CCLRC
Daresbury Laboratory

MrBUMP - Automated Search Model Generation and Molecular Replacement

Ronan Keegan (r.m.keegan@dl.ac.uk), Martyn Winn (m.d.winn@dl.ac.uk),
Daresbury Laboratory.

CCP4

HTPX

The aim of Mr Bump

•An automation framework for Molecular Replacement.
•Particular emphasis on generating a variety of search models.

•Wraps Phaser, Molrep and Amore.
•Also uses a variety of helper applications (e.g. Chainsaw) and bioinformatics tools (e.g. Fasta, Mafft)
•Uses on-line databases (e.g. PDB, Scop)

•In favourable cases, gives “one-button” solution
•In unfavourable cases, will suggest likely search models for manual investigation

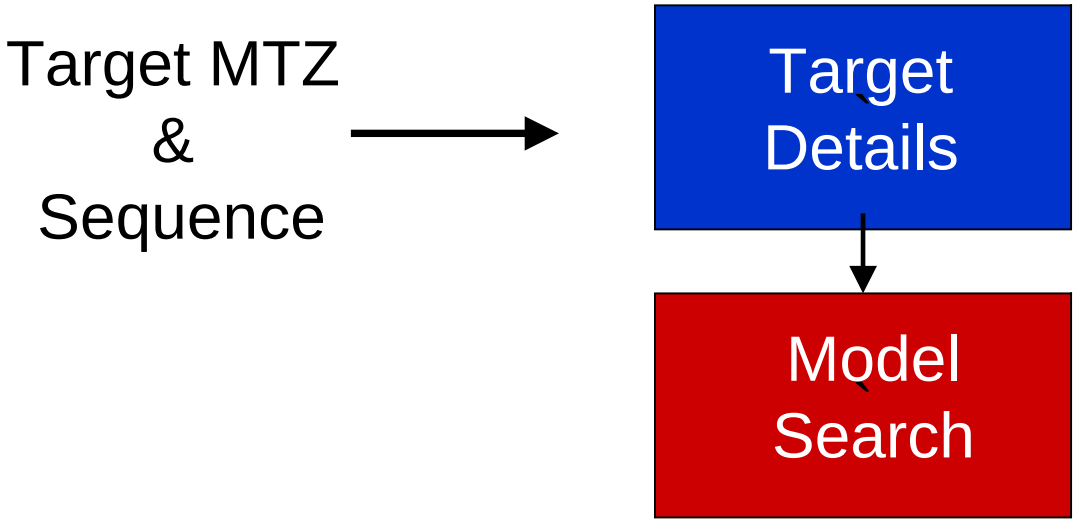
Target MTZ
&
Sequence

Target Details

Ronan Keegan
CCP4

Target Details

- Calculate:
  - Matthews Coefficient.
  - Estimated number of molecules in the a.s.u.
  - Molecular weight.
  - Extraction of any other relevant information.

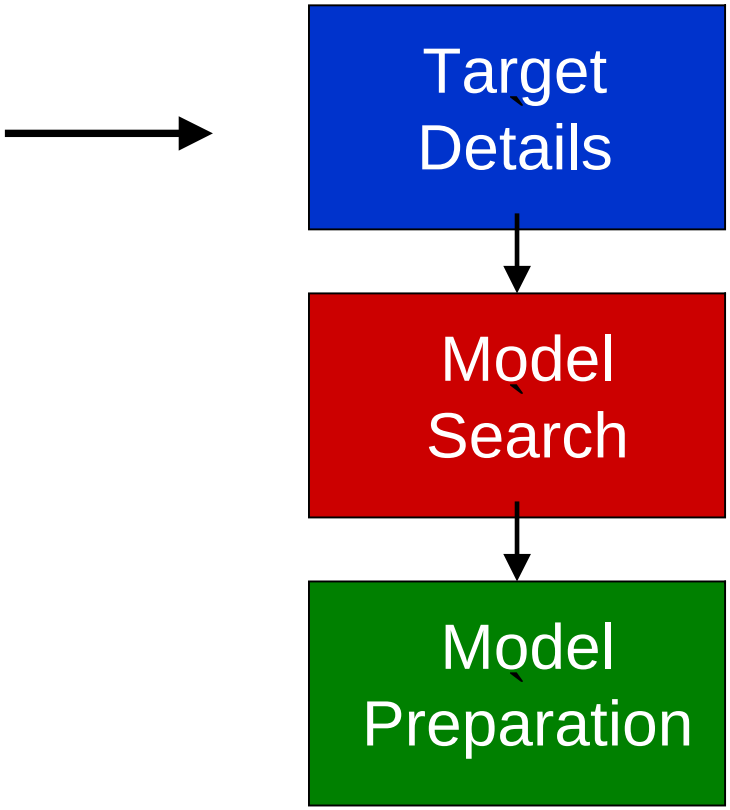


Ronan Keegan
CCP4

Model Search

- Template search:
  - Fasta sequence-based search using web or a local database of PDB sequences,
  - SSM secondary structure search using EBI webservice,
  - SCOP search (EBI) for domains within structures found in Fasta and SSM searches,
  - PQS search (EBI) for multimers based on structures found in Fasta and SSM searches.
- Multiple Alignment (Mafft or Clustalw) of all sequences with the target sequence and top scoring models are selected for further processing.

