PANEL

Data and Knowledge Bases for Genome Mapping: What Lies Ahead?

Panel Chair:

Nabil Kamel, Database Systems Research and Development Center (USA)

Panel Members:

M. Delobel, INRIA (France)
Thomas Marr, Cold Spring Harbor (USA)
Robert Robbins, NSF (USA)
Jean Thierry-Mieg, CNRS (France)
Akira Tsugita, JIPID, Science University of Tokyo, CODATA (Japan)

DNA is a long biopolymer that is found in the nucleus of every living cell. The DNA of a cell contains the genetic information that determines every physical aspect of an individual. This information is encoded in the sequence pattern of four different heterocyclic molecules called nucleotides, which are the building blocks of the DNA polymer. Such traits as sex, physical appearance, and details of metabolic function are determined by the sequence of these four nucleotides (adenine, guanine, cytosine, and thymidine) in an individual's DNA.

The Human genome consists of about 3 billion nucleotides arranged in 23 individual linear DNA molecules, the "chromosomes". As many as 90% of these 3 billion nucleotides may be non-informational and do not apparently encode any significant function or trait. Those DNA segments which do encode a meaningful function are known as genes. It is believed that humans have from 50k-100k genes ranging in length from less than 1000 nucleotides to over 200 million nucleotides. A class of laboratory techniques referred to as DNA sequencing methods is used to determine the exact order of the nucleotides in a DNA molecule, thus enabling the genetic information contained in the DNA to be deciphered. The $3 billion human genome initiative aims to map and sequence the entire human genome. The project will have to rely heavily on advanced data and knowledge base technology to efficiently perform its task. Furthermore, the ability to understand the biological data resulting from this project and to effectively make use of it will also have to rely on advances in current data and knowledge base technologies. Examples of such uses include rational drug design, introduction of new genetic therapy procedures, or even the design of new organisms.

The panel had its main focus on three main issues that are critical to genetic databases: heterogeneous database issues, map construction by sequence assembly, and database standards. Heterogeneous database issues are very important as the molecular biological databases being built and used for research in this area are distributed around the globe. Map construction involves a variety of theorem proving and knowledge deduction techniques, and finally, the issue of database standards needs to be addressed both as a distinct question and as it relates to other heterogeneous database issues. In summary, the panel had its primary focus on data and knowledge base requirements in support of the Human Genome Project and the ensueing use and integration of the expected information. The discussions emphasized the importance of addressing the heterogeneous database issues, knowledge base issues, and the question of database standards.