	A	0.857(ENS)	244	Monomer	No	1
2	A	0.846(ENS)	162	Monomer	Yes	1
3	A	0.861(ENS)	72	Monomer	Yes	1

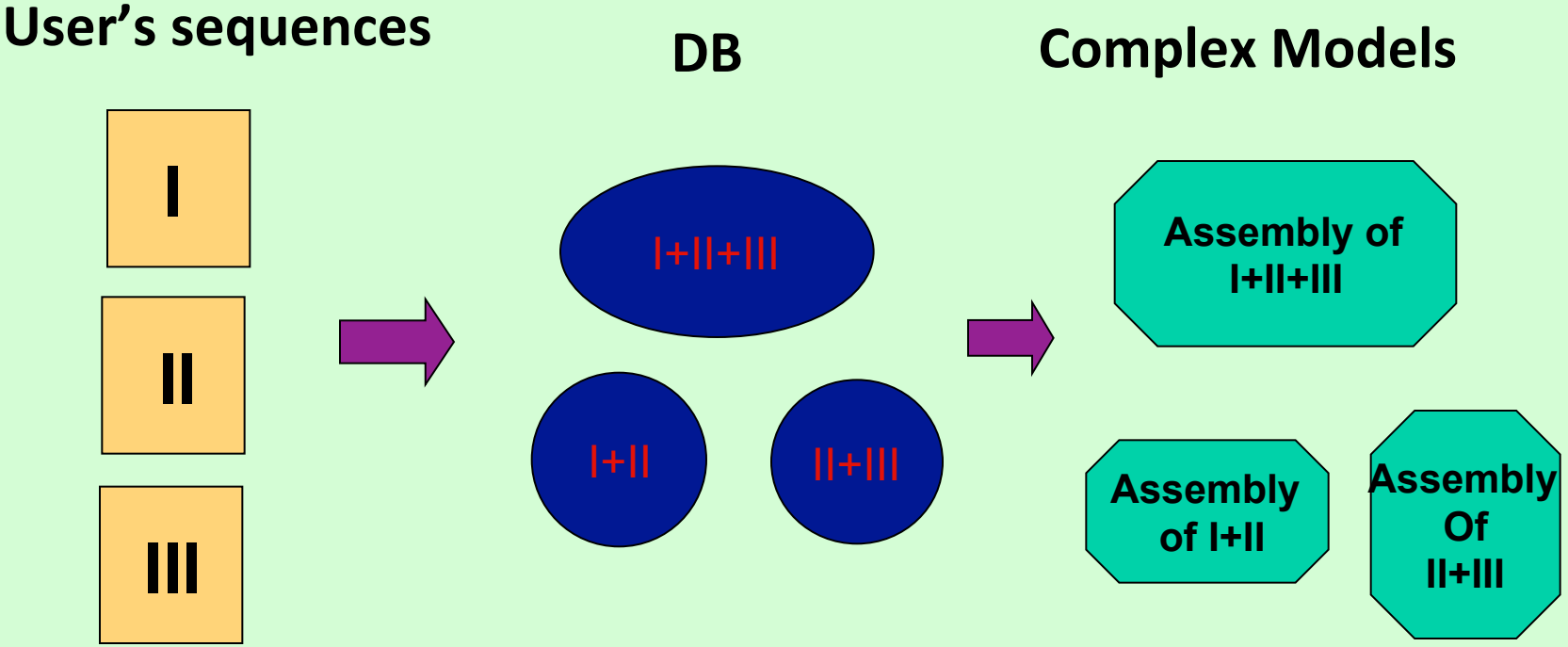


14

Organization for search models

Complex Models from multiple input sequences

BALBES will search the database for complexes of models containing all or some of the sequences if a user provides more than one sequences

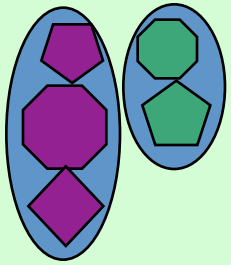


Organization for search models

Complex Models from two input sequences

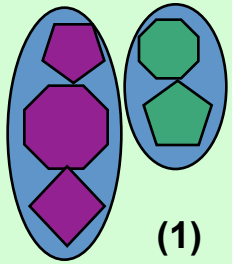
Homologues structure:

2b3t: hetero-dimer;
monomers are formed by
two and three domains.



Derived search models (and their priority):

assembly



(1)

Other homologues (1t43, 1nv8, 1zbt, 1rq0) are matching only one of two sequences. Priority rules applied to them are as in previous examples.

Note: If the system cannot find a good solution from assembly then it tries to solve using individual models and combine them. Individual models may come from different proteins.

