Data/Knowledge Bases for Biological Papers and Techniques

February 15, 1993

1 Project Information

This material is based upon work supported by the National Science Foundation under Grant No. IRI-9117030.
Principal Investigator: Kenneth Baclawski
Co-Principal Investigators: Robert Futrelle, Carole Hafner and Maurice J. Pescitelli
Students: Natalya Fridman, Beixing Li and Chendong Zou
Location: Northeastern University, College of Computer Science, Boston, Massachusetts
Period and Funding Level: August 1, 1992 to July 31, 1995; $680,578.

2 Introduction.

The point of the project is to find ways to use database management techniques to support biological research. While biology is a very large and diverse field, the primary output of the enterprise is its research literature. This literature consists of about 600,000 papers every year. Our objective is to develop database and text analysis techniques for making this literature more accessible. The goal is to be able to analyze, store and query research papers electronically.

The vast majority of these papers report experimental work and do so in a highly structured manner. There are now long-term AI-based efforts in progress to produce full knowledge bases from these papers[Fut89]. In the shorter term, there is important information which can be captured and stored in databases. Moreover, local laboratory versions of the database software can be developed. When integrated with other tools this software can provide working scientists with a laboratory notebook software to assist them in their everyday activities.

The purpose of this project is to extend current database and information retrieval technology so as to represent more of the content of research papers. Knowledge about biological research papers will be incorporated by a combination of data structures and query processing. The data structures are developed in a top-down fashion and build on the structures already in the text such as sections, paragraphs, figures, etc.

The first part of our project will focus on the text in the Materials & Methods sections of biological research papers. There are a number of reasons why we chose this particular section.

1. The section is an important one. Biology is a technique-driven rather than a theory-driven experimental science.

2. This section is easier to analyze than the rest of the paper. It has a more limited (although still very large) vocabulary. The language used is more stylized (although still complicated). The other parts of the paper use more complex syntactic and linguistic forms.

We propose to develop a database for the Materials & Methods sections of biological research papers. Initially, we will just be dealing with papers in the field of bacterial chemotaxis. But the system will be designed to be extendible to other branches of molecular biology. This database would then provide a prototype and testing ground for databases that can encompass more of the content of papers.

This report will begin with a small example of the kind of text that occurs in biological research papers. We then show how this text will be represented as data in a database. The structure of the database is then described along with some examples of the kind of queries that will be supported. The database and the query processor will incorporate domain knowledge that can be regarded as "value added" to the research papers. Some of the research challenges that we face are given next. We intend to prototype and test actual systems. In addition to the database itself, we will be developing a laboratory notebook software, and we discuss some of the features of the proposed system. Finally we end with a summary of some of the themes that run through our project.

3 Example of Text from a Research Paper

The following is a typical example of the kind of text that appears in biological research papers.

> **Immunoaffinity chromatography**. IgG was purified from mouse ascites fluid by DEAE-Affi-Gel Blue (Bio-Rad) chromatography (5) followed by precipitation in 50% ammonium sulfate at 0°C. Purified IgG (5 mg/ml) was dialized against 0.1M sodium bicarbonate, pH 8.5 ... [SS87]

This example comes from the Materials & Methods section of a paper dealing with bacterial chemotaxis, the study of how bacteria move in response to chemicals in their environment. Nearly every biological research paper has a Materials & Methods section. This section explains the materials, such as chemicals, kinds of bacteria, proteins and such that are used in the experiment. It also describes the techniques or methods that were employed.

The example is the beginning of a procedure called *immunoaffinity chromatography*. The first part of this procedure involves preparing a chemical called an antibody (IgG) that is used, in turn, to prepare the chromatography column.

Immunoaffinity Chromatography

Purify

input: mouse ascites fluid
output: IgG

Chromatography

type: DEAE-Affi-Gel Blue (Bio-Rad)
reference: (5)

Precipitate

medium: 50% ammonium sulfate
temperature: 0°C

Dialize

input: purified IgG
concentration: 5 mg/ml
antagonist: 0.1M sodium bicarbonate, pH 8.5

...

Figure 1: Example of Knowledge Frames for Materials & Methods Section

The example illustrates the high degree of elaboration that typically occurs in describing experimental procedures: the immunoaffinity chromatography procedure consists of a series of steps, beginning with purification of the antibody, and followed by its dialization. The purification itself consists of chromatography followed by precipitation, and so on. Each step can have its own parameters in addition to being elaborated into subsidiary procedures.

To illustrate how data will be structured in this database, we analyze the example text. In Figure 1, the text has been expanded into a kind of outline format. This format is called a "knowledge frame" representation. It shows that the overall procedure is an immunoaffinity chromatography procedure. Within this procedure there are many steps, with the first one being a purification step and the second being a dialize step. The purification step has two parameters: an input fluid and an output substance. The purification step is elaborated into two subsidiary steps, and so on.

Analyzing the text into this format is called *parsing*. It transforms text into data to be stored in the database. To parse the text, it is necessary to know what structures are possible for the text in the Materials & Methods section. Such a structural description is called a *schema*.

