βαλβεσ

A Molecular Replacement Pipeline

Garib Murshudov

Chemistry Department, University of York

Contents

1) Introduction
2) Organisation of BALBES
3) Search model preparation
4) Updating BALBES
5) Warnings: Twin
6) Conclusions

Introduction

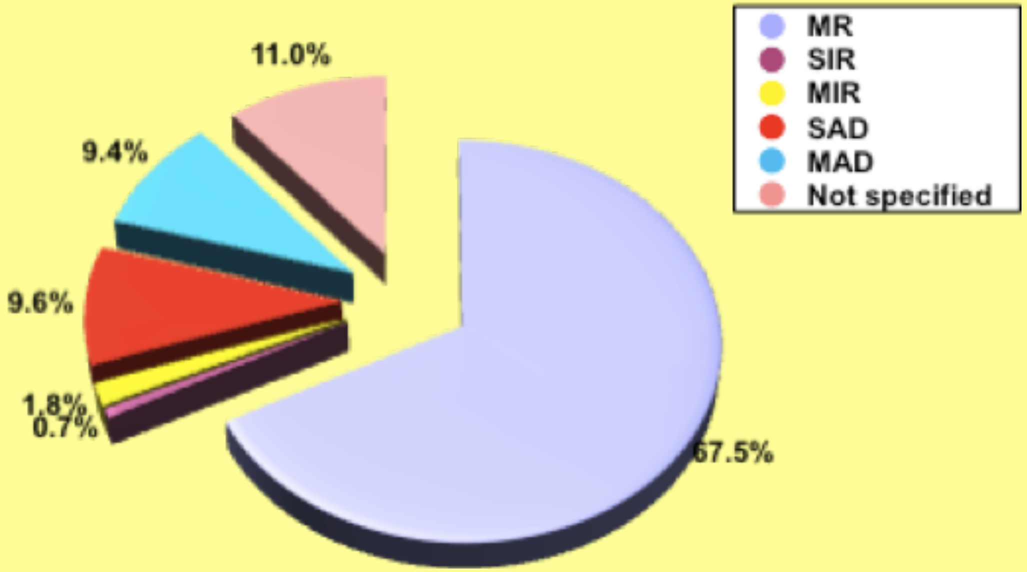


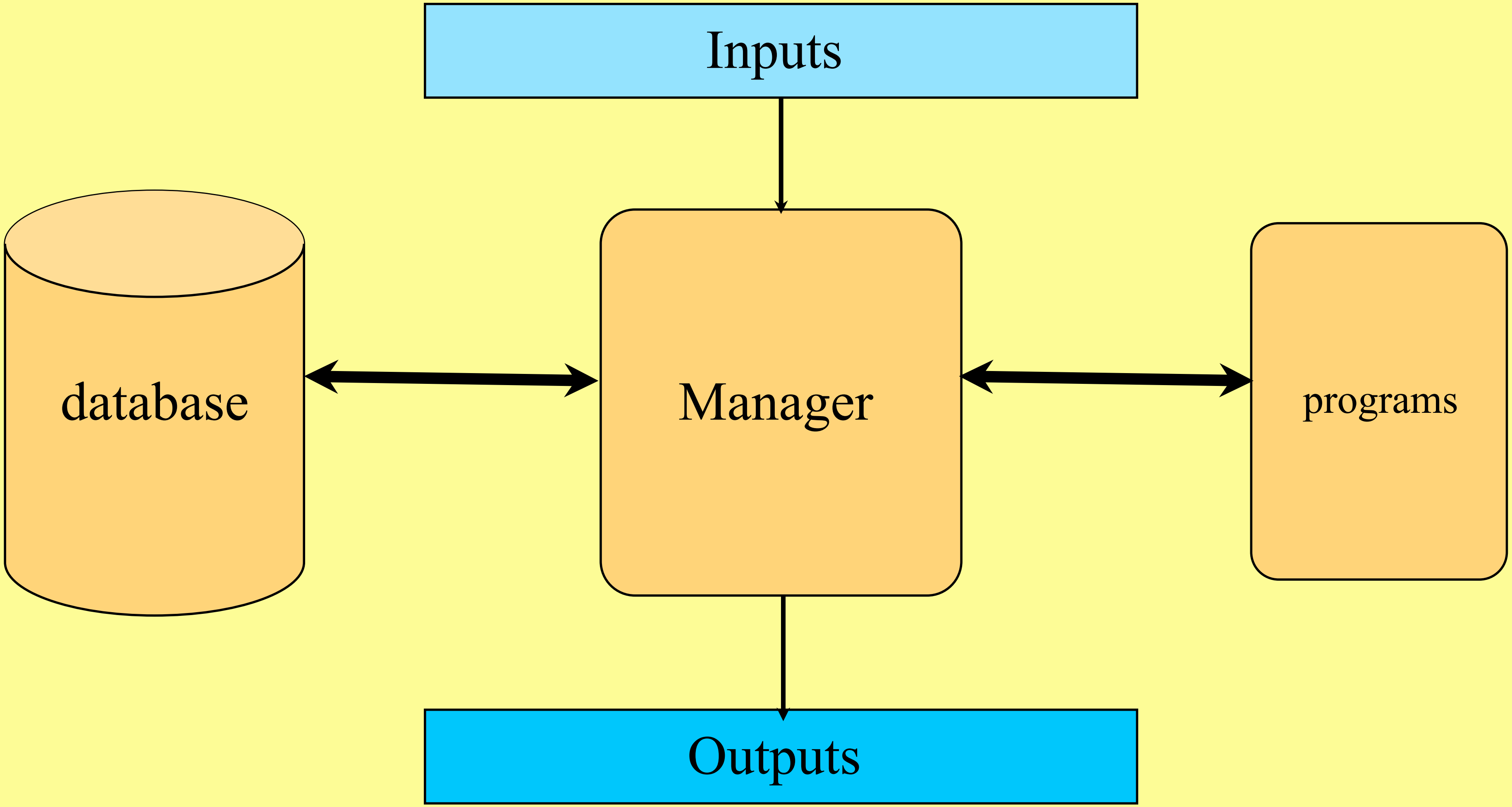
Diagram showing the percentage of structures in the PDB solved by different techniques

67.5% of structures are solved by Molecular Replacement (MR)

21% of structures are solved by experimental phasing

Organisation of BALBES

BALBES consists of three essential components

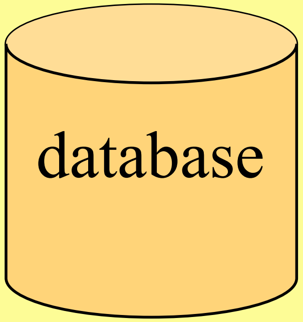


Manager

It is written using PYTHON and relies on files of XML format for information exchange:

1. Data
   - Resolution for molecular replacement
   - Data completeness and other properties
   - Twinning
   - Pseudo translation
2. Sequence
   - Finds template structures with their domain and multimer organisations
   - Estimates number of molecules in the asymmetric unit
   - “Corrects” template molecules using sequence alignment
3. Protocols
   - Runs various protocols with molecular replacement and refinement and makes decisions accordingly

Database



Chains . The internal database has around 35000 unique entries selected from more than 51,000 present in the PDB. All entries in the PDB are analysed according to their identity. Only non-redundant sets of structures are stored.

