Hans Journal of Computational Biology 计算生物学, 2011, 1, 5-10
http://dx.doi.org/10.12677/hjcb.2011.12002 Published Online December 2011 (http://www.hanspub.org/journal/hjcb)

Research of Protein Named Entity Recognition Based on SVMs*

Lejun Gong[1,2], Yaxing Fu[1], Xiao Sun[1], Jianming Xie[1], Shuangxin Yu[1]
[1]Department of Biological Science and Medical Engineering, Southeast University, Nanjing
[2]Faculty of Computer Engineering, Huaiyin Institute, Huai'an
Email: glj98226@163.com
Received: Sep. 8th, 2011; revised: Oct. 19th, 2011; accepted: Oct. 23rd, 2011.

Abstract: This paper describes an approach to identify protein named entity using Supports Vector Machines (SVMs), and selects four groups of features to do experiments for the protein corpus. Experiment results show the system performance of context features increases smaller than baseline system, and the combined feature of part of speech (POS) and word type is achieved 78.43% accuracy which is the best performance in all experiments. The research results show the combined feature of POS and word type play important roles in the protein entity recognition.

Keywords: Supports Vector Machines (SVMs); Protein Entity Recognition; Feature Selection

基于支持向量机的蛋白质命名实体识别的研究*

龚乐君[1,2]，付亚星[1]，孙　啸[1]，谢建明[1]，于双鑫[1]
[1]东南大学生物科学与医学工程学院，南京
[2]淮阴工学院计算机工程学院，淮安
Email: glj98226@163.com
收稿日期：2011 年 9 月 8 日；修回日期：2011 年 10 月 19 日；录用日期：2011 年 10 月 23 日

摘　要：发展一种利用支持向量机识别蛋白质命名实体的方法，选择四组特征对蛋白质语料进行识别实验。实验表明，与基线系统相比，上下文特征有较小的增幅，而当前词的词性及词形的组合特征获得了最好的性能，达到 78.43%的准确率。这一研究结果显示词性及词形特征在蛋白质实体识别中起着重要的作用。

关键词：支持向量机；蛋白质实体识别；特征选择

1. 引言

生命科学和技术的发展促进产生各种生物信息，而大量的生物信息散布在各种文献中，以文本的形式呈现。对这些生物医学文献进行加工和集中处理，可以从中提炼出更多的生物信息[1]。生物医学命名实体识的任务则主要是从生物医学文献中抽取生物医学实体[2]，例如，蛋白质、基因、DNA、RNA、疾病、化合物、药物名称等。这些实体的识别将对进一步发现它们之间的联系及相互作用有着非常重要的意义。

蛋白质是生命机器，是一类非常关键的生物医学实体。蛋白质参加绝大部分的生命活动，在生命活动过程中扮演极其重要的角色。蛋白质是基因功能的执行者，蛋白质机器运转失常会引发机体功能障碍，从而导致疾病。因此，生命科学中大量的生物医学文本与蛋白质关联，识别生物医学文本中的蛋白质名称是命名实体的主要研究任务。

当前，命名实体的研究的方法大致有三类，基于规则的方法、基于字典的方法及基于机器学习的方法。基于规则的方法需要领域专家建立规则库，而基于字典的方法存在着实体名称冲突和覆盖率不足的缺陷。随着语料库标注的迅速发展，基于机器学习的方法也

*基金项目：国家自然科学基金(60971099)。

迅速发展起来，本文研究如何将支持向量机(SVMs)[3]学习模型和蛋白质命名实体特征分析结合起来，发展蛋白质实体识别方法。

2. 材料与方法

2.1. 数据集

本文的实验数据来源于 AIMed(ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/proteins)中的蛋白质语料，采用 84 篇文摘作为测试数据。

