Abstract

Pharmaceutical companies spend billions of dollars annually on drug discovery research. In the process, they generate vast amounts of scientific data. Data warehousing could significantly shorten the research cycle that leads to a new drug. We propose a framework for the application of data warehousing to integrate a pharmaceutical company’s drug discovery data. We provide an analysis of the principal activities involved in drug discovery in the pharmaceutical industry, and a set of questions from which potential queries can be derived. This information is used to identify fact and dimension tables, and a dimensional data model for a drug discovery data warehouse is proposed.

1: Introduction

The goal of basic research in a pharmaceutical company is to identify chemical compounds that have efficacy as drugs against disease. To achieve this goal, pharmaceutical companies employ chemists who synthesize chemical compounds, and biologists who test them against various disease processes. Compounds that test positive are called leads. A lead is not a drug, however, and it can take years to get it through the research and development pipeline. The process of creating and testing thousands of compounds per year generates large amounts of scientific data. Within this ever-growing body of data resides considerable predictive potential, if it can be exploited effectively. The role of computational chemists, or, modelers, is to realize the predictive potential in a company’s scientific data. At present, modelers use operational databases, or build specialized, stand-alone databases for their work. Operational databases are built for transaction-processing, rather than analytical processing. They are volatile, poorly integrated and data in them is not standardized. Building stand-alone databases for each modeling application is labor-intensive and expensive, and does not integrate all of the data that could be used for multidimensional analysis.

A data warehouse is a read-only analytical database that is used as the foundation of a decision-support system. A data warehouse provides a subject-oriented, integrated, time-invariant, non-volatile store of business data for analytical processing[1]. The dimensional data model for data warehouses was introduced by Ralph Kimball[2]. It consists of two types of tables: fact tables, which are very large and contain the factual data about an enterprise, and smaller dimension tables which contain descriptive data that reflect the dimensions of the business. The dimensional model is often called a star-join schema. This data model enhances query performance by providing short, well-defined join-paths and allowing analysis of data by multiple dimensions.

The purpose of this study is to derive a model for a data warehouse for pharmaceutical drug discovery research, to aid in the discovery of active compounds. The rest of this article is organized as follows: Section 2 presents a description of the major activities involved in drug discovery, and the data they generate. Section 3 approaches warehouse requirements by looking at the types of questions that queries will be based on. Section 4 identifies important entities that will become tables in the data warehouse. Section 5 presents a star-join schema for the enterprise. Section 6 discusses the quality of our model and the appropriateness of data warehousing for this problem. Section 7 concludes our paper.

2: Enterprise description
2.1: Chemistry and chemical compounds

Pharmaceutical companies have departments devoted to the synthesis of chemical compounds. Compounds historically have been synthesized singly, however, technology has recently been developed to synthesize large numbers of compounds with related structures, all at once, in combinatorial chemical libraries[3]. Another source of compounds is isolation from natural substances. Finally, pharmaceutical companies can acquire chemical collections by purchasing them from other labs.

The simplest representation of a chemical compound is its molecular formula, which shows its composition. Structural formulas show connectivity. Many chemical compounds have a ‘handedness’ associated with them; this is represented by stereochemical notations. The issue of stereochemistry is important, since many pharmaceutically active compounds contain one or more stereo-centers. In order to compare compounds, a representation in three dimensions is required. Often, a compound can adopt different conformations. It is possible for a compound that is active against one disease process to be active against another, in a different conformation. It is therefore desirable for the data warehouse to store multiple conformers of the same compound.

Chemical compounds can be grouped into families by shared structural characteristics. These families can be based on overall structure or local structural features, called functional groups or substituents. Certain patterns, called pharmacophores are recurring themes in pharmaceutically active compounds. A compound can belong to more than one structural class. Other characteristics of compounds that play an important role in pharmaceutical activity are electrical charge, acidity, basicity, and hydrophobicity.

Combinatorial libraries fall into structural classes and substituent classes. When a library is made, the identity of each compound in it is unknown. If the library tests positive for activity against a disease process, the identity of the compound responsible is deduced using the scheme that was used to make the library, or by physical methods.

In order to design an effective data warehouse for pharmaceutical research, representations, attributes, and dimensions of chemical compounds must be chosen that: a) uniquely represent a compound’s composition, connectivity, 3D structure, and stereochemistry.

b) represent all significant conformers of the compound.

c) allow queries that group compounds into classes that share pharmacophores or other structural characteristics.

d) represent physical characteristics such as charge, acidity, basicity, and hydrophobicity.

2.2: Enzymology and assays

Biological processes are often carried out by enzymes[4], which are proteins that are specialized to bind to certain molecules called substrates. The site at which the substrate binds is called the active site. Often, the enzyme catalyzes changes in the substrate. An enzyme that is the focus of pharmaceutical research is often called a target. Other types of targets include receptors[5], which are biological macromolecules associated with cell membranes, and genetic material, which encodes the proteins in a cell.

