评述

基于机器学习的蛋白质相互作用文献挖掘方法研究进展

李满生, 常乘, 马洁, 朱云平*

军事医学科学院放射与辐射医学研究所, 北京蛋白质组研究中心, 蛋白质组学国家重点实验室, 国家蛋白质科学中心(北京), 北京 102206

* 联系人, E-mail: zhuyunping@gmail.com

收稿日期: 2016-06-20; 接受日期: 2016-07-07

国家重点基础研究发展计划(批准号: 2013CB910800)、国际合作计划(批准号: 2014DFB30010)、国家高技术研究发展计划(批准号: 2015AA020108)和国家自然科学基金(批准号: 61303073, 21475150, 21275160)资助

摘要 蛋白质相互作用是生物体内一类极其重要的分子活动. 自动挖掘、整合生物文献中的蛋白质相互作用有助于生物学的研究, 获得了人们的广泛关注, 成为生物文献挖掘领域的重要任务之一. 目前, 基于机器学习的蛋白质相互作用挖掘方法已经取得了很大进步, 对该领域的进展进行归纳总结将有助于方法的进一步优化和应用. 本文在对机器学习方法构建流程介绍的基础上, 进一步从机器学习的分类器、学习特征、方法评估以及挖掘系统 4 个方面对蛋白质相互作用文献挖掘进行系统总结, 并探讨了其发展前景.

关键词 蛋白质相互作用, 蛋白质相互作用挖掘, 文献挖掘, 机器学习

蛋白质相互作用(protein-protein interaction, PPI)是细胞中一类极其重要的生物分子活动, 这类活动参与细胞生命周期各个生物学过程. 从遗传物质复制到基因表达调控, 从细胞信号转导到物质新陈代谢, 从细胞分裂增殖到细胞衰老凋亡, 都是通过蛋白质分子间彼此的相互作用来完成. 因此, 蛋白质相互作用研究不仅有助于认识生命活动更深的层次, 同时对临床疾病的诊断和治疗也具有重要意义.

随着生命科学研究的发展, 研究成果文献也日益增多. 到 2015 年底, PubMed 数据库中的生物医学文献数量已经超过 2500 万篇, 并且正在以指数形式的速度快速增长^[1], 呈现出大数据的特征. 这些文献

中含有大量的 PPI 知识以及 PPI 最新的研究发现. 及时整理、补充这些新的发现无论是对于蛋白质网络前沿认识的掌握, 还是新分子通路假说的提出都十分重要. 另一方面, 目前蛋白质相互作用数据库中的 PPI 数据规模很大(表 1), 这些 PPI 数据多来源于高通量实验和预测方法, 大部分还没有相应的文献证据, 急需补充文献信息来进一步评估其可信度, 降低假阳性. 然而, 当前文献数量巨大且快速增长, 传统的手工阅读文献方式已经难以及时、高效地收集和整理这些信息, 从而使得分子生物学家的研究工作变得越发困难. 文献挖掘(literature mining)正是在“信息爆炸”的大数据时代掌握、利用海量文献资源, 挖掘和

引用格式: 李满生, 常乘, 马洁, 等. 基于机器学习的蛋白质相互作用文献挖掘方法研究进展. 中国科学: 生命科学, 2016, 46: 1235–1248

Li M S, Chang C, Ma J, et al. Progress of research in machine learning-based protein-protein interaction extraction. Sci Sin Vitae, 2016, 46: 1235–1248, doi: 10.1360/N052016-00142

表 1 蛋白质相互作用数据库资源统计

名称	蛋白质数目	PPI 数目	物种数目	文献数目	URL	版本
DIP	28497	81067	808	8002	http://dip.mbi.ucla.edu/dip/	2015-01
IntAct	91477	591277	894	14151	http://www.ebi.ac.uk/intact/	2016-05
MINT	25530	125464	611	5941	http://mint.bio.uniroma2.it/	2016-05
I2D	—	1279157	7	—	http://ophid.utoronto.ca	2.9
InnateDB	25061	366302	4	35647	http://www.innatedb.com	2015-12
BioGRID	64874	832938	61	46898	http://thebiogrid.org/	3.4.137
HPRD	30047	41327	1	453521	http://www.hprd.org/	Release 9
HINT	19314	53110	7	7769	http://hint.yulab.org	Ver. 3
STRING	9643763	184805933	2031	—	http://string-db.org/	10.0
Reactome	110090	155872	19	20708	http://www.reactome.org	Ver. 56
IID	68831	1566043	6	—	http://dcv.uhnres.utoronto.ca/iid	2016-03

整合文献信息的有效途径。

从 20 世纪 90 年代末开始, 研究人员逐渐开展了从生物文献中自动挖掘蛋白质相互作用关系的工作。发展至今, 人们已经开发了多种挖掘方法, 归纳起来可主要分为 3 类^[2,3]: (i) 基于共出现和规则方法 (co-occurrence and rule-based approach); (ii) 基于计算语言学方法 (computational linguistics-based approach); (iii) 基于机器学习方法 (machine learning-based approach)。人们最早使用共出现和规则方法^[4-8]从生物文献句子中识别蛋白质相互作用关系。蛋白质相互作用关系在生物文献中的描述具有一定的模式及表达规则, 例如, 蛋白质名称对共出现、句子包含特定的相互作用描述词汇(如 interact, bind, active)以及句子中的词序列呈现特定的排列等, 这些正是关系提取的依据。此类方法通常手工定义或是自动抽取规则, 对句子进行模式匹配来抽取句子中的蛋白质相互作用关系。计算语言学方法利用词性标注(part of speech tagging, POST)和句法分析(syntactic parsing)等技术, 根据语法规则构建句法结构从而判断蛋白质之间是否是相互作用关系, 随着自然语言处理(natural language processing, NLP)技术的进步该类方法逐渐获得重视和实践^[9-11]。基于机器学习方法能够整合基于规则方法以及自然语言处理方法挖掘研究过程中积累的知识, 使用它们作为有效学习特征来判断共现于句子中的一对蛋白质是否存在相互作用关系。近年来, 这类方法已经成为蛋白质相互作用挖掘方法的主流, 并且仍然是当前研究的热点。一系列基于机器学习的蛋白质相互作用挖掘方法取得了良好的效果^[12-14]。本文将重点综述基于机器学习的蛋白质相互作用挖掘方法的研究进展,

在介绍方法构建流程的基础上, 进一步对该类方法的两个核心问题——分类器构建和学习特征提取进行梳理, 对方法评估和现有的挖掘系统进行总结; 最后对基于文献挖掘蛋白质相互作用研究领域的方向提出一些展望。

