Scientific OLAP for the Biotech Domain

Nam Huyn

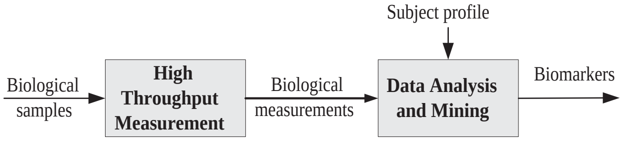
SurroMed, Inc.
2375 Garcia Ave, Mountain View, CA 94043, USA
phuyn@surromed.com

Abstract

On-line analysis of clinical study data has not benefited from recent advances in OLAP technologies. We examine the analysis requirements posed by the biotech domain that are not met by traditional OLAP. To accommodate these new requirements, we propose the concept of *Scientific OLAP* which applies more broadly to data analysis in controlled scientific experiments. We describe our experience implementing such a system for the support of biomarker discovery and we identify some key challenges that must be overcome before OLAP can be widely adopted in the biotech industry.

1 Introduction

A central mission among a growing number of biotech companies is to discover biological markers. A *biological marker,* or biomarker, is a "characteristic that is measured and evaluated as an indication of normal biological processes, pathogenic processes or pharmacologic responses to therapeutic intervention" [3]. For example, high levels of cholesterol in human blood have commonly been used as a biomarker for heart diseases. New biomarkers are being sought that enable diseases to be diagnosed more accurately or earlier than is currently possible. Thanks to breakthroughs in *high-throughput measurement* technologies in the last five years [5], tools such as gene chips, protein chips, and mass spectrometry are now widely available that are capable of detecting hundreds of thousands of gene products, proteins, and small organic molecules. These tools enable biotech companies to routinely generate, from tiny volumes of biological materials, very high volumes of measurement data that must be summarized, compared, and viewed efficiently.



While these data analysis tasks are critical to the success of these companies in biomarker discovery, OLAP tools (see [1] for an excellent survey of On-Line Analytical Processing) have not been widely adopted by the biotech industry. To understand why, let us take a closer look at the nature of data generated in *clinical studies,* i.e., controlled scientific experiments designed to answer specific clinical research or engineering questions such as drug efficacy, biomarker identification, and measurement method validation. Typically, the protocol for a clinical study specifies the following "ingredients": subject population, i.e., a well-characterized collection of subjects to be included in the study; biological samples, i.e., what kinds of samples (e.g., tissues, body fluids), how many and when they are drawn from the subjects; measurement methods, i.e., biological/chemical assays and instruments used to analyze the samples. For a clinical study aimed at evaluating drug efficacy, a view of the data schema might look like this:

subject	*draw*	*clinicalCls*	*drugCls*	m_1	m_2	...
John	1	Asthma	A	3.1	5.4	...
John	2	Asthma	A	4.6	5.3	...
Jane	1	Healthy	B	1.2	5.5	...
Jane	2	Healthy	B	1.7	5.6	...

Each row in this table corresponds to an observation, i.e., a biological sample with all its characteristics and measurements performed. The *draw* column represents the time point when the sample is taken, the *clinicalCls* (resp. *drugCls*) column represents the disease (resp. drug) group which the subject belongs, and the m_i's represent biological measurements. This example illustrates the fact that clinical study data have a natural multidimensional view, where observations

**Proceedings of the 27th VLDB Conference,
Roma, Italy, 2001**

are the facts of interest, *draw*, *clinicalCls* and *drugCls* are the dimensions, and the biological measurements are the target measures. While this view of clinical study data suggests that OLAP tools may be used for their analysis, a closer look reveals some fundamental differences between clinical studies and traditional OLAP applications:

- The subject population is carefully selected to minimize sampling biases, especially when the number of these participants is limited (typically in the 100's). Also, biological samples are drawn at carefully planned time points.
- Observations are linked to subjects, while in traditional data analysis applications, subjects are usually not tracked across transactions.
- The number of measurements made on each biological sample is several orders of magnitude larger than the number of samples, while in traditional OLAP applications, the number of facts usually far exceeds the number of target measures.
- An important goal of data analysis in clinical studies is to validate hypotheses following established scientific methods, e.g., to validate drug efficacy in clinical trials or to evaluate assays in biomarker discovery.

