



## CCP4 SOFTWARE SUITE: HISTORY, EVOLUTION, CONTENT, CHALLENGES AND FUTURE DEVELOPMENTS

## CCP4 SOFTWARE SUITE: HISTORIA, EVOLUCIÓN, CONTENIDO, RETOS Y PERSPECTIVAS DE FUTURO

**Eugene Krissinel**

Science and Technology Facilities Council, United Kingdom  
[Eugene.krissinel@stfc.ac.uk](mailto:Eugene.krissinel@stfc.ac.uk)

**Citation/Cómo citar este artículo:** Krissinel, E. (2015). "CCP4 Software Suite: History, Evolution, Content, Challenges and Future Developments". *Arbor*, 191 (772): a220. doi: <http://dx.doi.org/10.3989/arbor.2015.772n2006>

**Copyright:** © 2015 CSIC. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial (by-nc) Spain 3.0 License.

Received: September 12, 2014. Accepted: February 13, 2015.

**ABSTRACT:** Collaborative Computational Project Number 4 (CCP4) in Protein Crystallography is a public resource for producing and supporting a world-leading, integrated Suite of programs that allows researchers to determine macromolecular structures by X-ray crystallography, and other biophysical techniques. CCP4 supports the widest possible researcher community, embracing academic, not for profit, and for profit research. The primary aims of CCP4 include development and support of the development of cutting edge approaches to experimental determination and analysis of protein structure, with integration of them into the suite for worldwide dissemination. In addition, CCP4 plays an important role in the education and training of scientists in experimental structural biology. In this paper, we overview CCP4's 35-year long history and (technical) milestones of its evolution. We will also consider how a particular structure of CCP4 Suite and Collaboration has emerged, its main functionality, current state and plans for future.

**KEYWORDS:** CCP4; Collaboration; Crystallographic Computing; Macromolecular Crystallography; Software Suite.

**RESUMEN:** "Collaborative Computational Project Number 4 (CCP4)" en Cristalografía de Proteínas es un recurso público líder mundial, encaminado a producir y mantener un conjunto integrado de programas que permite a los investigadores determinar estructuras macromoleculares mediante cristalografía de rayos-X, así como por otras técnicas biofísicas. CCP4 va dirigido a la más amplia comunidad científica posible, abarcando investigaciones en el ámbito académico, tanto sin ánimo de lucro como con él. Sus objetivos principales incluyen el desarrollo y soporte de metodologías punteras para la determinación y análisis de estructuras de proteínas, integradas en un conjunto bien definido para facilitar su fácil difusión mundial. Además, CCP4 juega un papel importante en la formación y entrenamiento de científicos en biología estructural experimental. En este artículo, ofreceré una visión de conjunto de la larga historia e hitos técnicos de CCP4 (35 años) y consideraré cómo ha surgido la particular estructura colaborativa de CCP4, sus características más notables, estado actual y perspectivas futuras.

**PALABRAS CLAVE:** CCP4; cristalografía macromolecular; colaboración; computación cristalográfica; conjunto de programas.

## HISTORICAL OUTLOOK

CCP4 was established in 1979 as collaboration between individual research groups in order to facilitate exchange of software tools and new developments, and also to provide a forum for discussing new experimental techniques and solutions in the rapidly developing area of macromolecular crystallography. Obviously, implementation of these goals in pre-WWW era was more difficult and based primarily on meetings in-person and other traditional means of communication. At that time, the PDB contained only 53 structures, and it is truly remarkable that the need of a coordinated collaborative effort was already recognized. Back then, every structure solved was a real research on its own, and every piece of experience gained and every tool developed were invaluable for research community. The project was initially supported by the Science and Engineering Research Council (SERC), which later became the Central Laboratories of the Research Councils UK (CLRC) and now Science and Technology Facilities Council UK (STFC), CCP4 today's host.

In early stages, CCP4 was mainly a repository for various independent programs, developed by practicing crystallographers to aid data processing and structure solution. Most of CCP4 codes from that time were written in Fortran 77. Today, nearly all development is done using C++ and Python, however, Fortran is still used for few on-going projects, and, probably, remains overall dominating in the Suite.

As a collection of independent computer codes, developed in different styles and deposited at generosity by their many authors, CCP4 should have been relatively difficult to use and maintain. In general, structure solution in macromolecular crystallography (MX) is mathematically an ill-posed problem, where experimental data (number of reflections reduced by noise factor) is barely sufficient to recover all atomic coordinates. Because of that, the solution is often reached through unique pathways of computations and data manipulations, resulting in branched scenarios of computational jobs and programs used. Fortunately, it was recognised early enough that efficient operations are possible only if programs communicate via standard data file formats. With this sort of communication, individual programs could be combined in new ways by linking through Unix shell or VMS command languages. This philosophy of CCP4 organisation and use pattern has fully formed by the end of 1980s, with the establishing of MTZ format for reflection data (*mtzlib* library, originally deve-

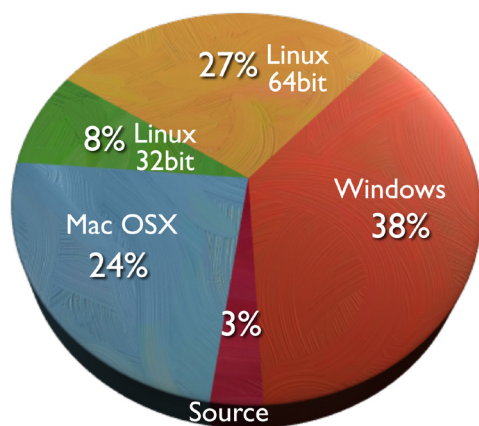
loped by Sandra McLaughlin and Howard Terry in 1989, and then advanced by Eleanor Dodson and Martyn Winn (CCP4 Documentation, <http://www.ccp4.ac.uk/docs.php>), CCP4 MAP format for density map data (*maplib*, David Agard, Phil Evans and John Campbell [(CCP4 Documentation, <http://www.ccp4.ac.uk/docs.php>)] and PDB format for coordinate data (Berman *et al.*, 2000).

Emerging of common formats has naturally resulted in the consolidation of associated functionality. CCP4 core libraries (written in Fortran and C) have mainly formed by the end of 1990s, and the process continued into 2000s, when advanced high-level C++ libraries such as *Clipper* (Cowtan, 2003) and *MMDB* (Krissinel *et al.*, 2004) were developed. These libraries made a foundation for many new CCP4 projects, notably *Coot* (Emsley and Cowtan, 2004), *Buccaneer* (Cowtan, 2006), *CCP4 MG* (Potterton *et al.*, 2002), *SSM* (Krissinel and Henrick, 2004), *PISA* (Krissinel and Henrick, 2007) and others. Other historical milestones for CCP4 projects are the release of Graphical User Interface (*CCP4i*) (Potterton *et al.*, 2003), and development of automatic structure solution multi-component pipelines [*Xia2* (Winter, 2010), *Buccaneer* (Cowtan, 2006), *MrBump* (Keegan and Winn, 2008), *Balbes* (Long *et al.*, 2008) and *Crank* (Ness *et al.*, 2004)]. These developments finalised the formation of CCP4 as a Software Suite with consolidation infrastructure and GUI, where interdependency of various components makes it different from a mere set or repository of independent programs.

