Enhanced Browsing System for Electronic Theses and Dissertations

Venkat Srinivasan, Mohamed Magdy, Edward A. Fox

Virginia Tech, Blacksburg, VA

{svenkat, mmagdy, fox}@vt.edu

Outline

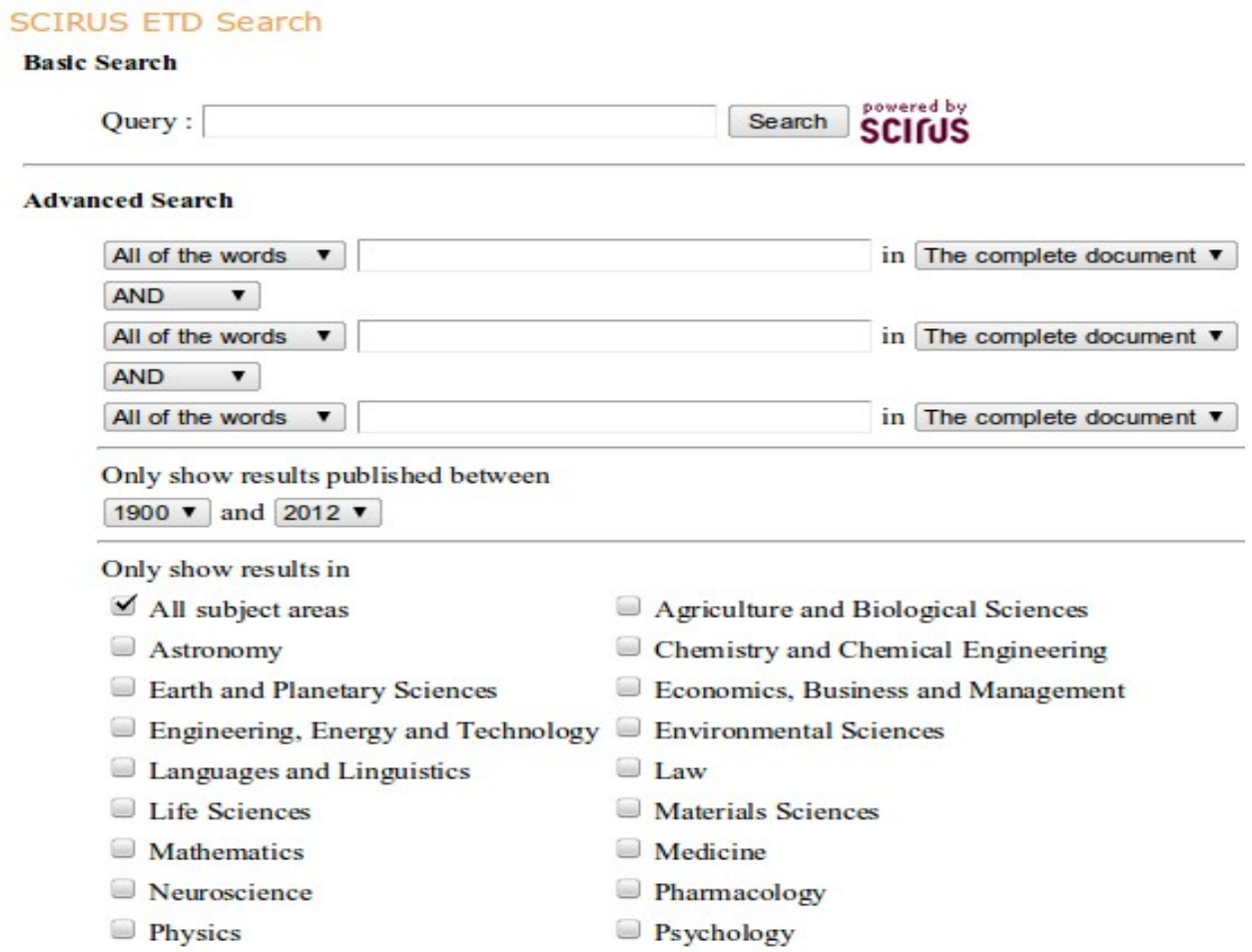
- Motivation
- Research Question
- Related Work
- Goals
- Methods
- Results
- Demo

