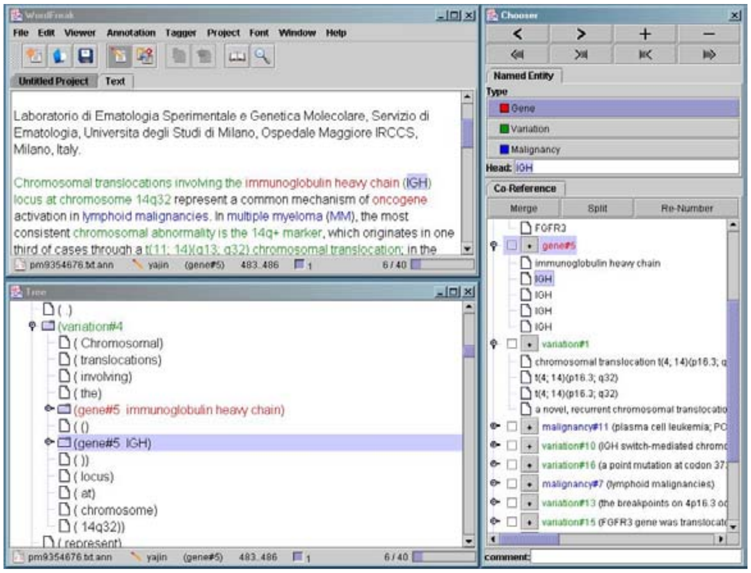
EIA-0205448: “Mining the Bibliome: Information Extraction from the Biomedical Literature”
University of Pennsylvania
Investigators: Aravind Joshi, Susan Davidson, Mark Liberman, Mitch Marcus, Martha Palmer
Biomedical collaborators: Peter White (CHOP, eGenome), Paula Matuszek (GSK)

Websites: http://www.ldc.upenn.edu/myl/ITR
http://www.cis.upenn.edu/~mamandel/term.htm

Our goal is qualitatively better methods for automatically extracting information from the biomedical literature, relying on three techniques: high-accuracy parsing, shallow semantic analysis, and integration of existing databases. An initial step is to create annotated corpora in collaboration with biomedical researchers: two test cases are gene variations in pediatric oncology, and inhibition of CYP450 enzymes.

The screen shot below shows part of a scientific abstract that we have annotated for entities and relations relevant to a project (based at Children’s Hospital of Pennsylvania) to construct a database of cancer-associated mutational events.



Laboratorio di Ematologia Sperimentale e Genetica Molecolare, Servizio di Ematologia, Universita degli Studi di Milano, Ospedale Maggiore IRCCS, Milano, Italy.

Chromosomal translocations involving the immunoglobulin heavy chain (IGH) locus at chromosome 14q32 represent a common mechanism of oncogene activation in lymphoid malignancies. In multiple myeloma (MM), the most consistent chromosomal abnormality is the 14q+ marker, which originates in one third of cases through a t(11; 14)(q13; q32) chromosomal translocation; in the

pm9354676.txt.ann yajin (gene#5) 483 486 1 6 / 40

(.)
(variation#4
(Chromosomal)
(translocations)
(involving)
(the)
(gene#5 immunoglobulin heavy chain)
(()
(gene#5 IGH)
( ))
(locus)
(at)
(chromosome)
(14q32))
(represent)

pm9354676.txt.ann yajin (gene#5) 483 486 1 6 / 40

🟥 Gene
🟩 Variation
🟦 Malignancy
Head: IGH

FGFR3
gene#5
immunoglobulin heavy chain
IGH
IGH
IGH
IGH
variation#1
chromosomal translocation t(4; 14)(p16.3; q
t(4; 14)(p16.3; q32)
t(4; 14)(p16.3; q32)
a novel, recurrent chromosomal translocatio
malignancy#11 (plasma cell leukemia; PC
variation#10 (IGH switch-mediated chromo
variation#16 (a point mutation at codon 37
malignancy#7 (lymphoid malignancies)
variation#13 (the breakpoints on 4p16.3 oc
variation#15 (FGFR3 gene was translocat

The many millions of biomedical publications available in electronic form contain a vast quantity of scientific information. Researchers would like access to this information structured in terms of well-defined relations (like “inhibition” or “mutation”) among entities of interest (like “gene”, “compound” or “cell line”). Recent techniques from computational linguistics can make more of this information accessible to biomedical researchers; at the same time, large existing biomedical databases promise an unparalleled depth of semantic support to make linguistic analysis more accurate.

To experiment with new textual information-extraction techniques, training and testing corpora are needed. To this end, we’ve developed or adapted software tools that allow human experts to annotate biomedical texts for relevant entities and relations (“entity tagging”), to mark syntactic structure (producing a “tree bank”), and to indicate shallow semantic structure, such as co-reference relations and predicate-argument relations (producing a “proposition bank”). We’ve defined compatible forms of “stand-off markup” (that is, annotations kept separate from the texts annotated), in order to allow multiple independent forms of text annotation to be created and used in an integrated way. We’re producing annotation specifications and training materials for syntactic and semantic annotation of biomedical text, and an annotation crew is testing these tools and specifications by multiple annotation of a set of initial test sets, in order to ensure adequate inter-annotator agreement. Once enough annotated material has been created, automatic taggers and parsers are trained for each task, and the annotators’ job becomes correcting the programs’ output, or creating initial material for a new application or a new dimension of analysis.

All of the software developed in this project is available under open-source license, and all annotated texts and resulting databases will be published for the use of other researchers.