Handling figures in document summarization

Robert P. Futrelle
Biological Knowledge Laboratory
College of Computer & Information Science WVH202
Northeastern University
Boston, MA 02115, USA
futrelle@ccs.neu.edu

BioNLP.org Diagrams.org

Importance

Figures occupy a major percentage of the scientific literature.

Biology literature (PuMed): 100M figures totaling 4TB.

The percentage in Biology papers is about 50% (!).

So summarization should include figures.

(Focus here is on diagrams = line drawings)

Obstacles

There are no parsable corpora, e.g., diagrams are in raster format.

Even when vectorized, parsing diagrams is difficult.

Vectorization good enough for parsing is difficult.

Summarization involves an interplay between text and figures.

Figures as a percentage of papers - an example

Explicit references: 25 sentences, 707 words = 8%
Implicit discussions: 28 sentences, 790 words = 9%
Figure captions: 83 sentences, 1477 words = 17%
Figure areas (equivalent words): 2119 words = 24%

TOTAL devoted to figures 5093 words = **58%**

Details

Paper analyzed was PDF version of:

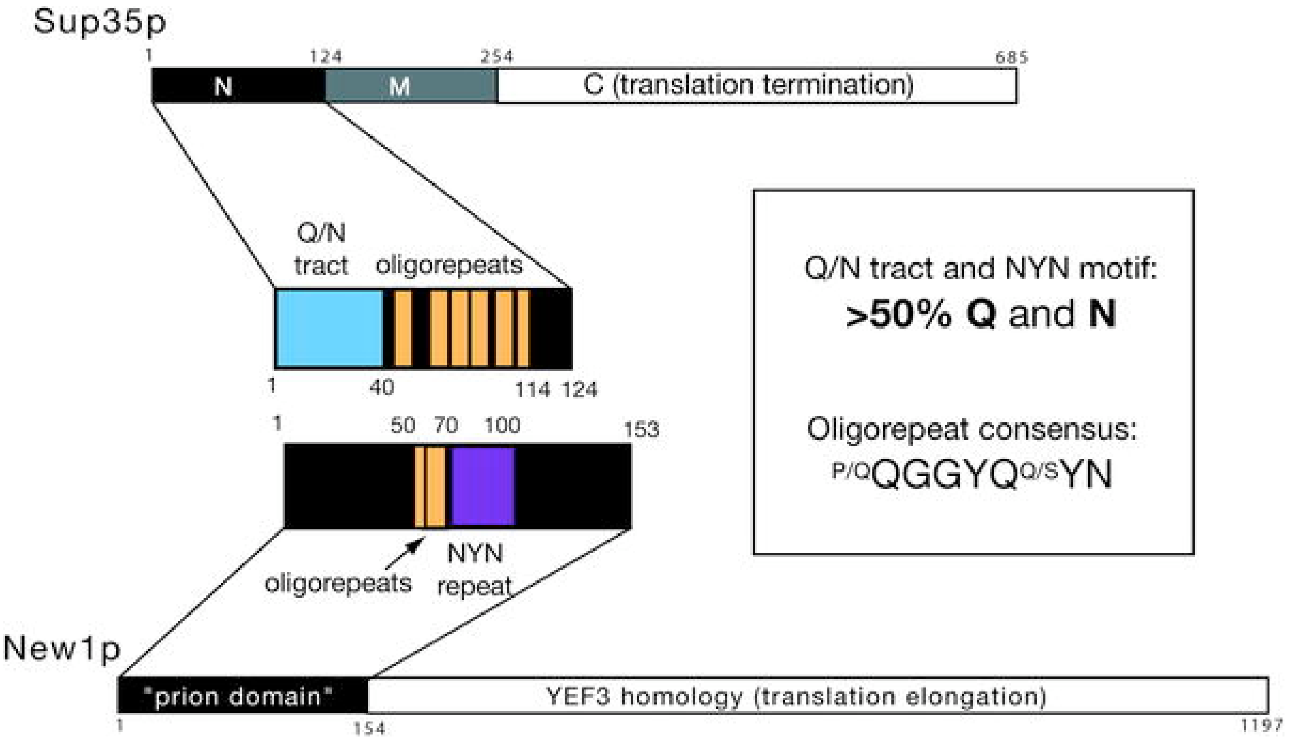
PLoS Biology **2**(4) 2004 "Dissection and Design of Yeast Prions" by Osherovich, et al

Area of figures measured in words that would occupy that space.