Assays are experiments in which an enzyme’s consumption of substrate is measured over time. If a test compound is present that has inhibitory activity against the enzyme, a decrease in turnover of substrate is observed. This decrease is commonly expressed as the compound’s IC50 or its Kᵢ. IC50’s and Kᵢ’s are derived by different mathematical formulas, and differ in ways that have consequences for drug discovery. Many legacy sources of assay data will report either one or the other, and although they have the same units, they are not directly comparable. The assays described above are all carried out in vitro, (in an artificial environment). Many compounds that show activity against a target in vitro show little or no activity in the whole organism. For this reason, in vivo (in life) assays with the same target are also often done. A compound can inhibit an enzyme via different mechanisms. These distinctions have consequences for drug design and compound activity data should represent them.

Unfortunately, data in legacy databases often does not distinguish between different types of inhibition.

The data warehouse must store information about different types of assays and targets (enzyme, receptor, etc) as well as indicating when a chemical compound is found to be active in an assay, what type of activity is observed, and to what degree (value of IC50 etc). Since there are a variety of measures of activity, and they mean different things, the data warehouse should be able to store them all, and to prevent meaningless comparisons between semantically unlike measurements.
2.3: Structural biology

The areas of biology that impact on drug discovery are too numerous to describe here, however, the study of the three-dimensional structures of enzymes and receptors generates data that is vital to the data warehouse. Of particular interest is the three-dimensional map of the active site. One of the queries against the data warehouse that will be run repeatedly is: “Which compounds in our database look like they might bind to this active site?” Structural data may represent one of the greatest challenges to the data warehouse. In order to exploit this data, a format must be found in which three-dimensional comparisons can be made, and algorithms must be written to transform data to that format.

The two most common methods for obtaining the 3D structure of a target are X-Ray crystallography[6] and nuclear magnetic resonance (NMR) spectroscopy[7]. The result of both of these methods is a set of x, y, and z coordinates for every atom in the target. NMR studies produce a family of structures, similar to the conformers of a chemical compound. Certain other experiments, such as site-directed mutagenesis (changing the chemical nature) of specific regions of target proteins can generate partial structural data.

2.4: Computational modeling

Modelers do the work that the data warehouse is designed to facilitate—experiments to gain predictive intelligence from existing data of many kinds. Modelers are likely to be the heaviest users of the data warehouse. The three experiments described below form the basis of the queries that modelers will probably run most often:
1. Homology Modeling[8]: Given a target whose 3D structure is known, predict the 3D structure of another target whose composition and/or function is similar.
2. Target-Compound Docking[9]: Given the 3D structures of a target and a compound, try different ways of fitting the compound into the active site, to see if the compound is likely to interact with the target.
3. Similarity Searching[9]: Compare a compound that is known to interact with a target with other compounds to determine if they may interact with the target.

The files that modelers use to represent the three-dimensional structures of compounds and targets are complex, and are the focus of active research and refinement. Typically, the output of structural comparison programs are a set of scores[9]. For example, if an entire collection of compounds were tested for similarity against a probe compound, the result would be a score, e.g. from zero to 1, for every species in the collection.

3: Target users and their questions

Potential users of the data warehouse are modelers, chemists, biologists, enzymologists, and managers. The following questions form the basis for queries that these users may wish to run against the data warehouse.

Enzymologists: An enzymologist has a new assay and has discovered no active compounds, however, the three-dimensional structure of the enzyme’s substrate is known. Q1. What subset of compounds in the database is structurally similar to the substrate?
If the X-Ray crystal structure of the enzyme is known: Q2. What subset of compounds looks like it might fit into the enzyme’s active site?
Perhaps an enzymologist has a well-established assay, with some actives already discovered.
Q3. What compounds are there in the database that are structurally or functionally similar to the known actives?

Managers: Managers will want to run queries that help to evaluate the performance of the enterprise.
Q4. In which assays have we been getting a lot of leads? In which assays have we been getting very few?
Q5. Which programs are producing a lot of drug leads? Which are producing few?
Q6. What compounds do we have that are structurally or functionally similar to the competition’s leading drug?
Q7. Which compounds are likely to cause side-effects by interacting with targets that are essential to human health?
Q8. We have just bought a sizable collection of compounds from an academic lab. Which assays should they be tried in first? Can we load the whole set of compound data for hundreds or thousands of new compounds and have them systematically tested for:
   a. Similarity to known inhibitors.
   b. Similarity to known substrates.
   c. Docking into the active sites of targets.
What if we could run these queries before we spent tens of thousands of dollars buying this collection of compounds?
Q9. We add 200 new compounds to our collection a week. Can they be automatically tested as in Q8 when they are loaded?
Q10. On the average, how many compounds did we have to assay for a target before we found a lead? For
each target? This year? Last year?

