BindingDB: A Web-Accessible Molecular Recognition Database

Xi Chen[†], Ming Liu[#] and Michael K. Gilson[*]

Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD, 20850, USA

Abstract: This paper presents an initial description of the BindingDB, a public web-accessible database of measured binding affinities for various molecular types (http://www.bindingdb.org). The BindingDB allows queries based upon a range of criteria, including chemical similarity or substructure, sequence homology, numerical criteria (e.g. $\Delta G^o < 5$ kcal/mol) and reactant names (e.g. "lysozyme"). Principles of Human-Computer Interactions are being employed in creating the query interface and user-feedback is being solicited. The data specification includes significant experimental detail. A full dictionary has been created for isothermal titration calorimetry data in consultation with experimentalists and data dictionaries for enzyme-inhibition and other measurement techniques are being developed. Currently, the BindingDB contains several data sets of broad interest, such as antigen-antibody binding and cyclodextrin/small-molecule binding. However, it is anticipated that online deposition by experimentalists will ultimately contribute a larger flow of data. We are actively developing software and file specifications to facilitate such deposition.

INTRODUCTION

Molecular recognition, the noncovalent binding of specific molecules to each other, is fundamental to biology. As a consequence, there is an extensive and useful body of experimental data on biomolecular binding. The data generated by conventional experiments are usually published in scientific journals. However, editorial policies prevent including full experimental details, even though such details may be useful to other researchers in assessing the quality of a measurement, reproducing a study, or even fitting a different reaction model to the same data. Moreover, the search capabilities for published binding data are very limited. For example, basic text searches can be performed with publication databases, such as the Science Citation Index, but publication databases do not allow searches over protein sequence, chemical substructure, or ranges of binding affinity. Furthermore, the development of high-throughput measurement technologies is now leading to volumes of data that overwhelm the publication capabilities of print journals. Electronic publication and effective methods for search and analysis are essential if such data are to be shared among scientists and used effectively.

This paper describes the BindingDB, a public, web-accessible database that archives measured binding affinities of molecular systems ranging from the purely chemical to the biological. The BindingDB is expected to be useful in a range of applications, including the development of theoretical models of molecular recognition; the creation of self-assembling chemical systems; and the development of drug-candidates targeted to specific proteins. Our aim is to call this nascent resource to the attention of scientists who may be interested in contributing experimental measurements to the database or in using the database for their own research.

Few other public databases have functions that overlap with the BindingDB. The pharmaceutical and biotechnology industries maintain databases with binding affinities, but these are not available to the general research community and are relatively restricted in the scope of molecular systems they include. A few public databases do provide general binding pair information, but they are not, to date, focused on affinities or experimental conditions and are limited in scope to naturally occurring biomolecules; see, for example the Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu/) and the Biomolecular Interaction Network Database (http://bioinfo.mshri.on.ca/BIND/) [1]. In contrast, a central aim of BindingDB is to provide public access to measured binding affinities and to include data for both naturally occurring and synthetic molecules.

A workshop held in August, 1997, led to a strong expression of support for the development of a binding database focused on binding thermodynamics. The participants represented academic, industrial, and government laboratories and included experts in the measurement of binding affinities, in bioinformatics, and in computational analysis of noncovalent binding. The recommendations of this workshop (http://www.bindingdb.org/update/workshop_rep1b.html) continue to guide the design and management of the BindingDB. In particular, the following guidelines are being followed.

Content

The database should include binding affinities for a variety of molecules, including biopolymers such as

[*] Address correspondence to this author at the Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD, 20850, USA; Tel.: 301-738-6128; FAX: 301-738-6255; e-mail: chen@umbi.umd.edu

[†] Current address: Vertex Pharmaceuticals, Inc., 130 Waverly St., Cambridge, MA 02139, USA

[#] Current address: Novascreen Bioscience Co., 7170 Standard Drive, Hanover, MD 21076, USA

1386-2073/01 $28.00+.00

proteins, nucleic acids and carbohydrates, as well as small organic molecules such as drug candidates and synthetic host and guest molecules. Measurements in both aqueous and non-aqueous solvents should be included. Experimental details, including methods and raw data, should be accommodated but not required for an entry to be acceptable.

Query

A wide range of queries should be supported, including text-based and numerical Structured Query Language (SQL) queries, chemical substructure searches, Basic Local Alignment Search Tool (BLAST) sequence searches, and combinations of these types.

