基于关系相似性的蛋白质交互自动识别

封二英 牛 耘 魏 欧 蔡昕烨
(南京航空航天大学计算机科学与技术学院 南京 210016)

摘 要 针对目前蛋白质交互关系识别主要以单句为依据、因标注数据缺乏而导致训练集规模小等不足,提出一种以关系相似性分析为框架、基于大规模文本的蛋白质交互关系自动识别方法。首先通过对大规模生物医学文本数据库的自动搜索获取描述蛋白质对的句子集合,然后分别从单词、短语结构、依赖关系3个角度抽取特征,建立向量空间模型来表示一对蛋白质之间的关系,最后根据两个向量之间的相似性对关系作出判断。所需训练数据直接取自现有蛋白质交互网络,无需任何额外的人工标注。实验表明,基于关系相似性的蛋白质交互关系自动识别取得了较高的精度(F-score 74.2%)。

关键词 蛋白质交互关系,关系相似性,句法分析,空间向量模型
中图法分类号 TP391 **文献标识码** A

Protein-protein Interaction Identification Based on Relational Similarity

FENG Er-ying NIU Yun WEI Ou CAI Xin-ye
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract Current protein-protein interaction (PPI) identification systems use single sentences as evidence, and often suffer from the heavy burden of manual annotation. To address these problems, a new relational similarity-based approach using large-scale text as evidence was proposed. First, description of PPIs is obtained by automatic searching of the whole PubMed database. Then, three types of features including lexical features, phrases, and dependency relations are extracted to build the vector space model of PPI. Finally, similarity between vectors is measured to classify the relationship between two proteins. In this method, training data is taken from existing PPI databases and no extra annotation work is needed. Results of the experiment show that this approach achieves high F-score (74.2%).

Keywords Protein-protein interaction, Relational similarity, Syntactic analysis, Vector space model

蛋白质作为生命活动的体现者并非孤立存在,而是通过相互的交互作用完成细胞中的大部分过程。蛋白质交互(Protein-Protein Interaction,PPI)网络的建立一直是研究生物过程中关注的核心问题。因而由领域专家手工收集的PPI数据库,如生物分子相互作用数据库(Bio-molecular Interactions Database,BIND[1])、人类蛋白质查询数据库(Human Protein Reference Database,HPRD[2])等纷纷建立。然而,随着生物医学文献的迅速增长,手工收集PPI信息远远不能满足研究的需要。目前,大量PPI信息仍隐藏在各种生物医学文献中。作为当前国际上最权威的生物医学文本数据库,PubMed由美国国立医学图书馆建立[3],收录了1966年以来世界70多个国家和地区出版的3400余种生物医学期刊的超过1800万篇生物医学文献。如何从这些文本中自动挖掘出PPI信息对于PPI网络的建立有着重要的意义。在此背景下,基于自然语言处理的PPI自动识别技术正在快速发展并已取得了很大的进展。然而,由于问题的复杂性,目前PPI自动识别系统的精度还远不能满足实际应用的要求。而以单句为依据、因标注数据缺乏而导致训练集规模小是影响现阶段PPI识别精度的两个重要因素。针对这些问题,本文提出以关系相似性分析为框架、以大规模文本库中对蛋白质对的描述为依据的PPI识别方法来克服现有技术的不足,提高PPI识别的精度。

1 相关工作

模式匹配和机器学习是PPI自动识别的主要方法[4,5-10]。模式匹配法首先根据蛋白质交互在句子中的常用描述方式建立一组模式。一个模式通常由单词和词性的序列组成,其中标出了形成交互关系的两个蛋白质的位置。抽取PPI时,如果一个模式与一个句子相匹配,则提取出句中相应位置的蛋白质作为存在交互关系的蛋白质对。模式匹配法简单直观,但难以处理复杂的句子。同时人工建立模式的巨大开销常导致模式的覆盖面很有限。因而一些基于规则的PPI抽取方法也涌现出来。例如Fundel等[6]提出的规则基于句子句法结构中的依赖关系;Temkin等[7]利用带有语法产生规