Motivation

- Current ways of exploring ETDs (and ETD collections) limited in functionality
  - Only full-text and/or metadata based search interfaces available
  - Complete ETD is served up as a result of a user query

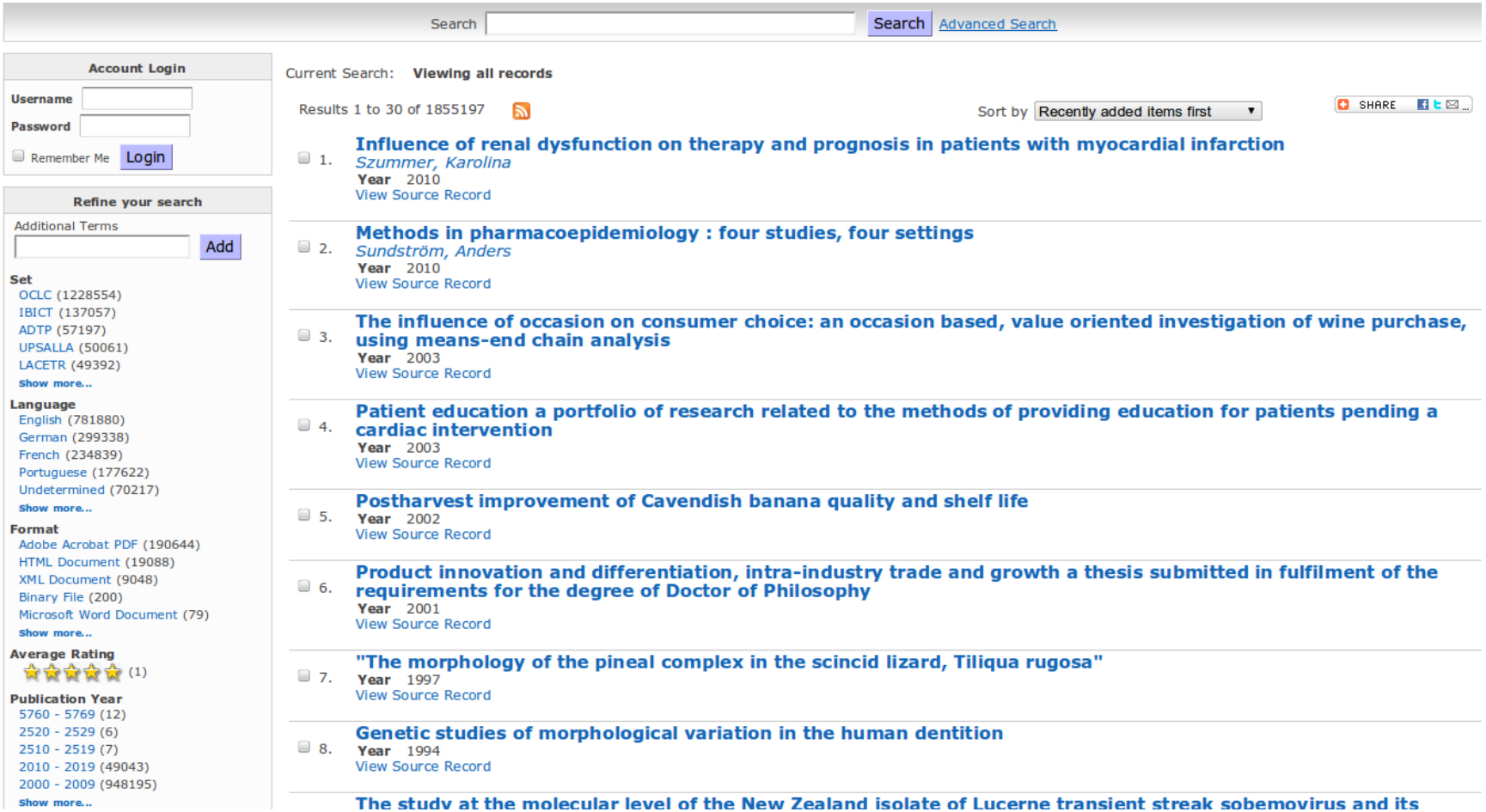
Motivation

- Scirus ETD Search (Interface at NDLTD website)



