Discovery informatics: its evolving role in drug discovery

Brian L. Claus and Dennis J. Underwood

Drug discovery and development is a highly complex process requiring the generation of very large amounts of data and information. Currently this is a largely unmet informatics challenge. The current approaches to building information and knowledge from large amounts of data has been addressed in cases where the types of data are largely homogeneous or at the very least well-defined. However, we are on the verge of an exciting new era of drug discovery informatics in which methods and approaches dealing with creating knowledge from information and information from data are undergoing a paradigm shift. The needs of this industry are clear: Large amounts of data are generated using a variety of innovative technologies and the limiting step is accessing, searching and integrating this data. Moreover, the tendency is to move crucial development decisions earlier in the discovery process. It is crucial to address these issues with all of the data at hand, not only from current projects but also from previous attempts at drug development. What is the future of drug discovery informatics? Inevitably, the integration of heterogeneous, distributed data are required. Mining and integration of domain specific information such as chemical and genomic data will continue to develop. Management and searching of textual, graphical and undefined data that are currently difficult, will become an integral part of data searching and an essential component of building information- and knowledge-bases.

Brian L. Claus
***Dennis J. Underwood**
BMS Pharmaceutical Research Institute
PO Box 80500
Wilmington
DE 19880-0500, USA
tel: +1 302 467 5129
fax: +1 302 467 6852
*e-mail: dennis.underwood@bms.com
e-mail: brian.claus@bms.com

▼ There is a revolution taking place in the pharmaceutical industry. An era where almost continuous growth and profitability is taken for granted is coming to an end. Many of the major pharmaceutical companies have been in the news lately with a plethora of concerns ranging from the future of the biotechnology sector as a whole through to concerns over the availability of drugs to economically disenfranchised groups in the developed and the developing world. The issues for the pharmaceutical sector are enormous and will probably result in changes of the healthcare system, including the drug discovery and development enterprises. Central challenges include the impact that drugs coming off patents have had on the recent financial security of the pharmaceutical sector and the need for improved protection of intellectual property rights. Historically, generic competition has slowly eroded a company's market share. There is a constant battle between the pace of discovering new drugs and having old drugs going off patent, providing generic competition opportunities to invade their market share. Recent patent expirations have displayed a much sharper decline in market share making new drugs even more crucial.

The 1990s were a decade where the pharmaceutical giants believed they could sustain growth indefinitely by dramatically increasing the rate of bringing new medicines to market simply by increasing R&D spending and continuing to use the same research philosophies that worked in the past. It is clear from the rapid rise in R&D expenditure and the resultant cost of discovering new drugs that the 'old equation' is becoming less favorable (Fig. 1). There is a clear need to become more efficient in the face of withering pipelines and larger and more complex clinical trials. For large pharmaceutical companies to survive, they must maintain an income stream capable of supporting their current infrastructure as well as funding R&D for the future. The cost of drug development and the low probability of technical success call for improved efficiency of drug discovery and development and further investment in innovative technologies and processes that improve the chances of bringing a compound to market as a drug. Already there has been quite a change in the way in which drugs are discovered. Large pharmaceutical companies are diversifying their drug discovery and development processes: They are relying more on the inventiveness of smaller biotechnology

1359-6446/02/$ – see front matter PII: S1359-6446(02)02433-9

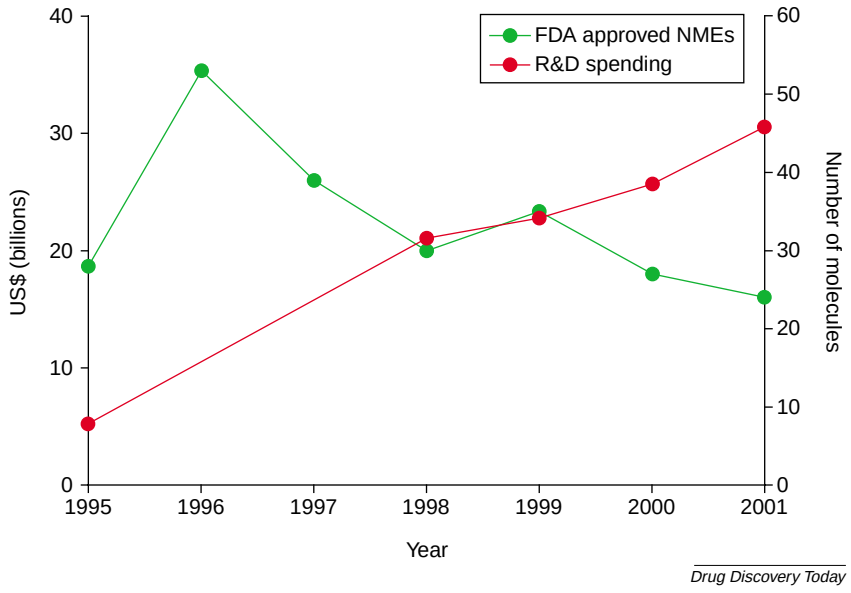

Figure 1. Graph of increasing R&D spending versus new chemical entities (NCEs) launched per year. Interestingly, the number of FDA approved new molecular entities (NMEs) have diminished even though the R&D budget across the industry has increased. There could be many reasons for this including increased scrutiny by the FDA and the typical cyclical nature of the industry, in which it can take 10 years from discovery to a drug.

companies and they are licensing technology, compounds and biologicals at a faster, more competitive rate. To meet crucial timelines they are outsourcing components of R&D to contract research organizations (CROs) enabling greater efficiencies either by providing added expertise or resources and decreasing development timelines. The trend towards mergers and acquisitions, consolidating pipelines and attempting to achieve economies of scale, is an attempt by large pharmaceutical companies to build competitive organizations. Although this might help short-term security, future ongoing success may not be ensured solely with this strategy.

