CCP4 - Software for Protein Structure Solution

Ronan Keegan, CCP4 Group

Research Complex at Harwell

What is CCP4? – History

- One of several CCPs set up in the UK to advance and support scientific software development
- CCP4 (Collaborative Computational Project Number 4) was set up in the late 1970's to bring together the leading developers of software in the field of protein X-ray crystallography in the UK
- The aim was to assemble a comprehensive collection of software to satisfy the computational requirements of the relevant UK groups

CCP4 today

- Funded by the BBSRC and MRC and coordinated by the STFC as part of the Computational Science and Engineering Department (CSED)
- Industrial funding
- Core group based at the Research Complex at Harwell (RCaH) next door to Diamond

CCP4 usage


Greenland
Iceland
Canada
United States
Mexico
Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Suomi Finland
Sverige Sweden
Norge Norway
Россия Russia
Україна Ukraine
Kazakhstan
Mongolia
China
S Korea
Afghanistan
Iran
Iraq
Pakistan
India
Saudi Arabia
Algeria
Libya
Egypt
Mauritania
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Kenya
DR Congo
Tanzania
Angola
Namibia
Botswana
Madagascar
Venezuela
Colombia
Brasil
Brazil
Perú
Bolivia
Chile
Argentina
Indonesia
Papua New Guinea
Australia
New Zealand

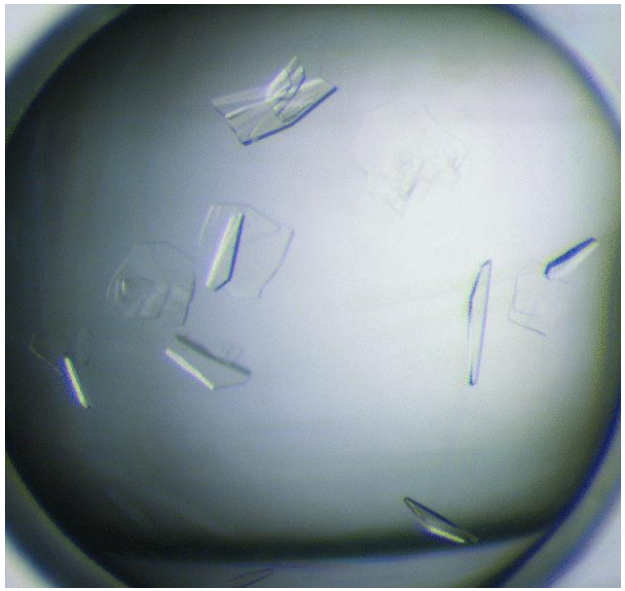
CCP4 Software Suite

- Provide a comprehensive suite of software for the determination of protein structures from X-ray diffraction images
- Support software developers at several universities throughout the UK and further afield
- Rapid pace of development of new and existing software
- Suite is updated and released to a global user base approximately once every 18 months

A very rough guide to X-ray crystallography for protein structure solution

- Proteins/DNA/RNA are the fundamental building blocks of all life
- X-ray crystallography is the most commonly used method for determining the atomic level detail of these structures

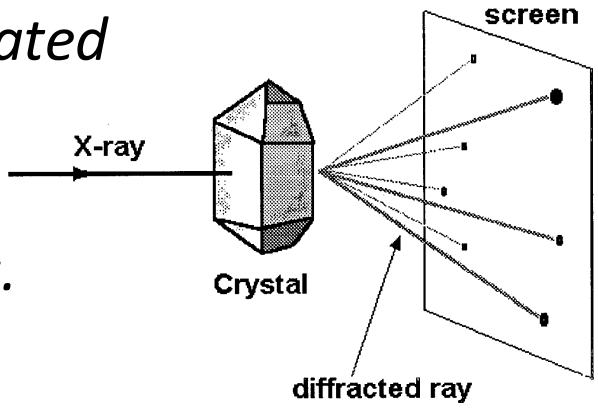
  - *Step 1: Protein expression, purification and crystallisation*

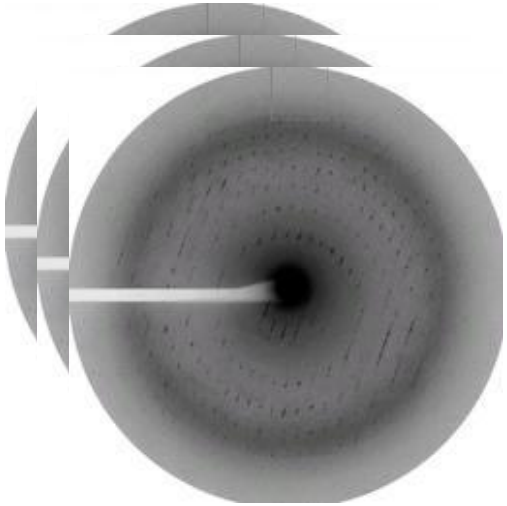


A very rough guide to X-ray crystallography for protein structure solution

- Proteins/DNA/RNA are the fundamental building blocks of all life
- X-ray crystallography is the most commonly used method for determining the atomic level detail of these structures

