Enhanced CACTVS Browser of the Open NCI Database

Wolf-Dietrich Ihlenfeldt,[‡] Johannes H. Voigt,[†,#] Bruno Bienfait,[†,§] Frank Oellien,[‡] and Marc C. Nicklaus*[,†]

Laboratory of Medicinal Chemistry, Center for Cancer Research, National Cancer Institute, National Institutes of Health, NCI/Frederick, 376 Boyles Street, Frederick, Maryland 21702, and Computer Chemistry Center and Institute for Organic Chemistry, University of Erlangen-Nuremberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany

Received June 14, 2001

A Web-based, graphical user interface has been developed to conduct rapid searches by numerous criteria in the more than 250 000 structures of the Open NCI Database. It is based on the chemistry information toolkit CACTVS. Nearly all structures and anticancer and anti-HIV screening data provided by NCI's Developmental Therapeutics Program have been included. This data set has been augmented by a large amount of additional, mostly computed, data, such as calculated log *P* values, predicted biological activities, systematically determined names, and others. Complex boolean searches are possible. Flexible substructure searches have been implemented. The user can conduct 3D pharmacophore queries in up to 25 conformations precalculated for each compound. Numerous output formats as well as 2D and 3D visualization options are provided. It is possible to export search results in various forms and with choices for data contents in the exported files, for structure sets ranging in size from a single compound to the entire database. Only a Web browser is needed to use this service, with a few plug-ins being useful but optional.

INTRODUCTION

Recently we have presented an overview over the 250 000-compound Open NCI Database,[1] which is the publicly available part of the half-million structures collection assembled by the National Cancer Institute (NCI) in the course of NCI's 45 year long efforts of screening compounds against cancer and, more recently, AIDS. This undertaking, managed by NCI's Developmental Therapeutics Program (DTP),[2] had as one of its "byproducts" computer databases[3−8] of structures and other information, the open part of which was made available to the public[9] a few years ago. One of the results obtained for this data set in our comparative study was that the Open NCI Database had the largest absolute number of unique compounds among the evaluated eight large chemical databases. It contains about 200 000 compounds that were only found in the NCI data set but in no other database. This alone makes the Open NCI Database a valuable resource wherever large chemical structure collections are needed, be it in the context of drug development, for testing and educational purposes, or for any other chemical information application. The fact that this database is entirely free and fully in the public domain, i.e., not subject to any usage restrictions, obviously adds to its value.

The original files are, for the most part,[10] bulk ASCII files for downloading in SD file format[11] containing the connection tables, plus a number of ASCII files containing other textual information, such as names, biological activities, SMILES strings, etc.[12] (A set of downloadable files that have to some extent been processed to combine, or add, information has been made available on a download[13] page of our server.) This set of files is not very easily searchable in a uniform way by the multitude of criteria one may want to apply to this data collection. Furthermore, whatever software one chooses for implementing a more easily and more powerfully searchable version of the Open NCI Database, one is always at risk of being hampered by platform and licensing limitations if one wants to make the result most widely available. These drawbacks are eliminated by creating a Web-based user interface for searching the database and analyzing and displaying results of such searches. Furthermore, hyperlinks to additional services allow immediate further processing of individual structures or hit sets in a great variety of ways.

We have therefore implemented a Web-based graphical user interface for Open NCI Database structures and data, named the Enhanced NCI Database Browser. This service is based on the chemical information toolkit CACTVS,[14] a very general and flexible chemistry information handling tool. The version of the service presented in this paper is Release 2 of the Enhanced NCI Database Browser. A prior release, now discontinued, was put in service a few years ago. The current version is characterized by a slightly larger structure set but mainly by the addition of a large number of new data fields (both directly derived from the structures and externally computed); by a substantially expanded set of search criteria; by several additional display capabilities; by the inclusion of up to 25 (computed) conformations for each compound with concomitant 3D pharmacophore search capabilities; by enlarging the set of output formats offered;

* Corresponding author phone: (301)846-5903; e-mail: mn1@helix.nih.gov.
† Laboratory of Medicinal Chemistry.
‡ Computer Chemistry Center.
§ Present address: ChemCodes Inc., 1300 Englert Dr. Ste. G, Durham, NC 27713.
# Present address: Schering-Plough Research Institute 2015 Galloping Hill Road, K15-L-0300 Kenilworth, NJ 07033.

10.1021/ci010056s
Published on Web 01/28/2002

by greatly expanding the list of links to additional Web-based services for further processing; and by addition of a hit list manager for storage and later reutilization of hits lists. By its very design, this service simultaneously fulfills the task of being a valuable tool in our own drug development efforts as well as making the Open NCI Database structures available to the public in a flexibly searchable way, augmented with a large body of additional information. Features that set it apart from other comparable Web-based systems[15] include, but are not limited to, the 3D search capabilities; links to more than two or three external services; flexible data set export with a large choice of file formats and data contents, for compound sets in size from one structure to the entire database; server-side query result storage; and background processing of large searches.

The large number of implemented search criteria, display options, data fields, output formats, etc. preclude a comprehensive description of each individual feature of the service. Enumeration of all the possible ways the user can combine these features to conduct a great variety of queries would be even less feasible. We therefore have to limit ourselves to highlighting a number of the salient capabilities that may be of most interest and value and to listing other possibilities in a cursory way. The probably easiest way to learn more about the service is simply to use it. At the time of this writing, the URL of the U.S. mirror is http://cactus.nci.nih.gov; the URL of the German mirror http://www2.ccc.uni-erlangen.de/ncidb2.[16]

DATA, SOFTWARE, AND HARDWARE

Data. The data set implemented as of the time of the main development work of the Web site (early 2001) is the totality of the various data sets made publicly available by DTP on their Web site.[12] This includes the currently latest data set with cancer screen results, release August 2000.[17] The last anti-HIV structures/results data set released, and thus included in our service, was from October 1999.[18] The total number of structures offered in the service is 250 251. The DTP Human Tumor Cell Line Screen biological data file contained cancer screen data for a few hundred more entries for which structures were not available on the DTP site, and these records of the Open NCI Database were therefore not included in our data set.

It is important to note that no information about absolute stereochemistry and double bond geometry (E−Z) of any compound is contained in the original connection tables that gave rise to the structure files downloadable from DTP. This affects about 43% of the Open NCI Database. Any structure returned by our service that shows a well-defined stereochemistry is therefore a CACTVS-selected stereoisomer, typically determined by the default rules of the program based on a rapid pseudoenergy estimation, and is to be interpreted and used with the appropriate reservation.

Practically all other information publicly released by DTP for Open NCI structures has been incorporated into our service. This includes the various anticancer and anti-HIV cell-based assay results, compound names for those entries (45 228 or ca. 18%) that do contain names in the original tables, and experimental log *P* values for 3576 compounds.

The program ACD/Names v. 4.0 (ACD Labs, Toronto) was used to calculate IUPAC names. It was able to calculate a name for 220 292 of the quarter-million structures. This added an additional name to many of the approximately 45 000 structures that already had a name (or multiple names), but, mainly, it assigned one or more names for the first time to the large majority of structures in the Open NCI Database. A total of 226 283 structures now have a name, searchable as a whole or by name fragments, thus less than 10% of the structures remaining without a name at this time.

While the service can accept any legal SMILES string[19] as an input format (including complete support of SMARTS for substructure queries), the new release has been modified to generate Unique SMILES[20] as the output format for exported structures. It should be noted that the canonicalization rules implemented are those of the original 1989 publication;[20] modifications of these rules appear to have occurred in the meantime but apparently have never been published.[21]

3D coordinates are available for each structure in the database. However, none of these coordinates are experimentally derived. [Analysis of the overlap of 249 071 structures of the October 1999 version of the Open NCI Database with the organic part of the Cambridge Structural Database (CSD) showed that about 3000 of the NCI molecules are also present in the CSD,[14] but none of the CSD coordinates was used in our service.] The 3D coordinate generation program CORINA[22,23] v. 2.4 was used to generate the baseline 3D conformation stored in our service for each structure. The program Catalyst (Molecular Simulations Inc., San Diego, CA) was utilized to calculate up to 25 additional conformers for each compound. Version 4.0 of Catalyst was used, and the FAST Search option together with an energy cutoff of 10 kcal/mol was selected. This yielded a total of 2 307 037 different conformers for 219 982 compounds, that Catalyst was able to process successfully. The average number of conformers per compound was 10.5. Catalyst conformations could successfully be incorporated in the service for 211 857 compounds. They are in addition to the one baseline CORINA conformation.

