中国科学: 生命科学 2010 年 第 40 卷 第 9 期: 805 ~ 819
SCIENTIA SINICA Vitae www.scichina.com life.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述

蛋白质相互作用信息的文本挖掘研究进展

李满生[①], 刘齐军[①②], 李栋[①], 刘培磊[②], 朱云平[①*]

① 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, 北京 102206;
② 国防科学技术大学计算机学院, 长沙 410073
* 联系人, E-mail: zhuyp@hupo.org.cn

收稿日期: 2010-03-12; 接受日期: 2010-03-26
国家重点基础研究发展计划(批准号: 2006CB910803, 2006CB910706 和 2010CB912700)、国家高技术研究发展计划(批准号: 2006AA02A312)、国家重大科学研究计划(批准号: 2008ZX10002-016 和 2009ZX09301-002)、国家自然科学基金(批准号: 30800200)和蛋白质组学国家重点实验室课题(批准号: SKLP-Y200811 和 SKLP-O200811)资助项目

摘要 蛋白质相互作用是生命活动中一种极其重要的生物分子关系, 对此领域的研究不仅具有理论意义, 还具有较强的应用价值. 近年来, 随着研究的深入, 各种蛋白质相互作用的生物医学文献激增, 挖掘其中的蛋白质相互作用关系成为人们面临的一大挑战. 当前, 已提出了多种文本挖掘方法, 对分散于生物医学文献中的蛋白质相互作用信息进行结构化或半结构化处理. 对这些工作进行分析, 总结出基于生物文本挖掘蛋白质相互作用信息的一般流程, 从蛋白质命名实体的识别、蛋白质相互作用关系的提取和蛋白质相互作用注释信息的提取 3 个子任务进行阐述, 同时介绍了生物文本挖掘领域的评测会议和一些挖掘蛋白质相互作用相关信息的工具. 最后, 对该领域存在的一些重要问题进行分析, 并预测了未来可能的发展方向, 以期对该领域相关研究提供一定的参考.

关键词
蛋白质相互作用
文本挖掘
命名实体识别
关系提取
注释信息提取

后基因组时代, 蛋白质相互作用(protein-protein interaction, PPI)研究越来越受到人们的重视, 它是一种研究蛋白质功能的重要方法. 从遗传物质复制到基因表达调控, 从细胞信号转导到新陈代谢, 从细胞增殖到细胞凋亡, 蛋白质相互作用都发挥了非常重要的作用. 所以蛋白质相互作用研究不仅有利于从系统角度理解各种生物学过程, 揭示疾病的发生机制, 还可以帮助人们寻找新的药物靶标, 为新药研发起到积极的作用.

随着研究的不断深入, 报道蛋白质相互作用信息的文献激增. 美国国立医学图书馆的 Medline (http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=pubmed)数据库已存储了超过 108000 种杂志的 19000000 篇文献摘要或全文. 其中, 大量文献不仅指出了发生相互作用的成对蛋白质, 甚至还包括了蛋白质相互作用的亚细胞定位(sub-cellular location)、生物学过程(biological process)和生物学功能(biological function)等详细信息, 它们成为构建结构化蛋白质相互作用信息库所必要的数据基础. 但是, 文献数目巨大且增长迅速, 人们通过手工阅读文献往往难以及时、高效地发现其所关心的蛋白质相互作用信息. 文本介绍的挖掘方法(text mining)是一种解决这种“信息爆炸”问题的有效途径. 当前, 一些蛋白质相互作用数据库, 如 MINT[1]和 IntAct[2,3]已开始尝试利用文本挖掘技术搜集蛋白质相互作用数据, 自动提取蛋白质相互作用注释信息, 提高研究人员获取蛋白质相互作用信

英文引用格式: Li M S, Liu Q J, Li D, et al. Generating functional annotations for protein-protein interactions (in Chinese). SCIENTIA SINICA Vitae, 2010, 40: 805—819, doi: 10.1360/052010-133

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

息的效率, 并为通过高通量实验和计算方法预测得到的蛋白质相互作用数据提供相应的文献依据, 提高蛋白质相互作用的可信度. 不仅如此, 文本挖掘蛋白质相互作用还能够减少重复实验带来的资源浪费, 而且挖掘得到的蛋白质相互作用有详细的生物学实验支持, 真实可靠. 因此, 基于文本挖掘方法得到的蛋白质相互作用数据逐渐成为分子相互作用数据库十分重要的信息来源.

本文总结了生物文本挖掘中蛋白质相互作用信息提取的总体流程, 从蛋白质命名实体的识别、蛋白质相互作用关系的提取和蛋白质相互作用注释信息的提取 3 个子任务的方法及研究现状进行了阐述, 并介绍了生物文本挖掘领域的评测会议和一些挖掘蛋白质相互作用相关信息的工具. 最后, 分析了该领域当前存在的一些重要问题, 并预测了未来该领域可能的发展方向.

1 蛋白质相互作用信息的文本挖掘

蛋白质相互作用信息的文本挖掘流程一般包括 5 大模块(module): 语料选择与预处理模块(preprocessing module)、蛋白质命名实体识别模块(protein named entity recognition module)、PPI 关系提取模块(protein-protein interaction extraction module)、PPI 注释信息提取模块(annotations extraction module)和 PPI 信息可视化模块(visualization module)[4](图 1).

在以上流程中, 主要关注命名实体识别、相互作用关系提取和注释信息提取 3 个方面. 这 3 个方面彼此间较为独立, 都存在各自的一些语料库和文本挖掘方法, 但是都采用统一的评价标准对其方法进行性能分析.

1.1 文本挖掘方法的评价标准

为了衡量和比较蛋白质相互作用信息提取方法的性能, 人们给出了相应的实验语料(corpus, 文本挖掘研究中所使用的数据集), 还提出了相关的评价指标.

作为文本挖掘方法客观效果评价的金标准, 语料按照功能分为训练语料、开发语料和测试语料 3 种. 一般使用规模较小的、经过专家注释过(annotated)的训练语料和开发语料来训练系统以及调整系统的效果. 而测试语料是文本挖掘系统最终发布时文本挖掘的数据来源, 作为文本挖掘系统的开放测试和应用推广. 在生物医学领域, 文本挖掘系统的测试语料一般是大型的文献数据库, 如免费为学术研究提供全部文章摘要的 Medline 和提供其收录的生物医学文献全文的 BioMed Central(http://www.biomedcentral.com/info/about/datamining/). 这些测试语料的数据规模过于庞大, 得到的挖掘结果无法完全验证结果的正确性以及信息挖掘的完整性, 成为文本挖掘在生物文献领域应用中的一大难题. 一般从训练语料和开发语料上做完整性和准确性的评价. 另外, 可以将从测试集挖掘得到的数据与已有的数据库相应的数据做覆盖度比较.

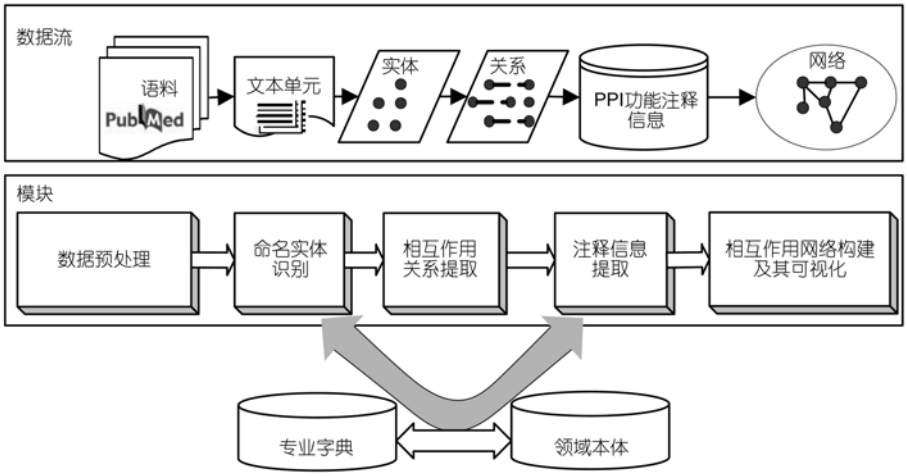


图 1 蛋白质相互作用信息提取流程

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在评价指标上, 通常利用 3 个常用的统计学指标来评价蛋白质相互作用信息提取结果: 准确率(precision)、召回率(recall)和两者的加权几何平均值(*F*-measure, *F* 值). 准确率反映了蛋白质相互作用信息提取的准确性, 召回率反映了信息提取的完整性, 另外, *F* 值用于综合评价蛋白质相互作用信息提取方法的效能. 对其定义如图 2 所示, 这 3 个指标可以分别应用于蛋白质命名实体识别, 蛋白质相互作用关系提取以及蛋白质相互作用注释信息提取任务中.

