In Search of Protein Locations

Catherine Blake[1,2] **Wu Zheng**[1]
clblake@illinois.edu wuzheng2@illinois.edu
[1] Graduate School of Library and Information Science
[2] Computer Science and Medical Information Science
University of Illinois, Urbana Champaign, IL, USA

Abstract

We present a bootstrapping approach to infer new proteins, locations and protein-location pairs by combining UniProt seed protein-location pairs with dependency paths from a large collection of text. Of the top 20 system proposed protein-location pairs, 18 were in UniProt or supported by online evidence. Interestingly, 3 of the top 20 locations identified by the system were in the UniProt description, but missing from the formal ontology.

1 Introduction

Identifying subcellular protein locations is an important problem because the protein location can shed light on the protein function. Our goal is to identify new proteins, new locations and new protein-location relationships directly from full-text scientific articles. As with many ontological relations, location relations can be described as a binary predicate comprising two arguments, Location(X, Y) indicates that X is located in Y, such as Location (CIC-5, luminal membrane) from the sentence: ClC-5 *specific signal also appeared to be localized close to the* luminal membrane *of the intestinal crypt.*

Identifying protein subcellular locations has been framed as a classification task, where features include sequences, motifs and amino acid composition (Höglund, et al, 2006) and protein networks (Lee et al., 2008). The SherLoc system (Shatkay et al., 2007) includes text features the EpiLoc system (Brady & Shatkay, 2008) represents text from Medline abstracts as a vector of terms and uses a support vector machine to predict the most likely location for a new protein. Classification accuracy varies between species, locations, and datasets.

We take an alternative strategy in this paper and propose a bootstrapping algorithm similar to (Gildea & Jurafsky, 2001). The proposed system builds on earlier work (Zheng & Blake, 2010) by considering a larger set of seed terms and by removing syntactic path constraints.

2 Approach

The proposed bootstrapping algorithm is depicted in Figure 1. The system identifies lexico-syntactic patterns from sentences that include a given set of seed terms. Those the patterns are then used to infer new proteins, new locations, and new protein-location relationships. The system thus requires (a) an existing collection of known entity pairs that participate in a location relationship (called the seed terms) (b) a corpora of texts that report location relationships and (c) a syntactic path representation.

Our experiments use seed protein-location relationships from the UniProt knowledge base (www.uniprot.org). The complete knowledge base comprises more than 80,000 protein names for a range of species. The system uses the location and the location synonyms from the UniProt controlled vocabulary of subcellular locations and membrane topologies and orientations (www.uniprot.org/docs/subcell release 2011_2). The system also used a list of protein terms that were created by identifying words that immediately precede the word *protein* or *proteins* in the TREC collection. Two-thirds of the top 100 proteins in the TREC collection were used as seed terms and the remaining 1/3 were used to evaluate system performance.

The system was developed and evaluated using different subsets of the Genomics Text Retrieval (TREC) collection (Hersh, & Voorhees, 2009). Specifically 5533 articles in JBC 2002 were used for development and ~11,000 articles in JBC 2004 and 2005 were used in the evaluation.

The syntactic paths used the dependency tree representation produced by the Stanford Parser (Klein & Manning., 2003) (version 1.6.4).

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 101–102,
Portland, Oregon, USA, June 23-24, 2011.