  - *Step 2: Crystal rotated in X-ray beam to generate set of diffraction images.*

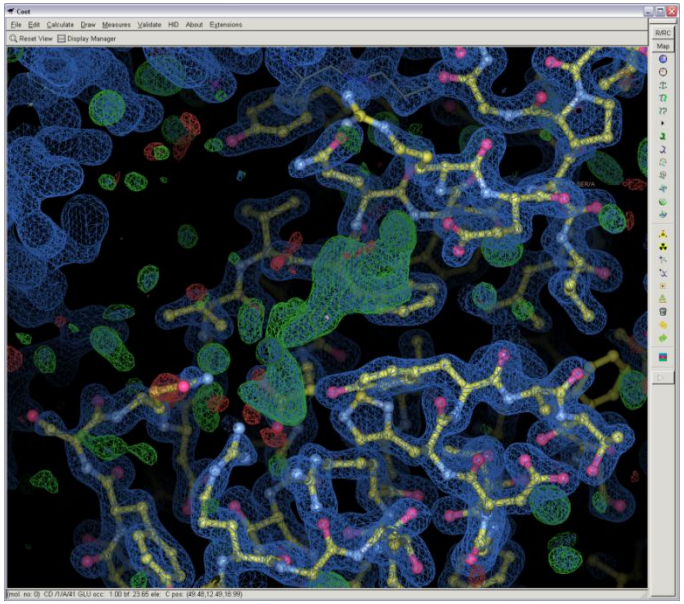




A very rough guide to X-ray crystallography for protein structure solution

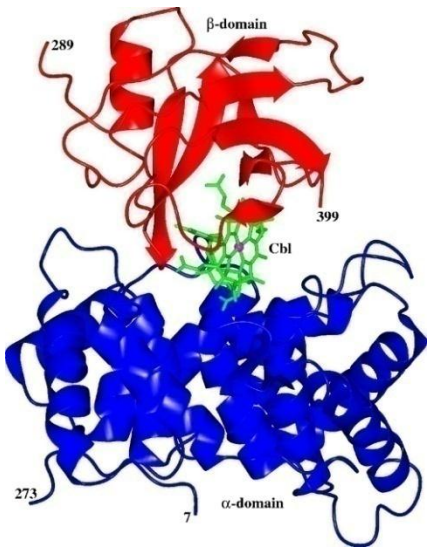
- Proteins/DNA/RNA are the fundamental building blocks of all life
- X-ray crystallography is the most commonly used method for determining the atomic level detail of these structures

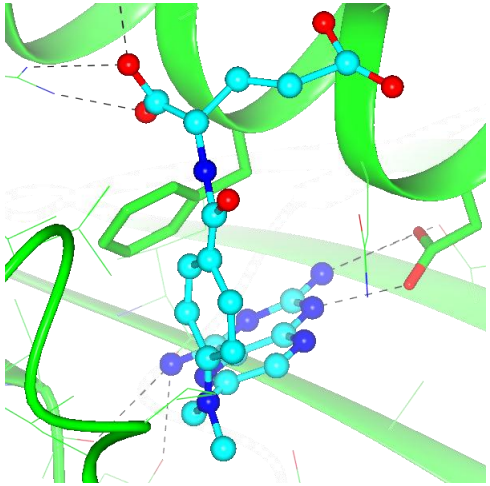
  - *Step 3: Diffraction images processed to generate electron density map for target protein. Model for structure built into map*

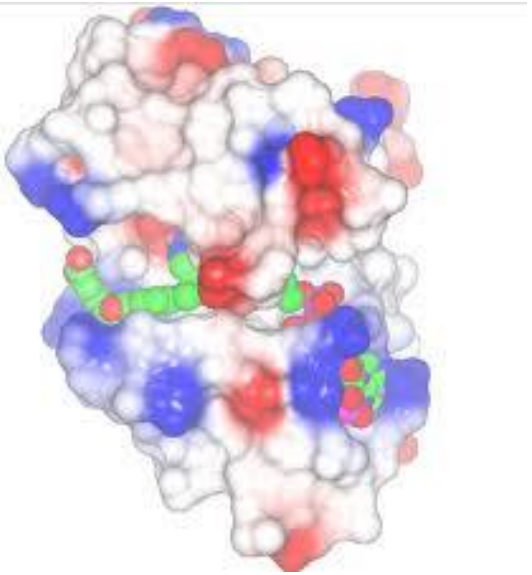


A very rough guide to X-ray crystallography for protein structure solution

- Proteins/DNA/RNA are the fundamental building blocks of all life
- X-ray crystallography is the most commonly used method for determining the atomic level detail of these structures