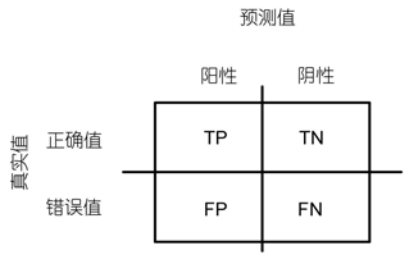
1.2　蛋白质命名实体识别

命名实体识别(named entity recognition)的任务定义为识别出文本中出现的专有名称和有意义的短语并加以归类. 在生物学领域的文本挖掘中, 命名实体识别就是从文本的句子中识别出生物实体的边界, 并判断其所属的类别. 蛋白质命名实体识别是生物命名实体识别任务中一项特定的、重要的研究内容, 是进行蛋白质相互作用关系提取和蛋白质相互作用注释信息提取两个任务的前提, 其识别效果将直接影响整个蛋白质相互作用信息提取系统的效果. 蛋白质命名实体识别任务可以分为两个子任务: 边界的确认和类别的判定. 这两个任务都与蛋白质名称的特征紧密相关, 除了不同于一些传统领域(如新闻领域)中的命名实体的数量稳定(如地名)和命名规范(如人名)的特征以外, 蛋白质名称还具有以下的命名特征: (ⅰ) 描述性的命名习惯: 许多蛋白质名称是描述性的, 由多个单词构成, 如“mitochondrial proton-transporting ATP synthase”, 名字较长, 难以确定其边界; (ⅱ) 具有包含(嵌套)关系: 蛋白质名称字符串中可能包含其他生物实体名子串, 如“epidermal growth factor”和“epidermal growth factor receptor”是两个不同的蛋白质名称, 这种情况也使蛋白质命名实体的边界难以确定; (iii) 非标准的命名习惯: 同一个蛋白名称可能有多种拼写形式, 如“immunoglobulin”, “immuno-globulin” 和 “immuno globulin”都是指同一蛋白. 另一方面, 功能无关的基因、蛋白质还可能出现同名的情况, 存在歧义; (iv) 缩写: 在生物医学文献中, 存在大量的蛋白质名称缩写词, 且缩写方法不规范, 有的是根据音节得到, 如“immunoglobulin 1”的缩写“Ig1”, 有的则用蛋白质全称中各单词首字母表示, 这种情况导致“TF”在不同的文章中可以是“transcription factor”和“tissue factor”的缩写.

以上这些问题使得蛋白质命名实体识别成为生物医学领域文本挖掘的一项重要的且具有挑战性的工作.

(1) 蛋白质命名实体识别相关语料. 语料是用于进行生物命名实体识别的数据集. 当前人们为生物命名实体识别任务发展了一些语料, 可用于对生物命名实体识别方法进行训练和评价, 概述如表 1.

在表 1 中, GENIA corpus[5]和 Medtag corpus[8]是应用最广泛的评测语料. GENIA corpus 是人类血细胞转录因子相关语料, 它包含了 2000 篇以“‘Humans’ [MeSH], ‘Blood Cells’[MeSH]和‘Transcription Factors’ [MeSH]”为关键词从 Medline 筛选的文章摘要, 该语料标注了 48 类生物实体. MedTag corpus 是一个由



TP: 从语料中挖掘得到的正确结果的数目; TP+FP: 从语料中挖掘得到的所有结果的数目; TP+FN: 语料中存在目标结果(包括正例和反例)的数目.

准确率(Precision):

$$Precision=TP/(TP+FP).....①$$

召回率(Recall):

$$Recall=TP/(TP+FN).....②$$

*F*值(*F*-measure):

$$F=[(1+\beta^2)(Precision\times Recall)]/(\beta^2\times Precision+Recall).....③$$

β^2是准确率和召回率的相对权重值, 如果将准确率和召回率的重要性同等对待, 则β值取1.

图 2　评价指标准确率、召回率和 *F* 值的定义

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 1　生物命名实体识别相关语料

名称	标注类型	规模	URL	发表年份
GENIA 语料[5]	48 种生物实体类型	2000 篇摘要	www.tsujii.is.s.u-tokyo.ac.jp/~genia/geniaform.cgi	1999
Medstract 语料[6]	基因, 名称缩写	49138 篇摘要	www.medstract.org/gold-standards.html	2001
Yapex 语料[7]	蛋白质	200 篇摘要	www.sics.se/humle/projects/prothalt/#data	2002
MedTag 语料(包括 AbGene, GENETAG)[8]	基因, 蛋白质	AbGene: 4000 个句子; GENETAG: 20000 个句子	ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedTag/medtag.tar.gz	2005
BioCreative 语料	基因, 蛋白质, GO 词汇	1000 个句子	www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/results	2004 2006
ProSpecTome 语料[9]	蛋白质	243 篇全文	http://textmining.cryst.bbk.ac.uk/ProSpecTome	2009

NCBI 提供的语料集合, 包括 AbGene, MedPost 和 GENETAG 3 个子语料库, 其中 AbGene 和 GENETAG 标注了基因、蛋白质名称, 而 MedPost 标注了文本中单词的词性标签.

(2) 蛋白质命名实体识别方法. 蛋白质命名实体识别是生物文本挖掘领域中的一个重要任务, 它是深入挖掘其他生物知识, 如蛋白质-基因、蛋白质-功能注释以及蛋白质-蛋白质等实体与实体之间关系的前提条件. 目前的生物命名实体识别的方法主要有基于字典(dictionary-based)、基于规则(rule-based)和基于机器学习(machine-learning)方法.

(ⅰ) 基于字典方法. 基于字典的生物命名实体识别方法是最早采用的一种识别方法, 采用基于基因和蛋白质名称字典的匹配方法搜索到相同或相似的字串, 用以从文本中识别出基因和蛋白质等实体. Proux 等人[10]首次尝试应用 FlyBase 数据库基因名称词典来识别 Medline 数据库语料中的基因. Torii 等人[11]使用生物医学用语词典识别基因和蛋白质两类实体, *F* 值达到了 88.7%. 当前可以用于生物命名实体识别的字典资源有 HUGO[12], NCBI Organism Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomy-home.html)和 Gene Synonym Lists[13]等. 由于基于字典的精确匹配方法准确性较高, 召回率较低, 为了提高系统的效果, 有研究尝试使用字典词条的变体扩充和模糊匹配来提高命名实体识别的召回率[14,15]. 该方法简单实用, 能够有效地寻找到字典范围中的蛋白质名称. 正是由于这一优点, 许多蛋白质相互作用关系抽取的研究往往使用基于字典的方法来识别实验语料中的蛋白质. 但由于字典的更新往往没有文献的更新速度快, 方法受限于字典的规模和质量, 在不同实验语料中的效果相差较大, 方法可移植性低.

(ⅱ) 基于规则方法. 一般情况下, 生物命名实体最常见且简单的组合规则是“大写字母”, “大写字母+数字”, “大写字母+连接符+数字”等. 因此, 根据预先定义的规则进行文本搜索即可获得生物命名实体. Yoshimasa 等人[16]采用启发式规则来最小化命名实体名称的歧义性和变化性, 实现了命名实体名称的标准化并进而提高了查找字典的效率. 基于规则方法的优点是, 规则可以按照需求灵活地加以定义和扩展. 但是由于生物实体的命名规则多样, 并不断涌现出新的形式, 手动的分析目标领域的文本并产生相应的规则需要花费大量时间. 方法在产生规则的语料中能获得较好效果, 但是普适性较差. 定义规则对于领域知识的依赖性很强, 需要专家参与. 另外, 由于基因和蛋白质等生物命名实体有相似的命名规则, 所以利用规则虽然较容易确认命名实体在文本中的边界, 但却不容易判断命名实体的类型, 需要上下文相关分析等复杂手段, 因此基于规则的方法较难达到理想效果.

