Decision Support Systems in Drug Discovery

Jian Shen*, Valery Polyakov, Ramesh Rachapudi and Tony You

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

Abstract: Decision support systems have been widely used in drug discovery owing to the complexity of data and information involved. In this article, we review both commercially available packages and publicized in-house made systems. All selected systems should be able to handle chemical structures and to organize information for decision makers, more specifically for medicinal chemists. Although we do not rank these system, pros and cons of these systems will be discussed.

INTRODUCTION

In an organized drug discovery, compound progression from high throughput screening (HTS) to preclinical study involves numerous decisions made by scientists, medical and patent professionals, and research managements. Unlike many engineering or business decisions where solid data can derive a conclusion, decisions in drug discovery are usually made without complete data or with conflicting data. For example, facing a compound with good binding affinity but marginal bioavailability in one animal model, one can decide 1) making another analog for binding and bioavailability testing; 2) submitting the same compound for testing the same animal model, 3) submitting the same compound for testing a new animal model or 4) abandoning the series and start another promising one. A decision is usually made based not only on the data but also on other factors such as knowledge about chemical modification, perception of data quality, feasibility of another animal model, personal experience of past successes and failures, chemical patentability, budget constraint, etc. Because this kind of decision requires highly skilled decision makers and multiple criteria to be considered, and the process itself is so complicated that no one can precisely describe it, it is called semistructured or unstructured decision [1,2].

For semistructured or unstructured decisions, human intelligence is still unmatched by the most sophisticated computer technologies such as Deep Blue [3], an IBM's massively parallel computer system that was designed to play chess at the grandmaster level. The information technology, however, can facilitate the decision-making by integrating data from various data sources, extracting relevant information and present it to decision makers in an understandable way, and providing multiple-criteria decision-making models. Without a computerized system, these functions have to be carried out manually by the decision makers. For example, before selecting a promising series of compounds for further study based on HTS data, a scientist needs to perform at least the following tasks: 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 from other computers; 5.) Grouping similar structures together with their activities. These tedious, repetitive and less intelligent-demanding tasks could take a substantial portion of effort and time in a decision process. Human errors are difficult to avoid during the data-structure integration and transformation. In contrast, a computerized decision support system (DSS) handles these tasks much more efficiently and accuratly than human does.

A generic DSS is an information system consisting data source, data integration, data warehouse, data transformation. The data source includes various databases, spreadsheets and text documents [1,4,5]. The data integration is computerized processes that extract relevant data from the data source and merge them into the data warehouse. The data warehouse or data mart, the foundation of DSS, is a read-only analytical database designed for specific decision-making. Finally, the data transformation or user interface is a set of applications that allow users to analyze data, manipulate data and generate information. Because the DSS can provide decision makers with relevant data, underlying information and projected trends, it has been widely used in many business and operational processes for years. However, those systems are mostly designed for structured decisions where the decision criteria are simple, and less skilled decision makers are required.

Many chemical database software [6-14] developed in 1980s might be considered as primitive DSSs in drug discovery. They were used to store both chemical structures and biological data and provide some analytical functions such as querying, filtering and sorting for different type of data. MDDR [15] (MACCS-II Drug Data Report) is a well-known example. The databases, however, ware designed primarily for storing structures and general browsing, not for supporting decision-making. Their capacities had reached the limit when enormous HTS biological data and various ADMET assays became available in later 1990s. Separate databases [16-20] were used to store the biological data in many pharmaceutical companies. The fragmented data sources and legacy databases left by the pharmaceutical industry mergers and acquisitions created a huge obstacle for researchers to access complete and relevant data. The lack of data integration mechanism and standardized information retrieval prolonged many decision processes in compound progression.

To be competitive in discovery informatics and be able to take the real advantage of HTS, genomics and

*Address correspondence to this author at the Aventis Pharmaceuticals Inc, 1041 Route 202-206, P.O. Box 6800, Bridgewater, NJ 08807-0800, USA; E-mail: jian.shen@aventis.com

1389-5575/04 $45.00+.00 © 2004 Bentham Science Publishers Ltd.

combinatorial chemistry [21], many pharmaceutical companies, specialized software developers and academic labs independently or jointly started to tackle the information bottleneck. This stirred a wave of information technology exploration in the enterprise-wide data integrations. Many medium-to-large biotechnology and pharmaceutical companies have developed some kind of information technology platforms to handle the increasing demand for data and information. Some early project-wide data integration and modeling systems such as Weblab Medichem [22] emerged aiming at the decision makers in medicinal chemistry.

In this review, we focus on a number of current DSSs used or can be used by medicinal chemists to support their day-to-day decisions. These decisions usually involve the selection of a compound or a series of compounds, from many others, for further biological testing, synthesis or development. Multiple-criteria including data and models are needed to justify the decisions. We include both commercially available information technology as well as a few in-house developed systems. Some of them may not fully fit the description of generic DSS. But all of them should be able to handle at least 2D chemical structures. They should be able to generate information beyond the raw data.

It is impossible for us to cover all information technologies related to the DSS in drug discovery due to the technical scope and our experience. In addition, there are a number of obstacles for an objective review of the DSS in medicinal chemistry. First, most real world DSSs are proprietary technologies and have rarely been revealed to public. Second, there is almost no user's consensus to a DSS, which is the ultimate judgment of a good system. Third, the DSS is intrinsically an ever-changing system as data and models change in addition to user's demands. Thus, many of our comments could be out-dated when this article is published.

CURRENTLY USED DSS IN DRUG DISCOVERY

There are basically three types of DSS technology used in the compound progression: web, client/sever and desktop. In a web-based system, all components including data integration, data warehouse and data transformation are installed on a powerful web server computer or computers. A decision maker interacts the DSS through general web browsers such as Internet Explore (Microsoft, Inc.) from any desktop computers. In the second system, a set of applications for data transformation is installed on the user's desktop (client) while the rest of the components resides on a sever computer. The third technology is simply a stand-alone system on a PC, which includes both the data warehouse (or data mart in small scale) and data transformation. The data integration can be done by other technology unless the amount of data is small, a typical case for many small drug discovery organizations.

