硕 士 学 位 论 文

基于模板核和扩展特征的蛋白质关系抽取

Template and Expand Features for Extracting Protein-Protein Interaction

作 者 姓 名：刘 昊

学科、 专业：计算机系统结构

学 号：20809426

指 导 教 师：王 健

完 成 日 期：2011 年 5 月

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

摘　　要

以往的蛋白质关系抽取系统主要提取出能够标示是否存在关系的典型特征，同时引入经典的已分类特征构建出属于自己的 PPI 系统。这样的系统在改善分类效果的同时，对特征的提取、分析、融合过程要花费大量时间，在实际应用中的分类效果并不明显。因此需要在原系统基础上考虑如何才能在不牺牲精度前提下建立优化的特征集，以此提高系统抽取的效率。

本文采用了一种优化特征的方法进行蛋白质关系抽取，其主要思路是考虑到目前的主流蛋白质关系抽取系统存在如下缺点：

（1）抽取特征较多，且提取过程各不相同，如何进行组合最为恰当还没有提出具体方法，有的时候两个表现优异的特征进行融合后反而会降低整体抽取精度。

（2）有效特征的组合大大降低了系统效率，对系统效率的分析和细节的精度调整并没有进行详细的叙述。

（3）实际抽取蛋白质关系的测试集中可能存在训练集中不包含的全新蛋白质关系和特征，单纯引入训练集的特征训练并不全面，要考虑在庞大语料集中提取更泛化的特征。

针对上述问题本文试图通过对主流特征的筛选和优化来对蛋白质关系抽取系统进行构建，在借鉴前人研究基础上，引入了一种新的模板特征方法，模拟人工标注原理进行抽取词序列模板，通过多个核的融合共同提高分类效果。在原有特征提取基础上构思一种扩展特征的方法以提取精简优化特征向量，该方法能够在扩展语料集中自动找到最符合蛋白质关系抽取标准的合适特征群落，将这一特征群落进行标准化，在保持原有实验精度的前提下大大提高了关系抽取效率。

该方法能够很好的解决图核方法不能处理复杂长难句的不足，通过模板快速匹配出分类结果，同时引入了句法分析和关键词标准化减少句子中的冗余词噪音，在庞大语料集中提取的扩展特征辅助提高了实验效率。在图核、模板特征和扩展特征的共同作用下，实验在 Aimed 语料的实验 F 值达到了 63.1%。

关键词：蛋白质关系抽取；模板；图核；句法分析器；扩展特征

Template and Expand Features for Extracting Protein-Protein Interaction Abstract

Before protein relationship extraction systems mainly extract the typical characteristics that indicate the existence of relations, while introducing the classic characteristics that have been classified to build their own PPI systems. In the same time of improving the classification results, such a system consumes a lot of time to complete feature extraction, analysis and integration process, and in practical applications the effect of the classification is not obvious. Therefore, how to establish the feature set based on the original system without sacrificing accuracy is needed to be considered, in order to improve the extraction efficiency of the system.

In this paper, a feature optimization technique is applied to extract protein-protein interaction. The main idea is based on the consideration of following disadvantages of current methods:

(1) Redundant features are extracted with various methods, and how to combine the most appropriate method has no specific, sometimes fusion of two outstanding methods unexpectedly reduces the overall extraction accuracy.

(2) Effective combination of features greatly reduces the system efficiency; the details of adjustment and analysis of system efficiency and accuracy are not described.

(3) Test set may contain absent features in train set, so features in train set is not comprehensive. Generalized features should be extracted in a more wide range corpus.

In response to these problems, the paper attempts to build protein relation extraction system by sifting and optimizing the mainstream feature. Drawing on the previous studies, it introduces a new template features methods to simulate artificial mark principle for extracting word sequence template, and to improve classification results by multiple nuclear fusion. Based on the original feature extraction, design a kind of expansion characteristics of method to extract vector optimization of concise features. In the original feature extraction based on an extended characteristics of the design method to extract concise characteristic vector optimization, the method can automatically find appropriate feature community which most conforms to the standard of protein relation extraction. Standardizing the features community, and in the premise of maintaining the original experiment precision, it greatly improves the efficiency of the relationship extraction.

This method can perfectly solve the shortage of dealing with the complex long sentence, by classified results of the matched template. Meanwhile, it can reduce the noise of redundancy in sentence to improve the effectiveness of the experiment through the semantic analysis and

standard key words. With the co-work of the graph kernel, templates features and extend features, the F value reach 63.1% in the corpus of Aimed.

