Summarizing Key Concepts using Citation Sentences

Ariel S. Schwartz and **Marti Hearst**
EECS and SIMS
University of California at Berkeley
Berkeley, CA 94720
sariel@cs.berkeley.edu, hearst@sims.berkeley.edu

Citations have great potential to be a valuable resource in mining the bioscience literature (Nakov et al., 2004). The text around citations (or *citances*) tends to state biological facts with reference to the original papers that discovered them. The cited facts are typically stated in a more concise way in the citing papers than in the original. We hypothesize that in many cases, as time goes by, the citation sentences can more accurately indicate the most important contributions of a paper than its original abstract.

One can use various NLP tools to identify and normalize the important entities in (a) the abstract of the original article, (b) the body of the original article, and (c) the citances to the article. We hypothesize that grouping entities by their occurrence in the citances represents a better summary of the original paper than using only the first two sources of information.

To help determine the utility of the approach, we are applying it to the problem of identifying articles that discuss critical residue functionality, for use in *PhyloFacts* a phylogenomic database (Sjolander, 2004).

Consider the article shown in Figure 1. This paper is a prominent one, published in 1992, with nearly 500 papers citing it. For about 200 of these papers, we downloaded the sentences that surround the citation within the full text. Some examples are shown in Figure 2.

We are developing a statistical model that will group these entities into potentially overlapping groups, where each group represents a central idea in the original paper. In the example shown, some of the citances emphasize what the paper reports about the structural elements of the SH2 domain, whereas other emphasize its findings on interactions and others focus on the critical residues.

Often several articles are cited in the same citance, so it is important to untangle which entities belong to which citation; by pursuing overlapping sets, our model should be able to eliminate most spurious references.

The same entity is often described in many different ways. Prior work has shown how to use redundant information across citations to help normalize entities (Wellner et al., 2004; Pasula et al., 2003); similar techniques may work with entities mentioned in citances. This can be combined with prior work on normalizing entity names in bioscience text, e.g, (Morgan et al., 2004). For a detailed review of related work see (Nakov et al., 2004).

By emphasizing entities the model potentially misses important relationships between the entities. It remains to be determined whether or not relationships must be modeled explicitly in order to create a useful summary.

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 134–135,
New York City, June 2006.

References

A. A. Morgan, L. Hirschman, M. Colosimo, A. S. Yeh, and J. B. Colombe. 2004. Gene name identification and normalization using a model organism database. *Journal of Biomedical Informatics*, 37(6):396–410.

P. I. Nakov, A. S. Schwartz, and M. Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shiptser. 2003. Identity uncertainty and citation matching. *Advances In Neural Information Processing Systems*, 15.

K. Sjolander. 2004. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinf.*, 20(2):170–179.

B. Wellner, A. McCallum, F. Peng, and M. Hay. 2004. An integrated, conditional model of information extraction and coreference with application to citation graph construction. In *20th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Waksman G, Kominos D, Robertson SC, Pant N, Baltimore D, Birge RB, Cowburn D, Hanafusa H, Mayer BJ, Overduin M, et al., *Abstract Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides.*
Nature. 1992 Aug 20;358(6388):646-53. [PMID: 1379696]

Three-dimensional structures of complexes of the SH2 domain of the v-src oncogene product with two phosphotyrosyl peptides have been determined by X-ray crystallography at resolutions of 1.5 and 2.0 A, respectively. A central antiparallel beta-sheet in the structure is flanked by two alpha-helices, with peptide binding mediated by the sheet, intervening loops and one of the helices. The specific recognition of phosphotyrosine involves amino-aromatic interactions between lysine and arginine side chains and the ring system in addition to hydrogen-bonding interactions with the phosphate.

Figure 1: Target article for summarization.

Binding of **IFNgamma R** and **gp130 phosphotyrosine peptides** to the **STAT SH2 domains** was modeled by using the coordinates of **peptides pYIIPL (pY, phosphotyrosine)** and **pYVPML** bound to the **phospholipase C-gamma 1** and **v-src kinase SH2 domains**, respectively (#OTHER_CITATION, #TARGET_CITATION).

The ligand-binding surface of the **SH2 domain** of the **Lck nonreceptor protein tyrosine kinase** contains two pockets, one for the **Tyr(P) residue** and another for the **amino acid residue** three positions C-terminal to it, the +3 amino acid (#OTHER_CITATION, #TARGET_CITATION).

Given the inherent specificity of **SH2 phosphopeptide interactions** (#TARGET_CITATION), a high degree of selectivity is possible for **STAT dimerizations** and for **STAT activation** by different ligand-receptor combinations.

In fact, the **v-src SH2 domain** was previously shown to bind a **peptide pYVPML** of the **platelet-derived growth factor receptor** in a rather unconventional manner (#TARGET_CITATION).

Figure 2: Sample citances pointing to target article, with some key terms highlighted.