RIMS
RESEARCH INFORMATION MANAGEMENT SYSTEM

Projects

Action Due Now
All My Projects

Project Detail

Proposal
Award
Progress
Lead PI Info
Msg to Proj Mgr

Other Items

FAQs
My Preferences
Change Password
Logout

Help

Award Information - ER63601-1021466-0009501

ID: ER63601-1021466-0009501
Principal Investigator: Herbert J. Bernstein 631-244-3035
Co-PIs:
Institution: Dowling College
Title: Local System Support for PDB Biological Unit Search and Display
SC Division: SC-23.2
Program Manager: Roland F. Hirsch 301-903-9009
Research Areas:

Project Progress

Most recent report of results to date:

The BIOMOL grant is for “Local System Support for PDB Biological Unit Search and Display” to augment Rasmol's [Bernstein 2000] [Sayle, Milner-White 1995] existing macromolecular display functions with new capabilities by taking advantage of recent increases in local computing power in order to move functionality that is now scattered among various local and remote systems into one local package. The grant has been funded for three periods, 1 Sep 2003 -- 31 August 2006 (Period I), 1 Sep 2006 -- 31 August 2009 (Period II) and 1 Sep 2009 -- 31 August 2011 (Period III). Note that Period III consists of an original year of funding and a one year no-cost extension. Period III is the final period.

Progress in Period III (2009 -- 2011):

The major focus of the work for Period III is on issues relating to archiving of diffraction images, e.g. by integration of CIF and NeXus software. In addition work is continuing on support for identification of complexes. We are pleased to report significant progress in both areas. In October 2009, in collaboration with our colleagues in the BIOIHDF group, we published “Unifying Biological Image Formats with HDF5” in CACM, 52:10, 42-47, reprint appended. Following up on that, in January 2010 we were able to demonstrate interoperable representation of imgCIF data in HDF5 files via NeXus. We demonstrated that capability at the “HDF5 as hyperspectral data analysis format,” Workshop, ESRF, Grenoble, FR, 11-13 January 2010. The images in those files could be read and displayed directly by HDF5 tools without the need for back-conversion. This has resulted in a reconsideration by the NeXus group of how best to handle issues such as axis representations to avoid the need to repeat the discussion already resolved in the development of imgCIF. Those discussions continue.

A new tiff2cbf utility has been added to CBFlib, allowing diffraction data from additional classes of detectors to be processed by XDS, which reads imgCIF data.

The library has also been modified to facilitate adaptation to the new method-based approach to CIF being pursued by the IUCr.

CBFlib is now a robust and fully accepted infrastructure component for support of synchrotron diffraction images, with API interfaces for C, Fortran, Python and Java. CBFlib is well-accepted as a Debian package and we have made progress in acceptance by the Fedora project. The only remaining issue for use by the Fedora project is that CBFlib includes James Hester's PyCIFRW package, and that causes a license conflict for the Fedora project.

The code done thus far has been released, and a further update is expected at the close of Period III.

In the thread of research relating to complex identification, in April 2010, our paper (with P. A. Craig) on "Efficient molecular surface rendering by linear-time pseudo-Gaussian approximation to Lee-Richards surfaces (PGALRS)," appeared in J. Appl. Cryst., V. 43:2, 356-361. A copy of a preprint of the paper is appended. That work has continued with a collaborative effort with L. C. Andrews on improvements to the NearTree algorithm, both in the form of a more balanced search and more aggressive balancing of the tree that have resulted in significant performance improvements in the timing both of identification of atoms in proximity to a surface and in rendering of that surface. One of the important issues in understanding complexes is correctly and efficiently identifying probable interfaces between components that are not explicitly bonded. Recently obtained results indicate that application of the PGALSR algorithm with the new NearTree algorithm can provide such an efficient algorithm by subtracting the surface of the union of the complex components from the unions of the surfaces of the complex components. Additional presentations for the 2011 meeting of the American Crystallographic Association in New Orleans and the 2011 International Union for Crystallography Congress in Madrid on this research are in preparation and will be reported after completion of Period III.

Work continues on both threads during the current no-cost extension, and we continue to seek other sources of funding to continue the work. There is a significant possibility of funding from the NIH that would allow the work on complexes to continue in collaboration with Paul Craig at RIT. In collaboration with others, we continue to seek funding on issues relating to archiving of diffraction images.

For further background on both threads of work, see the report on progress in Period II, below.

