OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003

Effective Integration of Protein Data through Better Data Modeling

ADRIANE CHAPMAN, CONG YU, and H.V. JAGADISH

ABSTRACT

Protein data, from sequence and structure to interaction, is being generated through many diverse methodologies; it is stored and reported in numerous forms and multiple places. The magnitude of the data limits researchers abilities to utilize all information generated. Effective integration of protein data can be accomplished through better data modeling. We demonstrate this through the MIPD project.

INTRODUCTION

THERE IS A PROLIFERATION of data sources in biology. Each research group and each new experimental technique seems to generate yet another source of valuable data. This data is not represented in any standard format. Usually it is not possible to define a tightly-specified standard format that is general enough to anticipate the needs of these new data sources. Even when open standards such as XML are used to represent data, they are frequently in the form of customized, source-specific, schemas. Moreover, schemas themselves change frequently, as knowledge in the field evolves, and new attributes are found to be of importance. Researchers relying on the integration of data from multiple such sources need help.

Even researchers conducting experiments, and therefore quite likely interested in a comparatively limited class of sources of data, need help. Since experiments are expensive to conduct, reuse of data is desirable whenever possible, for instance by patching together information derived from multiple previous experiments conducted for possibly different purposes. Effectively performing such data integration requires good metadata annotation with respect to experimental conditions and similar other information for each data set in question. However, such annotations are frequently missing. Even when present, they are frequently incomplete and never standardized.

Some standards for meta-data specification are beginning to emerge. For instance, MESH is used widely to annotate medical literature, and UMLS has been proposed as the next step beyond it. Drug ontologies have been developed based on chemical components and on functional characterization. While development of a standardized domain-specific ontology is of value, there is much information that such ontologies are not likely to capture. For instance, details of the experimental conditions, possibly considered trivial at the time of the experiment itself, may turn out to be crucial at a later time (Foster, 2002). No ontology is likely to have a priori captured such detail.

In addition to metadata regarding the environment, the experiment, and so forth, there is also considerable local metadata that could be associated with individual data items (or sets of data items). For instance, scientists may often wish to annotate specific readings, by way of explanation, or to record an insight not

Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan.

evident from just the numbers. Similarly, data can be of variable quality, due to experimental error of various sorts, and also because science progresses by advancing hypotheses not all of which are eventually substantiated. We should provide facilities to maintain data provenance to enable tracing the derivation of each item in a database (Buneman, 2002). We should also keep track of reliability quantitatively, through the association of probabilities, and similar other quantitative expressions.

At the University of Michigan, we have been studying these issues, and currently have partial solutions in place, based on our Timber XML data management project (Jagadish, 2002). Specifically, we are able to capture quantitative and qualitative reliability information associated with facts at any granularity (Nierman, 2002). We are also able to represent the experimental technique used to obtain the data, along with relevant environmental factors that may be important in future interpretation of the data.

Using the above as a basis, we have begun to address the problem of integrating the large amount of web accessible data available to the biological enterprise, focusing specifically on protein interaction data (Bader, 2000). We find that there is significant overlap in content among sources as well as innumerable links connecting the source contents to each other. We are developing new data representation and integration techniques that permit effective integrated representation of such disparate overlapping data, along with all of the environmental and reliability annotations mentioned above (MIPD).

ACKNOWLEDGMENTS

This research was supported in part by NSF under grant IIS-0208852 and by a Bioinformatics Pilot grant from Pfizer and from Howard Hughes.

REFERENCES

BADER, G., and HOGUE, C. (2000). BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics **16,** 465–477.

BUNEMAN, P., KHANNA, S., TAJIMA, K., et al. (2002). Archiving scientific data. Presented at the ACM SIGMOD, Madison, Wisconsin.

FOSTER, I., VÖCKLER, J., WILDE, M., et al. (2002). Chimera: a virtual data systems for representing, querying, and automating data derivation. Presented at the 14th International Conference on Scientific and Statistical Database Management, Edinburgh, Scotland.

JAGADISH, H.V., AL-KALIFA, S., CHAPMAN, A., et al. (2002). TIMBER: a native XML database. VLDB Journal **11,** 274–291.

MIPD (Michigan Protein Database). Available: www.eecs.umich.edu/db/mipd.

NIERMAN, A., and JAGADISH, H.V. (2002). ProTDB: probabilistic data in XML. Presented at the 28th VLDB Conference, Hong Kong, China.

Address reprint requests to:
Dr. H.V. Jagadish
Department of Electrical Engineering and Computer Science
University of Michigan
1301 Beal Avenue
Ann Arbor, MI 48108-2122

E-mail: jag@umich.edu