Web-Based

Although it is simple from the user's point of view, the web-based DSS can be very complicated to build. These systems can include multiple server computers and databases across different geographical sites. Sage *et al*. [23] described Lionbioscience' decision support system composed of multi-layer data processing. Its data warehouse SRS is able to extract data from both flat files and Oracle databases. The scalable DiscoveryCenter provides users with capabilities of standard chemical structure searching, property searching, Visualization capabilities. Users are able to share data and models. The authors illustrated how to decide a promising lead series for a hypothetical estrogen receptor project. Based on experimental data and computed property distributions of the corresponding virtual libraries, the best series can be easily identified.

While all results in the above example were displayed through web pages, it is not clear how much data manipulation and modeling can really be performed with the web interface. We recently found that Lionbioscience also provided a desktop tool [24] , the Lead Engine, specifically designed for Cheminformatics. Obviously, the functions of web-based user interface are not sufficient to meet the challenges in lead identification and lead optimization.

Instead of building a massive data warehouse such as Lionbioscience' SRS, IBM DiscoveryLink [25] is able to integrate multiple, heterogeneous data sources into a single virtual database. With one query, researchers can get a cohesive view of results for manipulation, comparison and analysis — while the data itself remains unchanged in separate databases across different platforms. In fact, Lionbioscience has an option to integrate DiscoveryLink technology. Several pharmaceutical companies have used these technologies to build customized enterprise-wide discovery DSS.

Figure **1** describes the architecture of TLC [26], for Target Lead Candidate, developed at Aventis with IBM DB2 Information Integrator. The top section is the user desktop (Client) where the web browser and required components are installed. These applications are generally part of the user desktop, which means they are not new applications that a scientist has to learn. The components in the background are built such that minimal changes are made to the foundation systems. The application sends the data request to the web server, which then transmits the data to the middleware layer that parses out the request and routes it to the appropriate foundation system. When the data are returned from the foundation system, the results are sent back to web server that then returns the results to the client application that made the request. The querying and reporting of the results is accomplished by a business intelligence tool called Brio [27]. The advantage with such architecture is that it is scalable, which means that one can add many foundation systems without disturbing or changing the structure of the existing databases. The user interface of the application presents the condensed version of the underlying foundation systems called data models. There are separate data models (or tables) for chemical, biological, logistical and analytical data all of which contain at least one field that is common to all of them such as a compound id or a batch id based on which data can be linked. When the result data are returned to the reporting tool, they are in the form of a series of rows that need to be arranged into tables and views to make sense of those data. On the web server, each project team will be assigned their own area to organize their queries and resulting views. When the queries are published to the

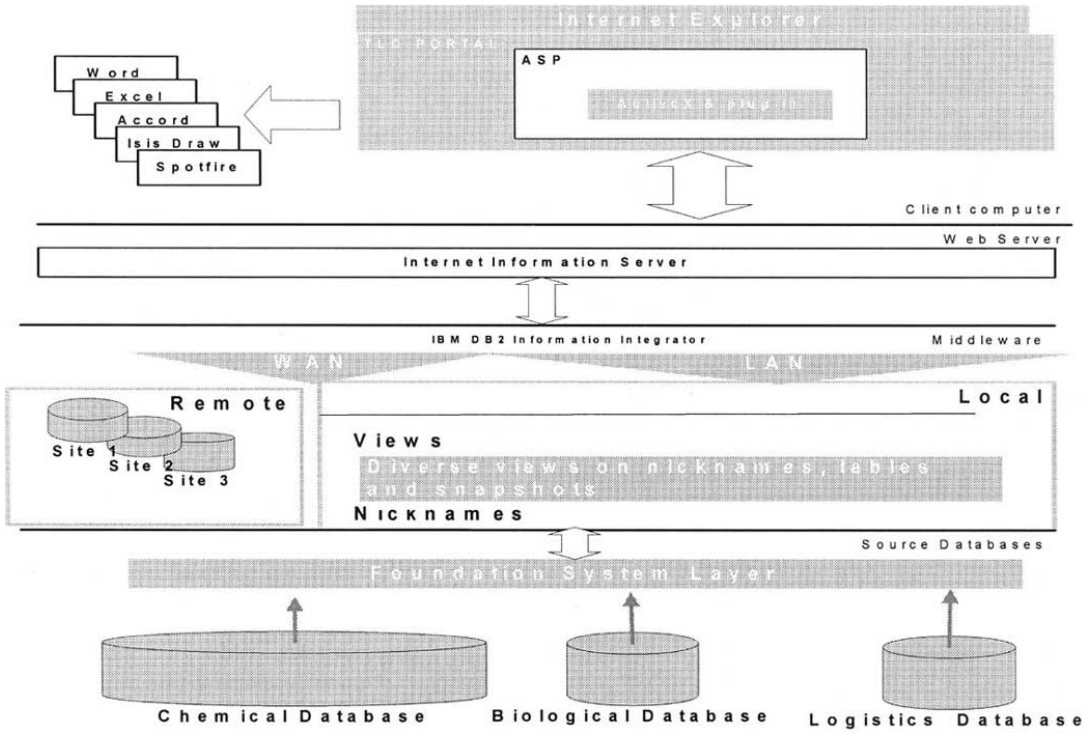

Fig. (1). The web-based TLC system of Aventis.

portal, they are shared among the specific project team, across project teams, across departments and across all global sites.

ChemEnlighten [28,29] (Tripos Inc.) is an intranet-based technology that runs in either Netscape 4.0.5 (or better) or Internet Explorer 4.0 SP1 or better. ChemEnlighten is designed to handle larger data sets generated by today's combinatorial chemistry and HTS program. Its Data Viewer allows tables to be viewed in a "spreadsheet" format that contains structures plus the associated property data or in a grid format that contains only structures. ChemEnlighten integrates visual access to large tables, the capability to generate a range of chemical metrics, and the ability to perform analysis and selections on the data in the table. Over 300 metrics are available for database analysis and filtering. The OptiSim algorithm is one of several techniques available for data filtering and subset selection. OptiSim works on combinations of metrics and mimics selections made using hierarchical clustering. The search in ChemEnlighten is available through UNITY. In order to take the full advantage of the system, one should acquire other software tools offered by Tripos.

