Agenda

Meeting title:	CCP4 Working Group 2 meeting
Date:	Wednesday 28th January 2015 **Time:** 11:00 – 16.00
Location:	LMB Cambridge, Francis Crick Avenue, Cambridge Biomedical Campus, Room 2A260
Circulation:	ccp4wg2@stfc.ac.uk
Present:	Jon Agirre (JA), Svetlana Antonyuk (SA), Charles Ballard (CB), Arnaud Basle (AB), Ben Bax (BB), José Brandao (JB), David Brown (DB), David Damerell (DD), Judit Debreczeni (JD), Eleanor Dodson (EJD), Paul Emsley (PEM), Phil Evans (PRE), Mike Hough (MH), Ronan Keegan (RK), Eugene Krissinel (EK), Andrey Lebedev (ALB), Andrew Leslie (ALS), Fei Long (FL), Ed Lowe (EL), Airlie McCoy (AMC), Stuart McNicholas (SMN), Rob Nicholls (RN), Martin Noble (MN), Ray Owens (RO), Harry Powell (HP), Andrew Purkiss (AP), Paul Rowland (PR), Khushwant Sidhu (KS), Ivo Tews (ITE), Jose Trincao (JT), Ville Uski (VU), Sameer Valenkar (SV), Pamela Williams (PW)
Apologies:	Gwyndaf Evans, Vilmos Fulop, Nick Keep, Garib Murshudov, Johan Turkenburg, Keith Wilson

Please arrive for coffee at 11:00, the meeting starts 11:30.
Directions http://www2.mrc-lmb.cam.ac.uk/about-lmb/how-to-find-us/

1. Approval of minutes from the London WG2 meeting 8/10/14
2. Chairs report (Ivo Tews)
3. CCP4 and Bioinformatics - extending capabilities (Ronan Keegan, Ivo Tews)
    a. Ray Owens (RCaH): The OPPF perspective – from protein to construct to crystal
    b. David Damerell (Oxford): The SGC perspective – designing constructs in SATURN
    c. Sameer Velankar (Cambridge): The PDBe perspective – structure mining
    d. Ronan Keegan: History and present capability in CCP4
4. Treating ligands in CCP4 - extending capabilities (Paul Emsley)
    a. Paul Emsley: activities of the ligands group and meetings
    b. Paul Emsley: present capability in coot, lidia and pyrogen
    c. Fei Long: ACEDRG release and updates, demo
    d. Rob Nicholls: analysis and validation of ACEDRG output
5. What's New in CCP4 Core Group, CCP4 6.5 download status, CCP4 Schools (Eugene Krissinel)
6. CCP4 Gui2 (Martin Noble)
7. CCP4 SW 2015 "Phasing" digest and feedback (Ivo Tews)
    a. Judit Debreczini: educating on ligands
8. CCP4 SW 2016 (Ivo Tews)
9. AOB
10. Take note of the date of the next meeting

13:00 – 14:00 Lunch will be served in the LMB canteen

Minutes

1. Approval of minutes from the London WG2 meeting 8/10/14

The minutes from the London WG2 meeting 8/10/14 were approved.

2. Chairs report (Ivo Tews)

WG2 welcomes new members who have signed up; both receiving the mailing list and in attendance. WG2 focuses on "practicing crystallographers" in its constituency; while giving programmers the opportunity to discuss new developments, WG2 thus offers the option to discuss with users / the community.

The Acta Cryst D71 (1) special issue now out and published on CCP4 web pages (http://journals.iucr.org/d/issues/2015/01/00/isscont s.html); we thank all authors who have contributed.

WG2 previously identified the issue of modernizing web pages; this is now followed up by Eugene.

WG2 was represented by the chair at the Exec meetings in Dec 11th 2014 and on Jan 6th 2015. The following proposal was made: that WG2 discusses proposals for the upcoming SW in their October meeting and takes these forward to WG1 for discussion; further, the function of WG2 to help promote Gui2 and providing an alpha testing environment was underlined; Gui2 is represented in this meeting.

The initiatives in the fields of ligands and bioinformatics have come forward and are represented in this meeting.

The next meeting focussing on the current programme developments is Coseners, planned for April 2015; the two main topics for the meeting are Gui2 and Ligands.

3. CCP4 and Bioinformatics - extending capabilities (Ivo Tews)

WG2 wanted to find out what capabilities in bioinformatics users require, and has invited a few speakers to explore developments; WG2 would like to understand whether bioinformatics can be a topic for a SW.

