Key Directions for Research Libraries

Dealing with Raw Unprocessed Data for the Biology Community

Eldon Ulrich, University of Wisconsin-Madison

The Biological Magnetic Resonance Data Bank (BMRB) is a repository primarily for atomic level nuclear magnetic resonance spectroscopic (NMR) data for proteins/peptides, nucleic acids, and small biologically relevant molecules. Metadata describing the molecule(s) studied, the experimental conditions and instrumentation used, and the primary publication relevant to the study are archived. This data provides insight on the three-dimensional structure, chemistry, dynamics, and thermodynamics of the systems studied. A large collection of raw unprocessed time-domain data, as collected from the NMR spectrometer, is also available, as well as educational material and links to a variety of web resources.

Many researchers use individual entries from the archive in comparative studies against the biological systems they are investigating or to bootstrap their own work on closely related systems. In the early stages of these projects, there is a need to understand the techniques required to study systems with similar characteristics, to identify who has investigated such systems, and to locate the kind of data and conclusions that have been derived in the past. Later in the project, tools to quantitatively compare local data to data taken from BMRB and other public repositories are needed. These tools may include file format converters and data visualization systems. At the point of publication (or even at the beginning of a project in writing a grant), researchers have a need to integrate the knowledge they have obtained into a broader context. A specific goal is to construct a focused list of relevant publications to be used as references for a publication or grant. This requires a literature search for information on related molecules or even seemingly unrelated molecules that have similar function or dynamic and thermodynamic characteristics.

The Gene Ontology project provides a controlled vocabulary that allows **efficient searching** for gene products with an analogous function and cellular location. **Query tools** capable of locating molecular similarity through sequence or three-dimensional conformational homology across various archives are needed. Applications that generate networks of sequence or functionally homologous molecules exist, but generally are not linked to publications. **Web based tools** that correlate authors and publications with the molecules studied and the techniques used through text and keyword searches are available. However, inconsistencies in molecular nomenclature and other ambiguities often lead to many extraneous hits. A more structured use of database accession codes in publications and a combination of the query approaches mentioned above, although computationally expensive, may provide the biological research community with more highly refined views into the published literature.

The field of **biological NMR** is being pushed forward by researchers who often make use of both the **raw unprocessed data** and the assigned spectral parameters available from BMRB to derive and test new methodologies. These include techniques for automating NMR data analysis protocols and, recently, novel methods for deriving three-dimensional structures from assigned chemical shift data. Many of these efforts involve the development of probabilistic profiles from high quality data sets constructed by careful selection of subsets of the public databases. Alternatively, methods are developed for identifying and 'correcting' aberrant data in the public archive to construct local databases of sufficient data quality and quantity. This raises the question of whether a public archive should make the effort to identify and mark data outliers and provide filtered or corrected versions of its repository.

The Collaborative Computing Project for NMR (CCPN) and **BMRB NMR-STAR data models** along with **IUPAC atom nomenclature rules** are providing **standards** for the uniform collection and archiving of biological macromolecular NMR data. The use of these **standards** has greatly improved the ability of scientists to integrate published data into their workflow and to develop novel software applications.

About the Author

Eldon Ulrich is director of the Biological Magnetic Resonance Data Bank (BMRB) at the University of Wisconsin-Madison. The project was initiated with John Markley in 1988 as an international repository for Nuclear Magnetic Resonance (NMR) spectroscopic data derived from studies of biological macromolecules and from small molecules of biological interest. Funding is provided by the U. S. National Library of Medicine and all project resources are freely available to the public. BMRB is a member of the Research Collaboratory for Structural Bioinformatics (RCSB) and of the World Wide Protein Data Bank (wwPDB). Eldon has been involved in the design of data models for biological NMR data and of web-based data deposition systems for these complex data sets. His research interests have included NMR investigations of protein and nucleic acid structures, and small molecule lipid membrane interactions.