Target MTZ
&
Sequence

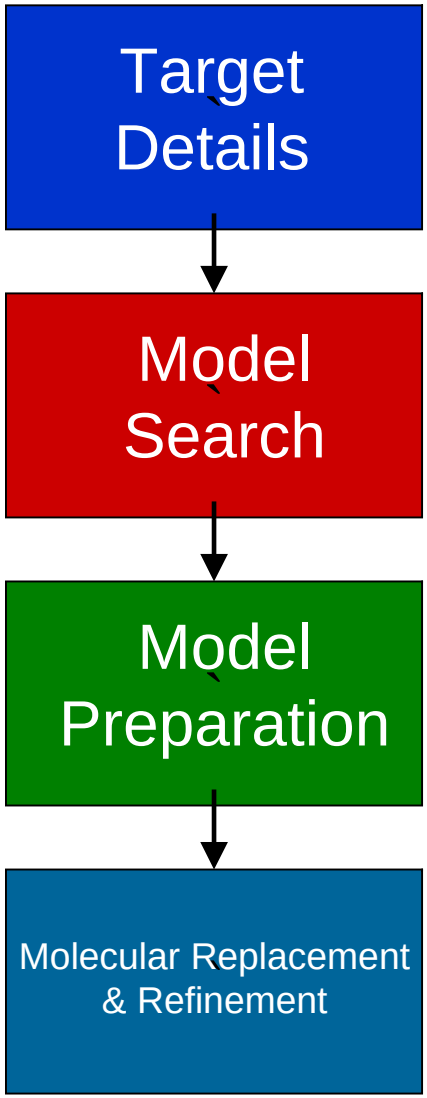


Ronan Keegan
CCP4

Model Preparation

- PDB files downloaded.
- SCOP results are used to extract domains from the PDB files.
- Multimer PDB files retrieved from PQS server.
- Search models prepared in three ways:
  - PDBclip – original PDB with waters removed, hydrogens removed, most probable confirmations for side chains selected and chain ID's added if missing.
  - Molrep – side-chain pruning
  - Chainsaw – more severe side-chain pruning

Target MTZ
&
Sequence



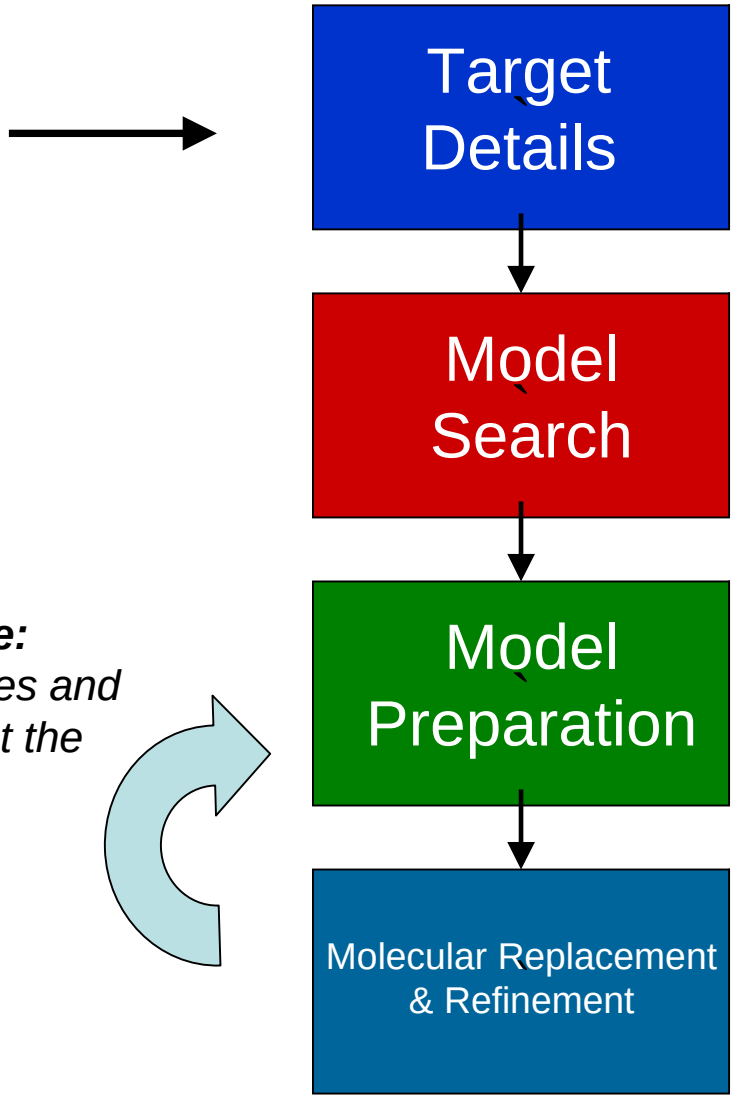
Ronan Keegan
CCP4

Molecular Replacement and Refinement

- search models can be processed with Amore, Molrep or Phaser or combinations of all three,
- an ensemble of the top search models is created as an additional input model to Phaser,
- Refmac used for restrained refinement,
- change in Rfree value during refinement is used to determine how good the resulting model is,
- marginal solutions are high-lighted.