Motivation

- VTLS ETD Search



Enhanced ETD Browsing http://zappa.dlib.vt.edu/etd

Motivation

- ETDs are typically long, and reading and comprehending them can take a long time

- Utility of ETDs as educational resources can be increased by developing tools that aid in their reading and comprehension

Research Question

- ETDs have structure:
  - Chapters, sections, sub-sections …
- ETDs contain different streams of information
  - Text, figures, tables …

- Can all this information be leveraged to improve reading and comprehension, and thereby the utility of ETDs?

Related Work

- Document segmentation
  - Dividing the document into segments (chapters, for example)
  - Typical methods involve identification of Table of Contents, and subsequent identification of various sections using this information
  - Several approaches exist, but give average-to-good results only within a specific genre of documents (not generalizable)

Related Work

- Enhanced document browsing
  - Allows for browsing of different features or dimensions of a document
  - Not much prior work
  - A known example is that made available by journal “Cell” (following slides) for browsing a sample research paper

Related Work

Cell's document browsing prototype

Summary | Introduction | Results | Discussion | Exp. Proc. | Data | References | Supp. Info. | Related Info. | Comments (2)

Cell, Volume 140, Issue 1, 49-61, 8 January 2010

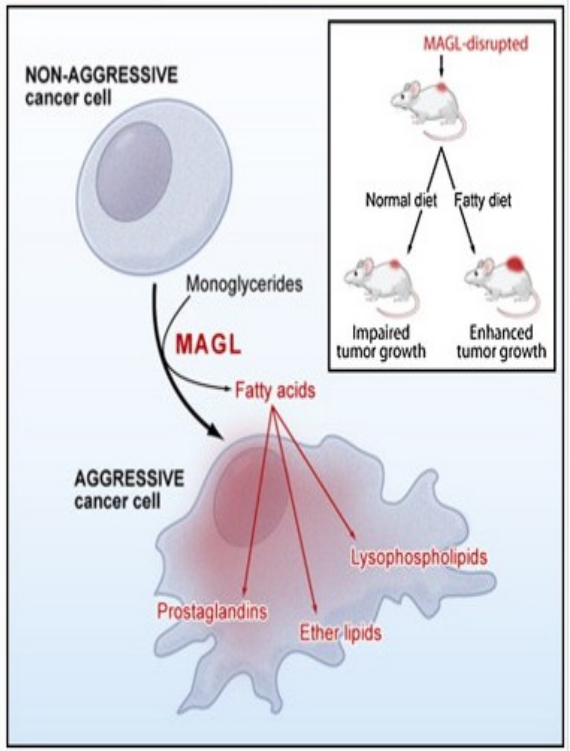
10.1016/j.cell.2009.11.027

Referred to by: Chewing the Fat on Tumor Cell Metabolism...

Referred to by: A New Age for MAGL

PDF 2.19 MB
Extended PDF 4 KB
Export Citation
Permissions

Graphical abstract | PaperFlick



Authors

Daniel K. Nomura, Jonathan Z. Long, Sherry Niessen, Heather S. Hoover, Shu-Wing Ng, Benjamin F. Cravatt ✉ See Affiliations

Highlights

- Monoacylglycerol lipase (MAGL) is elevated in aggressive human cancer cells
- Loss of MAGL lowers fatty acid levels in cancer cells and impairs pathogenicity
- MAGL controls a signaling network enriched in protumorigenic lipids
- A high-fat diet can restore the growth of tumors lacking MAGL in vivo