2.2. 蛋白质实体表示形式

生物医学文献中的蛋白质实体表示形式极其复杂。这些复杂性表现在实体既有单个单词形式的实体，单词的长度不一，并且大写和小写杂合在一起，例如：urokinase，Cactus，IkappaBalpha 等；也有多个单词组成的短语，例如：bradykinin B(1) receptor，Protein phosphatase 2A，这给蛋白质实体边界的确立带来了很大的困难；有些相同的词或者短语又可以表示不同类别的生物医学实体，例如 c-myc，IL-2 即可表示蛋白质，也可表示基因，要通过上下文才能判别出来；有些蛋白质实体拥有多个不同的书写形式，例如：Protein phosphatase 2A，protein phosphatase 2A，protein phos-phatase 2A 等表示同一蛋白质实体；文本中的蛋白质实体的缩写词占有很大的比重，如 PP2A(表示 protein phosphatase 2A 蛋白质实体)，PKC(表示 protein kinase C 蛋白质实体)。生物医学文献中的蛋白质实体由于存在这些复杂的表示形式，使蛋白质实体的识别成为富有挑战性的一项研究。

2.3. 蛋白质实体特征分析

本文针对蛋白质实体识别所面临的困难及所表现出的特性，构建一个基于机器学习的蛋白质识别系统。构建该系统时选择合适的实体分类特征集合非常重要，我们主要通过对训练语料的语言信息进行统计和分析，从训练语料中提取蛋白质实体特征。所选用的特征主要有以下几种：1) 单词特征；2) 词形特征；3) 词性特征；4) 上下文特征。

1) 单词特征

将生物医学文本中单词本身作为特征向量的组成部分。单词是文本的基本组成部分，同时也是组成生物医学实体的基本成分，有些单词本身就是蛋白质的名称。因此，将当前单词作为特征是符合客观问题实质的。从实验语料中抽取单词组成词汇表从而单词特征可用式(1)的正交编码的形式来表示。

$$\text{word}_i = \begin{cases} 1 \text{ 如果该词在词汇表} \\ \quad \text{中第} i \text{个位置} \\ 0 \text{ 其它} \end{cases} \tag{1}$$

2) 词形特征

由于蛋白质名称多数含有数字、大写字母、特殊符号等，将这些特征作为表面线索识别蛋白质实践证明有着较好的效果。在本文中主要使用表 1 所示的词形特征。采用的也是正交编码的形式以 15 位二进制数唯一标示某一个词的词形如式(2)所示。

$$\text{wordType}_i = \begin{cases} 1 \text{ 如果该词的词形在词形表} \\ \quad \text{中第} i \text{个位置} \\ 0 \text{ 其它} \end{cases} \tag{2}$$

3) 词性特征

在一般情况下，蛋白质实体可能是一个名词短语，这些短语包括的词性类别可以分为 NN(名词)，NNS(名词复数形式)，CC(连接词)，JJ(形容词)，IN(介词)等。

Table 1. Word type feature
表 1. 词形特征

序号	词形特征	举例说明
1	ALL_Captial	DNA
2	Init_Captial	P53
3	InitCaptialSecondLower	Cdc28
4	Letters	DNA,hairy
5	Include_Digit	IL-2, Cdc28
6	OneDigit	8,5
7	TwoDigit	53
8	Natural_Number	521
9	Letter_Digit	HIV-1
10	Include_hyphen	IL-2
11	GreekLetter	alpha, beta, kappa
12	Tail_hyphen	Receptor-
13	Initial_hyphen	-mediated
14	punctuation	, .:…
15	RealNumber	80.1

词性特征在识别生物医学命名实体能够提供更有帮助的信息，本研究中采用了斯坦福的词性标注器 stanford-postagger[4]对文本语料进行词性标注，该词性标注器采用了最大熵的模型进行词性标注，在生物医学文本词性标注中具有较高的性能。本文通过组建词性表，采用式(3)的 36 位二进制正交编码来唯一表示某一个词的词性。词性特征如表 2 所示：

$$\text{partOfSpeech}_i = \begin{cases} 1 & \text{如果该词的词性在词性表中第}i\text{个位置} \\ 0 & \text{否则} \end{cases} \quad (3)$$

