INTRODUCTION TO THE SESSION: "DISTRIBUTED AND INTELLIGENT DATABASES"

D. FRISHMAN
Munich Information Center for Protein Sequences/GSF,
Am Klopferspitz 18a, 82152 Martinsried, Germany
frishman@mips.biochem.mpg.de

"Come and take choice of all my library
And so beguile thy sorrow"
–William Shakespeare, "Titus Andronicus"

It is very characteristic for our time that so many research projects in biology result not only in publications, but also in creation of databases representing the results in a systematic and formalized way. Many subject areas are covered by more than one databank. While primary protein and nucleic sequence databases come in just two different flavors, at least 10 protein pattern and domain collections, several aggregations of DNA functional sites, nearly a dozen of protein structure-related compendia, and more than 30 highly specialized datasets of mutations and genetic disorders are now accessible. To make the chaos complete, genomic databases appear at an astonishing rate. Nine complete bacterial and one eukaryotic genomic sequences have been publicly released so far; another 70 genomes are on the waiting list. Associated with each of them are archives of functional assignments, metabolic pathways, genetic elements, etc. Finally, the massive amounts of data generated by the human genome project have the potential to overshadow all currently available databanks.

With so many heterogeneous data sources at hand, connectivity is of utter importance, and will be the main focus of this session. It is quite indicative that all contributions deal with different aspects of database interconnection. The DBGET/LinkDB system presented by Fujibuchi and co-workers allows the linkage biological databases distributed over a network. Both WWW and command-line user interfaces are provided. The system is open, and new databases can be easily included. LinkDb maintains not only the links provided in each database, but also computes reverse and indirect links. Simple indexing mechanisms allow for easy and frequent updates.

Schulze-Kremer describes ontologies suitable for knowledge sharing in molecular biology. He defines the representation problem in terms of "is a subset of" and "is a member of" relationships, thus stressing the role of interconnection on the conceptual level. His work also outlines the need for better communication between molecular biology, computer science, and linguistics.

The necessity of a common language in molecular biology is clearly seen in the work of Fukuda and his colleagues who address the problem of automatic extraction of protein sequence-related data from the literature databases. The flood of newly introduced words makes it impossible to rely on predefined dictionaries. The authors propose a set of rules based on context analysis and the knowledge of general tendencies in naming new objects to allow for very reliable protein name extraction.

Wu and Shivakumar developed a database of protein families (PROCLASS) that merges and adds value to the PIR superfamily classification and the ProSite collection of motifs and patterns. Consideration of both global relationships and local similarities limited to a specific domain or functional site allows to take into account domain organization of proteins. The authors suggest that PROCLASS could be an ideal tool for functional characterization of newly determined genomic sequences.