

PROTEIN NAME TAGGING FOR BROWSING SUPPORT, ACTIVE DATABASE CROSS LINKING, AND INFORMATION RETRIEVAL.

P. Lidén¹, L. Asker¹, G. Eriksson², K. Franzén² and F. Olsson²

¹*Virtual Genetics Laboratory AB; SE-171 77 Stockholm, Sweden;* ²*Swedish Institute of Computer Science; Box 1263, SE-164 29 Kista, Sweden;*

Whereas many applications of natural language processing for molecular biology focus on protein name tagging for the purpose high-level information extraction from large corpuses of scientific text, such as automatic identification of protein-protein interactions, high quality protein name tagging has a value in itself. The aim of this study was to design, implement, and evaluate a high-accuracy protein name tagger, and give proof-of-concept for some of the most basic applications of protein name tagging in an information retrieval setting, namely browsing support, active database cross linking, and enhanced query functionality. A combination of heuristics, dictionary look-up, syntactic analysis, and the application of a local dynamic dictionary were used to create a protein name tagger. This tagger outperforms a previously published similar system when benchmarked on a corpus of manually annotated Medline abstracts. In addition to evaluating the tagging performance, the implemented algorithm was used to add mark-up to a corpus of approximately 10000 Medline abstracts, which were indexed in a state-of-the-art information retrieval system. Indexing highlights many basic benefits of adding named entity mark-up such as protein names. One obvious benefit is that the search process is enhanced by the addition of a search field. Furthermore, the mark-up can be used for providing active hyperlinks between protein entities in presented documents and protein sequence databases, such as SwissProt, when both databases are indexed in the same information retrieval system. Efficient links can also be constructed in the opposite direction providing high precision retrieval of documents relevant for protein entries. Fast and accurate cross linking can be obtained by using an efficient implementation of the field based approximate cosine measure, which is a simple standard information retrieval technique for document similarity searching. This poster presents methods, results, implementation details, and features of a prototype system.