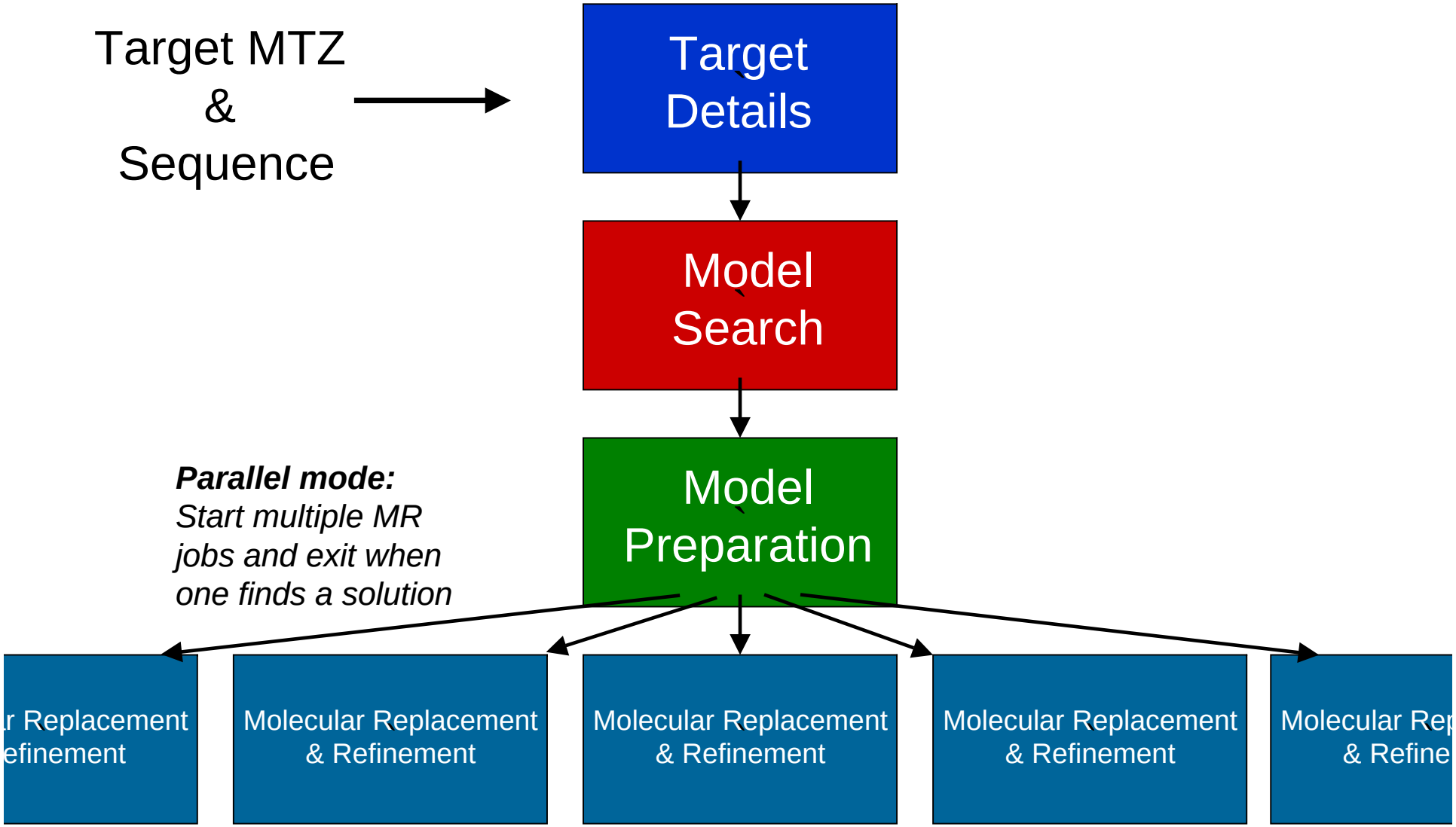
Ronan Keegan
CCP4

Target MTZ
&
Sequence

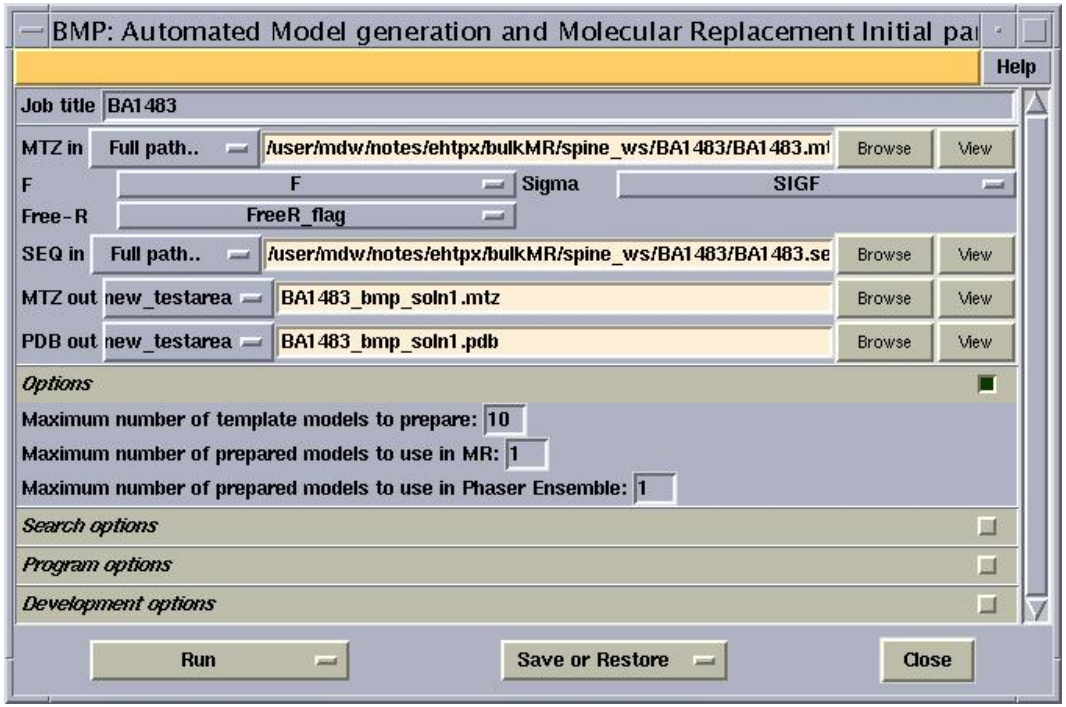


Serial mode:
Check Scores and exit or select the next model

Ronan Keegan
CCP4


Target MTZ
&
Sequence
Target Details
Model Search
Model Preparation
Parallel mode:
Start multiple MR jobs and exit when one finds a solution
r Replacement
efinement
Molecular Replacement & Refinement
Molecular Replacement & Refinement
Molecular Replacement & Refinement
Molecular Rep
& Refine

Testing and pre-release version of Mr BUMP January '06



- Tested against list of SPINE targets in July '05
- Announced at CCP4 study weekend and on CCP4 BB.
- Comes with CCP4 GUI .