Client/Sever

Pipeline Pilot [30,31] from SciTegic Inc. (SPP) is based on the concept of pipelining the virtual molecules, normally one by one, through the series of calculations or tests represented by separate components. Each component would add or remove some calculated properties to the molecule, i.e. molecular weight, LogP, etc, which would be carried with the molecule downstream and could be used for further analysis, for example channeled into a predefined visualization (*i.e.* Spotfire or Accord for Excel) so that the user can make a decision using also visual presentation of the data. Components and pipes can be laid-out visually, greatly simplifying analysis and comprehension of the whole process by a novice user. The created layouts could be saved as proprietary "protocols" and reused by the author or others using the same server later, or exported in an XML format and shared among the users of multiple servers. Thus, SPP is able to communicate with a web service, which makes it easy to plug-in virtually any application into SPP. Figure **2** illustrates the implementation of QikProp [32] as one of the SPP components.

SPP comes with multiple built-in components conveniently organized by their categories and sub-categories visually presented as folders and sub-folders. For example the DataReaders category contains 23 components that allow reading most popular molecular files formats (i.e. SD, MOL2 etc.), read data directly from databases (including ISIS[33]), and read a variety of files that might contain data about molecules (i.e. Excel, comma separated, XML, etc). DataWriters category contains components to write data in the above mentioned file formats as well as some other (i.e. Accord for Excel) formats.

The real strength of the application comes from its tools for data analysis, model building, learning, library enumeration, fingerprints and descriptor calculations. The library enumeration engine is extremely fast and accurate. It does not even fail stereo-chemical assignments that are the weakest point for most of enumeration software packages [34] with the exception of only few vendors like CombiLibMaker [35]. One can do both fragment- and reaction-based enumeration. Users can quickly predict desirable properties by utilizing Bayesian learning component, saving the model, and then applying it to the new data.

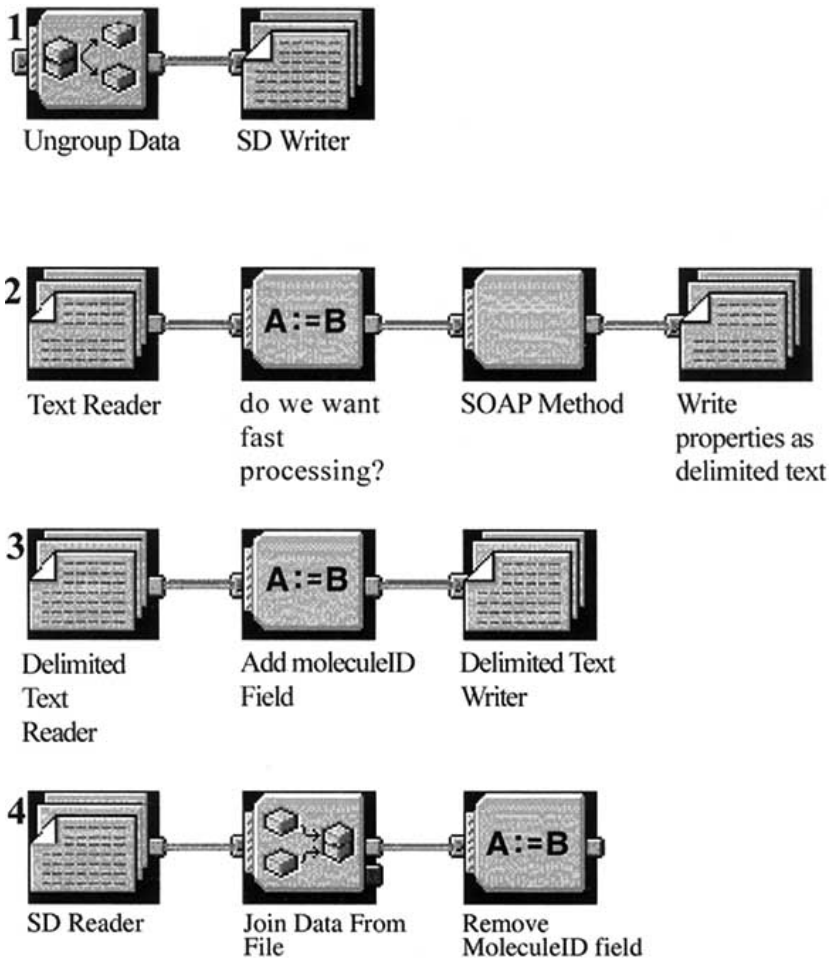

Fig. (2) . SPP sub-protocol that implements calling QikProp web service.

SPP offers ways to cluster the compounds using a maximum dissimilarity method [36]. This method randomly chooses the first center and selects other centers one by one, so that the new center is maximally different form the existing ones. When enough centers are selected, the rest of molecules are assigned to their nearest cluster centers. The SPP component provided for clustering allows clustering on different types of properties that can be either predefined in SPP or supplied by the user. Clustering on numerical properties vs. fingerprint data may lead to very different results. SPP also provides a component that allows viewing molecules grouped by clusters.

Spotfire's DecisionSite 7.0 for Lead Discovery [37-39] provides easy access to standard data repositories such as ISIS databases and IDBS ActivityBase, which is an Oracle database. Data in files such as excel spreadsheet and text file can also be imported directly. Compound structures and biological data from different sources can be merged through a unique key. However, the importing process has to be manually carried out one at a time. This could be cumbersome where multiple structure and activity databases need to be linked. Once data are loaded, the document can be shared with other users.

Spotfire has probably the most impressive data visualization and dynamic filtering capacities. Users have a lot of options to view and display multi-dimensional data. A unique structure viewer, which flashes up a structure when a user moves the mouse cursor over a data point in a scattering plot, has not been seen in other applications.