1 基于机器学习的 PPI 挖掘方法构建流程

基于机器学习的蛋白质相互作用挖掘方法基本流程如图 1 所示, 包括: (i) 语料预处理。标准语料中已经标注了蛋白质名称以及蛋白质相互作用关系。有些语料还需要经过预处理获得更丰富完整的注释, 如命名实体识别(named entity recognition, NER)、句子划分(sentence detection)、单词划分(tokenization)、词性标注(part of speech tagging, POST)、词形还原(lemmatization)、词干提取(stemming)、语块划分(chunking)以及句法分析(syntax parsing)等。(ii) 特征提取及筛选。基于上一步处理结果获取适合表达蛋白质相互作用关系的特征。(iii) 特征生成。包括特征的组合及转换等。由于自然语言的多样性及复杂性, 一种特征难以全面或正确表达句子中蛋白质相互作用关系, 需要通过特征的组合及优化来提高方法的判断能力。同时, 获取的文本特征向量需要全部转化为数字特征向量来满足分类器训练要求。(iv) 分类模型的训练及测试。这个步骤通常包括分类器参数的训练、优化及评估。

2 PPI 挖掘的核心问题

基于机器学习的蛋白质相互作用挖掘方法的核

图 1 基于机器学习的 PPI 挖掘方法构建流程

心问题,主要集中在机器学习分类器的构建以及蛋白质相互作用表示特征的提取筛选两个方面。

2.1 分类器选择

目前,不少机器学习分类器已经被应用于蛋白质相互作用挖掘工作中,包括贝叶斯(Bayesian model)模型^[15,16]、最大熵(maximum entropy, ME)模型^[17~20]、核函数方法(kernel method)^[12,21~30]、支持向量机(support vector machine, SVM)^[31~35]以及神经网络(neural network, NN)模型等。其中核函数方法及支持向量机两个分类器最为常用。

(1) 贝叶斯模型。贝叶斯模型通过贝叶斯公式,将一对蛋白质是否为相互作用关系的分类问题转化成求解给定一个特征向量时,这个特征向量对应的蛋白质对为相互作用的概率问题。Marcotte 等人^[15]最早使用了朴素贝叶斯分类器(naive Bayes classifier)来对含有蛋白质相互作用的生物文献进行分类。由于朴素贝叶斯假设样本的各个特征是相互独立的,然而描述蛋白质相互作用的句子特征之间往往是相互关联的,模型具有局限性。这时,考虑了特征间关联因素的贝叶斯网络(Bayesian networks, BNs)模型显然更适用。Chowdhary 等人^[16]利用 12 个特征构成向量,

也即贝叶斯网络的节点,用于计算句子中所含的候选相互作用三元组(两个蛋白质名称及相互作用词汇)为相互作用的概率。他们的研究证实贝叶斯网络模型确实优于朴素贝叶斯模型。另外贝叶斯网络适合处理存在缺失值数据,即使信息不完整,仍然可以进行分类。由于结合了先验知识的概率,贝叶斯网络模型对过分拟合有很好的鲁棒性。然而构造一个好的网络过程十分复杂,也成为该方法的难点所在。

(2) 最大熵模型。最大熵模型利用了最大熵理论,计算在特征限制下最优的概率分布。最大熵模型是自然语言处理领域一个较为成功的模型,它已经被广泛应用于解决分词、词性标注、句法分析以及实体关系(entity relation)挖掘等问题。这些问题均可转化为分类问题,进而可以形式化为估计某一类 y 在上下文 x 下发生的概率问题。蛋白质相互作用关系挖掘问题属于实体关系挖掘的一种,自然也可以使用该模型来解决。自 Kambhatla^[17]最先将最大熵模型应用于实体关系挖掘,随后已经有一系列基于该模型的蛋白质相互作用关系挖掘工作^[18~20]得到开展。Xiao 等人^[18]使用了最大熵模型整合了词、语法及语义特征,在 IEPA 数据集上得到了 88.0% 的精准率和 93.9% 的召回率。Sun 等人^[19]在 PPI 挖掘方法中分两步

使用了最大熵模型, 首先是估算包含有 PPI 信息的句子, 在此基础上计算句子中的每一对蛋白质相互作用的概率. Lee 等人^[20]在 BioNLP'09 Shared Task 提供的数据集上使用了词特征、句子特征和依存句法特征的组合训练了最大熵模型, 该模型获得了 91.36% 召回率. 最大熵模型的优点是易于利用各种有效的特征, 且训练时间短, 适用于大规模的数据集. 同时, 最大熵模型可以有效地利用少量特征标注样本来获得较强的泛化能力.

(3) 核函数方法. 核函数(kernel function)本质是一个内积运算. 它是一个从原空间到新的特征空间的映射, 这个映射可以把原空间中线性不可分的样本在另一个特征空间变成线性可分. 所以严格说来, 核函数并不能直接对样本进行分类, 需要结合线性判别分析(linear discriminant analysis, LDA)或者支持向量机实现分类功能. 根据构造方式的不同, 核函数可以分为平移不变核函数和旋转不变核函数. 而根据样本表示特征的不同, 可以分为特征核(feature kernel)、树核(tree kernel)、图核(graph kernel)及语义核(semantic kernel)等, 它们已经被广泛应用于蛋白质相互作用挖掘当中^[12,21]. 核函数又可以看做是计算一对样本相似度的函数. 基于树核的方法^[22-25]通过将包含蛋白质对的句子进行句法分析获得句法树之后, 计算核函数的值来确定两棵树之间的相似度, 从而判定蛋白质相互作用关系是否存在. 基于图核^[26,27]和组合核^[12,28,29]的方法与树核方法类似. Tikk 等人^[30]对 13 种核函数方法进行了比较分析指出, 特征集的选择比定义相似性计算的函数更有助于方法性能的提高. 因此, 在蛋白质相互作用挖掘研究中, 核函数方法通常用于有效特征的抽取. 相对于其他方法来说, 其计算速度还是一个关键问题.

(4) 支持向量机. 支持向量机是 Corinna Cortes 等人于 1995 年基于统计学习理论的 VC 维(Vapnik-Chervonenkis dimension)和结构风险最小化原则基础上提出的一种机器学习方法. 目前认为, 支持向量机具有较好的泛化学习能力, 同时, 通过结合核函数的映射还可以处理线性不可分问题, 因此应用广泛^[31-36]. Li 等人^[34]利用了句法特征训练了 SVM 模型, 并在该方法上开发了挖掘系统 PPIMiner. Jiang 等人^[35]通过整合词表征(word representation)、向量合成(vector composition)和特征选择(feature selection)来提高 SVM 分类模型的性能.

