

# A Molecular Replacement Pipeline

Garib Murshudov

Chemistry Department, University of York

# Introduction

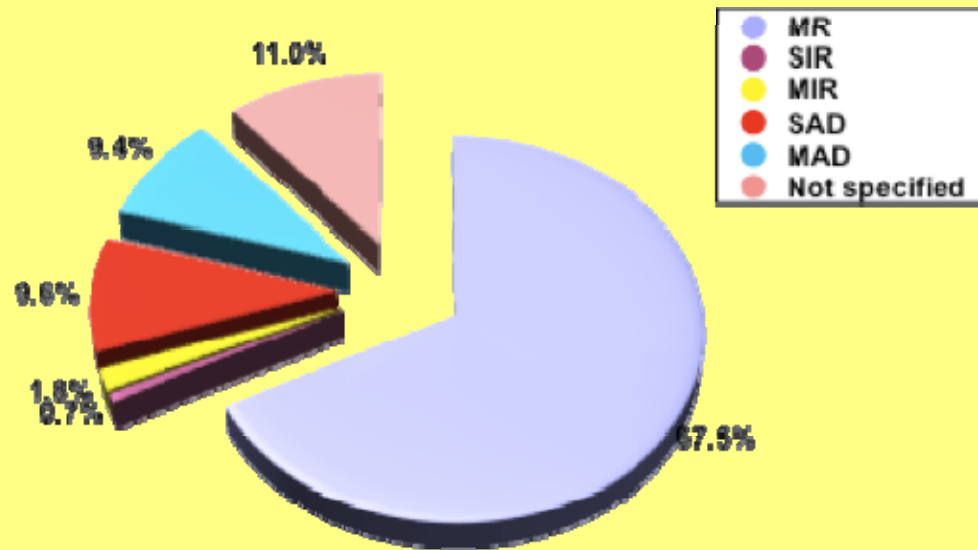


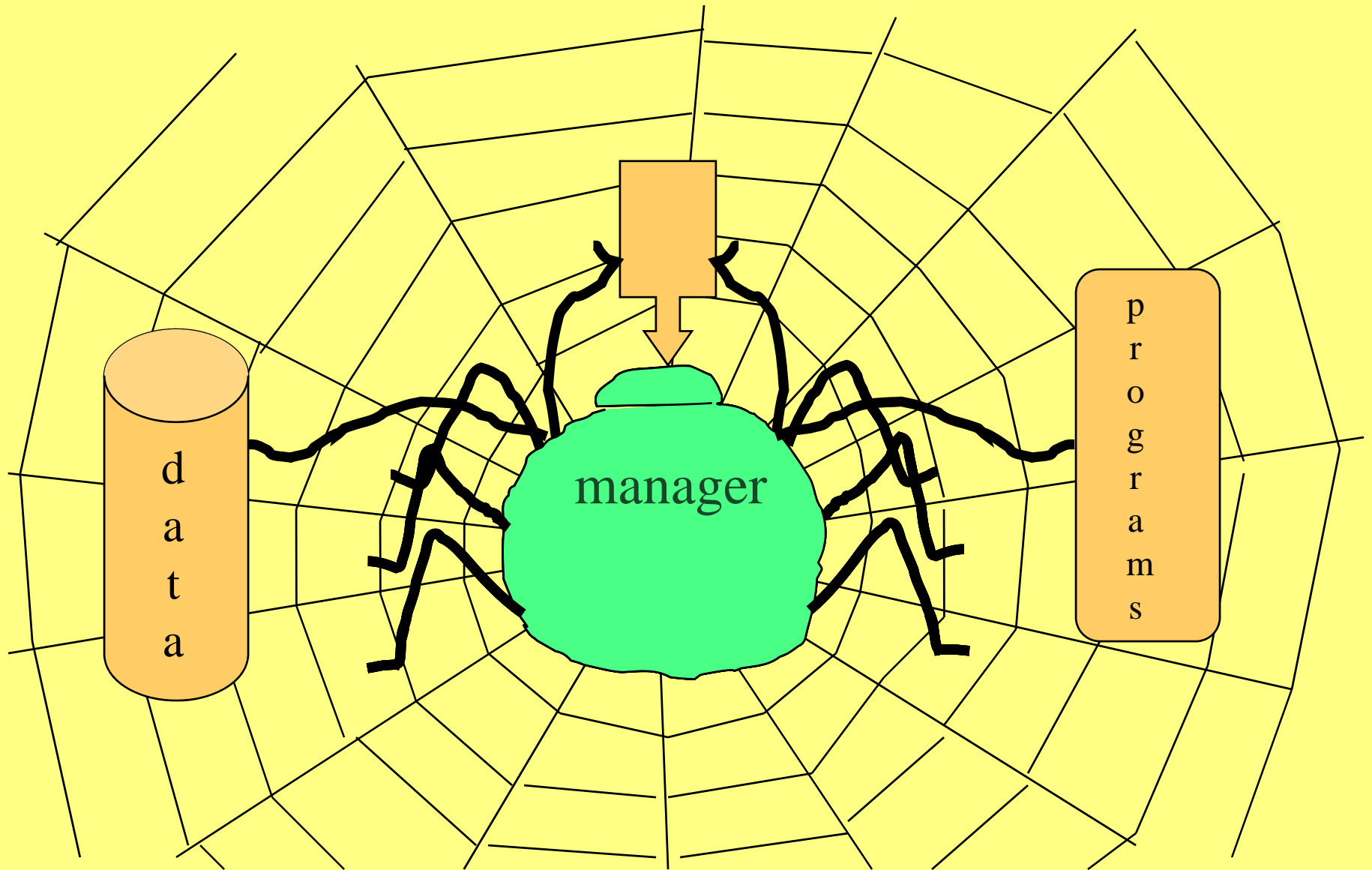
Diagram showing the percentage of structures in the PDB solved by different techniques