For chemists, DecisionSite can search both local and remote ISIS databases with substructure, similarity and list. ISIS/draw is called for drawing structures. The list search only allows for registration key IDs, which could spell trouble for databases using different structure ID other than the registration ID. It has been reported that structure searching using CAS SciFinder is also available, which adds another dimension of convenience. Because it is mainly a data visualization tool, DecisionSite does not have structural clustering and other cheminformatics functions. However, once the information is provided, DecisionSite is able to read and display it including non-numerical information such as ISIS key and maximum common substructure (MCS) [40-46] string.

Leadscope [47-52] is a software for systematic substructure analysis using a predefined hierarchy template library containing ~27,000 chemical features. The activity statistics associated with each feature are readily displaced with colors for quickly identifying promising features and compounds. It provides more options on searching based on Boolean combinations of chemical structures (exact, substructure and similarity), property ranges and textual searching.

Similar to Spotfire, Leadscope reads a single SD file with both structures and data or reads a SD file for structures and text file for data. However, non-numerical data field will be ignored. New data can be merged with existing structure but not with existing data. It can search both internal stored structures and external SD files. For efficient data integration and search, it is recommended to establish an internal structural database and update it regularly. This can be done on a server with Leadscope Enterprise.

Leadscope provides scaffold analysis and allows users to choose one of them to build a SAR table. This function can be very useful in focused library design. It also provides interactive filtering capability. However, it only provides online Help with a PDF file, which could be difficult for casual users.

The newest release of ClassPharmer Suite [53-57] from Bioreason emphasizes on structural classification and SAR model by using a recursive-partitioning (RP) [58-60] method. It takes SD file as main data sources. A user has the option to choose primary key field. It will report which record cannot be imported in a log file. It can perform substructure searches on external SD or SMILE files [9].

Bioreason technology produces chemical classes using a MCS-based approach. A unique feature is that a user can restrict the classification to essentially a partition of the compounds or allow compounds with more scaffolds or structural domains to be included in multiple classes described by the different scaffolds. This will enrich the information content of the classes to which a compound legitimately belongs to thereby facilitate multiple SAR studies.

It has many analytical functions such as sorting, interactive filtering and SAR tabling. The distribution charts for each class in a spreadsheet format provide basic statistical information, which can immediately strike user's eyes to identify promising classes. It provides a set of pharmacophore feature for users to build predictive model using the RP method.

Desktop

HAD [61,62] is a simple and economical solution initially developed to handle HTS hits selection. The system utilizes several existing software in chemistry and modeling to merge, organize structure, data, information and knowledge models. Everything is encapsulated into a single SD file and sent back to the users for exploitation and analysis, generally with Isis/base. Gradually, the system becomes a multifunctional information retrieval and analysis system.

The system relies on Unity databases, which can be automatically synchronized with other databases such as ISIS/host, providing structural data. The structure IDs and activity data are provided by a user through an email. The email received by the server triggers the automated process of data integration, information buildup and return of the generated SD file as seen in (Fig. **3**).

The data integration process includes searching multiple structure databases (Unity) for matching IDs and merging extracted structures with the activities. The information buildup includes computing commonly used molecular descriptors such as clogP and PSA, clustering structures based on MCS, and generating various flags to alert hazard chemicals or potential protease inhibitors. The whole process is carried out using software Sybyl in batch mode.

When the generated SD file arrives in the user's email inbox, s/he can import it directly into a pre-designed

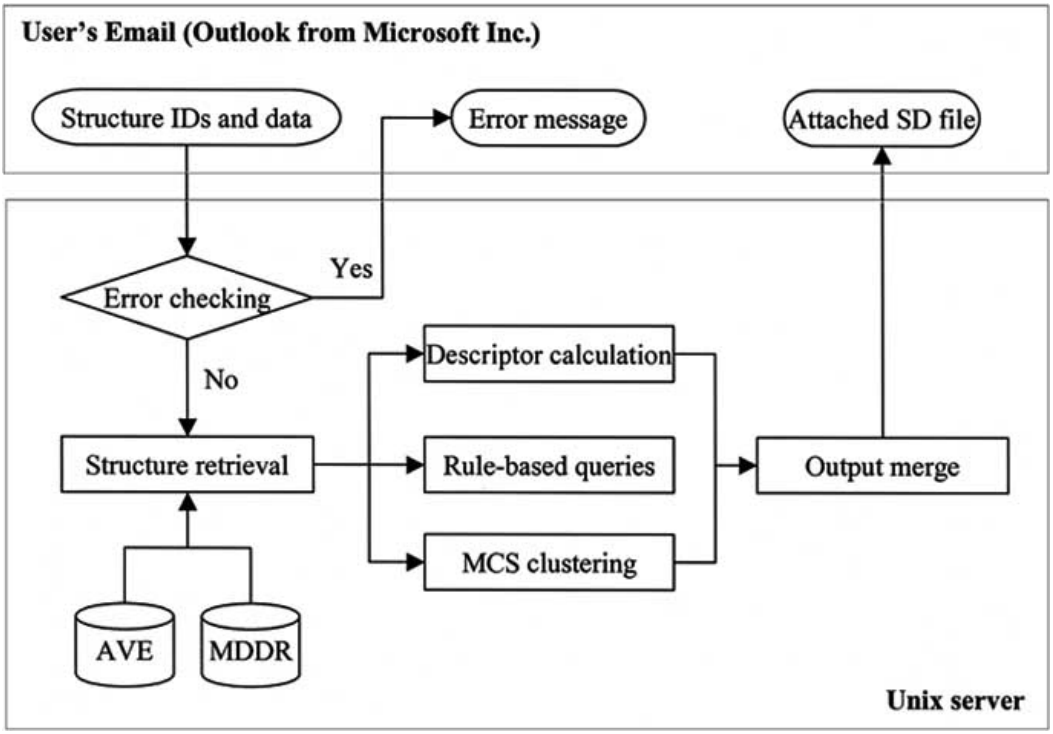

Fig. (3). A schematic view of the Email trigged data integration and information buildup process used in HAD.