到稿日期:2012-09-13 返修日期:2012-12-25 本文受教育部高等学校博士学科点专项基金项目(20103218120024),国家自然科学基金项目(61170043),校青年科创基金(NS2012073),国家自然科学基金青年科学基金项目(61202132)资助。
封二英(1985—),女,硕士生,主要研究方向为自然语言处理,E-mail:fengerying@nuaa.edu.cn;牛 耘(1974—),女,副教授,主要研究方向为自然语言处理;魏 欧(1974—),男,副教授,主要研究方向为软件工程;蔡昕烨(1983—),博士,讲师,主要研究方向为智能计算与生物信息。

则的句法分析器来识别 PPI。这些系统着眼于分析整个句子的语法特点，从而充分揭示句中成分之间的关系，能获得更高的准确性，但需要更高的计算能力和时间复杂度。

近年来，越来越多的 PPI 自动识别研究引入了机器学习的方法。机器学习的方法免去了人工建立模式或者规则所需的繁重努力，通过对一个训练集的学习自动建立分类模型来判定蛋白质之间的交互关系。例如 Bunescu 等[8]利用一个基于序列的核来计算两个句子的相似性。Niu 等[9]从单词、单词的句法关系、蛋白质的上下文等信息中提取特征来抽取 PPI 关系。唐楠等[10]使用基于多核的学习方法来进行蛋白质关系信息的抽取，融合了基于特征的核、树核以及图核。

受到自然语言处理领域信息提取技术(Information Extraction)的影响，目前这些 PPI 自动识别工作集中于以单句为研究单位，判别单句中出现的任何两个蛋白质之间是否存在交互关系。这种方式的优势在于能够在句子级别上提供蛋白质交互的描述和证据。然而，PPI 识别的核心目标是建立完整的蛋白质交互网络。因此，直接以目标蛋白质对为研究对象能更好地适应这种需求。并且，生物医学文献中的大量单句本身也因经常需要对多个蛋白质的相互作用进行描述而使用复杂的句法结构。根据对 PubMed 的数据统计[9]，在可能包含蛋白质交互信息的句子中，超过 40%的句子包含对 3 个以上蛋白质间作用的描述。因此，单个句子中可能隐含多种交互关系。但是，由于句法的复杂性，这些交互关系的描述多为间接描述。因此，如果仅依赖单个句子中的信息进行交互关系分析，往往难以得到确定的判断。而且，目前基于机器学习的方法所必需的训练集要求对一个句子中出现的每一对蛋白质标注其是否交互。对于含有 3 个蛋白质的句子就需标注 3 对蛋白质的交互情况，而对含 4 个蛋白质的句子蛋白质对数就增至 6 对。显然要得到这样的标注训练集代价高昂。正是由于此，采用这种方法的分类模型一般都建立在很小的训练集上。而这必然影响模型的推广应用的效果。

针对这些问题，本文以关系相似性研究为理论框架，提出一种基于大规模文本的 PPI 自动识别方法以克服现有方法仅依赖单句线索的不足。这种方法直接利用已有的 PPI 数据库作为训练集，因而避免了大量的人工标注的负担。实验结果表明基于关系相似性的蛋白质交互识别算法取得了较高的精度(F-score 74.2%)。在下面的描述中，第 2 节详细介绍了基于大规模文本的 PPI 抽取方法；第 3 节描述实验和结果；最后对方法进行了总结。