(ⅲ) 基于机器学习方法. 将命名实体识别作为词分类问题, 则可以使用机器学习的方法来解决. 机器学习方法需要经过特征选择(feature selected)、分类学习和结果验证 3 个主要步骤. 特征选择是机器学习方法识别命名实体中的重要环节, 常见的被选取的特征有字典特征、词性特征(part of speech, POS)、词形特征和上下文特征(context)等. 文本中的每个词都用以上的特征集向量表示, 然后基于这些特征进行分类学习. 当前, 分类学习所使用的方法主要包括支持向量机(support vector machine, SVM)、隐

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

马尔科夫模型(hidden Markov model, HMM)、最大熵马尔科夫模型(maximum entropy Markov model, MEMM)和条件随机域(conditional random field, CRF)等. Li 等人[17]使用 GENETAG 语料, 将识别任务分为了命名实体检测和命名实体分类两个阶段, 在这两个子任务中均采用了 CRF 方法. 该方法能够有效地减少训练时间, 更重要的是每个子任务都能够挑选到相关性更大的分类特征, 最终获得了 74.31%的 *F* 值. 机器学习方法的优势在于比基于字典的方法更灵活, 可以发现现有蛋白质名称字典中未包含的实体, 并且可以根据上下文语境特征更准确地判断实体类型, 同时也避免了基于规则方法构建规则的繁重任务. 然而, 机器学习方法的效果依赖于训练语料规模和质量以及特征的选取. 目前, 寻求有效的特征仍是命名实体识别研究的热点. 总体来说, 基于机器学习方法能综合各方面的特征来提高命名实体识别效果, 所以是目前生物实体识别的核心方法.

事实上, 许多工作都基于以上 3 种方法, 进行一定程度的综合从而提高命名实体识别系统的性能. 基于字典方法通常与其他识别方法结合使用, 如整合到基于规则方法中, 来弥补自身规模和质量有限的缺陷. Hanisch 等人[18]首先使用一个预处理的基因和蛋白质同义词词典识别得到潜在的基因或蛋白质名称共现的生物医学文本, 然后用基于规则的搜索算法来获取多个词构成的名称. 在其随机测试使用的小鼠和果蝇数据集中, 该方法获得的 *F* 值将近 0.8, 而在酵母菌的数据集中 *F* 值高达 0.9. 字典特征还经常作为机器学习方法分类学习的一个有效特征[19~21]. 基于规则方法则常常被整合到基于机器学习方法的后期处理过程, 如实体的合并等[19,22~24].

1.3 蛋白质相互作用关系提取

(1) 蛋白质相互作用关系提取相关语料. 与命名实体识别任务一样, 蛋白质相互作用提取同样需要构建相应语料来训练、测试各类提取方法的准确率、召回率和 *F* 值, 衡量系统的性能. 所以, 人们在蛋白质命名实体识别语料的基础上, 进一步将语料中的蛋白质相互作用关系标注出来, 构成蛋白质相互作用关系语料, 常用于蛋白质相互作用提取的语料如表 2 所示.

由于语料的多样性(diversity)和有偏性(bias), 缺乏统一的蛋白质相互作用注释框架, 往往使得使用不同语料的不同研究方法无法比较. 为此, Pyysalo 等人[31]使用共出现方法和 RelEx 方法对表 2 中 5 个常用语料(AIMed, IEPA, LLL05, BioInfor 和 HPRD50)进行系统分析, 揭示了语料间的主要相似性和不同点, 并提供了统一格式的语料和转换这些语料的工具(http://mars.cs.utu.fi/PPICorpora), 向蛋白质相互作用关系标记语料标准化迈出了重要一步.

(2) 蛋白质相互作用关系提取方法. 从 20 世纪 90 年代末开始, 逐渐开展了关于蛋白质相互作用关系的文本挖掘工作. 目前, 蛋白质相互作用提取方法主要可以归纳为 3 大类: 基于规则方法(ruled-based approach)、基于统计和机器学习方法(statistical and machine-learning approach)与基于计算语言学方法(computational linguistics-based approach).

(ⅰ) 基于规则方法. 蛋白质相互作用关系在生物文献中的描述都有一定的语法规则, 如共出现、词性标记(part of speech, POS)的排列以及特定的描述词汇(如动词 interact, bind)等, 这些正是关系提取的依据. 该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. 基于规则的方法相对简单, 在早期的蛋白质相互作用

表 2 蛋白质相互作用关系提取相关语料

名称	句子规模(个)	URL	发表年份
AIMed 语料[25]	1955	ftp://ftp.cs.utexas.edu/pub/mooney/bio-data	2005
IEPA 语料[26]	486	http://class.ee.iastate.edu/berleant/s/IEPA.htm	2002
LLL05 语料[27]	77	http://genome.jouy.inra.fr/texte/LLLchallenge/	2005
BioCreative-PPI 语料	1000	http://www.mitre.org/public/biocreative/resources	2006
SPIES 语料[28]	891	http://spies.cs.tsinghua.edu.cn	2005
BioInfor 语料[29]	1100	http://www.it.utu.fi/BioInfer/	2007
HPRD50 语料[30]	145	http://www.bio.ifi.lmu.de/publications/RelEx/	2007
PPI 语料[30]	–	http://mars.cs.utu.fi/PPICorpora/	2007

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

提取研究中经常应用. 1999年, Ng和Wong[32]定义了基于词形的5个规则, 如<A>…<fn>…<B>匹配规则, 其中A, B表示蛋白质名称, 而fn表示描述相互作用关系的动词及其不同词形, 对蛋白质相互作用关系提取进行了初步探索. Blaschke等人[33]利用14个特定的蛋白质相互作用相关动词来构建表达蛋白质相互作用的规则模式, 然后从文本句子片段中识别相互作用. 显然, 这样简单的规则难以产生令人满意的结果. 系统的效果依赖于预定规则的质量和规模, 而且对于训练语料的针对性较强, 所以缺乏鲁棒性和可移植性. 2003年, Albert等人[34]利用蛋白质名称和相互作用词汇来识别两个蛋白和一个相互作用词汇三者共同出现的句子, 在Medline数据库中找到了3308个蛋白质相互作用, 准确率为22%. Huang等人[35]为了避免手工构建规则的繁重任务和规则的不全面性, 尝试了自动构建蛋白质相互作用模式(pattern). 首先, 将训练语料进行词性标记(POS tagging), 然后基于这些词性标记, 从句子中自动提取出类似的模式, 基于这些自动构建的模式识别出蛋白质相互作用关系. 在选定语料的实验结果获得了80.5%的准确率和80.0%的召回率. 这样的效果在基于规则方法提取蛋白质相互作用关系的研究工作中较少见.

(ii) 基于统计和机器学习方法. 当一个句子中存在两个或两个以上蛋白时, 蛋白质相互作用关系的提取可以看成是判断句子中有无蛋白质相互作用关系的二值分类问题. 该方法需要已注释的蛋白质相互作用关系训练语料, 从中提取蛋白质相互作用关系的有效表示特征, 使用支持向量机(support vector machine, SVM)[36]、最大熵模型(maximum entropy, ME)[37]等分类方法进行训练, 进而用于开发语料的蛋白质相互作用关系提取. 较少有基于机器学习方法取得良好的效果, 因为缺少高质量的、同时满足自然语言处理及机器学习的语料(训练集)[38]. 有研究对多个蛋白质相互作用语料进行了比较和评估. Airola等人[39]在多个语料上使用全路径图核方法(all-paths graph kernel)获取蛋白质相互作用信息, 对AIMed, BioInfor, HPRD50, IEPA和LLL05 5个语料, 使用五倍交叉验证(cross-validate)实验取得的平均 F 值为56.4%. Miwa等人[40]采用丰富特征向量(rich feature vector)及考虑语料权重的支持向量机(SVM-CW)方法从多个语料中提取蛋白质相互作用, 取得了65.2%的 F 值, 是目前已报道的蛋白质相互作用关系提取技术的最好水平.

