HIV Structural and Biothermodynamics Databases: a Resource for the Pharmaceutical and Biotechnology Industry

Talapady N. Bhat*, Yadu B. Tewari, Henry Rodriguez, and Robert N. Goldberg

*Biotechnology Division (831), NIST, 100 Bureau Drive, Gaithersburg,MD 20899-8313, U.S.A.bhat@nist.gov

Abstract

Federal agencies, academia and industries have invested heavily in the development of structural and biothermodynamic data. However, the data are still largely distributed over several public and private archives leading to issues of inadequate interoperability. In order to maximize the investment made in these areas it is necessary to make it widely available to, and easily accessible to continued development and use by a broad community of researchers and industrialists. For this reason, during the last several years we have been working on developing and maintaining two databases; an integrated structural data resource for AIDS (http://xpdb.nist.gov/hivsdb/hivsdb.html) and a biothermodynamic data resource (http://xpdb.nist.gov/enzyme_thermodynamics).

Introduction

Bioinformatics is the collection, archival, annotation, evaluation, analysis, and distribution of recorded knowledge. It plays a critical role in the technological and industrial use of biological data. During the last several decades, both public and private sector have invested heavily in the development of 2-D and 3-D structural and biothermodynamic databases. The results of these efforts are largely held in several distributed private and public archives; thus making it difficult to access, share and analyze these data. For the last several years, NIST has been developing two integrated data resources; one for the AIDS structural data [1] and the other for the biothermodynamics data[2]. We have also used these databases to develop and test both information infrastructure and data mining tools such as chemical ontology and Chem-BLAST (Chemical Block Layered Alignment of Substructure Technique). Chemical ontology is established for all compounds using data dictionary driven tools that define inter compound relationships. Chem-BLAST is a technique to align and query chemical substructures using techniques similar to those of BLAST[3] that searches on proteins and DNA sequences.

HIVSDB

The human immunodeficiency virus encodes an aspartic protease that is responsible for post-translational proteolytic processing of the *gag* and *gag-pol* polyprotein gene products into mature functional proteins. Following these proteolytic activities, the viral particle undergoes a morphological transformation from non-infectious to infectious virions. Despite the widespread availability of drugs that bind to the active site of HIV protease and thus block proteolytic activity, treatment of AIDS is still a work in progress. In general, the active site is considered to be fairly rigid and the inhibitor fills this cavity by exhibiting chemical and geometrical complementarities to the amino-acid residues of the protein. Different drugs predominantly differ in the way they organize themselves inside the active site of the enzyme, as revealed by X-ray three-dimensional structure of the enzyme/drug complex. For this reason, - X-ray crystallographic data plays a major role in rational design of AIDS drugs. Efficient and accurate standardized indexing of inhibitor compounds and their components that target specific pockets of the enzyme is vital for rapid and reliable inhibitor structure comparison and drug design studies. HIVSDB is unique from other structural biology databases such as the PDB in that its holdings include structural data both from the PDB and those obtained directly from industrial and other laboratories. HIVSDB, unlike the PDB, also contains 2-D structural data of potent inhibitors obtained directly from the published literature.

A novel technique to annotate both 2-D and 3-D chemical structural data has been developed and illustrated using HIVSDB. This method first establishes standard dictionaries for the chemical taxonomy of fragments of compounds. Then it annotates all the compounds into a data-tree based on their chemical substructures. Novel search engines have been developed to compare, align and query different elements of the data-tree. These search engines allow a user to query both using text strings for chemical names and 2-D drawings of the substructures. The organization of substructures into several structural layers is utilized to provide a novel search engine - Chem-BLAST. A full description of the tools and techniques will be published elsewhere.

At present, work is underway to extend this database effort to include other AIDS related molecules such as Reverse Transcriptase and Integrase.

Biothermodynamic Data

Thermodynamic data play an essential role in chemical manufacturing technology. Examples include the production of food and pharmaceutical products such as fructose from corn syrup. The important and well-recognized fact is that these data are often distributed in various scientific publications many of which may not be available electronically. Moreover, the scattered nature of such data makes it hard for researchers to obtain reliable access for use by process engineers involved in the optimization of product yield, and the efficient utilization of energy. Thermodynamic property data for biological molecules play an especially important role in process calculations. The primary application of the data is for assessing the feasibility of reaction process for industrial decision making and optimizing product yields.

The biothermodynamic database provides comprehensive biothermodynamic data used by biochemists and biochemical engineers for assessing the feasibility of many types of biochemical processes and for optimizing product yields in industrial applications. These data have been obtained by searching through many of the published literature.

The biothermodynamics database provides a compilation of data on the thermodynamics of enzyme-catalyzed reactions. The data presented are limited to direct equilibrium and calorimetric measurements performed on reactions under *in vitro* conditions. This is the principal thermodynamic information that is needed to determine the position of equilibrium of a given reaction. The following information is given for each entry in this database: the reference for the data; the reaction studied; the name of the enzyme used and its Enzyme Commission number; the method of measurement; the conditions of measurement (temperature, pH, ionic strength, and the buffer(s) and cofactor(s) used); the data and an evaluation of it; and, sometimes, commentary on the data and on any corrections which have been applied. The absence of a piece of information indicates that it was not found in the cited paper.

The database may be queried mainly in two ways; either using predefined values or using user input values. The webpage is constructed using Perl scripts and the data are stored in MySQL. A user may also preview the data for all the elements prior to making queries. All the data have been examined and validated and users may view the citation of the original publication if they choose.

Unlike most other biological data, biothermodynamic data are rich in Greek characters, superscripts, subscripts. While the use of special characters is complicated in a database environment, efforts have been made to properly display these characters on the web.

The development and use of these two databases which have received over 3.7 million hits during the last nine months will be presented. The HIV structural database is through a collaboration among NIST, NIAID (Dr. Mohamed Nasr), NCI (Dr. Alexander Wlodawer) and the University of Rutgers (Drs Edward. Arnold & Kalyan Das).

[1] Prasanna, M.D., Vondrasek, J., Wlodawer, A., Bhat, T. N. Application of InChI to curate, index and query 3-D structures. *PROTEINS: Structure, Function, and Bioinformatics* **60**, 1-4 (2005).

[2] Goldberg, R.N., Tewari, Y. B., Bhat, T. N. Thermodynamics of enzyme-catalyzsed reactions - a database for quantitative biochemistry. *Bioinformatics* **20**, 2874-2877 (2004).

[3] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).