The new release of the Web service provides 3D pharmacophore search capabilities for these conformations. To the best of our knowledge, this was the first instance of such a capability being made available in a Web browser environment. If Catalyst conformers are present, the query will operate on those, otherwise the single CORINA baseline conformation will automatically be substituted. The available search functionality is patterned after what is provided by MDL's ISIS system (MDL Information Systems, Inc., San Leandro, CA). ISIS type query files, which can be conveniently drawn with the ISIS Draw software, can be uploaded to the Web interface. As of yet, no Java-based structure editors seem to be available that we could use to directly provide a comfortable input method for 3D queries on a Web page. Simple 3D constraints, such as interatomic distances and angles, can be input directly using a functionality offered by the service on the structure editor page, albeit not in the most user-friendly way at this time (for example, manual re-editing of the resulting query may be necessary). The 3D query facilities are intentionally limited to fixed-conformation search. Anything beyond this level cannot be reasonably implemented on a publicly accessible Web server. In addition, a maximum query execution time of 5 min is enforced for nonbackground searches on the public server.

If the time is exceeded, only the hits found so far are presented. It is however possible to continue the search, if the user is willing to wait for the initial response and resubmit the query to work themselves gradually through the database. CACTVS provides the number of rotatable bonds as a searchable criteria; this is however not necessarily the exact same set of rotatable bonds that Catalyst may have used for calculating the (up to) 25 conformers.

The software PASS (Prediction of Activity Spectrum of Substances)[24] in its version 1.4 was used to generate predictions of a large number of biological activities. These activities comprise toxicities, specific enzyme inhibitions, general pharmacological properties, etc. The predictions calculated by PASS are the probability of a compound being active (P_a) as well as the probability of it being inactive (P_i). These predictions of activity and inactivity probability, respectively, are separately searchable by probability ranges.

The original DTP data contain experimentally determined log *P* values only for a small fraction of the database (3576 compounds). Calculated log *P* data were therefore added for a large number of compounds, computed with two different programs, the first of them being the program KOWWIN (Syracuse Research Corporation, North Syracuse, NY), which has been found to be highly predictive in a comparative study.[25] KOWWIN was able to generate estimated log *P* values for 215 075 compounds. For 8914 compounds, log *P* values calculated with the ACD/Log P software v. 4.0 from Advanced Chemistry Development, Inc. (Toronto, Canada) were obtained from the company. Both sets of estimated values were incorporated into our service and are separately searchable.

Generally, we adopt an inclusive approach regarding calculated data to be added to the database. While we strive to use data computed with programs that are widely seen as reliable or have been shown to be highly predictive in published studies, no detailed analysis of the quality of the calculated data for the quarter-million NCI compounds can usually be performed. The data are provided with this clear proviso of their "as is" nature to the user. Inclusion of data of the same type (e.g. log P) from different sources is therefore not only not an unnecessary duplication but also a certain safeguard and control mechanism available to the user. Irrespective of their source, inclusion of any data into the site does not imply endorsement of specific companies or organizations or their products, nor is a warranty of any kind made. A strongly worded disclaimer on the home page of our server states this to the user.

Software. This service is based on the chemical information toolkit CACTVS.[26] CACTVS is a general and flexible chemical information handling tool. It was designed to handle any kind of chemical information by referring to an open set of descriptions of chemical data and data objects and uses loadable modules to define and extend its capabilities, instead of providing only a fixed set of functions and data it can operate upon. The CACTVS system provided a set of useful features for this project, such as its rapid scan chemistry database format, a flexible query language, and various Web-enabling support functionalities. Its use of a very high level scripting language (Tcl[27] with chemistry extensions) allowed for an efficient development of the CGI scripts for the Web interface as well as of the processing routines necessary for combining the various source data files into the primary database file. The execution speed of searches, etc. is nevertheless very satisfactory (usually in the range of seconds) since all computationally intensive data handling functionality is implemented via compiled C routines, which are linked to the high-level scripting language commands.

While the CACTVS system, as it existed before the project, already provided many of the information processing and information handling features which were needed for the NCI Database Browser, some additional development became necessary. The most extensive enhancement was the addition of the 3D structure query functionality to CACTVS. However, the implementation of this feature was rather straightforward and did not involve the development of query mechanisms fundamentally different from typical such techniques described in the literature. Another important enhancement that had to be implemented was the revision of the database scan file format to support large files. While the previous release had had a primary database file of 350 MB size, it became clear that hundreds of new data fields, plus conformer information would break the 2GB barrier even though only relatively few new structures were added. This necessitated the revision of the database scan file format to support large files. While this posed no problem on some platforms, such as recent versions of SGI IRIX, other platforms, such as GNU/Linux, have limitations that make it very difficult to exceed this file size. For this reason, modifications had to be introduced both into CACTVS and into various parts of the Linux installation that is currently used to host the U.S. mirror of this service to make handling of large database files possible and efficient.[28]

The tasks handled by CACTVS in this database project fall into two major areas. The first area, which was in fact the larger part of the application script development, was the design of about two dozen scripts needed to integrate the data from the various source files into the final database file. Source data were available in a variety of different formats, such as SD-files, text dumps of database tables, or even raw HTML Web page dumps which were used to establish reference IDs to other WWW structure databases. Data synchronization was a major part of this programming work, since the various third-party programs unfortunately tended to do unexpected things to the output for compounds they could not process. Popular strategies encountered were to simply omit structure records from the output or to write error messages into output fields, which were sometimes not trivial to recognize as such. More difficult to handle were unexpected structure manipulations such as the complete and seemingly random reordering of the atoms in the output files by MSI Catalyst, which had to be compensated for when other atomic data, such as the CORINA 3D coordinates, were merged with the Catalyst data.

After preprocessing of the raw input data, the implementation of the WWW interface and CGI backend proceeded in a straightforward way. The database Browser is operated from a single, highly compact CGI script of less than 2500 lines of Tcl code. This script handles all server operations, including execution of queries, presentation of result lists, compound visualization, and forwarding of structure information to external Web services.

Since the interface provides functions such as the continuation of searches, state must be held in some location. We