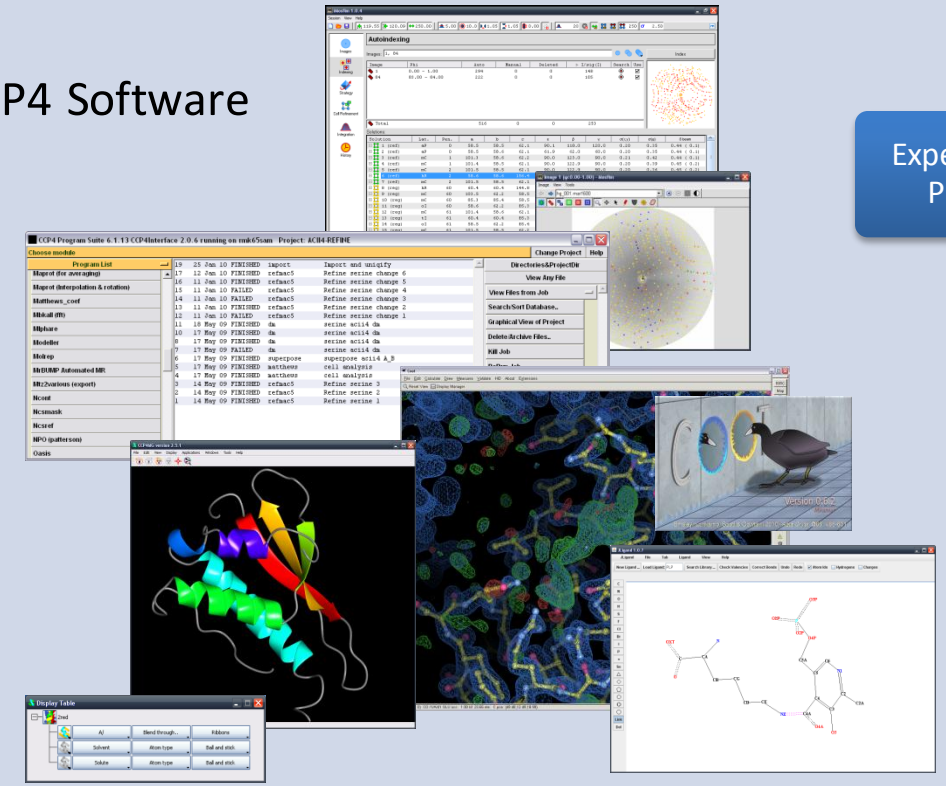






Diamond Software (*Alun Ashton*)