Key Words：extracting the protein protein interactions; template; graph kernel; syntactic analysis parser; Additional features

目　　录

1 绪论

1.1 研究背景

当前，随着基因制药，人类基因组识别，致病基因的研究日趋深化，生物医学领域的蛋白质种类以及相应的生物医学文献数目日益增多，如何才能准确定位新发现的蛋白质和蛋白质之间，蛋白质和疾病名之间，疾病名和基因之间的关系成为生物医学领域的重要研究内容。其中蛋白质关系抽取是生物医学自然语言处理领域的主要研究内容，现阶段通过人工方法对蛋白质关系标注很显然已经不能满足医学工作者的实际需要，因此必须找到合适的方法利用计算机自动抽取出生物医学文献中的蛋白质关系，进而深入发掘出蛋白质与蛋白质之间、蛋白质和疾病之间，疾病和基因之间的相互关系并根据新发现的关系构建关系网络，为下一步蛋白质关系聚类打下基础。

目前主流的蛋白质关系抽取系统有很多，其抽取的文本类型和对象各不相同，大致可以分为从文本角度进行抽取，从已标注蛋白质关系中聚类抽取，从生物临床实验中统计汇总三类。从文本角度抽取主要是从新出版和已出版的生物医学文献题目、摘要、全文中抽取蛋白质等相互作用关系，抽取的信息主要包括词特征、分析树特征、图特征等等，在人工监督或半监督抽取特征的基础上构建生物信息网络并发掘有价值信息。从已标注蛋白质关系中聚类抽取新的作用关系类似于生物医学网络聚类，其原理是依据能够标示蛋白质关系的重要特征常常聚类出现，考虑如果将以判明存在关系的蛋白质和未判明关系的蛋白质进行聚类构建蛋白质关系网络，则可以通过网络中的传递依赖关系找到最可能存在关系的蛋白质关系组，该方法最大的特点是泛化能力强，可以在标准数据中最大限度发掘出有价值蛋白质关系，但该系统所得到的蛋白质关系是否是真实情况下的蛋白质关系有待专家进行实验，而且大多数的关系网络过于稀疏，不能反映出蛋白质之间作用信息的全貌，要想反映信息全貌则要牺牲时间代价进行复杂网络的构建。还有一种最直接的蛋白质关系获取系统基于临床的信息获取，该系统每隔一段时间人工监督或半监督的获取新发现蛋白质及其在临床上的实时数据，通过不断更新维护关系数据库来维持蛋白质关系数据的健壮性和完整性，该系统由于现阶段获取信息受版权和隐私保护而难以推广，且维护造价昂贵，但构建的蛋白质关系最精确完整。目前使用最广泛的蛋白质关系自动抽取是基于生物医学文献文本中的监督或半监督抽取，本文将针对如何从生物医学文献中自动抽取蛋白质关系进行主要论述。

1.2 蛋白质关系抽取研究现状

蛋白质关系抽取研究进行至今，概括可分为词共现发展阶段和机器学习发展阶段。词共现方法主要依据包含蛋白质关系的句子中的特征词出现的概率、次数、频度、种类来统计分析出该句子是否包含蛋白质关系特征，词共现方法又可以分为基于单句的蛋白质抽取方法，基于句群的蛋白质关系抽取方法，基于数据库标示码的蛋白质关系抽取方法等。基于单句的蛋白质关系抽取方法是目前比较常见的方法之一，这种方法的目标是通过对词序列的分析找到能够表征蛋白质关系的词特征，这种方法在早期的 PPI 任务中经常用到，但这种方法的提取方式过于简单，不能剔除原始句子中的冗余词噪音（比如标点，无意义词性等），并且由于没有对句子进行句法分析，无法准确表示被标点分割开的两个蛋白质关系。

