

> University of Ulster

Bioinformatics researcher finds biological gold

Situation

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. Indeed, many scientists would agree with Nobel Prize winner Paul Berg, who champions the notion that every disease has a genetic component.

Critical Issue

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. Analyzing biological problems *in silico* will become a powerful complement to studying them *in vivo* and *in vitro*.

Bioinformatics first attracted notice when it was used in the Human Genome Project to isolate precious gene sequences from mostly junk DNA. 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 each other. Dr. Werner Dubitzky, professor of bioinformatics at the University of Ulster in Northern Ireland and member of the editorial boards 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.

At-a-glance

Country: United Kingdom
Industry: Scientific research
Date Founded: 1966
Company Type: Public
Students: 21,200 (2001)

Application

Microarray gene
expression research

Solutions Used

Clementine®



Solution

Dubitzky relies on cutting-edge commercial 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 the journal *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, for example, the genetic composition of a healthy individual and a patient diagnosed with cancer, a scientist can pinpoint which genes may be involved in cancer development.

Results

- Clusters and classifies genes more quickly and accurately
- Provides greater biological insight

Clusters and classifies genes more quickly and accurately

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. This is quicker and more discerning than current non-molecular methods in which tumors are classified by their appearance and clinical course. By rapidly and correctly classifying the tumor, physicians can diagnose patients earlier and provide the proper therapy, thus increasing their chances for survival.

Provides greater biological insight

After clustering, Dubitzky typically applies Clementine's decision tree to highlight which parameters (i.e., which genes are over- or underexpressed) are characteristic for the clusters. Compared to other decision tree software programs that are limited to two branches per node, Clementine resolves genetic relationships to a higher power by generating trees with varying numbers of branches per node. These trees can offer biological insight not attained by other methods. In

differentiating between acute myeloid leukemia and acute lymphoblastic leukemia, for example, 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 in 2003, and 800 other databases worldwide to integrate, data mining's role in bioinformatics looks to branch far into the future.

□ "I like Clementine because of its ability to quickly generate results and its intuitive user interface."

– Dr. Werner Dubitzky
Professor of Bioinformatics
University of Ulster

To learn more, please visit www.spss.com. For SPSS office locations and telephone numbers, go to www.spss.com/worldwide.

SPSS is a registered trademark and the other SPSS products named are trademarks of SPSS Inc. All other names are trademarks of their respective owners. © 2002 SPSS Inc. UUCS-0105

For more customer stories, visit www.spss.com/success.