4) 上下文特征

上下文特征是基于物以类聚的思想，考虑的是蛋白质实体的存在可能跟它前面的词或后面的词的特征有关。本文中上下文特征嵌入其它特征的表示法中，组成单词上下文特征、词形上下文特征、词性上下文特征。例如采用了上下文的单词特征正交编码的形式可用公式(4)来定义：

$$\text{wordContext}_{ki} = \begin{cases} 1 & \text{如果该词在}k\text{位置且在词汇表中属于第}i\text{个位置} \\ 0 & \text{其它} \end{cases} \quad (4)$$

其中，k 定义为与当前词相关的上下文词的位置，取负值表示上文，取正值表示下文，等于零时为当前词。本研究中，上下文窗口定义为 3，$k \in \{-1,0,1\}$，即当前词的前一个词、当前词及当前词的后一个词。词性上下文、词形上下文的定义与此类似。

2.4. 系统框架

本文采用了基于语料库的机器学习法进行蛋白质实体的识别，文本在投入到分类器中之前，需要进行预处理。首先采用启发式规则过滤文本中与蛋白质实体分类特征无关的符号，并对文本进行分句、分词、词性标注，抽取相应的单词信息、词形信息、词性信息，针对实验样本数据把这些信息汇集组成相应的表目，并做好标记。预处理完成后，就可以针对实验数据进行特征选择、特征抽取，生成特征文件，再把该特征文件投入到分类器中进行学习或预测。

支持向量机是近年来广泛使用的机器学习方法，已经成功应用于许多自然语言问题，如基本的短语块的识别[5]、词性标注[6]、命名实体识别。它有以下几个优点：a) 可以解决样本有限情况下的机器学习问题，目标是得到现有情况下的最优解；b) 算法最后转化为二次型寻优问题，得到全局最优解，可以避免神经网

Table 2. Part-of-Speech feature
表 2. 词性特征

序号	词性	描述	序号	词性	描述
1	CC	Coordinating conjunction	19	PRP$	Possessive pronoun
2	CD	Cardinal number	20	RB	Adverb
3	DT	Determiner	21	RBR	Adverb, comparative
4	EX	Existential there	22	RBS	Adverb, superlative
5	FW	Foreign word	23	RP	Particle
6	IN	Preposition or subordinating conjunction	24	VBP	Verb, non-3rd person singular present
7	JJ	Adjective	25	TO	to
8	JJR	Adjective, comparative	26	VBZ	Verb, 3rd person singular present
9	JJS	Adjective, superlative	27	VB	Verb, base form
10	LS	List item marker	28	VBD	Verb, past tense
11	MD	Modal	29	VBG	Verb, gerund or present participle
12	NN	Noun, singular or mass	30	VBN	Verb, past participle
13	NNS	Noun, plural	31	SYM	Symbol
14	NNP	Proper noun, singular	32	UH	Interjection
15	NNPS	Proper noun, plural	33	WDT	Wh-determiner
16	PDT	Predeterminer	34	WP	Wh-pronoun
17	POS	Possessive ending	35	WP$	Possessive wh-pronoun
18	PRP	Personal pronoun	36	WRB	Wh-adverb

络结构选择和局部极值问题；c) 算法将实际问题通过非线性变换到高维特征空间，在高维空间中构造线性判别函数解决非线性问题，可以提高泛化性能；d) 通过对二类问题的推广，可以解决多类分类问题。针对上述特点本文选用支持向量机作为分类器进行蛋白质实体识别的研究，系统框架如图 1 所示。

2.5. 蛋白质实体识别多分类分析

蛋白质实体的识别问题可以看作为分类问题，输入的是一组词序列，例如文本中的词序列“p38 stress-activated protein kinase”可用$\{w_i\}(i = 1,\cdots,n)$表示，针对文本中词序列中的文本符号 w_i 分配一个预先定义的分类标签 t_i，学习系统的任务就是预测每一个文本符号 w_i 的分类标签 t_i。由于蛋白质实体在文本中表示的复杂性，常以多词短语的形式出现，为确定蛋白质实体的边界，本文采用的 BIO[7]表示法对蛋白质实体进行分类，可以更好的区分蛋白质的边界。其中 B 表示蛋白质开始部分，I 表示蛋白质的中间部分，O 表示非蛋白质实体。词序列$\{w_i\}(i = 1,\cdots,n)$例如“p38 stress-activated protein kinase inhibitor reverses bradykinin B(1) receptor”可用分类标签 t_i(B、I、O)进行分类所对应的结果如图 2 所示。