(5) 神经网络模型. 神经网络是一种运算模型, 由多个神经元节点连接构成, 每个节点代表一种特定的输出函数, 称为激励函数(activation function). 神经元节点通常分为可见节点和隐节点. 根据包含隐节点的隐含层(hidden layer)规模, 可将神经网络分为传统神经网络和深层神经网络(deep neural network, DNN). 在传统神经网络中, 隐含层规模较小, 作为输入神经元节点的学习特征由人工确定. 传统神经网络模型难于训练, 只能实现较少隐含层的结构, 同时容易过拟合, 学习效果并不比其他方法更优, 曾一度陷于沉寂. 因此在蛋白质相互作用挖掘中仅有少量应用. 2003 年, Sugiyama 等人^[37]利用了由 3 个隐层(每层包含 3 个神经元节点)构建了浅层模型, 该模型获得了 76.2% 精确率及 75.4% 召回率的效果, 低于同时构建的 SVM 模型. 随着 Hilton 和 Salakhutdinov^[38]在 2006 年提出由多隐层构成的深度神经网络具有优异的特征学习能力, 同时利用逐层初始化(layer-wise pre-training)方法可以解决训练难度问题. 神经网络模型以“深度学习”(deep learning, DL)这个新概念重新受到研究人员热捧, 在各领域得到尝试. 最近深度神经网络也逐渐被应用到蛋白质相互作用挖掘研究中^[39-41]. Zhao 等人^[39,40]采用了一种神经网络模型——自动编码器(auto encoder, AE)构建了新的蛋白质相互作用挖掘模型. 该模型能够自动学习特征, 并能从文献中有效挖掘蛋白质相互作用, 最优综合性能达到 81.36%. 与前面提到的核函数方法、支持向量机等浅层学习相比, 深度学习能够进行自动特征学习, 规避了人工选择特征; 同时深度学习具有深层次的学习模型, 多层次实现了用较少参数表达复杂的函数, 从而能够更好地刻画包括生物文献等在内的大数据蕴含的丰富信息.

除以上介绍的方法外, K 近邻(K-nearest neighbor, KNN)^[10,37]和决策树(decision tree, DT)^[37]等分类器也被运用于蛋白质相互作用关系的挖掘任务当中, 但不是主流方法. 同时, 还有一些工作^[13,36,42]尝试组合多个分类器方法, 全面利用不同的特征来提高蛋白质相互作用挖掘方法性能.

2.2 特征提取

机器学习模型效果的好坏, 除了分类器的因素外, 还受到学习特征提取和筛选的影响. 目前蛋白质相互作用挖掘中最常用的几种特征是词汇特征(word

feature)、词性特征(POS feature)、逻辑特征(logic feature)、句法特征(syntax feature)、上下文特征(context feature)以及语义特征(semantic feature)。

(1) 词汇及词性特征. 词汇特征是指蛋白质名称之间的词及周围的词两个无序的词集, 也称词袋(bag of word, BOW). 蛋白质名称周围的词包括第一个蛋白质名称左边的 N 个词和第二个蛋白质名称右边的 N 个词. 这类特征还包含了关键词特征(keyword feature), 如相互作用词和连词等. 词性特征指以词的特点为根据划分的词类, 如名词、动词及形容词等. 在语料处理过程中, 分词及词性标注的效果会直接影响这些特征进而影响挖掘结果的准确性. 因此词性标注等自然语言处理相关工具的选择也很关键, 在后面章节将作介绍.

(2) 逻辑特征. 逻辑特征包含句子中两个蛋白质名称间距离(用间隔词的数目衡量)、蛋白质名词间是否含有相互作用词、其他蛋白质、否定词以及标点

符号等. 两个蛋白质名称间距离较为直观, 同时也是一个十分重要的学习特征. 一般情况下, 两个蛋白质之间的距离越短, 这两者就越可能相互作用^[12]. 但 Bui 等人^[33]在研究中指出, 在某些相互作用的表达模式下这一特征是失效的.

(3) 句法特征. 自然语言的语法分为词法(morphology)和句法(syntax). 句子各个组成部分及其安排的规律称之为句法, 获取这类有效学习特征的分析分为 3 个层面: 浅层句法分析(shallow parsing)、深层句法分析(deep parsing)和全解析(full parsing). 浅层句法分析也称语块分析(chunk parsing), 即将句子解析成较小的单元, 如名词短语(noun phrase, NP)、动词短语(verb phrase, VP), 并整合成主语(subject)、宾语(object)等成分, 得到清晰的句子主干, 方便进一步挖掘实体关系. 深层句法分析和全解析最终得到句子的句法树(syntax tree)(图 2). 依存句法分析(dependency parsing)是在蛋白质相互作用关系挖掘

图 2 语料处理及常用学习特征

A: 蛋白质相互作用挖掘任务处理的基本语料单元——句子; B: 蛋白质名称识别; C: 通过分词处理获得的词序列; D: 由词性标注工具获得句子对应的词性序列; E: 句法分析获取的短语结构树及语义依存树

中最常用的深层句法分析之一。依存句法分析可获得依存树(dependency tree)或依存图(dependency graph), 根据这类特征可以构建树核函数或图核函数, 也可以作为一类有效的特征直接用于分类器使用。由于蛋白质相互作用关系的描述常涉及反映关系的一些动词, 以这些动词作为中心词得到依存分析树, 更准确地呈现出复杂句子中的蛋白质相互作用关系的描述规则, 使蛋白质相互作用关系抽取方法取得更好的效果。Fundel 等人^[11]应用了依存分析方法产生句法分析树(依存树), 及句子各个成分之间的依存关系, 然后获取句子中的蛋白质对是否存在依存关系及其依存路径特征用于训练机器学习模型。该方法的精准率和召回率都达到了 80%。

(4) 上下文特征。上下文特征是一类较为宽泛的特征, 带有较强的主观性。通常这类特征通过对词、词频、近义词及同义词的统计来衡量。前面介绍的词汇特征和逻辑特征常常被归入此类。

(5) 语义特征。语义特征通常使用语义相似度(semantic similarity)来衡量。常用于蛋白质相互作用关系挖掘的语义特征包括两类, 两个蛋白质的语义相似度和上下文语义相似度^[13]。它们的基础都是词与词之间的语义相似度。蛋白质的语义相似度可以通过每个蛋白质周围的词集之间的语义相似度体现, 而上下文语义相似度则可以由组成句子的词集之间的相似度计算。词与词之间, 词集与词集之间的语义相似度计算方法多种多样, 其中根据本体(ontology)及依存树算法较为常见。

在蛋白质相互作用挖掘方法构建时, 单个特征往往难以全面描述蛋白质相互作用关系。所以常常将多组特征进行组合, 并对它们进行优化从而提高机器学习方法的性能。Yang 等人^[32]就通过支持向量机融合多类特征集, 包括词汇特征、逻辑特征以及句法特征来获得更优的蛋白质相互作用挖掘方法。