Access

The database should be a public resource accessible via the WWW and should be equipped with a user-friendly interface. Every effort should be made to promote direct deposition of new binding data by experimentalists.

Evaluation

The database should be evaluated and guided by the user community.

The present paper discusses the development of the BindingDB in terms of data content, user-interface, database design architecture, and plans for further development.

DATA CONTENT

A variety of techniques are currently used to measure binding affinities. Of these, the initial implementation of the BindingDB focuses on Isothermal Titration Calorimetry (ITC), a widely used technique that yields not only binding affinities but also enthalpies and entropies [2,3]. The data provided by ITC include the essential attributes of any binding experiment, so this initial work provides a solid foundation for extension to other measurement techniques as well.

The data dictionary for ITC data uses the relational model and currently comprises 29 normalized database tables. As summarized in Table 1, the database tables can be grouped into four categories: Reactants, Reaction Solution, Experimental Results and Details, and References. The data dictionary is now discussed in more detail.

Reactants

The binding reactants may be polymers, such as proteins; small organic molecules; or complexes of polymers

Table 1. Summary of data in the BindingDB

Binding reactants	
Monomer	Small organic molecules, including name(s), MDL Molfile, SMILES, molecular weight.
Polymer	Polymers (e.g. protein, DNA) including name(s), sequence, source organism, molecular weight.
Complex	Molecular complexes in terms of Monomer and Polymer components.
Other DB	Listings of a molecule in external databases such as GenBank and the PDB, including database name and molecule ID.
Reaction solutions	
Solution	Solution in which reaction was studied, including identity of pH buffer, pH of preparation, and pointers to solvent and solute data in the following two tables.
Solute	Solutes used in reaction solution, including name, source, purity, concentration.
Solvent	Base solvents (e.g. water, ethanol) used in reaction solution, including name, source, purity, mole fraction.
Experimental results and details	
ITC Results	Derived thermodynamic results and summary of experimental conditions, including ΔG, ΔH, ΔS, logK, Temperature, pH, pressure, ionic strength, stoichiometry parameter, uncertainties.
ITC Runs	Calorimetry runs used to establish a result in the **ITC Results** table. Includes cell reactant identity and concentration, cell volume, syringe reactant and concentration, and injection volume. Raw data may also be stored here.
Instrument	Instrument used in the measurements, including type, year, model, manufacturer.
Data fitting method	Data fitting method, data fitting software name and version.
References	
Article	Title, journal name, volume, year, page, abstract, keywords, PubMedID**.**
Authors	Names and contact information on article authors and database entrants.
BindingDB Entry	Entry date and title, measurement technique, comments, and searchable keywords.

and/or nonpolymers. These types are stored in the Polymer, Monomer and Complex tables, respectively. The molecules are identified both by name and by chemical structure: sequences for polymers, and Simplified Molecular Input Line Entry Specification (SMILES) strings and MDL Molfiles for nonpolymers. Additional attributes include molecular weight, molecule type, and database IDs for molecules known to be described in other databases, such as the Protein Data Bank [4,5].

Reaction Solution

The solution in which each reaction is studied is described in terms of the base solvent – e.g. 100% water or water/ethanol in a 50% mole fraction—along with any buffer or other solutes, such as salts or reducing agents.

Experimental Results

These tables include the thermodynamics results – free energy, enthalpy and entropy of binding – along with key experimental conditions such as pH and temperature. Binding data must be interpreted according to a specific reaction model, such as $A+B \rightarrow AB$. We initially planned to capture a wide range of reaction models in a single, general data representation, but the resulting data dictionary proved to be complex and difficult to query efficiently. In the current data dictionary, a separate table is reserved for each reaction model and two reaction models are currently included: $A+B \rightarrow AB$ and $2A+B \rightarrow A_2B$. This approach is efficient and straightforward, and it is expected that the vast majority of binding data will be accommodated by a small set of specific models. Another table stores detailed information, potentially including raw data, for the ITC measurements that were analyzed to generate the thermodynamic results. Additional information, such as the data fitting method and the instrument description, are also included in this category.

References

Tables in this category provide publication information, including a PubMed reference ID, related to each data entry. For data that are not yet published, BindingDB will be used as the journal name. In addition, information on the person who entered each data set is recorded for data confirmation purposes.