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

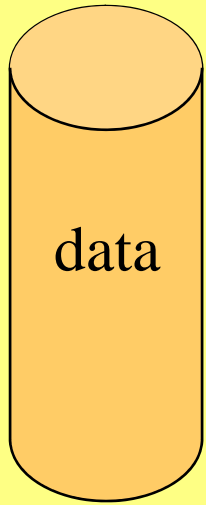
- Finds template structures with their domain and multimer organisations
- Estimates number of molecules in the asymmetric unit
- “Corrects” template molecules using sequence alignment

## 3. Protocols

- Runs various protocols with molecular replacement and refinement and makes decisions accordingly

# Manager

makes decisions on different protocols according to search models available in the internal database.



# Database

**Chains** . The internal database has around 20000 unique entries selected from more than 45,000 present in the PDB. All entries in the PDB are analysed according to their identity. Only non-redundant sets of structures are stored.

**Domains**. About 8000 of them have two or more domains. The DB contains 30000 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

# Molrep

- Molrep is highly automated molecular replacement program. With given data, pdb and sequence (if available) it tries to solve molecular replacement problems.
- It automatically decides amount of molecule you expect in the asymmetric unit
- It uses packing function to remove false solutions
- It does anisotropic correction when it is necessary before starting molecular rotation and translation search
- It can handle pseudo-translations
- It can search small fragment in the electron density
- It can do multi-copy search for cases when there are several domains or several subunits.
- It can do fit of Xray model to EM and EM to Xray using full point group constraints

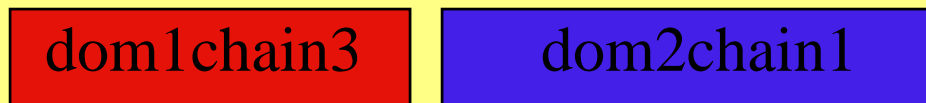
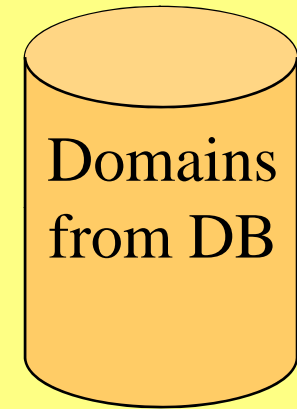
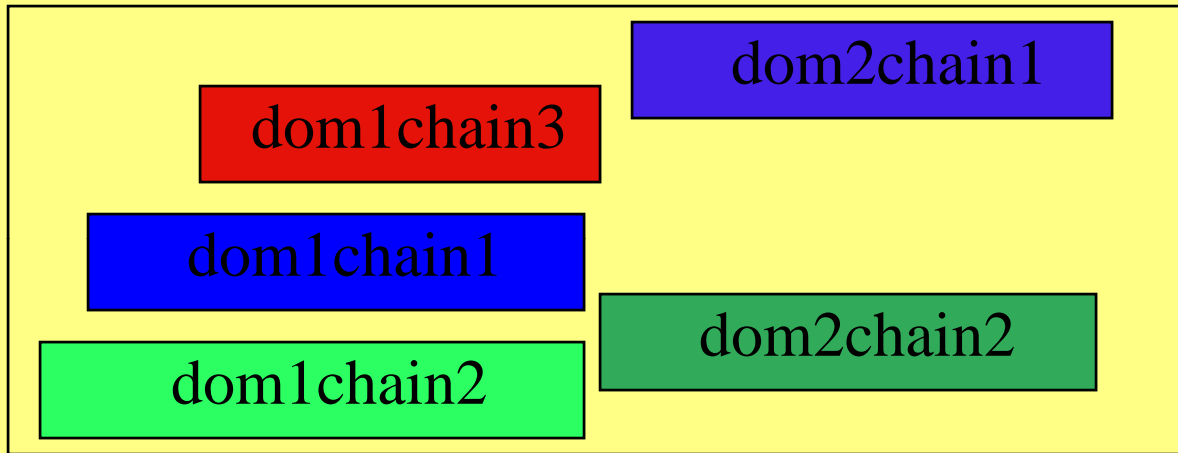
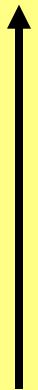