Explicit/Implicit sentences example

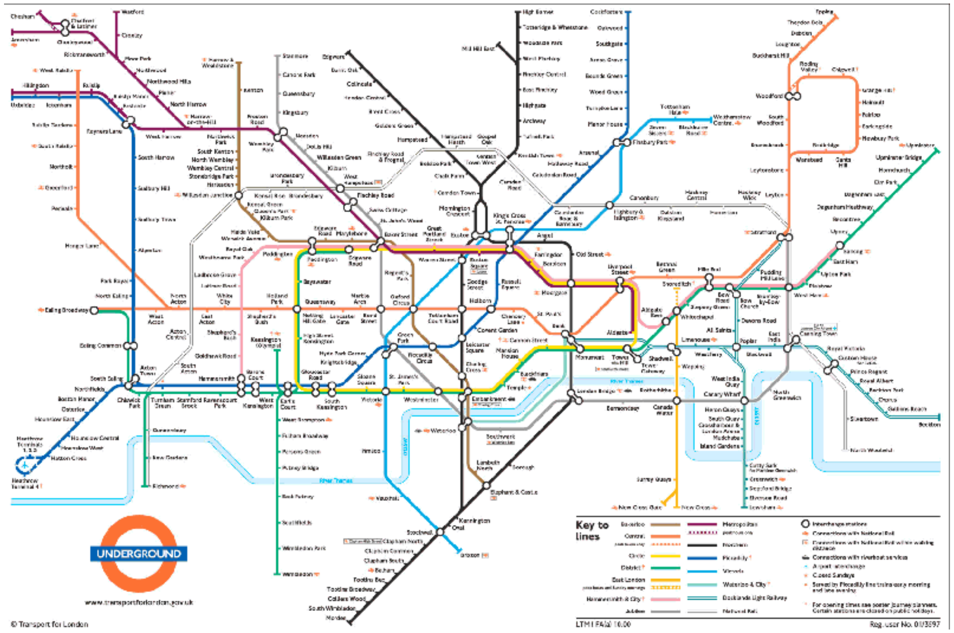
Explicit: The charged middle domain (M) is not required for prion behavior, but modulates the efficiency of chaperone-dependent prion transmission (Liu et al. 2002; L.Z.O., unpublished data) (Figure 1).

Implicit: Two distinct regions in the N domain have previously been implicated in Sup35p aggregation: a Q/N-rich tract (residues 1–39) (DePace et al. 1998) and an oligopeptide repeat (residues 40–112) that consists of five and a half degenerate repeats of the consensus sequence P/QQGGYQQ/SYN (Liu and Lindquist 1999; Parham et al. 2001; Crist et al. 2003).

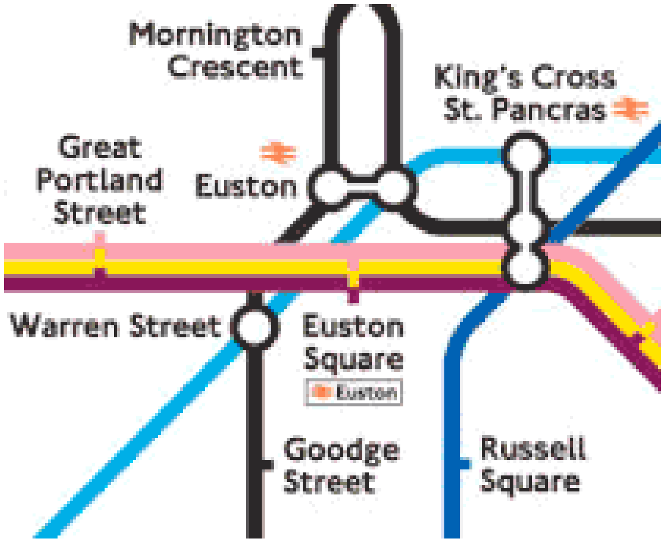


Example1: Shrinking to "thumbnails" -- informative and indicative summaries

A thumbnail that attempts to be informative, but with unreadable text:



A thumbnail that is (only) indicative but useful as a guide to the full figure:



