Data Mining Technologies Inc.

Knowledge Discovery From Databases

Data Mining for Genetics Research

White Paper

Data Mining Technologies Inc.
Melville, NY 11714
(631) 692-4400 x100
e-mail mgilman@data-mine.com
www.data-mine.com

April 2006

Data Mining Technologies Inc.
Knowledge Discovery From Databases

.

Nuggets® for Biogenetic Research

Science has progressed to the point where new technologies can be harnessed to analyze and study proteins and genes and ultimately to understand their behavior.

Nuggets® offers the potential to provide new insight in the field of Bioinformatics and drug research with a new technique which can search for patterns in the DNA sequence and related data which will assist pharmaceutical companies and genetic researchers in finding new cures for diseases, uses of proteins expressed by genes, and gene functionality.

Bioinformatics is defined as the mathematical, statistical and computing methods that attempt to solve biological problems using DNA and amino acid sequences and related information. The greatest achievement of Bioinformatics presently is the Human Genome Project. But the surface has only been scratched. Now that we have the blueprint the question is what does it mean. This opens open the next phase of the quest and offers opportunities for new analytical methods to be used. The questions are deep and the water is murky. Existing statistical methods have been applied often with less success than hoped for. Many ad hoc methods have been developed with varying degrees of effectiveness. The field is wide open and ready for new ideas.
There are two major divisions in the study to be pursued. One is morphological, i.e. the study of form and shape and how that relates to cell and gene functionality. The other is analyzing the informational structure of the genes and the DNA molecule as a data processing topic. In a broad sense it can be viewed as the application of computer technology to the management of biological information. Specifically, it is the science of developing computer databases and algorithms to facilitate and expedite biological research. Bioinformatics is being used largely in the field of human genome research by the Human Genome Project that has been determining the sequence of the entire human genome (about 3 billion base pairs) and is essential in using genomic information to understand diseases. It is also used largely for the identification of new molecular targets for drug discovery.

Fields of Study in Biogenetics
Scientists today have uncovered multiple whole genomes and can look for differences and similarities between all the genes of multiple species. Thus, scientists can draw particular conclusions about species and general ones about evolution. This is known as the study of comparative genomics. Also, technologies exist to measure the relative number of copies of a genetic message (levels of gene expression) at different stages in development or disease or in different tissues. These technologies are known as DNA microarrays. Other, more direct and larger scale methods of identifying gene functions and associations will grow in significance; this is the field of functional genomics. Then, a shift of emphasis of sequence analysis from genes to gene products will lead to scientific attempts to catalog the activities and characteristics of interactions between all gene products. This area is known as proteomics. And finally, scientists are also attempting to crystallize and predict the structures of proteins in the field of structural genomics.

How can Nuggets® contribute to the fields of biogenetic research? Firstly, taking the study of Proteomics as an example, the following is a step-by-step overview of the analytical processes that involves to the use of Bioinformatics.

An Example: Proteomics

Proteomics is the study of proteins. Genomics is the study of DNA and the processes that will lead to the creation of proteins. Combined, both studies will lead researchers to better analyze and understand gene expressions. When cells receive signals such as a growth factor, their response is normally at the protein level. In this example, cells will up-regulate genes needed for cell division. Technologies have been developed to monitor genome-wide transcription responses of cells at the mRNA level. One method of protein analysis using 2D gel electrophoresis is performed to extract samples based on molecular weight and charge. Using gels to deliver reproducible results, proteins are extracted from each culture samples. Then, the gels are fixed and stained with a fluorescent type dye and scanned. Expression levels are measured by the size of each feature on the gel. At this stage, Bioinformatics software is used to find differences between a control sample and a treated sample. This will provide information about the abundance of proteins that were up or down-regulated. This will provide valuable information on which proteins that should be studied further.

The isolated proteins are extracted from the gels and treated with enzymes to cut amino acid chains into shorter polypeptides that are about 5 to 10 amino acids in length. These fragments are separated by a capillary electrophoresis process and analyzed using a method called rapid-throughput mass spectrometry to determine the sequence of polypeptide fragments, their mass and post-translation modifications. These amino acid sequences are then compared with an existing sequence database to find matches.

