Over the last decade or so a great deal of the CCDC’s code base and programs have been replaced with newer and more advanced software written in C++ using the CCDC’s C++ Toolkit. It is now the turn of the statistical analysis software within the CSD System to get the upgrade treatment!

The CSD is frequently used for investigation of intramolecular and intermolecular geometric structural parameters as well as other data types such as space group, colour, morphology and unit cell dimensions. The existing program Vista has provided statistical analysis tools for such data since its development in 1994. As the software requirements of users and the CSD System itself have evolved, more emphasis has been placed on 3D visualisation of data and closer interactivity between analysis tools. To provide a more flexible and extensible framework for statistical analysis of CSD data, a new set of functionality has been developed. This set of tools incorporates and extends the functionality previously contained in Vista, and additionally provides a highly interactive interface in which the data spreadsheet, histograms, scatterplots and the 3D visualiser are all inter-connected.

**Features**

**Dynamic presentation of data**

These new statistical analysis tools are available within the popular Mercury interface and this makes the connection between the analysis of data and the 3D visualisation of structural fragments simple and dynamic. All views on the data within the software are inter-connected so you can select a set of points within the spreadsheet and they will be highlighted in any plots and charts that are displayed (see Figure 1).

**Numerical & statistical analysis**

Alongside the facility to generate histograms, polar histograms, scatterplots, polar scatterplots and heat plots there are a range of options for numerical and statistical analysis. Descriptive statistics for distributions such as mean, variance, standard deviation, skewness and kurtosis can be calculated. Tools are also available for performing Principal Components Analysis (PCA) which can aid analysis of multivariate data by transforming the original variables into a smaller set of uncorrelated variables. In addition there are options for determining correlation and covariance between descriptors as well as significance tests to determine whether two samples have the same underlying mean or not.

**Crystallographic options**

A number of crystallography-specific functions are available within this set of statistical analysis tools. This includes the ability to re-format a histogram of hydrogen-bonding angles using the well-known conical correction method as championed by Kroon & Kanters (1975). Another area of importance when dealing with crystallographic data is the subtle issue of topological symmetry. Essentially there are multiple ways of mapping the atoms and bonds of a topologically symmetric search fragment onto the atoms and bonds of each search hit in a crystal structure. An option is presented in the software to combine these multiple parameters and treat them as a single distribution for all plotting and analysis functions.

**How to get it**

The new data analysis features within Mercury are available for all users with a registered copy of the Cambridge Structural Database System. To get the updated version of Mercury please visit the CCDC website at http://www.ccdc.cam.ac.uk/. The new tools are fully documented and there are several tutorials available to illustrate their use. Documentation can be accessed through the program interface or via the CCDC website. If you have any queries or comments about the software, please email support@ccdc.cam.ac.uk.

Dr Pete Wood, Research and Applications Scientist

---

**Figure 1** – Illustration of the data analysis interface including the spreadsheet and two plot types.

**Figure 2** – Scatterplot of H...O distance (Å) against H...O=C angle (°) for alcohol to ketone hydrogen bonds with O-H...O angle (°) shown using a colour scale.
The CCDC played host in February 2011 to the first executive committee meeting of the newly established Young Crystallographer’s General Interest Group of the European Crystallographic Association (ECA GIG-YC, http://gig1.ecanews.org). This group has been recently established to represent the interests, needs and concerns of European Young Crystallographers. As part of its charitable work the CCDC is pleased to be able to support such initiatives for young scientists around the world.

Dr Pete Wood, Research and Applications Scientist

Supporting young scientists in Europe

CCDC have charitable status meaning that all income earned by us must be applied to the benefit of science. As part of this objective we are keen to enable use of the Cambridge Structural Database System (CSDS) in academic institutions that would like to make use of our software, but who have limited or no access to the CSDS. An example of this is the use of the CSDS for training in Information and Communication Technologies (ICT) at the University of Kinshasa in the Democratic Republic of Congo, a project named Science Training Laboratory. Our invitation to take part in this initiative came from Professor Zéphyrin Yav also from the University of Kinshasa.