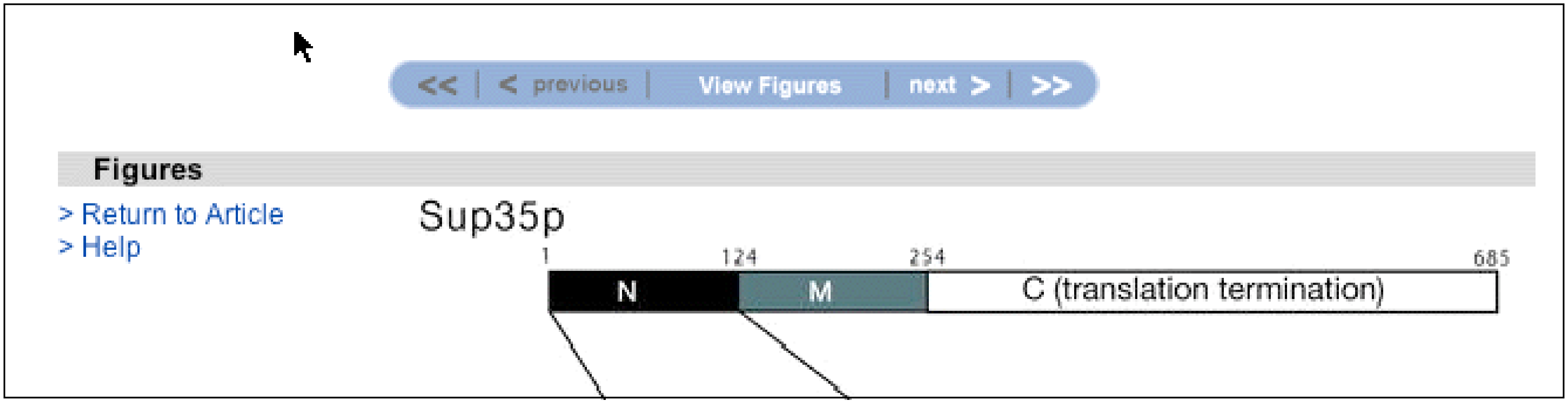
(See the Gallery item: "The London Underground Map" on Diagrams.org)

Example 2: Figures as document surrogates - "View Figures"

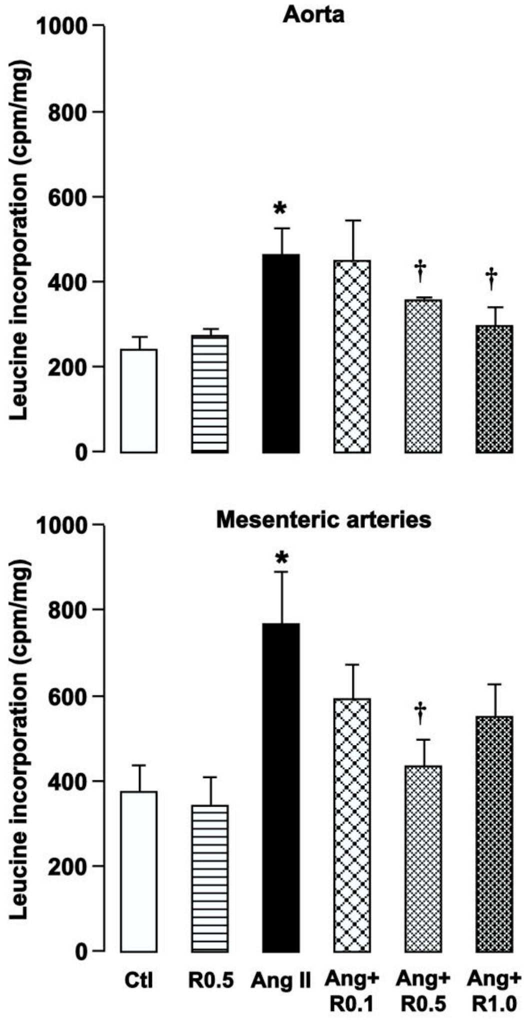
The earlier figure (our slide #2) was taken from one frame of a slide show presentation available for every paper in the PLoS journals.

Each figure includes the figure caption and a citation to the containing paper. Below is a portion of such a page.

Authors could tailor figures and captions to this new presentation mode, so that a reader could peruse a paper without "reading" it.

