MS/MEng Thesis Project:
Finding and Understanding Chemical Structures in Documents
Professor Randall Davis
CSAIL

Indexing documents by keywords is a familiar and powerful technology in today's web-driven world. But what if the information you're seeking is in a graphic, and is never described in words? Put slightly differently, full-text indexing has proven to be a powerful foundation for information retrieval; what if we could read and understand (and then index the information in) diagrams as well?

Consider publications in chemistry and biology, for example, with their numerous diagrams of chemical compounds, reaction paths, etc. The example below (from [Jenkins08]) shows a synthesis pathway, with the accompanying text indicating only that "In this mechanism, cysteine adds to C6 of the pyrimidine, generating anion **7**. Expulsion of the leaving group gives **8,**" never naming the intermediate compounds. What if you happened to be interested in the compound numbered **8?** Traditional keyword search could never find it in [Jenkins08].

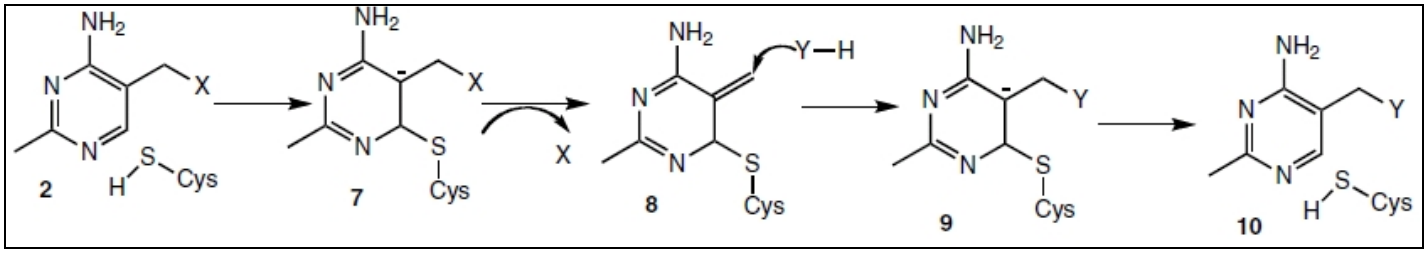

Fig 1: Synthesis pathway from [Jenkins08]

Indexing chemical and biological literature by the information in figures as well as text will open up new sources of information, enabling more effective search and retrieval. Searching for a chemical structure rather than a name also enables inexact matching, based on a measure of structure similarity.

But how should we specify the structure to search for? Compounds have names, but they are often long and unwieldy. It's far more natural to draw the structure. We ought, for example, to be able to draw this structure (Prozac)

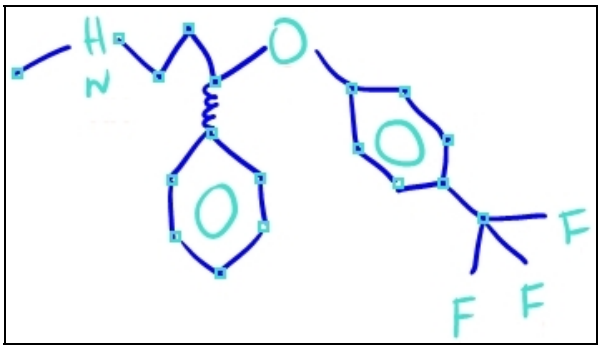

Fig 2: Prozac

and search based on *that*. With a system like this we could for example ask, "Are there any compounds like this [sketch] in recent journal articles discussing birth defects?"

We have created a program capable of understanding hand-drawn chemical structures of the sort shown in Fig. 2 [Ouyang07], and now want to build a system capable of scanning chemical literature, "parsing" the diagrams so that information in them can be indexed as well. The thesis here will be in designing and developing software capable of interpreting chemical diagrams of the sort found in online literature, building on the technology we have already created for understanding hand-drawn structures.

In the longer term this work could lead to understanding and indexing information in figures in a wide variety of literature.

Candidates need to be very good programmers; familiarity with basic image processing techniques is an advantage, as is experience with diagram understanding and chemistry. RA support available for the right candidate. Contact davis@csail.mit.edu.

[Jenkins08]
Jenkins et al., Mutagenesis studies on TenA: A thiamin salvage enzyme from Bacillus subtilis, *Bioorganic Chemistry*, Feb 2008, pp.29-32.

[Ouyang07]
Tom Ouyang, Randall Davis, Recognition of Hand-Drawn Chemical Diagrams, *Proc AAAI 2007*, pp. 846-851.