

Protein Names Precisely Peeled Off Free Text

- Authors: Sven Mika and Burkhard Rost
- Affiliation: Department of Biochemistry and Molecular Biophysics, Columbia University
- Source: Bioinformatics, Vol.20 Suppl. 1 2004, pages i241-i247

Error in page i244, right column, the paragraph right below equation (2)

Refer to HTML version at

http://cubic.bioc.columbia.edu/papers/2004_nlprot/paper.html



Message



- Introduced a system (NLProt) that combines a pre-processing dictionary and rule based filtering step with 4 separately trained SVMs to identify protein names in the biology literature.
- The authors believe that their system can achieve very good performance, with a precision of 75% and recall of 76%
- The system provide a good structure for taking advantage of different approaches. However if remove the bias, performance decreases a lot.



Outline

- Introduction
 - Problems in extracting protein names
 - Related work
- NLProt system design
- Testing and results
- Comments and discussion

Introduction

- **Problem setting:**

Information Extraction – Named Entity Recognition

- **Problems in extracting protein names**

Lack of common standards and fixed nomenclatures: the same entity may be referred to different names, the same name may refer to several different entities

Variant structural characteristics: maybe extremely short or extremely long, lack of explicit marking, common inclusion of modifiers in the name

Unclear status as name: often derive from regular noun phrases



Past approaches: Pros and Cons

- **Dictionary-based:** requires a lot of human labor, can't tag novel names, easy to use, high precision
- **Rule-Based:** Require a lot of human analysis, easy to support and expand
- **Machine Learning:** easy to tune to new domains, performance is low
- **Combination**



NLProt system design: overview

- Dictionary of protein names
- Filtering: common words and chemical compounds
- 4 different SVMs



System design: dictionary



- ◆ SWISS-PROT + TrEMBL
- ◆ Use this protein name dictionary to derive an input value for SVM4



System design: filtering

- ◆ Dictionary of common words: Webster dictionary, Dictionary of medical terms, species names, tissue types
Use this common dictionary for filtering in first step
- ◆ Chemical names
Remove chemical compounds based on a list of endings

System design

- ◆ Generating samples

environment 1				centre				environment 2			
a 6-fold decrease in high mobility group protein (HMG) could											
(1)	(2)	(3)	(4)	(A)	(B)	(C)	(E)	(5)	(6)	(7)	(8)
human Rad51 amino acid residues required for Rad52 binding											
(1)	(2)	(3)	(4)	(A)	(E)	(5)	(6)	(7)	(8)		

System Design

◆ Filtering samples

Set of regular expressions

Name is followed by 'cell(s)' or 'cyte(s)'

Name ending similar to chemical compound (list of 130 4-letter endings)