Ray Owens (RCaH): The OPPF perspective – from protein to construct to crystal

The main focus is to provide tools for users to guide construct design, and feed these designs into cloning and expression.

OPTIC database takes information on sequences and target notation from available web resources. (Uses Blast alignment) and presents this information in an accessible way. The primer designs are then rewritten as DNA sequences, and these are presented to a Vector selection tool. There are over 100 vectors to choose from (LIC cloning, against defined set of vectors with defined set of properties); the final selected construct sequences are again stored in the OPTIC DB.

The software has a one-stop approach to use tools that are already available. For the to OPPF targets, it provides a one-to-many relationship; all information is kept in OPTIC database, and relates back to information generated in the laboratory; the experimental results are recorded in **PIMS**; the crystal then links back with all this information. FUTURE needs are:

- Improve visualisation of sequence annotations
- Links between experimental results and data collection results into xtalPIMs
- Group users can apply for support and training
- Not all the tools used are freely available

RO is interested in working with others, there is presently a single developer at OPPF.

David Damerell (Oxford): The SGC perspective – designing constructs in SATURN

SGC is a non profit public-private partnership. David works with Brian Marsden in Bioinformatics. All this distributed to SGC associated sites. The software is all web service driven (web socket).

SATURN is the purpose built database launched at the beginning of SGC – OPTIC equivalent. The SGC also use a PIMS equivalent called **SCARAB** (commercial, by license).

The software is used to simplify and also annotate construct design (HT design approach). In analysis mode, checks are: sequence alignments, domain boundaries, secondary structure prediction, phylogenic trees. It then generates constructs, the **sequence editor** was demoed (calculation of DNA and melting temp etc.) **Data tool** was demoed, example construct plugin.

Implementation uses some third party libraries and requires licenses (e.g. psi-pred). It uses an "obscure" (DD) language called HAXE, but Java script convert is clean and provides an exit route; scripting tools use Python, Java/C++, etc.; easy to add other plugins from new developers. GLmol (slow), IView, ActiveICM (dies), server or stand-alone used to create presentations and annotations, generic visualization tool.

Future needs are viewing and summary tools for sequence annotation and structure.

Sameer Velankar (Cambridge): The PDBe perspective – structure mining

Part of the wwPDB: PDBe, BMRB, PDBj, RCSB-PDB

Sameer described the new tools they are developing for accessing structural information and annotations, to be released soon.

Common problems on **PDB searches** are inaccuracies. How reliable are the data? How complete is the annotation? Missing information in data entries are also a problem. To improve searches, wwwPDB have to **improve the data entries**.

The **clean-up mmCIF** project aims to standardize vocabularies, to provide ligand binding data and connectivity. Problems are often trivial (lower-case and mixed-case problems, protein names, component information and composition information) but time-consuming to fix. All entries will provide information about quality and assemblies (PISA), providing value added data. Extended citation tools will be provided.

Better **sequence search** utilities will be provided later this year, using FASTA or Gene3D structure based alignment (mapping 3D structures onto Cath and Pfam databases). Problems exist with respect to cross-referencing: primary citation (figures, full text); citations where the pdb entry is mentioned. Aim to provide up-to-date information: collaboration with UniProt to have most-up-to-date sequence info; include sequence annotation (variants etc.); Cath & SCOP have been updated; EC numbers are being reshuffled.

Ligand environment enquiries will be extended, to answer, e.g.: Which residues does this ligand bind to? Which ligands bind to these residues?

For MR, the question is what the **best available model** is: coverage (uniprot) and validation.

More non-structural biologists visit than structural biologists who often find the information returned too complicated; thus there is a need for new **visualization tools:** e.g. interactive topology diagrams, sequence viewer, new validation related components

Implementation & future: all data are available through the API (effectively the web pages), and are available to external developers (in Jml/JSmol/PyMOL), plugins; query system – finds relevant data and supports categorization (facets like amazon), finds best structure; use of the unique proteins and grouping (molecules tab); group on compounds or on sequence families; handle larger structures and create a server to provide a selected part (poly-ala, best chain, best environment); integrate access to different servers from one query, e.g. PDB, EMDB, PROSITE, his tags, etc; value added data mainly focused on ligand.

There were several questions from attendees

MN: say you query for a kinase, do you get kinase or all associated domains?

Sequence families / domains at the moment, but … coming

RK: API is Jason

PEM: Do authors update their entries?

… not much!

AB: What is "quality"?