3 评估语料、指标及竞赛

为了准确评估蛋白质相互作用挖掘方法, 比较不同方法性能的优劣, 构建统一的标准语料及衡量指标必不可少。同时, 生物医学文献挖掘领域近年来也组织了针对蛋白质相互作用挖掘的评测竞赛, 以促进蛋白质相互作用挖掘方法的交流和提高。

3.1 评估语料

目前, 主要存在着 IEPA^[43], LLL05^[44], AIMed^[45], HPRD50^[11]及 BioInfer^[46] 5 个可用于蛋白质相互作用挖掘评估的金标准数据集(表 2)。

LLL05 语料集有基础(basic)和丰富(enrich)两个版本, 基础版本包含了句子, 蛋白质名称及位置信息, 分词及位置信息, 相互作用关系和方向。而丰富版本较基础版本添加了句法分析信息。AIMed 数据集的语料是使用“Human”为关键词从 MEDLINE 数据库搜索的摘要集, 然后从中提取了 1955 个句子, 注明了蛋白质名称及蛋白质相互作用。HPRD50 语料集由 HPRD 数据库中摘取的 50 篇摘要构成, 包含 163 对相互作用。BioInfer 语料集中包含的注释信息除了最基本的基因/蛋白质名称及相互作用关系外, 还定义了一个特有的关系类型本体(relation type ontology)用于标定蛋白质相互作用类型。Pyysalo 等人^[47]将表 2 中 5 种语料进行了格式的统一, 方便了后续的研究人员使用。同时, 他们使用了共出现方法和 RelEx 方法对这 5 个标准语料进行系统分析, 揭示了语料间的相似点和不同点。

3.2 评估指标

精准率(precision)、召回率(recall)及 F 值(F-value) 3 个统计学指标常用于衡量蛋白质相互作用挖掘方法性能的指标。它们的计算方法如式(1)~(3)。通常在文

表 2 蛋白质相互作用挖掘评估语料集

语料集	IEPA	LLL05	AIMed	HPRD50	BioInfer
摘要数	303	45	225	50	836
句子数	486	77	1970	145	1100
实体类型	蛋白质	基因/蛋白质	蛋白质	蛋白质	基因/蛋白质
正例数	335	165	1068	163	2534
反例数	482	166	4832	270	7132
方向信息	有	有	无	无	有
发布日期	2002	2005	2005	2006	2007, 2008

献挖掘方法中, 同时获得较高的精准率和召回率是一种理想情况, 实际常常是精准率高、召回率就低, 或者召回率高、精准率就低. N 倍交叉验证(N -fold cross validation, CV)实验通常会被用来检验挖掘方法性能的稳定性.

精准率=正确挖掘 PPI 数目/语料中 PPI 总例数 (1)

召回率=正确挖掘 PPI 数目/语料中 PPI 正例 (2)

F 值=2×精准率×召回率/(精准率+召回率) (3)

3.3 评测会议

与蛋白质相互作用挖掘相关的评测会议主要有 3 个: LLL’05(Learning Language in Logic Challenge)^[48], BioNLP-ST(BioNLP Shared Task)^[49] 以及 BioCreAtIvE^[50].

LLL’05 会议在 2005 年举办了对基因/蛋白质相互作用关系挖掘方法的评测. 该会议为参赛方法提供了包含 77 个句子, 165 对相互作用关系语料集. 最好的方法获得了 0.54 的 F 值^[48].

BioNLP-ST 评测会议在 2009 年针对生物分子事件(bimolecular event)的挖掘方法进行评测^[51]. BioNLP-ST 2011^[52]的主要任务集中在传染病(infectious diseases, ID)挖掘、表观遗传学和翻译后

修饰(epigenetics and post-translational modifications, EPI)挖掘以及支撑性任务——实体关系挖掘方法的评测上.

BioCreAtIvE 评测会议(<http://www.biocreative.org>)自 2003 年起, 平均每 2~3 年都会根据生物文献挖掘发展的需要, 提出不同文献挖掘主题的评测任务, 并且每年都举办评估竞赛. 蛋白质相互作用挖掘在 BioCreAtIvE II (2006)即作为一个重要的评测任务提出. 纵观参与 BioCreAtIvE II 评测的方法, 大多数采用的是机器学习方法, 如朴素贝叶斯、最大熵及 SVM, 其中 SVM 方法使用最多^[53].

4 PPI 挖掘支撑工具及系统

4.1 PPI 挖掘支撑工具

蛋白质相互作用挖掘是一个系统工程, 从文献原始数据下载到蛋白质相互作用关系获取需要经过句子发现、句子分词、词性标注、句法分析、蛋白质名称识别及句子特征提取等多个过程, 需要相应的自然语言处理工具. 以下是常用的与生物文献挖掘相关的自然语言处理工具(表 3).

表 3 自然语言处理相关工具

任务	工具名称	URL	发表时间
词性标注	GENIA Tagger ^[54,55]	http://www.nactem.ac.uk/GENIA/tagger/	2005
	MedPost ^[56]	https://mmtx.nlm.nih.gov/MedPostSKRTagger.shtml	2004
	Stanford Tagger ^[57]	http://nlp.stanford.edu/software/tagger.shtml	2003
	Junk Tagger ^[58]	http://www.jaist.ac.jp/~kazama/junk_tagger.html	2001
	TnT Tagger ^[59]	http://www.coli.uni-saarland.de/~thorsten/tnt/	2000
	Tree Tagger ^[60,61]	http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/	1994
	Brill Tagger ^[62]	http://www.ling.gu.se/~lager/mogul/brill-tagger/	1992
句法分析	C&C CCG Parser ^[63]	http://svn.ask.it.usyd.edu.au/trac/candc	2009
	Enju ^[64]	http://www.nactem.ac.uk/enju/	2008
	Pro3Gres Parser ^[65]	https://files.ifi.uzh.ch/cl/gschneid/parser/	2008
	GENIA Dependency Parser, GDep ^[66]	http://www.cs.cmu.edu/~sagae/parser/gdep	2007
	Stanford Dependency Parser ^[67]	http://nlp.stanford.edu/software/lex-parser.shtml	2006
	Berkeley Parser ^[68]	https://github.com/slavpetrov/berkeleyparser	2006
	McCCJ Parser ^[69-71]	https://github.com/BLLIP/bllip-parser	2005, 2010
	Bikel ^[72,73]		2002
蛋白名称识别	Stanford Parser ^[74]	http://nlp.stanford.edu/software/lex-parser.shtml	2002
	Gimli ^[75]	http://bioinformatics.ua.pt/support/gimli/doc/index.html	2013
	GeneTUKit ^[76]	http://www.qanswers.net/GeneTUKit	2011
	BANNER ^[77]	http://cbioc.eas.asu.edu/banner/index.html	2009
	CBR-Tagger ^[78]	http://moara.dacya.ucm.es/	2009
	ProMiner ^[79]	www.scai.fraunhofer.de/prominer	2005
	ABNER ^[80]	http://pages.cs.wisc.edu/~bsettles/abner	2005

表 3 所列的词性标注工具中, MedPost 及 GENIA Tagger 两个词性标注工具是专门针对生物医学文献开发的. 其中, GENIA Tagger 具有很好的鲁棒性, 在不同的测试数据集上均能获得 97%~98% 的准确度.