ISIS/base template to make a data mart. All functions of ISIS/base can be used to interact with the data and information. The other option is to use Excel for data analysis since it can read SD file. A unique feature in this system is that users can sort chemical structures according to their resemblance. This is especially useful when a set of diverse structures needs to be examined one by one. The structural organization is done by sorting MCS string, which is generated during MCS clustering and saved as a part of molecular records. In addition to using the ISIS/base for data, information analysis, one can directly link it to Spotfire.

Others

There are a number of DSS-like systems reported but with little technical details. Neurogen [63] first revealed its Accelerated Intelligent Drug Discovery (AIDD) platform that integrates all major drug discovery technical components, such as combinatorial/medicinal chemistry, high throughput pharmacological screen, computer aided molecule design and research informatics, to accelerate the processes of lead identification and optimization. It combines the power of massive parallel data processing to highspeed compound synthesis and biopharmaceutical screen to support the decision process in new drug discovery projects [64,65]. Neurogen also transferred the technology to Pfizer's R&D.

Bristol Myers Squibb [66] has implemented SMART-IDEA (Structure Modeling and Analysis Research Tool – Integrated Data for Experimental Analysis), a part of a wider decision-support technology platform that expedites the drug discovery. The system uses Tripos' MetaLayer technology to integrate experimental data with a comprehensive suite of analytical tools that provide a virtual workbench for Bristol-Myers Squibb's in-house chemists and biologists. Through its enhanced computational and predictive capabilities and ability to analyze and visualize data on multiple compounds, SMART-IDEA helps shift the drug discovery process from one that has traditionally been sequential to one that is 'parallel'. Further improvement [67] will integrate Tripos' FormsBUILDER technology, a forms-based querying tool for retrieval and browsing of research data, into the SMART-IDEA application. This form-based searching technology enables scientists to easily customize their own forms and combine queries requesting chemical structures with more traditional data.

Molecular Operation Environment (MOE) [68] from Chemical Computing Group is a nice lightweight tool that allows addressing a lot of issues related not just to Molecular Modeling and Chemoinformatics, but also Bioinformatics. MOE could run on the large number of platforms including different flavors of UNIX, Linux, and Windows.

What elevates MOE to the lever of a DSS, is its ability to access multiple data sources (i.e. its own proprietary format, databases, SD files, etc), convert the data into its internal representation, perform required calculations (i.e. descriptor, properties, or even docking), and even visualize the results in order to simplify the decision making process. The whole process could be customized and automated by using a Scientific Vector Language (SVL). SVL allows manipulating small molecules, proteins, and nucleic acid chains. SVL is an interpretable language that is chemically aware, which makes it a good choice for writing scientific calculations. There is a fairly large scientific community that creates applications in SVL. Nevertheless, as an interpretable language, it has a lot of deficiencies related to the limits on its performance that is especially noticeable if one would attempt to encode some quantum chemistry algorithms or demanding visualization routines. In addition to that, it is almost impossible to find professional software engineers familiar with SVL for the large-scale application development. The company might be much better off introducing a C-like language (i.e. Java or C#) and creating chemically aware libraries for that language.

Although anybody can do some simple operations in MOE, mastering it requires much steeper learning curve then SPP and the product might not be suitable for the most of users outside *in-silica* departments.

CHALLENGES IN DSS DESIGN

Users (Decision Makers)

There are a number of challenges in the selection, design and implementation of a successful DSS for medicinal chemistry. The foremost is the acceptance of the decision makers, mainly the medicinal chemists, who have been using various solutions in the decision processes. It is not unusual to see different opinions about a new DSS. A power user may consider it unnecessary because s/he already has other tools to do the same tasks more effectively. Conversely, a casual user may feel the system is too complicated to learn and use efficiently. A new DSS should make majority of users more efficient and effective in the day-to-day research activities

The data integration, query and user interface are primarily related to the efficiency of a DSS. The goal for this part is to save user's time on non-scientific tasks. A new technology, even if it has fast data processing speed, does not necessarily save the user's time. Different options need to be carefully compared to match user's requirements and skills. For example, we implemented email as the requesting protocol in HAD because of the convenient drag-and-drop operation to attach a SD file to it.

The effectiveness of a DSS is conjugated with the presentation of data, information and knowledge. The key is when and how many. A decision maker may not want to see all the data simultaneously but just a main activity, i.e. IC_{50} or EC_{50}. Too much of data can confuse users. Many biological data have large discrepancies. Historical data can be reported using different units and under different assay conditions. There is an array of computed molecular properties one can feed into a DSS. The adequate amount of data, information and knowledge models in a DSS can only be determined by the system refinement in accord with the responses from users and organizations.

Technology

The Web-based DSSs have been widely implemented in many pharmaceutical companies for drug discoveries. Their main function seems to be on the data integration side, which is the bottleneck in many organizations. Little

desktop support and one place for data are the greatest advantages. For a research organization with multiple geographical sites, this kind of system is a necessity. For these systems to become a true DSS for medicinal chemistry, functions in data transformation such as structure clustering and sorting needs to be enhanced. The web portal technology may be used to add more analytical tools for the end users. Compared with other two technologies, the web-based system is usually not specifically designed for medicinal chemistry except Weblab Medichem [22], which is no longer marketed. Most of them do not provide independent clustering of chemical structures, an important data transformation process that the decision makers can benefit from.

Client/server-based DSSs are commercially available and require less IT supports on both the server and desktop. They mainly focus on data analysis and often provide their own unique proprietary technologies that medicinal chemists find useful in real situations. They can usually connect only to one data source at a time. Unless a user is familiar with all data sources and conversions, preparing needed data with acceptable format may prevent some decision makers from using the DSS. Thus for commercial packages, customization to link internal information systems is needed for medicinal chemists. Some programs cannot merge new data with the existing data internally. These weaknesses need to be improved in the future.

Storing project-wide data only on a server computer has the advantage of sharing data and information. This requires a highly reliable and scalable server and backup system. The level of file permissions may also need to be considered if multiple users are allowed to access the same file. The setting will be difficult for mobile computing or presentation of results at a place where no network exists. Of course, these problems disappear when a user can save the data on the front-end.