One of the most valuable assets of a pharmaceutical company is its experience in drug discovery and development. Of particular importance is the data, information and knowledge generated in medicinal chemistry, pharmacology and *in vivo* studies accumulated over years of research in many therapeutic areas. This knowledge is based on hundreds of person-years of R&D and yet most companies are unable to effectively capture, store and search this experience. This intellectual property is enormously valuable. As with the other aspects of drug discovery and development, the methods and approaches used in data-, information- and knowledge-base generation and searching are undergoing evolutionary improvements and, at times, revolutionary changes. It is imperative for all data- and information-driven organizations to take full advantage of the information they are generating. We assert that those companies that are able to effectively do this will be able to gain and sustain an advantage in a highly complex, highly technical and highly competitive domain. The aim of this overview is to highlight the important role informatics plays in pharmaceutical research, the approaches that are currently being pursued and their limitations, and the challenges that remain in reaping the benefit of advances.

There has been much time, money, and effort spent attempting to reduce the time it takes to find and optimize new chemical entities (NCEs). It has proven difficult to reduce the time it takes to develop a drug but the application of new technologies holds hope for dramatic improvements. The promise of informatics is to reduce development times by becoming more efficient in managing the large amounts of data generated during a long drug discovery program. Further, with managed access to all of the data, information and experience, discoveries are more likely and the expectation is that the probability of technical success will increase (Fig. 2).

Informatics paradigm shift

There is an important paradigm shift that has begun to take place not only in the pharmaceutical industry but also in other domains that either generate or is required to use data and information as part of their business processes. This shift is a natural outcome of our reliance on computers for input, storage, access to data and information and communication. It is becoming easier to store data but harder to access the stored data in a flexible, complete, yet manageable way. To compound these complexities, decision-making requires that all relevant data are accessible and available even though the data-types could be dramatically different. One only has to consider a trivial example of trying to find a photograph in an electronic collection that only has the date and location recorded as annotations; how do you search for a particular photograph of a sailboat out of a collection of photographs of coastlines, beaches and ships? Images and textual data are different data-types. The relevance of this example to the pharmaceutical industry is apparent. The kind of data generated in the course of developing a drug includes diverse data-types

including chemical, genetic, biochemical, pharmacological, histological, physiological and textual data (Table 1). Many organizations recognize a highly integrated data environment as key to using the data effectively; however, few have experienced the benefit of having brought this data environment to fruition. But as described above, the challenge is to build an environment that integrates different data-types, is manageable, is flexible and has a significant lifetime.

Many data environments and browsing capabilities were designed with the aim of archiving rather than providing meaning and context to the data and information and enabling insight into future studies and experiments. Systems designed in this manner are severely limiting to future development. In a real sense, an archival system is designed around a simple question and answer paradigm whereby the data returned is an exact, literal response to the question asked. As an example, consider the challenges of trying to use a data access system that is not tolerant of spelling inconsistencies and that does not have the ability to substitute synonyms for keywords in the search phrase or that only allows keywords joined through Boolean constructs. Efficient and thorough searching of archival systems requires a detailed understanding of the data structure of the repository before being able to phrase an appropriate question. Such data environments do not facilitate insight and are uninspiring. A different, more accessible, paradigm is to provide a natural language interface that is interpretative and can recognize syntax and context. Such systems have the ability to provide contextual assistance in asking the question. As a further advance, the data system could return answers within a contextual framework and facilitate exploration of the answer through natural linkage hierarchies and further refinement of the question or, importantly, further related questions. However, this kind of data and query environment, although technically feasible, is not generally available to researchers in the pharmaceutical industry. Data environments and browsing systems remain rigid with confining interfaces, data sources unreachable and data-types indecipherable. If systems such as these were implemented within the pharmaceutical industry, this would amount to a paradigm shift in the manner and efficiency of data and information management.