CCP4 Software Suite is distributed free for academic and not for profit research, with all source codes open to public. Up until end of 1990s, CCP4 was distributed only for Unix and VMS operating systems, and it was only in 2003 when first Windows release took place. Today, CCP4 is distributed for use on Mac OSX (64-bit), all Linuxes (32 and 64-bit) and MS Windows (32-bit with 64-bit in tests and preparations). As follows from data in Figure 1, by now MS Windows has become a major computing platform for CCP4 users. Table 1 lists to 10 countries in the download statistics for CCP4 release 6.4.0.

Today, CCP4 represents probably the most comprehensive collection of software for macromolecular crystallography. However, the reader would like to be aware that similar functionality is available from other sources, notably PHENIX (Adams *et al.*, 2010), SHELX (Sheldrick, 2008), Global Phasing Ltd. (Global Phasing, <https://www.globalphasing.com>), Arp/wArp (Langer *et al.*, 2008) and others.

**Figure 1.** CCP4 6.4.0 distribution statistics by operating system (from Google Analytics). “Source” relates to source code distributions, where users choose to compile and build the Suite on their systems



**Table I.** CCP4 6.4.0 distribution statistics by country (top 10 listed; data from Google Analytics)

Country	Total Downloads	Download %
United States	3223	21.98%
China	1881	12.83%
Japan	1510	10.30%
United Kingdom	1402	9.56%
Germany	931	6.35%
India	847	5.78%
South Korea	598	4.08%
France	440	3.00%
Brazil	390	2.66%
Italy	320	2.18%

## PROJECT STRUCTURE

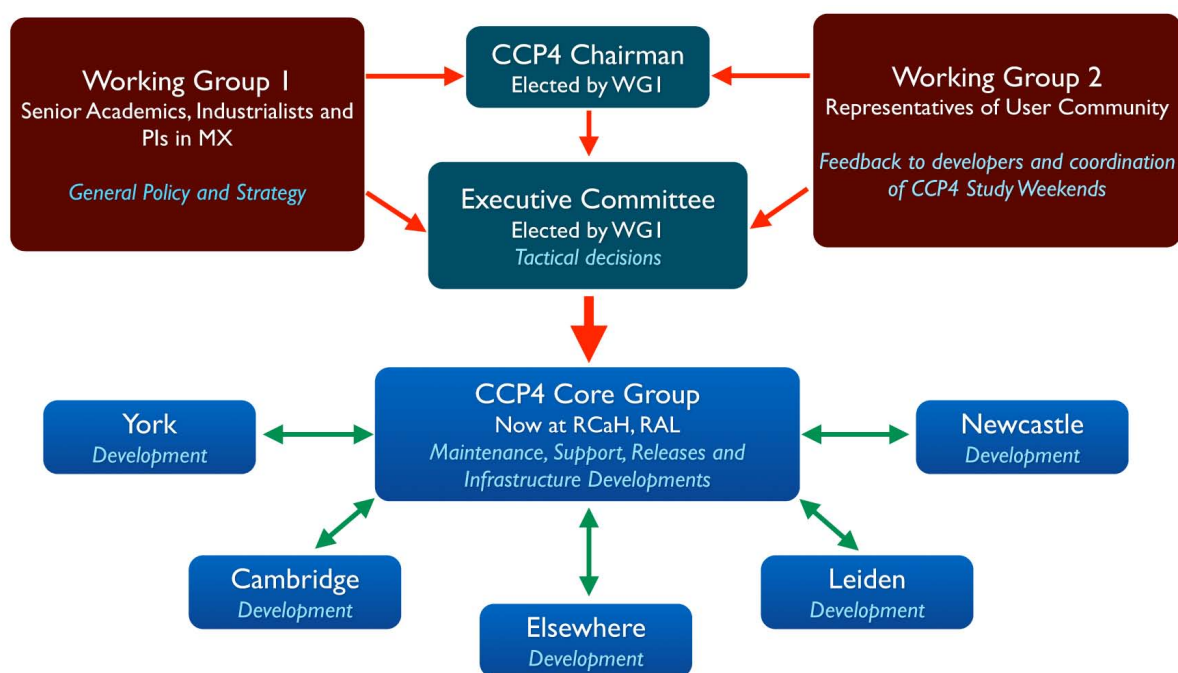
Today, after 35 years of existence, CCP4 represents a mature software project with user base estimated at ca. 20,000 researchers worldwide. It remains a public resource, available on free licences for academic and not for profit research. Although CCP4 always welcomes software depositions from research community, the reality shows that crystallographic computing has now advanced to the

state where competing with established software becomes hardly possible. As a result, CCP4 sees a relatively low level of deposition requests during last decade. Most of active projects represent an on-going development, or are initiated by CCP4 itself. In this situation, it is important to maintain a close link with end users and appropriate project structure in order to achieve most efficient use of resources. In CCP4, this is achieved with the hierarchical project structure (*cf.* Figure 2).

The supreme body in CCP4 hierarchy is Working Group 1 (WG1), which consists of about 130 PIs in the area of structural biology in the UK. WG1 assembles once a year during the CCP4 Study Weekends, which are traditionally held in the first week of January. Every 5 years, WG1 elects CCP4 Chairman, who leads CCP4 Executive Committee, also elected by WG1. Every year, CCP4 Chairman reports to WG1 on current state of the project and takes strategic directions on preferred way of further developments. Working Group 2 (WG2) is an additional body, consisting of 15-20 researchers, working actively in MX, who are willing to provide their feedback on CCP4 current state and developments to CCP4 Chairman and Executive Committee. WG2 meetings normally hold 3-4 times a year. In addition to general feedback, WG2 also suggests topics for CCP4 Study Weekends and elects its organisers.

CCP4 Executive Committee plays central role in current CCP4 management. The committee consists of 10-14 leading academics and 1-2 industrial representatives, all elected by WG1. The Committee meets a few times a year to take reports from CCP4 Core Group and individual developers, funded by CCP4, and also to take all needful decisions and corrective actions, such as initiation of new projects, allocation of resources and recruitment authorisations. One important function of Executive Committee is to prepare CCP4 grant proposals and otherwise look for ways of funding.

Most of CCP4 Executive decisions are realised through the activities of the CCP4 Core Group, located in the Research Complex at Harwell (RCaH) in the Rutherford Appleton Laboratory, UK. The Group is mainly responsible for the maintenance and distribution of CCP4 Software Suite, regular releases and updates between the releases, but also is engaged in software development and outreach activities. Most of coordination with CCP4 external developers, placed in other UK and EU universities and institutions, also goes through the Core Group.

