Subject: Final Report

DOE Award No.: DE-FG02-93ER61718

Project title:

"Efficient, Intelligent Systems for Navigating the Biological Literature"

Principal Investigator: Robert P. Futrelle

Grant Period: 9/15/93 - 9/14/96

Date prepared: April 4, 1997

Organization:

Northeastern University
Division of Research Management
423 Lake Hall
360 Huntington Ave.
Boston, MA 02115

DOE/ER/61718--T1

*****

The biological literature is huge and increasingly moving to electronic form. By developing a variety of new techniques, it should be possible to take advantage of this huge and growing electronic store. Computers should allow us to use the literature with greater efficiency and insight to disseminate information and to advance scientific understanding.

Though there is a great deal of research and development effort focused on electronic text, e.g., the Digital Libraries initiative, little attention has been paid to the diagrammatic content of documents. However, it is common knowledge among biologists, and scientists in general, that the figures in documents are of critical importance. Little work has been done to develop principles and systems for analyzing, representing, and indexing and searching the diagrammatic content of electronic documents. This has been the main thrust of this research project. Because of our unwavering focus on this problem over the last ten years, and our determination to make progress, and with the help of a number of able students, we have made unparalleled strides on diagram understanding.

The primary work in the world on the analysis of graphics in documents has been focused on low-level issues relating to scanning legacy documents (hardcopy) and trying to discover the graphics elements in them. A smaller group, ourselves included, have looked toward the future, assuming that scanning hardcopies will fade in importance, and assumed that we should figure out how to deal with diagrams available in electronic form as graphics files. Graphics files, as opposed to image files, have lines, curves, polygons, text, etc., represented as discrete objects, as they are originally generated in drawing and graphing applications. This has been our focus, the starting point for all our analysis.

Using our Diagram Understanding System we have been able to automatically analyze (parse) a score of complex data graphs and gene diagrams, accomplishing something that no other research group (or commercial product) has been able to achieve. These diagrams were ones drawn directly from published papers, not ones we made up. The success depended on a two-fold strategy:

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MASTER

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

**Portions of this document may be illegible
electronic image products. Images are
produced from the best available original
document.**

- Using constraints over graphic object collections as the computational paradigm instead of the conventional (string-based) parsing paradigm.
- First writing software that succeeded in analyzing complex diagrams and only then abstracting away to "discover" the overall software architecture. All regularities in the successful code were then pulled together so that they became uniform methods with uniform notation. This avoided imposing an inappropriate solution style from above.

The results have been published and additional papers are in preparation. The papers will be submitted to various journals and meetings, including *J. Computational Biology*.

One remarkable result of our effort is that an entire diagram can now be parsed in a few seconds, using our techniques and extrapolating to the newest, soon-to-be-available processors, e.g., the 533MHz PPC chip from Exponential. In principle then, our techniques are capable of analyzing every one of the several million different diagrams that are published in the literature each year, and doing this as fast as they are published. Doing this in practice would require having the diagrams in graphic format and having all the requisite grammars written to cover a wide variety of diagram types. But still, the accomplishment is heartening.

I used funds from the grant to partially support my sabbatical visits to the University of Edinburgh and Cambridge University as a Fulbright Senior Scholar. At the former, I was a guest of the Human Communication Research Centre where I participated in a research group working on the cognitive and logical aspects of diagrams. At Cambridge, I shared an office with Dennis Bray (one of the authors of the famous book *Molecular Biology of the Cell*). My primary contacts at Cambridge were people working in natural language parsing in the Computing Laboratory (the University's CS department).

During my sabbatical, I was able to develop many new and exciting ideas about diagram analysis and intelligent knowledge repositories, and it is these ideas that will occupy me for some time. My overall impression with the UK workers, especially in Edinburgh, is that their CS work is very logic-oriented and very far in spirit from what is needed to analyze diagrams successfully. They seem so wedded to a logic-based view that it is difficult if not impossible for them to break out of it when needed. The workers at Cambridge were more flexible, but working on the cutting edge of natural language parsing left them so deeply into the details that it was not easy to talk to them about the radically different demands of diagram parsing.

The most important new idea that I developed while in the UK was the Knowledge-Based Authoring Tool (KBAT). This will allow authors to create knowledge-based versions of their papers as they write them and prepare them for publication. (see Futrelle & Fridman paper). This will avoid having knowledge experts building systems that attempt to reconstruct the knowledge in papers after the fact, by parsing and other techniques. The underlying knowledge in the paper would apply uniformly to text and graphics. So far, this has resulted in one paper, an internal

university proposal, and a preliminary proposal to the NSF Challenges in CISE program.

The happy ending to all this is that we are out in front with approaches to scientific diagram analysis that are new and that work. The major results have been published and additional work will be submitted shortly. We are confident that we have made a substantial contribution to the design of electronic information systems for scientific research.

The personnel supported on this grant were,

Nikos Nikolakis, a CS graduate student, who brought the Diagram Understanding System to a full, productive state, as part of his doctoral dissertation. He received his PhD for this work in January, 1997. His work was absolutely excellent.

Michael Beresis, was involved with our initial experiments in translating diagram parse results into a knowledge-based form.

Dr. Arthur Miller (part-time) served as a resource person for biology/biochemistry expertise. His PhD is in enzyme biochemistry, and he also received an MS in Computer Science from Northeastern University.

Chris Hopkins, one of our top CS undergraduates, worked on two projects. One was a gene-drawing program that was a first step towards building a KBAT system, and the other, a simple object store that could save and restore objects from Lisp, Dylan, C++, or any mix of them.

Publications under this grant

Futrelle, R. P., Zhang, X., & Sekiya, Y. (1994). Corpus Linguistics for Establishing the Natural Language Content of Digital Library Documents. In N. R. Adam, B. K. Bhargava & Y. Yesha (Eds.), *Digital Libraries. Current Issues* (pp. 165-180). Berlin Heidelberg: Springer-Verlag.

Futrelle, R. P., & Nikolakis, N. (1995). Efficient Analysis of Complex Diagrams using Constraint-Based Parsing. In *ICDAR-95 (Intl. Conf. on Document Analysis & Recognition)*, (pp. 782-790). Montreal, Canada.

Futrelle, R. P., & Fridman, N. (1995). Principles and Tools for Authoring Knowledge-Rich Documents. In *DEXA 95 (Database and Expert System Applications) Workshop on Digital Libraries*, (pp. 357-362.). London, U.K.

Futrelle, R. P., & Nikolakis, N. (to be submitted, 1997). Context-Based Constraint Grammars for Visual Parsing. *Pattern Analysis and Machine Intelligence*.

Futrelle, R. P., & Nikolakis, N. (to be submitted, 1997). Spatial Indexing for Diagram Analysis. *Pattern Recognition*.

Nikolakis, N., & Futrelle, R. P. (to be submitted, 1997). Diagram Analysis using Context-Based Constraint Grammars. *Computer Vision, Graphics and Image Processing -- Image Understanding*.