We have added support to CBFlib for interoperation with NeXus, HDF5 and TIFF, and are collaborating with others in seeking funding to make a common data framework suitable for archiving of images a reality for the community. In the molecular complexes thread we developed a new fast algorithm for identification of surfaces in 2009 and 2010 and have just further improved the speed of the algorithm, making it considerably faster than what had previously been considered the fastest algorithms. We are now applying it to the identification of complex interfaces.

Progress in Period II (2006 -- 2009)

The major focus of the work in Period II was on surface identification for assembly of large complexes and on building infrastructure for management of diffraction images. Significant progress was made in the refinement and application of our new algorithm for approximations to Lee-Richards surfaces to create a workable tool for identification of surface atoms and residues. The software for support of diffraction images was extended, with wider acceptance of CBFlib by the community. The collaboration between this project and members of the community in that reporting period resulted in the addition of support for Fortran and Java applications to the existing support for C, C++ and Python applications in CBFlib.

The presentation of complxes can be done as independent models in the current style used by the Protein Data Bank in generating large structures from symmetrically related components. The presentation can also be done as a consolidated entry thereby avoiding reuse of atom serial numbers in WPDB, mmCIF and, if the total number of atoms is less than 100,000, in traditional PDB format. That work took us into issue of robust management of both data and metadata, closely related to our work with the IUCr (see http://arcib.dowling.edu/cifiucr) on upgrades to the capabilities of CIF, which draws on the same software base (CBFlib) that we use for our work on synchrotron image data. The current mmCIF support in RasMol uses a modified version of CBFlib, which was augmented to support the handling of maps in RasMol, and the changes have been fed back into the main line of CBFlib development.

Earlier we had been doing connected component analysis by passes through the covalent bond network and the hydrogen bond network. We came to the conclusion that we could achieve a significant simplification of this logic by switching to use of a more general water-probe atom mediated “surface-bond” analysis, since atoms coordinated by such pseudo-bonds are almost certainly part of the same biologically related complex.

We started with the classical approach of identifying surfaces by looking atom-by-atom for molecule-solvent interactions. However, in May 2007, while looking at the graphical representation of electron density maps and considering the solvent-bounded areas of such maps, we observed that there is a strong similarity between the surface obtained by probing a surface with a solvent atom (a Lee-Richards surface) and the surface obtained by selecting a appropriate contour level in the electron density. The advantage of contouring an electron density (either an experimental density or a model density created from pseudo-Gaussian atoms) is that contouring a surface in an electron density is much less computationally intensive than a classic Lee-Richards surface calculation. More importantly, this approach makes it much easier to classify atoms as being on the surface or being buried. Thus one can identify surface-surface interactions by first identifying the surface atoms of the components in isolation and then focusing on those atoms that are classified as surface in isolation but as buried in the complex. In 2007-2008, we reduced those observations to practice and found a robust combination of scale factors and spreads for the pseudo-Gaussians that reliably identify surface atoms. In 2008-2009 we integrated that algorithm into RasMol, partly under funding from this project and partly under funding from an NIH grant on molecular graphics scripting languages (see http://sbevsl.sourceforge.net). A paper on this work was prepared and published during Period III.