**Figure 2.** Structure of CCP4 Collaboration


#### CCP4 EVOLUTION AND MAIN FUNCTIONALITY

Over years, CCP4 showed a relatively stable growth and advancement. Thus, in 1994, CCP4 included 68 program components (Collaborative Computational Project Number 4., 1994), which increased to about 200 in 2011 (Winn *et al.*, 2011), with many programs retired and obsoleted in the meantime. We shall now give an overview of CCP4 evolution and main functionality. Note that this overview is not intended to be a comprehensive guide on CCP4 Software Suite. Neither does it pretend to describe or even mention the majority of over 200 various software components currently found in the Suite; the most relevant information on those should be looked for in CCP4 documentation.

One can identify few major lines of development of the CCP4 Suite, each one originating from a specific set of reasons and circumstances.

First line of development answers to scientific needs and theoretic advancements. The set of 68 programs from 1994 looks very basic by today's standards. Since that time, CCP4 received advanced data processing software, *Mosflm* (Leslie, 2006); scaling programs *Sca-*

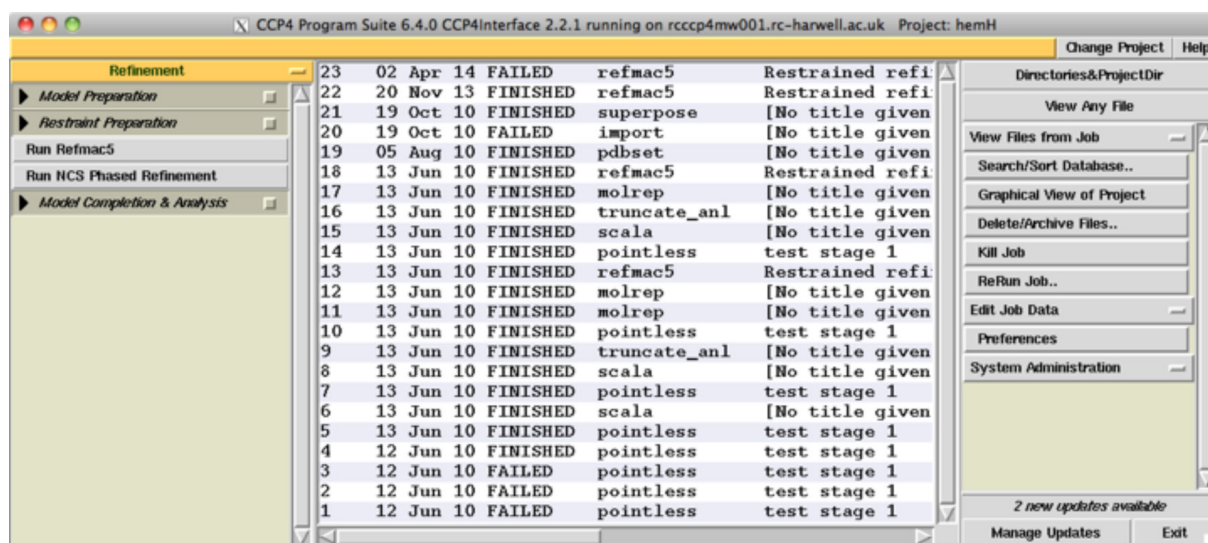
*la*, *Pointless* (Evans, 2006), *Aimless* (Evans and Murshudov, 2013); powerful molecular replacement packages *Molrep* (Vagin and Teplyakov, 1997) and *Phaser* (McCoy *et al.*, 2007); automatic model builder *Buccaneer* (Cowtan, 2006); state-of-the art refinement program *Refmac* (Murshudov *et al.*, 2011); assembly analysis software *Pisa* (Krisinel and Henrick, 2007); and many other programs delivering cutting-edge techniques in crystallographic computing. Owing to better algorithms, CCP4 users were able to cope with increasingly more complex cases, such as incomplete data sets, noticeable radiation damage, poor and low-resolution diffraction, twinned crystals, which, quite possibly, would be dismissed as unsolvable 10-15 years ago.

Second line of development is directed toward improvements in user experience. In 1990s, most crystallographic tasks in CCP4 were indeed performed with shell scripts. For extra convenience, the pattern of input keywords has been standardised, so that users would have to learn the "CCP4 language". This mode of operation was commonly accepted at that time, as most computing jobs were done from command prompt in terminal. Most of CCP4 programs

can still be used in this way. However, the concept of graphic computing has captured nearly all aspects of computations by the end of 1990s and, in answer to this trend, CCP4 released *CCP4i*, a Graphical User Interface (GUI) (Potterton *et al.*, 2003) (see Figure 3). For the first time in crystallographic computing, *CCP4i* has introduced the concept of *CCP4 Project*, which is a rationalised presentation of computational jobs and produced data. In cases, when solving

a structure takes hundreds of computational jobs and many unsuccessful trials, *CCP4 Projects* make it easier to identify the particular way that lead to solution. Another important feature of *CCP4i* is that it suits both experienced user (by uncovering shell scripts it generates, which then can be used as templates in own scripting), and users biased to purely graphical way of computations, as there is no necessity to use the command-prompt mode.

**Figure 3.** CCP4 Graphical User Interface *CCP4i* (Potterton *et al.*, 2003)



The use of GUI has become prevailing in crystallographic computing from around 2005, which only reflects the general trend. Most of young crystallographers are no longer comfortable with working in Terminal; some of them even are not aware of such mode of operation. This situation is particularly prominent among MS Windows users, which make a considerable fraction (see Figure 1). Answering to this trend, CCP4 has undertaken a number of projects on graphical front-ends for various applications. Among most known are *iMosflm* (Battye *et al.*, 2011) for interactive data processing with *Mosflm* (see Figure 4); *Coot* (Emsley and Cowtan, 2004) for interactive model building and completion (see Figure 5); *CCP4 MG* (Potterton *et al.*, 2002) for advanced visualisation of macromolecular structures; *JLigand* (Lebedev *et al.*, 2012) for crystallographic restraints generation; *ViewHKL* (Krissinel and Evans, 2012) for visualisation of reflection data (see Figure 6); *QtRView* (unpublis-

