Small and Large Molecule Structure Databases Under the Same Roof

This special edition of the CCDC Newsletter celebrates the opening of the CCDC’s first office in the USA. The New Jersey based operation is enthusiastically endorsed this unprecedented public-private partnership between the preeminent global data resources for archiving small and large molecule experimental structures.

Today, CCDC and RCSB PDB scientists are busy strengthening existing collaborations and working towards development of new tools for understanding how drugs act on their protein targets. The move to Rutgers also gives CCDC applications scientists access to considerable expertise in materials science represented within the School for Arts and Sciences and the School for Engineering.

Professor Helen Berman, Director of the RCSB PDB, and Dr. Colin Groom, Executive Director of the CCDC, first began discussing the possibility of the CCDC coming to Rutgers at the American Crystallographic Association meeting in Hawaii, last year. The Center for Integrative Proteomics Research is a new 75,000-square-foot facility dedicated to fostering interdisciplinary structure-function studies of complex biomolecular phenomena. Center members include internationally recognized Rutgers faculty, leading research groups focused on computational chemistry, structural biology, mechanistic enzymology, and bioinformatics. In addition to housing the RCSB PDB and the CCDC, Rutgers Proteomics serves as the headquarters of the BioMaPS Institute for Quantitative Biology, with its interdisciplinary Ph.D. program in Computational Biology and Molecular Biophysics, and operates core facilities for NMR spectroscopy, mass spectrometry, and cryo-electron microscopy.

I would like to join Helen and everyone at the Center for Integrative Proteomics Research in welcoming the CCDC to Rutgers and to looking forward to more exciting collaborative research with the CCDC at the interface of structural chemistry and biology.

Stephen Burley currently serves as the Director of the Center for Integrative Proteomics Research and as a Professor in the Department of Chemistry and Chemical Biology and at Rutgers, The State University of New Jersey. He is also a Member of The Cancer Institute of New Jersey. Stephen previously served as a Distinguished Lilly Research Scholar in Lilly Research Laboratories and before this was the Chief Scientific Officer of SGX Pharmaceuticals, Inc., was the Richard M. and Isabel P. Furlaud Professor at The Rockefeller University, and an Investigator in the Howard Hughes Medical Institute. Stephen received an M.D. degree from Harvard Medical School and a D.Phil. in Molecular Biophysics from Oxford University, leading to post-doctoral work with Gregory A. Petsko at the Massachusetts Institute of Technology and William N. Lipscomb at Harvard University.

The Center for Integrative Proteomics welcomed the CCDC to Rutgers with a lunchtime mixer session. Pictured left to right are Stephen Burley, David Kimball (Associate Vice President for Translational Science Research and Professor at the Ernest Mario School of Pharmacy, Rutgers), Paul Davie (General Manager of CCDC Inc, USA) and Colin Groom (Executive Director of CCDC, UK)

The outside of the new Center for Integrative Proteomics building at Rutgers University, which houses the RCSB PDB. The sculpture is called “Synergy” and is based on the structure of collagen that the Director of the RCSB PDB, Prof. Helen Berman, has worked on extensively.
The aim of CCDC Inc. is the delivery of closer interactions and meaningful scientific collaborations with our many users and partners in the region. Here Dr. Terry Stouch provides an expert view on value that the CSD brings to drug discovery – a taster from a series of articles that he will be developing with us over the course of this International Year of Crystallography.

It goes without saying that three-dimensional molecular conformation sets the stage for efficacious drug / protein interactions and consequently is key to effective drug design. Unfortunately, the calculation of small molecule conformation is often thought to be a solved problem. It often appears – at least to the naked eye - to be easily determined with any of a large proliferation of molecular modeling programs. In fact, thanks to decades of research, some of the more advanced force fields can provide routinely dependable results. However, that is not always the case.

Drugs and other bio-active and naturally occurring molecules can be functionally complex. Not all chemical functions are easy to model. This can be true even when the functional group is isolated. The difficulty increases when multiple functional groups are brought together in close association as is often found in bioactive molecules. Another problem for modeling is that bioactive molecules do not necessarily interact with target biomolecules in their lowest energy state; higher energy conformations can be important. It is not enough for force fields to determine the lowest energy conformations; they must also accurately determine the higher energies that occur as a conformation climbs a torsional profile from low energy to the maximum barrier energy. If torsion angles are not properly determined and the conformation is not that of the bioactive state, then the interactions between the drug and the protein will not be properly represented and the interaction energies between the two will be in error. Even small changes in energy are important. The difference between a good compound and a useless one might be as little as a kcal/mol. In that torsional barrier can be several kcal/mol and that the torsional profile can go from lowest energy to the barrier maximum in as little as 60 degrees, even a small error in a torsion angle can be problematical. 

I’ve found the Cambridge Structural Database (CSD) to be invaluable for helping me understand molecular conformation and for validating force field results. Within the hundreds-of-thousands of crystal structures housed in the CSD, often a particular functional group or combination of functional groups might be found hundreds of times. This allows the construction of empirical torsional profiles that can provide probabilities of occurrence and, consequently, relative energies of different conformations. Energetically reasonable conformations can be assessed, not just that of lowest energy. An important aspect of this data is that the structures are from condensed phase and variations in torsion angles are often seen when different crystal packing environments are sampled. This gives information on the pliability of the particular torsion. Crystal packing interactions and their impact on molecular conformation is much more indicative of a protein/drug complex than is modeling in vacuum.