The desktop DSS is suited for decision makers who have less data demand in specific processes such as evaluation of HTS actives or a series of compounds with multiple experimental datasets in lead optimization. In fact, the majority of medicinal chemists only need to search but not to analyze the enterprise-wide compound collection. Most desktop DSSs are able to meet this demand. They are less likely to be affected by network outage or server crash. Other advantages include mobile computing and sharing results with other scientists. Most client/server DSSs also provide stand alone desktop versions. Like client/server DSSs, desktop DSSs lacks ability to integrate complicated chemical and biological data automatically unless an additional customization similar to HAD is implemented.

Some desktop DSSs can be quite demanding in terms of PC's memory. This could be additional cost one needs to consider for a wide range implementation. Users should also be aware that all systems need some time for data integration and processing depending on the number of structures and datasets. It is good to give users an indication of how long the process needs to complete, such as a window time bar provided by ClassPharmer Suite and SPP, or simply an email reply by HAD. Thus, the users can plan other research work and will not be frustrated waiting for results.

The division of the three architectures is not important to the users and their boundaries are not so distinct and sometimes even change. For example, most client/server DSSs also provide stand alone desktop versions. The desktop HAD (Isis/base) relies on email (a client/server) and sever computer for data integration and population. The web-based Lionbiosceince technology now offers a "thick" client called Lead Engine to support medicinal chemistry. Ultimately, a good DSS will use the necessary technology to satisfy the decision makers.

Chemical Structure Processing

Processing chemical structure may be the most important task in the DSS. Due to the diversity of chemical structures and non-standard input by human, errors are very difficult to avoid in many early DSSs. Frequently encountered problems include that a SD file cannot be read or the key ID is mismatched. These problems may be small for a computational chemist, but discourage casual users such as most medicinal chemists. Without a quick fix or an alternative plan, a real decision process can be delayed, which is not tolerable in a competitive drug research environment. Thus, a casual user should avoid using a new DSS unless it has been thoroughly tested.

Retrieve and Merge New Data

Ideally, a decision maker should retrieve all needed data with just a few keystrokes or mouse clicks. Most commercial packages, however, are still requiring users to prepare their own chemical structure files from other structure sources except Leadscope, which can make copy of structure database on Leadscope Enterprise server. Although the task is not very difficult for most medicinal chemists, the automation of the task made a DSS more user-friendly.

The more challenging task is to integrate all biological data associated with a set of query structures and present them to the decision makers in a DSS. There seems no technical obstacle to retrieve various data, but the standardization of the data remains difficult. Many historical scientific data are obtained with different assay conditions, stored with different units. Because substantial resources are needed for any implementation, questions such as "Are they compatible?" and "Does a researcher really want to see all of them?" have to be answered first.

An alternative is to use data reduction. Instead of pulling out all biological data associated with a compound, **A**, at the beginning, one only queries whether **A** has been active in any historical biological assays. The information can easily be flagged with **A** and other relevant compounds in a DSS. When a decision maker narrows down a few promising compounds including **A**, s/he can then request for exact data associated with fewer compounds using existing information systems. Building a very complicated DSS can thus be avoided.

Data Reduction

The concept of data reduction can also be used to convert non-searchable and non-sortable data into searchable and sortable data. Many biological data are non-numerical such as ">30" or "0.01, 0.02". Many DSSs simply cannot read in

such data. We had a situation where a set of compounds has to be selected from about 2000 compounds for advanced studies based on rankings of multiple selectivity assay data, which has the above two kinds of data.

A solution we derived is to export the raw data into a SD file and covert the non-numerical data into the acceptable format. A number of rules has been set such as converting ">30" to "60" and averaging multiple data. The application has been developed as a Web friendly protocol as described in SciTegic help allowing for repeated uses. In this approach, the original data retains their integrities while the advanced analysis can be easily performed.

Clustering

Chemical structure clustering is a must-have feature in a good DSS due to its usage in HTS and SAR. The clustering transforms individual compound data into the information about the activities or properties of a series of compounds. The underlying hypothesis is that each member shares a common structural component that is responsible for activities or properties. Most drugs are indeed developed from a series of similar compounds. While the computer-clustering results may not be agreed with everyone, it does provide a way of organizing structures in many cases, which speed up compound selections.

Many DSSs offers MCS-based or scaffold-based clustering because its results are similar to what most medicinal chemists can generate manually but with greater efficiency. Comparing with fingerprint-based methods, Azzaoui [69] shows that MCS-based clustering accumulates more true hits in top ranked classes.

Each program may use different algorithm for MCS clustering with wide range of parameters for final tuning. For similar compounds, the results should not be significantly different. We have tested one set of 33 structures using both Bioreason and HAD (based on Tripos' MCS). The former yields four classes with 1,2,5 and 25 in each. The later yields three classes with 1,2 and 30 each. The last class with 30 structures can further be divided in two subsets, 4 and 26 corresponding to the classes with 5 and 25 ones, respectively. Only one compound has discrepancies between the two methods, which can be accepted by different medicinal chemists.

Other Functions

Sorting and filtering are also needed in a DSS for medicinal chemistry. Sorting is an important way to turn data into information. For example, sorting numerical activities allows users to identify a set of promising compounds quickly. Sorting structures allows users to identify closely related structures. Filtering can be used to remove unwanted structures thereby focus on promising ones. A filter can be applied before data integration or dynamically controlled by users as we see in many commercial packages.

Models

While the overall decision in drug discovery is data-driven, models play an indispensable role in many processes. Incompleteness of data, conflicting data and high cost of experiments have placed models as a component in many decision processes. A well-known example is the Lipinski's rule of five [70], which is used to estimate the oral bioavalaibility. Almost every DSS we reviewed has this model. However, how to apply the model is the decision maker's choice. One may relax the rule for widening the scope of structures. Others want to apply more stringent lead-like rules [71] to increase the chance of success based on statistics.

An ideal drug-like compound should also 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 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 [72,73].

There are increasing demands for integrating these prediction systems into the DSS to support the decisions that ones have to make without sufficient data. We have shown that a knowledge-based protease inhibitor prediction system [74] can be implemented in HAD. The core of the prediction system is sets of 2D structural queries, which are validated by retrieving over 90% of annotated protease inhibitors in MDDR. The query results are displayed as flags indicating whether a structure matches known inhibitors of four classes of enzymes, serine, cysteine, aspartate and metallo protease. The usefulness of the prediction system has been demonstrated in a lead selection process.