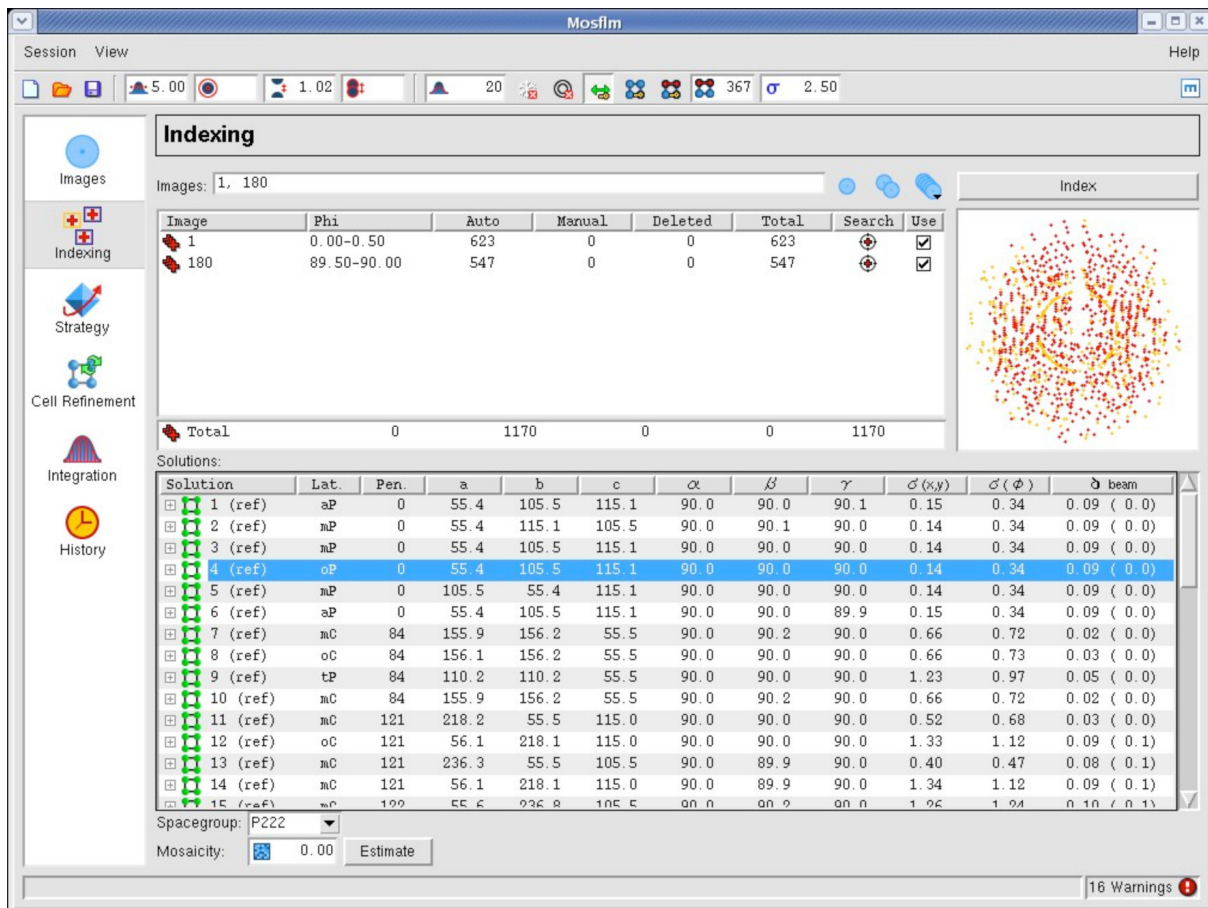
hed), a generic graphical viewer for CCP4 programs' output and log files; *QtPISA* (unpublished), a graphical front-end to *Pisa* software (Krissinel and Henrick, 2007). Providing more graphical front ends in CCP4 is recognised as an important factor for keeping up with younger generation of crystallographers and structural biologists.

Third line of development aims at simplifying the structure solution process for end users by means of advanced automation. Over last decade, MX field attracted many SB researchers not specialised in crystallography. For them, the multi-stage data processing and structure solution with many specialised programs represents a considerable challenge, involving a steep learning curve, many unsuccessful trials and, often, arrival at not best possible results. On another side, considerable number of practical cases may be solved following one of several predefined, albeit

relatively complex, routes. Nowadays, about 75% of structures are solved by molecular replacement, and this number continues to grow, owing to the exhaustion of apparently limited (Chothia, 1992) fold space,

and the ever increasing accumulated knowledge in the PDB. It is therefore possible to automate structure solution in most cases, although the ultimate one-button solution is never reachable.

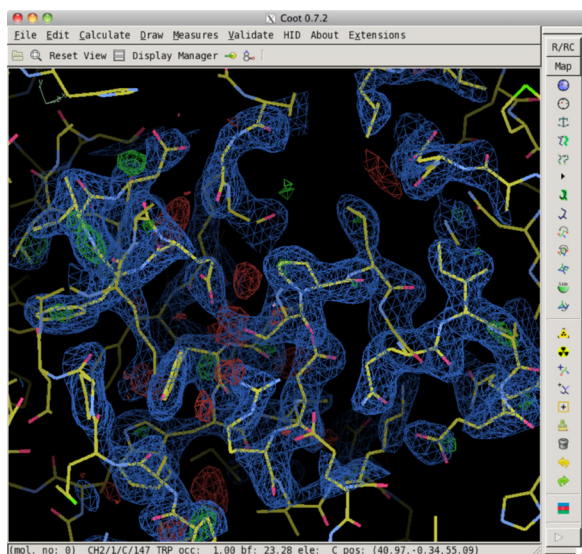
Figure 4. iMosflm GUI.



CCP4 automation covers all principal stages of practical structure solution. *Xia2* (Winter, 2010) is acknowledged as a high-quality tool for automatic data processing, and is used on many beamlines worldwide. This pipeline represents a sophisticated expert system, which can use either *Mosflm* (Leslie, 2006) or *XDS* (Kabsch, 1993) as the basic processing engine with a knowledge base allowing for intelligent decision-making and feedback. The processed reflection data may be attempted for automatic molecular replacement (MR) using either *MrBump* (Keegan and Winn, 2008) or *Balbes* (Long *et al.*, 2008). These pipelines perform

all set of computations, typical for MR, in automatic mode, including initial search for appropriate models and multi-domain fitting. As the primary MR engine, *Balbes* uses *Molrep* (Vagin and Teplyakov, 1997), while *MrBump* can use both *Molrep* and *Phaser* (McCoy *et al.*, 2007). Only basic information: sequence and merged reflection file (e.g. one produced by *Xia2*) is required in most instances. *Balbes* will also take advantage of using multimeric complexes (using data from *Pisa*), if such complexes are likely to be present in crystal. Overall, the success rate of automatic MR is measured at 70% and higher.

Figure 5. Coot version 0.7.2.

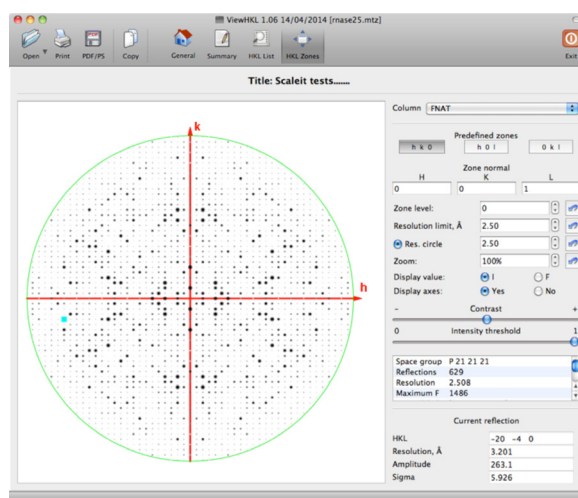


In case of anomalous diffraction, automatic experimental phasing (EP) with *Crank-2* (Ness *et al.*, 2004) may be also tried. This pipeline performs the whole cycle of experimental phasing using same basic information as MR pipelines, plus specification of anomalous scatters. *Crank-2* has multiple regimes of operations and may use either *SHELX* (Sheldrick, 2008) or *Crunch-2* (de Graff *et al.*, 2001) for heavy atom location.