- Good deal of interest and one report of a user using it to solve a structure that they hadn't been able to solve since 2002.

Pre-release version available from:

www.ccp4.ac.uk/martyn/BMP/BMP.php

Ronan Keegan
CCP4

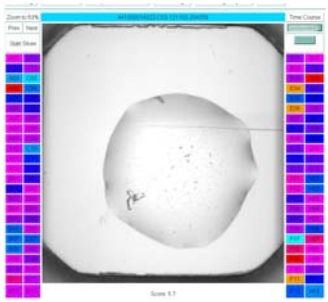
MrBUMP and e-HTPX

e-HTPX – an e-science resource for High Throughput Protein Crystallography

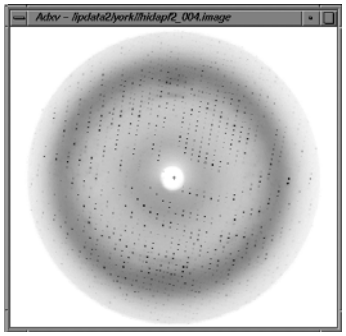
Reasons for e-HTPX

The high volume of data coming from structural genome projects has generated a demand for new methods to obtain structural and functional information about biological proteins and macromolecules. This has led to a demand for high throughput techniques to determine the structure of important proteins.

