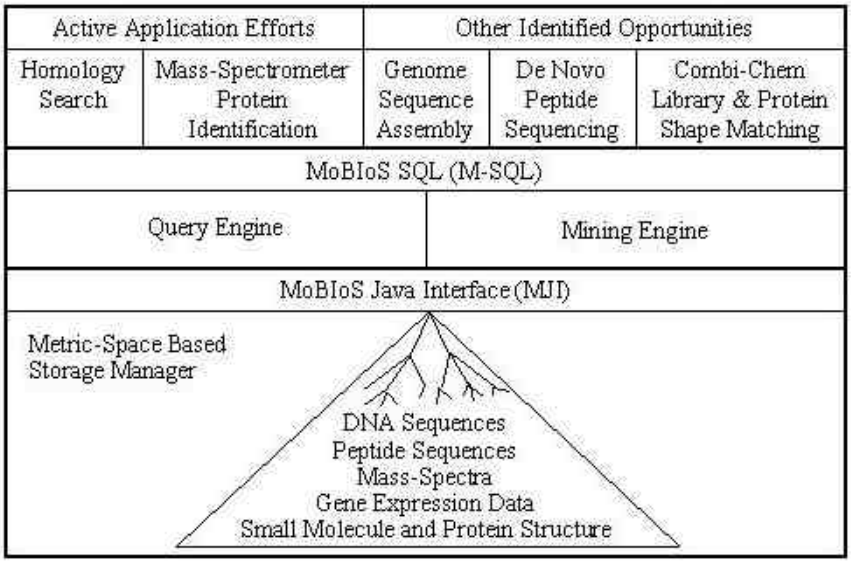
MoBIoS: The Molecular Biological Information System

Daniel P. Miranker

Department of Computer Sciences
University of Texas
miranker@cs.utexas.edu



Across life sciences, exploration of key data comprises sequentially scanning all of the data to locate information that matches even the most basic patterns; (O(n)). This includes genomic and proteomic sequence data (BLAST searches), proteomic mass-spectra and the combinatorial chemistry libraries critical to rational drug design. Evangelical pundits have adopted the computer metaphor and claim that the amount of biological data is growing exponentially. They are wrong. It is growing faster. The gulf between computer processing power and data analysis problems in life sciences is growing at an accelerating rate.

MoBIoS is a next generation database engine which embodies a solution to this most basic problem in life-science data management; how to locate matching data in a fast, scalable manner; (O(log n)) The technical accomplishment is similar, by precise analogy, to moving rows of customer information from a text file into a relational database (e.g. Oracle). For large data sets this can mean orders of magnitude speed improvements.

Conventional databases use index mechanisms which exploit the ordinal nature of business data. For example, a list of customer records may be alphabetized or ordered by customer id. In either case, it is understood ahead of time that there is a mapping of data to numbers and the numbers can be used to address the data. For MoBIoS, database index methods are being developed that rely only on an ability to measure the similarity of two data objects relative to each other. No absolute reference coordinates are used. Thus, the methods can be applied when there is no geometric interpretation of the data. (When the data has no meaning in Euclidean space.) The general problem, known as metric-space indexing has been well studied for problems that fit in main-memory. MoBIoS is the leading effort in the development of database management techniques applying metric index methods to large-scale problems.

MoBIoS is implemented in Java starting with the Mckoi open-source relational database engine. A metric-space index was added to Mckoi as well as new biological class libraries. Embodying these developments in a general purpose database structure further enables MoBIoS to uniquely serve as a life science data warehouse and at the same time will allow application products may be built by SQL application developers using commodity SQL and Web-based application development platforms.

There is precedent for this expansion of database architecture. Relational databases do not handle 2- and 3-dimensional data types. (e.g. geographic data, aircraft telemetry). Both IBM and Oracle sell “spatial extensions” to their relational products. These extensions comprise special index structures and libraries of new data types and form the basis of Geographic Information Systems (GIS).