Once phases are successfully calculated with either MR or EP, a coordinate model may be built into electron density. This can be also done automatically using *Buccaneer* software (Cowtan, 2006) (note: *Buccaneer* is used automatically by *Crank-2* and optionally by *MrBump*). *Buccaneer* performs statistical chain tracing by identifying connected alpha-carbon positions using a likelihood-based density target. This approach is advantageous in low-resolution area. In high-resolution limit (up to 3.5 Å), one can use *Arp/wArp* software from EMBL-Hamburg (Langer *et al.*, 2008). Starting from 2013, *Arp/wArp* is co-distributed with CCP4.

Fourth line of development is associated with in-house operations, software maintenance and distribution. Up until 2010, CCP4 used maintenance infrastructure developed in 1990s. In-house setup was based on CVS source code repositories, open for reading but with write privileges restricted to CCP4 developers, staff and close associates. Build script was generated from a single configuration file, which has grown to a considerable size over years. MS Windows version of CCP4 was compiled with MS Visual Studio,

Figure 6. CCP4 reflection file viewer *ViewHKL*



and by 2010, the size of the corresponding MVS project was already critical. The Suite was undergoing 1-2 releases a year, and intermediate changes (bug fixes) would need to be downloaded from CCP4 ftp area and applied manually. The distribution was done in form of tarballs on Unix platforms, and in form of a (rather big) MSI file on Windows.

Software maintenance practices, formed in CCP4 in 2000s, were becoming less suitable with the growing size of the Suite. Particularly weak points were the hardly manageable configuration script; absence of individual builds for program components and libraries; and a cumbersome update procedure, based on personal communication with developers and users. By 2010, most of Core Group's effort was directed on tests, preparation of releases, bug fixes and updates.

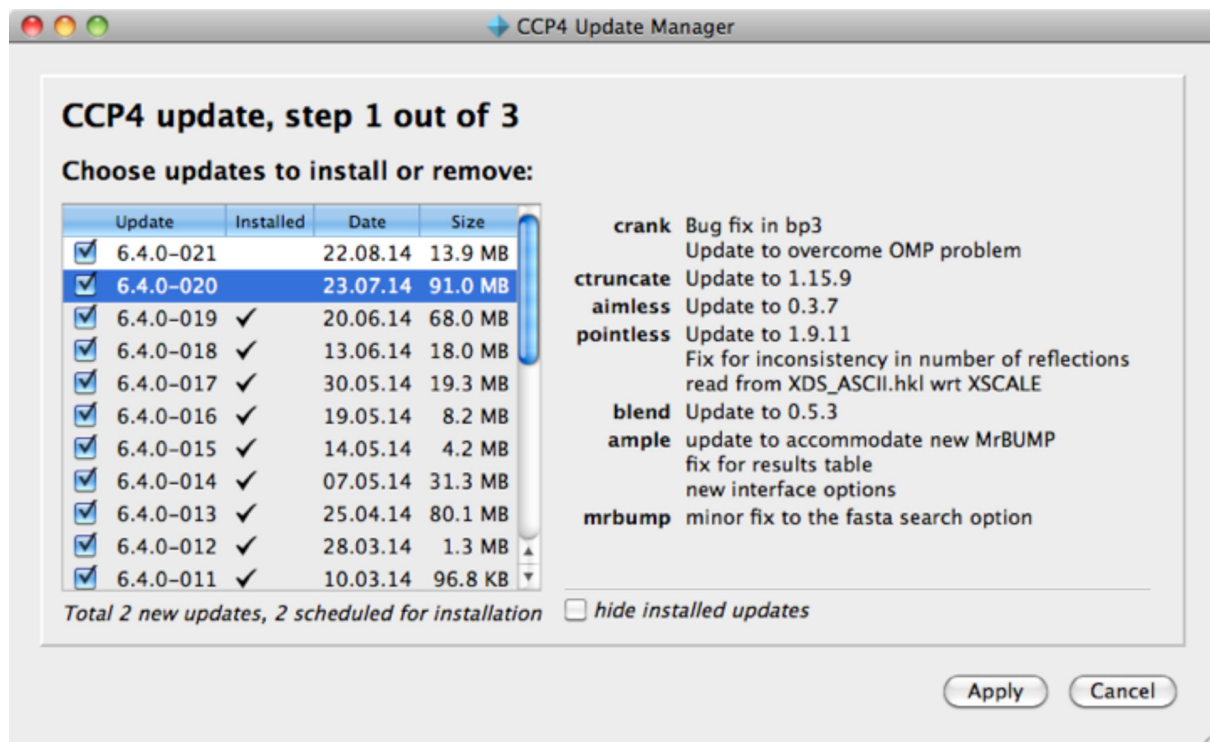
In 2010-2013, CCP4 in-house infrastructure has undergone a complete modernisation. The Suite was organised in *BAZAAR* version control system (Canonical Ltd), which is particularly suitable for use by teams of remotely distributed developers. In new repositories, the Suite was reorganised in a set of projects, each one with own build system, so that they can be built, installed and developed independently of each other. The overall build system was based on *JHBuild* (Linux GNOME Project), which automates source-code checkout from *BAZAAR*, compilation and building. Now, various CCP4 components may be built and installed independently (in case of interdependency of CCP4 program compo-

nents, all necessary dependencies are resolved and built by *JHBuild* automatically). The development environment was complemented with nightly builds and tests on a number of computational platforms, which is useful for spotting bugs.

The distribution mechanism was also modernised by the addition of custom graphical installer *CCP4 Se-*

*tup Manager*. The installer allows a user to configure CCP4 setup, after which download and installation proceeds automatically. A modified version of installer, *CCP4 Update Manager* (see Figure 7), is now used for automatic download and installation of updates, which are used for bug fixes and delivering new versions of software components between the releases.

Figure 7. CCP4 Update Manager



The performed modernisation of CCP4 in-house infrastructure has established a new, considerably more efficient, mode of interaction between CCP4 developers and software life cycle. Now, CCP4 developers work on synchronous set of code, exchanged between their desktops and *BAZAAR* and only inform Core Group when an update of their project needs to be issued. Owing to high level of automation achieved, the maintenance burden of making releases and updates by Core Group has been decreased considerably (by a factor of 3).