# Search models



Input sequence

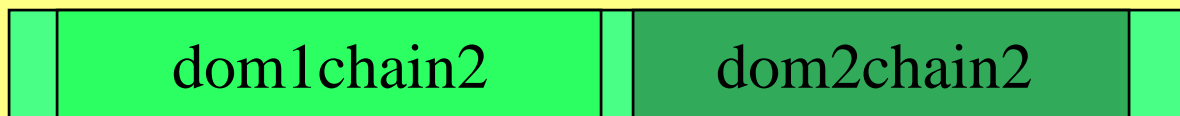
score



Best multi-domain model

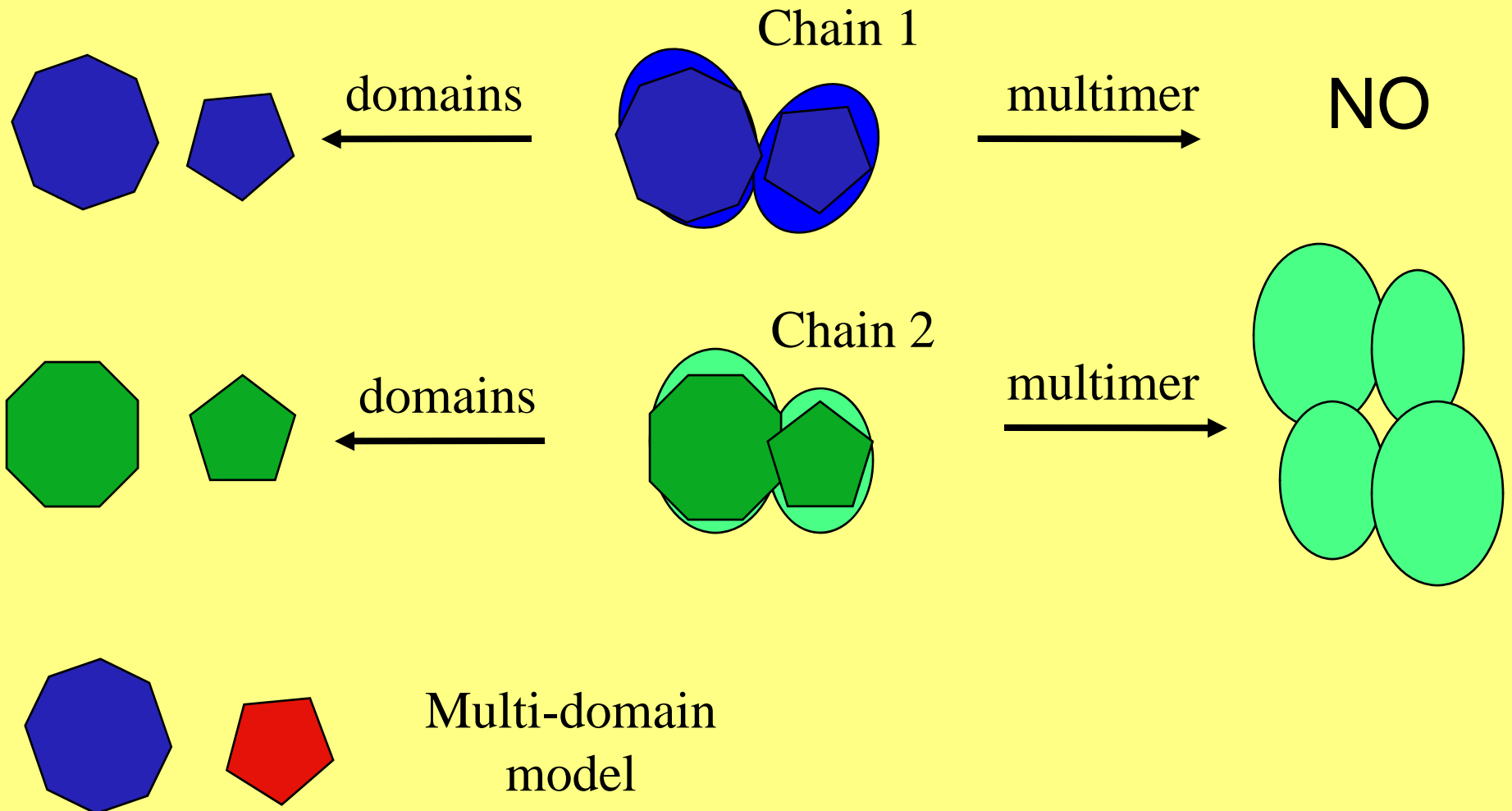


Full chain models



# Model preparation

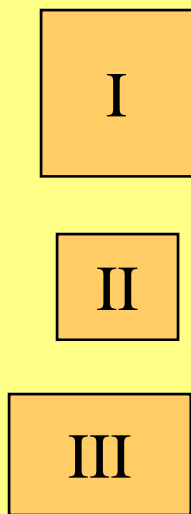
All models are corrected by sequence alignment  
and by atomic accessible surface area



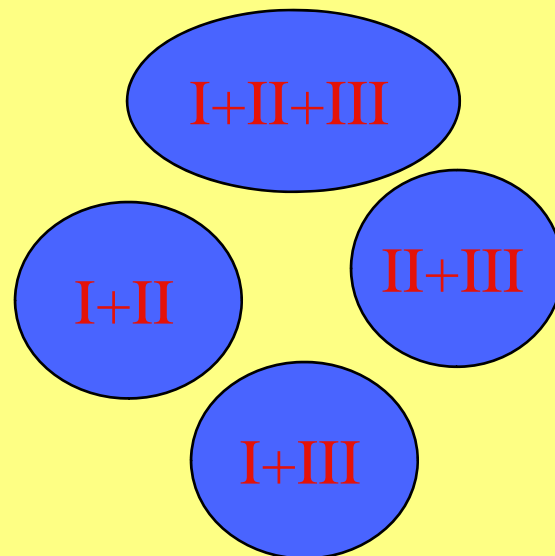
# Heterogeneous Search Models

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

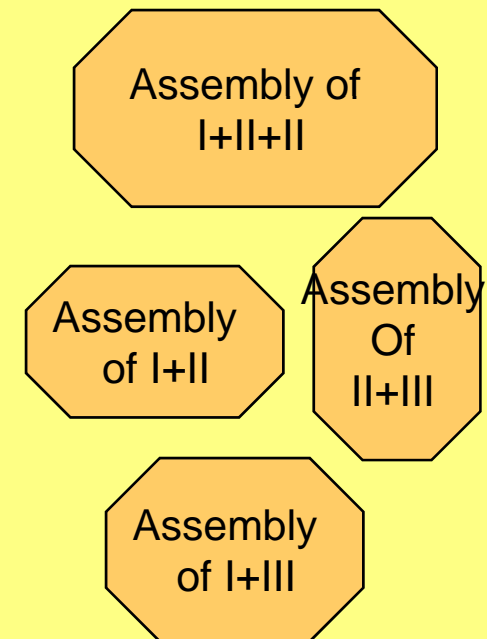
User's sequences  
models



DB



Search

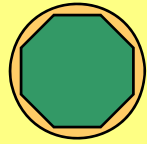


# Example 1: 2dwr

## Homologues

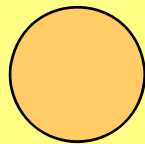
