

EXTRACTING AND STRUCTURING SUBCELLULAR LOCATION INFORMATION FROM ON-LINE JOURNAL ARTICLES: THE SUBCELLULAR LOCATION IMAGE FINDER

Robert F. Murphy Zhenzhen Kou Juchang Hua Matthew Joffe William W. Cohen
murphy@cmu.edu zkou@andrew.cmu.edu juchangh@andrew.cmu.edu mjoffe@andrew.cmu.edu wcohen@cs.cmu.edu
Departments of Biological Sciences and Biomedical Engineering and Center for Automated Learning and Discovery
Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA/U.S.A

ABSTRACT

Previous applications of information extraction methods to articles in biomedical journals have predominantly been based on interpreting article text. This often leads to uncertainty about whether statements that are found are attempts at reviews or summaries of data in other papers, conjectures or opinions, or conclusions from evidence presented in the paper at hand. The ability to extract information from the primary data presented in an article, which is often in the form of images, would allow more accurate information to be extracted. Towards this end, we have built a system that extracts information on one particular aspect of biology from a combination of text and image in journal articles. The design and performance of this system are described here, along with conclusions about possible improvements in the scientific publishing process that we have drawn from our implementation process.

KEY WORDS

Knowledge and Information Retrieval, Multimedia Databases, Data and Text Mining, Location Proteomics

1. Introduction

Biomedical research has undergone a major paradigm shift from consisting primarily of projects in which an individual investigator studies many aspects of a single gene, protein or process to increasingly consisting of projects in which teams of investigators study a single aspect of all genes, proteins or processes in a given cell type, tissue or organism. The successful completion of various genome projects, with their focus on obtaining the sequence of all genes in a particular organism, led this paradigm shift. In general, the results from these projects are objective (at least in the sense that the criteria for decisions are clearly specified independently from the data), systematic, widely useful, and well-suited to delivery via structured databases. While remarkable insights into a wide range of biological phenomena were achieved before the advent of genomics (and of course such insights continue to be achieved), results of traditional biological research are most commonly

communicated via journal articles in which raw data, methods, processed results and conclusions are mixed. In addition, the writing styles, vocabularies, and assumptions used for interpretation vary widely from paper to paper.

Thus, there is a dramatic contrast in the ease with which results from the two paradigms can be organized and communicated. This has created a critical need for approaches that can bridge between the systematic, structured information in biological databases and the idiosyncratic, unstructured information in journal articles. This is often posed as a need for automated annotation of gene and protein sequences, but there are at least two significant ways in which the general need differs from the specific approaches taken to sequence annotation. The first is that the need extends to extracting information from literature about biological phenomena at the molecule, cell, tissue and organism level that do not relate directly to sequence. The second is that most prior annotation work has focused on extracting information from the *text* in abstracts (or more rarely, journal articles), but not from supporting published *data* that is often in the form of images.

To illustrate the initial feasibility of addressing these broader needs, we have developed a prototype system that can extract structured information from images and text in journal articles. The focus of this system is on one class of images, those produced by fluorescence microscopy, that capture information about the distribution of proteins and other biological macromolecules inside cells. It builds on our prior work demonstrating the feasibility of fully automated recognition of the distributions characteristic of the major structures that comprise a eukaryotic cell. The work to date not only provides a usable resource for biologists, but also reveals the most difficult challenges for building systems for other categories of biological figures. The work further suggests some alterations in scientific publishing practices that could facilitate automated interpretation without putting undue demands on authors or interfering significantly with the traditional appearance of articles.