(iii) 基于计算语言学方法. 计算语言学方法利用语法规则和句法分析(parsing)等技术, 根据句法结构判断蛋白质间是否存在相互作用关系. 语法规则在前面已作介绍. 句法分析有浅层句法分析(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(NP)、动词短语(VP), 整合成主语(subject)、宾语(object)等, 得到清晰的句子主干, 方便进一步的关系挖掘. 深层句法分析和全解析最终得到句子的完整句法树(syntax tree). 常用的深层句法分析有基于中心语驱动短语结构语法的句法分析(HPSG parsing)和依存分析(dependency parsing)等. Matsuzaki等人[41]基于中心语驱动短语结构语法进行深层句法分析, 开发了能够基于特殊语义关系的语法特征的搜索引擎, 对Medline查找蛋白相互作用关系. Fundel等人[30]应用自然语言处理中的依存分析产生句法分析树(依存分析树), 然后基于语法规则, 提取蛋白质相互作用关系, 方法的准确率和召回率都达到了80%. 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, 为判断蛋白质相互作用关系取得更好的效果.

综上所述, 在蛋白质相互作用关系提取方法中, 基于规则的方法可以按照需求灵活地定义和扩展规则, 对特定的语料简单有效, 但是手动地分析目标领域的文本并产生相应的规则需花费大量时间, 并且需要有专家参与, 方法的可移植性差. 机器学习方法能够通过对已经标注好的语料进行自动地训练模型, 提取实体相互作用模式, 进而对关系进行判断, 避免了大量的人工定义规则, 但是机器学习方法强烈依赖于训练集规模和质量以及特征值选取. 基于计算语言学方法不需要大规模的训练语料却具有较好的移植性, 对于简单句子中的相互作用关系较为有效, 但处理复杂句子时方法的效果依赖于句法分析工具的准确性. 在提取蛋白质相互作用关系时, 常结合使用几种方法, 从而提高效果. Huang等人[35]为了避免手工构建规则的繁重任务和规则的不全面性, 从构建好的语料出发, 使用机器学习方法提取蛋白质相互作用的描述规则, 从而在确保方法准确率的同时

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

提高方法的召回率.

(3) 蛋白质相互作用关系挖掘的信息存储. 随着文本挖掘技术和方法研究的深入, 对蛋白质相互作用关系挖掘任务的尝试逐渐突破了小规模语料和方法学研究的层次, 应用到实际生物文本挖掘中. 虽然一些方法在小规模测试语料上性能良好, 但在实际大规模文献上挖掘结果的完整性和准确性难以得到验证[42], 所以这些方法通常用于开发在线工具, 协助人们从现有的文献数据库 Medline 等挖掘相关领域的蛋白质相互作用. 蛋白质相互作用数据库为了保证收录数据的质量, 常常收集有文献支持的蛋白质相互作用数据作为其核心数据集, 由于这样的数据集需要人工判读文献, 数据集规模都较小. 现有的蛋白质相互作用数据库收集高通量实验和计算方法预测的数据集外, 还正在逐渐地整合文本挖掘来源的蛋白质相互作用数据集, 给研究者提供有效的信息(表 3).

IntAct 和 MINT 蛋白质相互作用数据库主要的数据来源为文献, 为协助 BioCreative Ⅱ 的 PPI 提取任务评测, 不仅提供了开发算法所用的训练集, 还提供了评估文本挖掘工具所需的测试集. IMEx(http://disber.net/imexdrupal/)是一个蛋白质相互作用数据库联盟, 包含 DIP, IntAct, MINT, Mpact, MatrixDB, MPIDB, BioGRID 等数据库. 它收集数据的原则是 PPI 数据必须来自实验证实或是文献来源的可靠数据. 目前, 以上这些具有文献支持的数据正被广泛地运用于物种蛋白质相互作用网络的构建和网络特性分析等工作中[54,55].

1.4 蛋白质相互作用注释信息提取

完成蛋白质相互作用关系的提取, 可以有效地构建蛋白质相互作用网络. 然而, 这样得到的蛋白质相互作用网络是简单叠加的静态网络, 无法准确地反映细胞系统内动态的蛋白质相互作用网络. 因为蛋白质相互作用是一个生物分子事件(biological molecular event), 具备时空(spatial-temporal)特性, 在一定的细胞过程、一定的细胞位置(亚细胞定位)条件下才会发生. 为了准确构建蛋白质相互作用网络或者进行不同蛋白质相互作用间的比较分析和分类(classification), 还需更深入提取这些与蛋白质相互作用密切相关的注释信息(如相互作用的类型、方向、检测方法及功能注释等). 在 BioCreative Ⅱ[56]评测会议的蛋白质相互作用提取任务中, 要求在获取蛋白质相互作用及蛋白质名称命名标准化信息后, 检索蛋白质相互作用的检测方法以及相互作用证据等注释信息. Wang 等人[57]使用机器学习的方法, 从文本中挖掘蛋白质相互作用的检测方法. 目前, 这类提取蛋白质相互作用注释信息的研究工作较少. 2009 年, 生物事件提取(bio-event extraction)作为一个全新的公开评测任务被提出(BioNLP'09[58]), 说明生物事件的挖掘提取是当前及未来生物文本挖掘研究领域

表 3 蛋白质相互作用数据库资源及数据源情况

数据库名称	蛋白数目	PPI 数目	数据格式	数据来源	URL
DIP[43]	21891	69171	PSI-MI	文献获取	http://dip.doe-mbi.ucla.edu/dip/Stat.cgi
MIPS[44]	>900	>1800	PSI-MI	文献获取	http://mips.helmholtz-muenchen.de/proj/ppi/
BIND[45]	–	188517	PSI-MI	高通量实验、文献获取	http://bond.unleashedinformatics .com/
HPRD[46]	27081	38806	PSI-MI	文献获取	http://www.hprd.org/
MINT[47]	30300	82442	PSI-MI	高通量实验、文本挖掘	http://mint.bio.uniroma2.it/
IntAct[48]	63142	209203	PSI-MI	文献获取	http://www.ebi.ac.uk/intact/main.xhtml
CCSB-LIT[49]	2192	4067	PSI-MI	文本挖掘	http://interactome.dfci.harvard.edu/
COCIT[50]	3737	6580	PSI-MI	文本挖掘	–
MPI-LIT[51]	940	746	PSI-MI	文献获取、数据整合	http://www.jcvi.org/mpidb/interaction.php?dbsource=MPI-LIT
STRING[52]	2590259	–	PSI-MI	基因组信息推断、高通量实验、共表达信息推断和现有知识推断	http://string-db.org/
BioGRID[53]	548207	335416	PSI-MI	数据整合、文献获取	http://www.thebiogrid.org/

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的重要方向. 生物医学信息抽取领域对于生物实体关系动态注释信息的关注, 将从狭义的生物事件提取(生物实体关系提取)扩展到广义的生物事件提取(如蛋白质的磷酸化、事件的时空条件信息等). 蛋白质相互作用信息的提取正是这个评测任务的重要部分, BioNLP'09 生物事件提取评测会议为蛋白质相互作用提取提供了一个新的视角, 促进了蛋白质相互作用注释信息的重视和深入挖掘研究.