USER INTERFACE

Data Deposition Interface

Data are collected by the BindingDB staff from previously published papers, and are deposited via client-server software built with Oracle Developer Forms 6.0 (Fig. **1**) Data are initially stored in a slightly denormalized version of the full database that allows for a simpler deposition process. Data in this intermediate database are then checked and moved into the main database by a set of Procedural processing Language/Structured Query Language (PL/SQL) scripts. A data submission can be either "new" or "revised". In revision mode, the user can retrieve previously entered data and change it as needed. Upon submission of the revised data, the old data are labeled as obsolete and linked to the revised data. Obsolete data are not retrieved by queries but can be accessed via links from the up-to-date version of the data.

Because the current input tool provides a direct connection to the intermediate database, it allows the depositor to access and reuse existing data during the deposition process, making for greater efficiency and accuracy. On the other hand, the software must be installed on the user's own computer, which complicates distribution to data entrants at other institutions. We have therefore developed a web-based deposition interface that will be broadly accessible to the scientific community; this effort is described in the section on plans for further development. (see note added in proof).

Database Query Interface

As summarized in Table 2, the BindingDB currently provides the following four query methods, along with several levels of results reporting.

Express search (http://www.bindingdb.org/index.html) provides a single form field for exact keyword, searches.

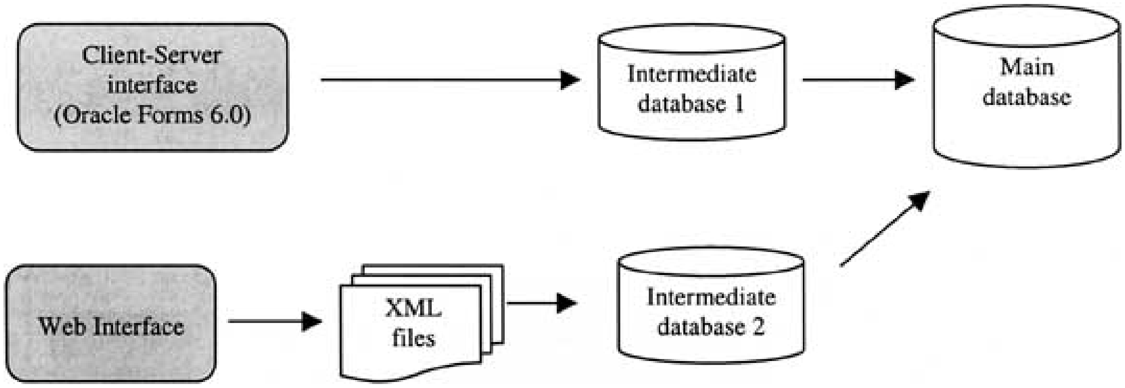

Fig. (1). The BindingDB data deposition process.

Table 2. Current Query Capabilities of the BindingDB

Query Options	
Express Search	Text query by reactant name, substrate name, author name and entry keyword.
Experimental Data Query	Query on numerical or textual data with up to two specified attributes that include reactant names, author names, and ranges of ΔG, ΔH and -TΔS.
Chemical (Sub)Structure Search	Search Monomers for exact match, substructure match, or molecular similarity.
Homology Search	Search BindingDB with BLAST for protein or nucleic acid sequences homologous to query sequence.
Result Presentations	
Summary Page	Reactant names, ΔG, ΔH, TΔS, pH, Temperature, link to Detail Page
Detail Page	Details of selected binding reaction, including reactant names, structures, sequences, and sources; measurement technique; entry date; thermodynamic data (ΔG, ΔH, TΔS, logK) with error estimates; pH; Temperature; ionic strength; references; and link to Further Details.
Further Details	ITC run data, data fitting method, instrument, article abstract, etc.

Currently a number of textual descriptors are accessible via this query method, such as reactant name, author name and entry keyword and additional textual information will be added in the future.

Experimental data query (http://www.bindingdb.org/bind/dbsearch/index.html) allows users to search for entries possessing one or two selected attributes, including reactant name and author name, as well as ΔG, ΔH and TΔS values within selected ranges.

(Sub)Structure search (http://www.bindingdb.org/bind/chemsearch/marvin/index.html) provides exact, substructure and similarity search capabilities against the small molecules (Monomers and Monomers in Complex) in BindingDB and returns the measurement entries that involve the matching molecules. We use JChem1.1 from ChemAxon, Inc. for the chemical draw tool and search engine [6].