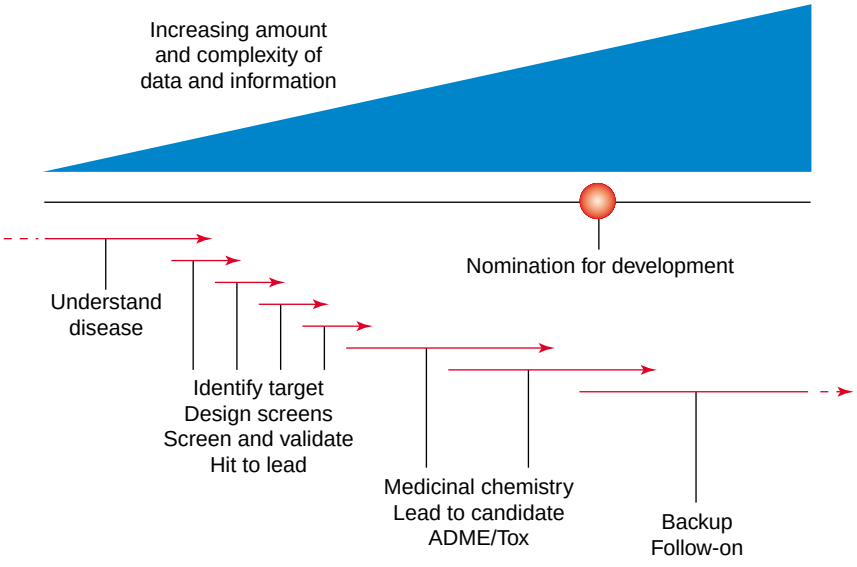

Figure 2. The amount and complexity of data and information increases dramatically during the discovery phase of drug development. The need for integrating the data and information increases as the drug discovery project progresses and is especially important when compounds transition from discovery to development (nomination for development) and in later backup and follow-on programs. Backups are compounds that are being moved into development behind previous candidates that have different ADME/Tox properties. Follow-ons occur later in the process and generally have a different chemical structure or chemotype.

Definitions

For the purposes of this review, we define data, information and knowledge in the following way: Data servers (or sources) are 'low-level' physical repositories of numbers, characters, images, and other such representations corresponding to variables, observables, and their transformations (Fig. 3). (For the purposes of this overview, data and observables are equivalent and are used interchangeably.) We consider documents and text as data. In general, data sources can contain different data-types (e.g. numerical, textual and graphical). The manner in which the observables in the data servers are organized, represented, stored and accessed can be relational, hierarchical, object-based, sequential, indexed, or organized in any other way, provided that the data model(s) and relationships are exposed and interpreted by external processes. Deriving relationships between data produces expressions that encapsulate information. Information servers are the encapsulation of methods, algorithms, and systems by computing systems that produce information from data. This can be done by refining, transforming, analyzing, and integrating

Table 1. Examples of data-types used within the pharmaceutical industry[a].

	Classes	Data types
Chemistry	Structural and physical descriptors	Chemical structure, compound name, synonyms and identifiers, partition coefficients (e.g. logP), molecular weight (MW), topological and topographical descriptors (e.g. atom-pairs, pharmacophore fingerprints), conformations and indices, synthetic route, purity, spectral properties, inventory location and quantity.
Biology	Assay methods, enzymology, biochemistry, pharmacology	Inhibition constants (e.g. K_i, IC_{50}, % change at concentration), Hill slope, titration curve, assay descriptors (e.g. assay protocol, buffers), reagent descriptors
	Mechanism	Enzyme nomenclature, cofactors, cellular compartmentalization, post-translational modification, proteomics, genomics
Structural biology	Methods	Experimental method (e.g. X-ray, NMR), crystallization conditions (e.g. protein purity, buffers, additives), purification method, expression system (e.g. vector, insect, bacterial), crystal form, space group, refinement parameters (e.g. resolution, R-factor)
	Structural descriptors	Domain structure, structural characteristics (e.g. family, RMS differences)
ADME/Tox	Assay methods, pharmacokinetics and dynamics	Free fraction, protein binding, renal clearance, P450 inhibition, cytotoxicity, assay descriptors (e.g. assay protocol, buffers, species)
	Toxicity	Physiology, pathology, histology, species comparison, mechanistic basis
Genomics, proteomics, transcriptomics		Nucleotide and protein sequence data, gene location, gene structure, SNPs, expression profiles, tissue distribution, functional annotation
Text	Scientific literature, patents, reports, memos	Text, tables, graphs, photographs, bibliographies, appendices, clinical data

[a]This table highlights some of the different data-types that are essential to the drug discovery and development process and is not an attempt to be comprehensive.

data (i.e. data processing), and then identifying patterns and relationships in the data using study, experience, or instruction as computer methods and processes. Metadata comprise a component of information because they provide contextual meaning to the data elements. That is, metadata are data about data and, therefore, are explicitly managed data. Metadata can include descriptive information about the context, quality and condition, or characteristics of the data. Knowledge differs from data or information in that new knowledge can be derived from existing knowledge using logical inference or rules. If information is data plus meaning, then knowledge is information plus processing. In this sense, data sources can include databases, information-bases, and knowledge-bases.