Domains. The DB contains 35000 domain definitions Loops and other flexible parts are removed from the domain definitions.

Multimers of structures (using PISA)

Hierarchy is organized according to sequence identity and 3D similarity (rmsd over Ca atoms).

Programs

programs

MOLREP - molecular replacement

Simple molecular replacement, phased rotation function (PRF), phased translation function (PTF), spherically averaged phased translation function (SAPTF), multi-copy search, search with fixed partial model

REFMAC

Maximum likelihood refinement, phased refinement, twin refinement, rigid body refinement, handling ligand dictionary, map coefficients

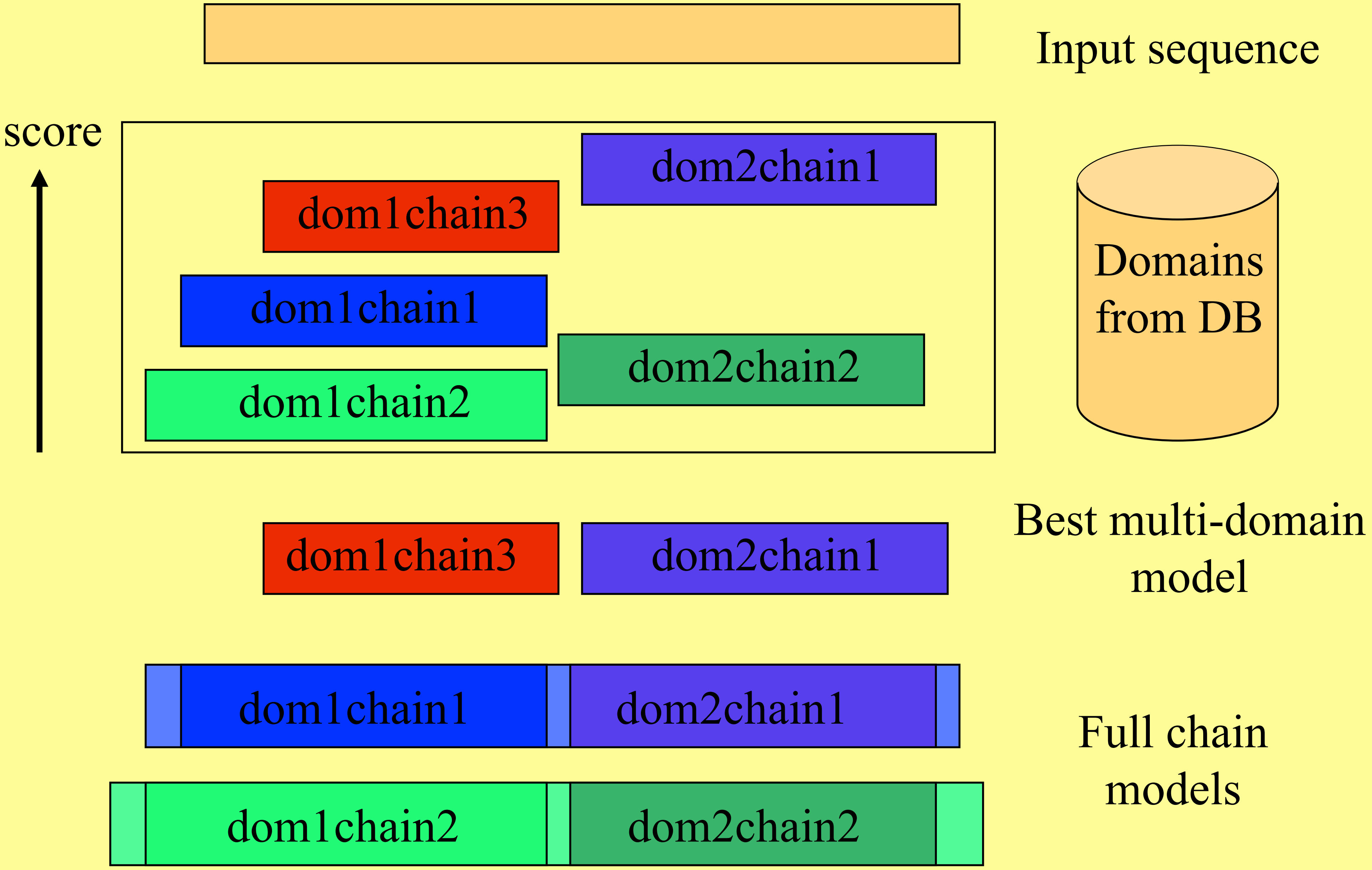
SFCHECK

Optical resolution, optimal resolution for molecular replacement, analysis of coordinates against electron density, twinning tests, pseudo translation

Other programs:

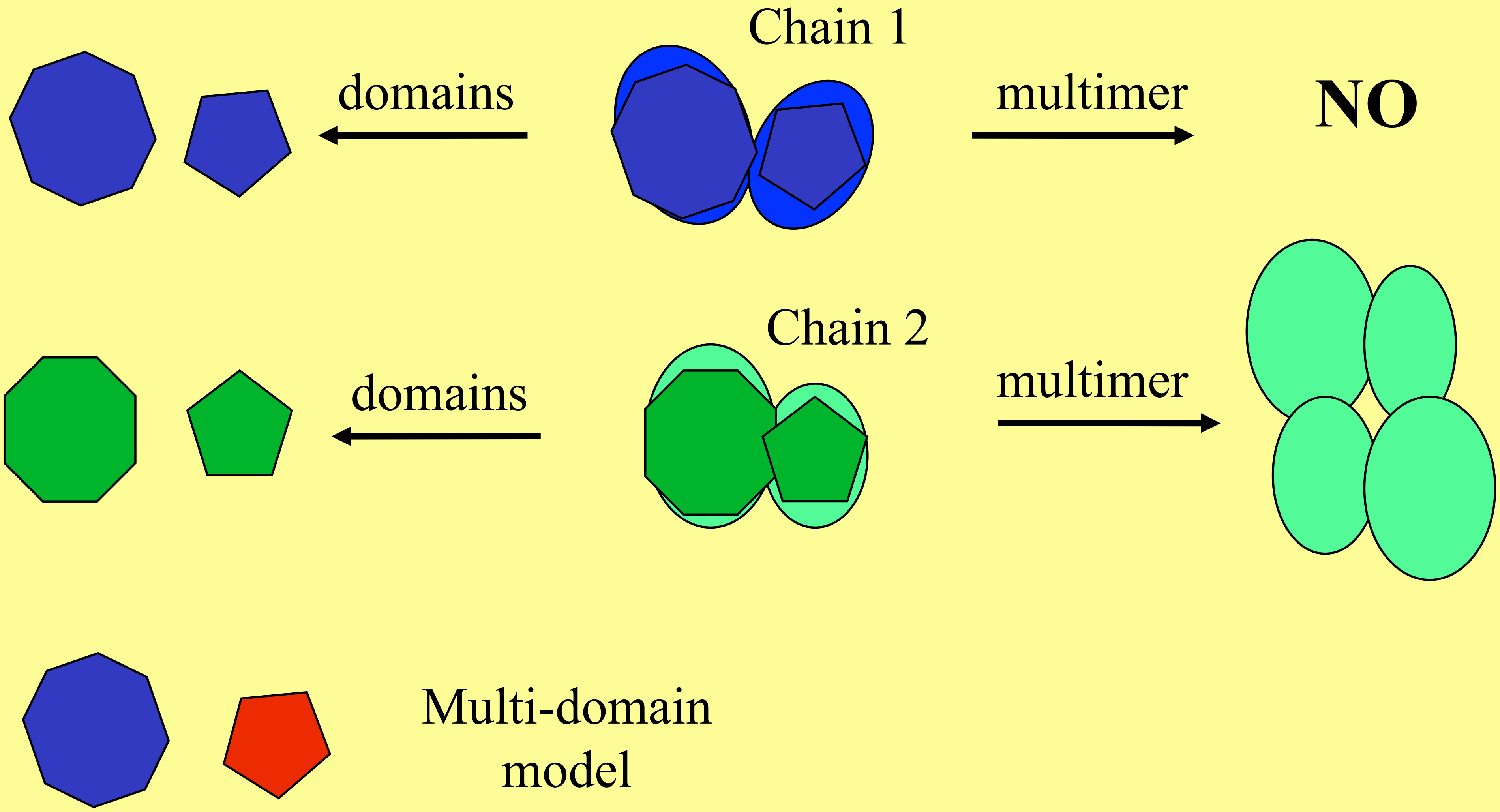
Alignment, search in DB, analysis of sequence and data to suggest number of expected monomers, semiautomatic domain definition

