Some Challenges in Integrating Information on Protein Interactions and a Partial Solution

Professor H. V. Jagadish
University of Michigan

Independently constructed sources of (scientific) data frequently have overlapping, and sometimes contradictory, information content. Current methods of use fall into two categories: force the integration step onto the user, or merely collate the data, at most transforming it into a common format. The first method places an undue burden on the user to fit all of the jigsaw puzzle pieces together. The second leads to redundancy and possible inconsistency. We propose a third: deep data integration. The idea is to provide a cohesive view of all information currently available for a protein, interaction, or other object of scientific interest. Doing so requires that multiple pieces of data about the object, in different sources, first be identified as referring to the same object, if required through "third party" information; then that a single "record" be created comprising the union of the information in multiple matched records, keeping track of differences where these occur; and finally by tracking the provenance of every value in the dataset so scientists can judge what items to use, and how to resolve differences.

In this talk, I will describe our experiences with this approach in MiMI (http://mimi.ncibi.org). I will also discuss barriers to domain scientist use of the system, and my thoughts regarding how to make systems truly "usable".

H. V. Jagadish is a Professor of Computer Science and Engineering at the University of Michigan in Ann Arbor. After earning his PhD from Stanford in 1985, he spent over a decade at AT&T Bell Laboratories in Murray Hill, N.J., eventually becoming head of AT&T Labs database research department at the Shannon Laboratory in Florham Park, N.J. He has also served as a Professor at the University of Illinois in Urbana-Champaign and as the Shaw Visiting Professor at the National University of Singapore. Professor Jagadish is well-known for his broad-ranging research on information management, and has over 150 major papers and 33 patents. He is a fellow of the ACM ("The First Society in Computing") and a trustee of the VLDB (Very Large DataBase foundation). Among many professional positions he has held, he has previously been an Associate Editor for the ACM Transactions on Database Systems (1992-1995), Program Chair of the ACM SIGMOD annual conference (1996), and Program Chair of the ISMB conference (2005).

19th International Conference on Scientific and Statistical Database Management (SSDBM 2007)
0-7695-2868-6/07 $25.00