2 基于大规模文本的 PPI 判定

2.1 关系相似性

自然语言处理领域的关系相似性研究为准确分析文本中存在的关系描述提供了统一的框架。关系相似性研究如何用相似性刻画和分析词对(word pair)所蕴含的关系，Medin 等[11]对关系作了如下描述：关系是带两个及以上参数的谓词(predicate)(比如，X 和 Y 冲突，X 比 Y 大)，用于表达物体之间的抽象的联系。关系相似性注重于考察词对在大规模文本库中的上下文之间的联系，通过对其进行统计分析，提取分布规律来刻画关系特征，然后建立计算模型，衡量词对间关系的相似性以判断关系的类型。近年来，关系相似性的研究在自然语言处理领域受到越来越多的关注[12,13]。在 PPI 识别中，蛋白质交互作用表达的正是两个物体(蛋白质)之间的一种典型的语义关系，体现在对关系的文本描述中的多个层次上，具有鲜明的关系相似性分析方法所适用的问题结构特征。因此，本文提出从关系相似性角度开展蛋白质交互作用自动识别的方法，将交互关系在不同语言分析层次上的特征纳入统一框架，并充分利用大规模文本库中包含的描述关系的丰富上下文信息来提高对交互关系的自动识别精度。

关系相似性分析的主要模块包括：关系获取、关系表示以及相似性计算，如图 1 所示。关系获取是指得到包含目标关系的两个成分的有代表性的文本描述的集合。这些描述可以是短语、句子或段落等。比如，Turney[14]选取了包含目标成分(X,Y)的 128 种短语(如 X of Y,Y for X,X to Y)的文本集合作为目标关系的描述。而 Nakov[15]则采用包含目标关系的句子的集合。与 Nakov[15]类似，本文采用包含目标蛋白质对的句子集合作为这对蛋白质的描述。然后，要根据这些描述建立有效的关系表示机制。本文采用向量空间模型来表示一对蛋白质间的关系。关键问题是选取能有效刻画目标关系的重要特征作为向量的维。最后要设计恰当的相似性比较算法对关系作出判断。下面详细阐述关系相似性研究框架下的 PPI 自动识别方法。

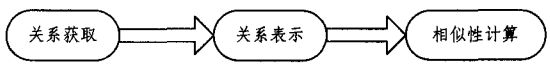

图 1 关系相似性分析过程

2.2 系统框架结构

PubMed 数据库收录了超过 1800 万篇生物医学文献摘要，其是建立 PPI 网络的重要信息来源。现有的 PPI 抽取工作都是建立在对 PubMed 一个很小的子集的分析上，这无论对基于规则或基于监督的机器学习方法都造成了很大的局限。与此不同，本文提出的方法针对每个蛋白质对(一个蛋白质对为一个 PPI 候选对象)在整个 PubMed 范围内收集其全部描述作为此蛋白质对的签名档。这样每个签名档包含了目标蛋白质对的较完整的关系的表达。然后针对签名档提取重要特征，建立起目标蛋白质对的向量表示，进而对其关系作出判断，整个过程如图 2 所示。

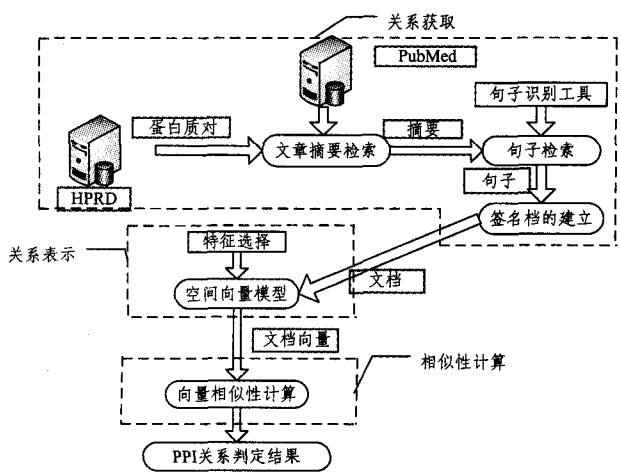

图 2 PPI 判定过程

2.3 获取目标蛋白质对的签名档

首先将目标蛋白质对在 PubMed 数据库中搜索时所包含的组成该对的两个蛋白质的句子作为此蛋白质对的签名档。

由于 PubMed 数据库不提供直接检索句子的功能，签名档的获取分两步进行。

1)搜索包含目标蛋白质对的文献摘要

通过 PubMed 数据库的应用程序接口查询包含目标蛋白质对的文献摘要。以目标对中的两个待考察的蛋白质为查询条件，利用检索命令(esearch)搜索包含这两个蛋白质的摘要的编号(PubMed ID)。再把所得编号作为参数通过调用获取命令(efetch)来取得实际的摘要文本，esearch 和 efetch 命令的参数说明可参考文献[16]。实验中通过脚本语言编写的应用程序对目标蛋白质对集合中的所有蛋白质对实现了自动搜索，获得了相应的摘要文本。这样，目标蛋白质对有由若干摘要组成的文本集合与之对应。