Fifth line of development lies in the web application domain. Up until 2007, crystallographic software was mainly used in form of local setups, either on personal desktops/workstations, or on laboratory-wide facilities such as computational clusters. Such setups

were always more convenient and appropriate, by allowing most interactivity and direct access to local data. However, the situation started to change with progressing automation of the Suite and increased use of PDB and other bioinformatics resources, which are mainly available on-line due to considerable volume of data involved. For example, one of sources for getting structural data on small chemical molecules was *MSDchem* tool, available on-line from 2004 (Golovin *et al.*, 2004) (now *PDBeChem* from PDBe at EBI). *SSM* web-server (Krissinel and Henrick, 2004) (now *PDBeFold*) was used for protein structure analysis and homology searches in the PDB, useful for finding search models for MR. Web-server *Pisa* (Krissinel and Henrick, 2007) gives access to the database on macromolecular complexes. *SSM* and *Pisa* were probably



first CCP4 web-application projects, conducted jointly with the European Bioinformatics Institute (the EBI).

First CCP4 automatic pipelines, *MrBump* (Keegan and Winn, 2008) and *Balbes* (Long *et al.*, 2008) presented crystallographic community with a considerable simplification of structure solution process. However, they require considerable computation resources and sizeable databases of search models, which need to be regularly updated for best performance. For example, *Balbes* database contains about 40,000 search models, which occupy ~ 3GB of computer memory in compressed state, and a typical *Balbes* job can take a few hours, sometimes a week, of continuous calculations. It is therefore clear that automated pipelines are well suited for deployment as web application, excusing ordinary users of necessity to have access to additional computational resources and maintain local databases. Currently, CCP4 hosts web-servers for both *MrBump* and *Balbes*, which are used by few hundred researchers and solve about 100 structures every month. The growing popularity of these services proves concepts for putting more CCP4 functionality on the web and future plans on cloud computing.

#### CCP4 OUTREACH AND EDUCATION ACTIVITIES

CCP4 invests considerable effort and resources in the education and training of scientists in structural biology worldwide. The need for such training is apparent and ever increasing, with the increasing number of researchers solving macromolecular structures in course of their studies in situation when MX university courses are not common. CCP4 conducts its outreach and education activities in several different ways.

CCP4 Annual Study Weekend is a large conference, attracting, on average, from 400 to 500 participants, mostly senior PhD students and postdocs, but also all other researchers. Traditionally, Study Weekends take place in first week of January in the UK, usually on a university campus. The meeting consists mainly from lectures, delivered by distinguished academics on chosen aspect or topic of MX, and "lunchtime bites" which include author software demonstration. The topics are chosen by CCP4 Working Group 2 such that all MX course is covered in 5 years, which is a typical training time-frame for a young researcher.

Few times a year, CCP4 conducts International Crystallography Schools and Workshops. They represent intense 5-9 day long practical courses for 20-30

selected students in various locations, chosen such as to suit students from particular regions. A typical student is a PhD trainee with an MX project in final studies, or a young postdoc. Annual events of this type include CCP4 Schools at Advanced Photon Source (Argonne, IL), Japan (Okinawa Institute of Science and Technology, Osaka University, Tokyo University, Fukuoka University and others), South America (Instituto Pasteur Montevideo, Uruguay, University of San Carlos, Brazil), but also in other locations across Europe and worldwide. School programme includes all stages of structure solution process, from data processing to phasing, model building, completion and validation. In addition, if course location adjoins a synchrotron facility (such as APS in Argonne), the school starts from 1-2 days of practical data collection, using crystals from students' own projects. In all cases, students are requested to come with own data, which they struggle to solve in their home labs. The course time is equally split between lectures and practical sessions, where students work on their data. As a rule, a few new structures are solved during each school.

CCP4 sponsors other workshops, conferences and conference sessions, too. This includes well-known meetings such as BCA Summer Schools, iUCR Crystallographic Computing Schools, Gordon Conferences and others.

Last, but not the least, form of CCP4 outreach and education activities is the CCP4 Bulletin Board (CCP4BB), which is a world-wide Internet forum for all MX-related topics, in both experimental and computational areas, MX job advertisements, conference and workshop announcements and similar materials. CCP4BB has ~8000 active subscribers and passes ~100 messages daily.

Figure 8. Rita and Cornforth Award Medal for CCP4



In 2011, CCP4 has received the Rita and John Corforth Award and Medal from Royal Society of Chemistry “for providing a resource that underpins macromolecular structural chemistry worldwide and for exemplar team-ethos over many years” (Royal Society of Chemistry). The award statement explicitly mentions that “the community nature of CCP4 is nowhere more apparent than on its widely used bulletin board. With queries and contributions from crystallographers around the world, from within and beyond the CCP4 team, this resource is probably the most complete repository of knowledge about the theory and practice of macromolecular crystallography”.

### CURRENT CHALLENGES AND FUTURE DIRECTIONS

The 35-year long history of CCP4 presented the Project with a number of challenges, coping with which was an essential part of CCP4 activities and resulted in the formation of CCP4 Software and Collaboration as it is now. Over this remarkably long, for a software project, life time, CCP4 has seen and adopted many changes in both theoretical developments, computing platforms, practices, frameworks and environments, accumulated effect of which represents now a challenge on itself. Periodically, obsolete and no longer supported dependencies in CCP4 need to be revised and replaced. One example is the *BLT* module in *CCP4i*, which is no longer supported and has to be maintained by CCP4.

Evolution of computational platforms represents an ever-going challenge for CCP4. Every new release of OS such as Windows, Mac OSX or a Linux flavour, every new version of compilers spots problems in few-million line size CCP4 code, which needs addressing. Thanks to the new CCP4 in-house infrastructure with nightly builds, most of such problems are now uncovered in the course of in-house builds and tests, but still, many of them are reported by users. CCP4 Core Group maintains a e-mail based help desk, where bug reports and technical problems may be submitted. On average, CCP4 Helpdesk deals with 3-6 reports weekly.

The evolution of computing platforms presents CCP4 also with another challenge in form of users oriented primarily to graphical style of computations. *CCP4i*, developed more than a decade ago, barely answers to modern GUI standards. In answer to this challenge, CCP4 is developing a new graphical interface, GUI-2, which is based on pyQT technologies and serves many purposes. Firstly, it will eliminate dependency on obsolete Tcl/Tk framework, which is already not supported on many platforms. Secondly,

it will harmonise CCP4 Software technically by providing *Python* wrappers for all program components, by this, effectively, converting the whole Suite into a library of *Python* classes. This will simplify scripting and pipelining for CCP4 developers and advanced “old-style” users. Thirdly, it will simplify user actions by better organisation of CCP4 projects and eliminating the need of using explicit file names. Finally, it will enhance user experience by using a modern and advanced graphical framework.

