生物信息学中的文本挖掘方法

邹 权[1]， 林 琛[1]， 刘晓燕[2]， 郭茂祖[2+]

(1. 厦门大学 信息科学与技术学院，福建 厦门 361005；
2. 哈尔滨工业大学 计算机科学与技术学院，黑龙江 哈尔滨 150001)

摘 要：从两个角度讨论应用于生物信息学中的文本挖掘方法。以搜索生物知识为目标，利用文本挖掘方法进行文献检索，进而构建相关数据库，如在 PubMed 中挖掘蛋白质相互作用和基因疾病关系等知识。总结了可以应用文本挖掘技术的生物信息学问题，如蛋白质结构与功能的分析。探讨了文本挖掘研究者可以探索的生物信息学领域，以便更多的文本挖掘研究者可以将相关成果应用于生物信息学的研究中。

关键词：生物信息学；文本挖掘；机器学习；蛋白质相互作用；文献检索

中图法分类号：TP18 **文献标识码**：A **文章编号**：1000-7024 (2011) 12-4075-04

Text mining in bioinformatics

ZOU Quan[1], LIN Chen[1], LIU Xiao-yan[2], GUO Mao-zu[2+]

(1. School of Information Science and Technology, Xiamen University, Xiamen 361005, China;
2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Text mining methods in bioinformatics are discussed from two views. First, three problems are reviewed including searching biology knowledge, retrieving the reference by text mining method and reconstructing databases. For example, protein-protein interaction and gene-disease relationship can be mined from PubMed. Then the bioinformatics applications of text mining are concluded, such as protein structure and function prediction. At last, more methods and applications are discussed for helping text mining researchers to do more contribution in bioinformatics.

Key words: bioinformatics; text mining; machine learning; protein-protein interaction; information retrieval

0 引 言

文本挖掘是用计算机算法及程序对自然语言进行理解、分析，是人工智能应用领域的重要研究方向。早在计算机出现的时候，就有了自然语言处理及文本挖掘的研究，随着机器学习、数据挖掘等算法研究的不断深入，目前文本挖掘方法在自动文摘、自动问答、Web 关系网络分析、指代消解等问题上都取得了较好的效果。

生物信息学是伴随着人类基因组计划的进行和完成而兴起的一门交叉学科，是利用计算机、统计等信息学方法预测和解决遗传相关的生命学科问题，其中最主要的是数据的存储、检索和分析。美国国立生物技术信息中心 (national center for biotechnology information，NCBI) 专门为存储相关生物数据建立了多个数据库，其中包括存储 DNA、蛋白质的序列数据库 (如 dbEST、dbSNP)，有存储疾病相关数据的 OMIM 数据库，也有存储基因芯片数据的 GEO 数据库，还有存储生物、医学相关文献的 PubMed 数据库。

在规模日益增大的数据库中检索用户偏好的知识需要用到文本挖掘的方法，因此有研究者试图用计算机相关的算法和程序在 PubMed 中检索自己感兴趣的论文，如寻找不同蛋白质相互作用关系。随着对遗传密码的破解，研究者逐渐的发现生物序列，特别是蛋白质序列，与人类的语言在构成本质上有着很强的相似性，因此除了直接将文本挖掘应用于生物信息学研究进行文献检索外，越来越多的研究者开始把蛋白质序列当作特殊的“文本”，利用已有的文本挖掘方法对其进行分析，从而对蛋白质的结构和功能进行预测。本文从以上两个方面出发，总结生物信息学研究中用到的文本挖掘方法，目的在于让生物信息学研究者了解文本挖掘，也希望更多的文本挖掘研究者能够将好的方法应用于生物信息学研究中。

收稿日期：2011-01-18；**修订日期**：2011-03-18。

基金项目：国家自然科学基金项目 (61001013、60932008、61001143)。

作者简介：邹权 (1982－)，男，黑龙江佳木斯人，博士，讲师，研究方向为生物信息学；林琛 (1982－)，女，福建厦门人，博士，讲师，CCF 会员，研究方向为数据挖掘与 Web 社会网络；刘晓燕 (1963－)，女，黑龙江哈尔滨人，博士，副教授，研究方向为生物信息学；+通讯作者：郭茂祖 (1966－)，男，山东夏津人，博士，教授，研究方向为计算生物学、机器学习与新型计算模型。E-mail：maozuguo@hit.edu.cn

1 生物信息学中的文献挖掘

文本挖掘技术的发展,对生物文献检索,尤其是生物信息数据库的构建起了很大的作用。自然语言领域的著名国际会议 ACL(annual meeting of the association for computational linguistics)2005 与生物信息学领域的著名国际会议 ISMB(annual international conference on intelligent systems for molecular biology) 2005 曾就生物文献检索问题展开专门的 Workshop,讨论生物信息学中相关问题的文献发掘问题。其中,蛋白质相互作用预测与基因功能与疾病的关系预测是最主要的两个应用主题。

1.1 蛋白质相互作用预测

预测蛋白质之间的相互作用网络是生物信息学和系统生物学中的重要研究课题。以往的研究是人工从文献中记录蛋白质的相互作用关系,但随着生物文献指数级别的增长速度,需要有程序自动地从 PubMed 的摘要中识别蛋白质对间的相互作用。然而蛋白质的命名没有统一的规则,有很多蛋白、基因的名称是在一起混用的。因此,从文献摘要中识别蛋白的名称,进一步识别存在相互作用关系是利用文本挖掘寻找蛋白相互作用的关键问题。