Medicinal chemists have used SAR table in lead finding and lead optimization for a long time. The table can be generated easily with commonly used software such as ISIS/base once a core structure and substitution sites are defined. Several new DSSs offer even better solutions by automatically defining MCS as a core and corresponding substitution sites. Users also have options to choose different MCSs in subsets as a core. These functions facilitate the knowledge building process.

Many DSSs provide statistical modeling tools for users to build statistical models or quickly recognize good (predictive) structure-activity relationships. Bayesian learning component offered by SPP probably aim at power users. In contrast, causal users should have no difficulty to understand pharmacophore features predicted by the recursive partitioning of Bioreason's ClassPharma. These models had and will continue to have greater influence on the decisions of what compounds to make by medicinal chemists.

CONCLUSIONS

The complexity of drug discovery requires both human intelligence and better information technologies. How to apply both in every stage of the research becomes the challenge to scientists and technology professionals. The DSS intends to replace tedious and time-consuming human

work in data integration and data transformation. The basic components of a DSS for medicinal chemistry should include data acquisition, chemical structure clustering, searching, filtering and sorting. Models and advanced data mining technologies will appear more in DSSs. The implementation and performance of a particular DSS depends not only on the software, but also on user's computer skill, criteria of decisions, existing information systems, technical support and the cost. The utilization of a suitable DSS should increase the research productivity for medicinal chemists.

ACKNOWLEDGEMENT

The authors thank Dr I. Morize for carefully reading the manuscript and Drs. T. Garyantes, R. Pearlstein and S. Reiling for helpful discussions.