Ronan Keegan
CCP4

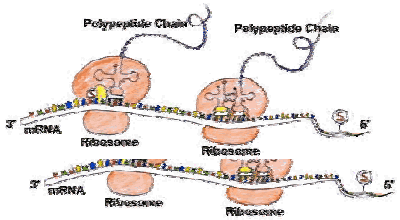


Crystallisation

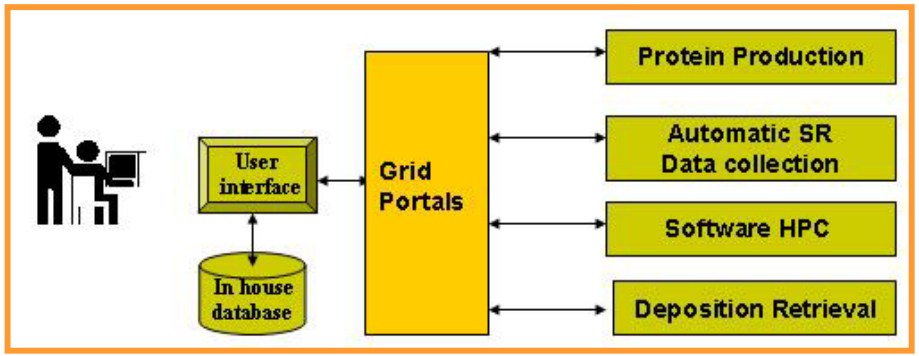


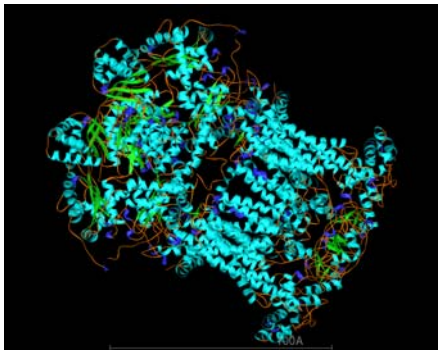
Data Collection

Phasing



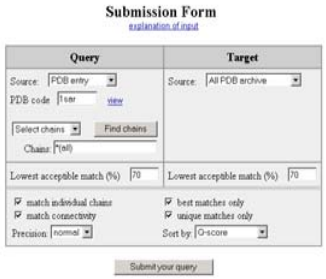
Protein Production



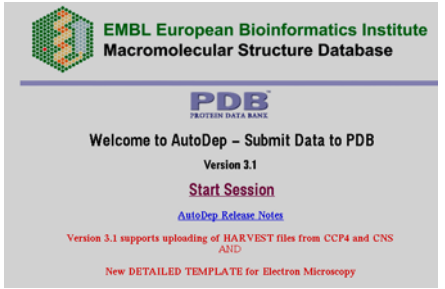


Protein Structure

Target Selection



Structure analysis



Deposition

Ronan Keegan
CCP4

e-HTPX System Architecture – Portal Architectures

- Two portal architectures have been designed in order to suit different research environment needs.
  - Service site portal – portal is hosted by the synchrotron, minimizes the requirements placed on the user. (web-browser is the only requirement).
  - Client site portal – portal is installed and managed at the client institution.
- Both implementations are equivalent clients to the underlying web services.

Service-Site Architecture

• e-HTPX portal is hosted at the synchrotron site and acts as the client interface to the SRS services.

Advantages:

• Client access is centralised through a known / single point of entry (a gateway server). Allows service site firewall administrators to implement IP-recognition methods via defined ports.

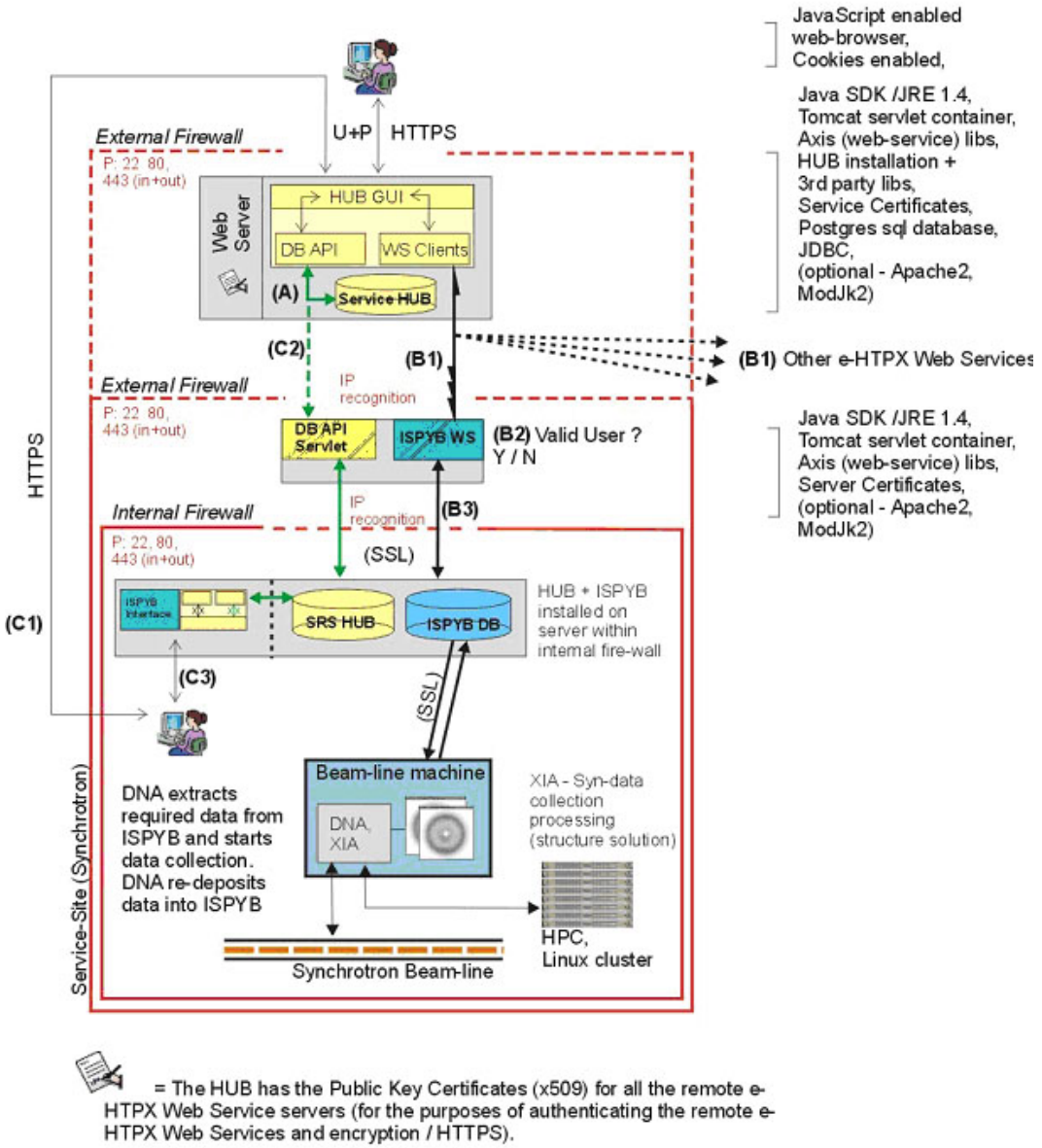
• No requirements placed on client.

