HAD: An Automated Database Tool for Analyzing Screening Hits in Drug Discovery

Jian Shen*

Aventis Pharmaceuticals Inc., 1041 Route 202-206, P.O. Box 6800, Bridgewater, New Jersey 08807-0800

Received April 10, 2003

Collecting, organizing, and reviewing chemical information associated with screening hits are human time-consuming. The task depends highly on the individual, and human errors may result in missing leads or wasting resources. To overcome these hurdles, we have developed a decision support system, Hits Analysis Database (HAD). HAD is a software tool that automatically generates an ISIS database file containing compound structures, biological activities, calculated properties such as clogP, hazard fragment labels, structure classifications, etc. All data are processed by available software and packed into a single SD file. In addition to search capabilities, HAD provides an overview of structural classes and associated activity statistics. Chemical structures can be organized by maximum common substructure clustering. The ease of use and customized features make HAD a chief tool in lead selection processes.

INTRODUCTION

Chemical information including structure, physiochemical properties, and biological activities are vital to modern pharmaceutical researches. With advances in genomics, high-throughput screening (HTS), and combinatorial chemical synthesis, it is increasingly difficult to retrieve, integrate, and analyze all data from various sources. The situation becomes critical in lead discovery where a few promising compounds are derived from experimental or computational screenings of million compounds. Decision on progression of an active compound (hit) is made based on biological activities, structural features, potential side effects, pharmacokinetic properties, etc. Without a consistent and standard way of collecting and organizing these data, the decision process can be delayed, and human errors are therefore difficult to avoid.

A number of software vendors or pharmaceutical companies are developing Web-based or client-sever based technologies[1−3] to deal with the challenge. These tools can link multiple databases to perform data retrieval and certain analyses. Several drawbacks, however, limit their acceptance in hit analysis. First, these integrated technologies are designed to accommodate a wide range of data inquires with limited functionality. In particular, most software lacks a convenient structural analysis tool warranted in lead finding. Second, despite user-friendly user interfaces, most tools require extensive training for lab scientists who review and analyze those hits. Frequent software updates or revisions make training and technical support a constant workload. Third, there is great financial cost associated with the initial purchase and installation, subsequent customization, and maintenance for software and hardware. On the other hand, existing chemical modeling software with similar or better functionality has not been fully utilized. Last but not least, the reliability of a new system could take years to accomplish due to the complexity of chemical information. In today's competitive drug discovery environment, quick deployment of a reliable and practical system to meet research demands is of paramount importance.

* Corresponding author e-mail: jian.shen@aventis.com.

Aiming to speed up the lead selection process, we have developed a simple yet versatile decision support system (DSS), Hits Analysis Database (HAD).[4] HAD is a computer application that automatically generates a local ISIS database containing compound structures, screening activities, calculated properties such as clogP, hazard fragment labels, and structure classifications. All data are processed by the available software and packed into a single SD file, which can be imported into the ISIS database and other desktop applications. In addition to search capabilities, HAD provides an overview of maximum common substructure (MCS) classes and associated activity statistics. Diversified chemical structures can be reorganized according to their structural resemblance. The ease of use and sufficient functionality make HAD a widely used computer application in our lead selection processes.

DESIGN AND METHOD

A DSS for lead selection should help a user identify promising lead series, discover potential problems and solutions, and make a recommendation to advance or to stop the compound progression of a chemical series. The primary design objective of the DSS is to streamline the existing nonstandard manual operations in collecting and organizing hit data. The typical manual operations before hit evaluation include the following:

1. Receiving HTS hit IDs and activity data in the form of spreadsheet;
2. Retrieving chemical structures from the corporate compound database;
3. Merging the structure and activity data together in the form of a spreadsheet or database;
4. Collecting other data including calculated properties;
5. Grouping similar structures together with their activities.