2)搜索包含目标蛋白质对的句子

接下来是对摘要文本集合进行处理，找出其中包含目标蛋白质对的句子。这个过程如下：

①对摘要文本集合进行句子识别。这里使用的是伊利诺州大学 urbana-champaign 分校认知计算研究组开发的句子识别工具[17]。

②在第①步得到的句子集合中搜索所有包含目标蛋白质对的句子，这些句子描述了两个目标蛋白质之间的关系。本步结束时，目标蛋白质对有由若干句子组成的集合与之对应，形成它的签名档。

建立蛋白质对的签名档后，即可利用其中的上下文信息对目标蛋白质对之间的关系进行判断。

2.4 建立蛋白质对的向量空间模型

每个目标蛋白质对的签名档包含了较完整的关系的描述。与之前工作以单句为依据不同，本步将从整个签名档中提取重要线索建立目标对的向量空间模型。向量的每一维对应于从签名档中提取的一个特征。关系相似性研究的实践表明词法结构和句法结构是刻画关系的重要组成部分。本文分别从这两个层次上对蛋白质的签名档进行了分析，并提取其中的重要线索作为识别 PPI 的依据。

• 单词特征。抽取至少出现在 25 个签名档中的单词作为最终的特征词，共 6592 个特征。其特征值取 0(该词未出现在签名档中)或 1(该词出现在签名档中)。

• 短语结构特征。这里需要对签名档中的句子进行浅层句法分析。所使用的是 Apache OpenNLP 的句法分析工具[18]。该工具依照句法相关关系进行句子分块划分，分成名词短语、动词短语等块结构，如图 3 所示。

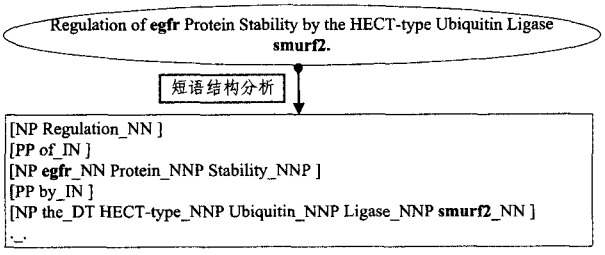

图 3 Apache OpenNLP 短语结构分析输出结果示例

在句法分析结果的基础上，找出包含目标蛋白质名字的短语，然后抽取该短语前后各相邻的两个短语的中心词作为特征。如果在该短语位置的前后出现的是标点符号则不抽取特征。例如在上例中，对于目标蛋白质对的两个成分 egfr 和 smurf2 抽取的特征是：Regulation、of、by、smurf2、Stability、by，如图 4 所示。选取至少出现在 25 个签名档中的这样的中心词作为最终的特征词，共 770 个特征，其特征值取 0 或 1。

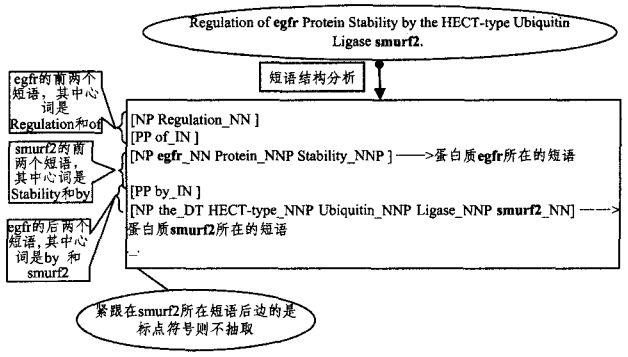

图 4 短语结构特征提取示例

• 依赖关系特征。对签名档中句子的成分进行依赖关系分析，这里使用的分析工具是 Stanford Parser[19]。该工具分析句子的语法结构，指出哪些词之间存在依赖关系，以及依赖关系的类型(主语、谓语等)，如图 5 所示。

