SESSION TITLE: Distributed and Intelligent Databases

P. ARGOS
European Molecular Biology Laboratory,
Mayerhofstr. 1, 69012 Heidelberg, Germany

H.-W. MEWES, D. FRISHMAN
Martinsried Institute for Protein Sequences,
Max-Plank-Institute for Biochemistry
Am Klopferspitz 18a, 82152 Martinsried, Germany

Presently molecular biological databases are rapidly increasing both in number and information content; however, their interpretation is rapidly becoming a bottleneck in molecular biology. An outstanding example is afforded by amino acid sequence collections and their doubling every few years with entries now exceeding 100 000; and yet there is no comprehensive database multiply aligning them as families with conserved motifs. The information in the banks has not been integrated since the originators (often scientists in a given field) have focused on their specific specialty such as tertiary protein structures or only immunoglobulin sequences. Mere collection was sufficient when the information content was relatively small, the opposite of the present status given the dominance of genome sequencing projects and commercial pharmaceutical biotechnology. Large databases need classification, interpretation, and analysis. Furthermore, they need synthesis for optimal use of information necessary for new discoveries of peviously unknown relationships. The integration required is extensive and should span across genetic mapping and sequence data, evolutionary development, protein structure and function, and cellular processes and organization.

The session will focus on recent advances in delivering high quality information to the biological community. New approaches include interconnecting existing databases distributed over many computer sites for simultaneous information retrieval through Internet tools, knowledge-based expert systems, interface languages, and multiserver management as well as methods that make the primary databases more intelligent and yield newly synthesized and interpreted databases, exemplified by careful clustering of the available data, incorporation of experimental data, functional and structural characterization of whole genomes using novel techniques such as structure prediction, threading, and pattern recognition. It is clear that major discoveries in molecular biology will be severely inhibited without automatic access and association of available information.