The University of Kinshasa has poorly equipped laboratories thus it is impossible to teach chemistry using practical hands-on experiments. The use of ICT in teaching has great potential for the university; simple experiments can be carried out and complimented with ICT computer simulated experiments. This not only reduces costs but also enables the knowledge to be disseminated to a larger number of people. In addition to other software providers, CCDC have been lucky enough to be involved in Science Training Laboratory. We have provided software and teaching materials for free and these resources have been used in training 60 academics, 100 secondary school teachers, 900 students and 450 pupils on the use of the CSDS. In addition researchers in the university's Chemistry Department whose work includes isolating new molecules from Congolese medicinal plants have also been trained in searching the CSD. Furthermore a number of examples developed using the CSD (e.g. 3D visualisation, introducing fundamental structural concepts) have been included in revised courses for science and medicine masters and graduate students.

Science Training Laboratory is a collaborative project between Kinshasa University (Democratic Republic of Congo), Cologne University (Germany), Catholic University Leuven (Belgium) and Cambridge Crystallographic Data Centre (United Kingdom). It is jointly funded by the British Council and the DElPHE programme.

Dr Susan Henderson, Marketing and Communications Scientist

Over the coming months you can come and meet us at various meetings and events around the world.

<table>
<thead>
<tr>
<th>Date</th>
<th>Conference, meeting or event</th>
<th>Venue</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>28 May 2011</td>
<td>ACA</td>
<td>New Orleans, LO, USA</td>
<td>Exhibition, Talks</td>
</tr>
<tr>
<td>5 Jun 2011</td>
<td>9th International Conference on Chemical Structures</td>
<td>Noordwijkerhout, NL</td>
<td>Talk</td>
</tr>
<tr>
<td>26 Jun 2011</td>
<td>ICCOSS XX</td>
<td>Bangalore, India</td>
<td>Attendance</td>
</tr>
<tr>
<td>17 Jul 2011</td>
<td>Gordon Research Conference: Computer-aided drug design</td>
<td>West Dover, VT, USA</td>
<td>Talk</td>
</tr>
<tr>
<td>22 Aug 2011</td>
<td>IUCr Congress</td>
<td>Madrid, Spain</td>
<td>Exhibition, Talks, User Forum</td>
</tr>
<tr>
<td>28 Aug 2011</td>
<td>Fall ACS</td>
<td>Denver, CO, USA</td>
<td>Talk</td>
</tr>
</tbody>
</table>

Supporting the scientific community

ECA GIG-YC committee members Peter Wood, Laura Roces and Alexander van Driessche at the CCDC.

Professor Zéphyrin Yav pictured on his recent visit to CCDC.
Protein crystallography –
getting your ligands right

One of the ironies of protein crystallography is that whilst the ligands usually represent the main reason for solving a structure, they are also the most difficult part of the structure to solve. The reason for this is that protein crystals commonly diffract to around two Angstroms. This means that one is not able to determine the positions of individual atoms from the electron density (see Figure 1). In order to be able to create atomistic models from electron densities protein crystallographers make use of restraint libraries. Proteins, consisting only of 20 amino acids, have well established restraints based upon data derived from the Cambridge Crystallographic Database (CSD)\(^1\). Restraint libraries for ligands, which can have much greater chemical diversity than proteins, are less trivial to come by and are often generated manually. In these cases the quality of the ligand restraints libraries depends on the person generating them.

Mogul\(^2\) is a tool that can be used either to validate ligand geometries or to help generate ligand restraints. Mogul provides easy access to information on the preferred values of bond lengths, valence angles and acyclic torsion angles, using data derived from the CSD. The preferred geometric values of a molecule supplied in either 2D or 3D format can be obtained by the click of a button. Mogul searches are very fast and in cases where the CSD does not contain exact hits searches are automatically generalised to obtain information from fragments with similar chemistry.