Search models



Model preparation

All models are corrected by sequence alignment and by accessible surface area



Chain 1

domains

multimer

NO

Chain 2

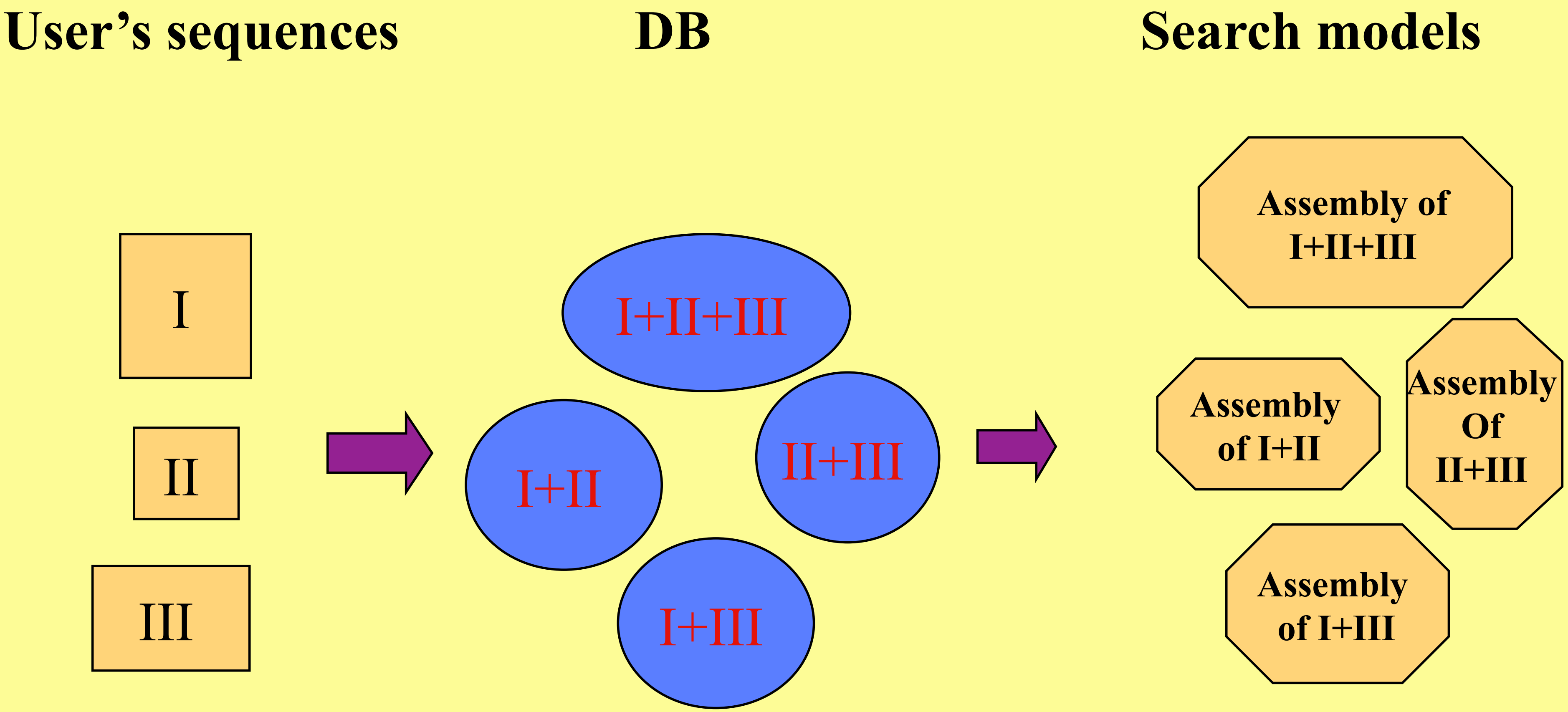
domains

multimer

Multi-domain model

Heterogeneous Search Models

If a user provide several sequences, BALBES will search the database for complexes of models containing all or most of the sequences.

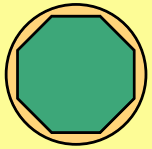


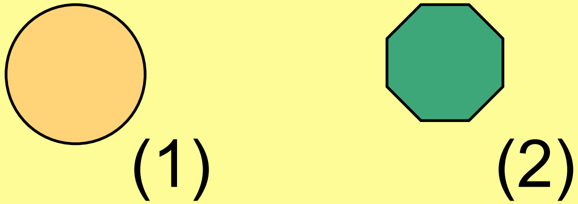
Example 1: 2dwr

Homologues

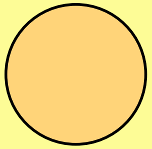
Derived search models (and their priority)

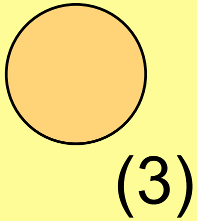
2aen: monomer and one domain definition associated with it.
Identity = 82%