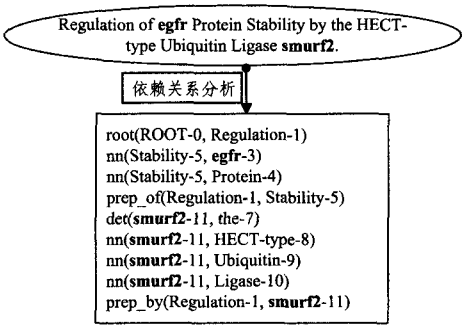

图 5 Stanford Parser 依赖关系分析输出结果示例

图中，prep_of(Regulation-1，Stability-5) 代表 Regulation 和 Stability 之间存在由 of 连接的介词关系。分两种情况抽取依赖关系特征。

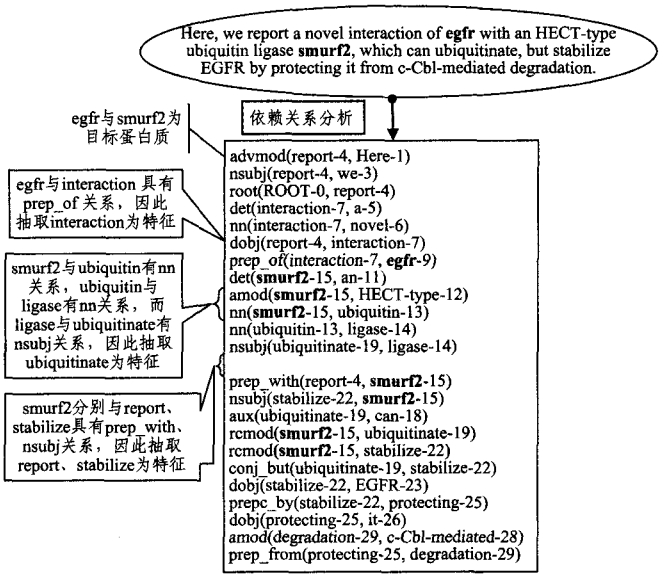

图 6 依赖关系特征提取示例

1. 抽取与目标蛋白质具有直接的 nsubj 或 dobj 或 prep-*(*代表任一介词)关系的词作为特征。

2. 抽取与目标蛋白质具有间接的 nsubj 或 dobj 或 prep-*关系的词作为特征。首先找出与目标蛋白质具有 nn 关系的

名词，不妨设该名词为 A，然后找出与 A 有 nn 关系的名词 B，如果没有词与 B 具有直接 nsubj 或 dobj 或 prep-* 关系，则继续寻找与 B 有 nn 关系的词，继续这样寻找下去，直到找到一个单词，该单词与单词 C 之间存在直接的 nsubj 或 dobj 或 prep-* 关系，此时抽取该单词作为特征。两种情况下抽取特征过程如图 6 所示。选取至少出现在 25 个签名档中这样的词作为最终的特征词，共 926 个特征，其特征值取 0 或 1。

2.5 相似性计算

本步通过比较蛋白质对向量之间的相似性来对目标蛋白质对是否存在交互关系作出判断。这里根据两个向量的余弦值来衡量它们的相似性，如下所示。余弦值越大，两个向量的夹角越小，相似度就越高。相反，余弦值越小，两个向量的相似度就越低。

$$\vec{r}_1=\langle r_{1,1},\cdots,r_{1,n}\rangle,\vec{r}_2=\langle r_{2,1},\cdots,r_{2,n}\rangle$$

$$\text{cosine}(\theta)=\frac{\sum_{i=1}^{n} r_{1,i}\cdot r_{2,i}}{\sqrt{\sum_{i=1}^{n}(r_{1,i})^2\cdot\sum_{i=1}^{n}(r_{2,i})^2}}=\frac{\vec{r}_1\cdot\vec{r}_2}{\sqrt{\vec{r}_1\cdot\vec{r}_1}\cdot\sqrt{\vec{r}_2\cdot\vec{r}_2}}=\frac{\vec{r}_1\cdot\vec{r}_2}{\|\vec{r}_1\|\cdot\|\vec{r}_2\|}$$