本文使用了 BIO 表示法对蛋白质实体进行分类涉及了三类情况，标准的 SVMs 分类器只是针对两类样本进行分类，解决这个问题构建一个多类支持向量机通常有两种方法，一种是 one-vs-rest，另一种是 one-vs-one。

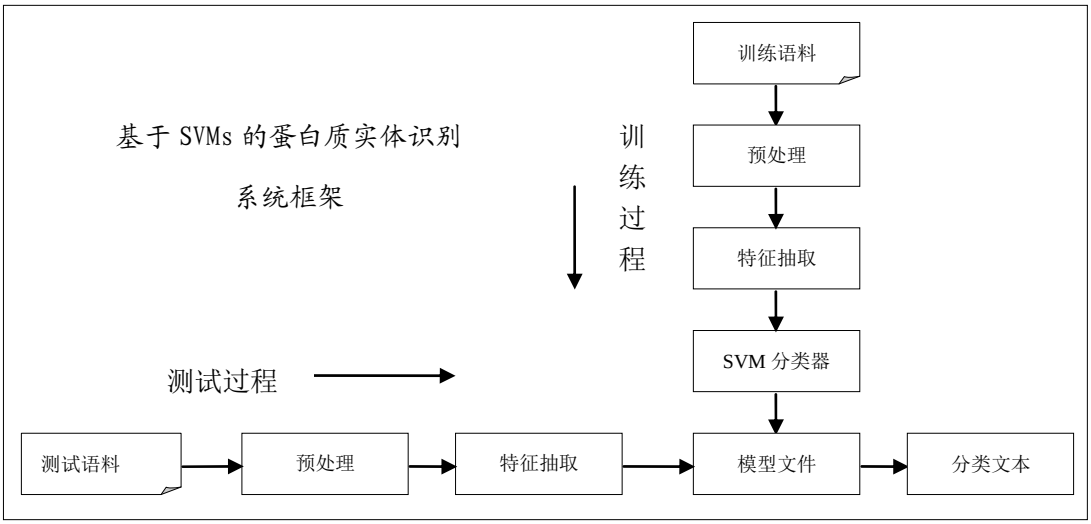

Figure 1. System architecture of protein entity recognition based on SVMs
图 1. 基于 SVMs 的蛋白质实体识别的系统框架

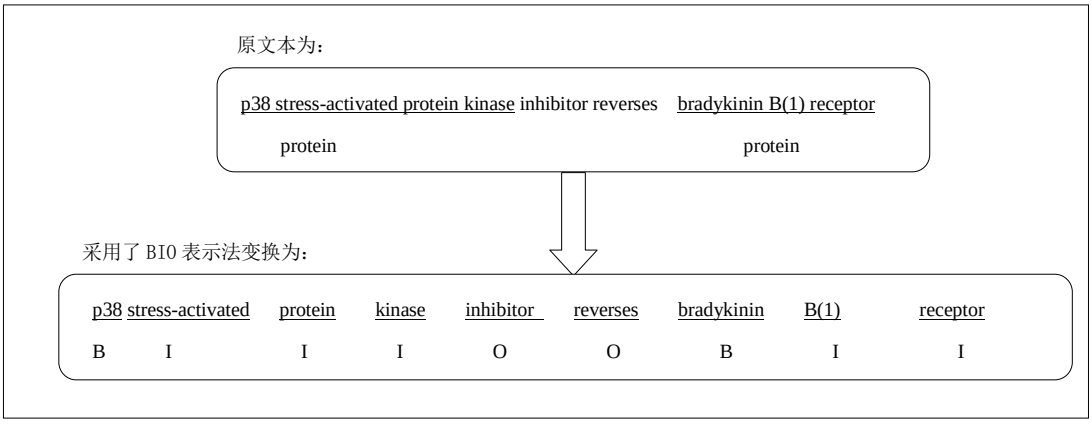

Figure 2. Boundary determination of protein entity by BIO format
图 2. 使用 BIO 确定蛋白质实体边界