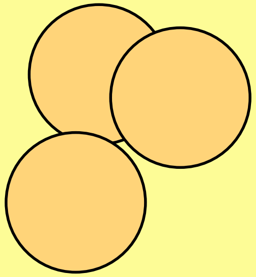


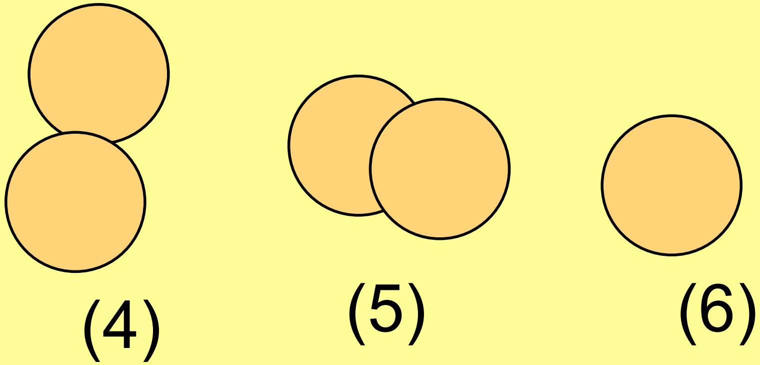
1kqr: monomer, no domain definitions
Identity = 45%





1z0m: dimer, no domain definitions
Identity = 25%





Example 3: 2gi7 Derived search models (and their priority)

Homologues

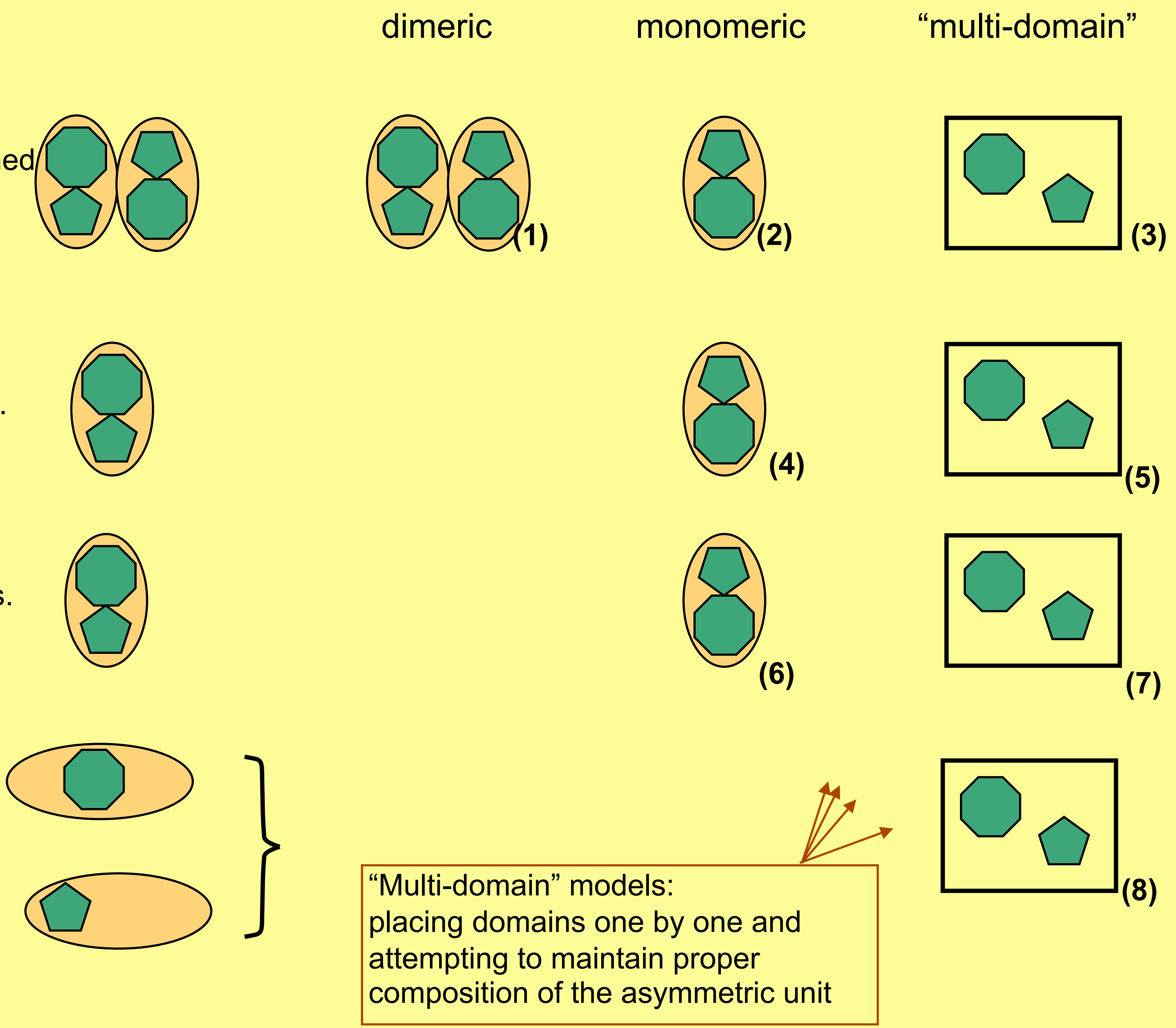
1p7q: homo-dimer;
each monomers is formed
by two domains.
Identity = 45%

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

Example 4: assembly (two sequences are submitted)

Assembly models

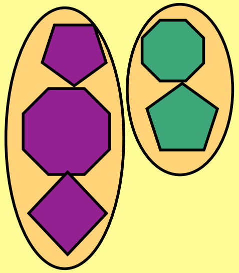
In case when two or more sequences are submitted attempt will be made to find hetero-oligomer matching all or some of these sequences.

If found, such hetero-oligomers will be first models to try.

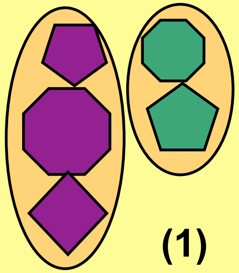
Homologues structure:

Derived search models (and their priority):

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



assembly



(1)

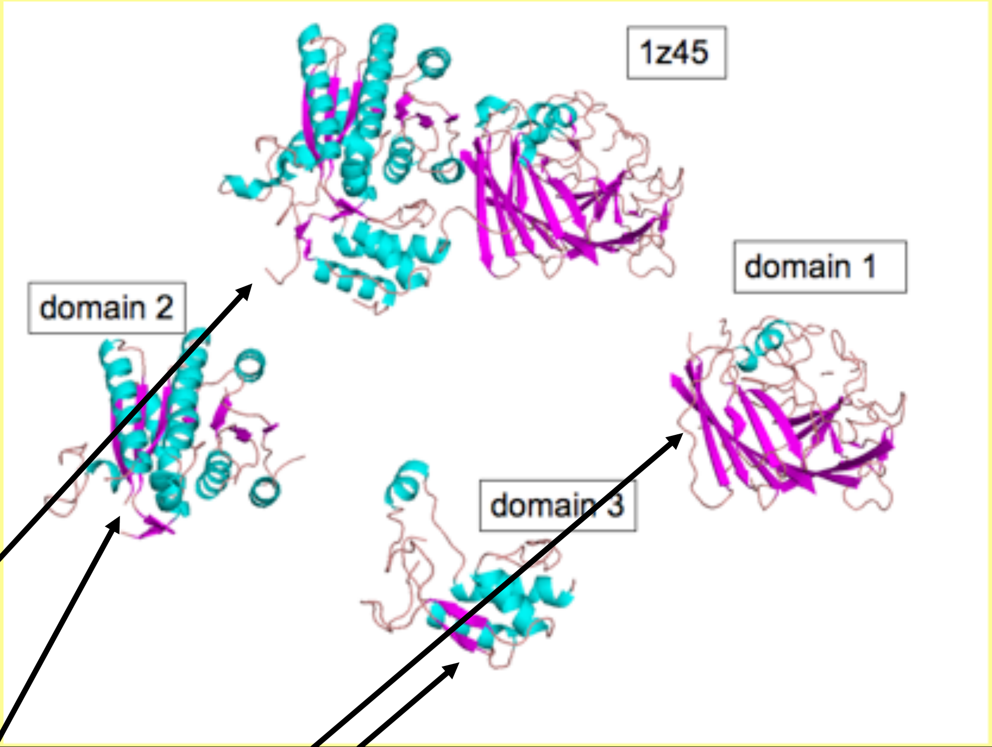
Other homologues (1t43, 1nv8, 1zbt, 1rq0) are matching only one of two sequences. Priority rules applied to them are as in previous examples.