2 Scientific OLAP for Clinical Studies

These differences translate into data analysis requirements that are not found in traditional OLAP and that turn out to be very important for the domain of clinical studies. Generally, these requirements include more rigorous and richer types of data analysis using established statistical methods, more stringent notions of comparisons, the need to qualify results to minimize chances of making the wrong inference based on a limited number of observations, and the ability to handle large numbers of target measures. We propose the concept of *Scientific OLAP* as an extension of traditional OLAP that accommodates these unique requirements, which we describe below. For further details, we refer the reader to [2].

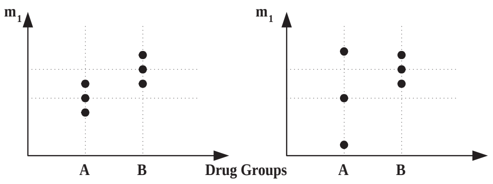
Rank-based aggregation Notably missing from traditional OLAP and SQL systems are the MEDIAN operator and the more general PERCENTILE operator. However, in many experimental sciences and in biology in particular, summarizing data using medians and percentiles is the **norm**, for good reasons. First, measurable biological entities, such as the concentration of many proteins expressed in human serum, often are not normally distributed. For these biological entities, MEDIAN gives a more accurate summary than AVG. Furthermore, measurements often are noisy and error-prone, which make MEDIAN a more robust operator against outliers. A partial solution has recently appeared in some commercial systems where SQL is extended with a family of functions, called *analytic* functions, that provides better support for analytical processing. An example is the RANK analytic function which computes the ranking for each row in a rowset, relative to a row-dependent group of rows. To illustrate how this function works, consider the view from Section 1 and let us call this view *observations*. The following query:

```
SELECT subject, clinicalCls, m1,
  RANK() OVER
  (PARTITION BY clinicalCls ORDER BY m1)
FROM observations WHERE draw =1
```

computes the ranking in m_1 of all observations at time 1 within each clinical group. This ranking will be useful for computing the median in m_1 for each clinical group. The RANK analytic function may be used to implement medians and percentiles, but the lack of true rank-based aggregations makes the implementation of many statistics commonly used in clinical studies both cumbersome and inefficient.

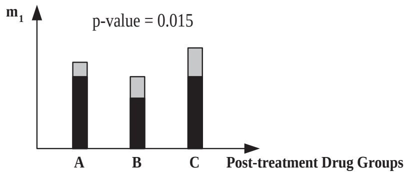
Multiple-group comparison A common question in clinical studies is whether or not several groups of observations differ with respect to some measures in any *significant* way and not by chance. For instance, to study the effect of several drugs on human subjects, a separate group of subjects is often recruited for each drug, and in order to ensure that no bias has been introduced in the drug group assignment, it is important to verify that the drug groups exhibit no significant differences before any drug is administered. Another common example of group comparison arises in studies for diagnostic markers where a battery of measurements is performed on subjects that belong to different disease groups and where measurements that show significant differences between the groups are to be identified.

Support for multiple-group comparisons in traditional OLAP systems is typically limited to using first order statistics such as the mean. However, as the following figure illustrates, these statistics are no longer sufficient to detect subtle but important differences.



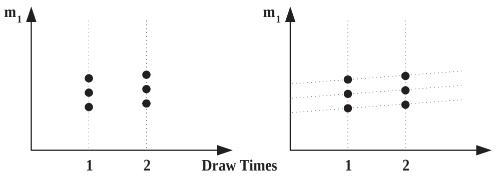
In this figure, the differences in means of measurement m_1 between drug groups A and B are identical in both graphs. However, since the values are more scattered in the right graph than in the left graph, intuitively the difference on the right should be less significant than the difference on the left.