In fact, I do routinely rely on force fields. However, when I question their results or spot a potentially problematical functional group or a tricky combination of functional groups, I go to the CSD to see what real data teaches. Often the experimental data validates the force field, providing additional reassurance. Otherwise, I rely on the CSD data when I set the conformations for my drug design studies. Although force field validation and conformation determination have been my principal use of the CSD, this database has lots more to offer, particular in terms of understanding atomic interactions. The X-ray crystal structures of small molecules contain a wealth of information valuable to drug discovery research. Unfortunately, it has been underutilized by the drug design community. Drug discovery research would benefit from routine incorporation of this data into its decision making. The chemical software industry would provide a substantial service by providing convenient access to the CSD data.

Terry R. Stouch, PhD
Science For Solutions, LLC
touch@gs.com

Terry Stouch, PhD, has 30 years in drug discovery research in both large pharma and biotech during which time he helped place 9 compounds into human clinical trials. He consults for large and small pharma, chemical software companies, and molecular and biomolecular database initiatives on issues of drug design, molecular property prediction, molecular and biomolecular structure, pharmaceutical data evaluation and modeling, and scientific software design. He is Senior Editor-in-Chief of the Journal of Computer-Aided Molecular Design, current Chair of the Computers in Chemistry Division of the American Chemical Society, and Fellow of the AAAS and IUPAC.
Did you know there is a fundamental difference between a CCDC deposition number and the associated refcode? The deposition number is a direct reference to the specific files, exactly as deposited. The DOI links back to these freely available files. Data extracted from the deposited files is further validated and curated to prepare an entry for use with the various CSDS and WebCSD tools. The refcode identifies this entry.

Roughly 60,000 new structures will be assigned deposition numbers this year and throughput is on the rise. Moreover, there are many more deposition transactions annually due primarily to pre-publication corrections by depositors and to post-publication reconciliation.

Did you know that one of the most common reconciliations involves the author list? The manuscript’s author list mightn’t be precisely known at pre-publication deposition. The principle investigator’s name should be included with any deposition. It makes reconciliation easier for all involved.

Even a flawless deposition may be complicated and labor intensive to curate, particularly if it is a complex and/or precedent setting structure. Even naming can present challenges. A well-chosen name can convey chemical nuances that an auto generated name may lack. Time permitting, an Editor might sift through the associated published manuscript, search the literature, or otherwise generate a better name. According to Editor Seth, it is helpful when authors populate the _chemical_name_ systematic field of the deposited .cif. It makes the curation process more efficient, plus the scientist has an opportunity to help mold the conventions of nomenclature.

Preparing the best possible, error-free and properly documented .cif is an important part of our symbiotic relationship. It’s nice to know that Deposition Coordinators will reconcile author lists and Editors will name every structure. But if we prepare our depositions correctly, they shouldn’t need to. Given the opportunity costs for human intervention, I want my CCDC collaborators spending their time guarding against the unnoticed natty issues that were beyond my expertise. I want to use my last line of defense wisely.

Greg Ferrence is Professor of Inorganic Chemistry at Illinois State University in the USA and recently spent a three month sabbatical at the CCDC in the UK.

CSD entry and private communication MITGUT, deposited with the CSD by Prof. Ferrence.

2014: The International Year of Crystallography

The crystallographic community is buzzing due to the fact that the International Year of Crystallography (IYCr) is now upon us! Crystallography is at the heart of everything we do at the CCDC so it will come as no surprise to you that we are official partners of the IYCr.

We were delighted to be involved in the Opening Ceremony of the year, held at UNESCO headquarters in Paris, France, on 19th and 20th January, in front of more than 800 attendees. A special highlight for us was the fact that our own Dr Juliette Pradon (pictured) gave a presentation at the ceremony, talking about our collaboration with the University of Kinshasa in the Democratic Republic of the Congo, Africa.

Looking forward to what’s left of 2014, we have many events planned including attendance at many major crystallographic conferences where cutting edge research being carried out at the CCDC will be presented and workshops on the CCDC’s software products will be provided. We look forward to seeing you at one of these events, or speaking with you at our stand at one of the many associated exhibitions at these events, see: http://www.ccdc.cam.ac.uk/events. Learn more about the International Year of Crystallography at http://www.iycr.org.

Juliette presenting at the IYCr Opening Ceremony.
Crystallography in Chemical Education

Judith Currano MSc. is the Head of the Chemistry Library at the University of Pennsylvania, and has recently joined the CCDC’s Board of Governors. Here, as a librarian and educator, she gives us a taste of the work she does instructing and helping students to get the best and most efficiently derived results from their searches of the Cambridge Structural Database.

I introduce the Cambridge Structural Database to all of the first-year inorganic and organic chemistry graduate students during a course in chemical information that is required for the University of Pennsylvania’s PhD in Chemistry. Being an expert in chemical information and not in crystallography, I aim to teach my students three basic skills: to use the CSD text search options to retrieve the crystal structures of substances published in known journal articles; to retrieve a list of crystal structures having specified structural elements; and to determine structural parameters of substances, such as bond lengths, angles, and distances from specified points, as well as methods of statistically analyzing these parameters using Excel and Mercury. The students will go on to use the database in a variety of ways; some will only retrieve the crystal structures of known substances, while others are more interested in general features of structure in substances containing particular elements or structural features. Upon joining a research group, many will learn more in-depth skills tailored to their specific field of study from their faculty advisors and advanced students and post-doctoral fellows in the lab.

By the time I introduce CSD, students in my classes have already spent a large amount of class time learning and practicing substructure search techniques, and they are able to easily adapt these skills to CSD searches. Given the nature of the material in CSD, bond types are much more specific than in most of the other databases the students encounter; for example, they must determine whether a substance is liable to be represented with pi or delocalized bonds, rather than just single, double, or triple bonds. For this reason, I encourage them to make frequent use of the “any bond” option or to determine how a similar, known substance has been represented in the database. Overall, they are pleased with the degree of specificity that they can introduce in their substructure searches, particularly for organometallic species and coordination compounds, which tend to be represented inconsistently through other information sources.