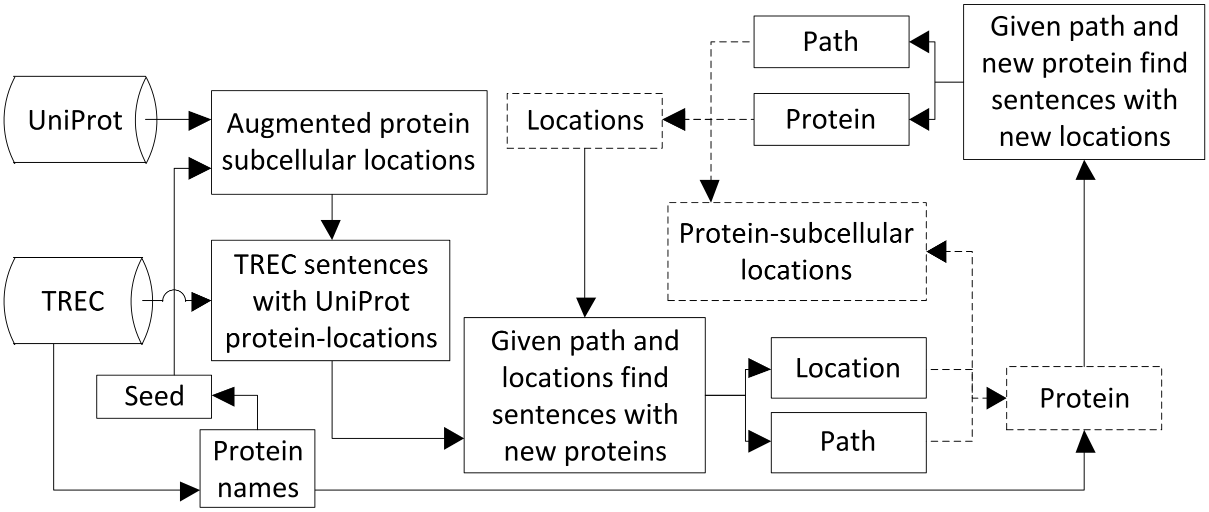
UniProt
TREC
Augmented protein subcellular locations
TREC sentences with UniProt protein-locations
Seed
Protein names
Locations
Given path and locations find sentences with new proteins
Location
Path
Protein
Path
Protein
Given path and new protein find sentences with new locations
Protein-subcellular locations

Figure 1 – The Bootstrapping approach used to generate new proteins, subcellular locations and protein location pairs. Inferred proteins and locations are depicted with a dashed line.

3 Results

The system identified 792 new proteins in the first iteration. All but 3 of the most frequent 20 proteins were in UniProt. All proteins in the test set were identified, but only 10 were in the top 100 proteins.

The system identified just over 1,200 new protein-location pairs after the first bootstrapping step. We evaluated the twenty most frequent pairs. Two erroneous proteins in the previous step caused two protein-location pair errors. UniProt reported 13 of the remaining 18 protein-location pairs. The five remaining pairs, were supported by online sources and in sentences within the collection.

The system identified 493 new locations after the second bootstrapping step and we evaluated the top 20. Sentences in the collection suggest that 9 of the new locations are in fact locations, but that they may not be subcellular locations and that 8 proposed locations are too general. Interestingly, 3 of the top 20 locations identified by the system are mentioned in the UniProt definitions, but are not included in the control vocabulary as a synonym, which suggests the need for automated approaches such as this to supplement manual efforts.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant IIS-0812522. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Brady, S., & Shatkay, H. 2008. EpiLoc: a (working) text-based system for predicting protein subcellular location., Pac Symp Biocomput (pp. 604-615).

Gildea, D., & Jurafsky, D. 2001. Automatic labeling of semantic roles. Computational Linguistics, 99(9): 1-43.

Hersh, W., & Voorhees, E. (2009). TREC genomics special issue overview. Information Retrieval, 12(1), 1-15.

Höglund, A., Dönnes, P., Blum, T., Adolph, H.W., & Kohlbacher, O. 2006. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics, 22(10):1158-1165.

Klein, D., & Manning., C.D. 2003. In Accurate Unlexicalized Parsing (pp. 423-430). Paper presented at the In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003).

Lee, K., Chuang, H.-Y., Beyer, A., Sung, M.-K., Huh, W.-K., Lee, B., et al. 2008. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. Nucleic Acids Research, 36(20), e136.

Shatkay, H., Höglund, A., Brady, S., Blum, T., Dönnes, P., & Kohlbacher, O. 2007. SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data Bioinformatics, 23(11), 1410-1417.

Zheng, W., & Blake, C. 2010. Bootstrapping Location Relations from Text. American Society for Information Science and Technology, Pittsburgh, PA.