BLAST search (http://www.bindingdb.org/bind/blast/index.html) uses BLAST [7] to search BindingDB's entries for entries involving polymers homologous to a user-specified sequence and returns a BLAST report containing a line for each homologous sequence. Binding reactions corresponding to each listed sequence can then be accessed and reviewed.

The result of a query is a set of binding reactions that are presented in a summary page displayed on the user's browser. Each row of the table summarizes one reaction, with information on the reaction model, the names of the reactants, thermodynamic results, and the pH and temperature. Each row also contains a “More” link that transfers the user to a page presenting the details of the specific reaction, including citation data, solution description, and chemical description of the reactants. Small organic molecules ("Monomers") for which chemical structures are stored are displayed in a Java viewing applet, and polymer sequences are provided. The user can drill down from this page to further experimental details such as ITC run data, and instrument make and model.

Interface Design

In addition to functionality, usability is a key focus of our interface design. A new online database can be difficult to learn and potential users might become frustrated or lost in the process of depositing data or executing complex queries, especially with a poorly designed interface. The computer science field of Human Computer Interactions (HCI) aims to use the principles of human cognition to guide the design of user-friendly interfaces [8]. Application of existing HCI principles and exploration of the factors that make for a usable database site maximize the value of the binding database to the user.

HCI principles dictate that interface design be targeted toward the expected types of users. Users of the BindingDB are expected to be professionals who know the task concepts, but arrive with minimal knowledge of the new interface and query procedures. To overcome this limitation, short instructions, dialog boxes and informative feedback regarding the each task are provided before, during and after the task. An online tutorial will also be made available. Designing the web site to have a hierarchy and representation similar to other relevant scientific sites further shortens user learning time; the present model uses organizational techniques found in the current Protein Data Bank (http://www.rcsb.org/pdb/). In order to accommodate both novice and expert users, both simple one-step queries and complex queries are accommodated. It is also essential to display results in ways that are convenient and revealing. Examples include allowing the user to choose how query results are sorted in the summary page, using fonts and formats that focus attention on the main results while unobtrusively providing explanatory information, and presenting certain types of data graphically. Finally, it are important to minimize the frustration and human error associated with data entry. This can be achieved by minimizing the steps required by the user, providing informative feedback on the deposition process, allowing easy reversal of actions, and minimizing the need for the user to manage large amounts of information at once.

A staged design procedure has been followed [9]. First, a low fidelity paperboard edition of the user interface was developed and then modified based on comments from within our development group. Second, a prototype online version of the interface was created and posted publicly. Comments and suggestions, primarily in the form of email messages, were collected from ITC experimentalists, computational chemists and graduate students in Chemistry and Biology. The interface was modified accordingly and the resulting interface can be viewed and tested at http://www.bindingdb.org. More formal studies will be conducted in other formats, such as questionnaires, organized discussions, and direct observation of users accessing the database (see note added in proof). The results of these studies will guide further enhancement of the interface.

DATABASE ARCHITECTURE

The database is constructed with a 3-tier architecture, as diagrammed in Fig. (**2**). The first tier is the user interface, which runs on the user's computer within a web-browser. The second tier is an application server, the functional module that processes the data by receiving inquiries from the user's computer, formulating queries and sending them to the database, and collecting and formatting the results of queries for display at the user interface. The application server runs at the server (database) site. The third tier is a database management system (DBMS) that stores the data and performs queries received from the application server. The DBMS is also at the database site, but can run on a different computer than the application server. The three-tier design is useful because the modularity allows one tier to be replaced or modified without affecting the other tiers. In addition, separating the application server from the DBMS allows the computational workload to be distributed over multiple computers as usage rises [10].

The DBMS itself is an integrated system of heterogeneous databases that stores and organizes the binding data and supports a variety of query types. It currently comprises four major components.

1. A core relational database, managed by Oracle Database Server (Oracle, Redwood, CA) provides the central physical storage for the primary experimental and molecular data described in Table 1. This relational database contains all deposited information in tabular form.

2. Chemical structure fingerprints [11] (http://www.jchem.com/doc/admin/GenerFP.html) for small molecules listed in the Monomer table are stored in the Oracle database, permitting integration of chemical substructure and relational database searches. The fingerprints are generated and searched with JChem 1.1 (ChemAxon, Hungary).