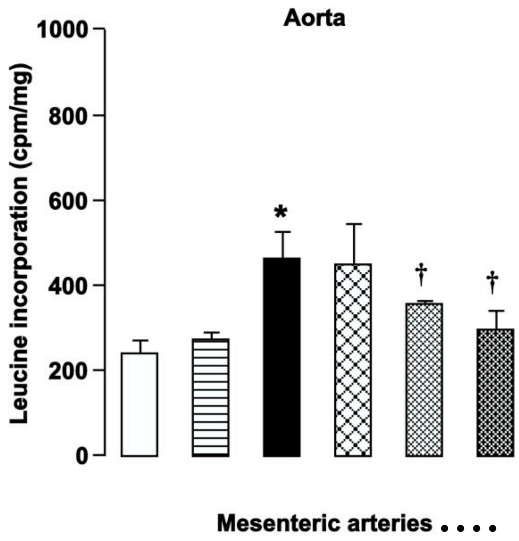

<< | < previous | View Figures | next > | >>
Figures
> Return to Article
> Help
Sup35p
1
124
254
685
N
M
C (translation termination)

Example 3: Extraction of subfigure components



Many figures have a number of similar components, any of which could form an indicative summary.

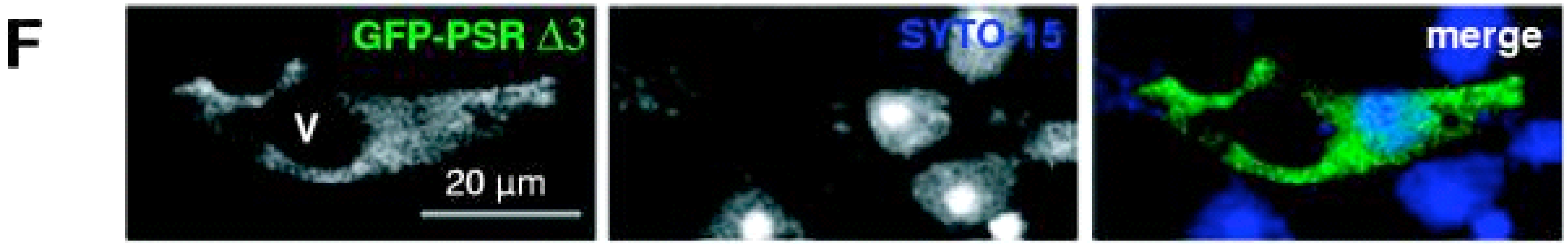
One approach would be to extract the top graph on the left, but to include the label of the bottom one to indicate its omission:



BMC Cardiovascular Disorders 4(6) (2004) "Signaling of angiotensin II-induced vascular protein synthesis in conduit and resistance arteries in vivo" by Daigle, et al

Example 4: Text and figures

Text can occur within diagrams, in captions and in body text that explicitly or implicitly references a figure. The text within the micrographs below would have to be found via OCR. "V" is a typical deictic reference, "pointing" at a region.



PSR localisation in living animals Single optical sections of GFP expressing cells of living hydra after transfection with GFP-PSR (A and B), PSR-GFP (C), NLS-mutants ΔNLS1-2 (D) and ΔNLS3 (E and F). Left hand panels show GFP, middle panels nuclear staining with SYTO-15 or membrane staining with FM-464 as indicated. Right hand panels are merged images. V indicates vacuoles typically seen in living hydra cells.

Note that the phrase: "as indicated" requires that text within the figure be examined.

BMC Cell Biology (5)26 2004 by Cikala.et al "The phosphatidylserine receptor from Hydra is a nuclear protein with potential Fe(II) dependent oxygenase activity"

Text and figures 2

The full and proper summary of a document will have to involve a simultaneous and integrated summary of the text and the figures.

This is a tall order, so an approach which summarizes the two separately is a way to begin. But there are a variety of simple approaches that might work. For example, including the first sentence in a paper that mentions a specific figure, along with all or part of the figure is a good heuristic. If only a portion of a figure is extracted, e.g., part "A", then only the figure title, if any, and the portion of the figure caption explicitly describing "A" could be retained.

Prospects for automation

Diagrams in particular need to be vectorized and the text within them analyzed by OCR. This is a work in progress in our group.

Vectorized diagrams need to be parsed to reveal their structure. We developed a successful parsing system and are now redeveloping it as a Java-based system (not a simple port!).

Neither of the tasks above is simple or quick. But as we make progress, we will be able to investigate techniques for automating the various summarization strategies described here.

Discussion

We have described work-in-progress. We certainly have no automated system for summarizing figures and their associated text in Biology paper (and there is no other system that can do this, to our knowledge).

Further details and references are contained in the Workshop paper.