Thus, in order to support group comparisons which are clearly more stringent in clinical studies, summaries in OLAP must include not only the group averages but also some second order statistics such as the variance within each group and some measure of how significant the differences are. OLAP front-end tools that support richer visualization are also needed. For instance, the effects of drugs A, B, and C on measurement m_1 might be summarized in the following plot:



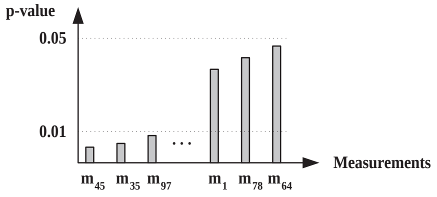
using something called the *p-value* to measure the probability that the drug effects are identical by chance (i.e., the smaller the p-value is, the more significant the difference becomes). Many statistical tests can be used to measure how significant a difference is. Commonly used ones, such as the *ANOVA F-statistic* and the *Kruskal-Wallis statistic* (see Analysis of Variance in [4]), can be implemented using standard SQL aggregations with the help of the `RANK` analytic function.

Repeated observations So far, in comparing groups of observations, we have ignored that observations from different groups may be related to each other. For example, some measurements are made on the *same* subject albeit at *different* time points, e.g., before and after treatment. These related observations, or *repeated observations*, exemplify what is known in classical statistical testing as *repeated measures.* If we ignore these relationships, we may fail to detect small but significant group differences, as the following figure illustrates:



On the left, the difference between groups is not statistically significant, since the difference in mean is small compared with the variance within each group. But on the right, with the additional knowledge that the observations are paired, intuition tells us that the difference should be significant, since the observed values consistently increase as we move from one group to the other, albeit in very small amounts. Note that repeated observations are distinct from time series for which trend analysis is supported in many OLAP systems, since a time dimension is not required. Also, traditional multidimensional models have no provisions for capturing the concept of repeated observations. Otherwise, significance testing is not difficult: statistics commonly used for comparing groups of repeated observations include the *Paired T-Test* and the *Wilcoxon signed-rank statistic* (see [4]), which can be implemented using standard SQL aggregations and the `RANK` analytic function.

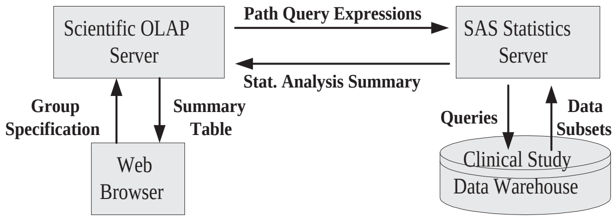
Scaling with the number of measures In Section 1, we used a multidimensional view of the clinical study data where each measurement is treated as a separate target measure. Since the number of measurements in typical clinical studies is extremely large (say in the 10,000's), this view is not practical: in order to visualize the summary statistics for all the measures using traditional OLAP front-end tools, one would have to sequence through a large number of screens! A better alternative is to represent the measurement type as a dimension. Thus, *slicing* on a particular measurement would show a summary of the comparative statistics for that measurement. This representation also allows us to compare all the measurements side by side in the same plot, to use common OLAP operations such as *dicing* to view only those measurements whose difference satisfies a user-specified significance threshold, and to rank the measurements according to their level of significance. An extended OLAP front-end tool might visualize the significance of the measurements in one chart as shown in the following plot:



which would help quickly reveal the important measurements.

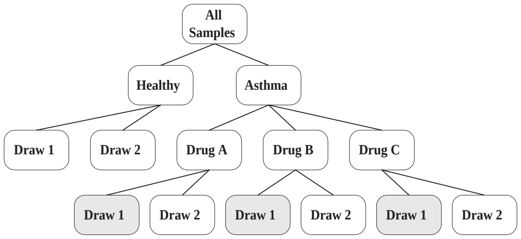
3 Implementation Experience

The following figure depicts an on-line data analysis tool we implemented directly on top of a relational database, which we use routinely to analyze clinical study data:



In this implementation, comparative statistics, included in most analysis result summaries, are evaluated in a statistics server, separate from the database server. While some statistics could have been implemented using straight SQL, the use of an established statistical computation engine to compute them is purely for acceptance reasons, that is, at least until a database extension certified for statistical analysis is available. This decoupling results in processing inefficiencies mainly due to high volumes of network traffic and to the inability to take advantage of query optimization: for instance, instead of relying on the relational engine to optimize the execution of an aggregate query, the data is aggregated on a group-at-a-time basis. Also, because aggregate view materialization is not used, every new view request is evaluated against the base data, which results in further delay in processing the request.