不同的句法分析器对于句法特征提取有重要的影响, Miyao 等人^[81]评估了 McCCJ^[69-71], Stanford^[67], Berkeley^[68]和 Enju^[64]等几个句法分析器对蛋白质相互作用关系挖掘效果贡献, 结果显示 Enju 分析器优于其他分析器性能. Miwa 等人^[82]对 Bikel^[72,73], Stanford, GDep^[66], McCCJ, C&C CCG^[63]和 Enju 6 个分析器在生物事件提取上进行的比较分析也得到了同样的结论. BioNLP Shared Task 2011 提供了 Berkeley, C&C CCG, Enju, GDep, McCCJ, Pro3Gres^[65]及 Stanford 这 7 个句法分析器处理的测试语料. 而 BioNLP Shared Task 2013 则提供 McCCJ, Stanford 及 Enju 3 个处理的语料, 可见 McCCJ, Stanford 及 Enju 这 3 个分析器在生物文献挖掘中得到肯定.

蛋白质名称识别是从文献数据中挖掘蛋白质相互作用关系最基本的一个环节, 其效果的好坏直接影响蛋白质相互作用关系挖掘的结果. 目前, 采用条件随机域(conditional random field, CRF)算法的 Gimli 工具^[75]在标准数据集上获得了 87.17%F 值. 另外, 其提供命令运行的方式让用户可以自行选择学习特征以及参数来训练自己的模型, 同时还能将程序整合到自己的挖掘平台当中. 良好的性能及灵活性使该工具成为当前生物命名实体识别最好的解决方案.

4.2 PPI 挖掘系统

随着蛋白质相互作用挖掘方法研究的深入, 对蛋白质相互作用关系挖掘任务的尝试逐渐突破小规

模测试语料和方法学研究的范畴, 扩展到生物文献挖掘的实际应用当中. 基于前面介绍的分类器和学习特征已经开发了一批软件或在线工具, 用于协助人们从现有的文献数据库(如 PubMed, PubMed Center 等)中挖掘相关领域的蛋白质相互作用. 较早出现的 iHOP (information hyperlinked over proteins)^[83,84]针对生物实体识别及实体关系挖掘. 研究人员能够通过它挑选包含目标蛋白质的句子, 然后将句子中包含的蛋白质相互作用进行网络可视化. Yang 等人^[85]整合基于特征、卷积树的核函数来实现 PPI 挖掘, 并在此基础上开发了 PPIExtractor 工具. 该工具能够进行蛋白质名称标准化及对 PPI 挖掘结果评分. 表 4 列出了常用的一些 PPI 挖掘系统.

从蛋白质相互作用挖掘系统的发展趋势来看, 它们越来越重视应用场景, 针对性地考虑了文献数据的特点以及用户使用的更多需求. 例如, 根据生物数据更新速度快的特点, 实现数据自动化更新; 对于文献数量多的特点, 进行数据并行化处理; 为方便人工筛选, 整合了蛋白质名称标准化功能^[85,86], 即将挖掘到的蛋白质名称与现有数据库 ID 进行对应, 同时对挖掘结果进行打分排序^[85,87,88], 并实现蛋白质相互作用网络可视化功能^[83-85,89].

5 总结与展望

目前, 蛋白质相互作用挖掘研究已经取得了长足的进步, 人们已经从多个角度对基于机器学习的蛋白质相互作用挖掘方法进行了研究和改进.

(1) 机器学习技术的发展影响着蛋白质相互作用挖掘研究方向. 除了本文总结的监督式学习方法外, 半监督式学习方法^[10,94,95]也已经被引入到该领

表 4 蛋白质相互作用文献挖掘系统

系统名称	使用方式	URL	发表时间
iHOP ^[83,84]	Web server/ services	http://www.ihop-net.org/UniPub/iHOP/	2004
PPIExtractor ^[85]	Standalone	http://202.118.75.18:8080/PPIExtractor	2011
PPI Finder ^[86]	Web server	http://liweilab.genetics.ac.cn/tm/	2009
PolySearch ^[87]	Web server	http://wishart.biology.ualberta.ca/polysearch/	2008
PIE the search ^[88]	Web server	http://www.ncbi.nlm.nih.gov/IRET/PIE/	2012
PPLook ^[89]	Web server/ Standalone	http://meta.usc.edu/softs/PPLook/	2010
ReactionExtractor ^[90]	Standalone	—	2012
PPInterFinder ^[91]	Web server	http://www.biominigbu.org/ppinterfinder	2013
PCorral ^[92]	Web server	http://www.ebi.ac.uk/Rebholz-srv/pcorral/	2013
BioQRator ^[93]	Web server	http://www.bioqrator.org/	2014

域. 由于对生物语料进行人工标注除了耗时费力还需要专业的背景知识, 代价很高. 已有标注语料对于全部生物医学文献而言规模很小. 半监督式学习方法可利用少量的标注语料结合大量的无标注语料来提高分类器的泛化学习能力, 突破标注语料规模及偏性的限制, 有可能逐渐成为该领域未来的研究热点之一. 另外, 无监督式学习方法^[96]、深度机器学习方法^[39-41]也逐渐被应用到蛋白质相互作用挖掘研究当中, 推动着该领域的发展.

(2) 蛋白质相互作用挖掘发展得益于自然语言处理技术的进步. 目前已经有较好的工具进行分句、分词、词性标注、短语分块以及句法分析等基础工作, 从而能够更好地解析和表示文本数据的语法特征. 词汇特征、词性特征、逻辑特征还有句法特征等已被证实能够有效地表达蛋白质相互作用关系. 值得指出, 句法特征在蛋白质相互作用挖掘方法性能的提高中作用最为突出.

(3) 蛋白质相互作用挖掘的内涵正逐渐得到扩展. 除相互作用关系以外, 蛋白质相互作用方向^[97]、蛋白质相互作用类型、相互作用的实验信息以及蛋白质与其他生物实体关系的挖掘正受到越来越多的关注.