对蛋白质相互作用信息的理解并不统一. Duan 等人[59]使用了蛋白质状态(protein states)和状态转换(state transitions)的词汇来描述蛋白质的相互作用, 并构建了基于这些词汇的 LiveDIP 数据库. Ratsch 等人[60]认为, 一个蛋白质相互作用可以表示为一个具有前后条件的生物事件. 蛋白质相互作用(PPI)是一个发生在特定条件下的生物事件(event), 具有蛋白质相互作用方式(how)、所处的生物学过程(when)、亚细胞定位(where)等属性. 为了进一步明确作为生物分子事件的蛋白质相互作用的注释信息的具体内涵, 需要构建一个包含以上信息的蛋白质相互作用本体. PSI-MI(molecule interaction)标准[61]虽然是目前广泛采用的描述蛋白质相互作用数据的本体, 但是其注重于实验信息描述, 不适合蛋白质相互作用信息的文本挖掘. 基因本体(gene ontology, GO)[62]是目前生命科学领域中最全面的基因和蛋白质功能注释本体, 其相关的概念同样适用于蛋白质相互作用功能信息的注释. 将这两个本体整合起来, 将能更好地描述蛋白质相互作用. 一个完善的蛋白质相互作用本体能更好地表达蛋白质相互作用信息, 进而能更好地促进蛋白质相互作用注释信息的深入挖掘.

另一方面, 在生物医学信息抽取领域, 已有较多基于本体的文本挖掘工作, 核心的任务是将本体的受控词汇(controlled vocabularies)映射到文本上, 从而得到基因或蛋白质的功能注释信息. 文本挖掘借助具有良好框架的本体, 能将散布在文本中的非结构化生物知识结构化[63], 从而将表示领域知识的生物文献、本体和生物数据库统一起来[64,65]. Doms 和 Schroeder[66]使用正则表达式方法从生物医学文献数据库 PubMed 的摘要中进行 GO 受控词汇的提取, 为检索到的基因和蛋白质相关文献提供一个基于基因本体的树形层次化视图. Muller 等人[67]收集了 33 类生物领域本体, 包括细胞、基因、生物学功能、调控、疾病和药物等, 然后给这些本体的词汇构建相应的正则表达式, 在小规模的语料上测试了将这些正则表达式表达的词汇匹配到生物医学文献的效果. Huang 等人[68]整合了来自 GO, MeSH, LocusLink 和 OMIM 生物本体和词表资源构成词汇集, 归纳成了基因、蛋白质、疾病等 7 类生物实体, 然后自动学习这些词汇的模式(pattern), 将模式匹配到文本时采用打分的方法控制挖掘结果的可靠性. 这些方面的工作虽然没有直接对于蛋白质相互作用注释信息进行提取, 但是提取基因和蛋白质的功能注释信息对于蛋白质相互作用注释信息的提取具有提示、支撑和促进的作用.

综上, 基于文本挖掘方法的蛋白质相互作用信息提取工作大多集中于对相互作用关系的提取, 虽然人们开始关注蛋白质相互作用信息的挖掘并开展了一些工作, 但是并没有明确蛋白质相互作用信息挖掘的范畴及具体的挖掘对象, 更缺乏蛋白质相互作用信息描述的本体, 大多数研究并没有进一步挖掘蛋白质相互作用的注释信息. 基于本体的文本挖掘工作关联了基因或蛋白质与其注释信息, 这部分工作可以借鉴到蛋白质相互作用注释信息的提取工作中, 为重构(re-construction)条件特异蛋白质相互作用网络(动态生物网络)提供更为可靠的数据.

1.5 生物文本挖掘方法的评测会议

为了评估和促进生物文本挖掘领域的信息检索和信息抽取任务的技术发展水平, 近年来该领域的一些组织举行了很多生物文本挖掘方法的评测会议. 会议提出明确的评测任务, 这些评测任务是生物文本挖掘中的一些热点研究问题或挑战性任务. 通过提供公共的训练语料集和测试语料集, 可以很好地比较当前各个任务的研究情况. 生物命名实体识别的主要评测会议有 JNLPBA2004 和 BioCreative Ⅰ (2004), 而实体关系抽取的评测会议主要有 LLL'05 和 BioCreative Ⅱ (2006).

JNLPBA2004 以 GENIA 为训练语料, 对 protein, DNA, RNA, 细胞系和细胞类型 5 类实体进行识别. 由测评结果可知, CRFs 是该评测中取得最好效果的模型, CRFs 系统只使用了很少种类的特征, F 值就达到了 69.8%.

BioCreative 至今已经举办了 3 次评测, BioCreative Ⅰ (2004)主要的评测任务是基因名称识别(gene name identification)、基因名称标准化(gene normalization)

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

和自动提取基因的 GO 注释. 该评测以 GENETAG 为训练语料识别基因名称, 然后将基因名称和特定的基因数据库关联(基因名称标准化), 同时关注文本中基因的 GO 注释. BioCreative Ⅱ(2006)和 BioCreative Ⅱ.5 (2009)除了继续关注基因名称标识和基因名称标准化, 还提出提取文本中的蛋白质相互作用关系任务. BioCreative Ⅱ的蛋白质相互作用关系提取任务分为 4 个子任务[56]: (ⅰ) 获取蛋白质相互作用相关文章; (ⅱ) 蛋白质相互作用对的提取及命名标准化; (ⅲ) 识别用于蛋白质相互作用检测的方法; (ⅳ) 检索提供蛋白质相互作用证据的文章. 在 BioCreative Ⅱ评测任务(ⅱ)中, 效果最好的方法准确率为 37%, 召回率为 33%.

LLL'05 首次提出了生物实体关系提取的挑战任务, 该评测要求学习规则以从 Medline 数据库摘要中提取基因或蛋白质的相互作用关系. 会议为生物实体关系抽取提供了 77 个句子的训练语料.

BioNLP'09 第一次提出生物事件抽取的公开评测任务, 有 42 个研究小组参加, 最终有 24 个小组提交了最终的结果. BioNLP'09 主要关注基因或蛋白质事件, 并将生物事件的提取划分为 3 个子任务: 核心事件的提取(如蛋白质的磷酸化)、事件信息的富集(如生物事件的定位信息)和否定或推测类型事件的识别. 该评测会议为生物事件的文本挖掘提供了一个基于 GENIA 事件本体(GENIA event ontology)框架上的语料集 GENIA 事件语料[69]. 该语料包含了 GENIA 语料的一半, 由 1000 篇 Medline 摘要构成, 有 9372 个句子的 36144 个事件被注释.

表 4 显示了评测会议关注的一些热点研究问题和挑战性任务, 及其指定的一些标准的评测语料.

这一系列评测会议反映了生物文本挖掘领域研究的动态, 其评测结果提示了文本挖掘技术在实际应用中的水平, 并指明了该领域一些潜在的挑战, 对于文本挖掘应用于生物领域的研究具有促进作用.

1.6 用于蛋白质相互作用信息提取的工具

基于以上命名实体识别和蛋白质相互作用关系提取方法, 已开发了多种工具用于完成相应的任务, 协助研究人员完成蛋白质相互作用信息的挖掘(表 5).

大多数的生物医学文本挖掘工具都是基于以上挖掘任务的一种或多种方法, 以下介绍一些相关系统.

LingPipe 是一个自然语言处理的软件包, 其中包含了文本主题分类(topic classification), 命名实体识别(named entity recognition)等多个常用自然语言处理的子模块. 它的命名实体识别模块提供了基于规则和基于统计等方法来识别基因、蛋白质名称, 该系统提供 GENETAG 语料训练的命名实体识别模型, 可以由用户自己定义训练语料来训练得到更具有针对性的模型, 提高了系统的可移植性.

ABNER 是一个生物命名实体识别工具, 它使用了条件随机域(CRFs)方法来学习生物命名实体的特

表 4 生物文本挖掘(Bio-TM)领域的重要评测会议分析与比较 [a)]

评测会议名称	关注领域	主要评测任务	评测语料
TREC Genomics track (03~07)[70]	Bio-IR	文本检索和分类	TREC Genomics Track 语料
JNLPBA2004[71]	Bio-NER	生物命名实体(蛋白质, DNA, RNA,细胞系和细胞类型)识别	GENIA 语料
BioCreative Ⅰ (2004)[72]	Bio-NER Bio-IE	任务 1: 基因名称识别 任务 2: 基因名称标准化 任务 3: 基因功能注释提取	GENETAG 语料 BioCreative 语料
LLL'05 (http://www.cs.york.ac.uk/aig/lll/lll05)	Bio-IE	生物实体关系提取	LLL05 语料
BioCreative Ⅱ (2006)[2]	Bio-NER Bio-IE	任务 1: 基因名称标记 任务 2: 基因、蛋白质名称标准化 任务 3: 蛋白质相互作用关系提取	GENETAG 语料
BioNLP'09 (http://www.cs.york.ac.uk/aig/lll/lll05)	Bio-IE	任务 1: 核心事件提取 任务 2: 事件信息富集 任务 3: 否定和推测事件识别	GENIA 事件语料

a) Bio-IR: 生物医学信息检索; Bio-NER: 生物医学命名实体识别; Bio-IE: 生物医学信息提取

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

征, 该系统提供了 JAVA 程序的 API 使用基于 BioCreative 语料和 NLPBA 语料所训练的模型, 来识别文本中的基因, 蛋白质, DNA, RNA, 细胞类型和细胞系等实体, 在两个语料上评估的 F 值分别是 69.9%和 70.5%.