Organization for search models

An example

THREE SEQUENCES ARE SUBMITTED

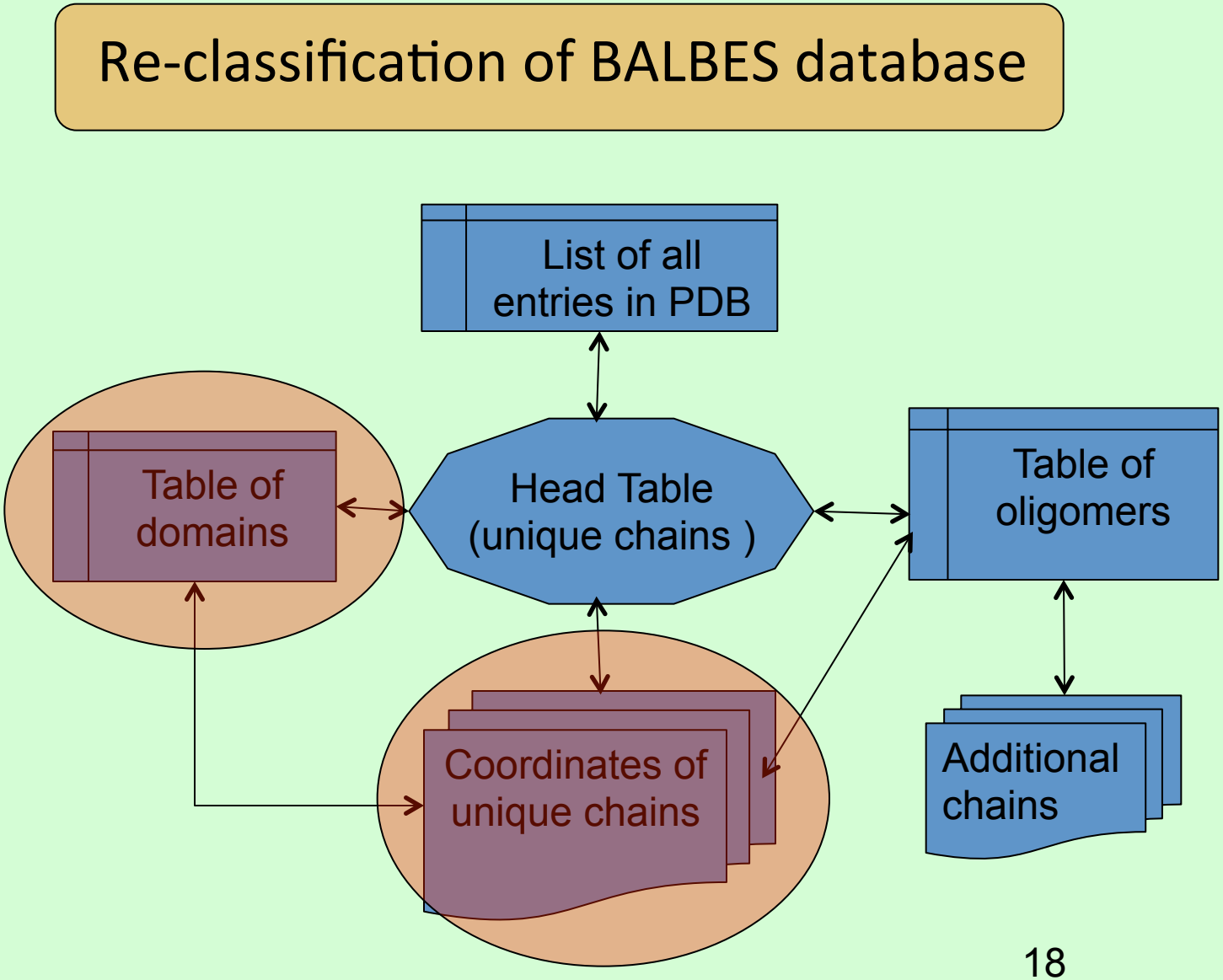
Sequence 1: 199 residues ,sequence 2: 241 residues, sequence 3: 215 residues

Assembly model 1: 2vlmD + 2vlmE
Assembly model 2: 2eyrB + 2eyrA

Sequence 1--> models: 2vlmD, 1ogaD, 2cdfA, 3mv7D, 2wbjC
Sequence 2 --> models: 1ogaE, 2axhA, 3kxfP, 2po6D, 2eyrB
Sequence 3 --> models: 1enfA, 1esfA, 1fnuA, 2g9hD, 3ea6A

Final Results	
Models	1ogaE+1enfA+1ogaD
R_ini/R_fin	0.388/0.250
Rfree_ini/Rfree_fin	0.404/0.310

Improvements for the database



Acknowledgements

BALBES Development:

Alexey Vagin University of York

Garib Murshudov MRC LMB

Web server support:

Ville Uski STFC RAL

Ronan Keegan STFC RAL

Charles Ballard STFC RAL

PISA:

Eugene Krissinel STFC RAL

ARP/wARP development team