Advances in data systems: system architecture, metadata and data access

Drug discovery and development is a highly competitive, rapidly changing area that increasingly depends on information technology to store, exchange, and mine data and information. The current application of HTS, directed parallel synthesis and combinatorial chemistry and the generation of ADME/Tox data on an ever increasing scale has led to a vast increase in the amount of available data for drug discovery and development. Management and access to data and the generation of information and knowledge from this data are rate-determining at many steps of the drug discovery process. Data heterogeneity (i.e. different data-types) also serves to make the process of using the data more complex. The optimization of compounds can be considered as a process of collecting and understanding the results of experiments and measurements from different areas such as medicinal and physical chemistry, biology and ADME/Tox from which new compounds are made and new experiments performed. The complexities of this process place a demand on the capabilities of informatics systems to store, access and integrate data in a high performance manner. Within this challenge, there are many opportunities for improvement including data collection, data integration and methods for accessing, refining and mining the data.

It is essential to capture all aspects of an experiment and typically this is done through notebooks. Once in a notebook, however, it is difficult to retrieve the data and the experimental conditions. The ideal, therefore, is to facilitate the capture of the information and data electronically and to provide ways of searching this data. There are efforts to develop electronic notebooks but the regulatory agencies have been slow to accept these as validated and verified data and information sources. However, once systems and sources have been shown to be 'trustworthy' their acceptance is likely to follow. An apparent limitation has always been computational power and storage capacity (disk). Historically, computer processor and storage limitations have compromised the kind and amount of data and information that can be stored.

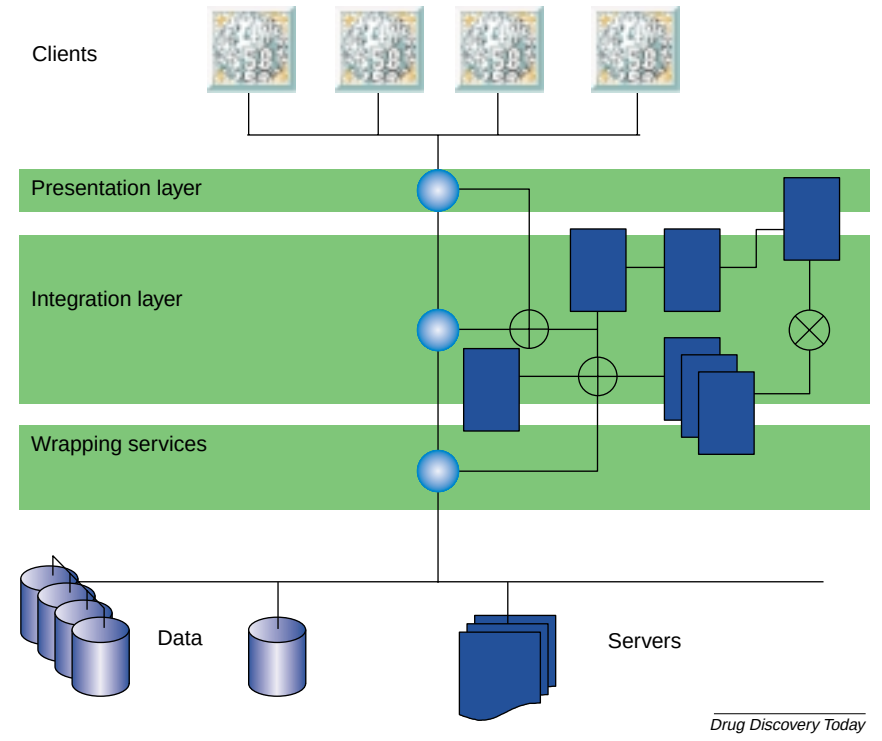

Figure 3. Heterogeneous and distributed environment. The clients at the top of the hierarchy access the data and information via a metadata rich 'integration layer'. Access to the underlying data structures and computer and application servers is mediated through the 'wrapping services', which maintains the information necessary to bind and direct the queries to the correct source.

As a simple example, consider a company having a compound collection of 50,000 compounds tested in 20 assays. To store the compound identifiers, assay identifiers and single value assay results as relative activities would require on the order of 10 megabytes of storage. This is not a lot of storage to achieve in today's standards, but in the past it has been a severe limitation. We are now in a position to store and access much more of the information necessary to fully capture an experiment and its results. However, the list of data are surprisingly long and varied (Table 1). To fully annotate an experiment and its results requires a chemical structure, the chemist who synthesized it, the date of synthesis and testing and the assay results, where the physical sample is stored, and a plethora of other vital information necessary for continued use of the compound and its data. Instead of representing activity with a single relative activity value we are able to store the following: derived data (e.g. IC_{50}, K_i, minimum inhibitory concentration), the normalized raw data (e.g. percent change at a concentration of compound), the controls, references and standards, the plate map, the raw data (e.g. counts), the experimental protocol(s) and annotations (assayist, date, and so on) and the relationships between different experiments (Table 1). All of this can be saved for many hundreds of thousands of compounds tested in replicate across many different assays. The storage requirement for this is multiple gigabytes of information. Consider adding to this other crucial pieces of information required to optimize a drug such as ADME/Tox, patent estate and literature, memos and reports, and storage estimates quickly skyrocket into terabytes and access to all of this data becomes difficult.

