阅读总结：

（1） 本文主要介绍了一种蛋白质间相互作用的信息抽取方法，并没有涉及到信息学中的互信息。

（2） 本文浅层句法分析以及句子结构分析的作用是为了去掉冗余信息，以此提高抽取的准确率。

（3） 本文的贡献就是提出了一种去掉冗余信息的方法（浅层句法分析+句子结构分析），采用的信息抽取的核心方法还是最大熵机器学习算法。

（4） 本文在并列结构分析时采用了有限状态自动机进行分析，这个方法用于文本处理很不错，值得学习。

融合浅层句法分析的蛋白质互作用信息抽取方法

钱伟中，王娟，傅翀，秦志光

电子科技大学 计算机科学与工程学院

计算机应用研究 2011 年 3 月

摘要：针对传统基于机器学习方法在蛋白质互作用信息抽取中的缺陷，提出融合浅层句法分析的信息抽取方法，该方法将候选的句子进行浅层句法分析，包括：短语切分、同位语切分、并列结构分析、句子切分。经过该步骤，句子被划分为多个单独的语法单元，然后对每个语法单元采用基于最大熵的分类方法进行蛋白质互作用信息抽取。该方法在 BC-PPI 语料库中获得了 62.1%的 F1 性能。比较实验结果标明，该方法能有效减少误判和漏判，提高信息抽取的性能。

蛋白质互作用（PPI）是生物医学信息抽取研究的重要内容。

抽取 PPI 对的方法性能主要受如下三个因素影响：

A、 蛋白质实体提取方法的性能；（实体识别）

B、 上下文特征选择；（特征提取）

C、 复杂句子结构对性能的影响。（同位语从句、定语从句、并列句等）

PPI 对抽取方法主要有三类：

A、 基于自然语言处理的方法（定语语法模式——需要大量语义资源，抽取性能较低）

B、 基于模式匹配的方法（提取 PPI 文本模式——需要大量文本模式，降低系统性能）

C、 基于机器学习的方法（不依赖于具体句子模式——具有较强泛化性）

蛋白质互作用对的定义：三元组（protein1，iword，protein2）。其中 protein1，2 为蛋白质名，iword 为两者的交互词。

1.1 分析过程

（1）输入候选句

（2）对句子进行词性分析（MedPost 词性分析器）

（3）对句子的词性分析结果进行浅层句法分析，划定短语边界（Ramshaw 等人提出的基于转换的浅层句法分析器）

（4）对句子结构进行同位语规则分析

（5）对句子结构进行并列结构分析

（6）对句子结构进行从句结构分析

1.2 同位语分析

模式 1 NC1 NC2

模式 2 NC1 Keywords NC2（本文 Keywords=such as|including）

1.3 并列结构分析（有限状态自动机）

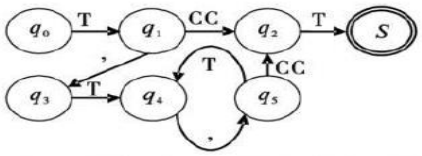

图1　并列结构识别的有限自动机

其中：q_0 为起始状态，S 为终止状态，状态之间的有向连接线代表状态迁移，状态线匹配的标志为集合 $S=\{T, CC, ', '\}$ 中的元素。其中，T为任意的名词短语，CC为连词。如果待匹配的名词短语序列具有一条从起始状态至终止状态的连接，则表明该序列满足名词并列结构条件。

1.4 从句结构分析（利用限定词 that、which 等）

思想：本文提出的融合浅层句法分析的信息抽取方法主要思想是在进行信息抽取前，将句子进行语法分析，去**掉冗余信息，缩小搜索范围**，以提高文本抽取性能。

在生物文献蛋白质互作用对抽取任务中，蛋白质互作用信息抽取是一个二元分类任务。任意的机器学习算法,如支持向量机、最大熵模型等，都能用于蛋白质互作用对的识别。

最大熵算法的词法特征

本文采用的词法特征如表 2所示。

表 2　词法特征

特征名	描述
Pr	蛋白质名称中的词
Bw	蛋白质对之间的词
W f	第一个蛋白质名称之前三个词
W a	第二个蛋白质名称之后三个词
K	交互词及位置