After surveying a number of options, we selected the chemical database ISIS/base (MDL information System, Inc.) as the desktop application and user interface for data collection and analysis. ISIS/base has been widely used for handling chemical structures and associated data in our

10.1021/ci034067s CCC: $25.00
Published on Web 07/11/2003

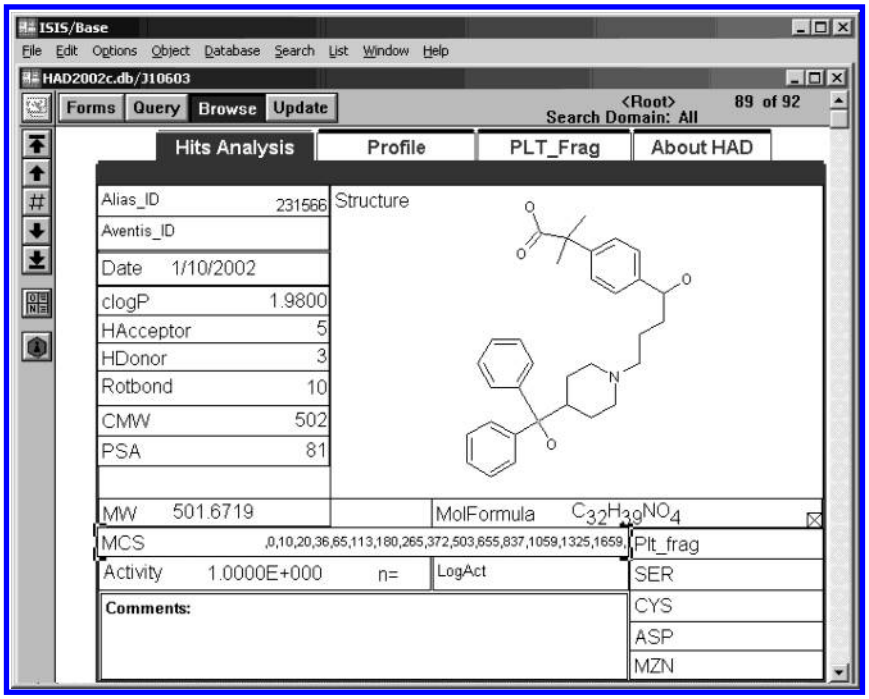


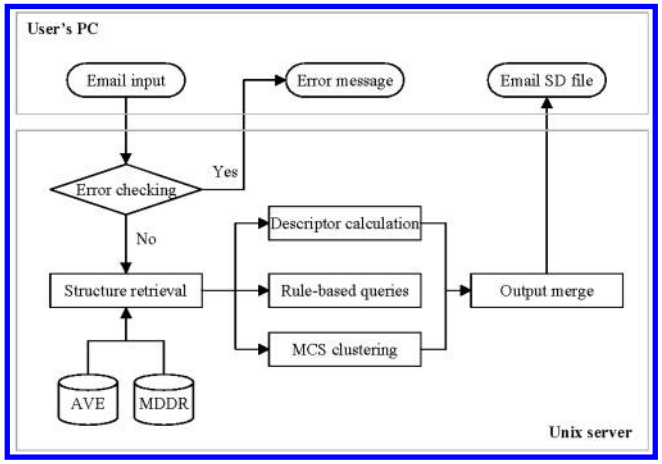
Figure 1. The desktop window of HAD/Isis showing a MDDR (MDL information System, Inc.) compound in Hits Analysis form. Other button-activated forms provide additional experimental input as well as a lookup table for unwanted fragments and some hints on using HAD.

organization. Thus, a favorable user acceptance is expected because of the familiar interface and functionalities. Furthermore, the database can be linked to Spotfire[5] for data and structural visualization and connects to other chemical modeling and spreadsheet applications through a SD file (a well adopted chemical data format). Financially speaking, there is no additional cost for the acquisition, training, and support of ISIS/base and processing software.