Summary

Tumor cells display progressive changes in metabolism that correlate with malignancy, including development of a lipogenic phenotype. How stored fats are liberated and remodeled to support cancer pathogenesis, however, remains unknown. Here, we show that the enzyme monoacylglycerol lipase (MAGL) is highly expressed in aggressive human cancer cells and primary tumors, where it regulates a fatty acid network enriched in oncogenic signaling lipids that promotes migration, invasion, survival, and in vivo tumor growth. Overexpression of MAGL in nonaggressive cancer cells recapitulates this fatty acid network and increases their pathogenicity—phenotypes that are reversed by an MAGL inhibitor. Impairments in MAGL-dependent tumor growth are rescued by a high-fat diet, indicating that exogenous sources of fatty acids can contribute to malignancy in cancers lacking MAGL activity. Together, these findings reveal how cancer cells can co-opt a lipolytic enzyme to translate their lipogenic state into an array of protumorigenic signals.

Enhanced ETD Browsing http://zappa.dlib.vt.edu/etd

Related Work

Cell's document browsing prototype

Summary | Introduction | Results | Discussion | Exp. Proc. | Data | References | Supp. Info. | Related Info. | Comments (2)

Activity-Based Proteomic Analysis of Hydrolytic Enzymes in Human Cancer Cells

To identify enzyme activities that contribute to cancer pathogenesis, we conducted a functional proteomic analysis of a panel of aggressive and nonaggressive human cancer cell lines from multiple tumors of origin, including melanoma (aggressive [C8161, MUM2B], nonaggressive [MUM2C]), ovarian (aggressive [SKOV3], nonaggressive [OVCAR3]), and breast (aggressive [231MFP], nonaggressive [MCF7]) cancer. Aggressive cancer lines were confirmed to display much greater in vitro migration and in vivo tumor-growth rates compared to their nonaggressive counterparts (Figure S1 available online), as previously shown (Jessani et al., 2002,Jessani et al., 2004,Seftor et al., 2002,Welch et al., 1991). Proteomes from these cancer lines were screened by activity-based protein profiling (ABPP) using serine hydrolase-directed fluorophosphonate (FP) activity-based probes (Jessani et al., 2002,Patricelli et al., 2001). Serine hydrolases are one of the largest and most diverse enzyme classes in the human proteome (representing ~1%–1.5% of all human proteins) and play important roles in many biochemical processes of potential relevance to cancer, such as proteolysis (McMahon and Kwaan, 2008,Puustinen et al., 2009), signal transduction (Puustinen et al., 2009), and lipid metabolism (Menendez and Lupu, 2007,Zechner et al., 2005). The goal of this study was to identify hydrolytic enzyme activities that were consistently altered in aggressive versus nonaggressive cancer lines, working under the hypothesis that these conserved enzymatic changes would have a high probability of contributing to the pathogenic state of cancer cells.