(4) 与蛋白质相互作用挖掘相关的竞赛已经成为 BioNLP-ST, BioCreAtIvE 等主流生物文献挖掘评测会议每一届竞赛的重要组成部分. 同时蛋白质相互作用相关语料数据集也在不断的丰富当中, 为挖掘方法的评估提供了重要的支撑.

(5) 已有的挖掘系统在蛋白质相互作用关系提取任务中发挥越来越重要的作用, 提高了分子生物

学家获取信息的效率.

蛋白质相互作用信息挖掘的现实需求正驱动着该领域的快速发展. 然而, 在对实际数据进行处理时, 现有方法和工具的自动挖掘结果仍存在较高的假阳性. 这是因为蛋白质相互作用信息挖掘的过程是一个系统工程, 需要经过文献数据的筛选、蛋白质名称识别、蛋白质名称标准化、蛋白质相互作用关系挖掘等多个步骤. 每个步骤的方法都会给结果带来一定程度的假阳性. 基于机器学习的蛋白质相互作用挖掘方法本身, 训练语料的偏性及分类器的泛化能力限制也是造成假阳性的重要因素. 总之, 自然语言存在着复杂性、不确定性等因素, 给机器学习方法带来了巨大的挑战. 基于机器学习的蛋白质相互作用文献挖掘研究领域还需要着重关注和解决以下问题:

第一, 从方法学本身进一步提高挖掘工具的性能. 尝试更多新的机器学习方法(如深度学习)及分类器的组合, 同时发掘更多新的有效学习特征及特征组合.

第二, 开发适合实验生物研究人员应用的工具. 如前所述, 蛋白质相互作用的自动挖掘结果的假阳性较高. 为了快速获得高质量的蛋白质相互作用数据, 需要通过人工的参与. 因此, 生物信息人员需要与实验生物研究人员充分沟通、明确需求, 以构建更高效实用的交互式辅助挖掘平台, 利于使用者对挖掘结果进行筛选处理. 同时, 需要考虑到生物文献巨大的数量及指数式增长速度, 强化挖掘平台的性能, 使其在数据整合一体化、更新自动化及处理并行化方面进一步提高.

参考文献

- 1 Sayers E. A General Introduction to the E-utilities. In: Entrez Programming Utilities Help. Bethesda (MD): National Center for Biotechnology Information (US), 2010
- 2 Zhou D, He Y. Extracting interactions between proteins from the literature. *J Biomed Inform*, 2008, 41: 393-407
- 3 Li C, Liakata M, Rebholz-Schuhmann D. Biological network extraction from scientific literature: state of the art and challenges. *Brief Bioinform*, 2014, 15: 856-877
- 4 Blaschke C, Andrade M A, Ouzounis C, et al. Automatic extraction of biological information from scientific text: protein-protein interactions. In: International Conference on Intelligent Systems for Molecular Biology, 1999, Aug 6-10, Heidelberg. Palo Alto: AAAI Press, 1999. 60-67
- 5 Ono T, Hishigaki H, Tanigami A, et al. Automatic extraction of information on protein-protein interaction from scientific literature. *Genome Inform*, 1999, 10: 296-297
- 6 Ng S K, Wong M. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Inform*, 1999, 10: 104-112

- 7 Huang M, Zhu X, Payan D G, et al. Discovering patterns to extract protein-protein interactions from full biomedical texts. *Bioinformatics*, 2004, 20: 3604–3612
- 8 Hao Y, Zhu X, Huang M, et al. Discovering patterns to extract protein-protein interactions from the literature: Part II. *Bioinformatics*, 2005, 21: 3294–3300
- 9 Temkin J M, Gilder M R. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 2003, 19: 2046–2053
- 10 Erkan G, Ozgur A, Radev D R. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In: *EMNLP-CoNLL*, 2007 Jun 28–30, Prague. Association for Computational Linguistics, 2007. 228–237
- 11 Fundel K, Küffner R, Zimmer R. RelEx—Relation extraction using dependency parse trees. *Bioinformatics*, 2007, 23: 365–371
- 12 Yang Z, Tang N, Zhang X, et al. Multiple kernel learning in protein-protein interaction extraction from biomedical literature. *Artif Intell Med*, 2011, 51: 163–173
- 13 Li L, Zhang P, Zheng T, et al. Integrating semantic information into multiple kernels for protein-protein interaction extraction from biomedical literatures. *PLoS One*, 2014, 9: e91898
- 14 Papanikolaou N, Pavlopoulos G A, Theodosiou T, et al. Protein-protein interaction predictions using text mining methods. *Methods*, 2015, 74: 47–53
- 15 Marcotte E M, Xenarios I, Eisenberg D. Mining literature for protein-protein interactions. *Bioinformatics*, 2001, 17: 359–363
- 16 Chowdhary R, Zhang J, Liu J S. Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics*, 2009, 25: 1536–1542
- 17 Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In: *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, 2004 Jul 21–26, Barcelona. Association for Computational Linguistics, 2004. 178–181
- 18 Xiao J, Su J, Zhou G D, et al. Protein-protein interaction extraction: a supervised learning approach. In: *Proc Symp on Semantic Mining in Biomedicine*, 2005 Apr 10–13, Hinxton. Citeseer, 2005. 51–59
- 19 Sun C, Lin L, Wang X, et al. Using maximum entropy model to extract protein-protein interaction information from biomedical literature. In: *Advanced Intelligent Computing Theories and Applications*, 2007 Aug 21–24, Qingdao. Berlin: Springer, 2007. 730–737
- 20 Lee H G, Park S Y, Rim H C, et al. A maximum entropy-based bio-molecular event extraction model that considers event generation. *J Inform Proc Syst*, 2015, 11: 248–265
- 21 Tikk D, Thomas P, Palaga P, et al. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol*, 2010, 6: e1000837
- 22 Zhou G, Qian L, Fan J. Tree kernel-based semantic relation extraction with rich syntactic and semantic information. *Inform Sci*, 2010, 180: 1313–1325
- 23 Qian L, Zhou G. Tree kernel-based protein-protein interaction extraction from biomedical literature. *J Biomed Inform*, 2012, 45: 535–543
- 24 Ma C, Zhang Y, Zhang M. Tree Kernel-based Protein-Protein Interaction Extraction Considering both Modal Verb Phrases and Appositive Dependency Features. In: *Proceedings of the 24th International Conference on World Wide Web Companion*, 2015 May 18–22, Florence. ACM, 2015. 655–660
- 25 Li L, Guo R, Jiang Z, et al. An approach to improve kernel-based protein-protein interaction extraction by learning from large-scale network data. *Methods*, 2015, 83: 44–50
- 26 Zhang Y, Lin H, Yang Z, et al. Neighborhood hash graph kernel for protein-protein interaction extraction. *J Biomed Inform*, 2011, 44: 1086–1092
- 27 Airola A, Pyysalo S, Björne J, et al. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 2008, 9: 1–12
- 28 Li L, Yang L, Huang D. Protein-protein interaction extraction based on ensemble kernel. *J Chin Inform Process*, 2013, 1: 86–92
- 29 Miwa M, Sætre R, Miyao Y, et al. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *Int J Med Inform*, 2009, 78: e39–e46
- 30 Tikk D, Solt I, Thomas P, et al. A detailed error analysis of 13 kernel methods for protein-protein interaction extraction. *BMC Bioinformatics*, 2013, 14: 1–20
- 31 Mitsumori T. Extracting protein-protein interaction information from biomedical text with SVM. *IEICE Transact Inform Syst*, 2006, 89: 2464–2466
- 32 Yang Z, Lin H, Li Y. BioPPISVMEExtractor: a protein-protein interaction extractor for biomedical literature using SVM and rich feature