Internally, we have checked the quality of 100s of ligands deposited to the Protein Data Bank (PDB) using Relibase\(^+\), Mogul and the Uppsala Electron Density Server (EDS). A conservative estimate from this study is that roughly 20% of ligands in the PDB have geometries so severely distorted that they cannot be rationalised by the protein straining the ligand. To remedy this situation the CCDC is collaborating with the PDB in their efforts to offer depositors improved validation procedures during data deposition.

Although it is useful to be able to validate the geometries of ligands, one would prefer to get them correct in the first place. One of the reasons for creating Mogul was to help protein crystallographers generate better ligand restraints libraries. Recently Global Phasing\(^3\) have released their Grade\(^4\) software which uses Mogul to automatically build ligand restraint dictionaries. As part of validating Grade Global Phasing have compared restraints dictionaries produced for amino acids to the Engh & Huber\(^5\) protein restraints and the results show an impressive correlation. In addition to the collaboration with Global Phasing CCDC are also collaborating with CCP4\(^9\) who are working on an interface to Mogul in Coot\(^6\). The Mogul interface in Coot will be based on the work by Judit Debreczeni at AstraZeneca, where the Mogul results are visualised in the context of the protein structure and the electron density. Bond length and angle values in the refinement dictionary are then updated based on values suggested by Mogul. AstraZeneca are not the only pharmaceutical company making extensive use of Mogul. At GlaxoSmithKline protein crystallographers have been making use of Mogul to help them with quality assurance of the small molecule parts of their structures for several years.

Mogul is a very well established and validated piece of software. It has been used by academic and industrial users for nearly ten years. Now is the ideal time for the protein crystallographic community to reap the benefits of it.

Dr Tjelvar Olsson, Research and Applications Scientist

References
8. O. S. Smart el al. Grade, Global Phasing Ltd. Cambridge, United Kingdom, 2011
What is the CSD System?
The CSD System is the name given to the Cambridge Structural Database (CSD) and its corresponding suite of search, retrieval and analysis tools.

I have started using the CSD System but I seem to have the choice of two search tools (ConQuest and WebCSD): why do I need both?

Although they are both search tools, ConQuest and WebCSD are subtly different; ConQuest is a locally installed desktop tool for searching the CSD while WebCSD is the online portal to the CSD.

When would you suggest I use WebCSD?
WebCSD is best suited to simple searches that require rapid results e.g. to assess whether a newly solved crystal structure has been previously determined. Available as either a public or private server, WebCSD’s ease of deployment means it can be easily accessed anywhere. Its simple interface makes WebCSD an ideal tool for teaching: no prior knowledge of the interface is required to make good use of the software. WebCSD is a perfect ideas generator.

And when is it most appropriate to use the locally installed tools?
ConQuest provides a comprehensive range of text, numeric and geometric search functionality, including the ability to define intermolecular geometry searches. All types of searches (e.g. text, numeric, substructure and geometric) can be combined in a Boolean fashion, so that highly specific search criteria can be satisfied. If you have a geometric search to run, or a combination of different search criteria to satisfy then ConQuest is the best search tool.

What can I do with WebCSD that I can’t do with ConQuest?
As a web-based tool WebCSD can be easily kept up to date, both in terms of the software and the database it searches on. WebCSD receives regular automatic data updates and remains the only way to access the latest pre-edited “CSD X-Press” entries. Another unique feature of WebCSD is the ability to run molecular similarity searches, i.e. to find similar CSD structures to a query molecule.

What are WebCSD’s limitations?
3D searches are not yet possible. Currently there are no query or hitlist management tools, and no advanced analytical tools. The visualisation features provided in the embedded visualiser are fairly limited (however users can view structures in Mercury, if installed locally).

So what does the future hold for WebCSD?
We intend to start expanding functionality, starting with 3D searching. Our aim is to eventually offer all the key functionalities provided by our desktop software.

For further details about WebCSD and the CSD System, please refer to our website www.ccdc.cam.ac.uk

Mr Ian Thomas, Scientific Software Developer