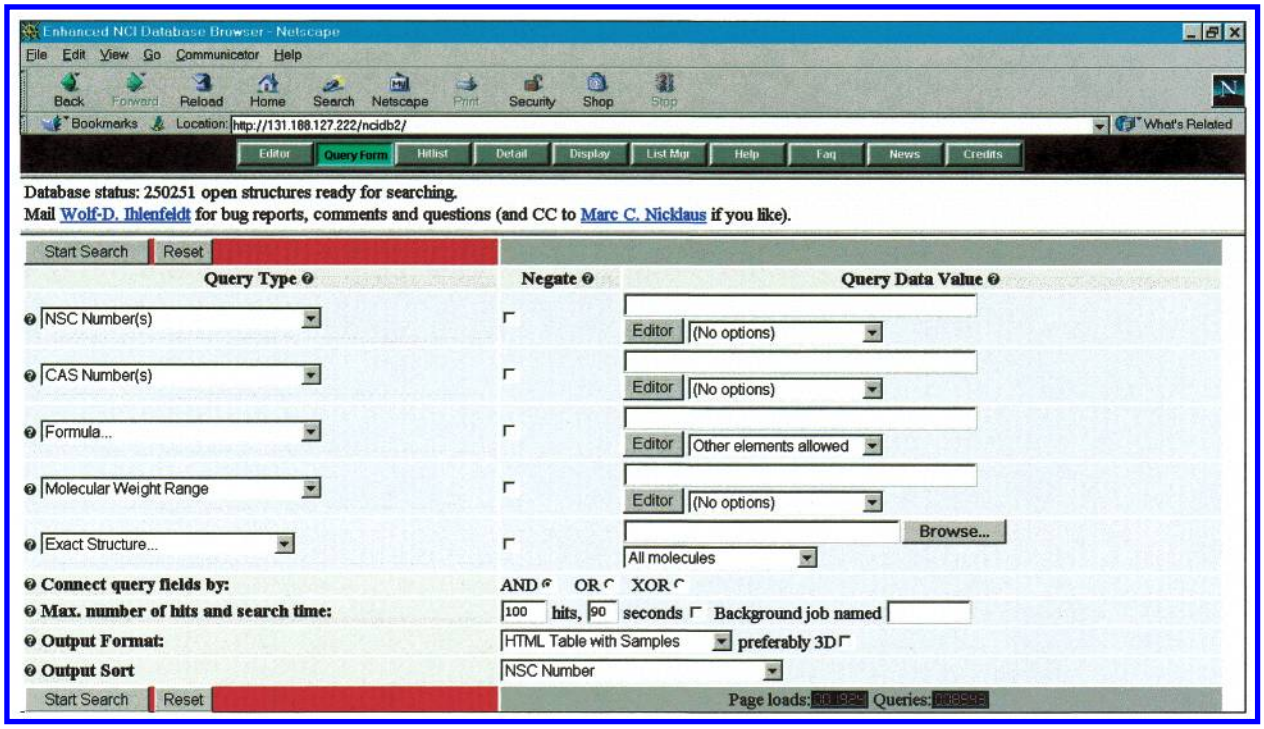

Figure 1. Query form in its initial status, without any query specified.

decided to keep this information completely within the response pages, so no connection and status tracking on the server side was needed. Status data are stored, depending on the context, in hidden WWW form fields or in the parameter section of synthetically generated URLs. Standard database operations which do not require permanent, personalized result storage do not necessitate, and thus do not use, user identification by cookies or other means. Cookies are used only if the user wants to store and manipulate personalized hit lists on the server, for example, by generating unions, intersections, etc. of previously selected compound sets, and for background searches. This is only done for the user's convenience; even these operations can be performed without accepting cookies. Temporary files containing query results, etc. are stored on the Web server in a directory which is routinely cleaned by a cron job. Because of the design of the scan file, startup of the CGI interpreter and opening of the database file for scanning and retrieval is near instantaneous—so far it has not been necessary to resort to a more complex processing model, such as a constantly running database server which is contacted by the CGI scripts executed as result of a user query.

Hardware and Operating Systems. While some of the preprocessing of the data and the calculations of the various estimated properties etc. were performed on several additional platforms (Windows 95/98 PCs, SGI systems running under IRIX), the actual service on the U.S. side with its Web-based Graphical User Interface, the CACTVS database file comprising the entirety of the data, and all the search and display generation functions are currently running on an Intel Dual Pentium III (500 MHz) server with 1 GB of RAM, under the GNU/Linux operating system. The European mirror, as of the time of this writing, is using a 800 MHz Pentium III Dual processor system with a RAID Array and 1 GB of RAM. Even the previously used, less powerful European server, an SGI Origin 200 (180 MHz, 256MB RAM), offered an acceptable performance. The database builds were also performed using the Linux platform. However, this is not a necessity since the binary CACTVS database files are platform-independent and binary compatible. Internally, all data are represented with the XDR binary encoding standard. Database file are interchangeable between platforms with different byte ordering (such as Intel and MIPS).

SERVICES AND CAPABILITIES

Query Screen. Figure 1 shows a screen shot of the Query screen, which greets the user when the service is initially accessed. A button bar at the top of the window allows the user to switch between the different screens ("panes") of the service, such as the Query Form, the Hitlist pane etc. without multiple browser windows being opened. Switching between various panes is made possible by extensive use of JavaScript. One of the most challenging problems during the development of the interface was the preservation of the contents of form fields on the various window panes upon switching back and forth. For this purpose, the top level frame set contains a JavaScript machinery to save and restore form contents without having to rely on the operation of the browser cache.

Query Types. The left part of the Query Form offers the user five separate query fields. The first four of those are identical in terms of the options and suboptions they offer. Table 1 gives a list of the query options currently available. A "..." notation indicates, both in Table 1 and the service, that suboptions are available. For some search options, the suboptions can themselves number in the range of 20 or more. In the case of PASS searches, a separate selector window is opened with more than 500 suboptions (see below). Space limitations as well as the likelihood of future changes make it impractical to list the entire tree of suboptions here. The reader is instead referred to the service itself.

Table 1. Query Types Currently Implemented[a]

NSC Number(s)
CAS Number(s)
Formula...
Record Number(s)
Molecular Weight Range
Number of Atoms Range
Number of ESSR Rings Range
Number of H-Bond Donors Range
Number of H-Bond Acceptors Range
Number of Rotatable Bonds Range
Number of Catalyst Conformers
Drug Likeness Prediction...
PASS Prediction Range...
logP Range...
Name Search...
Complexity Range
AIDS Screen Result
Yeast Screen Level
Stereocenters
Data Availability Constraints...
LIQCRYST Number(s)
Random Set
Full Structure...
Similarity Search...
Transformation Search...
Functional Group Search...
Substructure and/or 3D Search

[a] Three dots ("...") denote that suboptions are available (not shown here).

One specific query type worth mentioning separately are the Data Availability Contraints. By selecting this query type with one of its several subtypes (such as "Any tumor screening [data]" or "[Any] CAS number"), one can search for structures that have *any* value of the specified type in the database, irrespective of the *value* of this data type. In combination with the *Simply Count Hits (entire DB)* output format, one can use this feature to quickly derive various statistics on our data sets.

Query Data Values. Query Data Values are entered in the right part of the Query Form. These values can be a number, a number range, a comma or white-space separated series of numbers or number ranges, a "yes" or "no" value, predefined textual values etc., depending on the query type chosen. Typically, selection of a query type will make the system generate a brief message in the Query Data Value line guiding the user as to what are appropriate entries for this query type. Data input is both checked by JavaScript client-side functions and the server-side CGI script, and appropriate error messages are generated if necessary. The fifth, bottommost, of the query field rows is geared toward structure and substructure searches, in particular the 3D pharmacophore search (see below). This field allows the user to upload and submit a file with a structure definition, a 3D pharmacophore query, etc. Its most useful function is the import of more complex structure queries in the MDL ISIS query format or SMARTS. With the exception of some more complex R-group searches involving R-group logic, the full set of ISIS capabilities is supported.

PASS Searches (Predicted Biological Activities). When the user selects the query type "PASS Prediction Range...," a separate selector popup window appears in which the user can scroll through the more than 500 possible predicted activities. A specific activity has to be selected, and the type of prediction (probability of activity or inactivity, respec-