- sets. *J Biomed Inform*, 2010, 43: 88–96
- 33 Bui Q C, Katrenko S, Sloot P M. A hybrid approach to extract protein-protein interactions. *Bioinformatics*, 2011, 27: 259–265
- 34 Li M, Munkhdalai T, Yu X, et al. A novel approach for protein-named entity recognition and protein-protein interaction extraction. *Mathemat Problems Eng*, 2015, 2015: 10
- 35 Jiang Z, Li L, Huang D. A general protein-protein interaction extraction architecture based on word representation and feature selection. *Int J Data Mining Bioinform*, 2016, 14: 276–291
- 36 Yang Z, Lin Y, Wu J, et al. Ranking support vector machine for multiple kernels output combination in protein-protein interaction extraction from biomedical literature. *Proteomics*, 2011, 11: 3811–3817
- 37 Sugiyama K, Hatano K, Yoshikawa M, et al. Extracting information on protein-protein interactions from biological literature based on machine learning approaches. *Genome Inform*, 2003, 14: 699–700
- 38 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, 313: 504–507
- 39 Zhao Z, Yang Z, Luo L, et al. Deep neural network based protein-protein interaction extraction from biomedical literature. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015 Nov 9–12, Washington. IEEE, 2015. 1156
- 40 Zhao Z, Yang Z, Lin H, et al. A protein-protein interaction extraction approach based on deep neural network. *Int J Data Mining Bioinform*, 2016, 15: 145–164
- 41 Li L, Guo R, Jiang Z, et al. Improving Kernel-based protein-protein interaction extraction by unsupervised word representation. In: 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2014 Nov 2–5, Belfast. IEEE, 2014. 379–384
- 42 Miwa M, Sætre R, Miyao Y, et al. Combining multiple layers of syntactic information for protein-protein interaction extraction. In: Proceedings of the Third International Symposium on Semantic Mining in Biomedicine, 2008 Sep 1–3, Turku. Turku Centre for Computer Science (TUCS), 2008. 101–108
- 43 Ding J, Berleant D, Nettleton D, et al. Mining MEDLINE: abstracts, sentences, or phrases. In: Proceedings of the Pacific Symposium on Biocomputing, 2002 Jan 3–7, Lihue. World Scientific, 2002. 326–337
- 44 Nédellec C. Learning language in logic-genic interaction extraction challenge. In: Proceedings of the 4th Learning Language in Logic Workshop (LLL05), 2005 Aug 7, Bonn. Citeseer, 2005. 31–37
- 45 Bunescu R, Ge R, Kate R J, et al. Comparative experiments on learning information extractors for proteins and their interactions. *Artif Intell Med*, 2005, 33: 139–155
- 46 Pyysalo S, Ginter F, Heimonen J, et al. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 2007, 8: 50
- 47 Pyysalo S, Airola A, Heimonen J, et al. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 2008, 9: S6
- 48 Hutchison L J, Chakravarty P, Kawchuk L M, et al. LLL'05 challenge: genic interaction extraction-identification of language patterns based on alignment and finite state automata. In: Proceedings of the 4th Learning Language in Logic workshop (LLL05), 2005 Aug 7, Bonn. Citeseer, 2005, 72: 38–45
- 49 Kim J D, Pyysalo S. Bionlp shared task. In: Encyclopedia of Systems Biology. Berlin: Springer, 2013. 138–141
- 50 Hirschman L, Yeh A, Blaschke C, et al. Overview of BioCreative IV: critical assessment of information extraction for biology. *BMC Bioinformatics*, 2005, 6: S1
- 51 Kim J D, Ohta T, Pyysalo S, et al. Overview of BioNLP'09 shared task on event extraction. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, 2009 Jun 4–5, Boulder. Association for Computational Linguistics, 2009. 1–9
- 52 Pyysalo S, Ohta T, Rak R, et al. Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. *BMC Bioinformatics*, 2012, 13: S2
- 53 Krallinger M, Leitner F, Rodriguezpenagos C, et al. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol*, 2008, 9: S4
- 54 Tsuruoka Y, Tsujii J I. Bidirectional inference with the easiest-first strategy for tagging sequence data. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, 2005 Oct 6–8, Vancouver. Association for Computational Linguistics, 2005. 467–474
- 55 Tsuruoka Y, Tateishi Y, Kim J D, et al. Developing a robust part-of-speech tagger for biomedical text. *Adv Inform*, 2005: 3746: 382–392
- 56 Smith L, Rindflesch T, Wilbur W J. MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*, 2004, 20: 2320–2321
- 57 Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the