随着对蛋白质关系抽取方法研究的深入，大量边缘学科的研究理论和研究成果被引入到蛋白质关系抽取中，其中 Vapnic 在 1995 年提出的支持向量机系统（Support Vector Machine）因其具有解决小样本、非线性和高维模式识别中表现出许多特有的优势，而受到广泛的重视，使用支持向量机的蛋白质关系抽取系统普遍可划归为机器学习方法[1]。进而人们提出了基于句法分析的蛋白质关系抽取方法，这种方法以目前主流的句法分析器为基础，在预处理阶段将原始句子进行句法分析，找到核心句式结构，剔除掉句子内部的无意义词，精简了句子的结构，在降噪的基础上进行特征提取，这种方法能够很好的避免冗余词噪音对特征提取的影响，同时在句法分析中能够找到两个蛋白质之间的句法关系，并将其作为特征引入后期分类，精度得到普遍提高，但这种方法的处理过程复杂，且并没有统一的句法分析器和特征提取核方法，多个实验室各自提出了自己的句法分析器和核方法，如何组合能够优势互补达到更好的分类效果是我们面临的难题，采用了多种方法后我们还必须考虑所消耗的时间资源代价能否得到性能的大幅提升。

早期的 PPI 抽取方法侧重以单句为分析单位进行抽取，抽取对象也集中在词元信息，词结构信息上，Giuliano 等人在 2006 年设计了一种分析方法能够抽取出包含蛋白质关系句子的浅层语言信息，该方法通过设计一个联合核函数来合并两类不同的信息源：出现蛋白质关系的句子和实体关系的上下文信息，该方法首次尝试利用上下文句法特征进行关系抽取，并在当时取得了很好的实验效果，但该方法并没有提出一个成型的句法分析器，虽然给出了句法分析的具体算法，分析效果基于规则添加，不能进行扩展和推广[2]。后期的研究者偏向设计更合理的句法分析器，其中 Sætre 等人提出将深层句法分析器和依存句法分析器进行融合共同提高 PPI 抽取效果的方法，该方法通过两类不同来源的句法分析器分别对同一个句子进行分析形成分析树，尝试在树中找到两个蛋白质之间的路

径关系，并在此基础上引入了 SVM 进行特征分类，使得实验效果得到进一步提高[3]。不过该方法并没有对句子结构进行细节优化，特别是对于长难句的处理上时间消耗太大，句法分析的效果也没有进行横向对比，后期各个实验室分别提出了自己的句法分析系统，分析原理各不相同，Miyao 在其基础上对各个句法分析器在 PPI 抽取的效果进行了比较分析，以找到更合理的分析器组合，最大限度提高分类效果[4]。研究发现使用深度和依存树可以更好的提高分类精度，同时如果对分析器进行特定领域的语料再训练可以明显改善 PPI 抽取效果，不过该方法对如何在分析树基础上的特征提取没有进行详细描述。Airola 在 2008 年提出了句法分析器基础上了新的特征抽取方法，该方法引入了图核信息，可以在句法分析树基础上构建出图核 G 矩阵，并通过 G 矩阵计算出两个句子的结构相似度，进而进行 PPI 抽取，该方法能够更好的提取出结构化特征，并形成了特征向量，给后期研究者提供了思路[5]。但该方法并没有针对词信息进行抽取，仅仅提取了结构信息，Miwa 在 2009 年对图核信息进行了进一步研究，引入了三类核方法：词袋、最短路径、图核[6]。三种核和两种句法分析器的合并使用能够使实验精度达到目前的最好效果，不过该方法由于使用的子模块众多，各个模块之间的融合细节调整很费时间，我们无法在其基础上进行实验还原。

1.3 待解决的问题

通过上述分析我们发现，目前主流的蛋白质关系抽取方法普遍将早期的词共现信息作为特征引入到机器学习系统中，通过句法分析、聚类、支持向量机训练和测试等方法得到关于句子中是否存在蛋白质作用关系的判定特征。几乎所有的蛋白质关系抽取系统都普遍使用了句法分析器和核方法提取特征向量，侧重于提取词结构信息，对词义本身信息和序列信息的提取仅仅使用了词袋方法，统计了与蛋白质的位置关系，出现的频率，词元三种信息，这三种信息几乎毫无变化的引入到后期的训练和预测程序中，词序列中包含了更多信息并没有进行提取，这样分类效果往往不尽如人意，这类特征在统计理论中可以解释为自变量特征，其变化的趋势和下一步变化状态受多方面因素制约，必须对自变量特征进行演化加工，使其表示成更具有分类趋势的因变量特征进行分类计算[7]。其次，能够标示蛋白质是否存在关系的信息实际上并不是通过人工肉眼观察就能够概括的，必要时还可以通过发掘隐藏信息的方法来构建出隐藏特征序列。