We use molecular modeling packages Sybyl and Unity (Tripos Inc.) to populate HAD with the relevant structures and data. These Unix-based packages have been known to handle small molecules and chemical structures very well. A nice feature of Sybyl is its MCS clustering, which categorizes chemical structures according to their structural resemblance. Besides visually appealing, this function is very useful in validating hits and establishing structure−activity relationship (SAR).

HAD data generation and workflow is shown in Scheme 1. The system is automated with Unix scripts and can be started by a user's e-mail request, which specifies the project ID, compound ID, and associated activity such as IC_{50}. When the process is finished, the user will receive an e-mail with an integrated SD file containing all requested structures and associated data. This SD file can be imported into a custom designed ISIS/base template or Excel spreadsheet for browsing and analysis.

Scheme 1



The HAD process retrieves compound structures from the Unity database, which is updated automatically from our global corporate database (ISIS). Searching multiple databases is possible, and redundant structures can be eliminated. Over 95% of the CPU time is spent on MCS calculation, which varies depending upon the number of compounds, the structural diversity, the least number of compounds in each cluster, etc. Most HTS hits, up to several thousand compounds, can be completely processed within a few hours to over a night. A faster algorithm for MCS clustering is certainly warranted. Although a SGI Origin computer with 12 R10000 processors is used for HAD nonexclusively, a single R10000 processor has ample power to handle all internal requests.

USAGE AND APPLICATIONS

The desktop HAD retains all ISIS/base functions including searching and sorting, see Figure 1. The data fields are designed to meet the requirements for our lead selection process. In addition to structure and activity, it provides molecular descriptors such as PSA[6] (polar surface area) and those used in Rule of Five.[7] These descriptors have been shown to be useful in predicting pharmacokinetic properties of compounds. The other fields are described in the following sections.

Unwanted Fragments. A compound with potential toxic or hazard chemical fragments should be removed from lead consideration. A list of these unwanted fragments has been

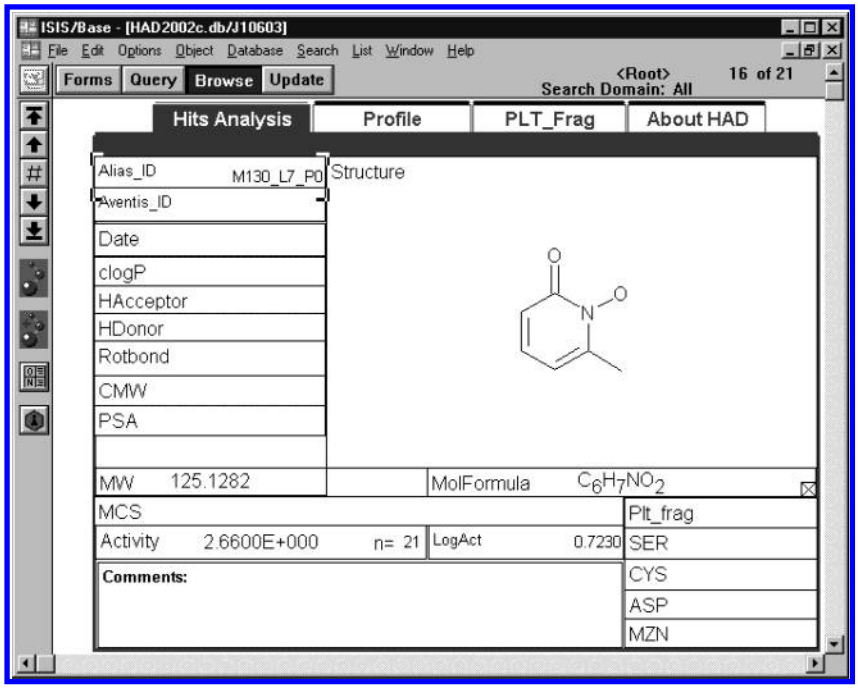


Figure 2. A window view of HAD showing a core structure associated with 21 actives from a screening.

determined and ranked by internal experts. During the HAD process, these fragments are searched against every retrieved structure. The most hazardous hit is flagged in the field of "PLT_frag" with an ID number. A button-activated lookup table will display all IDs and associated chemical names. Because an analogue with an unwanted fragment may still be useful to establish SAR within a series, HAD only alerts users to the information rather than removes those compounds from hits. The numerical field allows filtering out all or a subset of unwanted fragments easily.