Adding to the challenge of storage is the issue of data heterogeneity. A well designed and well defined data environment is crucial to ensure future access from rapidly evolving search engines and browsers. In addition, well constructed data sources ensure that these data are available to data reduction and data mining, the purpose of which is to extract relationships and information from many sources. It is also likely that novel methods will have different data requirements than existing methods. The ability to store more complete sets of data increases the likelihood that the relevant data will be available when new methods are implemented. As datasets grow and become more complete, the methods used to analyze them need to shift from methods that only deal with portions of the data to methods to analyze the entire collection of data at one time. This provides context for the experiment and

the resultant data. It also becomes possible to look for anomalies in the data and identify tests that need to be repeated or to identify results that are yielding particularly interesting information.

There are essentially two primary data architectures commonly used for addressing the needs of querying and integrating heterogeneous and distributed data generated in drug discovery and development. The first approach solves the data integration and heterogeneity problem by centralizing data in a single repository, a warehouse. The second approach uses a design that acknowledges a heterogeneous and distributed data environment and integration of data occurs through methods and 'wrappers' that understand the data structure of each individual database and provide methods for joining or integrating the data for each query. This approach is loosely termed database 'federation'. As always, there are variations on each of these approaches that gray the distinction between them, such as using multiple data-marts (smaller databases similar to a data warehouse but optimized for a single data type).

Data storage and heterogeneity are both informatics challenges. However, the most crucial need is to be able to query and integrate these data and information and to be able to access ancillary information which provides context to the data. Without context, the manner in which the data were generated, the details of the methods used, the reagents used, and other observables from the experiment, the data would have no true meaning or value. These requirements are clearly a test for any discovery informatics team.

Centralized versus distributed

In a warehouse approach the data are integrated in a single repository which provides a single source for user searches and data-mining. The main purpose of this design is to provide a single, integrated data structure that is well understood and can be accessed in a controlled manner. This presumably has advantages in terms of data management within a large, multifaceted organization. Further, this approach deals with data heterogeneity and data validation at the point of loading the warehouse and performs the data integration at the storage level. Frequently, this is accomplished using ETL (extract/transform/load) tools or scripts to migrate data into the warehouse. The limitations of this approach are the lack of linkage between the warehouse and the underlying data sources (disconnected context); generally only a subset of the data are uploaded into the warehouse for performance reasons. Another severe limitation is the inability of this model to deal with data heterogeneity. Typically, different data-types have different storage requirements that are usually addressed through specialized repositories that are designed to optimize management and access. These limitations argue strongly for a distributed data architecture that keeps the data repository close to the source of the data and recognizes that data are heterogeneous.

In a database federation approach, the data architecture and the information environment remain connected (linked) and the contextual meaning of the data is maintained. This approach necessitates the implementation of a middle-ware layer that encapsulates a metadata/information environment composed of methods and interpreters that understand data relationships, and provides a means of 'wrapping' the different, 'low-level' data sources. The approach covers the processes of formulating a query across complex, heterogeneous data-types, parsing components of the query and formulating sub-queries, optimization and running of the sub-queries across the various data and information services. Integration of the results of the queries and caching of the query and results at the intermediate data and information manager occur at the middleware layer. This approach presents the data and information retrieved by interacting directly with the data and information manager. Data hierarchy is maintained throughout this process and enables retrieval without re-formulation of the query. It is important to realize that the federated architecture can integrate warehouses as components depending on the requirements (Fig. 3).

This architecture is congruent with a distributed data environment in which data are generated and stored in highly specialized databases and data storage areas, including file systems, and is optimized for writing and reading the type of data that is generated. Thus data are not restricted to conform to a single set of standards as in a warehouse but can be stored in the most efficient manner for each type of data. Standards are imposed at the level of communication and defining how interactions take place without restricting content. This is analogous to the design of the Internet which is a collection of public networks adhering to standards such as TCP/IP (Transfer control protocol/Internet protocol) and URLs (Uniform Resource Locators). In drug discovery applications, the types of data generated and stored include those listed in Table 1.

The view into the data and information services provided by these systems and methods enables the user to sort, cluster and associate data that have fundamentally different data-types and, therefore, provides a means for developing hypotheses. It is these complex associations between data that enable the distillation of information and hypotheses. Encapsulation of these processes into methods, algorithms and scripts provides knowledge-bases that can run independently, continuously distilling information and knowledge from the current and new data.

It is crucial that the user interface to any data environment is easily configurable to meet the needs of a diverse user community. It is unlikely that a single interface will meet every need, situation or task. The true goal should be a metadata driven architecture that can reorganize the presentation of information depending on the questions asked. This model is heavily dependent on the metadata. We have seen hints of approaches that tailor views that are necessary to deal with the information explosion in World Wide Web delivery platforms such as portals and configurable content pages. This is also seen in stand-alone tools such as Themescape (now Micropatent Europe; http://www.micropat.com/). However, this is not generally a feature common in systems used by the pharmaceutical industry.