我们考虑能否设计一种核方法提取出尽可能多的因变量信息和隐藏特征序列信息，和图核方法互补共同提高 PPI 抽取效果。设想能够从句子中提取出更多的词义和词序列信息，这些信息本身就包含了能够表征蛋白质作用关系的语义，句子中如果包含这样的

词序列则蛋白质更倾向于产生关系，这种方法特别适合用于复杂句子的处理，可以将其化简为一个词序列，并在序列中找到合适的词模板。

1.4 本文工作

以往的蛋白质关系抽取系统都倾向于提取出能够区分关系的典型特征，同时引入经典的已有特征构建出属于自己的 PPI 系统，这样的系统确实可以改善分类效果，不过提升幅度并不明显，同时对特征的提取、分析、融合过程要花费大量时间，在实际应用中的作用不大，因此我们需要在现实基础上考虑如何才能在不牺牲精度前提下尽量优化和精简已存的特征，并以此来提高系统分析的效率。

本文采用了一种优化特征的方法进行蛋白质关系抽取，其主要思路是考虑到目前的主流蛋白质关系抽取特征较多，且提取过程各不相同，特征之间的组合往往可以提高整体的抽取精度，但是如何进行组合最为恰当还没有提出具体方法，同时特征之间的组合也并不是都可以提高实验结果，有的时候两个表现优异的特征进行融合后反而会降低整体抽取精度。针对上述问题本文试图通过对主流特征的筛选和优化来对蛋白质关系抽取系统进行构建，在原有特征提取基础上构思一种扩展特征的方法以提取精简优化特征向量，该方法能够在扩展语料集中自动找到最符合蛋白质关系抽取标准的合适特征群落，将这一特征群落进行标准化来进行测试集的预测分类，在保持原有实验精度的前提下大大提高了关系抽取效率。

本文在借鉴前人研究基础上，引入了一种新的模板核方法，模拟人工标注原理进行抽取词序列模板，通过多个核的融合共同提高分类效果。该方法能够很好的解决图核方法不能处理复杂长难句的不足，通过模板快速匹配出分类结果，同时引入了句法分析和关键词标准化减少句子中的冗余词噪音，进而提高了实验效率。在图核和模板特征的共同作用下，我们在 Aimed 语料的实验 F 值达到了 63.1%。

1.5 本文的结构

本文从蛋白质关系抽取的实际出发，详细阐述了基于图核特征和模板特征结合的蛋白质关系抽取原理。文章分为五个部分，包括了研究背景介绍，相关算法说明，单一语料参数设定，扩展特征提取说明，交叉实验，具体章节安排如下：