CCP4 Software



Crystallisation

Data Collection

Data Processing and Reduction

Experimental Phasing

Molecular Replacement

Density Modification

Model Building

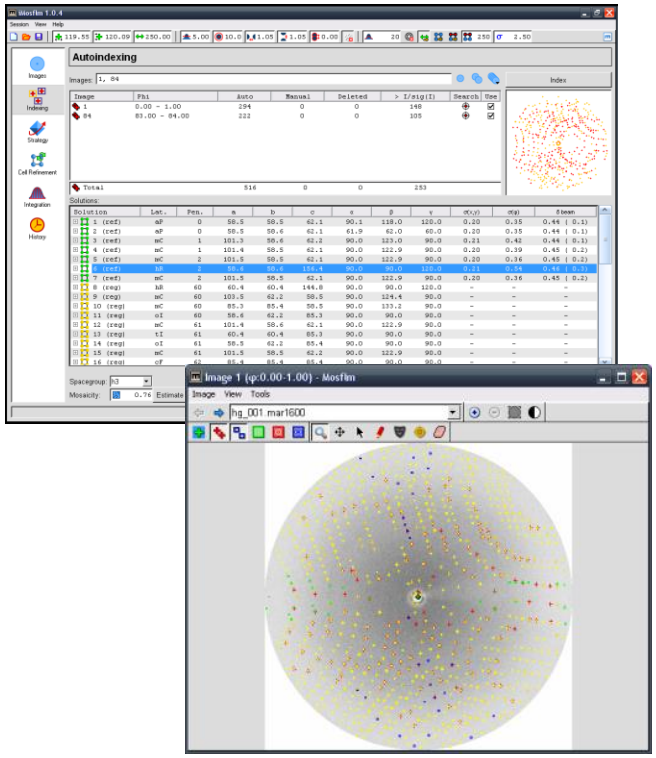
Refinement

Structure Analysis

Deposition

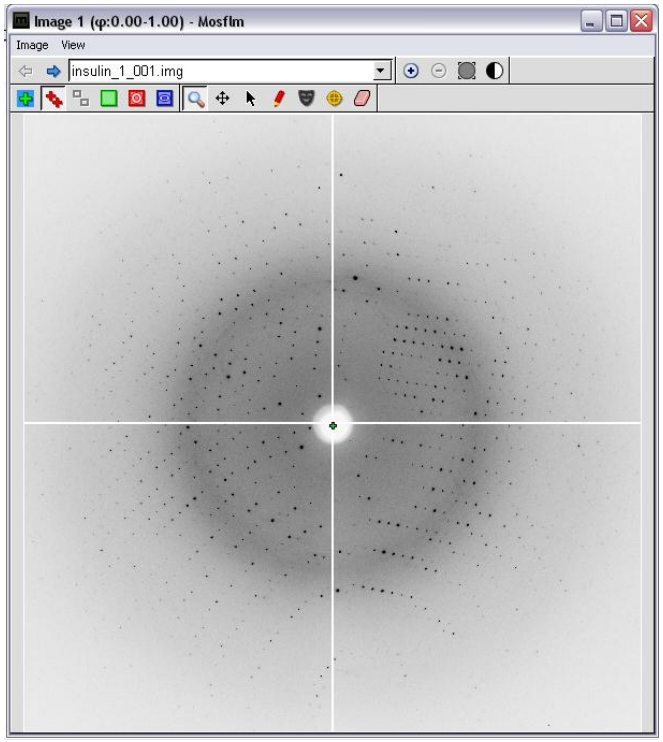
Processing of X-ray images

- Currently: Mosflm
  - Jointly developed by CCP4 and the MRC/LMB
  - 40 year old package (old but reliable)
  - Not designed for high speed and high throughput on modern beamlines
- For a CCD data set, data size for full collection sweep on a single crystal can vary between 1 and 10 GB
- For new Pilatus detectors, data size can be between 5 and 20 GB but collected at a much faster rate



Automatic data processing: Xia2

- Developed originally as part of the e-HTPX, an e-Science pilot project to develop technology for high throughput protein crystallography and now supported by Diamond.
- Automates the data processing procedures – wraps Mosflm as well as other similar packages
- Currently deployed at Diamond and available to users around the world through the CCP4 suite



BioStructX

- Pan-European project to coordinate and support access to emerging methods and infrastructure in structural biology
- CCP4 along with Diamond and the MRC/LMB in Cambridge are involved in Work Package 6: X-ray Data Integration (Synchrotrons, software developers and detector manufacturers)
- Aim is to produce a new data integration package compatible with modern pixel array detectors such as the Pilatus systems installed at Diamond (3x6M, 1x2M @ 25 Hz currently)

HecKLer

Andrew Leslie (MRC/LMB)

Gwyndaf Evans (Diamond)

David Waterman (CCP4)

Graeme Winter (Diamond)

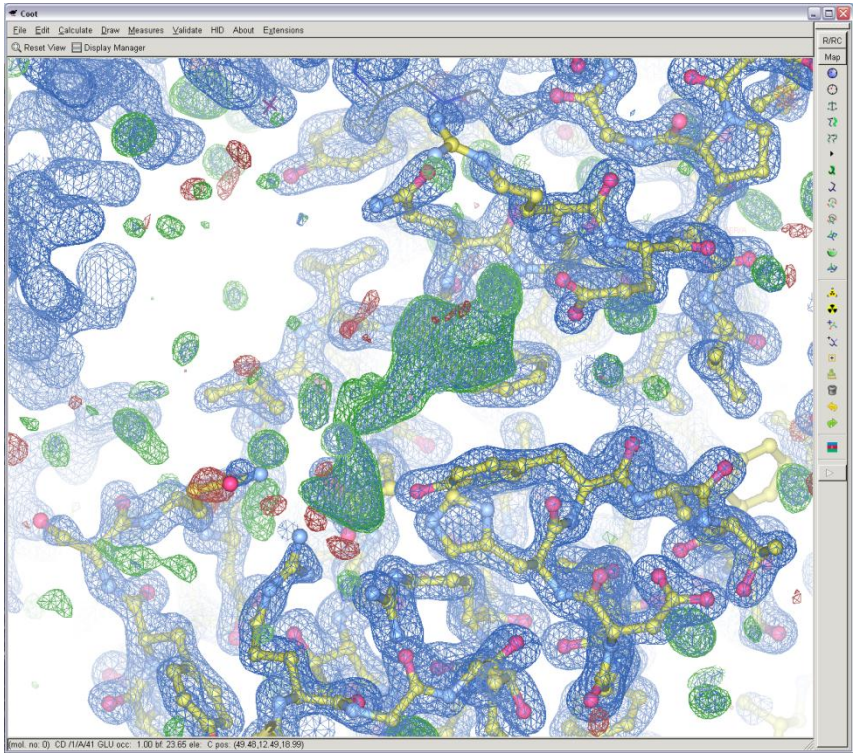
- High speed to keep pace with high rate of data collection (up to 100 Hz)
- Typical datasets could include several thousand images totalling as much as 20 GB in size
- Deal with poor crystal samples – high mosaicity, overlapping spots, etc.
- Unified with FEL data collection software efforts to produce single package
- Needs to be highly portable and parallel for running on GPU and CPU machines