3 实验设计及结果分析

3.1 实验数据

在现有工作中，PPI 识别的目标是出现在同一个句子中的任意一对蛋白质。与此不同，本文从实际构建 PPI 网络的需求出发，直接以目标蛋白质对为研究对象。实验中所用数据全部来自于真实的 PPI 网络，全部训练数据可直接从已有 PPI 数据库得到而不需进行任何额外的人工标注。其中，有交互关系的蛋白质对来源于专家人工收集的目前最大的人类 PPI 数据库 HPRD[2]。将其中被包含在 PubMed 的 1 篇以上摘要里的蛋白质对作为有交互关系的集合。这样的蛋白质对是 1417 对。实验中根据生物信息领域常用的方法，对 HPRD 中的蛋白质进行随机组合来产生无交互关系的蛋白质对，并确保随机组合产生的蛋白质对没有出现在 HPRD 的 PPI 集合中。这样产生的蛋白质对有 67714 对。其中 1397 对在 PubMed 中至少有一个句子同时包含组成该对的两个蛋白质。综上所述，实验的数据集包含有交互关系的蛋白质对 1417 对，无交互关系的蛋白质对 1397 对。

3.2 实验设计

实验中对数据集中的每个蛋白质对都生成了一个基于其签名档的向量，并采用留一交叉验证法（leave-one-out cross-validation）进行测试：把数据集中的一个蛋白质对作为测试数据而将其他 2813 对作为训练数据，如此循环 2814 次，从而使得每个蛋白质对都被测试一次。在进行相似性计算时，采用和 Turney[14] 以及 Nakov[15] 相同的方法，即最近邻分类（one-nearest neighbour）将与测试数据最相似的那一对训练数据中的蛋白质的标签（即是否存在交互关系）指定给测试数据。实验结果采用精确度（Precision）$P=TP/(TP+FP)$、召回率（recall）$R=TP/(TP+FN)$、F-Score（$F\text{-}Score=2P\times R/(P+R)$）3 个指标进行评价。

表 1 显示了有交互关系的蛋白质对的判断结果，表 2 显示了无交互关系的蛋白质对的判断结果。

表 1 有交互关系的蛋白质对的相似性计算结果（%）

	精确度	召回率	F-Score
单词特征 (1)	72.9	74.6	73.8
(1)+短语结构特征 (2)	72.8	75.6	74.2
(2)+依赖关系特征	72.7	75.4	74.1

表 2 无交互关系的蛋白质对的相似性计算结果（%）

	精确度	召回率	F-Score
单词特征 (1)	75.9	74.2	75.0
(1)+短语结构特征 (2)	73.9	71.0	72.4
(2)+依赖关系特征	73.7	71.0	72.3

可以看出，无论对于有交互关系或无交互关系的蛋白质对，基于关系相似性的 PPI 识别方法都取得了较高的 F-Sore。并且，识别的精确度和召回率都稳定在较高的值。说明 PPI 的文本描述存在着共性，而从大规模文本中提取的单词和句法层次上的特征能有效表示这些共性，从而建立起好的相似性计算模型对关系作出正确的判断。在增加了短语结构特征和依赖关系特征后，对于有交互关系的蛋白质对而言精确度几乎没有影响，而召回率略有提高。对于无交互关系的蛋白质对，精确度和召回率都有所下降。这大概是高维稀疏向量由于维数的增加而对相似性计算产生了一些干扰而造成的。