2aen: monomer and one domain definition associated with it.

Identity = 82%



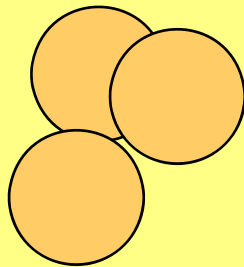
1kqr: monomer, no domain definitions

Identity = 45%

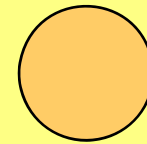


1z0m: dimer, no domain definitions

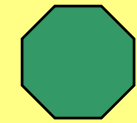
Identity = 25%



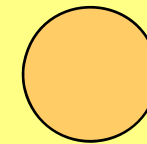
## Derived search models (and their priority)



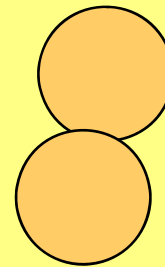
(1)



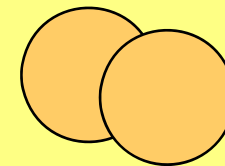
(2)



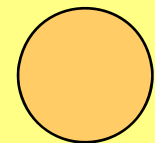
(3)



(4)



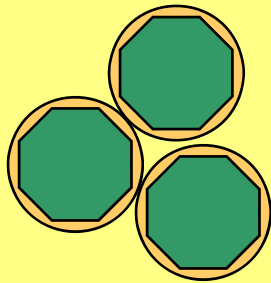
(5)



(6)