Structure Solution on the Beamline

- Currently practice at Diamond:
  - Automated data acquistion
  - Automate X-ray image data processing
- A desirable goal is to produce a high quality structure for the target at the beamline:
  - To provide rapid feedback on crystal quality
  - Enable high-throughput of protein structure determination
  - Enable high speed assessment of ligand binding in a series of crystals for drug discovery
- Recent work has seen efforts to automate the structure solution step

DIMPLE – Drug candidate assesement

- High throughput assessment of crystals containing potentially bound drug candidates
  - Primary use of synchrotron facilities by Pharma industry
  - 100s of crystals
  - Speed critical

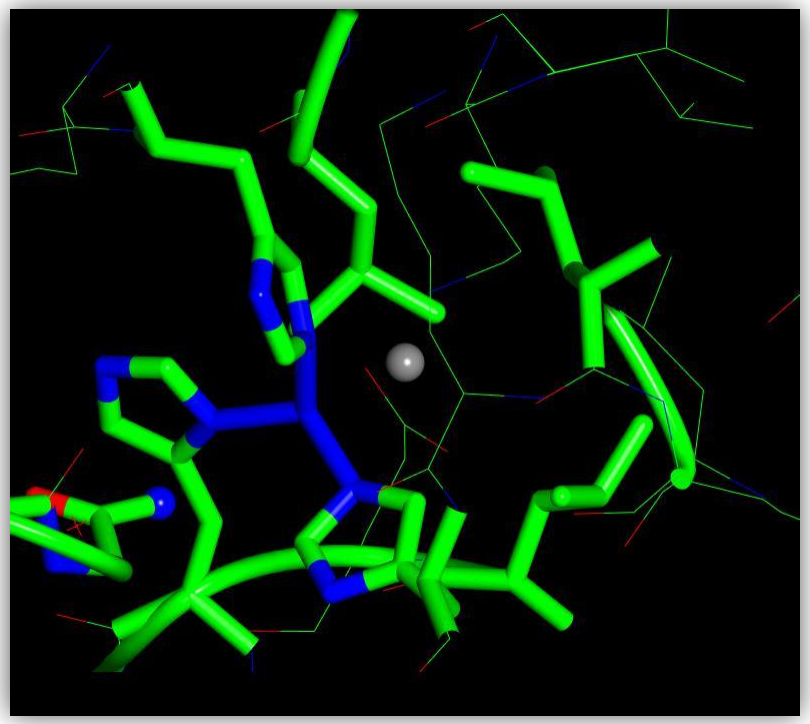


The Phase Problem

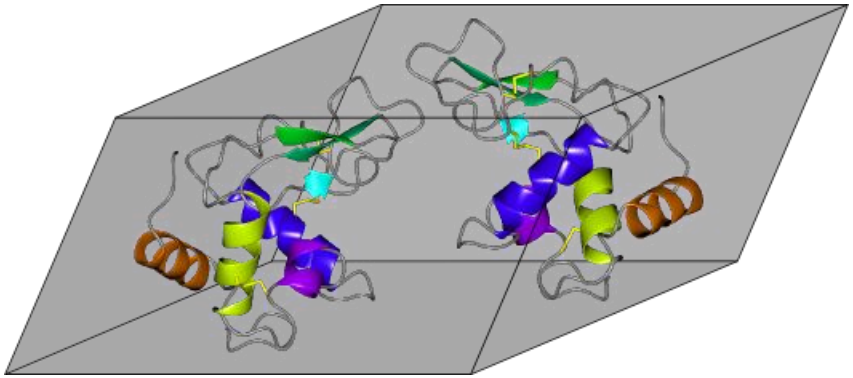
- Diffraction image spot intensities and phase information are required to construct electron density map of target protein
- Phase information not given by diffraction images. Must be derived from other techniques –
  - *Experimental Phasing:* Based on comparison of X-ray data from two or more slightly different crystal structures
  - *Molecular Replacement*: Phases taken from similar, related, proteins called homologues
  - Can't use direct methods given large number of atoms