由于支持向量机模型（SVM）被广泛用于目前基于机器学习算法的 PPI 识别系统中且取得了很好的效果，因此本文也建立了 SVM 分类器（SVM-light[20]）来对基于大规模文本的蛋白质交互关系向量进行分类。实验采用了 5 折交叉验证法。将数据平分成 5 等份，在每一次实验中选择其中的一份作为测试集，其余的 4 份作为训练集。最后求 5 次的平均值作为最终实验的结果值。表 3 所列为 SVM 对有交互关系的蛋白质对的分类结果。

表 3 SVM 分类结果（%）

	精确度	召回率	F-Score
单词特征 (1)	81.9	54.2	65.2
(1)+短语结构特征 (2)	84.8	53.4	65.6
(2)+依赖关系特征	86.8	53.0	65.8

SVM 分类器取得了很高的精确度，进一步说明所用特征的有效性。在单词特征的基础上增加短语结构和依赖关系特征后，精确度都有所提高。采用全部 3 种特征达到了最高的精确度 86.8%。这表明 SVM 能够较好地处理高维并且可能不相互独立的特征。然而，与相似性算法相比，召回率下降比较明显，因而导致最终的 F-Score 降低了约 9 个百分点。总的说来，以大规模文本为依据、以相似性计算为方法的关系相似性模型在精确度和召回率上均取得了较好的效果，因此得到了最高的 F-Score (74.2%)。

结束语 本文从实际构建 PPI 网络的需求出发，以全新角度对蛋白质交互关系识别展开研究。直接以目标蛋白质对为研究对象，以大规模文本为依据提取特征建立关系相似性计算模型。实验中使用已有 PPI 网络中的真实数据，结果表明基于关系相似性的 PPI 自动识别取得了较高的精确度、召回率和 F-Score。与目前的 PPI 自动识别系统相比，本文中方法能更直接地用于 PPI 网络构建，能够充分利用已有的 PPI 网络数据而不需任何额外的人工标注，并且有助于弥补以单句为线索来判断交互关系的不足。在今后的工作中，准备采取积极的降维策略，以减少高维空间对相似性计算产生的干

（下转第 251 页）

法的收敛速度和求解精度得到了提高，是一种简单高效的改进方法。

参 考 文 献

[1] Karaboga D. An idea based on honey bee swarm for numerical optimization[R]. Kayseri: Erciyes University, 2005

[2] Karaboga D, Basturk B. On the performance of artificial bee colony(ABC) algorithm[J]. Applied Soft Computing, 2008, 8(1): 687-697

[3] Irani R, Nasimi R. Application of artificial bee colony-based neural network in bottom hole pressure prediction in underbalanced drilling[J]. Journal of Petroleum Science and Engineering, 2011, 78(1): 6-12

[4] Horng M H. Multilevel thresholding selection based on the artificial bee colony algorithm for image segmentation[J]. Expert Systems with Application, 2011, 38(11): 13785-13791

[5] Öztürk C, Karaböga D, Görkemli B. Artificial bee colony algorithm for dynamic deployment of wireless sensor networks[J]. Turkish Journal of Electrical Engineering and Computer Sciences, 2012, 20(2): 1-8

[6] 罗钧，李研. 具有混沌搜索策略的蜂群优化算法[J]. 控制与决策，2010，25(12)：1913-1916

[7] Wu Bin, Fan Shu-hai. Improved artificial bee colony algorithm with chaos[A]// Computer Science for Environmental Engineering and Ecoinformatics, 2011[C]. Berlin: Springer, 2011: 51-56

[8] Rajasekhar A, Abraham A, Pant M. Levy mutated artificial bee colony algorithm for global optimization[A]// IEEE International Cnference on Systems, Man and Cybernetics, 2011[C]. Anchorage: IEEE, 2011: 655-662

[9] Guo Peng, Cheng Wen-ming, Liang Jian. Global artificial bee colony search algorithm for numerical function optimization[A]// 7th International Conference on Natural Computation, 2011[C]. Shanghai: IEEE, 2011: 1280-1283

[10] 暴励，曾建潮. 一种双种群差分蜂群算法[J]. 控制理论与应用，2011，28(2)：266-272

[11] Boettcher S, Percus A G. Extremal optimization: Methods derived from co-evolution[C]// Proc. of the Genetic and Evolutionary Computation Conf. San Francisco: Morgan Kaufmann. 1999: 825-832