# Example 2: 2chp

Homologues

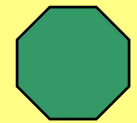
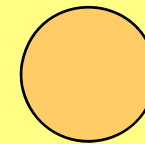
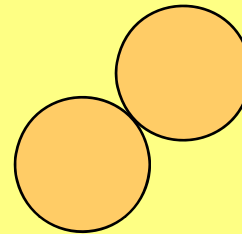
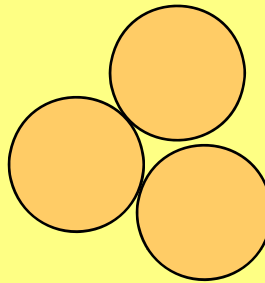


1ji5: trimer  
Identity = 50%

1jig: trimer  
Identity = 49%

1n1q: trimer  
Identity = 48%

Derived search models  
(and their priority)



(1)

(2)

(3)

(4)

(5)

(6)

(7)

(8)

(9)

(10)

(11)

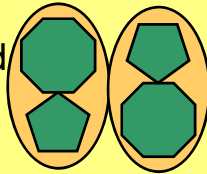
(12)

# Example 3: 2gi7

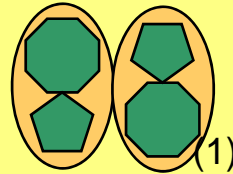
## Derived search models (and their priority)

### Homologues

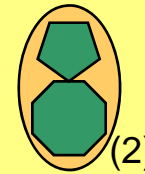
1p7q: homo-dimer;  
each monomers is formed  
by two domains.  
Identity = 45%



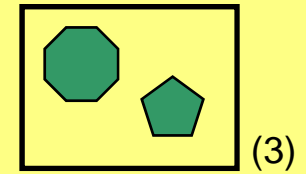
dimeric



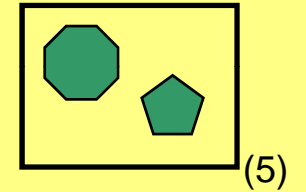
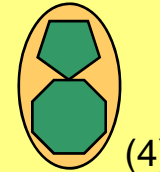
monomeric



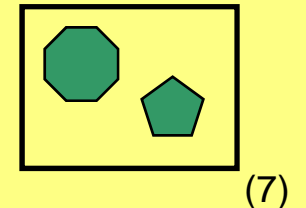
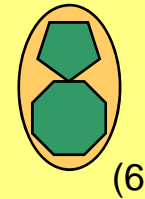
“multi-domain”



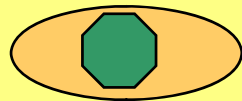
1ufu: monomer  
formed by two domains.  
Identity = 45%



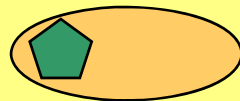
2d3v: monomer  
formed by two domains.  
Identity = 46%



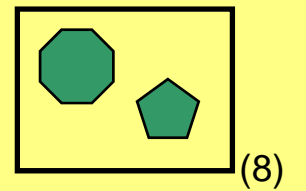
xxxx: contains  
domain 1  
Identity = 42%



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)

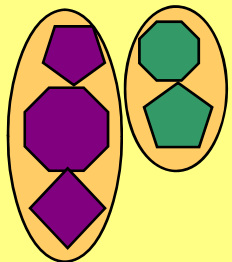
## Assembly models

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:

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

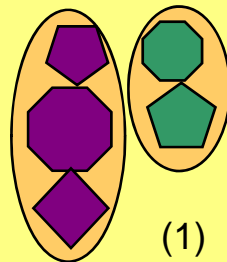


Derived search models (and their priority):

assembly

monomeric

“multi-domain”



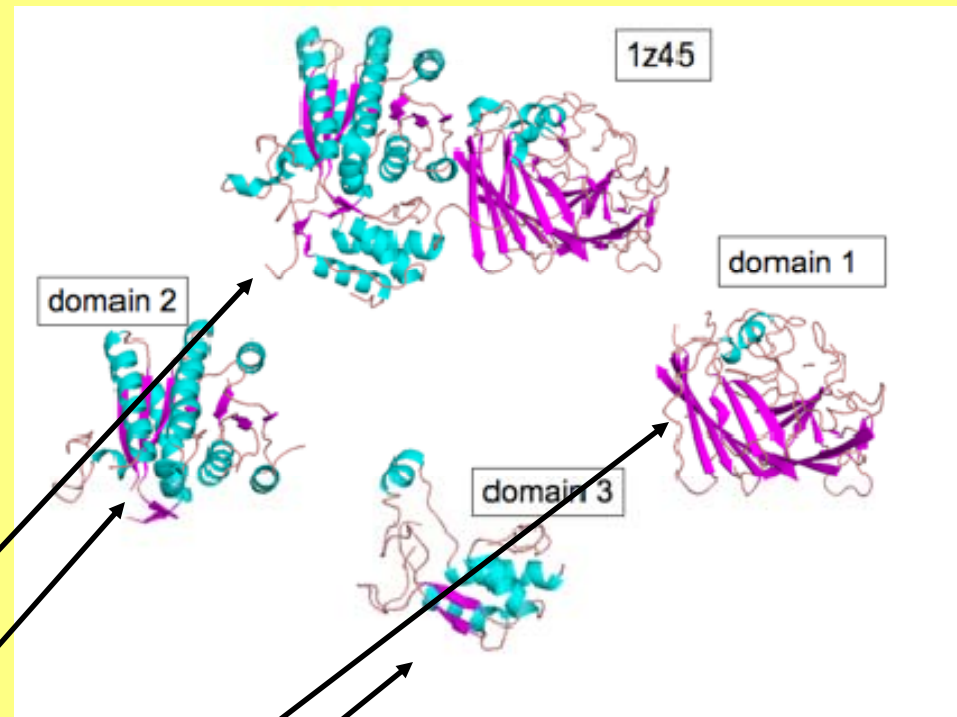
Other homologues (1t43, 1nv8, 1zbt, 1rq0) are matching only one of two sequences. Priority rules applied to them are as in previous examples.

