

Enabling Integration of Biological Datasets

Xuan Zhang* Kaushik Sinha* Ruoming Jin† Gagan Agrawal*

*Department of Computer Science and Engineering
Ohio State University, Columbus OH 43210

{zhangx,sinhak,agrawal}@cse.ohio-state.edu

† Department of Computer Science
Kent State University, Kent OH 44242

{jin}@cs.kent.edu

Bioinformatics data is growing at a phenomenal rate. Besides the exponential growth of individual databases, the number of data depositories is increasing as well. The number of database entries in DBCat, a catalog of biological data sources, reached 500 in year 2000 and continues to grow. Because of the complexity of biological concepts, bioinformatics data usually has complex data structures, and therefore cannot be easily captured with the relational model. As a result, various flat-file formats have been used. Out of 111 biological databases studied by Kroger in 2003, 36% to 40% are implemented as flat-files collections.

While flat-file representations are easy for human interpretation, there are no established standards for their layouts. As a result, manually written parsers are widely used to extract data from them. This has limited the readiness of the data for data consuming programs, such as integration systems. Integration of data sources has become a critical issue in biological research in recent years, as it allows biologists to combine knowledge from multiple disciplines. Biological data is generally maintained heterogeneously and updated in an unsynchronized manner. Thus, whenever one data source changes its format, an integration system using manual wrappers has to be updated accordingly. This has become an extremely time-consuming and error-prone task. Furthermore, when a new data with a flat-file representation is found, it cannot be incorporated into the integration system until a wrapper is written specifically for it. This again requires considerable programming effort, and is a often the detriment in incorporating data from a new source.

We have designed a three-step approach for improving the accessibility of data in flat-file formats and reducing the human involvement in integration process. The list of steps involved is also shown through Figure 1.

The layout of a flat-file bioinformatics dataset is first learned semi-automatically in two steps. The first step is to infer the delimiters used by the dataset using a metric we have defined, d_score . This is based on the information of frequency and position of token sequences and provides a superset of the actual delimiters. The incorrect delimiters can be prone manually by the users. In the second step, the layout descriptor of the dataset is generated from the correct delimiter set, based on the relative order of the delimiters.

From the layout descriptor, a parser is created automatically. This process was based on our previous work on the wrapper generation system.

To integrate the dataset parsed, or to use it as part of an available data processing tool, we need to assign a label or name to every attribute. This process is called *schema mining*, and comprises the rest of the steps in Figure 1.

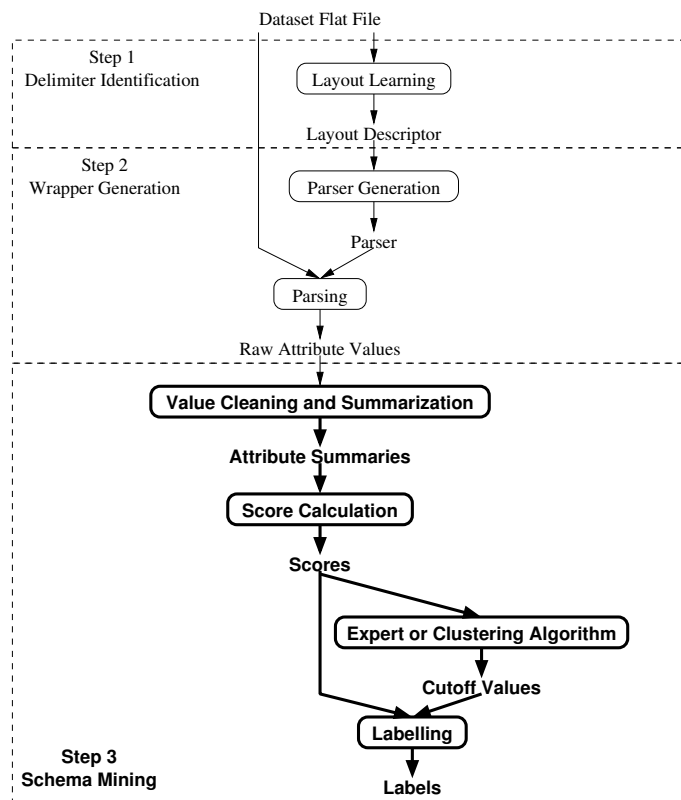


Figure 1: Overview of Our Approach for Biological Data Integration