3. ASCII flat files external to the Oracle database are used to store protein and DNA sequences forsequence homology searching. Each data record is in FASTA format and is assigned the appropriate Polymer ID from the BindingDB so that it can be cross-referenced

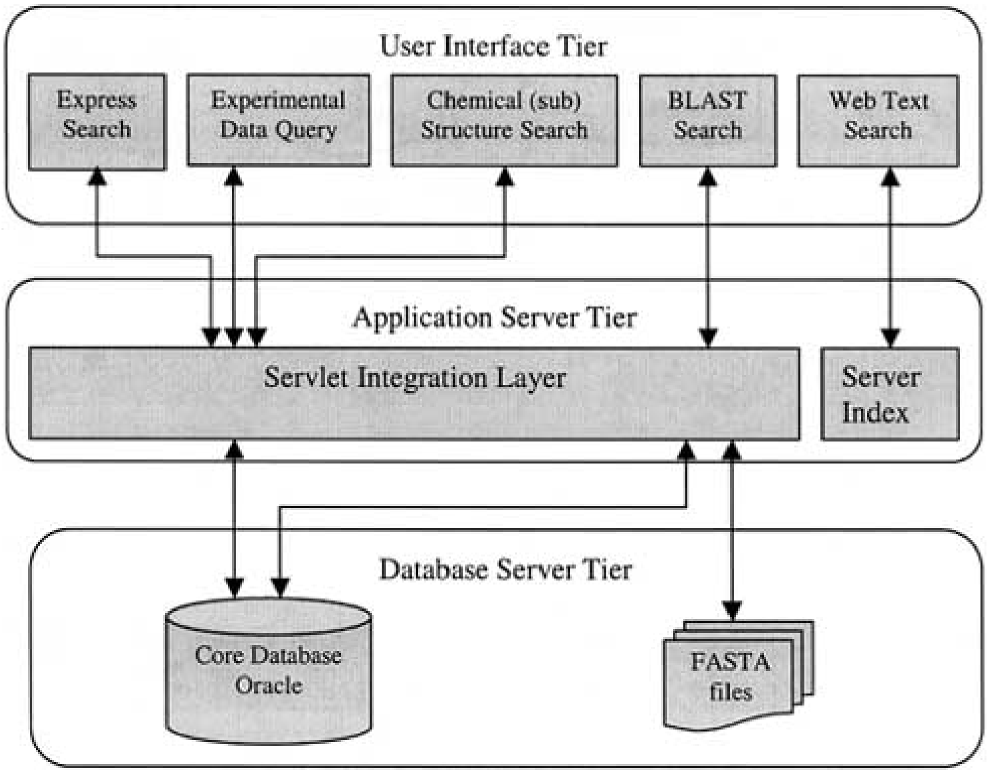

Fig. (2). The integrated query interface of the BindingDB.

with the corresponding record in the relational database. Storing sequence data outside the relational DBMS permits the use of search criteria not supported by standard SQL queries. Here, a BLAST homology search over the flat files yields a list of Polymer IDs that are then used to retrieve the associated binding data from the relational part of the database.

4. The iPlanet web server is used to index the textual content of the web-site itself and provides support for text searches via its internal search engine.

A fifth component, XML(Extensible Markup Language) files of binding data, will be added in the near future as part our effort to facilitate the receipt and deposition of data from various external sources (see note added in proof).

Communication between the application server and the DBMS and among the constituent databases is accomplished by using the Java Servlet technique (http://www.javasoft.com/docs/books/tutorial/servlets/) (Servlet Integration Layer in (Fig. **2**). The BindingDB user-interface on the user's computer dispatches a search request to the application server. The Servlet program determines which database must be queried and formulates the query statements. Each database executes the query and returns results that satisfy the query. The Servlet program then formulates these results into a web page and sends it to the user's browser for display. Because the BLAST search technique is conducted on flat files outside of the Oracle DBMS, a special Servlet program coordinates and integrates the queries of the relational database and the external flat files. The Servlet Integration Layer will also permit other databases, such as Medline (http://www.nlm.nih.gov/medlineplus/medline.html), GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) and the PDB (http://www.rcsb.org/pdb/), to be integrated with the BindingDB

PROJECT STATUS AND DIRECTIONS

The development of the BindingDB has followed guidelines and suggestions provided by workshop participants from academia, industry and government. The current implementation accepts data from a client-server deposition tool into an intermediate database, transfers these data into the main database, and provides textual, numerical, chemical, and protein/DNA sequence query capabilities through a user-friendly interface that runs in a WWW browser. The data dictionary focuses on ITC characterization of binding, but is designed to be extended to other techniques. Continued development of the BindingDB focuses on accelerating the intake of high-quality binding data into the database, on extending and enhancing the query capabilities, and on improving the interface through analysis of the user experience.