The study of proteins also includes areas like protein-to-protein interactions, protein identification and quantification, protein function and annotation, and applications and drug discovery.

The interesting property of most large biological molecules is that they are polymers or ordered chains of simpler molecular modules called monomers. They all have the same “thickness” and way of connecting to each other. Each monomer molecule belongs to the same general class though they each have their own well-defined set of characteristics. Combined to form macromolecules, they will have different, specific properties. According to this scheme, the monomers in any given macromolecule of a DNA or a protein chain can be treated computationally as alphabetical letters put together in pre-programmed arrangements to carry messages or do work in cells.

The challenge

With the large amount of sequence and expression data available through the new analytical tools, scientists are faced with the challenge to sift through the data and decipher biological information, and to look for clues as to the function of cellular networks and underlying mechanisms in biology. Statisticians and scientists are constantly coming up with algorithms for mining gene expression data, and using large computer systems in helping them in this task.

Bioinformatics problems can involve complex interactions among thousands of variables. Most methods cannot deals with problems of this size. For example, the human genome contains about 30,000 different genes that interact in very complex non-linear ways. In addition, the methods used to analyze the genome sequence inherit statistical shortcomings such as requirements for linearity, independence and underlying assumptions about statistical distribution functions that are hard to prove. A new crop of homegrown methods has evolved to mitigate these

shortcomings. These methods often don't work well in practice or are limited to special classes of problems.

The Nuggets® Advantage

In tests of real world data such as that found on the University of California Irvine website http://www.ics.uci.edu/~mlearn/MLRepository.html) Nuggets® Enterprise on an Intel based PC and found that it has performed extremely well – it has shown results faster and with higher accuracy compared to competing methods such as Neural Networks and statistics and ad-hoc methods.

Nuggets® uses a proprietary methodology enabled by SiftAgent™, an artificial intelligence based data mining system developed by Data Mining Technologies. Using rule induction, Nuggets® has the ability to overcome difficulties associated with traditional methods of data mining.

Nuggets® has the ability to analyze thousands of variables, a feature not found in many data mining technologies available in the market, even in expensive software on large-scale mainframe computers. This is especially important in the field of biogenetics research because of the requirement to sift through enormous amounts of sequence data that have many variables with complex non-linear relationships and interactions. In addition, it can analyze a database automatically, formulating the questions and answering the questions by itself which can avoid the issue of determining what questions to ask of the data.

Inherent Parallelism of the Nuggets® Architecture

One of the exciting features of Nuggets® is that it is inherently parallel. What this means is that it can be loaded onto multiple Intel based servers to achieve gains in computational output almost linearly. This allows for significant advantages over other data mining solutions. A conceptual design of the architecture would look like the one as pictured in figure 1.0. The administrator console will have control over the number of servers connected via a TCP/IP network. The console is used to issue commands to run and stop the analytical processes, and to generate reports while the servers can work individually. Conceptually, the more servers plugged into the configuration, the more powerful the system becomes.

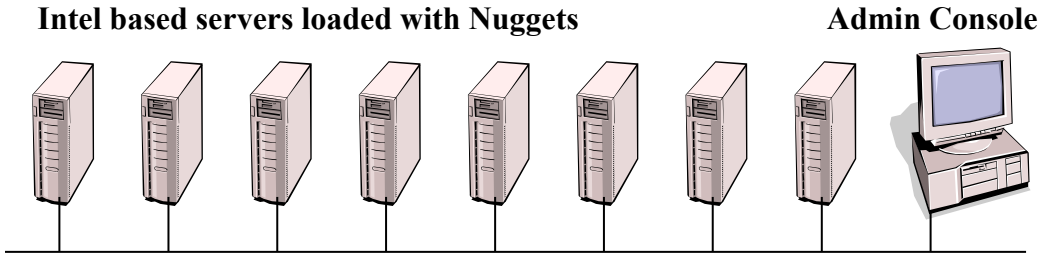

Figure 1.0

The Next Step

Data Mining Technologies intends to seek a hardware platform company as a partner to combine our abilities to produce a unique data mining based Bioinformatics system offering to biogenetics companies. The next step will is to seek a biogenetics company which would be interested in providing domain expertise in developing suitable project targets for proof-of-concept.