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

Only non- redundant structures are stored

All structures **newly** deposited to the PDB are tested

against the **old** internal database by using BALBES.

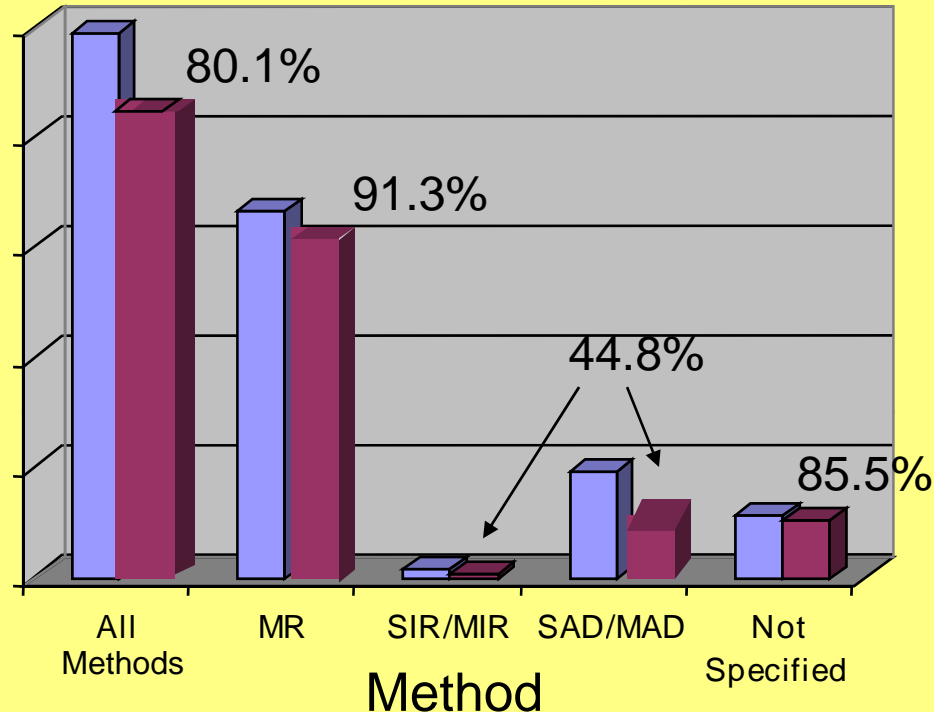
Only after that the DB is updated.

Updating and test 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 latest 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

When we added a new option: To check all potential space groups with full automatic molecular replacement. It was used more than we expected. We had to turn this option off temporarily. We have already updated the system and it is now available.