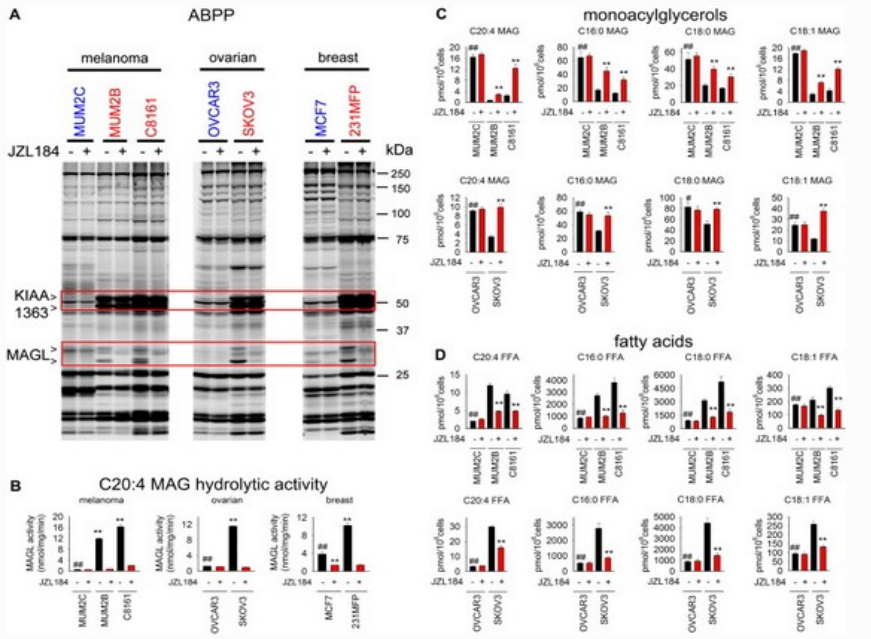
Serine hydrolase activities were identified from aggressive and nonaggressive cancer cell proteomes by enrichment with a biotinylated FP probe (Liu et al., 1999) and multidimensional liquid chromatography-mass spectrometry analysis (Jessani et al., 2005). Among the more than 50 serine hydrolases detected in this analysis (Tables S1, S2, and S3), two enzymes, KIAA1363 and MAGL, were found to be consistently elevated in aggressive cancer cells relative to their nonaggressive counterparts, as judged by spectral counting (Jessani et al., 2005,Liu et al., 2004). We confirmed elevations in KIAA1363 and MAGL in aggressive cancer cells by gel-based

Figure 1 | Figure 2 | Figure 3 | Figure 4 | Figure 5 | Figure 6 | Figure 7 | Figure S1 | (5) >

Show Caption | View High Quality | PowerPoint

Figure 1. MAGL Is Elevated in Aggressive Cancer Cells, where the Enzyme Regulate...