Structure Cluster. Calculated using Distill of Sybyl, the "MCS" field is a string of numbers presenting compound clusters at different levels. By sorting this field, structures can be rearranged by MCS series rather than by the default alphabetic order of registration IDs. Searching a specific number in the MCS field will retrieve all compounds associated with a common substructure. Therefore, a chemist can speed up the hit evaluation process by examining each structure series first and then removing nondrugable series and focusing on promising ones.

HAD stores a set of core substructures and activity statistics as additional records for browsing hit series and substructure searching. For example, Figure 2 shows that M130_L7_P0 is a core for a set of 21 compounds with an average activity of 2.66 (a user can modify the name of activity with a unit). Searching "130" (the core ID) in MCS field will retrieve all associated 21 compounds. A summary of core structure and activity statistics—including mean, maximum, minimum and standard deviation—for each chemical series can be obtained using the ISIS/base command "Save Structure−Activity Table". The saved MS Excel spreadsheet allows a user to identify promising chemical series with greater ease. In addition, users can cut and paste a core structure directly to other ISIS databases for substructure searching. This data-mining strategy is used for seeking similar compounds that are not in the screening set.

Rule-Based Predictions for Ligands. A biological activity against a therapeutic target is not the only criteria for a lead. An ideal drugable compound should minimize the interactions with other receptors thereby reducing potential adverse effects. Screening against a panel of receptors can reduce this risk. In addition to the cost, however, it is impossible to screen against all receptors. One solution looks for the compound history, i.e., its activities against other drug targets in the past. Due to the limited in-house data, the specificity of a chemical class is difficult to assess without a lengthy investigation. Alternatively, one can use an artificial intelligence (AI) system[8] to recognize chemical patterns that are known to cause certain biological activities. This approach has been used to develop several computer systems for predictions of toxicity and metabolism.[9]

Similarly, we have developed a rule-based protease inhibitor prediction system[10] and implemented it in HAD. The system recognizes over 90% of annotated protease inhibitors in MDDR. The field boxes SER, CYS, ASP, and MZN in Figure 1 display 1 when a structure matches queries for inhibitors of serine, cystein, aspartic, and metallo proteases, respectively. The detail of the system is beyond the scope of this paper and will be described elsewhere.

The system serves as a validation tool for protease-targeted projects. Flagged hits confirm experimental results. Nonflagged hits could be false positives or new scaffolds. The latter is of warranted interest in the lead discoveries. Conversely, the system alerts potential selectivity issues for nonprotease-targeted projects. Figure 3 shows that one active series in a GPCR-targeted screening has several predicted protease inhibitors. Confirmed by historical data, this series is removed from lead consideration.

Beyond HTS. Selecting compounds based on multiple criteria is a very common decision process in the various stages of drug discovery. HAD provides a convenient tool when structures and associated data are needed. For example, HAD can be used to select virtual HTS[11] (vHTS) hits. The vHTS has gained popularity in recent years due to the advancement of computer technology and demand for reducing expenses in drug development. Although vHTS