There have been some significant gains in introducing advanced informatics systems into the pharmaceutical industry but a greater emphasis on this strategy is needed. Early work at Merck (http://www.merck.com/) in collaboration with IBM (http://www.ibm.com) showed that a heterogeneous, distributed data and information environment provided access to data necessary for the efficient operation of drug discovery programs [1,2]. Key to this prototype was the integration of cheminformatics data within the environment. The outcome of this work is DiscoveryLink (formerly the known as the 'Garlic' project), which provides the basis of IBM's technology for managing and exploring data within many different domains including the pharmaceutical industry. A similar approach was developed within DuPont Pharmaceuticals (now BMS; http://www.bms.com/) specifically to provide access to data, information and data-mining tools for drug discovery. In this case, all discovery data, including HTS data, was successfully managed and accessed. These efforts used expert internal resources that were fully conversant in the needs and requirements of their respective drug discovery organizations. There have been significant efforts by third party developers, at times in collaboration with industry R&D, to develop marketable discovery informatics platforms; for example, Tripos (http://www.tripos.com), Accelrys (http://www.accelrys.com) and Lion BioSciences (http://www.lionbioscience.com). Although each provides useful functionality, none has matured to the point of providing a complete solution to the informatics challenge. The detachment of a software development company from the pharmaceutical company client inherently results in a limited understanding of the processes involved and the changing needs in drug discovery and development.

Advances in data mining, information discovery and knowledge-base development

One of the factors influencing the change in the way a scientist deals with data is the explosion of computing power in the past several decades. Computing power has historically followed Moore's Law [3], which predicts that computational power will double every 12–24 months. This trend was expected to end in the late 1990s. Fortunately, not only has this trend continued, but it has kept pace with the explosion in the rate of generation of chemical, biological and other data. Discovery informatics is a niche that attempts to leverage this information explosion and translate it into marketable products for drug companies. It is necessary for computational methods to deal with larger, more complex sets of data than ever before. More computing power facilitates all aspects of the data and information management process.

In general there are four approaches that are important to discovering information and knowledge from data: data searching and integration, data transformation, data reduction and data-mining. Data searching and integration have already been discussed here. Data transformation refers to any alteration to the data either to facilitate browsing or to enable data mining. Browsing oriented transformations include operations such as computing averages and standard deviations and dynamically converting results to different units to facilitate comparison between data points. Transformations to enable data mining refer to format conversions, calculations of fingerprints for chemical structures, descriptor assignment and other computational steps to prepare data for interpretation. Data reduction methods are crucial for dealing with rapidly expanding datasets. Clustering will be discussed as an example of a data reduction method. Data mining methods are then considered as they apply to various categories of problems: classification, pattern discovery, search and optimization and data extraction.

Data reduction methods were originally designed to lessen the size of the computational and data management problem by selecting representative data points and reducing the number of calculations that need be performed. This is a useful approach and is used intensively but it suffers from both a representation and a sampling problem. Data reduction methods are a guide through large datasets. Reducing the amount of data, provides a manageable dataset that, hopefully, contains sufficient 'truth' to allow further investigation of the data and decisions to made. Clustering as a data reduction method was originally touted as being able to facilitate computation by selecting representative members of similar compounds and only performing calculations on the selection. This allowed a larger set of compounds to be evaluated with a small number of calculations. The concept behind clustering is to organize a large set of data into smaller groups based on a similarity metric. The assumption, which is well validated by experience,

is that items ending up grouped together share some common properties. Because the items are organized via a similarity metric, the descriptors used to compute that similarity have a tremendous impact on the results. It must be said, however, that this selection process is highly dependent on the manner in which the items are described (the descriptors) and the way in which they are clustered.

There are two main categories of clustering methods, hierarchical and non-hierarchical [4]. Within each of those classifications are multiple methods such as 'k-means' and 'Jarvis-Patrick', as well as different metrics for computing similarity between objects, such as the Tanimoto coefficient, Dice metric, or geometric distance [5]. Frequently, different descriptors are used depending on the application. Topological descriptors include the following: Atom Pairs, Topological Torsions and '2.5D' descriptors [6], Daylight Fingerprints (http://www.daylight.com), MDL Keys (http://www.mdl.com). Other descriptors include 3D pharmacophores [7–10], physical property descriptors including Lipinski Filters [11], Molecular Weight, Polar Surface Area, clogP and rotatable-bonds [12,13]. Principal component analysis attempts to reduce the number of descriptors or variables in a system by determining any relationships between the descriptors. Analysis proceeds with a reduced set of descriptors.

However, clustering alone is unlikely to be the panacea hoped for. Minor modifications to molecules can radically change their biological activity. This does not render clustering worthless. It simply requires a shift in expectation. Instead of considering clustering to be the final calculation, it needs to be used as an initial sampling that enables testing or examination followed by further refinement of the selection. The aim is to build a good representative selection of the data that can characterize the diversity, depth and range of the data. Representation of diversity or some other means of coverage or completeness is a challenge that is ultimately subjective; it is difficult for methods and algorithms to replace the analytical abilities of an experienced scientist. The methods need to guide rather than replace judgement and need to facilitate rather than inhibit. These, in essence, are the real challenges for discovery informatics.

A commonly overlooked facet in computational methods for drug discovery is the data quality and the uncertainty associated with each data point. Methods that take data uncertainty into consideration will provide better insights than methods that consider all data as equal and exact; the methods are only as good as the data they use.

