

Software Tools to Support High-Throughput Materials Discovery

George Fitzgerald*, Max Petersen, and Michael Doyle
Accelrys, Inc., San Diego, CA 92121 (USA)
*gfitzgerald@accelrys.com

This presentation demonstrates how software tools for data mining, analysis, and statistical modeling can be used to facilitate the discovery of new materials in a high-throughput lab environment.

Owing to the volume of data created during high-throughput discovery, one of the challenges in this field is mining the data for useful knowledge that can help in the search for optimum materials such as catalysts or pharmaceutical compounds. Generally, the volume of data and the need for rapid analysis make it infeasible for a human researcher to examine each data point and make crucial decisions regarding the next set of experiments. One would prefer an automated system that can process large volumes of data and provide intelligent guidance as rapidly as possible. To increase the reliability of such a system, it should make maximal use of all the available data.

We present a new system that can mine a database for information such as structure and composition, and then analyze correlations with performance characteristics. Uniquely, the method can also incorporate: images; analytical data such as spectra; and sample processing steps. Further, the ability to mine this information from a central database makes it possible to update the reports in real time.

Models of activity may be constructed on the basis of, for example, recursive partitioning, neural nets, or genetic function algorithms. While some of these methods require a “data-rich environment” others such as recursive partitioning can be utilized even in cases where data are sparse. These provide qualitative and quantitative predictive tools which may be combined with optimization algorithms to identify optimal materials. This method can be automated to steer future experimental design on the basis of the predicted optimum.

Examples of the application include recursive partitioning models based on IR and NMR spectra and optimization of dyes using genetic algorithms.