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

Example of search: Multi-domain protein

This structure can be solved with multi-domain model.

PDB entry 1z45 has three major domains. One of the domains has also two subdomains. Domain 1 is similar to 1ek6 (seq id 55%). Domain 2 similar to 1yga (seq id 51%) and domain 3 is similar to 1udc (seq id 49%)



1z45 - isomerase
1ek6 - two domains of isomerase
1yga - another domain of isomerase
1udc - two domains of isomerase

All these proteins are although isomerases they have slightly different activities

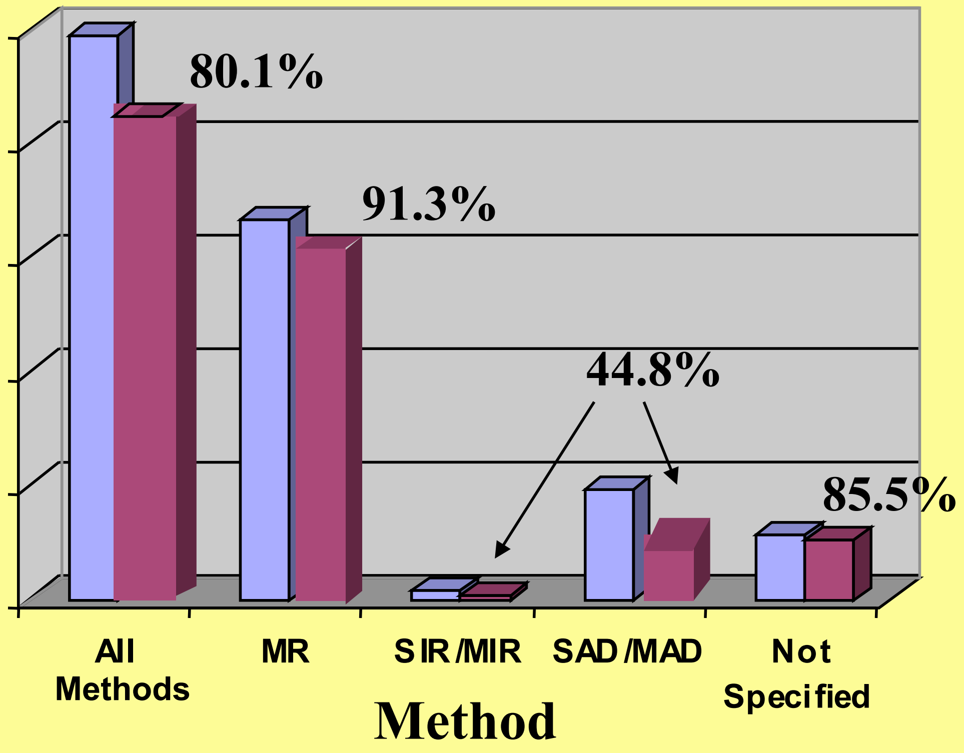
Updating and Calibrating the System

All structures newly deposited to the PDB are tested against the old internal database by using BALBES. Only after that the DB is updated.

Updating and tests are carried out every half a month.

automatically generated domains are checked manually
to make sure that automatic domain-definition transfer
does not introduce errors.

The success rate of the tests (Jan - Feb 2008)



N structures = 950

Blue: the number of structures originally solved by a given method

Magenta: the number of structures BALBES was able to solve

Note: the fraction of structures solved by MR = 67%
The success rate of our latest tests was more than 80%

Note that some of the structures solved by experimental phasing could be actually solved by MR!

Space group uncertainty

Balbes can check space group assumption. In this case it will do calculation in parallel for all potential space groups and at the end make decision. For example for if you give P222 then the program will test

$P222$, $P2_122$, $P22_12$, $P222_1$, $P2_12_12$, $P2_122_1$, $P22_12_1$, $P2_12_12_1$

Current version does not change the point group.

How to run BALBES:

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 and a sequence file)

Running BALBES from the command line:

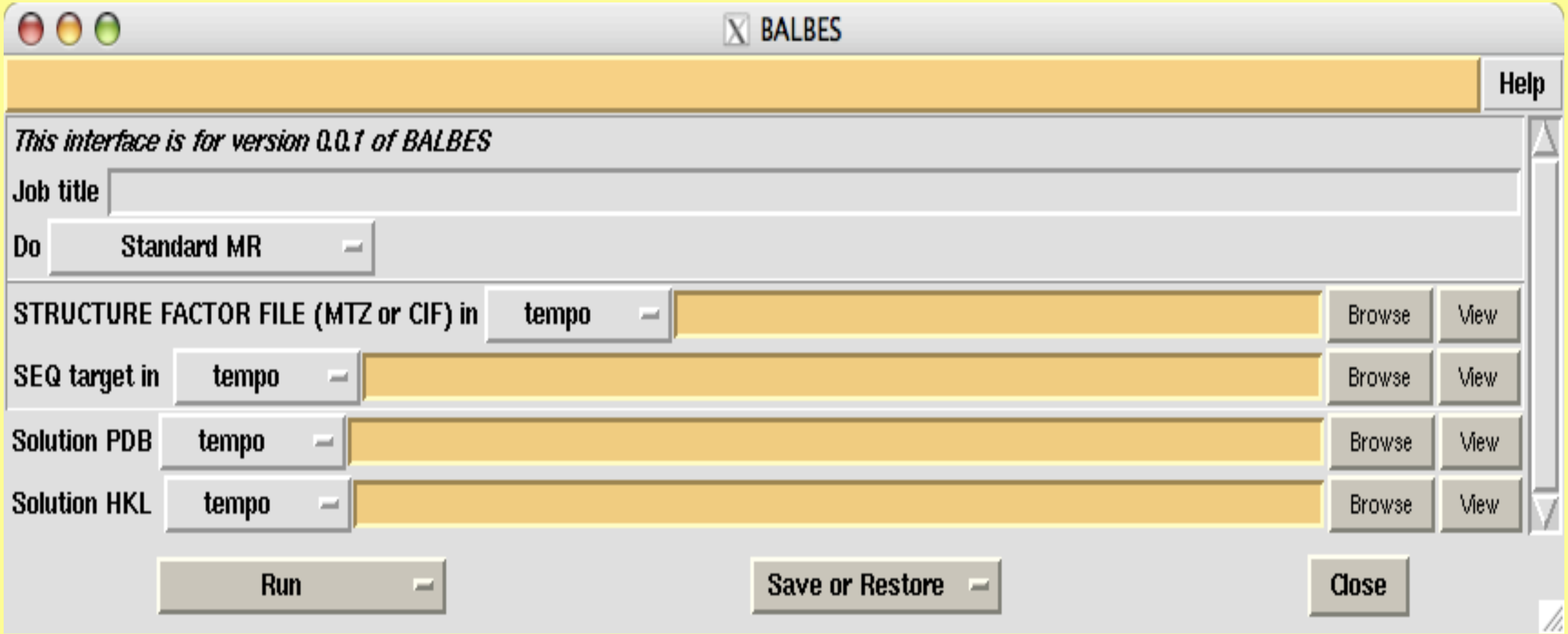
```
balbes –f structure_factors_file -s sequence_file –o output_directory
```

