XML Access for Proteomics Data

Philip Andrews[†] **Jason Atkins**[§] **H. V. Jagadish**[§*]
Jennifer McCormick[†] **Peter Ulintz**[†]

1 Preamble

Philip Andrews is a leading researcher in Proteomics. Over the past several years his laboratory at the University of Michigan Medical School has played a lead role in the characterization of the E. Coli proteome. The Prime (Proteome Research Information Management Environment) database system has been developed by this group and is currently one of many proteomics information sources available on the web.

H. V. Jagadish is a leading researcher in Databases. Over the past few years he has focused on the study of hierarchical information structures, with a specific focus on XML, and their use for the representation and access of heterogenous and autonomous information sources.

Almost a year ago, Professors Andrews and Jagadish initiated a collaboration in the area of Bioinformatics. This document is a brief description of the some of the thinking behind this joint project.

2 Broad Agenda

There are many autonomous and heterogeneous sources of data on any topic imaginable, and proteomics is no exception. Whereas there is considerable talk about centralized repositories, and there even is one dominant repository for some types of proteomic information (http://www.expasy.ch/sprot/), most scientists find it necessary to consult a number of different web sources to carry out their research. The idiosyncratic nature of these web sites makes it arduous for a human to perform this integration task, even when there are cross-links available.

Given the syntactic commonality provided by XML, the question is whether we can develop techniques that let individual sites retain their autonomy and yet permit an automated pulling together of information relevant to any given search criterion. The techniques necessary for doing this for the proteomics domain are likely to be similar to heterogeneous data querying techniques in other application domains. However, a few domain-specific twists are likely to be important, such as support for error annotation.

* Contact author and proposed workshop attendee. jag@eecs.umich.edu
§ Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor.
† Biological Chemistry Department, University of Michigan, Ann Arbor.

3 Current Status

3.1 System Development

We have developed an XML schema suitable for representing the data in the Prime repository. We have ported some of the data from Prime into our XML store, and intend to have all of Prime in XML in the future. We have also begun to develop XML schema wrappers for a few external sites of interest, with the notion that queries from our site can address information in foreign sites, with the wrappers doing the needed translation and remote server access.

3.2 Research

A central task for biological chemists is protein identification, namely placing tags on (visually dark) spots in a gel image. This task is difficult and error-prone, and often involves some educated guess-work. Additional assays can be run on selected spots, using mass spectrometers for example. However, these additional tests are expensive and laborious, and hence not routinely performed on every spot. The net result is a set of protein identifications that are not quite definite, and that go through a "data curation" procedure to ensure data quality before being made public.

We are currently studying this identification process as a central data modeling issue. What happens if two (or more) proteins overlap at a single spot on the gel? How can one record and disseminate probabilistic data – can it be useful to get information with incompletely (and possibly even incorrectly) identified proteins in some gel? What is the basis for developing such probability measures in the first place? What are reasonable error models – for instance we may wish to permit a fairly narrow Gaussian distribution around the expected molecular weight, but also have non-zero probability at multiples of this weight to allow for dimerization/polymerization. How can one represent all of this information in an XML context? (Our current solution is to use attributes as meta-data). How can one pose, and efficiently evaluate, aggregate queries to determine various frequencies central to estimating probability? How does all of this change as the science is better able to perform protein function identification rather than just structure identification?

We are actively investigating questions such as these, and at present have (partial) answers to at least some of them.