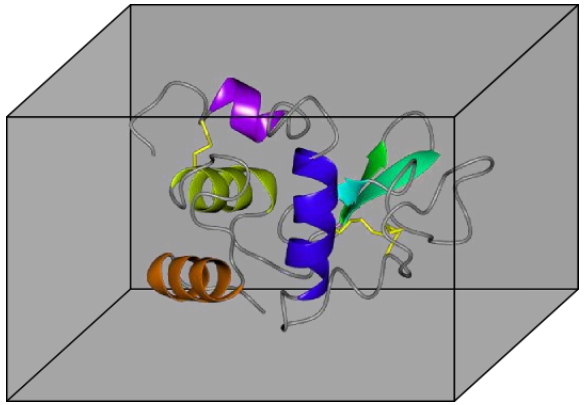
Experimental Phasing

- One method is heavy atom phasing:
  - There are only a few heavy atoms in the crystal unit cell
  - Therefore, their positions can be found by direct methods from intensity differences between native crystals and crystals with the soaked-in metals
  - Native phases can then be calculated from metal sub-structure phases
- Other methods such as anomalous scattering employ a similar approach



Molecular Replacement



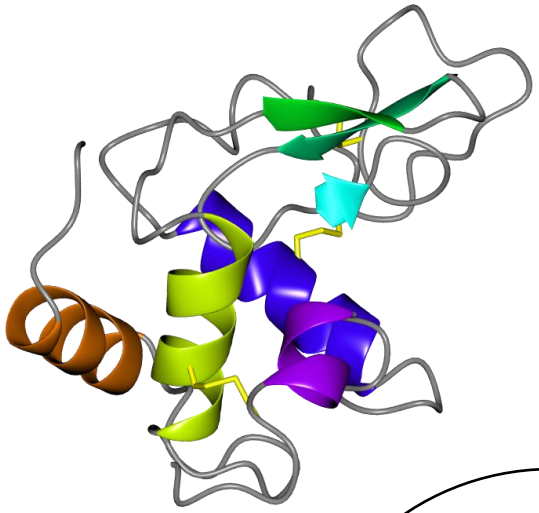


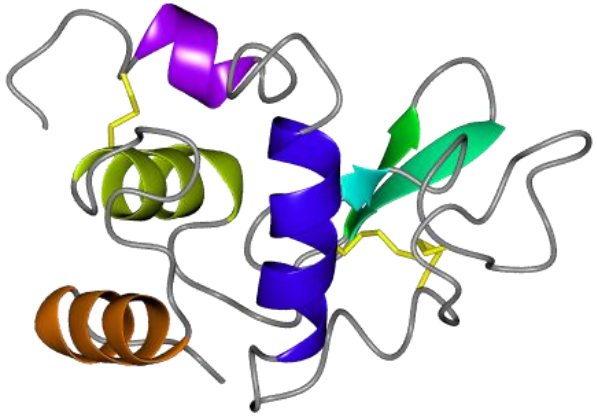
Previous crystal form

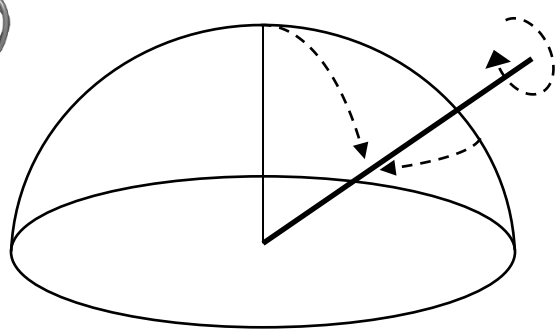
Current crystal form

$$\Re\left(\phi,\psi,\kappa,x,y,z\right)$$

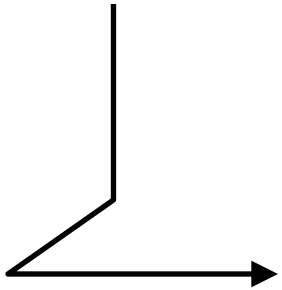
Separability







+



(φ, ψ, κ) (x, y, z)