The binding data themselves are the heart of the BindingDB, and it is essential to increase the amount of data users can access in the database. To this end, we are developing an XML format (document type definition or DTD) for transfer of binding data, along with software that will enter data received in this format into the BindingDB. This approach will create a clearly defined and general pathway for binding data to enter the BindingDB and will conveniently isolate the problem of creating a user-interface for data deposition from the problem of data transfer and entry. We will then establish a WWW data deposition interface that will generate XML files ready for transfer and deposition into the database. The establishment of a file format from which data can be transfered directly into the BindingDB will also make it straightforward for manufacturers of computer-controlled instruments – e.g. calorimeters -- to create software that will generate deposition-ready files, thus saving time and minimizing the opportunity for human error in the deposition process. The XML file specification provides a natural route for exporting data from the BindingDB to other applications and for integrating the BindingDB with other biomolecular and chemical databases, enabling complex cross-database queries. It also provides a convenient, human-readable backup format for the data. The development of stable and accepted data specifications and DTDs for the BindingDB may prove useful in the development of broader standards for describing and sharing binding data within the scientific community. Many binding measurements use techniques other than ITC, and we plan to extend the existing data dictionary accordingly. An early extension will be to the measurement of enzyme inhibition constants (K_i) because of their importance in drug development and because many such data are available. The existing query methods will be extended to traverse both enzymologic and calorimetric data and the user-interface will be modified accordingly.

The user-interface and query capabilities will be enhanced in other respects as well. Query functions will be provided that allow users to construct more complex queries from existing query elements, and graphical displays -- such as histograms of affinities -- will be provided. Users will be allowed to set up customized data repositories on the BindingDB web-site that reflect their own interests or that focus on data they have deposited. Finally, more formal user surveys and usability studies of the interface will be carried out. The results will guide the ongoing development of the BindingDB as a public, web-accessible resource for the scientific community.

NOTE ADDED IN PROOF

The BidingDB now accepts binding data via on-line forms, accessible via the homepage. Users are invited to deposit data, and also to fill-out the on-line survey. An XML DTD has been established and is used in the deposition process.

ACKNOWLEDGEMENTS

This project is supported by NSF grant DBI-9808318 and by the National Institute of Standards and Technology through the Standard Reference Data and Systems Integration for Manufacturing Applications programs. We thank Drs. J. Westbrook, P. Bourne, H. Berman, K. Breslauer, M. Doyle,

N. Hodge, D. Pilch, E.Plum, J. Rumble, F. Schwarz, and W. Windsor for many helpful discussions. We also thank Dr. Ferenc Csizmadia for facilitating access to the JChem software.

ABBREVIATIONS

SMILES	=	Simplified Molecular Input Line Entry Specification
ITC	=	Isothermal Titration Calorimetry
BLAST	=	Basic Local Alignment Search Tool
PL/SQL	=	Procedural processing Language/Structured Query Language
SQL	=	Structured Query Language
HCI	=	Human-Computer Interactions
DBMS	=	Database Management System
XML	=	Extensible Markup Language
DTD	=	Document Type Definition

REFERENCES

[1] Bader, G.D.; Hogue C.W. *Bioinformatics*., **2000**, *16*, 465.

[2] Breslauer K.J.; Freire E.; Straume M. *Methods Enzymol*., **1992**, *211*, 533.

[3] Fisher, H.F.; Singh, N. *Methods Enzymol.*, **1995**, *259*, 194.

[4] Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.; Meyer, E.F Jr.; Brice, M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.*, **1977**, *112*, 535.

[5] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. *Nucleic Acids Research*, **2000**, *28*, 235.

[6] Csizmadia, F. Journal of Chemical Information and Computer Sciences, **2000,** *40*, 323.

[7] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. *J. Mol. Biol.*, **1990,** *215*, 403.

[8] Shneiderman, B. *Design the User Interface*, Addison Wesley Longman; **1998**.

[9] Rettig, M. *Communications for the ACM*, **1994**, *37*, 21.

[10] Dickman, A. *Informationweek*. **1995**, *553*, 74.

[11] Flower, D.R. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 379.

Received: 5 August 2000 **Accepted: 28 December 2000**