one-vs-rest 对于 k 类的问题，将其中某一类的 n 个训练样本视为一类，其他训练样本归为一类，这样就需要构建 k 个二元 SVMs 分类器 $f_1, f_2, \cdots, f_k$。每个测试样本 x 都利用这 k 个分类器进行分类，得到 k 个函数值 $f_1(x), f_2(x), \cdots, f_k(x)$，识别类别为 $\hat{k} = \arg\max_k \left(f_1(x), f_2(x), \cdots, f_k(x) \right)$。

one-vs-one 对于 k 类问题，为每两类的组合构造一个分类器，这样共有 $K = k(k-1)/2$ 个分类器，采用投票机制对每一类分别打分投票 $v_1, v_2, \cdots, v_k$，每个测试样本 x 分别经 K 个分类器 SVM_{ij} 进行识别，如其属于第 i 类，则 $v_i = v_i + 1$，否则 $v_j = v_j + 1$，识别类别为得票值最多的一类即 $\hat{k} = \arg\max_k \left(v_1(x), v_2(x), \cdots, v_k(x) \right)$。

两种方法比较，据文献[8]报告 one-vs-one 比 one-vs-rest 效果更好。实验中采用的工具包 libSVM[9] 使用了 one-vs-one 解决多分类问题。

3. 结果与讨论

我们从该实验数据抽取出 695 个句子，936 个蛋白质，18,768 个单词的，形成词汇表含有 4330 个词汇。实验中使用十倍交叉验证法，采用准确率(ACC)来衡量系统的性能，其定义如式(5)，为使实验结果简洁明了，本文针对 2.3 节分析的蛋白质实体特征作标记(见表 3)，实验结果见表 4 所示。

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

其中 *TP* 为正确的肯定的分类标记数，*FN* 为系统错误的否定分类标记数，*FP* 为错误的肯定的分类标记数，而 *TN* 为正确的否定的分类标记数。

Pre_class 代表当前词之前一个词分类结果，体现当前词的分类与上文的词类特征有关，例如，如果当前词的分类标签为 B，那么该词的前面第一个词的分类标签一定为 O。由于当前词及下文的词类标记是将要预测的项目，因此本研究中只考虑上文的词类特征。本研究中以特征组合 Cur_lex + Cur_shape + Cur_loc 作为基线系统(baseline)，其它特征组合及实验结果如表 4 所示。

通过上面一系列的实验可知，基线系统性能为 69.70%的准确率，单组的词性特征及词形特征都远远高于基线系统，而单词特征稍低于基线系统；词性与词形的组合特征获得了实验中最好的效果达到了 78.43%的准确率，词形与单词的组合特征稍低于基线系统，而词性与单词的组合特征稍高于基线系统。结果表明，当前词的词性特征对系统的性能起着决定性的作用，其次是词形特征，单词特征最弱，这可通过两方面确立，第一，单组特征中词性特征性能最高，词形次之，最后是单词特征，第二，组合特征中词性与词形特征

Table 3. Feature token
表 3. 特征标记

上下文(窗口大小为 3)	单词特征	词形特征	词性特征	词类特征
Word_Pre (上文)	Pre_loc	Pre_shape	Pre_lex	Pre_class
Word_Cur (当前词)	Cur_loc	Cur_shape	Cur_lex	-
Word_Suf (下文)	Suf_loc	Suf_shape	Suf_lex	-