Disadvantages:

• Main Architectural weaknesses:-

a) increased service side responsibility for storing users project data in portal/hub database;

b) Interoperability with clients LIMS system.



Client-Site Architecture

• e-HTPX portal is hosted at the client site:
a) Client site responsibility for storing potentially sensitive data (especially relevant for industrial clients).

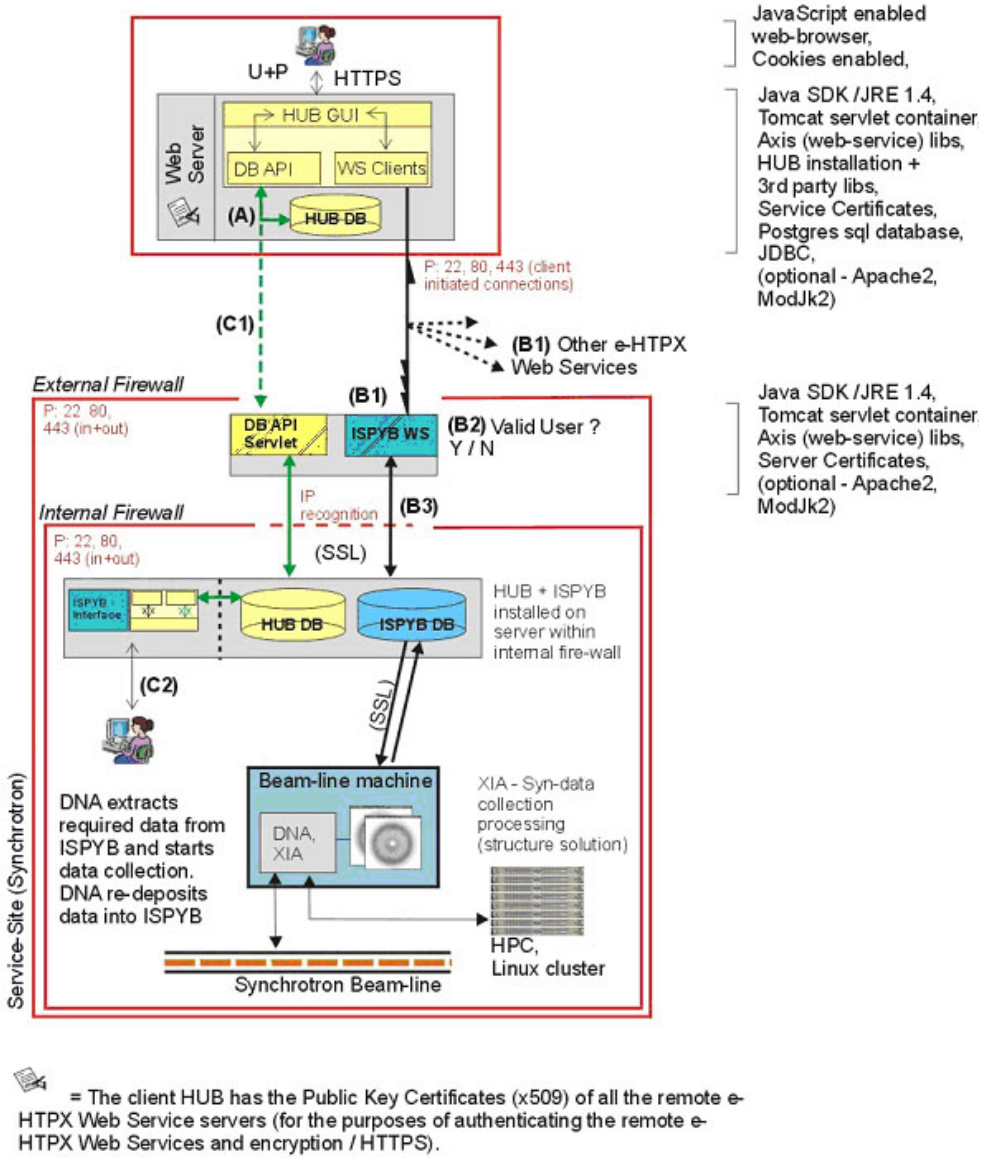
•b) Interaction with LIMS (laboratory information management system).