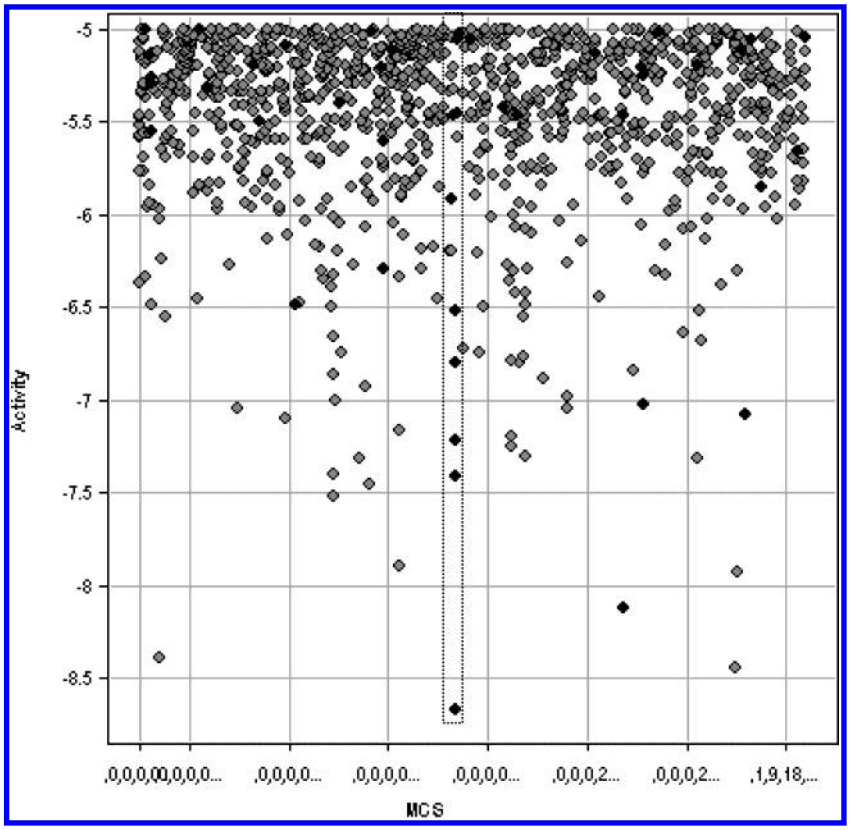


Figure 3. A Spotfire plot of logIC50 vs MCS for A GPCR targeted project. The highlighted black dots specify SER = 1. The dashed rectangular in the middle encloses a series of active compounds associated with a common substructure. The MCS label is too long to be displayed in the plot but can be viewed in Spotfire using a pointing device.

reduces the number of compounds for experimental screenings, the number of virtual hits may still be too large and often contains many nondrugable compounds, which should be removed before experimental confirmation. HAD is used here to save the cost of compounds and assays, as it does for HTS hits.

Advanced users, such as computational chemists, can also benefit from using HAD. All submitted data and gathered structures are saved under each project on the HAD server. Users can import these data into a range of Unix-based modeling software for advanced study, such as QSAR or computational docking. Unproductive transmitting, copying, formatting data and structures thus can be avoided. In fact, this is one of author's motivations for developing the system. As a convenient tool for structure−activity merging and organizing, HAD is used to study many SARs and to develop new tools such as the rule-based protease inhibitor prediction systems discussed above.

Performance. The ease of use as well as its simple and reliable performance makes HAD a main software tool for lead selection. Figure 4 shows the increasing usage of HAD for the past 18 months. The system is very stable yet easy to modify in order to meet new demands or environmental changes. During the development and implementation of HAD, Aventis chemical database underwent several changes, from separate databases to a single one, from multiple IDs to unified IDs. Each time, HAD is able to adopt the new standard quickly and go beyond. For example, a user can submit mixed compound IDs (new and historical IDs) to HAD and retrieve all relevant structures. While in other systems, a user has to separate these IDs and request them

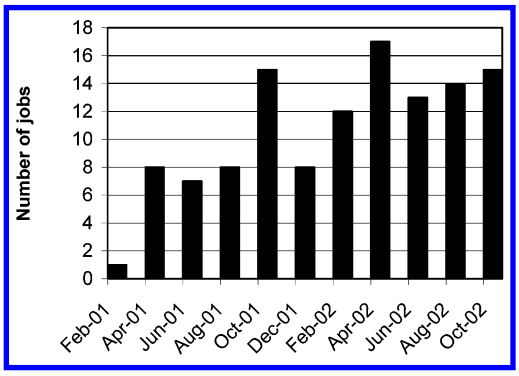


Figure 4. The number of HAD jobs completed during the last 18 months.

in different fields. In the future, we plan to implement other validated AI systems and historical data retrievals.