A harmonic average, combination of parameters, resolution dependent, still a problem under discussion!

DB: As a crystallographer, I may have structures that are not to be deposited and still want the analysis -> API over

AB: Good to find more information about structure prior to deposition

PEM: Users appreciate help in deposition. Harvest directory not used?

We will update the harvest file, and can improve validation if we have the information.

AP: Consistency of programs, expanding from co-ordinates only is impossible / difficult, calls for consistency in PDB format.

MN: Quick wins are to collect the state of the information, not using this is wrong; CCP4 should use interim solutions and should help users to navigate the interfaces between different sources of information.

SV: We would require software in ccp4 that generates sequence information.

Ronan Keegan: History and present capability in CCP4.

Using bioinformatics information as tools in the context of MR; presently we have various tools & no clear steer of how to do integrate this – no coordination: some tools have developed, some are distributed form third parties (e.g. ProSMART, PISA, BALBES, … Sculptor, Chainsaw, AMPLE, Mafft, Phmmer, HHPred, ClustalW); generally: sequence information is handled badly, with many format conventions; CCP4MG and COOT also have some sequence related capabilities.

The tools need to be consolidated, and made accessible via GUI2. Implementation is underway in Gui2; bio-pymol -> BioPyton helps.

It would still be required to initiate a sequence coordinated module within ccp4, as an immediate and not too ambitious goal. We may be able to use the library to improve MR and detection to improve success. An analysis should show what tools we have and what we need, then consolidate into wrappers, and expose tools in Gui2. We also need to be able to call web services directly and return information (example - Coot and CCP4MG to wwwPDB). In the longer range we need a wider range, e.g. tools for secondary structure, surface entropy, disorder, to feed back to a construct design pipleline.

MN: Gui2 has one internal held format to distribute to other formats, but there is a significant work package to understanding sequence information, we would like JalViews functionality but not the complexity – does this call for a sequence viewer in Gui2?

EJD: We should look at Gabor's work for PhaserMR, currently only highlighted for PHENIX applications.

4. CCP4 and Ligands (Paul Emsley)

WG2 wanted to find out what capabilities in ligands treatment CCP4 currently offer, and has invited a few speakers to explore developments; WG2 also explores the ligands topic for a SW.

Paul Emsley: activities of the ligands group and meetings

Paul outlined the current position with a Venn diagram. This is on-going work of the Ligands group, and one can sign up to this group by e-mail to Paul. Ligand project in ccp4 presently encompasses Libchk, Prodrg, Acedrg, coot, jligand, pyrogen, lidia and privateer. Two aspects of the current work are detailed here.

Building a new ligand/hetgroup

Input information as a Smile string, or by sketching; **Coot** offers **Lidia; Jligand** has capability to interpret a sketch and find LINKs between groups; the Smiles string is interpreted and the to produce a connectivity list which is fed to an appropriate force-field, encoded in **AceDRG** (ligand restraint generator) and **Pyrogen**, based on good small-molecule chemistry; atom types are recognised depending on bonding type.

Currently working on generation of a set of coordinates for the hetgroup. (See below - Martin in his GUI2 task **Make Ligand** uses **Rdkit** to select a reasonable conformer). When there is an existing similar hetgroup it is necessary to rename the atoms: COOT has a **match_residue_and_dictionary** script to do this.

Validating an existing ligand

Coot has a Blob fitting tool (coot/findligand), which looks at electron density; the REFMAC dictionary lists deviations from the assigned geometry; Coot and CCP4MG both display the current model, and Coot can analyse the hetgroup environment; Jon Agirre has developed tools for validating carbohydrates, and these results can be displayed via the coot interface.

Future needs are: automation, pipelining; what's that blob – interpretation, then do fitting; there is still a big issue over how to handle LINKs – very important for treating glycosylation; exploring different conformer fits against electron density and scoring these (see Martin Noble's presentation), and the scoring needs to take into account: density correlation, bad contacts and an energy score.

EJD: The discussion has identified the need for one central reference database of hetgroup dictionaries

PEM: Eugene's **SRS** is the way to do it.

Fei Long: ACEDRG release and updates, demo

Advantage of using COD – this is an open access database of small molecule crystal structures which has been mined to make a much more reliable set of atom types and chemical geometry. It provides the forcefield for ACEDRG.

Future work: curating and validating COD; more careful checking for chirality, planarity, and hydrogen positioning; extending atomtypes to cover **metals.**

Rob Nicholls: analysis and validation of ACEDRG output, demo

Can we trust the atom types that are generated?