• Web-Service polling methodology required to circumvent client firewall restrictions.

• Clearly separates web-service client code from web-service host code.

• Host server certificates required on hub to authenticate client institution. Similar approach to authenticating clients (x509 digital certificates).

• To increase service site security, requests may be optionally routed through a single gateway server. A number of methodologies to implement this:- a) Reverse proxy server (apache mod_security); b) Multiple actors on the SOAP path; c) Place web-services on gateway.



MrBUMP and e-HTPX

- Dedicated machine for running e-HTPX structure solution jobs submitted via the e-HTPX portal interface.
- MrBUMP service is made available as a web service with a WSDL description.

Ronan Keegan
CCP4

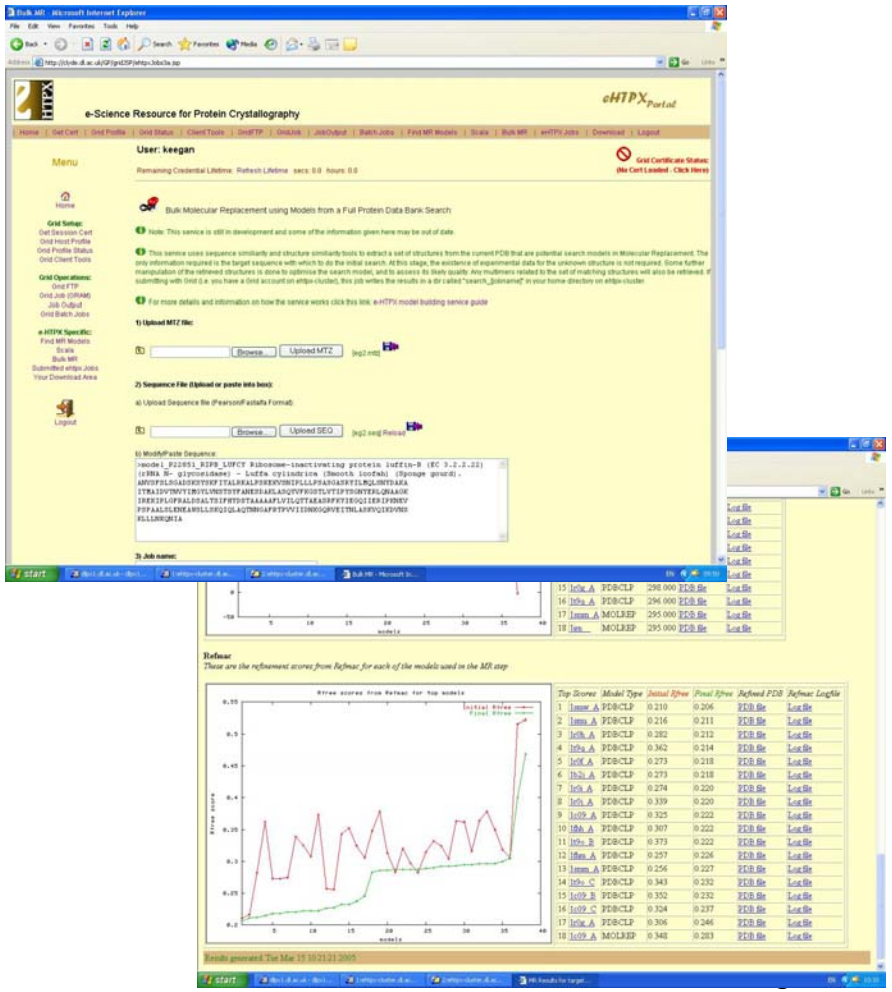
e-HTPX MrBUMP Portal

- The WSDL can be parsed by a user so as to incorporate the service into their own application.

- Alternatively, a remote user can access the service via the stand-a-lone interface provided by e-HTPX.