The complexity of drug discovery requires both human intelligence and better technologies. How to apply both in every stage of the research becomes the challenge to scientists and technology professionals. While HAD is used to collect, process, and organize data more efficiently in our lead selection, chemists' ability in pattern recognition and knowledge about chemicals are still fully utilized. The author believes that this balanced approach will lead to more productive drug discoveries.

CONCLUSIONS

HAD performs structure and knowledge retrieval, data organization and merging, and property calculation. Manual operation has been minimized to increase the productivity

in lead discovery. The familiar desktop interface and minimal training make this tool widely accepted by medicinal chemists. Meanwhile, users are able to analyze chemical information with state-of-the-art molecular modeling and data visualization software. Based on available software components, the decision support system is inexpensive to build and easy to maintain. The capacity and functionality of HAD has met current data demand for hits analysis and other compound selection processes in our organization.

ACKNOWLEDGMENT

The author thanks John Hunkins, Bruce White, Ramesh Rachapudi, Isabelle Morize, and Mary Windhorst for technical support and suggestions.

REFERENCES AND NOTES

(1) Brown, R. D.; Guner, O. F.; Hahn, M.; Li, H. *Web-Based Productivity Tools for Chemists: WebLab(Tm) Medchem and Diversity Explorer*; Book of Abstracts, 216th ACS National Meeting, 1998; CINF-051.

(2) Leach, A. R. *Web-Based Tools for Compound Selection, Library Design, and Compound Acquisition*; Book of Abstracts, 221st ACS National Meeting, 2001; CINF-017.

(3) Coles, S. *Developing HT Information Systems, a Modular Design*; Book of Abstracts, 224th ACS National Meeting: 2002; CINF-067.

(4) Shen, J. *Automated Database Tool for Analyzing Screening Hits*; Book of Abstracts, 221st ACS National Meeting, 2001; CINF-079.

(5) Demesmaeker, M. Decision Analytics in Life Science Discovery through Visual Integration of Chemical and Biological Information on the Desktop. In *Rational Approaches to Drug Design, Proceedings of the European Symposium on Quantitative Structure−Activity Relationships, 13th, Duesseldorf, Germany, Aug. 27-Sept. 1, 2000*; Prous Science: Barcelona, Spain, 2001; pp 506−511.

(6) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and Its Application to the Prediction of Transport Phenomena. 2. Prediction of Blood-Brain Barrier Penetration. *J. Pharm. Sci.* **1999**, *88*, 815−821.

(7) Lipinski, C. A. Drug-Like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharm. Toxicol. Methods* **2001**, *44*, 235−249.

(8) Luger, G. F.; Lewis, J.; Stern, C. *Brain Behav. Evol.* **2002**, *59*, 87−100.

(9) Marchant, C. A.; Combes, R. D. Artificial Intelligence: the Use of Computer Methods in The Prediction of Metabolism and Toxicity. In *Bioactive Compound Design*; Ford, M. G., Ed.; Bios Scientific: Oxford, 1996; pp 153−162.

(10) Shen, J.; Hong, J.; Morize, I. *Knowledge-Based 2D Structure Queries for Searching Protease Inhibitors. Abstracts of Papers, 223rd ACS National Meeting*; American Chemical Society: Washington, DC, 2002; COMP-225.

(11) Stahl, M.; Rarey, M.; Klebe, G. Screening of Drug Databases. *Methods Principles Med. Chem.* **2002**, *14*, 137−170.

CI034067S