There are three layers: public environment: ccp4 – acedrg, pdb/cif; developers environment: acedrg > acedrg tables (to ccp4) > generation of validation tools; a third aspect may need to be offline (industry).

Independent validation by systematic validation and statistical analysis, working at the level of the acedrg tables

Quality Control now depends on:

- weeding out duplicate structures
- statistical analysis – checking for outliers in bond distributions
- checking chemical sanity.

MN: What is going to impact the accuracy on current ligand refinement – torsions

5. CCP4 Gui2 (Martin Noble)

CCP4i2 now wraps a lot of extended crystallographic capability, and offers logical pipelines to carry out the various work flows (data processing, experimental phasing, molecular replacement, etc.)

The reports are much more accessible, well formatted and provide feed-back and guidance.

It provides a productive environment to do crystallography, and is being used routinely in Martin's group. A workshop in York for experienced and novice crystallographers gave valuable feedback and another is planned for March 5th/6th in York with 10-12 participants from different laboratories.

Martin also reported on his work to extend ACEDRG to provide a better starting model for a hetgroup; all linked, all connected, to refinement / display. Uses the connectivity table to generate a representation as a MOL file, which can then be fed to **Rdkit** to generate many conformers consistent with the restraints provided, and then scores them for minimum energy. While this procedure may not find the lowest energy conformer (especially for large hetgroups), it generates a reasonable starting point for density fitting.

Gui2 timetable for **rollou**t is (about 4 months in total), release planned for autumn 2015:

- alpha as new scientific capabilities scope freeze:1-2 month
- function freeze beta 1 month
- code freeze one further month to fix bugs

An extensive presentation is to follow at the developers meeting at Coseners.

N.B. The talk was unfortunately cut short by time constraints (apologies, ITE).

6. What's New in CCP4 Core Group, The Cloud, Meetings and Summer Schools (Eugene Krissinel)

CCP4-6.5 is out (released 18.12.2014) with new components: privateer validate (Jon/Kevin), Feckless (Phi), Acedrg LibG (Garib), Blend (James Foadi), Crank2 (Raj); several updates on other programs.

The response before the release was good from the CCP4 developers, but the release got delayed while waiting for updates from external packages (e.g. ArpWarp, Coot, Phaser), and it may be necessary in future to convert the release to "by date" rather than "by feature".

Testing is still an issue and requires an equal amount of time (if not more) than preparing a release; repair updates were required almost immediately. There have been the following downloads:

	CCP4-6.5.0	6.5.01	6.5.02
Mac OSX	908	633	87
linux32	821	483	80
linux64	689	424	70
windows	866	141	13

It is planned to phase the linux32 distribution out.

Workshops in Japan and at DLS were popular (4-fold oversubscribed!), well attended and over-subscribed. Ditto the Study Weekend

7. CCP4 SW 2015 "Phasing" digest and feedback (Ivo Tews)

WG2 congratulated Thomas and Airlie for an excellent programme.

It was discussed whether the meeting was in parts repetitive (JD); HP said in teaching repeating contents is advisable to bring the point home.

Representation of (all) present used / offered software packages for phasing (i.e. SHELX, CRANK2, SHARP, SOLVE) was deemed desirable; however there was a clear consensus in WG2 that the motivation for programming the CCP4 SW is not for promoting CCP4 software.

The speakers and organisers had colour badges. PE said it did not make a big difference, but also does not hurt, so we would like to keep this.

The lunchtime bytes were not well organised – some tutors lost their slots, and sometimes their lunches as well! It is hard to stop people running over time, perhaps some (fierce) chairing was needed.

8. Planning for CCP SW 2016, WG2 only (Ivo Tews)

This will be held in Nottingham on January 8-10 2016.

Three topics had been suggested at WG1, viz Bioinformatics, From Crystal to Structure, and Ligand issues.

Paul Emsley and Judit made a presentation on their ideas on how a Study Weekend discussing Ligands could be organised and the meeting accepted this proposal. There has been a great deal of discussion and dissatisfaction in the past over the piecemeal facilities within CCP4 for analysing ligands, and focussing on this topic with a SW should help repair this.

Martin Noble suggested that the topic "From Crystal to Structure" could be built around the science behind GUI2 and this has been noted as a good theme for the following Study Weekend.

9. AOB

None.

10. Take note of the date of the next meeting.

Possible dates are June (suggested dates: 24.6.2015), Locations suggested: London.

Google poll to agree on date and location to follow.