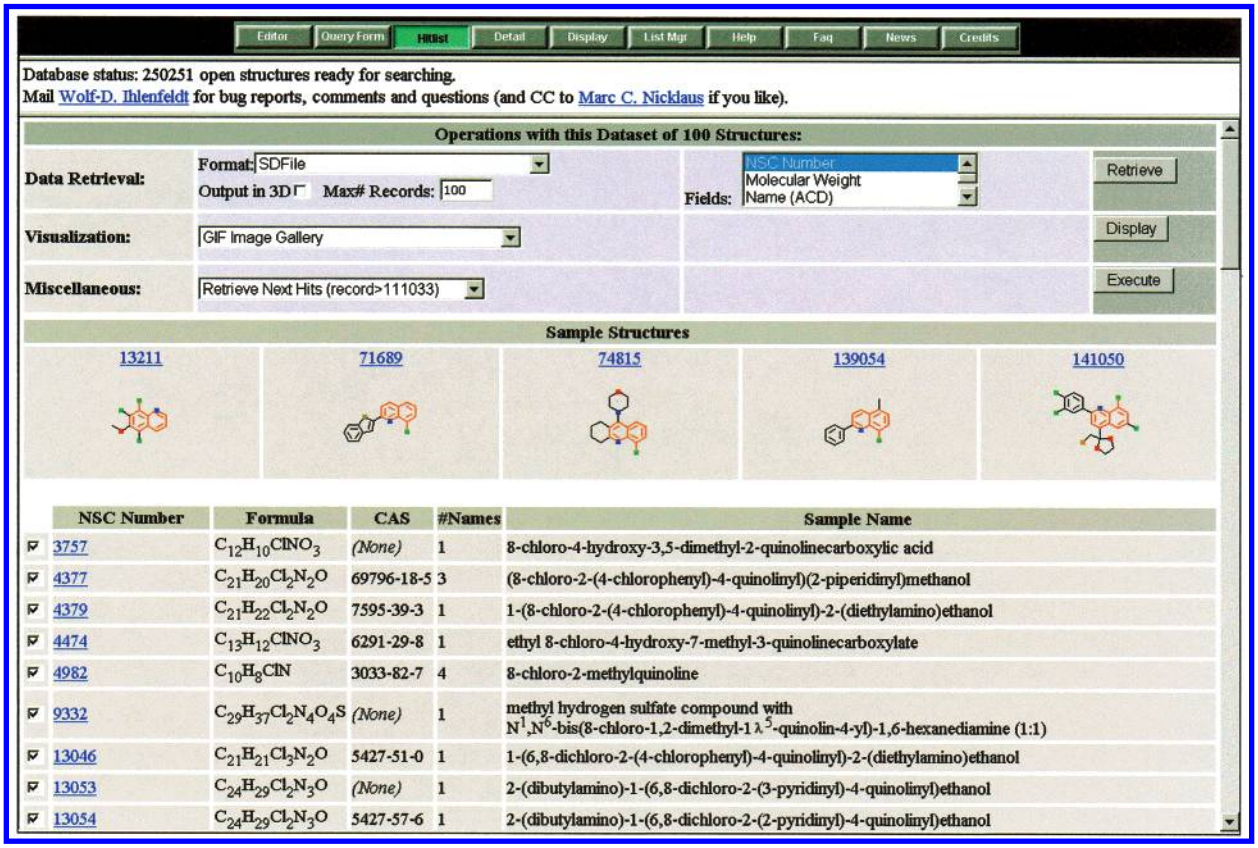

Figure 2. Hitlist pane. This data set was the result of a substructure search using 8-chloroquinoline. This substructure is highlighted in red in the sample structures (random selection) shown. All NSC numbers (NCI accession numbers) are live links to a page showing more detailed information (detail pane). Note the scrollbar on the right—this screenshot shows only part of the whole page.

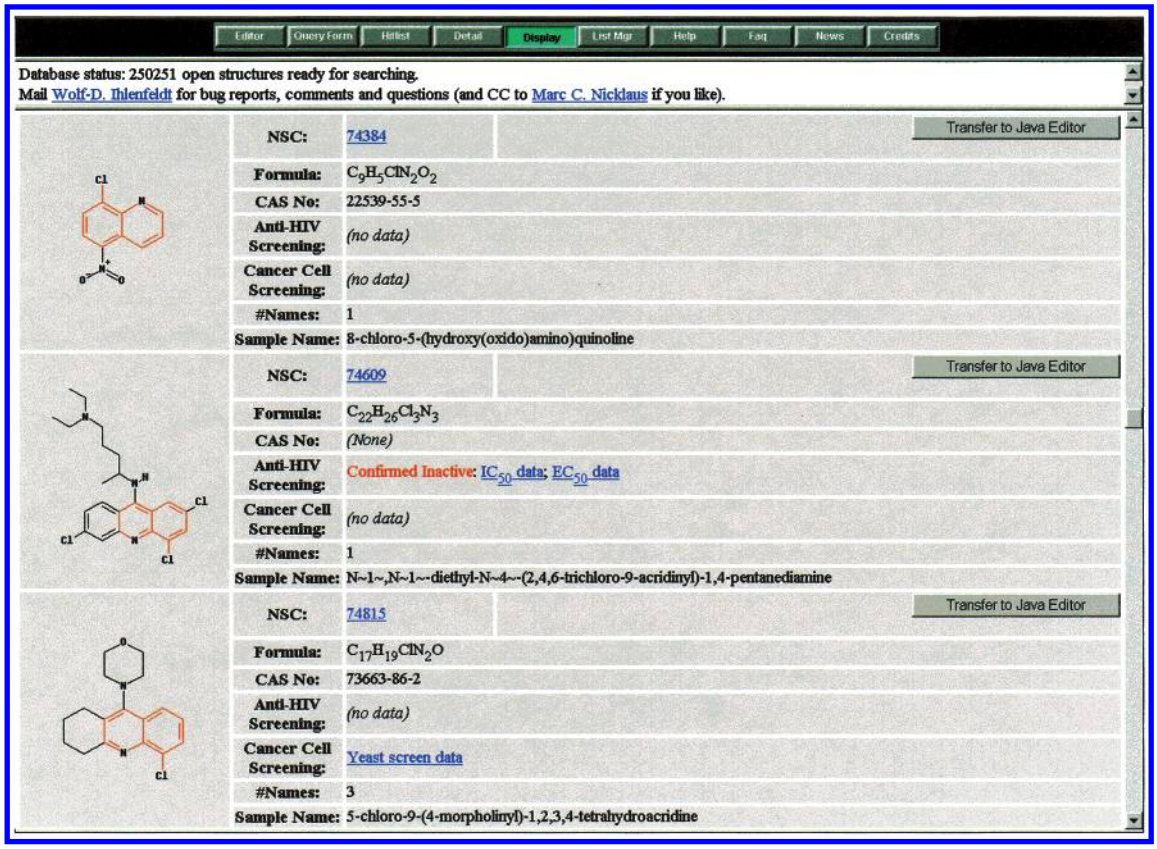

Figure 3. Image gallery. The search substructure (8-chloroquinoline) is highlighted in red in the GIF images. Note that some information is offered as a live link (blue underlined text).

tively) has to be specified. One should apply all the usual caveats to the result of a PASS search that are pertinent to SAR-type calculations of this sort. The PASS predictions can only be realistically used in a statistical manner for sets of compounds and should be treated as scientific "food for thought".

3D Pharmacophore Searches. To prepare a query for a 3D pharmacophore search, one can either use the Local Query Parameters area of the Editor pane, or one can create a query file in ISIS query format externally (using a program such as Catalyst or ISIS/Draw) and submit it to the service. Most of the additional features in query files are supported, such as exclusion spheres, centroids, points on lines, angles, planes, etc.

The best way to view the results of a 3D pharmacophore search is the Visualization option *Chime Display/All Conformers*. This will show all conformations, highlighting the one that was found to match the 3D query (Figure 5). (Once the search algorithm has found one match for a molecule, it will not look for additional conformers that could potentially also match the query.) Superimposition of the query onto the displayed conformers is planned for the future but not yet implemented.

This capability is not a replacement for full-fledged, dedicated 3D pharmacophore search programs. One of its main limitations is that it does not conduct any conformational search on-the-fly but uses a fixed set of precalculated conformers. On the other hand, this allows for a very rapid searching—these 3D searches in more than 2.3 million conformations are generally completed in the seconds to minutes range on our rather modest server hardware.

To test our implementation of 3D pharmacophore searches, we repeated, in a slightly simplified way, a search from previous work. For that study,[29] we had utilized the database program Chem-X to search for HIV-1 integrase inhibitors, using a three-point pharmacophore of oxygen, and oxygen or nitrogen, hydrogen bond acceptor centers, respectively. Specifying just the element types, and omitting an exclusion sphere that had been used in our previous study, we searched in the subset of 60 compounds from the Open NCI Database that had been assayed (out of several hundred that had been found by Chem-X). Out of those, 19 compounds had been found to be active at the level of 200 μM or below. The 3D search as implemented in this service identified 49 out of the 60 compounds as containing the pharmacophore. The 11 compounds that were not found were all in the subset of the 41 inactives, which can be interpreted that this search produced a lower rate of false positives than the original search. Repeating the search in the control set of 61 compounds that had been chosen by random from the Open NCI Database yielded only three hits.