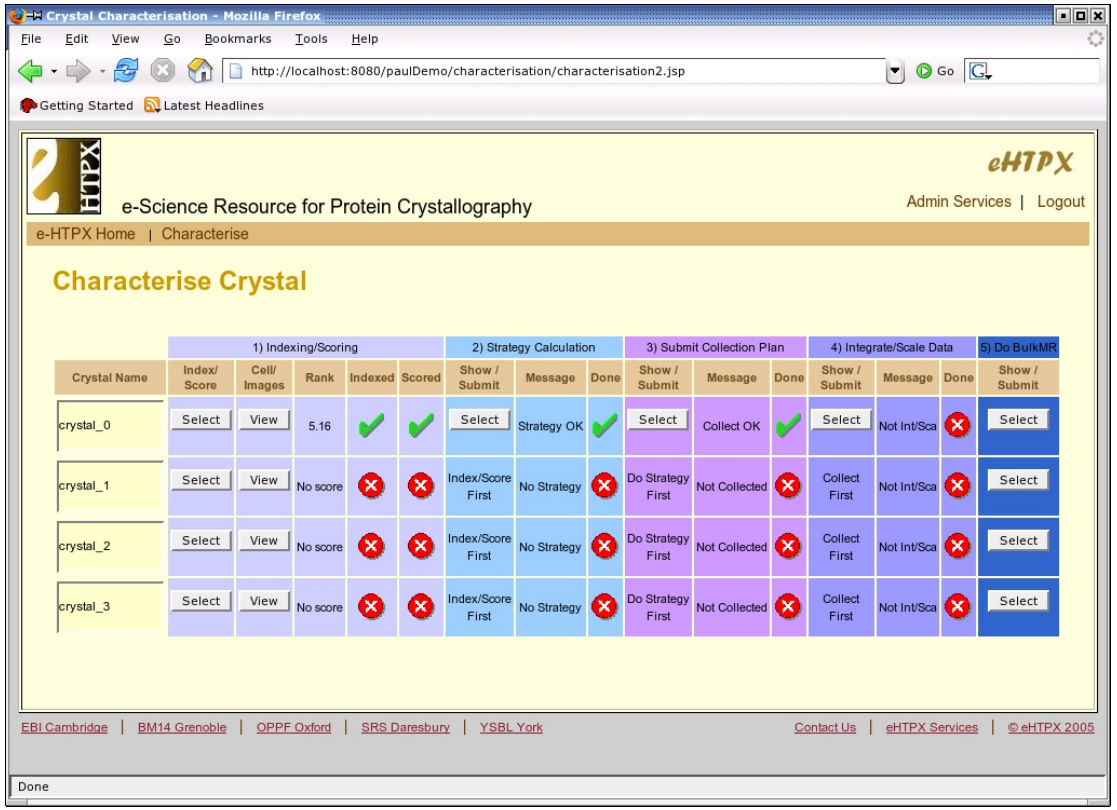
- Users upload their X-ray experiment diffraction data and invoke the service on the HPC resources provided at Daresbury Laboratory.

- Information is transferred in XML format using SOAP.



MrBUMP as part of e-HTPX workflow

- MrBUMP can also be accessed on the dedicated cluster as part of the e-HTPX workflow.
- Portal contains a top-level interface connecting with each of the web-services hosted at each of the sites involved.
- Allows the remote user to monitor the progress of their experiment and also to access the data processing facilities to process the collected data.



Ronan Keegan
CCP4

Future developments

- additional model building methods such as removal of parts of the main chain (e.g. by position or by B-factors),
- support for complexes,
- develop web-service version to allow CCP4i users to run jobs on e-HTPX/CCP4 clusters,
- simplified CCP4 GUI.
- Incorporate PISA multimer determining service (EBI).

Acknowledgements

- e-HTPX:
  - Dave Meredith, Graeme Winter, Michael Gleaves, Daresbury Laboratory.
  - Ian Berry, OPPF, Oxford.
  - Paul Young, YSBL, York University.
  - Oleg Dolomanov, EBI, Hinxton.
  - Ludovic Launer, BM14, ESRF.
  - http://www.e-htpx.ac.uk
- MrBUMP:
  - Martyn Winn, CCP4.
  - http://www.ccp4.ac.uk/BMP/BMP.html

e-HTPX Project

A distributed computing infrastructure for protein crystallographic structure determination.

- Project integrates a number of key services provided by UK e-Science, protein manufacture and synchrotron laboratories.
- Remote access to these services is implemented by a collection of Web services, each developed at the corresponding service-site.
- Client access to the web services is through the e-HTPX portal/hub
  - Easy to use interface
  - Hides user from complexity of underlying distributed computing infrastructure

Stage 1 – Select protein target

↓

Stage 2 – Crystallization of Protein

↓

Stage 3 – Data Collection (X-ray diffraction images, Scaling and Integration)

↓

Stage 4 – Structure Solution (HPC data processing to derive digital protein model)

↓

Stage 5 – Submit model into public database

Ronan Keegan
CCP4