To specify how data is to be aggregated and compared, we do not use the cube manipulation metaphor embodied in traditional OLAP front-end tools. Instead, the interface allows the user to recursively partition a given group of observations along any dimensions into subgroups, and to select arbitrary subgroups to analyze or compare. To illustrate this approach we call *dynamic group specification*, the following shows a hierarchy of groups of observations that the user obtained by first expanding the top node (representing the initial set of observations) along the *clinicalCls* dimension, and by expanding the remaining nodes along the *drugCls* and *draw* dimensions:



From this hierarchy, if the user wants to compare the different drug groups of asthma patients, he would select the nodes as highlighted in the figure. The advantage of our approach is two-fold: first, in order to view a particular aggregation summary, the user is not required to “navigate” through summaries for the intermediate aggregations, which may involve unnecessary computations; furthermore, since our group hierarchy does not require two nodes from the same level to be expanded along the same dimension, our method of group specification provides more flexibility than traditional OLAP systems. However, the lack of navigational capability is also a disadvantage, because it does not allow the user to follow a train of thought. Also, our approach does not scale well with dimensions that have a large number of distinct values. Finally, because the number of measures can be large, the summary of (comparative) analysis is shown in a table format, one row per measure, instead of a bar chart format used in a typical OLAP system.

4 Summary and Future Challenges

We proposed the concept of *Scientific OLAP* which extends traditional OLAP to accommodate the unique requirements posed by not only clinical studies, but possibly many other controlled scientific experiments as well. However, to provide an efficient implementation, several key challenges remain to be addressed.

Precomputing rank-based aggregations A common approach used in many OLAP systems to speed up aggregate queries is to use materialized subqueries to answer the original queries. However, most rank-based aggregate operators (e.g. MEDIAN) are not associative, and the use of materialized queries to optimize queries involving these operators is not obvious.

User-defined percentiles Medians and percentiles do not have a standard definition, especially for even-sized sets of values and bags. Short of providing a generic user-defined aggregation facility, it is not clear how to support all their variant definitions efficiently.

Custom comparative statistics Among the commonly used comparative statistics techniques, many are difficult to express as a composition of SQL aggregate queries. Implementing these techniques requires using sophisticated aggregation mechanisms that can be difficult to provide (see [2]).

Large scale visualization Traditional OLAP front-end tools provide a very limited form of visualization: bar charting. Comparing a large number of measures (say in the 10,000’s) requires using visualization techniques beyond bar charts that should be both intuitive and compact, and that can be implemented efficiently. The challenge is to identify such a powerful and general technique.

Finally, due to space limitations, this paper focused on OLAP for clinical studies only. We refer the reader to [2] for another challenging problem: mining clinical study data for biomarkers.

References

[1] S. Chaudhuri and U. Dayal. An Overview of Data Warehousing and OLAP Technology. In *SIGMOD Record,* 26(1), pp. 65–74, 1997.

[2] N. Huyn. Data Analysis and Mining in the Life Sciences. To appear in *SIGMOD Record,* 30(3), 2001.

[3] Biomarkers Definitions Working Group. *Biomarkers and Endpoints in Clinical Trials: Preferred Definitions and Conceptual Framework.* National Institutes of Health.

[4] R. L. Ott. *An Introduction to Statistical Methods and Data Analysis.* Duxbury Press, 1993.

[5] J. Ren. High-Throughput Screening of Genetic Mutations/Polymorphisms by Capillary Electrophoresis. In *Combinatorial Chemistry & High Throughput Screening*, 3(1), pp. 11–25, 2000.