MedScan 为生物实体关系提取系统, 基于自然语言处理方法实现. 它的关系提取框架分为 3 层: 文本预处理模块、自然语言处理模块和关系抽取模块, 关系抽取模块可以根据挖掘的实体关系类型更换, 从而使系统的应用更灵活.

近年来, 网络服务(web services)和集成可视化工具是蛋白质相互作用信息挖掘工具的发展趋势. ProteinCorral 和 EBIMed 提供了网络服务接口, 可以根据关键词或 PubMed ID 获取注释蛋白质名称或 GO 词汇的 Medline 摘要. ProteinCorral 是一个 web 服务软件, 它提供 3 种方法来挖掘两个蛋白质之间的相互作用关系, 分别为两个蛋白质名称共出现(co-occurrence, CO)方法、两个蛋白质名称及动词三者共出现(CO3)方法和自然语言处理判定(natural language processing, NLP)方法, 其中自然语言处理判断的准确性最高. 使用 PubMed ID 或蛋白质名称搜索, 可以得到以上 3 种方法标注出蛋白质相互作用关系的 PubMed 摘要. EBIMed 工具利用 UniProt 提供的蛋白质名称字典, 基因本体(GO)提供的功能注释词汇, MedlinePlus 提供的药物名称字典和 NCBI Taxonomy 提供的物种信息词汇, 将蛋白质和以上信息相联系. Ali BaBa 则可以根

表 5 蛋白质相互作用信息挖掘的相关工具

任务类型	工具			
	名称	URL	发表年份	平台/使用方式
蛋白质名称识别	AbGene-tagger[73]	ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe/AbGene/	2002	Linux
	Yapex-tagger[7]	www.sics.se/humle/projects/prothalt/#tagger http://ellis.sics.se: 8080/cgi-bin/Yapex/yapex.cgi	2002	Online
	LingPipe	http://alias-i.com/lingpipe/	2004	Java API
	GENIA-tagger[74]	www.tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/	2005	Linux
	ABNER[75]	http://pages.cs.wisc.edu/~bsettles/abner	2005	Java API
	ProMiner[18]	www.scai.fraunhofer.de/prominer	2005	Online
	BANNER[76]	http://banner.sourceforge.net/	2008	Java
相互作用关系提取	PreBIND[36]	http://bind.ca/	2003	Online
	Chilibot[77]	www.chilibot.net	2004	Online
	BioRAT[78]	http://bioinf.cs.ucl.ac.uk/biorat/	2004	Windows
	Intex[79]	http://cips.eas.asu.edu/textmining.htm	2005	Online
	MEDIE[80]	http://www-tsujii.is.s.u-tokyo.ac.jp/medie/	2006	Online
	Info-PubMed[80]	www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/	2006	Online
	ProteinCorral[81]	www.ebi.ac.uk/Rebholz-srv/pcorral/index.jsp	2008	Web Services
	PPI finder[82]	http://liweilab.genetics.ac.cn/tm/	2009	Online
注释信息提取	TerMine[83]	http://www.nactem.ac.uk/software/termine	2000	Web Services
	textpresso[67]	www.textpresso.org	2004	Linux
	GoPubmed[66]	www.gopubmed.com	2005	Online
	ONBIRES[68]	http://spies.cs.tsinghua.edu.cn: 8080 http://60.195.250.72/onbires	2006	Online
	MaxMatche[84]	http://dragon.ischool.drexel.edu/example/maxmatcher.zip	2006	Java API
	GOAnnotator[85]	http://xldb.di.fc.ul.pt/rebil/tools/goa	2006	Online
	OBO-Annotator	www.ohloh.net/p/obo-annotator http://sourceforge.net/projects/obo-annotator/	2008	Linux, Java API
	EBIMed[86]	www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp	2007	Web Services
	MedEvi[87]	www.ebi.ac.uk/Rebholz-srv/MedEvi/	2008	Online
集成网络可视化工具	MedScan[88]	www.ariadnegenomics.com/products/medscan.html	2003	Windows
	Cytoscape[89]	www.cytoscape.org/	2003	Windows
	ONDEX[90, 91]	www.ondex.org/	2005	Windows
	GeneWays[92]	http://geneways.genomecenter.columbia.edu/	2004	Windows
	iHop[93]	www.ihop-net.org/ http://www.ihop-net.org/UniPub/iHOP	2005	Web Services, Online
	AliBaba[94]	http://alibaba.informatik.hu-berlin.de/	2006	Online

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

据搜索关键词, 直接生成生物实体网络, 生物实体的类型有细胞、疾病、药物、蛋白质、物种和组织, 通过该软件, 可以方便地查询和可视化目标生物实体相关网络.

2 蛋白质相互作用信息提取存在的问题与挑战

蛋白质相互作用信息的文本挖掘研究在近年来取得了一定的进展, 在蛋白质命名实体识别和蛋白质相互作用关系提取等子任务上都开发了相应的评测语料和有效的解决方法, 但考虑到生物医学领域研究的需求, 几个方面的问题仍然值得在文本挖掘任务中深入探讨.

(ⅰ) 基因与蛋白质实体名称混淆不清. 在生物命名实体识别任务中, 即使能够很好地识别文本中命名实体的边界, 还要准确判断命名实体的语义类型, 到底该名称属于基因、蛋白质还是其他的生物实体. 但由于基因和蛋白质的命名特征相似, 准确判别其类型比较困难. 命名实体准确识别仍然是蛋白质相互作用信息挖掘准确性提高的首要因素.

(ⅱ) 蛋白质相互作用关系提取中的问题. 蛋白质相互作用关系是一个很大的范畴, 具有众多类型, 如物理相互作用(physical interaction)、遗传相互作用(genetic interaction, 又称功能相互作用)等, 另外还有在共表达(co-express)关系、共定位(co-location)关系等. 现有的蛋白质相互作用关系提取研究并未能充分区分以上的关系进行挖掘.

(iii) 蛋白质相互作用功能注释信息提取研究欠缺. 目前还缺乏一个可以较好地表达蛋白质相互作用必要信息的本体, 蛋白质相互作用信息文本挖掘工作多只关注蛋白质相互作用对的挖掘, 而忽略相互作用对应的功能注释信息(生物学过程、亚细胞定位及生物学功能)提取. 有部分工作关注了基于本体的蛋白质与其功能注释的关联挖掘, 孤立地理解蛋白质功能, 而非从相互作用的角度解读蛋白质的功能.

另外, 文本挖掘技术本身存在着许多挑战: 科学文本的复杂特性、语言表述的多样性(句法结构上的难度)和基因或蛋白质名称的多个同义词以及一直变化的科学词汇, 导致对文本语义的理解相当困难, 使蛋白质相互作用信息提取成为一项极具挑战的任务.

3 蛋白质相互作用信息提取研究展望

目前, 随着生物医学文献的剧增, 人们通过手工阅读文献已经难以及时、高效地获取信息. 文本挖掘方法是解决该问题的有效途径. 人们已经将文本挖掘方法运用于蛋白质命名实体识别、蛋白质相互作用关系的提取等任务, 还需要更深入地挖掘蛋白质相互作用注释信息. 为了进一步解决目前所面临的问题, 近期蛋白质相互作用信息的文本挖掘将可能有几个重要的发展方向.