## 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:

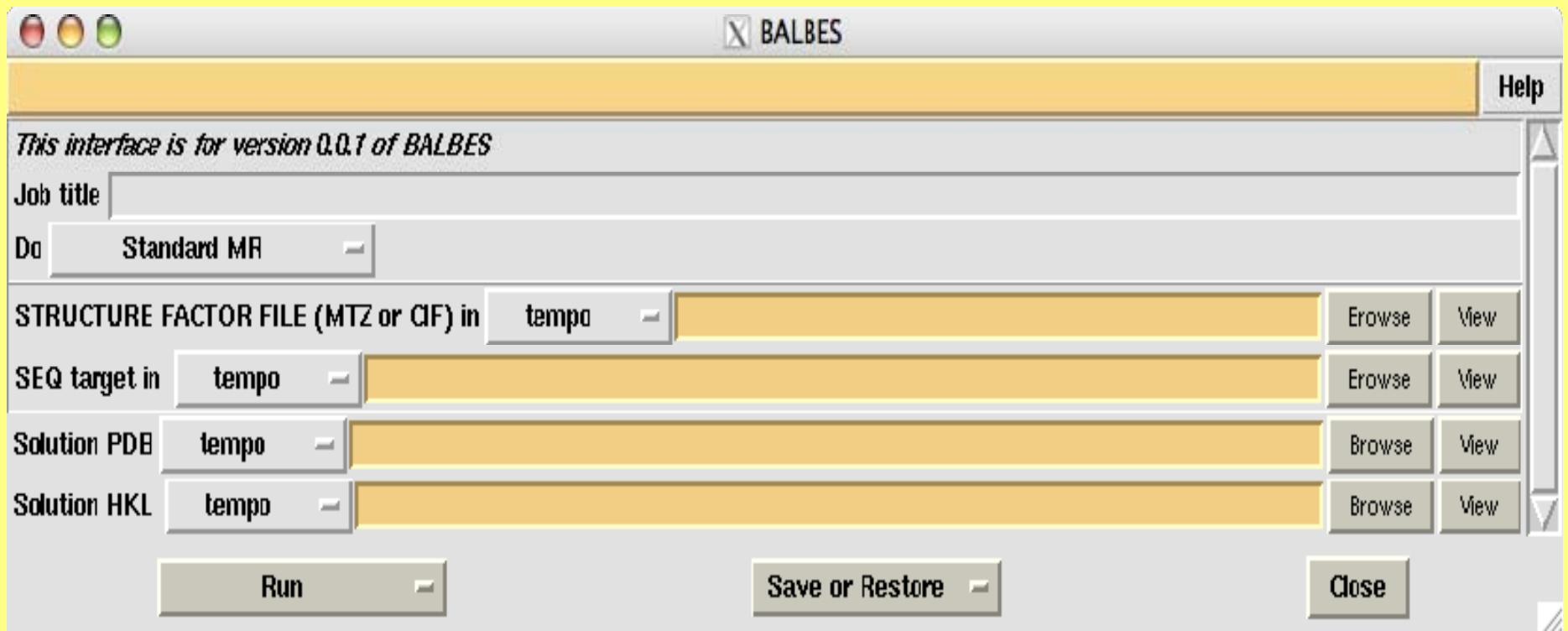
```
balbes -f structure_factors_file -s sequence_file -o  
output_directory
```

-f required

-s required

-o optional

# BALBES CCP4i interface



# BALBES Interface in Our Web Server

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

THE UNIVERSITY of York

York Structural Biology Laboratory



University | Chemistry | YSBL

Home

- ☒ Runnable Programs
- ☒ Downloads
- ☒ Dictionary

## Welcome to YSBL Software

**Any problems? - please contact [pyoung@ysbl.york.ac.uk](mailto:pyoung@ysbl.york.ac.uk)**

### Runnable Programs

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

Other Options - [Register](#), [Forgotten Password](#), [Change Password](#).

### Downloads

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

<a href="#">Balbes</a>	<i>an automated molecular replacement (MR) pipeline</i>
<a href="#">Molrep</a>	<i>an automated program for molecular replacement</i>
<a href="#">Refmac</a>	<i>a macromolecular refinement program</i>
<a href="#">Sfcheck</a>	<i>assessment of X-ray data and/or agreement between atomic model and X-ray data</i>
<a href="#">CCP4mg</a>	<i>an easy way to create beautiful publication quality images and movies</i>
<a href="#">Coot</a>	<i>a program for model building, model completion and validation</i>

### Dictionary

Download the Refmac [Dictionary](#).

wellcome trust

bbsrc  
biotechnology and biological  
sciences research council



# BALBES Interface in Our Web Server

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

THE UNIVERSITY *of York*

**York Structural Biology Laboratory**



University | Chemistry | YSBL

Home (Logout) > Login > Programs

Username: **garibM**

## Programs

Note: *You must have a CCP4 licence to run these programs.*

<b>Balbes</b>	An automated Molecular Replacement (MR) pipeline - Balbes integrates into one system all the components necessary for solving a crystal structure by Molecular Replacement
<b>ModSearch</b>	A program providing template models (including multimeric and domain information) for Molecular Replacement, using the sequence file supplied by the user - <a href="#">more info on ModSearch</a>
<b>Zanuda</b>	Structures with pseudotranslation: validation or correction of crystallographic origin - <a href="#">more info on Zanuda</a>
<b>Sfcheck</b>	Assessment of X-ray data and/or agreement between the atomic model and X-ray data
<b>Buccaneer</b>	Performs statistical chain tracing by identifying connected alpha-carbon positions using a likelihood-based density target.

# BALBES Interface in Our Web Server

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

THE UNIVERSITY *of York*

**York Structural Biology Laboratory**



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 precede each sequence)

Check Full Spacegroup:

Run ARP/wARP  
(on the Balbes solution):

Dissemination Level: Confidential

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



New version generates psuedo-NMR models automatically and uses them

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 its quality factor.