Chemists: A chemist has synthesized Compound X, which is active in assay A.
Q11. What other assays might it be active in?
Q12. For a set of compounds that are active in an assay, are there structural or functional similarities in the set?
Is there a correlation between degree of activity and a particular trait, e.g., charge, or hydrophobicity?

Modelers: Modelers will run most of the queries above, and will probably come up with new uses for the data warehouse. For the present, we can assume that their most important questions are related to target-compound docking (represented in Q2 and Q8), and similarity searching (represented in Q1, Q3, Q6, Q8, Q9, and Q12).

4: Identification of important entities

Compound: Compounds comprise the intellectual wealth of the pharmaceutical company, and the purpose of the data warehouse is to find out what they are good for. Some of the characteristics of compounds that should become attributes are: (compound#, molecular_formula, structural_formula, logP(a measure of hydrophobicity), conformers {a set}, 3D_structure, charge).

Structural Motif: Two other characteristics of compounds are the structural classes they belong to and the functional groups they contain. One reason we should model them is that we will often want to group compounds by these criteria. The distinction between global and local structural motifs is often rather arbitrary, especially with large, complex compounds, so it may be advantageous to combine these two characteristics into one table. Some of the data we may want to store about structural motifs are their structural formulas and descriptions.

Assay: The most important information about assays are the results when compounds are tested in them, such as IC50’s, Ki’s. (These numbers could also be modeled as attributes of compounds.) Other information we may want to store about an assay is data about the substrate; its structure, conformers, and properties in particular, so that we may compare them to compounds in the database. We may want to store the assay’s name, an identifying number, and a description. We will often want to group compounds by the assays that they are active in.

Target: We will want to run queries about compounds that might fit into active sites of targets. Clearly, the most important attribute of a target for this purpose is its three-dimensional structure. Some of the other data we may want to store about targets are: (target_name, target_type, protein_sequence, DNA_sequence, gene_locus, organism). We will often want to group compounds by the target that they are active against.

Program: Pharmaceutical companies usually group research activities according to what drug market they are directed at. These markets, or disease areas are often called programs. We would like to be able to group compounds by program, e.g. How many actives do we have in the diabetes program? Targets also get grouped by program. The company probably has assays running for different targets in the same disease process, so assays also get grouped by program. Some of the data we may want to store about programs are: (program#, program_name, organism).

Time: It is clear, from managers question Q10 that timestamps will be important, to aid in measures of performance of the enterprise. Timestamps are also important in identifying more recent additions to the compound collection, in the case of automated queries that should be run on all new compounds.

Library: Structural classes and functional groups have even greater significance for combinatorial libraries than for single compounds. The reason for this is the rationale behind creating a library. Usually, a library is created based on knowledge of a single active. A known active in a structural class may have a given set of substituents {A, B, C}. The library may be designed to generate all the possible combinations of substituents at those three sites. We will want to be able to query the database for libraries that should be tested in certain assays. We may also want to store data about how the library was made.

Lead: When a compound is found to be active in an assay, it becomes a drug lead. It may be useful to make a distinction between compounds and leads by creating a separate fact table. One reason for this is that the sources of data for compounds and for leads are different. Compound data is generated by chemists, and activity data by enzymologists. It is likely that the refresh schedules of these two tables will be different. Second, data about compounds and about leads are likely to have different access patterns. Finally, data about drug leads is more sensitive than data about
compounds, so the company may wish to have a higher level of security for lead data.

5: Dimensional data model

5.1: Fact tables

The central facts in our model are about compounds. We have chosen to model them in three different tables. Since some of the relationships between compounds and dimensions are many-to-many, the fact tables are not in second normal form. Specifically, one compound can be tested in many assays. One assay tests many compounds. Also, one compound can have more than one structural motif and one structural motif can be found in many compounds.

Figure 1. Star Schema for Drug Discovery Research

---

**Compound**: This table contains data about the compounds in the company’s collection. It can also store public domain data about compounds, for example, a compound published in a journal article that is of interest to the company for modeling purposes, or a virtual compound generated by computational predictions. The attribute ‘source’ can denote whether the compound is in the collection, is public, modeled, etc.

**Lead**: This table contains data for all compounds in the company’s collection that have been found to be active in one or more assays. The compounds represented in this table are also represented in the Compound table, hence Leads are a subset of Compounds. Public domain data can also be represented in this table. The lead key and compound key have the same domains, and can be used for a join between Lead and Compound tables.