-f required

-s required

-o optional

BALBES CCP4i interface



BALBES

This interface is for version 0.0.1 of BALBES

Job title

Do Standard MR

STRUCTURE FACTOR FILE (MTZ or CIF) in tempo

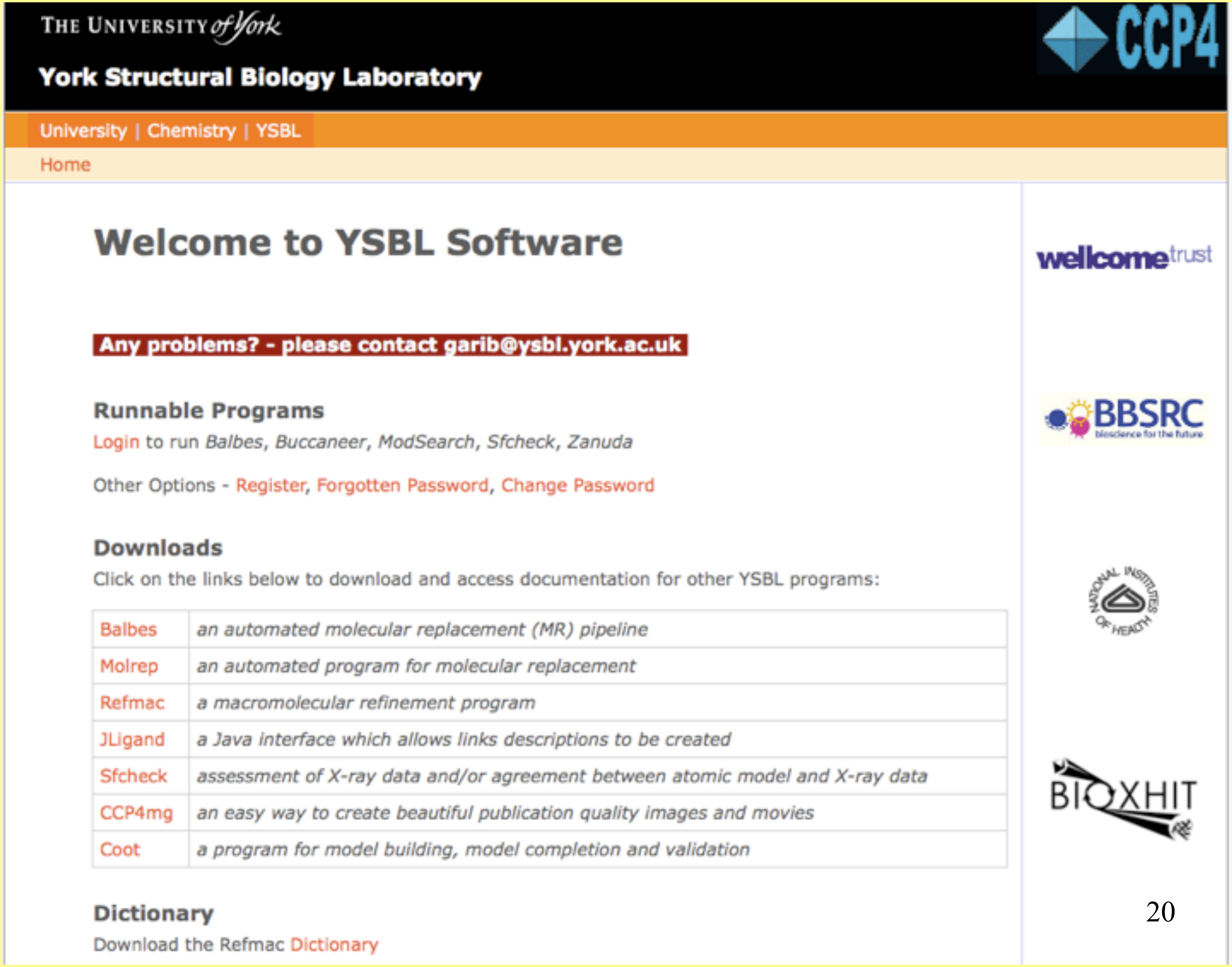
SEQ target in tempo

Solution PDB tempo

Solution HKL tempo

BALBES Interface in Our Web Server

(running using our Linux cluster) designed by P.Young



THE UNIVERSITY of York

York Structural Biology Laboratory

Welcome to YSBL Software

Any problems? - please contact garib@ysbl.york.ac.uk

Runnable Programs

Login to run *Balbes, Buccaneer, ModSearch, Sfcheck, Zanuda*

Other Options - Register, Forgotten Password, Change Password

Downloads

Click on the links below to download and access documentation for other YSBL programs:

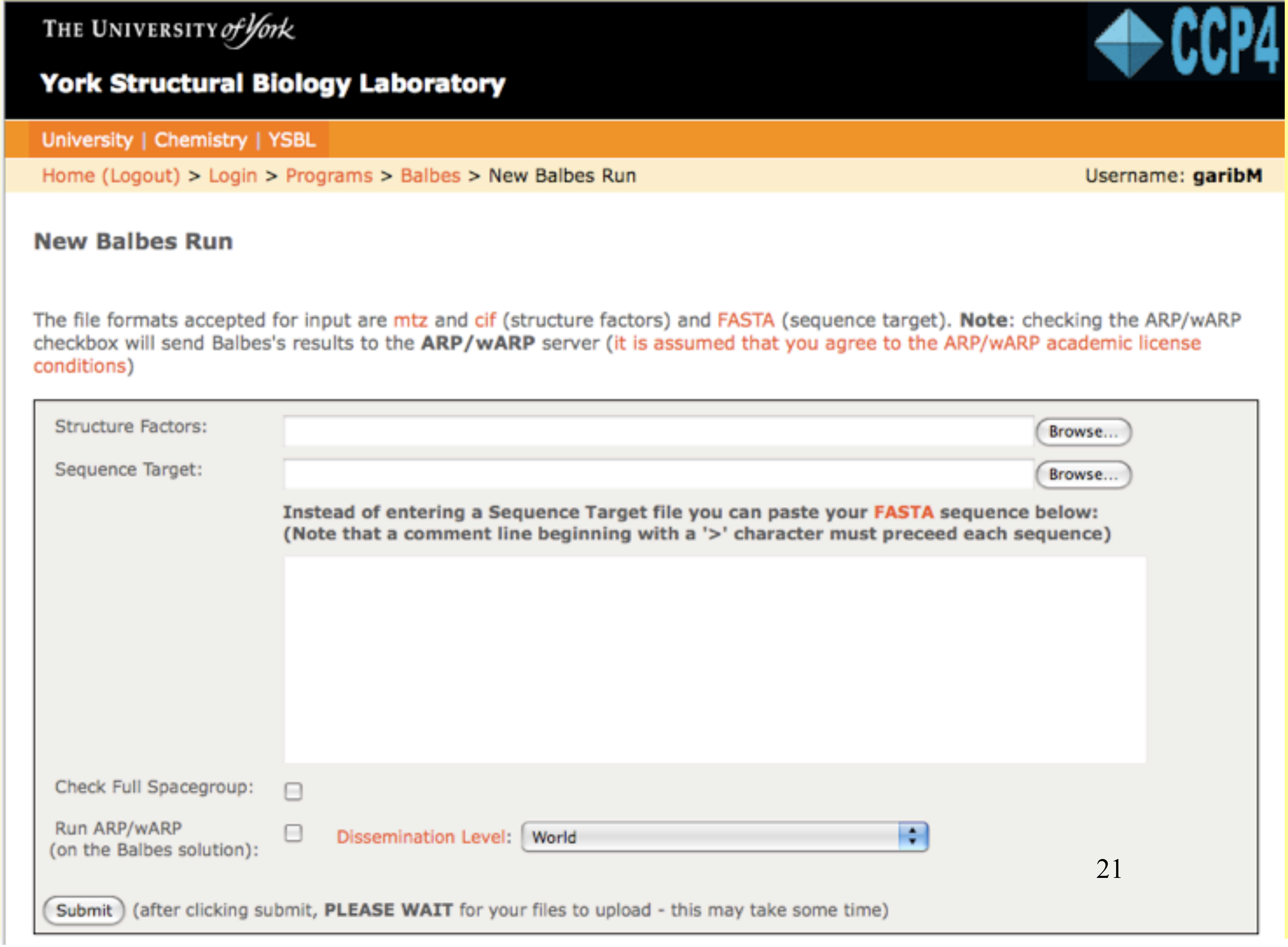
Balbes	*an automated molecular replacement (MR) pipeline*
Molrep	*an automated program for molecular replacement*
Refmac	*a macromolecular refinement program*
JLigand	*a Java interface which allows links descriptions to be created*
Sfcheck	*assessment of X-ray data and/or agreement between atomic model and X-ray data*
CCP4mg	*an easy way to create beautiful publication quality images and movies*
Coot	*a program for model building, model completion and validation*

Dictionary

Download the Refmac Dictionary

BALBES Interface in Our Web Server

(running using our Linux cluster) designed by P.Young



THE UNIVERSITY *of York*

York Structural Biology Laboratory

CCP4

University | Chemistry | YSBL

Home (Logout) > Login > Programs > Balbes > New Balbes Run

Username: **garibM**

New Balbes Run

The file formats accepted for input are mtz and cif (structure factors) and FASTA (sequence target). **Note**: checking the ARP/wARP checkbox will send Balbes's results to the **ARP/wARP** server (it is assumed that you agree to the ARP/wARP academic license conditions)

Structure Factors: [Browse...]

Sequence Target: [Browse...]

Instead of entering a Sequence Target file you can paste your FASTA sequence below:
(Note that a comment line beginning with a '>' character must preceed each sequence)

Check Full Spacegroup: □

Run ARP/wARP (on the Balbes solution): □ Dissemination Level: World

[Submit] (after clicking submit, **PLEASE WAIT** for your files to upload - this may take some time)

Complexes

In cases of complexes (more than one sequence) the system first tries assemblies (if available). If it can find good solution it stops. If it cannot find solution then it switches to individual sequence (with and without ensembles). For each sequence best solution is stored. The best among the best is fixed and program continues to search for the second, the third etc proteins. Again with and without ensembles.

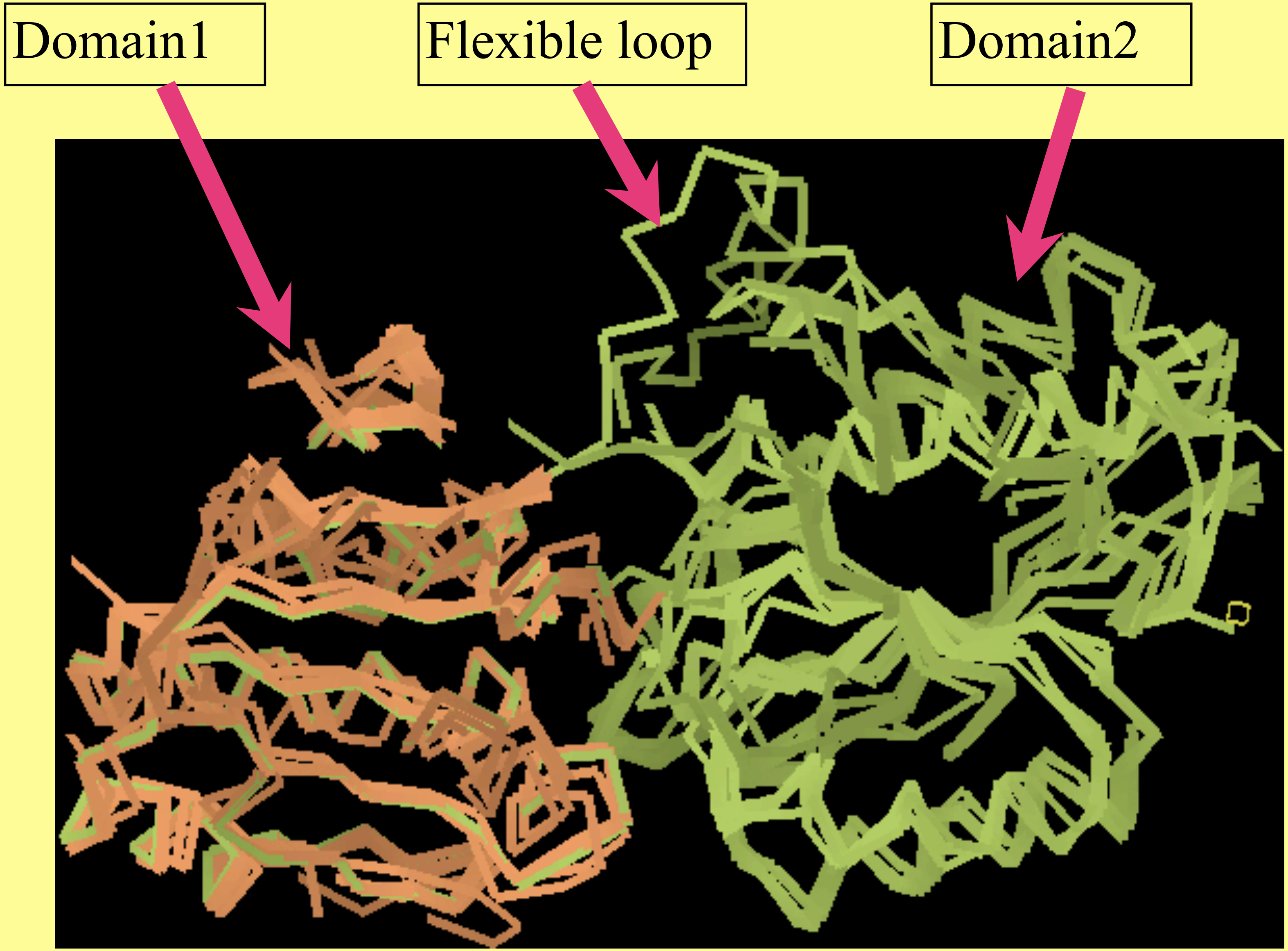
Moreover if space group is uncertain then the program will do all calculation for each potential space group candidate. Decision about space group is made at the very end of all runs (It may take some time).

