

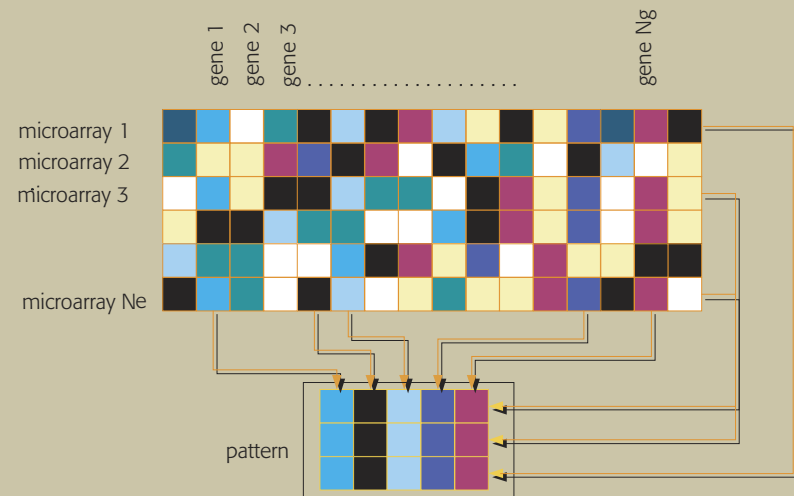
IBM Research has been exploring the field of computational biology for over twenty years. Following early work investigating quantum chemistry, more organized activity emerged with adaptations of combinatorial geometric hashing techniques to searching gene and protein sequences. One of the results of this effort was the identification of intrinsic coordinate systems for comparing moments of charge distributions, which were applied to the problem of similarity searches of molecular structures. Next, pattern and association discovery techniques were applied to annotation and functional analysis, gene identification, and gene array time-series analysis. Recognition of the increasing importance of large-scale computations in biology resulted in the commitment to develop a scalable architecture that could solve the most complex, computationally intensive scientific problems. This led to the development of Blue Gene, which has been the fastest supercomputing platform from early in its development.

IBM's broad research agenda includes pattern discovery, systems biology, computational neuroscience, protein folding, clinical informatics, and text mining. Recently, pattern discovery has provided insight into antimicrobial peptides and RNA interference. Techniques for identifying RNA precursors are leading to new ways of probing junk DNA and finding connections between junk DNA and coding DNA. Systems biology research is addressing statistical problems in gene expression array data, such as characterizing biological and instrument noise, mining high-throughput genomic and proteomic data to derive meaningful biological inferences, and techniques for reverse engineering, network analysis, and modeling of complex biological systems. In the area of protein structures and protein folding, researchers seek to understand the physics of the energy landscape of proteins and the kinetics of the protein-folding process. By developing fast algorithms for biomolecular simulations, which take advantage of the scale of computing enabled by Blue Gene, these simulation techniques are applied to the study of biological processes at the molecular level, including, for example, the structure and function of membrane proteins.

**BIOINFORMATICS AND MEDICAL INFORMATICS**

Bioinformatics and medical informatics have been a part of a revolution within biomedical research. Recently, application to the clinical domain is being driven by economic incentives, by changes in readily available information, ranging from universal availability of digital patient records, enabling the mining of clinical information, to benefits obtained by discovering new disease mechanisms and relationships, and by clinical decision intelligence tools that aid in designing effective treatments.

IBM's Genes@Work algorithm detects patterns of gene expression in DNA arrays. If the number of patterns in the data greatly exceeds the average number of patterns expected in random data (formula), then there may be an underlying biological effect.



$$N_{jk}(j,k,N_e,N_g,\delta) \sim \binom{N_g}{k} \binom{N_e}{j} \alpha^k (1-\alpha)^{N_g-k} [1-(1+\alpha)^{-1}]^k \delta^k N_e^{-j}$$

$$\alpha = j\delta^{j-1} - (j-1)\delta^j \quad p = 1 - \exp\{-N_{jk}\}$$

**DATA MINING**

Data mining, such as association discovery, inference rule and redescription mining, and information measure techniques have been applied to clinical data obtained through collaboration with research hospitals around the world. Text mining regularizes clinical records filled with diverse nomenclatures describing the same disease processes and treatments, and identifies connections in the clinical literature. These areas strike at the hardest barriers to the promised benefits of clinical decision intelligence.

**THE GENOGRAPHIC PROJECT**

One of IBM's most exciting challenges is the partnership with the National Geographic Society on the Genographic Project, a global research exploration to trace the migratory history of the human species. By looking at the genetic markers in the human population — data gathered from hundreds of thousands of participants — combined with knowledge from archaeology, anthropology, linguistics, etc., the project aims to develop a more comprehensive understanding of how the earth was populated as well as how the

connections and differences of the human species developed. IBM is participating in the development and validation of new techniques to analyze the vast volume of data that is being collected worldwide in this unprecedented genetic anthropology initiative.