## Complexes

In cases of complexes (more than one sequence) the system first tries assemblies (if available). If it can find 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 second, third etc proteins. Again with and without ensembles.

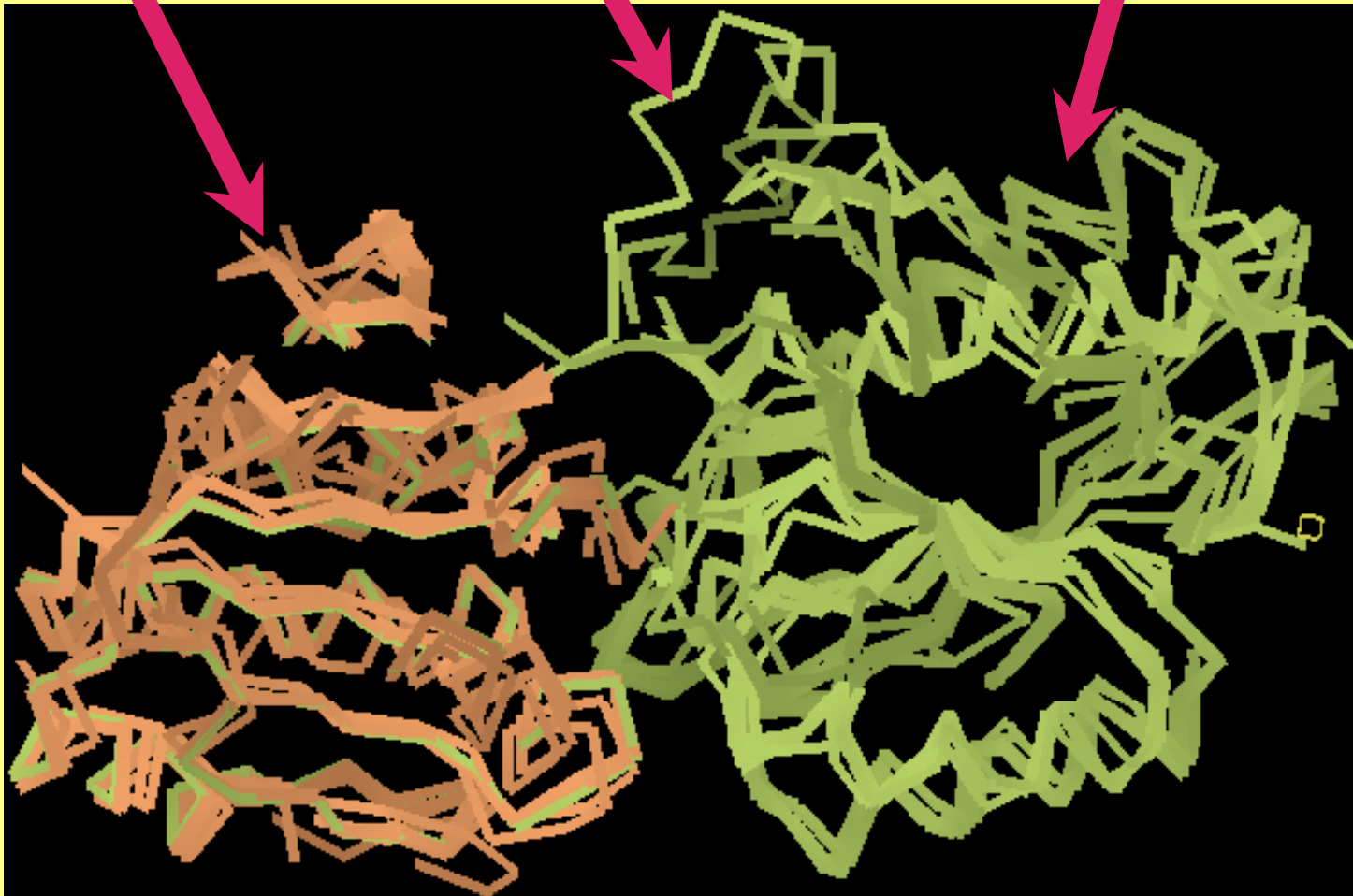
Moreover if space group is uncertain then the program will do all calculation for each potential space group candidate. Decision about space group is made at the very end of all runs (It may take some time).

## Complexes: Two domain example

Domain1

Flexible loop

Domain2



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.



# 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

## Team (YSBL, York)

Alexei Vagin

Fei Long

Paul Young

Andrey Lebedev

## Acknowledgements

E.Krissinel for PISA MSD/EBI, Cambridge

All CCP4 and YSBL people for support

ARP/wARP development team

Wellcome Trust, BBSRC, EU BIOXHIT, NIH for support

# The End

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/>