Next challenge, faced by CCP4, is also associated with the new type of user emerged over last decade, with main interest lying in biological aspects of their studies, rather than crystallography. Such a user does not usually have a deep knowledge and experience in practical aspects of crystallographic computations, and looks for automatic or semi-automatic ways to structure solution. In answer to this challenge, CCP4 keeps working on advanced automation. In this field, MR techniques make a good promise and are, in fact, limited only by the set of search models available. Jointly with the University of Liverpool, CCP4 develops a new MR tool, *AMPLE*, based on *ab initio* modelling for MR (Bibby *et al.*, 2012). *AMPLE* uses complex bioinformatics techniques for the rationalisation of hundred thousands models, generated by *Rosetta* (Shortle, Simons and Baker, 1998) or *Quark* (Xu and Zhang, 2012) software, and then using a reduced set of candidates for MR. *AMPLE* has been shown to provide solutions where no other method works, and new structures are being solved with it routinely (see, e.g., Bruhn *et al.*, 2014). However, the method needs further development in part of computational efficiency and increasing the model size.

Another appealing direction in MR, which will be pursued by CCP4 in next few years, is methods for difficult and non-conventional molecular replacement. This will improve efficiency of MR where homology of search model is poor, but additional data may be added from complementing techniques, such as Cryo-EM. The new methods will include advanced post-MR refinement and automated rebuilding of unbiased MR models.

Despite a remarkable progress in MX experimental techniques both in protein production and crystallisation, and X-ray source facilities, the overall quality of diffraction data seems to have a negative trend. This may be the result of researches tackling increasingly more difficult cases, or, given progress in MX software, trying to use data that would be discarded in earlier days. Whatever is the reason, the demand for

MX software working with low-resolution, low-quality, incomplete data is out there. In next few years, CCP4 will pursue a few developments in this direction. Starting from 2010, jointly with Diamond Ltd. and Lawrence Berkeley National Laboratory, CCP4 develops the new data processing software and framework DIALS. The ultimate purpose of this development is to produce the data processing package that would be able to cope with ever increasing volumes and rates of data collection, and developments in experimental methodology, from shutterless, fine-sliced rotation scans through to the thousands of randomly-oriented snapshots of serial crystallography. The new software will also utilise techniques of parallel processing using multiple CPU and GPU machines, facilitating not just speed, but highly accurate analysis based on a comprehensive underlying physical model. Next, CCP4 will conduct research into the analysis-driven data acquisition, which will deliver further automation and optimization of diffraction experiment. In essence, the research will exploit user-consented access to experimental data to systematically explore how early statistics from data analysis map onto the solvability of a structure, using appropriate prior knowledge of the problem in hand. This will result in the identification of the most reliable and robust metrics to use in various decision-making algorithms. The algorithms will be incorporated into a high-level control pipeline and direct feedback to Diamond beamline software and hardware will allow users to optimise their use of synchrotron beam time and if possible complete the structure determination. Another CCP4 endeavour in this group is represented by the development of tools for deriving accurate atomic models of biological macromolecules and their complexes using noisy, limited and incomplete data from multiple crystals. This will include the development of likelihood function for refinement against data from multiple crystals, advanced model for time-dependent radiation damage and maximum likelihood refinement against unmerged data and intensities.

As a common trend, advances in crystallographic software make it more complex, multi-component, using sizeable knowledge and data bases, and demanding more computing power. In most cases, CCP4 software can be comfortably used on a modern desktop or laptop computer. However, such a setup becomes less convenient if substantial data volumes involved or automated tools are frequently used. For example, multi-crystal data collection may result in 200-300GB of raw data, and execution times for CCP4 automated pipelines *MrBump* and *Balbes* are between few hours

and weeks; *ab initio* modeling with *AMPLE* requires access to supercomputing resources. This trend challenges CCP4 to develop a cloud framework for its users, which would provide access to sufficient data storage and computational resources for automatic data solution. Such a development is, indeed, planned. CCP4 Cloud is planned to provide a bridge between data-producing facilities, such as Diamond Ltd., CCP4 Projects, and virtualized computing resources. With this development accomplished, it will be possible for users to have access to their data from any desktop or mobile device, including popular tablet models, to directly import data collected at synchrotron beamlines, or upload data from other sources, into CCP4 Cloud Projects and have them processed in automatic or semi-automatic modes without leaving the Cloud. In difficult cases when automated solvers do not produce satisfactory results, Cloud Projects may be synchronized with ordinary CCP4 Projects on local devices for advanced processing in interactive manual mode.

## CONCLUSION

CCP4 remains a major source of software for macromolecular crystallography, which underpins molecular biology research worldwide. The Project exists for 35 years, a remarkably long time given the pace of changes in methods, computing hardware, languages and frameworks. Over all its history, CCP4 positioned itself as a public resource, devoted to the development and dissemination of novel methods in crystallographic computations, and providing a platform for communication and exchange of ideas and practices. This is reflected in the structure of CCP4 Collaboration, which is carefully designed for providing an efficient control and cooperation with MX research community.

While MX software development, maintenance and distribution remain the main focus of CCP4 activities, considerable effort is applied toward education and training. CCP4 outreach activities are vital for the successful project development as they put software developers and young users in close touch, exposing the real complexity of MX research to all of them.

Nowadays, MX represents a mature field, which is getting characteristic features of a technology, rather than a research field. Despite that, a number of challenges remain to solve, which range from purely technical problems like following the development of computer operating systems, to complex problems of getting best from users data, which is precious in many ways, automation, combination with other data

sources like Cryo-EM, free electron lasers, and many other aspects. The field remains a place for vibrant, active research and is worth attention of talented and enthusiastic young researchers, for whom CCP4 is always on a lookout.

## ACKNOWLEDGEMENTS

CCP4 is supported by Biotechnology and Biological Sciences Research Council UK, currently with BBS-RC Grant BB/L007037/1 "CCP4 Grant Renewal 2014-

2019: *Question-driven crystallographic data collection and advanced structure solution*". Industrial support through CCP4 licensing is vital for the project and makes many CCP4 projects and endeavours possible. Acknowledgements also spread to many thousands of CCP4 users, which make CCP4 Software popular and help keeping it on cutting edge of MX by providing their feedback, use cases and feature requests. The Project could not exist without generosity of software developers who have donated their talent and programs for the benefit of worldwide scientific community.