75%



Enhanced ETD Browsing

http://zappa.dlib.vt.edu/etd

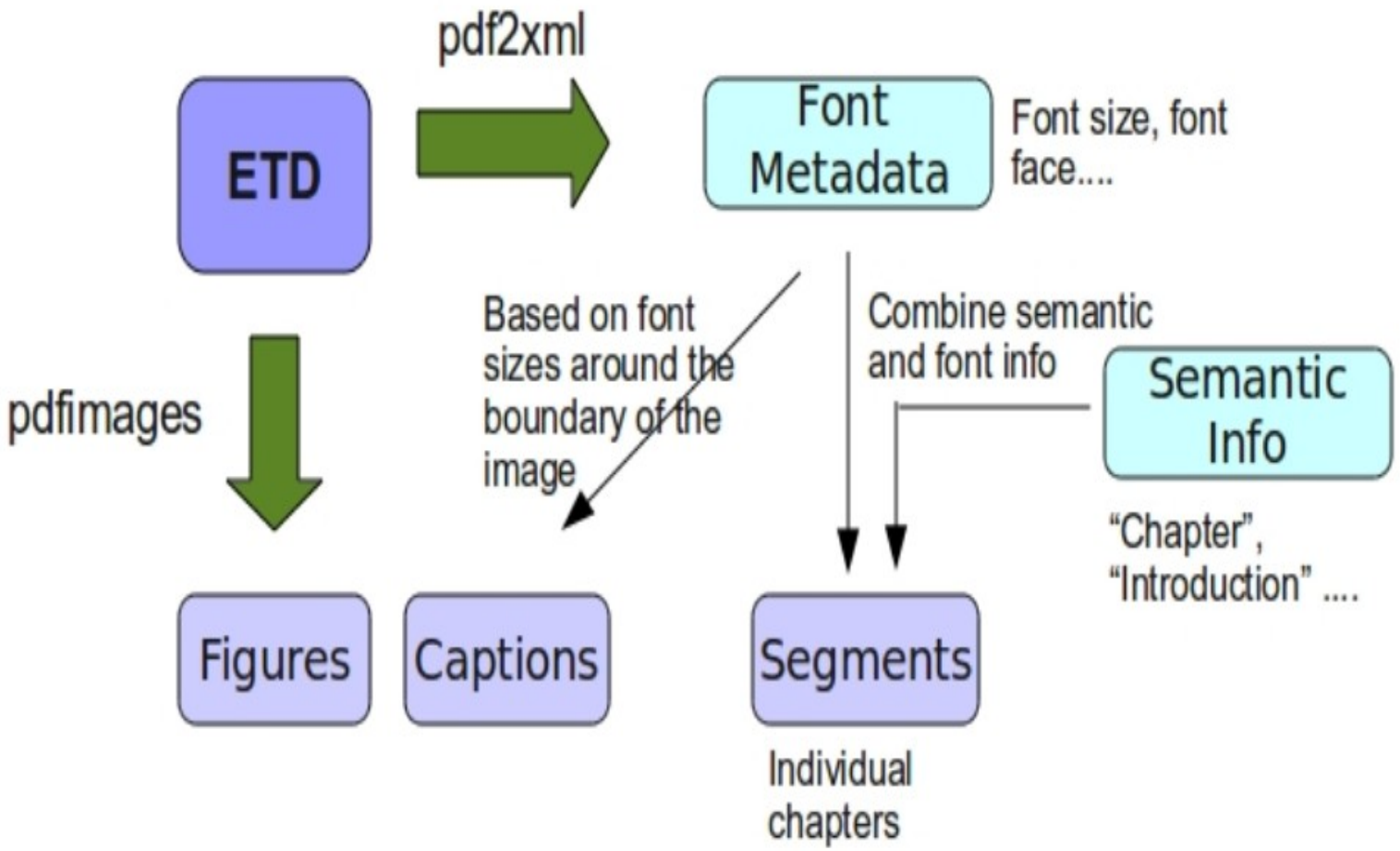
Goals

- Develop tools to extract individual chapters, images, etc. from ETDs
  - ETDs generally occur as PDF files. So we focus on developing tools to extract such information from PDFs.
- Develop a web-based prototype tool to lay out all this information in a form that aids reading, comprehension and navigation of ETDs

Methods

- Extracting individual chapters
  - Open source tool (pdf2xml) used to obtain font related metadata (font size, type, face etc.) from PDFs
  - Combine additional semantic information (keywords like “Chapter”, “Introduction”....) to identify chapter boundaries
- Extracting images and captions
  - Several Linux utilities used for extracting images (pdfimages, pnmtojpeg etc.) from PDFs
  - Captions extracted using output of pdf2xml

Methods


pdf2xml
ETD
Font Metadata
Font size, font face....
Based on font sizes around the boundary of the image
Combine semantic and font info
Semantic Info
pdfimages
"Chapter", "Introduction" ....
Figures
Captions
Segments
Individual chapters

Methods

- Web-based prototype tool to aid in reading, comprehension and navigation of ETDs
  - Content management system Drupal used to develop the prototype
  - Users can browse different dimensions, like chapters/sections, figures, references ....

Results

- Perl/Python based tools developed to extract individual chapters from ETDs
  - Out of 40 ETDs randomly selected for experimentation, the tools perfectly segmented 25 of them
    (Accuracy = 62.5%)
- Sources of error include:
  - Varied chapter beginning styles
  - Inconsistent font size/face usage

Results

- Perl/Python based tools developed to extract images and captions from PDFs
  - Out of 10 ETDs selected at random for experimentation, containing 91 images, 36 images were recovered
    (Recall = 25.2%)
- Sources of error include:
  - Limitations of open source tools used

Demo

http://zappa.dlib.vt.edu/etd

Future Work

- Better tools for extraction of information streams
  - Commercial tools like TET for extracting images and text
- More research, and development of better user interface
  - User studies required to understand the effectiveness of the methods developed
- Make a sizable number of ETDs available for enhanced browsing

THOUGHTS?