Boolean Operators. Boolean searches are possible by connecting all query fields (if used) with the operators AND, OR, or XOR. Negation of each field is possible. Since the user has the possibility of entering series or ranges of values for most query options, which corresponds to a logical OR between those values, rather complex queries can be submitted. Should these tools not suffice, then the use of the Hitlist Manager (see below) should be contemplated.

Search Options and Limits. Below the Query Type and Data Value sections, the user finds various selection capabilities which influence the search to be performed as well

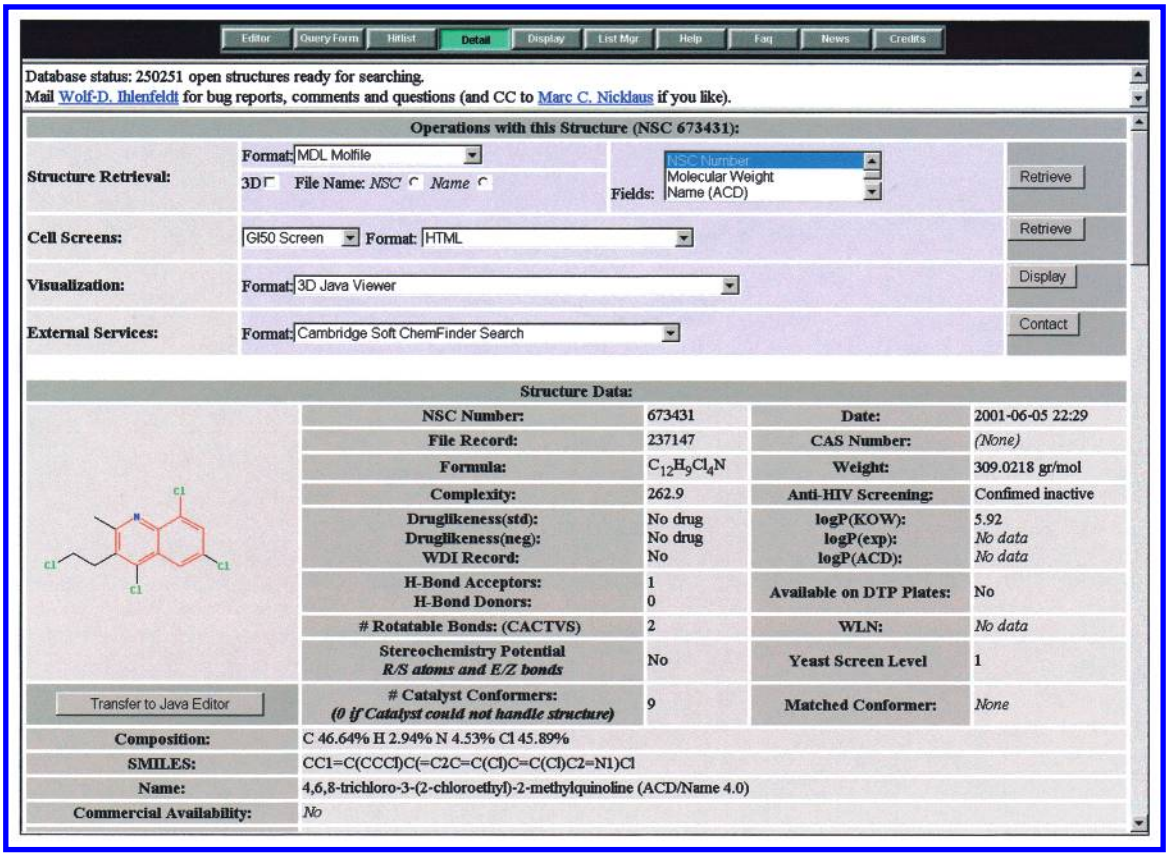

Figure 4. Detail pane. Note the size of the scrollbar—the page is much longer than this screenshot shows. Data not shown include anti-HIV screening results (if present), anticancer screening results (if present), and PASS predictions of biological activities.

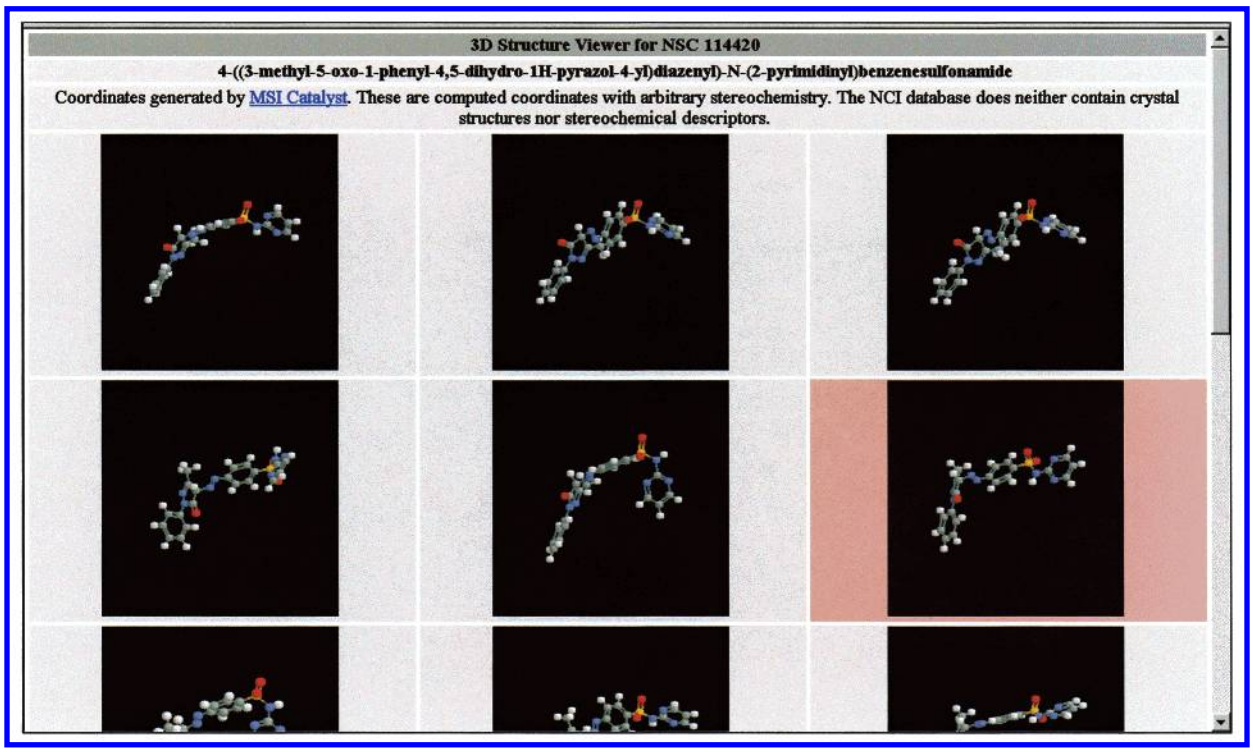

Figure 5. Display option: *Chime Display/All Conformers*. This screenshot shows (part of) the list of conformers of one of the compounds found with the HIV-1 integrase inhibitor 3D pharmacophore from Nicklaus et al. *J. Med. Chem.* **1997**, *40*, 920−929 (see text). The conformer that contained the pharmacophore is highlighted by a red background. Note that each of the conformers on this Web page is a true 3D representation in Chime format that can be individually manipulated by the user (rotated, moved etc.).

as the output returned. While a search may return just one single hit (for example, requesting structure data on a specific NSC number), oftentimes an entire hit list will be the result of a query. The default size of a hit list returned at a time is 100; the hard-coded limit up to which the user can increase the maximal hit list size is currently set to 10 000.