- 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, 2003 May 27–Jun 1, Edmonton. Association for Computational Linguistics, 2003. 173–180
- 58 Kazama J I, Miyao Y, Tsujii J I. A maximum entropy tagger with unsupervised hidden markov models. In: Proceedings of the 6th NLPRS, 2001 Nov 27–30, Tokyo. Citeseer, 2001. 333–340
- 59 Brants T. TnT: a statistical part-of-speech tagger. In: Proceedings of the Sixth Conference on Applied Natural Language Processing. Association for Computational Linguistics, 2000 Apr 29–May 4, Seattle. Association for Computational Linguistics, 2000. 224–231
- 60 Schmid H. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, 1994 Sep 14–16, Manchester. Citeseer, 1994. 44–49
- 61 Schmid H. Improvements in Part-of-Speech Tagging with an Application to German. In: Proceedings of the ACL SIGDAT-Workshop, 1995 Jun 26–30, Dublin. Citeseer, 1995. 47–50
- 62 Brill E. A simple rule-based part of speech tagger. In: Proceedings of the Workshop on Speech and Natural Language, 1992 Feb 23–26, Harriman. Association for Computational Linguistics, 1992. 112–116
- 63 Rimell L, Clark S. Porting a lexicalized-grammar parser to the biomedical domain. J Biomed Inform, 2009, 42: 852–865
- 64 Miyao Y, Tsujii J I. Feature forest models for probabilistic HPSG parsing. Comput Linguist, 2008, 34: 35–80
- 65 Schneider G, Hess M, Merlo P. Hybrid long-distance functional dependency parsing. Doctor Dissertation. Zürich: University of Zürich, 2008
- 66 Sagae K, Tsujii J I. Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles. In: EMNLP-CoNLL, 2007 Jun 28–30, Prague. Association for Computational Linguistics, 2007. 1044–1050
- 67 De Marneffe M C, MacCartney B, Manning C D. Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC, 2006 May 22–28, Genoa. Citeseer, 2006. 449–454
- 68 Petrov S, Barrett L, Thibaux R, et al. Learning accurate, compact, and interpretable tree annotation. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006 Jul 17–21, Sydney. Association for Computational Linguistics, 2006. 433–440
- 69 Charniak E, Johnson M. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005 Jun, Ann Arbor. Association for Computational Linguistics, 2005. 173–180
- 70 McClosky D. Any domain parsing: automatic domain adaptation for natural language parsing. Doctoral Dissertation. Rhode Island: Brown University, 2010
- 71 McClosky D, Charniak E, Johnson M. Effective self-training for parsing. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006 Jun 4–9, New York. Association for Computational Linguistics, 2006. 152–159
- 72 Bikel D M. Design of a multi-lingual, parallel-processing statistical parsing engine. In: Proceedings of the Second International Conference on Human Language Technology Research, 2002 Mar 24–27, San Diego. Morgan Kaufmann Publishers Inc., 2002. 178–182
- 73 Bikel D M. A Distributional Analysis of a Lexicalized Statistical Parsing Mode. In: EMNLP, 2004 Jul, Barcelona. Association for Computational Linguistics, 2004. 182–189
- 74 Dan K, Christopher D M. Fast Exact Inference with a Factored Model for Natural Language Parsing. Advances in Neural Information Processing Systems, vol. 15. MIT Press, 2002. 1–10
- 75 Campos D, Matos S, Oliveira J L. Gimli: open source and high-performance biomedical name recognition. BMC Bioinformatics, 2013, 14: 54
- 76 Huang M, Liu J, Zhu X. GeneTUKit: a software for document-level gene normalization. Bioinformatics, 2011, 27: 1032–1033
- 77 Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. In: Pacific Symposium on Biocomputing, 2008 Jan 4–8, Fairmont Orchid. Citeseer, 2008. 652–663
- 78 Neves M, Chagoyen M, Carazo J M, et al. CBR-Tagger: a case-based reasoning approach to the gene/protein mention problem. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, 2008 Jun 19, Columbus. Association for Computational Linguistics, 2008. 108–109
- 79 Hanisch D, Fundel K, Mevissen H T, et al. ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics, 2005, 6: S14
- 80 Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics, 2005, 21: 3191–3192
- 81 Miyao Y, Sagae K R, Matsuzaki T, et al. Evaluating contributions of natural language parsers to protein-protein interaction extraction. Bioinformatics, 2009, 25: 394–400

- 82 Miwa M, Pyysalo S, Hara T, et al. Evaluating dependency representation for event extraction. In: Proceedings of the 23rd International Conference on Computational Linguistics, 2010 Aug 23–27, Beijing. Association for Computational Linguistics, 2010. 779–787
- 83 Hoffmann R, Valencia A. A gene network for navigating the literature. *Nat Genet*, 2004, 36: 664–664
- 84 Fernández J M, Hoffmann R, Valencia A. iHOP web services. *Nucleic acids Res*, 2007, 35: W21–W26
- 85 Yang Z, Zhao Z, Li Y, et al. PPIExtractor: a protein interaction extraction and visualization system for biomedical literature. *Environ Technol*, 2013, 34: 1255–1266
- 86 He M, Wang Y, Li W. PPI finder: a mining tool for human protein-protein interactions. *PLoS One*, 2009, 4: e4554
- 87 Cheng D, Knox C, Young N, et al. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res*, 2008, 36: W399–W405
- 88 Kim S, Kwon D, Shin S Y, et al. PIE the search: searching PubMed literature for protein interaction information. *Bioinformatics*, 2012, 28: 597–598
- 89 Zhang S W, Li Y J, Xia L, et al. PPLook: an automated data mining tool for protein-protein interaction. *BMC Bioinformatics*, 2010, 11: 326
- 90 Czarnecki J, Nobeli I, Smith A M, et al. A text-mining system for extracting metabolic reactions from full-text articles. *BMC Bioinformatics*, 2012, 13: 172
- 91 Raja K, Subramani S, Natarajan J. PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. *Database*, 2013, 2013: bas052
- 92 Li C, Jimeno-Yepes A, Arregui M, et al. PCorral—interactive mining of protein interactions from MEDLINE. *Database*, 2013. 2013: p. bat030.
- 93 Kwon D, Sun K, Shin S Y, et al. Assisting manual literature curation for protein-protein interactions using BioQRator. *Database*, 2014, 2014: bau067
- 94 Li Y, Hu X, Lin H, et al. Learning an enriched representation from unlabeled data for protein-protein interaction extraction. *BMC Bioinformatics*, 2010, 11: 1
- 95 Song M, Yu H, Han W S. Combining active learning and semi-supervised learning techniques to extract protein interaction sentences. *BMC Bioinformatics*, 2011, 12: S4
- 96 Quan C, Wang M, Ren F. An unsupervised text mining method for relation extraction from biomedical literature. *PLoS One*, 2014, 9: e102039
- 97 Giles C B, Wren J D. Large-scale directional relationship extraction and resolution. *BMC Bioinformatics*, 2008, 9: 1

Progress of research in machine learning-based protein-protein interaction extraction

LI ManSheng, CHANG Cheng, MA Jie & ZHU YunPing

National Center for Protein Sciences (Beijing), State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China

Protein-Protein Interaction(PPI) is a kind of molecular event that plays a very important role in life activities. Automatic extraction and integration of PPIs from biological literature, which is one of the most important tasks of biological literature mining, can contribute to the research of molecular biology and has received widespread attention. Currently, machine learning-based PPI extraction methods have made great progress. It will help us to promote the improvement and application of these methods by reviewing their progresses. In this review, we firstly introduce the general construction process of machine learning-based PPI extraction method. Then, we summarize the classifiers, the features and the evaluation of machine learning-based PPI extraction method and PPI extraction tools. Finally, we discuss the prospects in this field.

protein-protein interaction, protein-protein interaction extraction, literature mining, machine learning

doi: 10.1360/N052016-00142