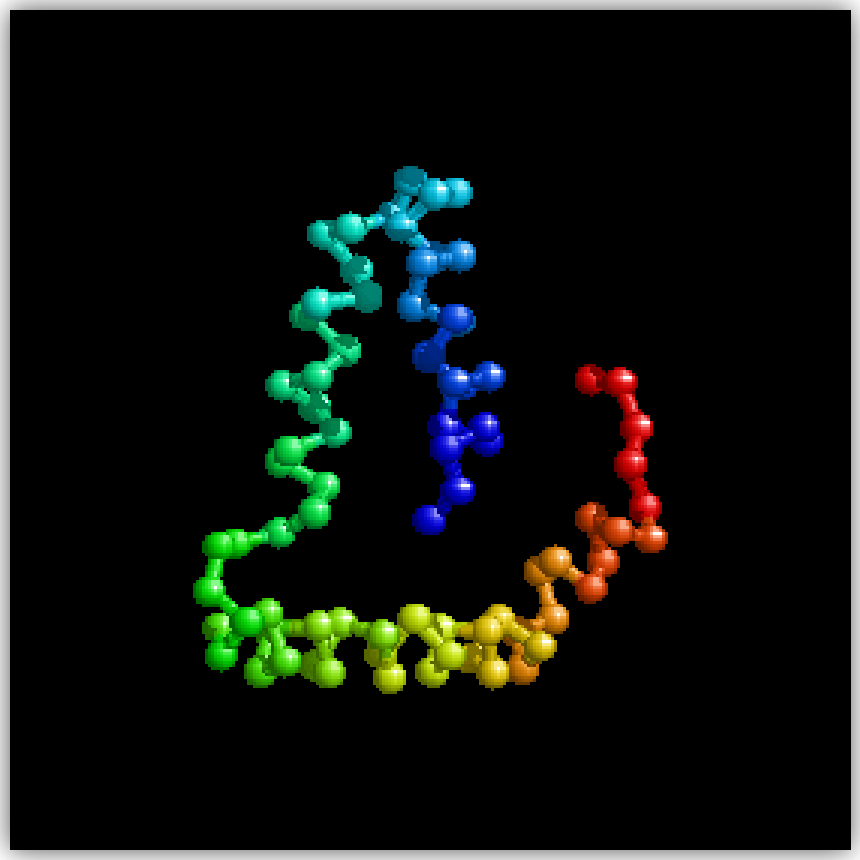
Experimental Phasing	Molecular Replacement
Essentially the only method for solving novel target structures	For solving complexes including known structures
More complicated experiment (considerably more wet-lab work beforehand and the handling of potentially hazardous materials)	Easier crystallisation - single native crystal needed
Requires no prior knowledge about the structure	Requires close known homologous structure and can be computationally expensive for marginal cases

Automation of traditional Molecular Replacement

- To improve the chances of success CCP4 has developed two automated pipelines for doing MR
- *BALBES* and *MrBUMP* take different approaches to the search for good homologues and their preparation for MR
- Both are computationally intensive, many trial models must be generated and tested (typically 40-50)
- To achieve a rapid solution cluster resources can be employed to batch farm the processing of individual search models
- *BALBES* is primarily used through its web interface with a cluster backend. Significant processing power can help to satisfy CCP4's large user-base.
- Both programs currently being deployed at Diamond

Or, new approaches remove the need for homologues...

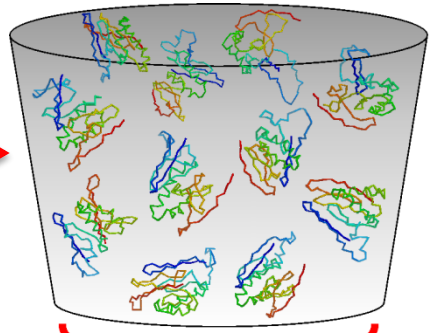
- The field of *ab initio* or *de novo* prediction of protein folds purely from their amino acid sequence has been developing rapidly in recent years (e.g. Rosetta)
- The CASP competition (Critical Assessment of Protein Structure Prediction) regularly puts forward unknown protein test cases
- Folds for proteins or protein domains as big as 120 amino acids can be successfully predicted on a regular basis



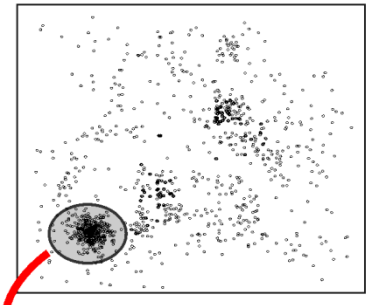
Combination with Molecular Replacement

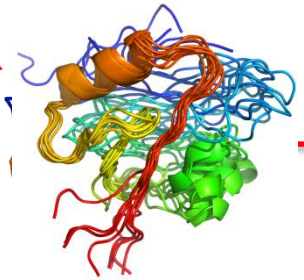
- Recent developments in protein structure solution have sought to exploit the generation of these models for use in molecular replacement
- Computationally intensive approach but can produce MR solutions in cases were no clear homologue exists
- Reduces the need for expensive and potentially hazardous Experimental Phasing experiments
- *AMPLE* is a new initiative funded by CCP4 and the BBSRC to utilise *ab initio* generated search models for MR