Background Searches. Background searches can be conducted by the user that exceed both the 10 000 structure

Table 2. Output Formats Currently Implemented

HTML Table with Samples
Plain HTML Table
Image Gallery
MDL Chime Gallery/2D
MDL Chime Gallery/3D
Simply Count Hits (entire DB)
NSC number file
CACTVS Hitlist (records)
Alchemy
Cactvs Binary
Cerius
CML
Compass
CTX
Gaussian Input
JCAMP/CS
Molconn-Z
Molfile
PDB
SMD 4
SDFile
SMILES
Sybyl2
VRML 2.0 (Default version)
VRML 2.0 (Custom version)
Xtelplot
XYZ
CAR
M3D

maximum and the 5 min time limit. For this, the user needs to assign a name to the search because the large hit list will be stored on the server instead of being transmitted directly to the user's Web browser. A login procedure with username and password is required to access this hit set, just as it is the case for any other operation with hit lists stored server-side through the Hitlist Manager (see below). Using this option, users can download, if so desired, the entire database with their own combination of included data fields added.

Download Page. Alternatively, a separate page is available on either server that makes available the structures used to build this service for bulk download in various formats, including structure files with DTP screening data added. This page is linked to from the service's Help page or can be accessed directly under http://<mirrorname>/ncidb2/download.html (with <mirrorname> standing for the name of the U.S. and German mirror, respectively).

Output Formats. Numerous output formats (currently 27) are provided. They are listed in Table 2. They cover a variety of different purposes: multirecord data set export, spreadsheet formats, single-compound models, visualization formats, and textual descriptions intended for human perusal. While some of the export formats are equally applicable to query results consisting both of an individual structure or a hit list, others obviously make sense only for a single compound. The system will adapt its format selection menus appropriately.

One option among the output formats is to just return a count of the hits, without the limitation in hit list size that is placed on all other output formats. This allows for convenient gathering of statistical information on the Open NCI Database as well as the data added for this service.

There are also other cases where the hit list size limit is silently temporarily adapted, for example, for structural similarity queries in combination with sorting on that field. For these queries, the whole database is searched, but only a limited result set is displayed, which cannot exceed the hit list size. This scheme ensures that the most similar compounds are displayed, regardless of the selected similarity threshold value, as long as the threshold value is exceeded.

No matter what the hit list size limit is, the service will maximally display 250 hits on the Hitlist pane at a time. A "Display Next Hits" button allows the user to step through the entire result set in chunks of 250 compounds.

Hitlist Manager. The user can store any result set on the server by selecting the "Store Hitlist" menu command from the "Miscellaneous" section of the Hitlist pane (Figure 2). Since this means that user-specific operations have to performed on the server, the user has to register. By means of the Hitlist Manager, one can retrieve existing result sets, annotate them, perform operations such as unions and intersections, and delete result sets.

Visualization Tools. The database interface implements a variety of structure visualization tools, mostly available through the Hitlist and Detail panes (Figure 3). A primary design goal was to allow the use of the NCI database Browser on any platform, with any web browser. While powerful chemistry display tools such as ChemScape's Chime plug-in are clearly useful in controlled, uniform environments such as company networks, the mandatory use of such plug-ins runs counter to the philosophy of a public service, since they are often available only for a very limited selection of platforms and force the user to perform a software installation. All basic information contained in the database and all retrieval options should be usable and accessible without relying on any platform-dependent plug-in. The primary display of structures from this database was therefore chosen to by means of GIF images, the lowest common denominator for all platforms and browsers. These GIF images are generated on the fly by the CGI script and annotated, for example, by drawing matching substructures with red bond lines. A so-called GIF Image Gallery (Figure 4) is the default visualization mode displaying an intermediate amount of information for each structure.

Advanced display options, including 2D and 3D displays with the Chime plug-in, are of course also supported but have to be explicitly requested by the user. As an alternative technology for the display of 3D structure data, we also support VRML. Finally, a ChemSymphony Java applet (courtesy of NetGenics, Cleveland, OH) with basic 3D display capabilities can be chosen as display option if neither Chime nor a VRML viewer are installed on the client computer.

Links to External Services. Another important feature of this database interface are its links to external services and databases. For many of the compounds contained in the open NCI database, additional data are available in a number of other Web-accessible databases. The Enhanced NCI Database Browser links to those data repositories where possible. Several schemes are used for this purpose. For some databases, such as LIQCRYST or NTP, we were able to obtain structure lists or other identifiers which made possible the construction of a cross-reference index during the database build phase. For these databases, a direct link is provided on the structure detail result pages, which will be inserted only if the NCI database knows that a corresponding record actually exists in the remote database.

For other databases, such as the chemistry information source ChemFinder by CambridgeSoft, no fixed link could be provided, since we did not have access to these services' structure lists. In these cases, a button is available which will invoke a CGI script on the NCI database server which will then initiate a search on the selected remote database. Where possible, selective and efficiently queried identifiers such as CAS numbers or full-structure SMILES are used for these searches, in preference to less queriable data such as compound names. The outcome of the search is by necessity unknown in these cases, and the presented result page must be interpreted by the user. Simulating the submission of a query form for initiating a query on a remote Web-accessible database is not always simple, especially when POST-forms are used in complex combinations with embedded, undocumented applets or plug-ins, hidden form fields of unknown meaning, and referrer-checking.

Related to the foregoing, the Enhanced NCI Database Browser provides a number of transfer links to remote services which perform some kind of computation on a selected structure. These services include, among others, 2D and 3D structure visualization, infrared and Raman spectra simulation, molecular orbital computations, Kohonen map neural network tools, and sample ordering from the DTP repository. These interfaces to external services generally operate in such a way that the original input forms of the respective services are displayed, with the fields specifying the structure on which to operate already filled in. In some cases, this involves preloading of structure editor applets with the compound from the NCI database. The specification of the remainder of the options of the various external services as well as the actual submission, by clicking on the execution button, are left to the user. We have developed a general CGI script for chemistry Web-page rewrites to facilitate the implementation of this feature. It takes as parameters the URL of a Web site, plus the names and desired new contents of form fields or applets that are to be manipulated. The script will request the page from the original server, rewrite it to change or preset identified fields, and emit it again for display on the client and for final submission to the external service by the user. This procedure can require some rather complex manipulations such as the proper adaptation of links on the page to images, CGI scripts, help files, etc. and the required recursive processing in case of nested frames.

Help Texts. An extensive help text is available to the user. By clicking on the "Help" button, a separate browser window is opened displaying this text. This is a deviation from the usual pane-switching behavior of our GUI, intended to allow the user to read the explanations while at the same time performing the operation on which he or she was seeking help. A Frequently Asked Question (FAQ) pane is also available, which discusses specific questions, as well as a News and a Credits pane with obvious functions.

Counters for page loads and queries submitted, placed at the bottom of the Query Form pane, tally up the usage of the site. Page loads essentially count the number of visits to the site. The numbers observed so far (on the U.S. server) indicate that each visit to the site leads to an average of approximately five queries being submitted. A Privacy Statement, linked to from the home page of the server, explains our policy toward user and usage data.

APPLICATION EXAMPLES

Three application examples shall be given that illustrate how this service can be used to help with drug development projects, to provide useful data sets for further user-side processing, and to answer scientific questions of various natures.

Example 1: Similarity Searching on an Active Drug Molecule. As an example, we choose Taxol as the drug molecule for which we want to find similar compounds in the NCI database that may have potential drug activity and want to learn more about them if literature is available.

1. On the Query Form pane, choose Query Type "Name Search..." Select "Name fragment" as the suboption [this is usually the safer suboption than "Exact name" since it will find names with additional characters in them, such as, in this case, "Taxol.RTM. (Registered Trademark)"].

2. Type "taxol" as the Query Data Value (search strings are not case-sensitive).

3. Click on either one of the Start Search buttons.

4. For the current data set, 20 hits should come up on the Hitlist pane.

5. Select the first one of the hits (NSC# 125973) by clicking on the NSC number, which is a live hyperlink.

6. This will bring up Taxol in the Detail pane.

7. Underneath the structure drawing of the molecule, click on the "Transfer to Java Editor" button.

8. In the Editor pane that should have come up, with the molecule shown in the JME Editor, click on the "Transfer to Query Form" button.