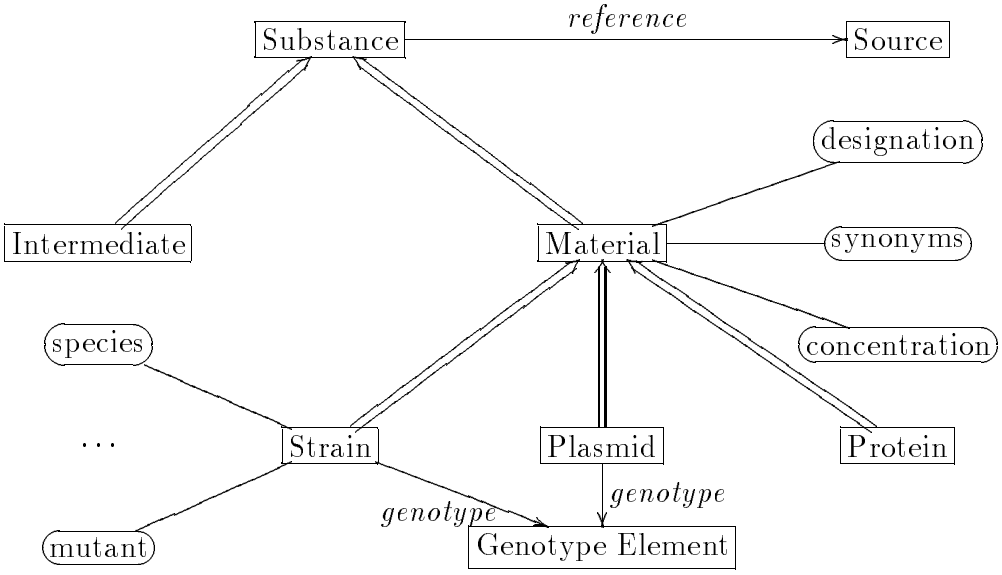

Figure 2: Part of the Structure of Materials & Methods Section

4 The Materials & Methods Database

The structure of the Materials & Methods Database can be described in many ways. Figure 2 gives a graphical representation. This shows only a small part of the entire schema of the proposed database. One can also describe the schema using ordinary English prose as follows:

```
The type substance is subclassified into intermediate and material, which
are disjoint.  A material may have one or more synonyms; may have a
concentration which is a number; must have a designation. It is uniquely
determined by the designation. The disjoint types protein, plasmid and
strain are derived from material.  A strain has a species, may have some
genotype elements as its genotype, ..., and may be a mutant.  A plasmid
may have some genotype element objects as its genotype.
```

The English prose given above was computer-generated. Using English descriptions allows us to communicate more easily with the end-users. Automatic generation ensures that all descriptions are consistent with each other at all times.

5 Queries

For the Materials & Methods database to be useful, it must support queries. The kinds of query that will be supported are closely related to the structure of the database. The following are some examples of the kinds of query that will be supported. These are actual queries taken from the Usenet newsgroup `bionet.molbio.methds-reagnts`.

- "Is there anyone out there who routinely cultures Lactococcus lactis? If so I was wondering if you could supply me with a recipe for the media used to grow this bug.[L.92]"
- "How do I amplify pBR322 plasmid using chloramphenicol?"
- "How do I use powdered silica to isolate DNA from agarose gels?"

Notice that the queries are in the form of a process step such as *culture*, *amplify* or *isolate* which has an input material (for example, pBR322) and possibly parameters (for example, chloramphenicol). A significant proportion of the queries in the newsgroup are of this form. Queries are structured in much the same way as the text in the database, except that query processing will involve making inferences involving domain knowledge about experimental techniques.

As new data structures are introduced, the query language will also be extended. The schema and query language will evolve together.

6 Challenges

There are many challenges in this project. Two of the most prominent are schema evolution and knowledge acquisition. These are both important research areas in databases and knowledge bases, but they have special significance in this particular case.

Biology is a very rapidly changing field. New materials and techniques are continually being introduced, and the understanding of existing materials and techniques keeps changing. Serving the needs of such a rapidly changing domain requires extending current design techniques. Because of the rapid pace of change, it is expected that the design phase will never end, and the end users of the system will always be involved in designing it. However, data already in the database must still be recognized after the design changes. Therefore, the system must support a form of schema evolution that allows for new schema versions while still supporting previous versions.

Another research challenge concerns natural language processing. Analyzing the text will require further development of natural language techniques. This will involve incorporating biological knowledge into the processing. The result will be a form of annotation of the text with "higher-level" or "value-added" information.

An alternative to analyzing the text is to capture this higher-level information at the source when the scientist is preparing a research paper. This could be done with a laboratory notebook software.

7 Laboratory Notebook Software

Laboratory notebook software could make it easy for scientists to submit Materials & Methods data to a central repository, thereby reducing the effort required to parse the English text in the paper. Such software could be integrated with traditional database software tools that

would support laboratory functions such as inventory and time tracking. Since the syntactic structure of the Materials & Methods section is relatively stylized, the laboratory notebook software could also be used to generate the text for this section of a paper. Ideally, the knowledge frames and the English would be generated simultaneously. By integrating these various functions and purposes, the proposed laboratory notebook software provide valuable services to working scientists on a daily basis.

8 Summary

To conclude, we discuss some of the slogans that characterize this project.

- "Everything is in English" The project is concerned with representing research papers as data in a database. The data will be highly structured English text. The description of the database structure is also in English prose, one can use the knowledge frames to generate text suitable for use in a research paper.
- "The system must evolve as biology does" The design of the schema will never end. But older versions of the schema cannot be ignored. Data in every version must be accepted.
- "The technology must scale up" Techniques that only work for small numbers of papers and a limited domain cannot be used. Performance will therefore be a issue.
- "The software must be integrated with other tools" Tools will be useful only if they are compatible with the day-to-day working habits of scientists. The two systems being developed by the project will be compatible with each other and with other tools, like word processors.
- "Systems should be prototyped and tested" Concepts will be tested by prototyping working systems and by having biologists use them.

References

[Fut89] R. Futrelle. An introduction to the Biological Knowledge Laboratory. Technical Report NU-CCS-89-15, Northeastern University, 1989.

[L.92] Mike L., 1992. Question posed to the Usenet newsgroup bionet.molbio.methds-reagnts.

[SS87] A. Stock and J. Stock. Purification and characterization of the CheZ protein of bacterial chemotaxis. *Journal of Bacteriology*, 1987.