In view of the importance of data and metadata management issues, a supplement to this DOE BIOMOL grant ER63601-1021466-0009501 was proposed and funded in 2007 for support of new software capabilities relating to the management of image data, especially image data from crystallographic experiments at synchrotrons. The objective of this supplement was to create a pool of software to assist in the conversion between existing synchrotron image formats and the emerging imgCIF standard. The work that led to that supplemental effort arose from a workshop series funded by DOE, NSF and NIH under a coordinated set of workshop grants. To summarize that work: The first workshop in the new series, on “Management of Synchrotron Image Data: imgCIF File System and Beyond”, was held on 22 July 2006 as part of the 2006 ACA meeting in Honolulu, Hawaii. That workshop concluded that further work on imgCIF was needed. The PI arranged to visit the Paul Scherrer Institut in Villigen, CH (home of the Swiss Light Source) from 9 to 14 January 2007. The PI worked at PSI with Miroslav Kobas, Eric Eikenberry and Wolfgang Kabsch to design an acceptable interface based on imgCIF to support use of the new SLS detector design with Wolfgang Kabsch's XDS processing software. In order to achieve this satisfactory result, two compression algorithms were added to CBFlib and new header padding was provided in binary sections with code written in both C and Fortran-90. Code for one of the compression algorithms and the Fortran-90 code were delivered in early February 2007; permissions were obtained from CCP4 and from J. P. Abrahams for inclusion of the second compression algorithm in CBFlib. Both compression algorithms are now included in current CBFlib releases. The visit to SLS was followed by an equally intensive visit to Andy Hammersley (the originator of the imgCIF format) and Jon Wright at the European Synchrotron Radiation Facility (ESRF) from 14 January to 20 January 2007. That visit concentrated on resolving issues in support of beam center definitions and binary arrays of real data, and on achieving clarity in axis definitions. This resulted in a new iteration of the CBFlib code in addition to the changes made for SLS. That code was used by Chris Nielsen of ADSC to resolve beam center issues in the first ADSC imgCIF-based test data collections at Diamond in early February 2007. Work continued collaboratively over the Internet. The PI also attended John Huffman's Crystal Grid raw data workshop in Bloomington, IN starting on 26 April 2007. That workshop went very well and subsequently participants have "push[ed] for continued adoption of standards for diffraction data (imgCIF and expanded metadata definitions) and to encourage the adoption of shared data repositories for research and educational use throughout the community." [email from John Huffman, 2 May 2007]. In 2007-2008, the infrastructure effort continued in work with Chris Nielsen of ADSC that resulted in "jiffies" to convert both to and from the ADSC detector formats. That code was been released in CBFlib. The original work on the SLS Dectris detector format also continued with a "jiffy" created to go from the original SLS "mini-CBF" format to full CBF format. Two final workshops were held, one at BNL on 22 May 2008 and one at the IUCR meeting in Osaka, JP on 26 August 2008. That last meeting appears to have been a tipping point after which use of the imgCIF format became accepted as at least a viable alternate format by the major detector equipment vendors. (See http://www.medsbio.org/meetings/Osaka_Aug08_imgCIF_Workshop_Report.html). In this reporting period, we added software for conversion from imgCIF to NeXus. That software was adopted by and improved by Peter Chang at Diamond Light Source. We created a Fortran interface to the entire CBFlib API in collaboration with Kay Diederichs of U. Konstanz, who now is maintaining the XDS package. In collaboration with Andy Arvai of Scripps, we augmented CBFlib to be able to deal with a new CBF variant format being produced by XDS.

Most recent products delivered:

Products delivered in the current reporting period:

Peer-Reviewed Publications:

H. J. Bernstein, P. A. Craig, “Efficient molecular surface rendering by linear-time pseudo-Gaussian approximation to Lee-Richards surfaces (PGALRS),” J. Appl. Crystallography 43:2, 2010, pp. 356 -- 361.

M. T. Dougherty, M. J. Folk, E. Zadok, H. J Bernstein, F. C. Bernstein, K. W. Eliceiri, W. Benger, C. Best, “Unifying biological image formats with HDF5,” CACM 52:10, 2009, pp. 42 -- 47.

Meeting Presentations:

L. C. Andrews, H. J. Bernstein, “A Pair-Based Approach to Structural Homology Using
Quaternion SLERP Averaging and Local Rotations,” poster T-100, ACA 2010, American Crystallographic Association meeting, 24 -- 29 July 2010, Chicago, IL.

H. J. Bernstein, “imgCIF, HDF5, NeXus: Issues in Integration of Images from Multiple Sources,” at “HDF5 as hyperspectral data analysis format” Workshop,11 -- 13 January 2010, ESRF, Grenoble, FR.

E. Zlateva, H. Bernstein, N. Darakev, G. McQuillan, J. Ihm, “Use of the Next Generation Dictionary Definition Language, DDLm, in CIF validation with CBFlib,” poster P-M056 (abstract W0122), ACA 2009, American Crystallographic Association meeting, 25 -- 30 July 2009, Toronto, Canada.

N. Darakev, H. Bernstein, J. Ihm, G. McQuillan, E. Zlateva, P. Craig, L. Grell, “New Molecular Graphics Movie Scripting Features under SBEVSL,” poster P-M057 (abstract W0124), ACA 2009, American Crystallographic Association meeting, 25 -- 30 July 2009, Toronto, Canada.

H. J. Bernstein, P. Craig, “A New Approach to Identification of Surface Residues,” poster P-T006 (abstract W0121), ACA 2009, American Crystallographic Association meeting, 25 -- 30 July 2009, Toronto, Canada.

Most recent notes concerning the project:

Progress in Period I (2003 -- 2006)

The focus during Period I of the project was on the addition of local support for crystallographic and non-crystallographic symmetry and the identification of biological units in the ensemble. In 2005-2006, the project was extended to include development of a new Wide Protein Data Bank (WPDB) format to overcome the 99,999 atom limitation of the traditional PDB format and to make some other improvements.