第 1 章，绪论，综述了论文所研究内容的背景以及研究现状，介绍了本文研究领域中的主要方法和论文的结构安排。

第 2 章，主要介绍了本蛋白质关系抽取系统中使用的句法分析器，分类器的使用原理，特征提取算法，测评方法。重点介绍了图核特征和模板特征提取算法。

第 3 章，详细介绍了在单一语料集下的蛋白质关系抽取参数设定实验，在实验中找到最合适的实验参数和提取算法，并在自分析过程中找到最佳的实验设定结果。

第 4 章，介绍了扩展语料特征的提取算法，在实验中证明了扩展特征对于蛋白质关系抽取的有效性，并通过实验调整最佳的扩展特征的参数。

第 5 章，在实验分析中对交叉语料进行实验，并和同类的实验方法进行对比，通过实验数据证明该方法的有效性。

2 相关知识及评测指标

2.1 句法分析

顾名思义句法分析就是指对句子中的词语语法功能进行分析，将词与词划分不同的层次，在已划分层次的句式结构中发掘出蛋白质存在关系的典型特征。随着文本挖掘技术的深化扩展，句法分析研究日益深入，不同实验室针对自身特点提出了各自的句法分析器，分析器大致可分为按词性划分的句法分析器，按关联对象划分的句法分析器，按距离划分的句法分析器。按词性划分的句法分析器是最常见的句法分析模式，该分析器模拟人工对句子进行句法分析，试图对句子中的词性自动划分出主语、谓语、宾语、补语，尽管这种句法分析器构思全面，具有很大的实际应用效果，但由于句子的成分变化复杂多样，靠机器自动的划分出句法信息困难很大，所以各实验室转而在其基础上构思更具有针对性的句法分析器。按关联对象划分的句法分析器正是在词性划分的基础上演变而来，该种句法分析器并不要求完整的划分出全句式结构，转而针对句子中的某个词或者某个词组进行句法分析，以该词或词组为对象分析出其他句式结构和该词或词组的关联关系，即标示出是否存在关系，对于蛋白质关系抽取任务来说该方法能够针对蛋白质关系进行准确处理，具备实用价值，但对于之后的特征提取还是存在很多噪音，其他任务也面临了同样的问题，因此后期的实验句法分析系统倾向于在解决了词与词关系的基础上引入距离信息，构建出按距离划分的句法分析器。按距离划分的句法分析器是目前主流的句法分析器，该分析器不仅在句式结构中标示了词与词间的作用关系信息，以连线标示，同时还在关系基础上标示了距离信息或者特征信息，距离信息就是该词通过与几个词的作用可以对关联词产生影响，特征信息就是连线两端的词构成了哪种词性词组，例如介词词组、补语词组等等。按距离划分的句法分析器能够有效的对包含蛋白质关系的句子进行分析并在预处理阶段筛选掉不含关系的冗余词，提高实验效率。

目前有很多种句法分析器用于输出基于不同图层的句法分析结构，这些结构分别包含了不同类型的有效特征，很多实验室都致力于研究句法分析器对 PPI 任务的效果，实验证明使用句法分析器不仅可以更深入发掘两个蛋白质之间的隐含信息，同时可以对分析树进行一定程度化简，从而提高后期处理的效率。本文中主要利用了两类句法分析器：

依存性句法分析器：依存性句法分析器将待处理句子看作词序列，用于找到待处理句子中词与词间的依存关系，构建出一个由词与词之间依存关系组成的依存语法树，本

文在依存句法分析器中选择了 Stanford 分析器，图 2.1 就是依存性句法分析器分析效果[8]。

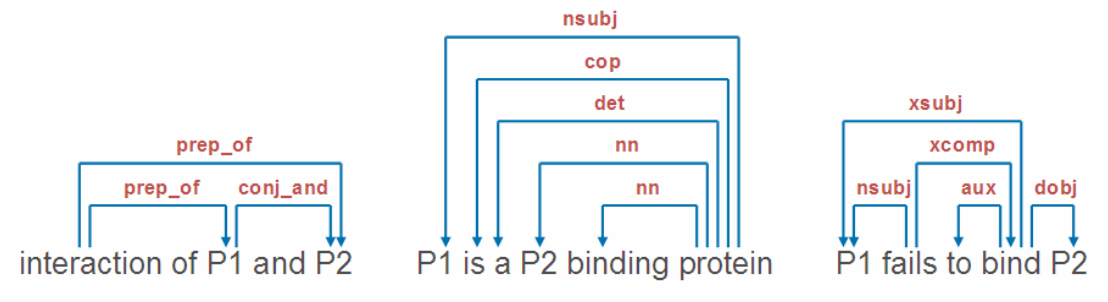

图 2.1　Stanford 依存分析器分析效果示例

Fig. 2.1　Example of effect analysis of standford parser

深度语法分析器：深度语法分析器也将待处理句子看作一个词序列，构建出图结构用以表现词与词之间的句法关系。深度语法分析器不同于依存性语法分析器，它更侧重于找到词语间的深层关系，本文在深度句法分析器中选择了 Enju 分析器，图 2.2 是深层语法分析器分析效果[9]。