(ⅰ) 随着生物命名实体名称的标准化研究深入, 命名实体(如基因和蛋白质)名称不断规范, 在此基础上进一步提高命名实体识别的类别判断准确性.

(ⅱ) 深入挖掘蛋白质相互作用信息, 如蛋白质相互作用关系类型、相互作用条件等. 在蛋白质相互作用对以及相互作用动词的提取研究成果基础上, 深入地挖掘蛋白质相互作用的功能注释信息, 为蛋白质相互作用网络和生物实体知识网络的构建提供更充分的信息.

(iii) 发展完善蛋白质相互作用信息本体. 蛋白质相互作用信息本体给蛋白质的相互作用的功能注释提供注释标准, 能更好地统一蛋白质相互作用注释, 也能促进蛋白质相互作用信息的挖掘. 蛋白质相互作用关系挖掘和基于本体的功能注释信息挖掘, 很有可能进行有效地结合, 从而达到丰富蛋白质相互作用信息的目的.

(ⅳ) 蛋白质相互作用与疾病、药物等信息的关系挖掘会得到生物学家的重视, 这方面的信息直接提示潜在的疾病标志物和药物靶标. 通过分析文本挖掘得到的蛋白质相互作用网络, 提示其中可能具有分子标记物或药物靶标意义的关键节点(蛋白质), 从而展开与分子生物学家的合作.

在“信息爆炸”的时代, 文本挖掘方法将会根据蛋白质相互作用信息提取的问题, 将非结构化的文本信息转化成结构化或半结构化信息, 从而提高信息获取的效率, 显示其在生命科学研究中的应用价值.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献

1 Krallinger M, Erhardt R, Valencia A. Text-mining approaches in molecular biology and biomedicine. Drug Discov Today, 2005, 10: 439—445

2 Krallinger M, Morgan A, Smith L, et al. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biol, 2008, 9: S1

3 Chatr-aryamontri A, Kerrien S, Khadake J, et al. MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. Genome Biol, 2008, 9: S5

4 Zhou D, He Y. Extracting interactions between proteins from the literature. J Biomed Inform, 2008, 41: 393—407

5 Kim J, Ohta T, Tateisi Y, et al. GENIA corpus-a semantically annotated corpus for bio-textmining. Bioinformatics, 2003, 19: 180—182

6 Pustejovsky J, Castano J, Cochran B, et al. Extraction and disambiguation of acronym-meaning pairs in medline. Medinfo, 2001, 10: 371—375

7 Franzen K, Eriksson G, Olsson F, et al. Protein names and how to find them. Int J Med Inform, 2002, 67: 49—61

8 Smith L, Tanabe L, Rindflesch T, et al. MedTag: a collection of biomedical annotations. Proc of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, 2005. 32—37

9 Kabiljo R, Stoycheva D, Shepherd A. ProSpecTome: a new tagged corpus for protein named entity recognition. Proc of The ISMB BioLINK, Special Interest Group on Text Data Mining, 2007, 19: 24—27

10 Proux D, Rechenmann F, Julliard L, et al. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. Genome Inform, 1998, 9: 72—80

11 Torii M, Hu Z, Wu C, et al. BioTagger-GM: a gene/protein name recognition system. J Am Med Inform Assoc, 2009, 16: 247—255

12 Povey S, Lovering R, Bruford E, et al. The HUGO gene nomenclature committee (HGNC). Human Genetics, 2001, 109: 678—680

13 Ari J, Jensen L, Ouzounova R, et al. Extracting regulatory gene expression networks from pubmed. Proc of the 42nd Annu Meeting on Association for Computational Linguistics, 2004, 191—198

14 Tsuruoka Y, Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition. J Biomed Inform, 2004, 37: 461—470

15 Tsuruoka Y, McNaught J, Tsujii J, et al. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics, 2007, 23: 2768

16 Yoshimasa T, John M, Sophia A. Normalizing biomedical terms by minimizing ambiguity and variability. BMC bioinformatics, 2008, 9: S2

17 Li L, Zhou R, Huang D. Two-phase biomedical named entity recognition using CRFs. Comput Biol Chem, 2009, 33: 334—338

18 Hanisch D, Fundel K, Mevissen H, et al. ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics, 2005, 6: S14

19 Lin Y, Tsai T, Chou W, et al. A maximum entropy approach to biomedical named entity recognition. Proc of BioKDD'04, 2004, 56—61

20 McDonald R, Pereira F. Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics, 2005, 6: S6

21 Mitsumori T, Fation S, Murata M, et al. Gene/protein name recognition based on support vector machine using dictionary as features. BMC Bioinformatics, 2005, 6: S8

22 Tsai T, Chou W, Wu S, et al. Integrating linguistic knowledge into a conditional random fieldframework to identify biomedical named entities. Expet Syst Appl, 2006, 30: 117—128

23 Tsai R, Sung C, Dai H, et al. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. BMC bioinformatics, 2006, 7: S11

24 Zhou G, Zhang J, Su J, et al. Recognizing names in biomedical texts: a machine learning approach. Bioinformatics, 2004, 20: 1178—1190

25 Bunescu R, Ge R, Kate R, et al. Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med, 2005, 33: 139—155

26 Ding J, Berleant D, Nettleton D, et al. Mining Medline: Abstracts, sentences, or phrases? Pac Symp Biocomput, 2002,7: 326—337

27 Nédellec C. Learning language in logic-genic interaction extraction challenge. Proc of the 4th Learning Language in Logic Workshop (LLL05), 2005, 31—37

28 Hao Y, Zhu X, Huang M, et al. Discovering patterns to extract protein-protein interactions from the literature: Part II. Bioinformatics, 2005, 21: 3294

29 Pyysalo S, Ginter F, Heimonen J, et al. BioInfer: A corpus for information extraction in the biomedical domain. BMC Bioinformatics, 2007, 8: 50

30 Fundel K, Kuffner R, Zimmer R. RelEx--relation extraction using dependency parse trees. Bioinformatics, 2007, 23: 365

31 Pyysalo S, Airola A, Heimonen J, et al. Comparative analysis of five protein-protein interaction corpora. BMC bioinformatics, 2008, 9: S6