9. The Query Form pane should come up, with the SMILES string for Taxol in the topmost Query Data value field.

10. Change the Query Type to "Similarity Search..." and select, e.g., "Tanimoto 95%" from the suboption popup menu as the specific similarity search you want to conduct.

11. Click on either one of the Start buttons.

12. For the current data set, this should produce a hit set of 47 compounds listed on the Hitlist pane.

13. One of these hits was cephalomannine (NSC# 318735); click on this NSC number to call up cephalomannine in the Detail pane.

14. In the External Services field, select Format "Medline CAS Number Search" and click on the "Contact" button to the right.

15. A PubMed page should show up in the Display pane with approximately 20 references that contain mentioning of cephalomannine (keyed by its CAS number).

16. Click on any of the individual article links to see the abstracts (in PubMed).

17. You can at any time go back to, e.g., the Hitlist pane by clicking on the appropriate button in the button bar at the top of the page, and continue with further, and different, analyses.

Example 2: Download All Structures in the Database That Have CAS Numbers Associated with Them and That Have No Stereogenic Centers and Include Names and SMILES Strings for All of Them. This is a data set that may be useful for further processing in various contexts. Exclusion of all compounds that possess one or more chiral

center and/or one or more E/Z double bond will ensure that this data set contains no possibly wrong stereoisomers (see above).

It is a good idea—for this search as well as in general—to first gain an impression on how large a hit list you may get from your search:

1. In the first (or any other) query row, choose "Data Availability Constraints..." as the Query Type and "CAS number" as the suboption; leave the Query Data Value at "yes."

2. In the second (or any of the remaining) query rows, choose "Stereocenters" as the Query Type, select "Potential for any stereochemistry?" as the suboption, and change the Query Data Value to "no" (or, equivalently, leave it at "yes" and click the "Negate" box on).

3. As the Output Format, choose "Simply Count Hits (entire DB)".

4. Click on either one of the "Start Search" buttons.

5. In the Status frame at the top, you should see the text displayed "Counted 81 058 hits."

Now we proceed to the actual downloading of these structures. Because this data set is larger than the hardcoded limit for interactive searches of 10 000 structures, you have to do this as a background search. You need to have established a user account and be logged in in order to be able to conduct background searches (see also step 11).

6. Assuming that you still have the query types and values from steps 1 and 2 in the Query Form, click the check box "Background job named" on, and enter a name (e.g., here, CAS_NOSTEREO) in the field next to it.

7. Change the Output Format to "HTML Table with Samples" or "Plain HTML Table".

8. Change the maximum number of hits field from its default of 100 (or whatever it is now) to either a value equal to or greater than the number of hits you expect, or to a blank, which stands for "infinite number".

9. The maximum time setting is disregarded for background searches.

10. Click on either one of the "Start Search" buttons.

11. If you are not yet logged in, or have not even established a User Id yet, the system will prompt you to do so at this point. Follow the instructions on the screen and in the Help page. Cookies are not mandatory but will make your work with stored hit lists more convenient. If you were not logged in yet but are now, click the Start Search button again on the Query Form.

12. In the Status frame, you should now see the message "Executing query as background job. Check your archives for results after a few minutes."

13. Click on the "List Mgr" button in the top button bar. The new search should appear at the bottom of the list of stored searches you may already have. (For this search, the results should appear within approximately a minute. For more complex searches, you may have to wait several minutes.)

14. Click on the radio button (either Selection 1 or Selection 2) for this search and then click on the Retrieve button.

15. The Hitlist pabe should come up, with the title "Operations with this Data Set of 81 058 Structures:".

16. In the Data Retrieval section, set the Format to "SDFile" (default), assuming this is the format you desire.

17. In the Fields selector, make sure that you have selected the following fields to be included in the SD file (use the Ctrl key to add to the selection):

NSC number
All names
CAS number
SMILES string

18. If you are interested in (calculated) 3D coordinates or in having the hydrogens being removed from the structures, check the appropriate check boxes.

19. Click on the Retrieve button.

20. After a short while, a standard Web browser dialogue box will appear; choose Save File with a file name and location of your choice.

Example 3: Screening the NCI Database for Potential New Structural Motifs, by Querying with Predicted Activities and Other Criteria. As an example, we want to obtain some new structural ideas for potential HIV protease inhibitors. We want to limit ourselves to molecules that have not yet been tested in the NCI anti-HIV screen, are of drug-like size, and avoid certain very common classes of compounds against this target.

1. On the Query Form pane, choose "PASS Prediction Range.." as the first Query Type. A separate PASS Selector window will pop up.

2. Select the activity type "Protease inhibitor" (this is the broadest applicable class), make sure that in the Query Probability Field, "Activity" is selected (default), and click on the "Transfer to Form" button. This should bring you back to the Query Form pane.

3. In the suboption field, you will see the internally encoded PASS activity type [something like E_PASS_DATA_PA(55)]—do not change this. The default value range for PASS activities is 0.7−1.0. To allow for structures that are more dissimilar from the PASS training set structures, we lower the lower bound from 0.7 to 0.5.

4. To exclude compounds that have already been through the NCI anti-HIV screen, choose "Data Availability Constraints..." as the Query Type in the second query row, select "AIDS screening" as the suboption, and click on the Negate check box for this query row (or, equivalently, change the default Query Data Value from "yes" to "no").

5. In the third query row, choose "Molecular Weight Range" as the Query Type and type "200−800" in the Query Data Value field.

6. Many of the known HIV protease inhibitors are peptide-like molecules. To reduce the number of such molecules in the hit set, we exclude a specific substructural motif: two consecutive amides. (Other strategies could be devised.) Click on the "Editor" button in the fourth query row. In the Java Molecular Editor, draw the appropriate structure—the backbone of a dipeptide—and click on the "Transfer to Query Form" button.

7. In the Query Data Value field, a SMILES string such as CC(NCC(N)=O)=O should appear. Click the Negate check box on for this query row.

8. To make sure that this search can retrieve all possible hits in the entire database, increase the maximum number of hits to 1000.

9. Leave the output format at, or set it to, its default value "HTML Table with Samples".

10. Click on either one of the "Start Search" buttons.

11. With the current data set, this search yields 410 hits, tabulated in the Hitlist pane. To prevent very large result sets overwhelming the Web browser, the Hitlist pane displays only sets of 250 structures at a time. Additional portions can be requested by clicking on the "Display Next Hits" button at the bottom of the page.

12. To get a better overview over the structures we obtained, select "GIF Image Gallery" as the option in the Visualization field, and click on the "Display" button to the right of it.

13. In the Display pane, a list of the structures, including 2D structure drawings and some additional data, should appear.

14. Scroll down this list to see what structures came out of this search. Each NSC number is, again, a live link that will bring up the Detail pane with the complete data available for each compound. Where available, display of screening data can be requested as a separate table in the Detail pane via another live link. (In this case, this can only be cancer screening data since we excluded anti-HIV screening in the search.)

15. Return to the Hitlist pane and click on the "Display Next Hits" button to see the remaining 160 structures.

16. Repeat steps 12−14 for this portion of the hit list.

(It should be noted that these are examples given to illustrate the capabilities of the system and would not necessarily be optimized searches for a real drug development project.)

CONCLUSION

We have implemented a public browser interface to the Open NCI Database structures and associated data, which is recognized as the largest publicly and freely accessible set of diverse structures. We have aimed at providing an open and very flexible access to this data store. It is primarily geared toward the expert chemist who wants to use the data for scientific studies.

One overarching design goal has been to develop a system that is as open as possible in terms of user access to the data. We have tried to realize all five levels of data accessibility that we feel should be a natural characteristic of all Web-based structure databases on the Internet, which is, after all, an open medium by definition.

1. Visualize—i.e., display of results on the screen, be it of structural information or of textual data in a tabular format.

2. Export—i.e., saving, to the user's computer, of individual structures, possibly together with associated data sets., data tables, etc.