Ensembles

In the new version the program first identifies domains for each sequence using alignment. Then for each domain it creates ensemble of molecules using internal domain database. Then using profile of sequence generated from these ensembles it realigns sequences to improve reliability.

Then for each ensemble it tries molecular replacement and refinement. Then takes the best “solution”, fixes it and tries to find more. When the score cannot be improved or maximum number of molecules expected is reached the program stops and gives (hopefully) solution with it quality factor.

Ensembles: Two domain example

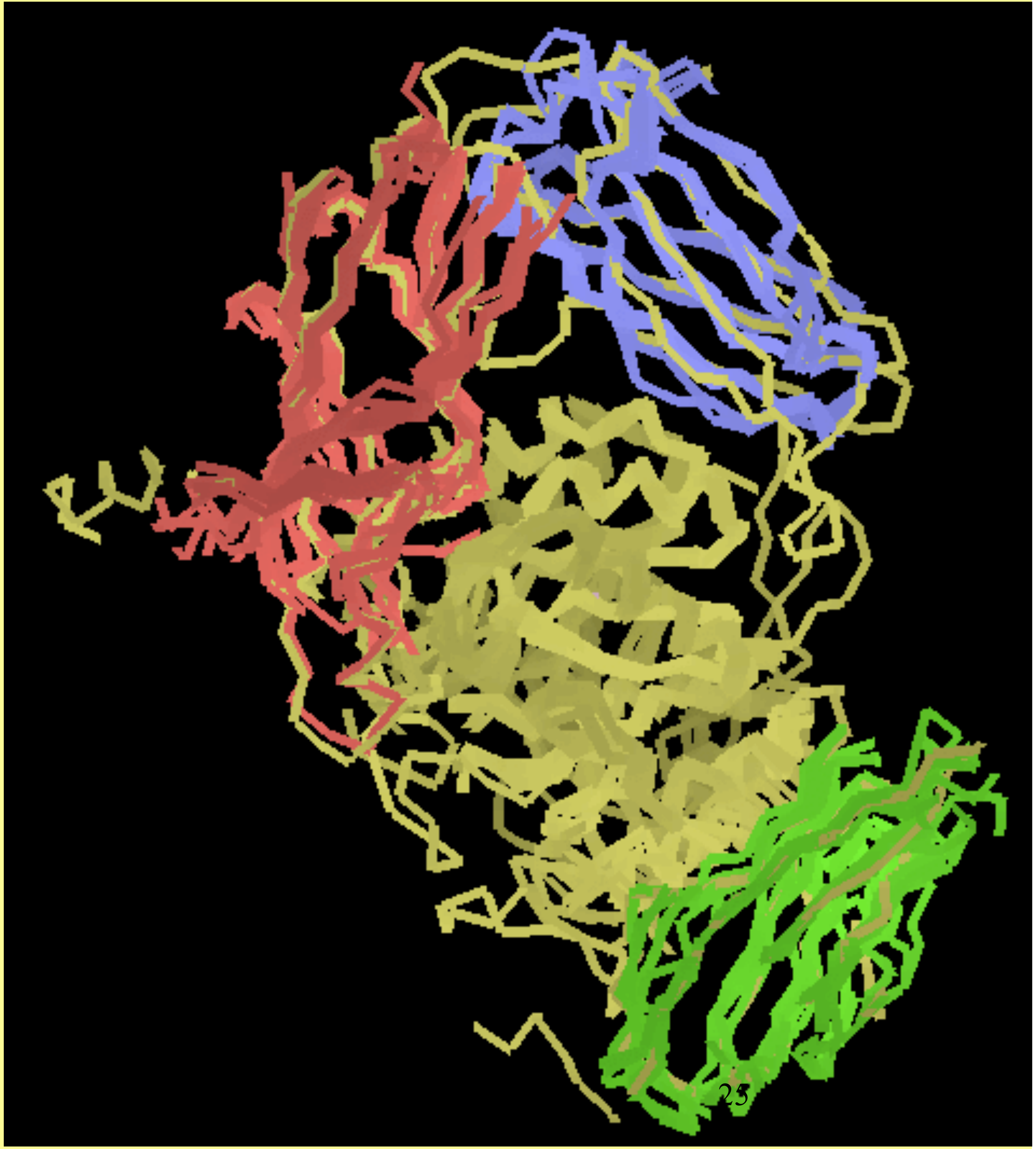


Domain1 and domain2 are used for MR. Flexible loops are not used if they are too small

Ensembles: Four domain example

Four domain protein with different domains. For each domain there are number of similar structures taken from BALBES's domain database.

During MR ensemble for each domain is tried and then solutions are combined to give final solution.



Refinement stage

Final decisions are made based on R-factors after refinement. Since we have similar structures we can use them in refinement. In the next version it will be added.

In refinement stage “jelly-body” refinement is used. It seems to increase success rate, especially for multidomain cases.

Future version will use more extensive search of space groups and decision on space group will be made after refinement.

Be careful: twinning

- Usually when R/Rfree are well below 50% then the structure is solved.
- When twin is present then it is no longer true. Twinning changes statistical properties of the data
- Best way of checking potential solution: refine and rebuild (arp/warp or buccaneer or coot) – if you can rebuild then everything is fine

Conclusions

1. Internal database is an essential ingredient of efficient automation
2. With relatively simple protocols, BALBES is able to solve around 80% of structures automatically
3. Interplay of different protocols is very promising
4. Huge number of tests help to prioritise developments and generate ideas
5. When there is twinning or other peculiarities then R/Rfree may not be reliable

People involved (YSBL, York)

Alexei Vagin
Fei Long
Paul Young
Andrey Lebedev

Acknowledgements

E.Krissinel for PISA MSD/PDBe, Cambridge

All CCP4 and YSBL people for support

ARP/wARP development team

Wellcome Trust, BBSRC, EU BIOXHIT, NIH for support

The site to download BALBES:

http://www.ysbl.york.ac.uk/~fei/balbes/

Webserver:

http://www.ysbl.york.ac.uk/YSBLPrograms/index.jsp

This and other talks:

http://www.ysbl.york.ac.uk/refmac/presentations/