Classification problems are well suited to decision tree methods. These methods use a training set of data to formulate a series of rules that can be applied to classify the training elements. Once derived, these rules can be applied to unknown elements to predict a classification or outcome. Decision trees have been applied to a large variety of problems including virtual screening and toxicity and property prediction. These methods typically breakdown when the unknown elements fall outside the parameters represented by the training set.

Pattern discovery and pattern recognition are methods that are gaining popularity with the increasing amounts of biological sequence data available. Pattern discovery is a class of methods that, in a deterministic or a probabilistic manner, determine the syntactical meaning within groups of associated objects. In other words the approach ascribes patterns or collections of descriptors, such as amino acids in protein sequences, to features such as protein class membership or other biological observables. Biological sequence data are very amenable to the concepts of pattern discovery and recognition [14]. With large numbers of sequences being available and categorized, it becomes possible to use pattern discovery to categorize orphan sequences and provide insight into function [15]. The use of pattern discovery with respect to drug discovery is not limited to sequence data; it also has uses for searching and finding relationships and correlations among other data types such as pharmacophores or chemical structure fingerprints. There have been a few attempts in the area of pattern recognition within chemical space and in combining with other data-types in an integrated space [16,17]. However, this an exciting area of development that promises new ways of building information from data, data-mining and knowledge discovery.

Drug discovery and development is in itself a search and optimization problem. Methods developed to address some of the needs in this area include pharmacophore mapping [18,19], docking simulations [20–22], linear and non-linear QSAR (quantitative structure–activity relationship) analysis, protein folding prediction, and combinatorial library design [23]. Interesting approaches in the application of optimization methods to the pharmaceutical industry include recursive partitioning [24] and model-building using genetic algorithms [25]. Generally there are four optimization and search approaches: regression methods that fit the data to model relationships, deterministic algorithms try all possible solutions, stochastic algorithms generate random solutions and select the best, and evolutionary algorithms evolve a potential solution set into a final solution. Each of these approaches have benefits and drawbacks. In general, regression methods, which include neural nets, recursive partitioning and other linear and non-linear methods, have a hard time fitting data that goes beyond the training set; that is, it is hard to generalize these

methods. Deterministic algorithms are guaranteed to find the best possible answer because they evaluate all solutions. Unfortunately, for difficult problems, deterministic solutions can be too time consuming to compute. Stochastic methods can be faster than deterministic algorithms and are useful as long as the initial assumption that the difference between the best stochastic solution and the best deterministic solution is small enough to be tolerable. Finally evolutionary methods, such as genetic algorithms, are an approach to solving some of the difficult optimization problems in the industry. These methods use concepts parallel to those in genetic evolution such as the generation of a population of potential solutions, crossovers and mutations during reproduction to yield new populations, and fitness functions to assess the suitability of the solution to the problem at hand. The hope with genetic algorithms is a performance increase over deterministic methods with an increase in solution quality over stochastic methods. Of course, the solutions found by genetic algorithms, as with any of the approaches outlined, are heavily dependent on the fitness functions used to assess the suitability of the populations and the descriptors used. A poorly defined fitness function or descriptor set applied to a poorly defined dataset will result in a meaningless result.

Textual data are a largely untapped source of information in the drug discovery arena. Latent Semantic Indexing [26,27] and Bayesian statistics [28,29] have been used to analyze, manipulate, and process documents and textual data to extract conceptual meaning and assess relevance to a user's query. Unfortunately, many companies are not capturing documents electronically and making them available for searching, nor are they extracting as much information as possible from the documents they are storing. To gain the fullest benefit from these documents, methods need to combine image analysis with the textual context to extract chemical structures, reactions, charts, graphs and tables.

It is not sufficient to make a single method or approach available. All of these approaches have many permutations. Each of these algorithms and metrics have subtle nuances that must be understood. This is why it remains necessary to have expert users who understand the intricacies of the methods to prevent misinterpretation of the results. It is appealing to release these methods generally within the organization, and this can be done in many cases. However, for new, cutting-edge methods, this is often difficult. Collaboration and teamwork remains the mainstay of drug discovery and development; it is difficult to become expert in all of the new technologies available.

Conclusions and future challenges

Discovery informatics and data-mining methods will be a crucial component to addressing the changing future of the pharmaceutical industry. Methods will need to handle larger and more diverse datasets. It is no longer sufficient to perform SAR (structure–activity relationship) studies using data from a single assay or program; it needs to be computed across all programs and assays with all compounds. It is no longer enough to restrict the data to activity and binding assay data; the methods need to incorporate other data sources including ADME/Tox and other *in vivo* results. Text data as an adjunct to the integration of experimental data are no longer appropriate, it needs to be considered in the first tier along with traditional data-types. Non-structured data, such as text and images have always posed difficulty for processing and searching and remains a challenge for integration and use, but also represents great potential for future advances. Pattern recognition might be able to make advances in areas other than biological sequence data. It might be applied to problems including toxiphore identification, bioisosteres, and molecular fingerprinting, as well as data quality assurance metrics such as identifying edge effects for plated assays, robot errors and errant reactions. It is important to realize that none of these problems can be solved without the data, especially without the context to the data. Analysis methods are only half the problem; data collection, querying and presentation remain challenges for the industry as well.