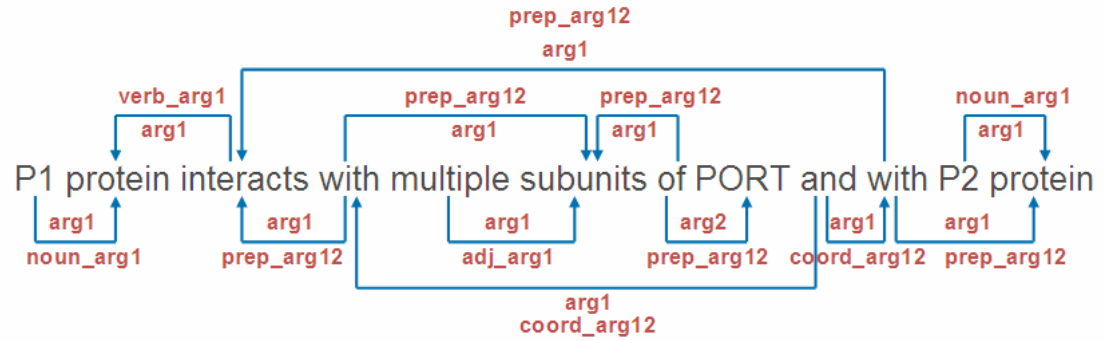

图 2.2　Enju 深度分析器分析效果示例

Fig. 2.2　Example of effect analysis of enju parser

通过上述两类分析器的分析，我们得到了句子的句法分析结构，两类分析器的分析原理不同，因此可以互相补充，后期的处理将依据两类分析器的分析效果进行集中实验。

2.2　支持向量机

支持向量机是统计学习理论文本分类的衍生成果，其目的是利用输入特征来构建多维空间向量，通过降维操作来进行二值或多值分类，达到多特征分类的目的。支持向量机在解决小样本集合、高维空间模式识别、非线性特征分类中具有很好的应用表现。SVM 分类器的选择方面我们采用了 L2-SVM 分类器[1]。

L2-SVM 是一种改进的 SVM 分类器，实际操作过程中我们要用到 LIBLINEAR 工具包，LIBLINEAR 是专为大规模线性分类设计的开源数据包[10]，该数据包支持逻辑回归和线性 SVM 分类，并为使用者提供了易于操作和再处理的命令行接口。大量实验已经证明该数据包中的 L2-SVM 分类器能够很好的应用在 PPI 抽取领域[5]。

2.3 图核特征提取方法

我们在分析器的分析基础上，对其中的有意义信息进行抽取，并最终计算出两个句子的相似度信息，这个过程就是核方法的运用过程。我们主要使用了两类核方法，图核方法和模板核方法[4]。

图核的主要目的是根据句法分析树的树状结构分析出两个蛋白质的距离信息，进而将句子表示为一个图结构，图核中主要使用了两类直接子图：分析结构子图（PSS）和线性顺序子图（LOS）。

分析结构子图（PSS）表现了一个句子的语法分析结构，PSS 中包含词顶点和链顶点，词顶点包含了词元信息和词袋信息，链顶点中包含了语法关联结构。最短路径是通过比较输入树的常见子树来计算出其相似度，计算分析树中两个蛋白质间的最短路径长度，如果最短路径穿过了该顶点则用“IP”进行标注，并将这个最短路径作为特征用于 PPI 抽取。

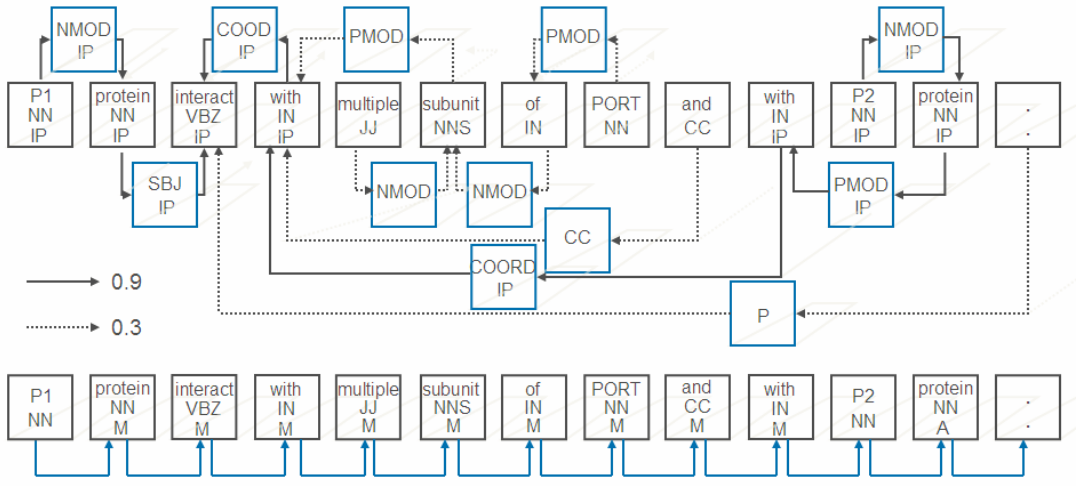

图 2.3　分析结构子图与线性顺序子图示例

Fig. 2.3　Example of phrase structure graph and deep parsing graph

线性顺序子图（LOS）表示了词在原句中的顺序位置信息，LOS 中只有词顶点，每个顶点中包含了词元信息，到目标词的相对位置信息和自身的词袋信息，LOS 是对 PSS