

ORACLE DATABASE 10G FUNCTIONALITY FOR BIOINFORMATICS

Tailoring Relational Database Management Systems (RDBMS) for Life Sciences has been winning considerable attention over the past two years. The advances in high-throughput experimental methods, such as genome sequencing projects or transcriptome profiling, has resulted in increasing data volumes and a greater range of data types. This growing complexity of discovery data requires increasingly sophisticated data management and analysis tools. Historically, RDBMS have been used simply to store data. However, as data volumes and complexity grow there is an increasing demand for powerful analytical tools to be embedded into RDBMS.

Integrating analytical capabilities into the RDBMS simplifies data management and increases analytical performance by minimising the overhead of moving data out of databases, enables data to remain in a secure, audited, scalable and highly available environment, and allows a single query to span many data types and data sources. Furthermore, many RDBMS offer workflow capabilities, which allow users to rapidly build flexible, automated bio-analytical pipelines.

A good example of a RDBMS that has integrated analytical capabilities that are specifically tailored to the requirements of the life sciences industry is Oracle Database 10g. Much of the analytical functionality is provided free with the database. However, some functionality is made available as separately priced options.

In this section, we briefly describe some of the more interesting analytical features of Oracle Database 10g. Oracle is a commercial RDBMS, but it is available in many companies, universities and academic research institutes. For more information about Oracle's efforts in life sciences, see http://www.oracle.com/technology/industries/life_sciences/index.html

Statistics

A variety of statistical functions are now available in the Oracle database. These include descriptive statistics, hypothesis testing, distribution fitting, correlations, Pareto analysis and cross tabs. Statistical functions in the database can be used in a variety of ways, for example, users can call Oracle's DBMS_STAT_FUNCS to obtain basic count, mean, max, min and standard deviation information for their data set; or users can determine the strength of relationships using hypothesis testing statistics such as a t-test, f-test or ANOVA.

Oracle Data Mining

Oracle Data Mining (ODM) provides data-mining algorithms that are embedded natively within the database. ODM provides a range of algorithms for clustering, classification and prediction.

In supervised learning, a target field or dependent variable is identified (eg drug response). The supervised learning technique then sifts through data trying to find patterns and relationships between the independent variables (such as drug, dose or administration method) and the dependent variable. For supervised learning, ODM has embedded in the database naive Bayes, adaptive Bayes networks and support vector machines.

In unsupervised learning, the objective is not indicated to the data-mining algorithm. Associations and clustering algorithms make no assumptions about the target field. Instead, they allow the data-mining algorithm to find associations and clusters in the data. You can use unsupervised learning to identify new disease classes for example. For unsupervised learning, ODM has embedded in the database association rules, hierarchical k-means clustering, O-cluster and non-negative matrix factorisation.

All of ODM's functionality is accessible via a Java-based application programming interface (API) and a PL/SQL API. In addition, ODM supports the mining of both structured data (numeric and

categorical data types) and unstructured data (text). Data preprocessing and automated transformation capabilities are provided.

ODM BLAST

Oracle has embedded in its database a version of BLAST similar to NCBI BLAST 2.0. The advantage of embedding BLAST functionality in the database is that it is possible to incorporate a BLAST search into a regular SQL query. This enables scientists to pre-filter their data prior to a BLAST search and then post-process the data prior to viewing the results. For example, users could perform a query that retrieves all sequences that meet a similarity threshold, and where the sequences were added after 'January 2002', the annotation included the word 'cardiac', and the individual was called 'John'. This filtering of sequence data can lead users to more rapidly retrieve the precise information that is of interest. In addition, it is possible to run automated batch processing of BLAST jobs within the database environment, making it easier to perform whole organism sequence comparisons for example.

Regular Expression Searches

Regular Expression Searches are supported in SQL and PL/SQL. Regular Expressions provide a powerful search and replace capability used by many in UNIX and Java environments. Support in the database allows these developers to write simple queries that previously would have taken multiple lines of SQL code. The Oracle Regular Expression Searches functionality can be used in the Life Sciences, for example, to identify Prosite motifs in protein sequences.

Text analysis

The Oracle Database provides a comprehensive information retrieval API called Oracle Text that allows developers to build text search applications. Oracle Text uses standard SQL to index, search and analyse text stored in the Oracle database, in files and on the web. Oracle Text can

perform linguistic analysis on documents, as well as search text using a variety of strategies including keyword searching, context queries, Boolean operations, pattern matching, mixed thematic queries, HTML/XML section searching and so on. It can render search results in various formats including unformatted text, HTML with term highlighting, and original document format. Oracle Text supports multiple languages and uses advanced relevance-ranking technology to improve search quality. Oracle Text also offers advanced features such as classification, clustering and support for information visualization metaphors.

Network analysis

The latest version of the Oracle Database includes a feature that enables users to model and analyse data as if it were in a graph representation. This functionality allows users to represent data as a series of nodes and links, which can then be traversed to find information of interest. This feature is provided with a range of functions that lets users perform analyses such as the shortest path between two points, tracing whether nodes can be connected, within-distance analysis, minimum cost spanning tree and nearest neighbour. This functionality has the potential to be used in many ways in the life sciences, including managing and analysing biological pathways and protein-protein interactions.

Table Functions

Many life sciences computational applications require extensive data processing with complex algorithms. Table Functions allows researchers to implement their own compute intensive algorithms in PL/SQL in the database or in C/C++ outside the database. These algorithms are then pipelined within the database for significantly improved performance on data analysis.

Summary

In summary, Oracle Database 10g has a range of analytical functionality embedded in it, which enables researchers to carry out many bioinformatics tasks within the database environment. Oracle Database 10g is a valuable product for the Life Sciences, as it enables analytical functionality of a variety of different biological and chemical data types, while also offering a scalable, secure, highly available data repository. The key advantage for software developers is that it simplified application development by allowing developers to invoke the database functionality and thereby saving in-house development efforts.

Reddy Gali

The Bauer Center for Genomics Research,
Harvard University
and

Susie Stephens
Oracle Corporation