REFERENCES

[1] Power, D. J. *Decision Support Systems: Concepts and resources for Managers*; Quarum Books: Westport, Connecticut, **2002**; 251.
[2] Kendall, K. E. K. a. J. E. *Systems Analysis and Design*; 5th ed.; Prentice Hall: Upper Saddle River, New Jersey, **2002**; 914.
[3] http://www.research.ibm.com/deepblue/home/html/b.html
[4] Kelly, S. *Data Warehousing in Action*; John Wiley & Sons: chichester, **1997**; 320.
[5] Vidette Poe, P. K. a. S. B. *Building a Data Warehouse for Decision Support*; second ed.; Pretice Hall PTR: Upper Saddle River, NJ, **1998**; 285.
[6] Charlwood, B. V.; Morris, G. S.; Grenham, M. J. In *Systematics Association Special Volume Series*, **1984**; pp. 201-208.
[7] Peeters, J.; Thielemans, T.; Van der Eycken, C.; Van Reet, S. In *Janssen Chimica Acta*, **1984**; pp. 9-11.
[8] Wipke, W. T.; Nourse, J. G.; Moock, T. In *Comput. Handl. Generic Chem. Struct., Proc. Conf.*, **1984**; pp 167-178.
[9] Anderson, E.; Veith, G. D.; Weininger, D. In *Report*, **1987**; pp 6.
[10] Martin, Y. C.; Danaher, E. B.; May, C. S.; Weininger, D. In *Journal of Computer-Aided Molecular Design*: Netherlands, **1988**; pp. 15-29.
[11] Lynch, M. F.; Rasmussen, E. M.; Willett, P.; Manson, G. A.; Wilson, G. A. In *Biochemical Society Transactions*, **1989**; pp 856-858.
[12] Chen, L.; Nourse, J. G.; Christie, B. D.; Leland, B. A.; Grier, D. L. In *Journal of Chemical Information and Computer Sciences*, **2002**; pp. 1296-1310.
[13] Pleiss, M. A. In *Tetrahedron Computer Methodology*, **1990**; pp 549-556.
[14] Ruff, P.; Storr, I. In *Laboratory Practice*, **1991**; pp. 55-56.
[15] Henry, D. R.; McHale, P. J.; Christie, B. D.; Hillman, D. In *Tetrahedron Computer Methodology*, **1990**; pp. 531-536.
[16] Elands, J. In *Jala*, **2001**; pp. 42-44.
[17] Bilow, J. In *Bioforum*, **2002**; pp. 746-747.
[18] King, R.; Elands, J. In *Modern Drug Discovery*, **2002**; pp. 21-22, 24.
[19] Chung, S. Y.; Wong, L. In *Trends in Biotechnology*: England: United Kingdom, **1999**; pp. 351-355.
[20] Furness, L. M.; Henrichwark, S.; Egerton, M. In *Pharmacogenomics*: England: United Kingdom, **2000**; pp 281-288.
[21] Warr, W. A. In *Journal of Chemical Information and Computer Sciences*, **1997**; pp. 134-140.
[22] Brown, R. D.; Guner, O. F.; Hahn, M.; Li, H. *Book of Abstracts, 216th ACS National Meeting, Boston, August 23-27*, **1998**; pp CINF-051.
[23] Carleton R.; Sage, K. R. H.; Nianish, S.; Rudy, P. *Molecular Informatics: Confronting Complexity*; The Beilstein-Institut Workshop: Bozen, Italy, **2002**.
[24] http://www.lionbioscience.com/solutions/chemistry/index_eng.html
[25] http://www.tripos.com/sciTech/inSilicoDisc/index.html
[26] http://www-3.ibm.com/solutions/lifesciences/pdf/Aventis.pdf
[27] http://www.hyperion.com
[28] Bradley, M. In *Abstracts of Papers - American Chemical Society*, **2001**; pp. CINF-016.
[29] http://www.tripos.com/sciTech/inSilicoDisc/index.html
[30] Tounge, B. A.; Reynolds, C. H. In *Abstracts, 36th Middle Atlantic Regional Meeting of the American Chemical Society*, Princeton, NJ, United States, June 8-11, **2003**; pp. 18.
[31] Stevenson, J. M.; Mulready, P. D. In *Journal of the American Chemical Society*, **2003**; pp. 1437-1438.
[32] Jorgensen, W. L.; Duffy, E. M. *Adv. Drug Deliv. Rev.,* **2002***, 54*, 355-366.
[33] http://www.mdli.com/
[34] Private communication: Internal evaluation of combi-chem enumeration software.
[35] http://www.softhome.com.tw/000004_SoftHome/soft/001 450_005360_CombiLibMaker1.htm
[36] Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. *Mol. Divers.,* **1996***, 2*, 64-74.
[37] Asher, B. *J. Mol. Graph Model.,* **2000***, 18*, 79-82.
[38] Ladd, W. M.; Lindstrom, M. J. *Biometrics,* **2000***, 56*, 89-97.
[39] Ahlberg, C. *Drug Discov. Today,* **1999***, 4*, 370-376.
[40] Cuissart, B.; Touffet, F.; Cremilleux, B.; Bureau, R.; Rault, S. *J. Chem. Inf. Comput. Sci.,* **2002***, 42*, 1043-1052.
[41] Willett, P. In *IMA Volumes in Mathematics and Its Applications*, **1999**; pp. 11-38.
[42] Wang, T.; Zhou, J. In *J. Chem. Inf. Comput. Sci.*, **1997**; pp. 828-834.
[43] Chen, L.; Robien, W. In *Journal of Chemical Information and Computer Sciences*, **1992**; pp. 501-506.
[44] Tonnelier, C.; Jauffret, P.; Hanser, T.; Kaufmann, G. In *Tetrahedron Computer Methodology*, **1990**; pp. 351-358.
[45] Brint, A. T.; Willett, P. In *Journal of Molecular Graphics*, **1987**; pp. 200-207.
[46] Sheridan, R. P.; Miller, M. D. In *Journal of Chemical Information and Computer Sciences*, **1998**; pp. 915-924.
[47] Cross, K. P.; Myatt, G.; Yang, C.; Fligner, M. A.; Verducci, J. S.; Blower, P. E. Jr. *J. Med. Chem.,* **2003***, 46*, 4770-4775.
[48] Blower, P. E.; Yang, C.; Fligner, M. A.; Verducci, J. S.; Yu, L.; Richman, S.; Weinstein, J. N. *Pharmacogenomics J.,* **2002***, 2*, 259-271.
[49] Yang, C.; Bakshi, B. R.; Rathman, J. F.; Blower, P. E. Jr. *Curr. Opin. Drug Discov. Devel.,* **2002***, 5*, 428-438.
[50] Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. *J Chem. Inf. Comput. Sci.,* **2002***, 42*, 393-404.
[51] Richon, A. *J. Mol. Graph Model.,* **2000***, 18*, 76-79.
[52] Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. Jr. *J. Chem. Inf. Comput. Sci.,* **2000***, 40*, 1302-1314.
[53] Bacha, P. A.; Gruver, H. S.; Den Hartog, B. K.; Tamura, S. Y.; Nutt, R. F. *J. Chem. Inf. Comput. Sci.,* **2002***, 42*, 1104-1111.
[54] Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. *J. Chem. Inf. Comput. Sci.,* **2002***, 42*, 1069-1079.
[55] Tamura, S. Y.; Bacha, P. A.; Gruver, H. S.; Nutt, R. F. *J. Med. Chem.,* **2002***, 45*, 3082-3093.
[56] Miller, D. W. *J. Chem. Inf. Comput. Sci.,* **2001***, 41*, 168-175.
[57] MacCuish, J.; Nicolaou, C.; MacCuish, N. E. *J. Chem. Inf. Comput. Sci.,* **2001***, 41*, 134-146.
[58] van Rhee, A. M.; Stocker, J.; Printzenhoff, D.; Creech, C.; Wagoner, P. K.; Spear, K. L. In *Journal of Combinatorial Chemistry,* **2001**; pp. 267-277.
[59] Chen, X.; Rusinko, A., III; Young, S. S. In *Journal of Chemical Information and Computer Sciences*, **1998**; pp. 1054-1062.
[60] Hawkins, D. M.; Young, S. S.; Rusinko, A. III In *Quantitative Structure-Activity Relationships*, **1997**; pp. 296-302.
[61] Shen, J. In *Abstracts of Papers - American Chemical Society*, **2001**; pp. CINF-079.
[62] Shen, J. *J. Chem. Inf. Comput. Sci.,* **2003***, 43*, 1668-1672.
[63] http://www.neurogen.com/technology.htm
[64] Manly, C. J. Managing laboratory automation: integration and informatics in drug discovery. In *Journal of Automated Methods & Management in Chemistry*, **2000**; pp 169-170.
[65] Manly, C. J. In *Abstracts of Papers, 225th ACS National Meeting, New Orleans, LA, United States, March 23-27, 2003*, **2003**; pp. COMP-344.
[66] http://www.accenture.com/xd/xd.asp?it=enweb&xd=i ndustries%5Chls%5Cpharma%5Ccase%5Cpharma_bristol.xml
[67] http://www.tripos.com/mediaRelations/pressReleases/2003/ 09_24.html
[68] http://www.chemcomp.com/

[69] Azzaoui, K. *Bioreason's 1st European User Meeting*: Strasbourg, **2003**.

[70] Lipinski, C. A. *J. Pharmacol. Toxicol. Methods,* **2000***, 44*, 235-249.

[71] Oprea, T. I.; Gottfries, J.; Sherbukhin, V.; Svensson, P.; Kuhler, T. C. *J. Mol. Graph. Model.,* **2000***, 18*, 512-524, 541.

[72] Greene, N.; Judson, P. N.; Langowski, J. J.; Marchant, C. A. *SAR QSAR Environ. Res.,* **1999***, 10*, 299-314.

[73] Marchant, C. A.; Combes, R. D. In *Bioactive Compound Design*, **1996**; pp. 153-162.

[74] Shen, J.; Hong, J.; Morize, I. In *Abstracts of Papers, 223rd ACS National Meeting, Orlando, FL, United States, April 7-11, 2002*, **2002**; pp. COMP-225.

Copyright of Mini Reviews in Medicinal Chemistry is the property of Bentham Science Publishers Ltd. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.