3. Dump—i.e., saving, to the user's computer, of entire hitlists as results of searches, all the way up to substantial subsets of the entire database (or even, if the user so desires, the entire database in a chunkwise manner).

4. Download—i.e., downloading of the raw data as bulk files, possibly in different formats.

5. Link—i.e., enabling direct linking, with the results obtained, to other services on the Web.

By implementing the wealth of query options described, in combination with the extensive set of data retrieval capabilities, visualization features, and links to external databases and searches, we have tried to create a new benchmark for structure databases on the World Wide Web. We believe that our approach to chemistry database interfacing is better fulfilling the above criteria in terms of its ability to deliver computer-readable data in many different formats to the chemist than most other structure databases on the Internet, which are generally designed almost exclusively for human eyes. Downloading single structures one by one in a proprietary structure editor format is not a viable solution for the generation of data sets for studies. We hope that this data set and its interface can become a primary source of test and study data sets used to develop structure−activity relationship and property prediction methodology.

Our plans for the future are to develop this database into a benchmark for structural properties. For example, we have already included two different log *P* computational methods which can be compared. We envision that more partners compute all kinds of descriptors on the structure set both to contribute to our public data set and to demonstrate the usefulness of their methodology. We also would like to see data mining tools and visualization techniques linked to this database in order to allow online analysis of various structure−activity relationships.

ACKNOWLEDGMENT

We thank Dan Zaharevitz of DTP, NCI, for making the data of the Open NCI Database publicly available by putting them up on the World Wide Web. We are thankful to Johann Gasteiger for his support in this project. The following persons and companies have gracefully allowed us to freely include their data or code in our service: Peter Ertl, Novartis; Advanced Chemistry Development (ACD Labs); Markus Wagener, Organon; Vladimir Poroikov and Dmitrii Filimonov, Russian Academy of Medical Sciences; and Volkmar Vill, Liquid Crystal Group Hamburg. The PASS project was supported by a Grant (RCI-Z064) from CRDF. The Computer Chemistry Center of the University of Erlangen-Nuremberg is thanked for hosting the German mirror of our service, as is the Center for Molecular Modeling at the Center for Information Technology, NIH, for hosting the server of the previous version of the service. The support of the Deutsche Forschungsgemeinschaft for one of us (W.D.I.) is acknowledged.

NOTE ADDED IN PROOF

DTP has very recently released a newly generated 2D structure file[30] for which the attempt has been made to reconstruct the correct stereochemistry for as many compounds as this was possible. This new structure set will be incorporated in the next major update of our Enhanced NCI Database Browser.

REFERENCES AND NOTES

(1) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI Open Database With Seven Large Chemical Structural Databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.

(2) http://dtp.nci.nih.gov.

(3) Milne, G. W. A.; Miller, J. A. The NCI Drug Information System. 1. System Overview. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 154−159.

(4) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P.; Hammel, M. J. The NCI Drug Information System. 2. DIS Pre-Registry. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 159−168.

(5) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P. The NCI Drug Information System. 3. The DIS Chemistry Module. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 168−179.

(6) Milne, G. W. A.; Miller, J. A.; Hoover, J. R. The NCI Drug Information System. 4. Inventory and Shipping Modules. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 179−185.
(7) Zehnacker, M. T.; Brennan, R. H.; Milne, G. W. A.; Miller, J. A. The NCI Drug Information System. 5. DIS Biology Module. *J. Chem. Inf. Comput. Sci*. **1986**, *26*, 186−193.
(8) Zehnacker, M. T.; Brennan, R. H.; Milne, G. W. A.; Miller, J. A.; Hammel, M. J. The NCI Drug Information System. 6. System Maintenance. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 193−197.
(9) http://dtp.nci.nih.gov/docs/dtp_data.html.
(10) Limited search capabilities in subsets of the NCI database are available at DTP's Web site, at http://dtp.nci.nih.gov/docs/dtp_search.html.
(11) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical-Structure File Formats Used by Computer-Programs Developed at MOLECULAR DESIGN LIMITED. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.
(12) http://dtp.nci.nih.gov/webdata.html.
(13) http://129.43.27.140/ncidb2/download.html.
(14) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, S.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach Toward Modularity and Flexibility. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 109−116.
(15) A comparison in tabular format of structure-based, searchable, small-molecule databases on the World Wide Web has been posted on our server at http://cactus.nci.nih.gov/ncidb2/chem_www.html.
(16) Should the URLs be found not to be valid any more, please contact the authors.
(17) http://dtp.nci.nih.gov/docs/cancer/cancer_data.html.
(18) http://dtp.nci.nih.gov/docs/aids/aids_data.html.
(19) Weininger, D. SMILES, A Chemical Language and Information-System 0.1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(20) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES .2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(21) Private communication from Daylight Chemical Information Systems, Inc.
(22) http://www2.ccc.uni-erlangen.de/software/corina.
(23) Sadowski, J.; Rudolph, C.; Gasteiger, J. The Generation of 3D-Models of Host-Guest Complexes. *Anal. Chim. Acta* **1992**, *265*, 233−241.
(24) Poroikov, V. V.; Filimonov, D. A.; Budunova, A. P. Comparison of Results of Prediction of a Spectrum of Biological-Activity of Chemical-Compounds by PASS Computer-System and by the Experts. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy* **1993**, 11−13.
(25) Mannhold, R.; Dross, K. Calculation Procedures for Molecular Lipophilicity: a Comparative Study. *Quant. Struct. Act. Relat.* **1996**, *15*, 403−409.
(26) http://www2.ccc.uni-erlangen.de/software/cactvs/index.html.
(27) Tool Command Language, an easily extendable scripting language; see, e.g., Osterhout, J. K. *Tcl and the TK Toolkit*; Addison-Wesley: ISBN 020163337X.
(28) The principal deployment platforms for the service at the time of this writing are GNU/Linux at the LMC and IRIX at the Computer Chemistry Center. Extending the CACTVS file handling for large files with 64-bit positioning and size was straightforward on IRIX with the XFS file system but necessitated quite a bit of patching on Linux. First, we had to use an inofficial prerelease Linux 2.4 kernel in order to obtain the necessary basic operating system support for files larger than 2GB. At the time of this work, semifunctional patches to provide the large file support (LFS) were available in the experimental kernel which is part of the SuSE 7.0 Linux distribution. Nevertheless, while the basic Ext2 file system with large file support worked more or less as expected, many operating system support routines, such as quotas, limit checks, etc. were found to not yet have been properly adapted. For example, we were unable to actually write files larger than 2GB except as root user. For the production of the database this was acceptable. However, to be able to read these large files as a nonroot user on a system which was configured to be reasonably secure, a small kernel patch had to be developed. Additionally, the C runtime library was found to contain a number of errors with respect to handling large files, which had to be circumvented. For example, the ftell64() function in the distributed C runtime library, needed to determine the current file I/O position, returned negative values when the file was positioned beyond the 2GB position, despite setting the appropriate compilation flags to activate the LFS execution environment. Even after solution of these problems, retrieval of individual structure records was initially found to be very slow and dramatically increasing in latency with increasing offset from the database file beginning [higher NSC/record numbers]. It turned out that the performance of the system call mmap64(), which affects memory paging, is much less optimized in early 2.4 kernel versions of Linux when compared with, e.g., IRIX/XFS. A workaround for this problem was achieved by certain changes in the format of the database file. The retrieval speeds on the Linux and the IRIX servers are now very similar and are typically in the 1-s range for any NSC number. It is hoped that future versions of Linux will solve these problems on the operating system level.
(29) Nicklaus, M. C.; Neamati, N.; Hong, H.; Mazumder, A.; Sunder, S.; Chen, J.; Milne, G. W.; Pommier, Y. HIV-1 Integrase Pharmacophore: Discovery of Inhibitors Through Three-Dimensional Database Searching. *J. Med. Chem.* **1997**, *40*, 920−929.
(30) A file at URL ftp://dtpsearch.ncifcrf.gov/dec01_2d.bin, available on the DTP web page http://dtp.nci.nih.gov/docs/3d_database/structural_information/structural_data.html.

CI010056S