## REFERENCES

- Adams, P. D., Afonine, P. V., Bunkóczy, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. and Zwart, P. H. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica*, D66, pp. 213-221. <http://dx.doi.org/10.1107/S0907444909052925>
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28, 1, pp. 235-242. <http://dx.doi.org/10.1093/nar/28.1.235>
- Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. and Leslie, A. G. W. (2011). iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallographica*, D67, pp. 271-281. <http://dx.doi.org/10.1107/S0907444910048675>
- Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. and Rigden, D. J. (2012). AMPLE: a cluster-and-truncate approach to solve the crystal structures of small proteins using rapidly computed *ab initio* models. *Acta Crystallographica*, D68, pp. 1622-1631. <http://dx.doi.org/10.1107/S0907444912039194>
- Bruhn, J. F., Barnett, K. C., Bibby, J., Thomas, J. M. H., Keegan, R. M., Rigden, D. J., Bornhold, Z. A. and Saphire, E. O. (2014). Crystal Structure of the Nipah Virus Phosphoprotein Tetramerization Domain. *Journal of Virology*, 88, pp. 758-762. <http://dx.doi.org/10.1128/JVI.02294-13>
- CCP4 Documentation [on line]. Available from <http://www.ccp4.ac.uk/docs.php>
- Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, 357, pp. 543-544. <http://dx.doi.org/10.1038/357543a0>
- Collaborative Computational Project Number 4 (1994). The CCP4 Suite: programs for Protein Crystallography. *Acta Crystallographica*, D50, pp. 760-763. <http://dx.doi.org/10.1107/s0907444994003112>
- Cowtan, K. (2003). The Clipper C++ libraries for X-ray crystallography. *IUCr Computing Commission Newsletter*, 2, pp. 4-9.
- Cowtan, K. (2006). The Buccaneer software for automated model building. *Acta Crystallographica*, D62, pp. 1002-1011. <http://dx.doi.org/10.1107/S0907444906022116>
- Global Phasing Ltd [on line]. Available from <https://www.globalphasing.com>
- de Graaff, R. A. G., Hilge, M., van der Plas, J. L. and Abrahams, J. P. (2001) Matrix methods for solving protein substructures of chlorine and sulfur from anomalous data. *Acta Crystallographica*, D57, pp. 1857-1862. <http://dx.doi.org/10.1107/S0907444901016535>
- Emsley, P. and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallographica*, D60, pp. 2126-2132. <http://dx.doi.org/10.1107/S0907444904019158>
- Evans, P. (2006). Scaling and assessment of data quality. *Acta Crystallographica*, D62, pp. 72-82. <http://dx.doi.org/10.1107/S0907444905036693>
- Evans, P. R. and Murshudov, G. N. (2013). How good are my data and what is the resolution? *Acta Crystallographica*, D69, pp. 1204-1214.
- Golovin, A., Oldfield, T. J., Tate, J. G., Velankar, S., Barton, G. J., Boutselakis, H., Dimitropoulos, D., Fillo, J., Hussain A., Ionides, J. M., John, M., Keller, P. A., Krissinel, E., McNeill, P., Naim, A., Newman, R., Pajon, A., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, G. J., Tagari, M., Tromm, S. Vranken, W and Henrick, K. (2004). E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Research*, 32, pp. D211-D216. <http://dx.doi.org/10.1093/nar/gkh078>
- Kabsch, W. (1993). Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *Journal of Applied Crystallography*, 26, pp. 795-800. <http://dx.doi.org/10.1107/S0021889893005588>
- Keegan, R. M. and Winn, M. D. (2008). MrBUMP: an automated pipeline for molecular replacement. *Acta Crystallographica*, D64, pp. 119-124. <http://dx.doi.org/10.1107/S0907444907037195>
- Krissinel, E. B., Winn, M. D., Ballard, C. C., Ashton, A. W., Patel, P., Potterton, E. A., McNicholas, S. J., Cowtan, K. D. and Emsley, P. (2004). The new CCP4 Coordinate Library as a toolkit for the design of coordinate-related applications in protein crystallography. *Acta Crystallographica*, D60, pp. 2250-2255. <http://dx.doi.org/10.1107/S0907444904027167>

- Krissinel, E. B. and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica*, D60, pp. 2256-2268. <http://dx.doi.org/10.1107/S0907444904026460>
- Krissinel, E. B. and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372, 3, pp. 774-797. <http://dx.doi.org/10.1016/j.jmb.2007.05.022>
- Krissinel, E. B. and Evans, P. (2012). Interactive graphical viewer and browser for reflection data (ViewHKL). *CCP4 Newsletter*, 48.
- Langer, G., Cohen, S. X., Lamzin, V. S. and Perrakis, A. (2008). Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols*, 3, pp. 1171-1179. <http://dx.doi.org/10.1038/nprot.2008.91>
- Leslie, A. G. W. (2006). The integration of macromolecular diffraction data. *Acta Crystallographica*, D62, pp. 48-57. <http://dx.doi.org/10.1107/S0907444905039107>
- Lebedev, A. A., Young, P., Isupov, M. N., Moroz, O. V., Vagin, A. A. and Murshudov, G. N. (2012). JIlgand: a graphical tool for the CCP4 template-restraint library. *Acta Crystallographica*, D68, pp. 431-440. <http://dx.doi.org/10.1107/S090744491200251X>
- Long, F., Vagin, A. A., Young, P. and Murshudov, G. N. (2008). BALBES: a molecular-replacement pipeline. *Acta Crystallographica*, D64, pp. 125-132. <http://dx.doi.org/10.1107/S0907444907050172>
- McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. and Read, R. J. (2007). Phaser crystallographic software. *Journal of Applied Crystallography*, 40, pp. 658-674. <http://dx.doi.org/10.1107/S0021889807021206>
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. and Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica*, D67, pp. 355-367. <http://dx.doi.org/10.1107/S0907444911001314>
- Ness, S. R., de Graaf, R. A. G., Abrahams, J. P. and Pannu, N. S. (2004). Crank: New methods for automated macromolecular crystal structure solution. *Structure*, 12, pp. 1753-1761. <http://dx.doi.org/10.1016/j.str.2004.07.018>
- Potterton, E., McNicholas, S., Krissinel, E. B., Cowtan, K. and Noble, M. (2002). The CCP4 molecular-graphics project. *Acta Crystallographica*, D58, pp. 1955-1957. <http://dx.doi.org/10.1107/S0907444902015391>
- Potterton, E., Briggs, P., Turkenburg, M. and Dodson, E. (2003). A graphical user interface to the CCP4 program suite. *Acta Crystallographica*, D59, pp. 1131-1137. <http://dx.doi.org/10.1107/S0907444903008126>
- Royal Society of Chemistry [on line]. Available from <http://www.rsc.org>.
- Sheldrick, G. M. (2008). A short history of SHELX. *Acta Crystallographica*, A64, pp. 112-122. <http://dx.doi.org/10.1107/S0108767307043930>
- Shortle, D., Simons, K. T. and Baker, D. (1998). *Proceedings of National Academy of Sciences of the USA*, 95, pp. 11158-11162. <http://dx.doi.org/10.1073/pnas.95.19.11158>
- Vagin, A. and Teplyakov, A. (1997). MOLREP: an Automated Program for Molecular Replacement. *Journal of Applied Crystallography*, 30, pp. 1022-1025. <http://dx.doi.org/10.1107/S0021889897006766>
- Waterman, D. G., Winter, G., Parkhurst, J. M., Fuentes-Montero, L., Hattne, J., Brewster, A., Sauter, N. K., and Evans, G. (2013). The DIALS framework for integration software. *CCP4 Newsletter*, 49, pp. 16-19.
- Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. and Wilson, K. S. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallographica*, D67, pp. 235-242. <http://dx.doi.org/10.1107/S0907444910045749>
- Winter, G. (2010). Xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography*, 43, pp. 186-190. <http://dx.doi.org/10.1107/S0021889809045701>
- Xu, D. and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, 80, pp. 1715-1735. <http://dx.doi.org/10.1002/prot.24065>