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

32 Ng S and Wong M. Toward routine automatic pathway discovery from on-line scientific text abstracts. Genome Inform, 1999: 104—112
33 Blaschke C, Andrade M, Ouzounis C, et al. Automatic extraction of biological information from scientific text: protein-protein interactions. Proc Int Conf Intell Syst Mol Biol, 1999: 60—67
34 Albert S, Gaudan S, Knigge H, et al. Computer-assisted generation of a protein-interaction database for nuclear receptors. Mol Endocrinol, 2003, 17: 1555—1567
35 Huang M, Zhu X, Hao Y, et al. Discovering patterns to extract protein-protein interactions from full texts. Bioinformatics, 2004, 20: 3604—3612
36 Donaldson I, Martin J, de Bruijn B, et al. PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics, 2003, 4: 11
37 Xiao J, Su J, Zhou G, et al. Protein-protein interaction extraction: a supervised learning approach. Proc Int Symp on Semantic Mining in Biomedicine (SMBM), 2005: 51—59
38 Kim S, Yoon J, Yang J. Kernel approaches for genic interaction extraction. Bioinformatics, 2008, 24: 118
39 Airola A, Pyysalo S, Bj rne J, et al. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinformatics, 2008, 9: S2
40 Miwa M, S tre R, Miyao Y, et al. A Rich Feature Vector for Protein-Protein Interaction Extraction from Multiple Corpora. Proc of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, 121—130
41 Matsuzaki T, Miyao Y and Tsujii J. Efficient HPSG parsing with supertagging and CFG-filtering. Proc of the 20th international joint conference on Artifical intelligence, 2007, 1671—1676
42 Rzhetsky A, Seringhaus M, Gerstein M. Seeking a new biology through text mining. Cell, 2008, 134: 9—13
43 Xenarios I, Salwinski L, Duan X, et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucl Acids Res, 2002, 30: 303
44 Pagel P, Kovac S, Oesterheld M, et al. The MIPS mammalian protein-protein interaction database. Bioinformatics, 2005, 21: 832
45 Gilbert D. Biomolecular interaction network database. Brief Bioinform, 2005, 6: 194
46 Prasad T, Goel R, Kandasamy K, et al. Human Protein Reference Database——2009 update. Nucl Acids Res, 2009, 37: D767
47 Ceol A and Aryamontri C. MINT, the molecular interaction database: 2009 update. Nucl Acids Res, 2010, 38: D532
48 Aranda B, Achuthan P, Alam-Faruque Y, et al. The IntAct molecular interaction database in 2010. Nucl Acids Res, 2010, 38: D525
49 Rual J, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature, 2005, 437: 1173—1178
50 Ramani A, Bunescu R, Mooney R, et al. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. Genome Biol, 2005, 6: R40
51 Rajagopala S, Goll J, Gowda N, et al. MPI-LIT: a literature-curated dataset of microbial binary protein--protein interactions. Bioinformatics, 2008, 24: 2622
52 Jensen L, Kuhn M, Stark M, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucl Acids Res, 2009, 37: D412
53 Breitkreutz B, Stark C, Reguly T, et al. The BioGRID interaction database: 2008 update. Nucl Acids Res, 2008, 36: D637
54 Wu J, Vallenius T, Ovaska K, et al. Integrated network analysis platform for protein-protein interactions. Nat methods, 2008, 6: 75—77
55 Gandhi T, Zhong J, Mathivanan S, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet, 2006, 38: 285—293
56 Krallinger M, Leitner F, Rodriguez-Penagos C, et al. Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biol, 2008, 9: S4
57 Wang H, Huang M and Zhu X. Extract interaction detection methods from the biological literature. BMC Bioinformatics, 2009, 10: S55
58 Kim J, Ohta T, Pyysalo S, et al. Overview of BioNLP'09 shared task on event extraction. Proc of BioNLP'09, 2009: 1—9
59 Duan X, Xenarios I, Eisenberg D. Describing biological protein interactions in terms of protein states and state transitions: the LiveDIP database. Mol Cell Proteomics, 2002, 1: 104
60 Ratsch E, Schultz J, Saric J, et al. Developing a protein-interactions ontology. Comp Funct Genomics, 2003, 4: 85—89
61 Hermjakob H, Montecchi-Palazzi L, Bader G, et al. The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nat Biotechnol, 2004, 22: 177—183
62 Ashburner M, Ball C, Blake J, et al. Gene Ontology: tool for the unification of biology. Nat Genet, 2000, 25: 25—29
63 Spasic I, Ananiadou S, McNaught J, et al. Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform, 2005, 6:

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

239
64 Cohen A. Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. Proc of the BioLINK2005 Workshop, 2005, 17—24
65 Ananiadou S, Kell D, Tsujii J. Text mining and its potential applications in systems biology. Trends Biotechnol, 2006, 24: 571—579
66 Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. Nucl Acids Res, 2005, 33: W783
67 Muller H, Kenny E, Sternberg P. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol, 2004, 2: e309
68 Huang M, Zhu X, Ding S, et al. ONBIRES: Ontology-based biological relation extraction system. Proc of the Fourth Asia Pacific Bioinformatics Conference, 2006, 327—336
69 Kim J D, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. BMC Bioinformatics, 2008, 9: 10
70 Hersh W, Voorhees E. TREC genomics special issue overview. Information Retrieval, 2009, 12: 1—15
71 Kim J, Ohta T, Tsuruoka Y, et al. Introduction to the bio-entity recognition task at JNLPBA. Proc of the International Joint Workshop on Natural Language Proc in Biomedicine and its Applications (JNLPBA-04). 2004, 70—75
72 Hirschman L, Yeh A, Blaschke C, et al. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC bioinformatics, 2005, 6: S1
73 Tanabe L, Wilbur W. Tagging gene and protein names in biomedical text. Bioinformatics, 2002, 18: 1124
74 Tsuruoka Y, Tateishi Y, Kim J, et al. Developing a robust part-of-speech tagger for biomedical text. Advances in Informatics-10th Panhellenic Conference on Informatics, 2005, 382—392
75 Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets. Proc of the international joint workshop on natural language Proc. in biomedicine and its applications (NLPBA), 2004, 104—107
76 Leaman R G G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput, 2008: 652—663
77 Chen H, Sharp B. Content-rich biological network constructed by mining PubMed abstracts. BMC bioinformatics, 2004, 5: 147
78 Corney D, Buxton B, Langdon W, et al. BioRAT: extracting biological information from full-length papers. Bioinformatics, 2004, 20: 3206—3213
79 Ahmed S, Chidambaram D, Davulcu H, et al. Intex: A syntactic role driven protein-protein interaction extractor for bio-medical text. Proc of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, 2005: 54—61
80 Ohta T, Miyao Y, Ninomiya T, et al. An intelligent search engine and GUI-based efficient Medline search tool based on deep syntactic parsing. Proc of the COLING/ACL Interactive Presentation Sessions, 2006, 17—20
81 Rebholz-Schuhmann D, Jimeno A, Arregui M, et al. Assessment of modifying versus non-modifying protein interactions. The Third International Symposium on Semantic Mining in Biomedicine, 2008
82 He M, Wang Y, Li W. PPI Finder: A mining tool for human protein-protein interactions PLoS ONE, 2009, 4: e4554
83 Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal of Digital Libraries, 2000, 3: 115—130
84 Zhou X, Zhang X, Hu X. MaxMatcher: Biological concept extraction using approximate dictionary lookup. PRICAI 2006: Trends in Artificial Intelligence, 2006, 4099: 1145—1149
85 Couto FM S M, Lee V, Dimmer E, Camon E, Apweiler R, Kirsch H, Rebholz-Schuhmann D. GOAnnotator: linking protein GO annotations to evidence text. J Biomed Discov Collab, 2006, 1: 19
86 Rebholz-Schuhmann D, Kirsch H, Arregui M, et al. EBIMed——text crunching to gather facts for proteins from Medline. Bioinformatics, 2007, 23: e237
87 Kim J, Pezik P, Rebholz-Schuhmann D. MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline. Bioinformatics, 2008, 24: 1410
88 Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for Medline abstracts. Bioinformatics, 2003, 19: 1699
89 Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 2003, 13: 2498
90 Koehler J, Rawlings C, Verrier P, et al. Linking experimental results, biological networks and sequence analysis methods using Ontologies and Generalised Data Structures. In Silico Biol, 2005, 5: 33—44
91 Kohler J, Baumbach J, Taubert J, et al. Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics, 2006,

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

22: 1383
92 Rzhetsky A, Iossifov I, Koike T, et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inform, 2004, 37: 43—53
93 Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics, 2005, 21
94 Plake C, Schiemann T, Pankalla M, et al. AliBaba: PubMed as a graph. Bioinformatics, 2006, 22: 2444

Progress of Literature Mining for Protein-Protein Interaction Information

LI ManSheng[1], LIU QiJun[2], LI Dong[1], LIU PeiLei[2] & ZHU YunPing[1]

[1]State Key Laboratory of Proteomics, Beijing Proteomics Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China;
[2]College of Computer, National University of Deference Technology, Changsha 410073, China

Protein-Protein Interactions (PPIs) play an important role in most cellular functions. Analyses of PPIs contribute greatly to the understanding of biological mechanism, which has a great of academic and realistic meaning. With the development of PPI research experiments, a great amount of PPI information has been deposited in biological literatures, which presents a significant challenge to extract and reconstruct them. To address this challenge, many literature mining methods have been developed. In this review, we firstly introduce the workflow of extracting the PPI information from the biomedical literatures. And then, we present the methods and tools developed to recognize the gene/protein names and to extract their relations within the workflow, together with the extraction of PPIs' annotation information. Evaluation conferences of various bio-literature mininig tasks are introduced. Finally, we analysis the existing problems and present the prospects in this field in order to offer some references and gift for researchers engaged in PPI information extraction.

Protein-Protein Interaction, Literature Mining, Named Entity Recognition, Relation Extraction, Annotation Information Extraction

doi: 10.1360/052010-133

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net