Rosetta / Quark generate 1000-2000 "decoy" models for our target sequence



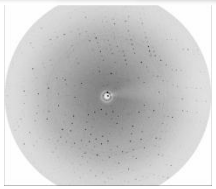
Clustering

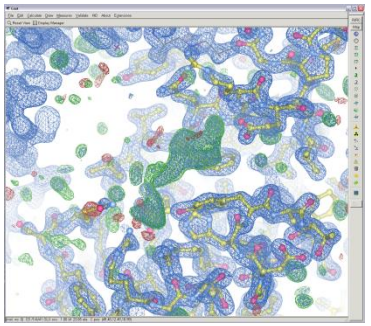




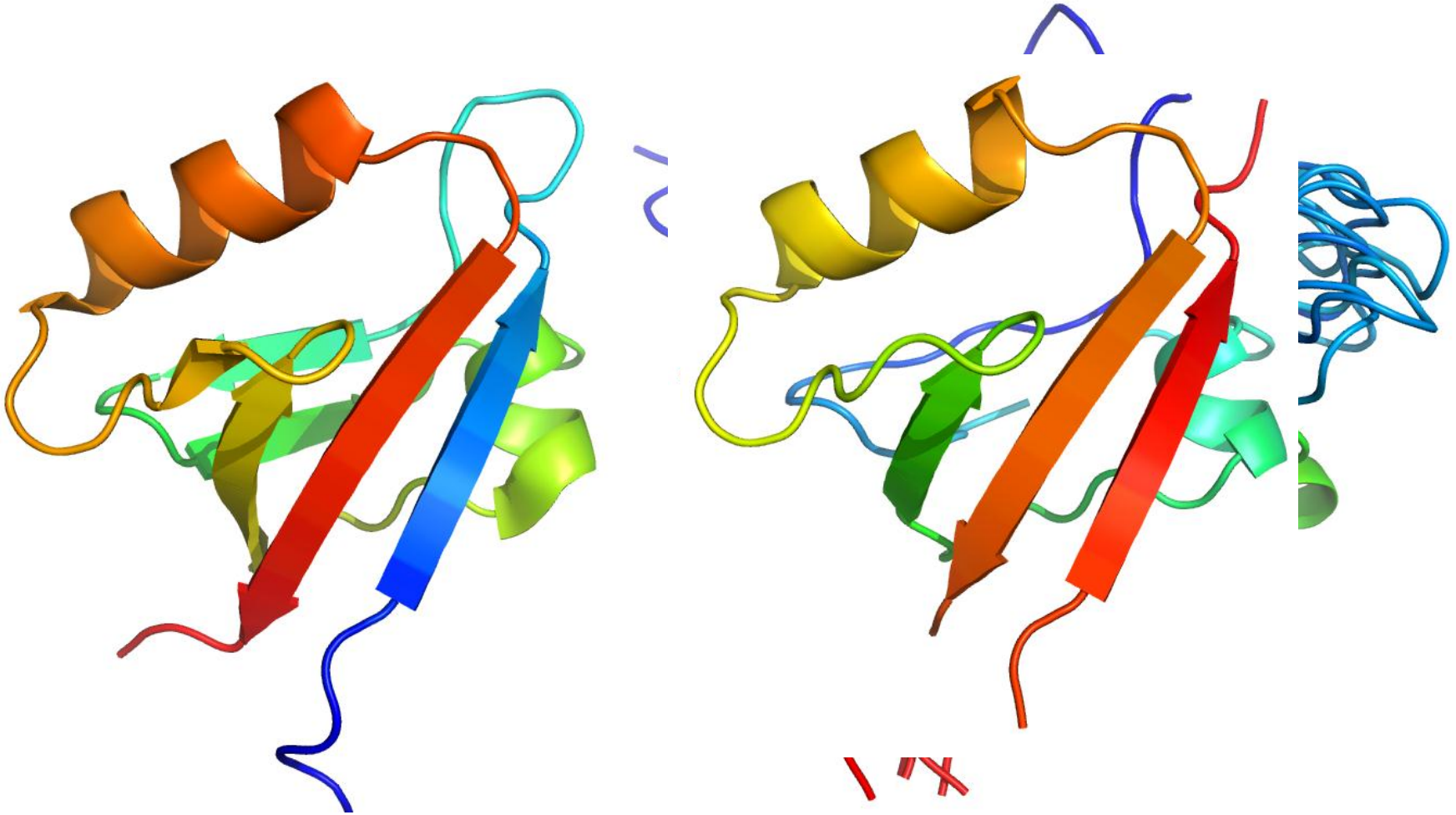
Cluster Centroid

Add side chains and perform molecular replacement with experimental data





Rosetta success (1r6j, a PDZ domain)



crystal structure

Model re-building unrelaxed Rosetta cluster

Model re-building relaxed Rosetta cluster

Summary

- Rapid developments in beamline and detector technology are putting a requirement on CCP4 to develop new, faster approaches to the processing of X-ray data images
- Several factors are also driving the development of improved and automatic downstream software:
  - High-throughput approaches to protein crystallography. Structure solution at the beamline
  - More difficult targets crystals
  - Protein crystallography is becoming a method used by an ever expanding field of users
  - Increasing need for centralised server-based processing facilities – infrastructure required

Acknowledgements

- *CCP4:* Martyn Winn, Eugene Krissinel, Charles Ballard, Andrew Leslie, Fei Long, Garib Murshudov, Andrey Lebedev
- *Biostructx:* David Waterman, Graeme Winter, Gwyndaf Evans
- *Diamond:* Alun Ashton
- *University of Liverpool:* Jaclyn Bibby, Daniel Rigden