References

1 L. M. Haas, *et al.* (2000) Prototype of integration of chemical and biological data in collaboration between Merck and IBM 1994; DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Systems Journal* 40 489–511 (available online at: http://www.research.ibm.com/journal/sj/402/haas.html)

2 Ricadela, A. (2001) Reinventing research. *Information Week*. March 1 (available online at: http://www.informationweek.com/828/research.htm)

3 Moore, G.E. (1965) Cramming more components onto integrated circuits. *Electronics* 38, 114–117

4 Barnard, J.M. and Downs, G.M. (1992) Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci.* 23, 644–649

5 Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy*, W.H. Freeman, San Francisco

6 Sheridan, R.P. *et al.* (1996) Chemical similarity using geometric atop pair descriptors. *J. Chem. Inf. Comput. Sci.* 36, 128–136

7 Spellmeyer, D.C. and Grootenhuis, P.D.J. (1999) Molecular diversity. *Annu. Rep. Med. Chem.* 34, 287–296

8 Blaney, J.M. and Martin, E.J. (1997) Computational approaches for combinatorial library design and molecular diversity analysis. *Curr. Opin. Chem. Biol.* 1, 54

9 Brown, R.D. (1997) Descriptors for diversity analysis. *Perspect. Drug Discov. Des.* 7–8, 31–49

10 Spellmeyer, D.C. *et al.* (1997)*Practical Application of Computer-Aided Design* (Charifson, P.S. ed.), pp 165, Marcel-Dekker

11 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 4–25

12 Grover, I.I. *et al.* (2000) Quantitative structure–property relationships in pharmaceutical research. (Part 1*). Pharm. Sci. Technol. Today* 3, 28–35

13 Grover, I.I. *et al.* (2000) Quantitative structure–property relationships in pharmaceutical research. (Part 2). *Pharm. Sci. Technol. Today* 3, 50–57
14 Rigoutsos, I. and Aris, F. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* 14, 55–67
15 Argentar, D.R. *et al.* (2000) Tuples, tables and trees: a new approach to the discovery of patterns in biological sequences. *Eighth International Conference on Intelligent Systems for Molecular Biology*, San Diego, CA, USA, 19–23 August 2000
16 Bradley, E.K. *et al.* (2000) A rapid computational method for lead evolution: description and application to α1-adrenergic antagonists. *J. Med. Chem.* 43, 2770–2774
17 Stanton, R.V. *et al.* (2000) Combinatorial library design: maximizing model-fitting compounds within matrix synthesis constraints. *J. Chem. Inf. Comput. Sci.* 40, 701–705
18 Miller, M.D. *et al.* (1999) SQ: a program for rapidly producing pharmacophorically relevant molecular superpositions. *J. Med. Chem.* 42, 1505–1514
19 Lemmen, C. and Lengauer, T. (2000) Computational methods for the structural alignment of molecules. *J. Comput. Aided Mol. Des.* 14, 215–232
20 Kearsley, S.K. *et al.* (1994) Flexibases: a way to enhance the use of molecular docking methods. *J. Comput. Aided Mol. Des.* 8, 565–582
21 Miller, M.D. *et al.* (1994) FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J. Comput. Aided Mol. Des.* 8, 153–174
22 Miller, M.D. *et al.* (1994) Advances in automated docking applied to human immunodeficiency virus type 1 protease. *Methods Enzymol.* 241, 354–370
23 Sheridan, R.P. *et al.* (2000) Designing targeted libraries with genetic algorithms. *J. Mol. Graph. Model.* 18, 320–334
24 Rusinko, A. III, *et al.* (2002) Optimization of focused chemical libraries using recursive partitioning. *Comb. Chem. High Throughput Screen.* 5, 125–133
25 Vaidyanathan, A.G. (2000) Distributed hierarchical evolution modeling and visualization of empirical data. WO067200
26 Hull, R.D. *et al.* (2001) Latent semantic structure indexing (LaSSI) for defining chemical similarity. *J. Med. Chem.* 44, 1177–1184
27 Hull, R.D. *et al.* (2001) Chemical similarity searches using latent semantic structural indexing (LaSSI) and comparison to TOPOSIM. *J. Med. Chem.* 44, 1185–1191
28 Vaithyanathan, S. and Dom, B. (1999) Model selection in unsupervised learning with applications to document clustering. *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)* (27–30 June 1999, Bled, Slovenia) (I. Brakto and S. Dzeroski, eds) Morgan Kaufman (available online at: http://www.almaden.ibm.com/cs/people/dom/external_publist.html)
29 Vaithyanathan, S. *et al.* (2000) Hierarchical Bayes for text classification. *PRICAI'2000 International Workshop on Text and Web Mining* (September 2000, Melbourne, Australia) pp.36–43 (available online at: http://www.almaden.ibm.com/cs/people/dom/external_publist.html)