**Library**: This table contains data about all combinatorial chemical libraries in the company’s collection. Since the representation of, and the theory behind combinatorial libraries is still evolving, this table does not represent all of the attributes that a company may want to store about them. This table is included in our model for future development.

The keys for all fact tables consist of all foreign keys, plus a unique key. (A schema that with this key structure is called a multistar schema[10].) This is because concatenation of foreign keys is not sufficient to uniquely identify each instance. Also, most companies have a number identifier assigned to each compound that, for dealings with government regulatory agencies, should remain associated with the compound in all references.

5.2: Dimension tables

These tables contain descriptive data about the enterprise and are categories that we may want to group chemical data in. Two dimension tables, Assay and Target, have foreign keys. This is because the Target dimension is an outboard table[10] for Assay and the Program dimension table is an outboard table for Target.

**Time**: The time dimension allows grouping of data by timestamps, and will aid in enterprise performance evaluation. It will also allow automated queries to be performed on the freshest data in the warehouse.

**Structural_Motif**: This dimension allows grouping of compounds, libraries, and leads by the structural motifs, both local and global, that they contain. Medicinal chemists think of compounds in terms of therapeutic classes, so this table contributes to the understandability of the warehouse from the user standpoint.

**Assay**: This dimension stores descriptive data about assays and allows data to be grouped by assay. It can
also store public domain data about assays and substrates. In the case of compounds and libraries, it will allow analysis of which ones should be tested in which assays. In the case of leads, it stores descriptive data about the assays that leads were found to be active in.

**Target:** This dimension stores the three-dimensional structures of disease targets so that they can be compared to those of compounds. It can also store public data about enzymes, receptors, and other targets. It allows grouping of assays, compounds, leads, and libraries by target.

**Program:** This dimension allows grouping of targets, assays, and factual data by the company’s programs. It will aid in managers’ evaluations of performance in different program areas.

### 6. Discussion

Data warehousing was developed with certain kinds of business applications in mind, specifically those that deal with numbers. The attributes at the core of drug discovery are not additive. This means that some of the power of data warehousing, such as drill-up and drill-down queries[1], and summarization[1] may be lost. However, the types of queries that are required for drug discovery research call for analysis of compound data by multiple criteria, which is what the dimensional data model facilitates.

The data about compounds at a given point in time reflects what is known about them at that time. The discovery that a compound is active in an assay changes the snapshot-in-time nature of the data. Updates are against the rules in the data warehouse[1]. This is why compounds have been modeled in two fact tables, Compound and Lead. The discovery that a compound is a lead is dealt with by a refresh of the Lead table. This doesn’t mean that the compound is deleted from the Compound table. The redundancy created by this approach is semantically desirable and may be tolerable, in the light of the relatively small size of the Lead table compared to the Compound table.

The best description of the relationship between a compound and a structural class that it belongs to is the object-oriented ‘isa’ relationship. Moreover, the best way of describing the relationship between a compound and the functional groups that it contains is the object-oriented whole-part relationship. The relationship between a library and its component compounds (not addressed in our model) is a whole-part problem too. In an object-oriented world, the deconvolution of a library into its parts could be a part of an object’s operations. Inheritance hierarchies correspond well to the native semantics of chemical structure families. A dimensional data model that incorporates object-oriented concepts could be a useful innovation for this problem.

There is one dimension that we have not modeled but that will appear in the result set of nearly every query we run, and will be the most frequently used Group_By criterion. A score is calculated every time a compound is compared to another compound, or docked into a target structure’s active site. Scores can be considered derived attributes. We may want to save scores to avoid the overhead of recalculating them. Also, although we have not modeled Score explicitly, users will need to know how to employ it, so it should be part of the GUI front end of the warehouse. Since the GUI will most likely be built based on the schema, it might be a good idea to add a virtual dimension called Score.

### 7. Conclusion

In this paper, we have derived a dimensional data model for a data warehouse for pharmaceutical drug discovery, using a generic enterprise description and a list of questions that users are likely to wish to pose to such a database. The central focus of the model are chemical compounds, which are represented in three different fact tables: Compound, Library, and Lead. The dimensions of the warehouse are Assays, Targets, Programs, Time, and Structural_Motif. The enterprise will differ from company to company, however, we feel that this model, with customization, can be applied to the real-world pharmaceutical drug discovery problem.

To the best of our knowledge, no pharmaceutical company has yet undertaken a data warehouse project for drug discovery. We hope this analysis shows that, although the problem is complex, it can be approached. If data warehousing and the dimensional data model are applied correctly, they can solve the problems caused by doing predictive computational chemistry in the legacy database environment. Most important, the integration of all of a company’s data about drug discovery in a single environment optimized for analytical query-processing could result in intelligence that shortens the research cycle that leads to a new drug.

### 8. References