Table 4. Performance of all the features of the system
表 4. 系统各项特征及其组合的性能

特征项目	ACC (%)	特征项目	ACC (%)	特征项目	ACC (%)
baseline	69.70	⊕Pre_class	69.70	⊕Word_Pre⊕Suf_loc	69.76
↓Cur_lex	69.66	⊕Pre_lex	69.70	⊕Word_Pre⊕Suf_shape	69.77
↓Cur_shape	69.74	⊕Pre_loc	69.70	⊕Word_Pre⊕Suf_lex	69.76
↓Cur_loc	78.43	⊕Pre_shape	69.72	⊕Word_Pre⊕Suf_lex⊕Suf_loc	69.77
Cur_lex	76.29	⊕Pre_shape⊕Pre_class	69.72	⊕Word_Pre⊕Word_Suf	69.74
Cur_shape	72.67	⊕Pre_class⊕Pre_shape⊕Pre_lex	69.74		
Cur_loc	69.64	⊕Word_Pre	69.76		

“↓”表示在基线的基础上减去相应的特征；“⊕”表示在基线的基础上加上相应的特征；Word_Pre 表示上文的所有特征；Word_Suf 表示下文的所有特征。

的组合获得了最好的效果，但是词形与单词特征的组合稍低于基线系统的性能。单词特征不明显究其原因是没有采用停用词特征，出现较多的冗余信息，致使单词本身的特征信息不显著。这一问题将在后续的工作中予以解决。

上下文特征的组合中，加入上文的特征性能稍弱于加入上下文特征的性能，上下文特征组合的性能与基线系统相比，没有得到大幅度的提高，而是缓慢增长，这一结果表明，上下文特征对系统的性能的提高起着一定作用，增幅不大的原因是上下文窗口过小，致使该特征不明显，后续工作将调整上下文的窗口，增大窗口大小使该特征显著提高系统的性能。

4. 结论

本文采用支持向量机针对蛋白质实体识别进行研究，特征选择主要采用了单词特征、词性特征、词形特征、上下文特征。实验中采用了 AIMed 语料中的蛋白质语料，实验结果表明单组特征中单词词性特征的效果最好，词形次之，单词本身的特征最弱；特征选择中词性与词形特征组合达到 78.43%的准确率，这也是实验取得的最好效果；上下文特征对基线系统的性能有较小幅度的提高，与基线系统相比，性能增幅不大，主要原因是上下文窗口设置过小，使该特征不明显；本文的研究表明词性及词形特征在蛋白质的识别中起着重要的作用。下一步的工作将是启用停用词特征、增加上下文窗口的大小、增加生物医学本体这一外部特征，将机器学习与生物医学本体技术有机结合起来进一步提高蛋白质实体识别的性能。

5. 致谢

本文由国家自然科学基金(60971099)资助。

参考文献 (References)

[1] P. Zweigenbaum, D. Demner-Fushman, H. Yu, et al. Frontiers of biomedical text mining: Current progress. Brief Bioinform, 2007, 8(5): 358-375.

[2] U. Leser, J. Hakenberg. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform, 2005, 6(4): 357-369.

[3] J. Kazama, T. Makino, Y. Ohta, et al. Tuning support vector machines for biomedical named entity recognition. In: Proceedings of the Workshop on Natural Language Processing in the Bio-Medical Domain at ACL, 2002: 1-8.

[4] K. Toutanova, C. D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), 2000: 63-70.

[5] T. Kudo, Y. Matsumoto. Use of support vector learning for chunk identification. Proceeding ConLL'00 Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning, 7: 142-144.

[6] T. Nakagawa, T. Kudoh and Y. Matsumoto. Unknown word guessing and part-of-speech tagging using support vector machines. In Proceeding of the 6th NLPRS, 2001: 325-331.

[7] L. A. Ramshaw, M. P. Marcus. Text chunking using transformation-based learning. In Proceedings of the ACL Third Workshop on Very Large Corpora, 1995: 82-94.

[8] C. W. Hsu, C. J. Lin. A comparison of methods for multiclass support vector machines. IEEE Transaction on Neural Networks, 2002, 13(2): 415-425.

[9] C. C. Chang, C. J. Lin. Training un-support vector regression: theory and algorithms. Neural Computer, 2002, 14(8): 1959-1577.

