Data Mining Advances
Bioinformatics Insight

A wealth of data flows through research organizations every day. This information holds, and often hides, valuable insight and new understanding. Leading organizations are now incorporating Clementine, the leading data mining workbench, into their operations — turning this deluge of data into faster, more accurate results.

Here’s how one researcher is gaining new insights into the genetic components of disease with SPSS Inc.’s Clementine.

Bioinformatics professor Werner Dubitzky finds biological gold with Clementine

Decoding the human genome ignited the biotech revolution. It also created one of the biggest hurdles the Information Age has ever faced: How do you make sense of three billion base pairs of human DNA? The answer promises to cure every disease with a genetic component, including cancer, asthma, diabetes and mental illness.

The new science — bioinformatics

This grand genomic dilemma has thrust the newborn discipline of bioinformatics, the use of computers and information technology to tackle biology problems, into the limelight. By 2004, life science companies will invest $6.5 billion in computer-aided methodologies such as artificial intelligence and data mining, according to market research firm Frost & Sullivan.

One of bioinformatics’ main goals today is the important task of assigning functions to the estimated 33,000 genes and determining how they interact with one another. Professor Werner Dubitzky, Chair of Bioinformatics at the University of Ulster in Northern Ireland and member of the Editorial Board of the Online Journal of Bioinformatics, Briefings in Bioinformatics and Physics of Life, is at the forefront of applying data mining analysis, a tool traditionally used in market research, to draw such conclusions from biological data.

Mining microarray data with Clementine

Dubitzky relies on cutting-edge software like SPSS Inc.’s Clementine, a data mining tool he first used on a project at the University of Ulster in 1996. Dubitzky says, “I like Clementine because of its ability to quickly generate results and its intuitive user interface.”

One of Clementine’s most important applications is analyzing the voluminous data generated from DNA microarrays. According to Nature, a single microarray experiment that examines 40,000 genes from 10 different samples, under 20 different conditions, produces at least eight million pieces of information. Microarrays provide a snapshot of gene activity for the entire genome by telling the investigator which genes are expressed for the condition under study. By comparing the genetic composition of a healthy individual and a patient diagnosed with cancer, for example, a scientist can pinpoint which genes may be involved in cancer development.

New insights into biology

In a typical microarray experiment, Dubitzky uses Clementine to cluster and classify genes that appear to underlie a particular tumor’s development, thereby providing a cancer “fingerprint” in a single experiment. By rapidly and correctly classifying the tumor, patients can be diagnosed and given proper therapy earlier, thus increasing their chances for survival.

After clustering, Dubitzky typically applies Clementine’s decision tree to highlight which parameters (i.e., which genes are over- or under-expressed) are characteristic for the clusters. These trees can offer biological insight not offered by other methods. For example, in differentiating between acute myeloid leukemia and acute lymphoblastic leukemia, Dubitzky discovered that the most important gene was one whose protein played a critical role in the transforming growth factor pathway. He also identified genes that were previously unknown to be related to tumor biology.

Future steps

One of Dubitzky’s future projects is to use Clementine to help model cell regulatory networks by identifying relationships between microarray data and information (i.e., gene function) already available in public databases. With data from the Human Genome Project available online and 800 other databases worldwide to integrate, Clementine’s future in bioinformatics looks to branch far into the future.

SPSS®