Name is in common dictionary

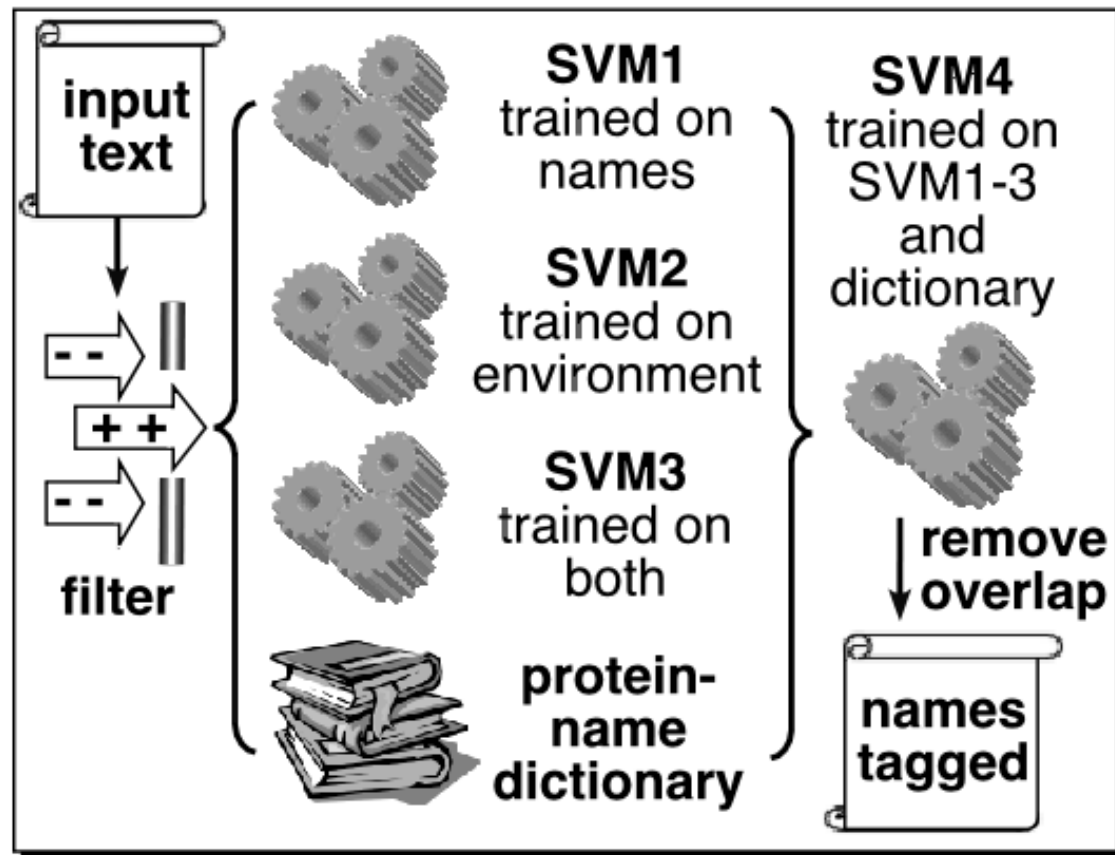
Name seems to be an author

Name is in parentheses following a filtered out word

Name is number followed by noun in plural form

System Design

- ◆ SVMs



Performance measure

- ◆ F-measure with equal weight on recall and precision: $2 * A * C / (A+C)$
- ◆ Precision (accuracy): $A=TP/(TP+FP)$
- ◆ Recall (coverage): $C=TP/(TP+FN)$
- ◆ Bias: reducing redundancy



Results

- ◆ Overall $F=75\%$
- ◆ Without bias $F= 50\% \sim 60\%$



Actual system



- ◆ NLProt

<http://cubic.bioc.columbia.edu/services/NLProt/submit.html>

- ◆ Yapex

<http://ellis.sics.se:8080/cgi-bin/Yapex/yapex.cgi>

Comments and Discussion questions

- ◆ Praise: provide a good structure to take advantage of different approaches using a integrated system
- ◆ Critics: After reducing the redundancy (Bias), performance is not that good
- ◆ Might include rules for identifying new protein names (improve coverage on novel names) , not just for filtering (improve accuracy)
- ◆ Vector construction: arbitrary number of frequent tokens, arbitrary weight on locations



Comments and Discussion questions

- ◆ Question: what could be done to improve the system?
 - Compiling a more comprehensive dictionary of protein name?**
 - Identifying working rules for tagging protein names?**
 - Try other machine learning techniques?**



References

- ◆ NLProt: extracting protein names and sequences from papers S Mika and B Rost 2004 Nucleic Acids Res. 2004 32(Web Server issue):W634-W637;
- ◆ Protein names precisely peeled off free text S Mika and B Rost 2004 Bioinformatics. 2004 Aug 4;20 Suppl 1:I241-I247;
- ◆ Protein names and how to find them. Franzen K, Eriksson G, Olsson F, Asker L, Liden P, Coster J. Int J Med Inform. 2002 Dec 4;67(1-3):49-61