[12] 齐洁，汪定伟. 极值优化算法综述[J]. 控制与决策，2007，22(10)：1081-1090

[13] 陈泯融. 基于极值动力学的优化方法及其应用研究[D]. 上海：上海交通大学，2008

[14] 骆剑平，陈泯融. 混合蛙跳算法及其改进算法的混合轨迹及收敛性分析[J]. 信号处理，2010，26(9)：1428-1433

(上接第 232 页)

扰，进一步提高识别的精度。

参 考 文 献

[1] Bader G D, et al. BIND-the biomolecular interaction network database[J]. Nucleic Acids Res., 2003, 31(1): 242-245

[2] Peri S, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans [J]. Genome Res., 2003, 13: 2363-2371

[3] U. S. National Library of Medicine. PubMed [OL]. http://www.ncbi.nlm.nih.gov/pubmed/

[4] Ono T, Hishigaki H, et al. Automatic extraction of information on protein-protein interactions from the biological literature[J]. Bioinfornatics, 2001, 17(2): 155-161

[5] Huang M L, Zhu X Y, Hao Y, et al. Discovering patterns to extract protein-protein interactions form full text[J]. Bioinformatics, 2004, 20(18): 3604-3612

[6] Fundel K, et al. RelEx_Relation extraction using dependency parse trees[J]. Bioinformatics, 2007, 23(3): 365-371

[7] Temkin J M, Gllder M R. Extraction of protein interaction information from unstructured text using a context-tree grammar [J]. Bioinformatics, 2003, 19(16): 2046-2053

[8] Bunescu R C, Mooney R J. Subsequence kernels for relation extraction[C]// Proceedings of the 19th Aunual Conference on Neual Information Processing Systems. Cambridge. MA, USA: MIT Press, 2005: 171-178

[9] Niu Y, et al. Evaluation of linguistic features useful in extraction of interactions from PubMed; Application to annotating known, high-throughput and predicted interactions in I²D[J]. Bioinformatics, 2010, 26(1): 111-119

[10] 唐楠，杨志豪，等. 基于多核学习的医学文献蛋白质关系抽取[J]. 计算机工程，2011，37(10)：184-186

[11] Medin D L, Robert L G, Gentner D. Similarity involving attributes and relations: Judgments of similarity and difference are not inverses[J]. Psychological Science, 1990, 1(1): 64-69

[12] Toch E, Reinhartz I, Berger I, et al. Humans, semantic services and similarity: A user study of semantic Web services matching and composition[J]. Journal of Web Semantics: Science, Services and Agents on theWorldWideWeb, 2011, 9: 16-28

[13] Bos J. A survey of Computational Semantics: Representation, Inference and Knowledge in Wide-Coverage Text Understanding [J]. Language and Linguistics Compass, 2011, 5(6): 336-366

[14] Turney P D. Similarity of semantic relations[J]. Computational Linguistics, 2006, 32(3): 379-416

[15] Nakov P, Hearst M A. Solving relational similarity problems using the web as a corpus[C]// Proceedings of ACL-08: HLT. Columbus, Ohio, USA: Association for Computational Linguistics, 2008: 452-460

[16] U. S. National Library of Medicine, Bookshelf [OL]. http://www.ncbi.nlm.nih.gov/books/NBK5500/#chapter1. ESearch

[17] University of Illinois at Urbana Champaign. Sentence Segmentation tool[OL]. http://cogcomp.cs.illinois.edu/page/tools_view/2

[18] Apache Software Foundation. Apache OpenNLP 1.5.2-incubating[OL]. http://opennlp.apache.org/index.html

[19] The Stanford Natural Language Processing Group. Stanford Parser[OL]. http://nlp.stanford.edu/software/lex-parser.shtml

[20] Thorsten Joachims. SVM-light Support Vector Machine[OL]. http://www.cs.comell.edu/people/tj/svm_light