Most entries determined by x-ray diffraction provided by the Protein Data Bank contain the crystallographic asymmetric unit, or unique fragments in the case of molecules such as viruses. In order to understand the biological significance of this data, it is important to apply both the crystallographic symmetry and appropriate non-crystallographic symmetry operations. EBI offered biological units in the PQS database, which are now offered directly by the PDB. However, the use of a shared server combined with the speed limitations of remote internet access limits the volume of data that can be efficiently provided. With appropriate extensions to the tools available to users on their local machines, many more choices can be made available in a cost-effective manner, allowing more timely insights by researchers and more effective training of students. By bringing full management of crystallographic and non-crystallographic symmetry operations into RasMol a better understanding of biological units is possible without the user having to retrieve information from multiple sources. Having symmetry and biological unit generation within RasMol also allows work with coordinate entries that have not yet been deposited in the PDB. In addition, when working intensively with a particular structure and needing to look at other entries, it is clearly desirable to provide local versions of search tools which interact smoothly and naturally with the display functions. In Period I, coding was started on almost all aspects of the project, and most of the major components were completed. The essential crystallographic symmetry code was completed first summer of the project and a first version of the noncystallographic symmetry (NCS) code was also in place early on and then reworked allow better memory management when working with very large numbers of NCS operations. The connected component code needed to infer biological activity was completed.

Near the end of the second year of the project, the grant was supplemented with an award to create a new Wide Protein Data Bank Format (WPDB). The macromolecular CIF (mmCIF) format was developed to capture all the detailed searchable relationships among data items and to overcome the field-width limitations of the old 80-column PDB format. While mmCIF has been of great value to the internal operation of the PDB and has simplified the creation of powerful search engines, users of the PDB have been reluctant to make the transition from the easily-read fixed-field PDB format to mmCIF, and most existing macromolecular software remains unable to read mmCIF directly. We addressed the issue by creating a new fixed-field wide PDB format (WPDB) that carries all the information provided by mmCIF using record formats very similar to those in the existing PDB format. By increasing the widths of many fields (especially atom number, atom name, chain identifier and coordinates), we overcame the major deficiencies of the old PDB format, including the 99,999 atom limitation and the idiosyncratic atom naming forced by the 4-column field width, especially for hydrogens. In addition, we added new record types to carry the new information from the mmCIF files (e.g. the more detailed information on secondary structure and biological function). We created simple external translation filter programs to go back and forth to mmCIF format and, where the data permitted, to the existing PDB format. The translation program from mmCIF to WPDB format is available both as source code and via a web server at http://biomol.dowling.edu/WPDB.

The WPDB format was integrated with RasMol for both reading and writing.

Personnel Involved:

PI/PD: Herbert J. Bernstein (Dowling College)

Research Collaborators: Lawrence C. Andrews (Micro Encoder, Inc., unfunded collaborator), Frances C. Bernstein (Bernstein + Sons, unfunded collaborator), Paul A. Craig (RIT, funded under an NIGMS grant)

Current Students: Kedian Jimenez, Barry LaPierre, Matt Rousseau.

Students who worked on the project in prior years: Niroshan Egodawatte, Clarice Chigbo, Ricky Chachra, Georgi Darakev, Damian Glinojecki, Parag Jain, John Jemilawon, Nan Jia, Stavros Louris, Petko Kamburov, Sagar Pilania, Stojan Regodic, Vencislav Stanev, Georgi Todorov, Isaac Awuah Asiamah, Nikolay Darakev, Jonathan Ihm, Gregory McQuillan, Elena Zlateva.

As software is developed it is added both to our open-source gforge server, http://blondie.dowling.edu and to sourceforge.

The RasMol code is available at
http://www.bernstein-plus-sons.com/software/rasmol
http://blondie.dowling.edu/projects/rasmol
http://www.sourceforge.net/projects/openrasmol

The CBFlib code is available at
http://www.bernstein-plus-sons.com/software/CBF
http://blondie.dowling.edu/projects/cbflib
http://www.sourceforge.net/projects/cbflib

The sourceforge site is proving to be very popular.

Other Project Information Sources:

Project URL:

http://biomol.dowling.edu

Related URL at institution:

http://blondie.dowling.edu

Contact: RIMSAdmin@science.doe.gov
Search OBER Abstracts Database Note: There is a delay in posting abstracts to allow for Program Manager review.