最初的研究是通过统计和计数的方法进行预测,人工构建出蛋白质名称的词典,然后找出出现两次以上词典中的元素且相距不远的摘要,进而判定相关的蛋白质存在相互作用。也有研究者利用动态规划的方法对蛋白质相互作用的语句进行比对,从而预测是否存在相互作用。

蛋白质相互作用预测在很长的一段时间都是生物信息学研究的热点问题,这也使得越来越多的文本挖掘与自然语言处理的研究者投身于该问题的研究。首先是对文献摘要进行更细致的文法分析,而不是简单的字典词数统计。Kim 等利用构造核的方法,把复杂的语义结构分析转化为求图中的最短路径[1]。类似的还有利用语法分析[2]、上下文无关文法分析[3]、本体分析[4]等信息检索手段分析文献摘要,从而挖掘蛋白质的相互作用。另外,集成学习[5]、贝叶斯网络[6]等机器学习方法也被应用与名称识别与相互作用识别的预测。

1.2 基因功能与疾病预测

蛋白质相互作用预测是在文本中寻找两个蛋白质且判断其存在相互作用。基因功能与疾病预测与其类似,即在文献中寻找基因的名称,同时寻找疾病的名称,进而判别该基因是否与该疾病存在关联。

这种识别大多分为 3 个步骤:首先是通过与词典的比对搜索相关论文摘要;然后根据词典中词语定位相关子句,有时向前后扩展,以保证准确率;最后利用文法分析方法或机器学习方法对事实进行判定。这种方法往往针对特殊的基因及疾病时,会取得较好的效果。Bui 等在 PubMed 中挖掘药物与 HIV 病毒变异之间的联系。Jiang 等利用 microRNA 命名的规则性,收集了近 3000 条不同 microRNA 与不同疾病之间的关系[7]。Cheng 等针对人类疾病、变异与药物作用效果之间的联系开发了文本挖掘系统[8]。Ivan 等则锁定的目标是人类与小鼠的小脑畸形研究[9]。Lars 等详细总结了相关的文献数据库、文献挖掘软件及功能[10]。

1.3 文献检索

生命科学文献浩如烟海,寻找相互作用的蛋白质对以及挖掘基因与疾病之间的关系只是两个主要的应用例子。还有很多的生命科学问题以及生物信息学问题需要利用文本挖掘方法,在 PubMed 等文献库中寻找答案。

挖掘生物文献解决相关问题时一般都要处理两个主要的问题,即实体名称识别(name entity recognition)和关系抽取(relation extraction)。所用的方法主要有:基于语言分析的方法、基于字典的方法、机器学习方法、统计方法。

另外,将 PubMed 数据库转化成 XML 关系数据库,利用短词匹配对论文、作者名称进行模糊搜索也是最近研究的热点问题[11]。

2 文本挖掘方法的应用

DNA 和蛋白质序列是有意义的遗传语言,被认为是生命的天书[12]。因此,有越来越多的自然语言处理与文本挖掘的计算方法被应用于生物信息学问题的研究,比如对蛋白质的频谱分析就源于自然语言处理中的词频统计。

2.1 蛋白质结构标定

蛋白质的结构决定功能,欲研究蛋白质的功能,首先要分析蛋白质的结构。蛋白质的结构分析主要是对给定的蛋白序列,标定出哪一段区域属于α螺旋、β片层,哪一段区域属于无序区(protein disordered region)。其中对α螺旋、β片层的预测即蛋白质二级结构预测。

如果把一段蛋白序列视为自然语言,则对区域的类型分析则类似于自然语言处理中的语法标定。首先,基于规则和统计相结合的方法被应用于蛋白质二级结构预测[13]。但随着统计的方法达到预测瓶颈后,有研究者陆续提出利用机器学习的预测方法,其中有基于人工神经网络[14]、基于支持向量机[15]以及基于最大熵的方法[16]。

类似的还有蛋白质无序区的预测。蛋白质无序区指在蛋白质空间结构中不具有稳定的或特定的三维结构的区域。包括人工神经网络[17]、支持向量机[18]、条件随机域[19]、随机森林等文本挖掘与机器学习方法都被应用于对其进行预测。其中,目前较常用的服务器地址如表 1 所示。

2.2 蛋白质功能预测

蛋白质的功能预测是生物信息学最基础的研究课题之一,其中包括蛋白质相互作用与相互作用位点预测、蛋白质亚细胞定位、跨膜蛋白的预测与分类、蛋白质功能分类、多功能酶的识别等。

蛋白质的序列最容易获得,它像自然语言一样具有许多复杂的规则,但不同的是人们很难总结和理解蛋白质序列的规则。因此,需要借助计算语言学以及机器学习的方法对由氨基酸序列所表达的"蛋白质语言"进行分析和预测,从而理解蛋白质序列所具有的功能。

蛋白质相互作用预测是蛋白质功能研究的最基础课题之一,许多研究者致力于根据两条蛋白序列预测其是否存在相互作用。目前已有很多机器学习方法应用于其中,包括支持向量机[20]、核方法[21]、决策树[22]、随机森林[23]、贝叶斯网络[24]、AR 模型[25],包括本体标注、样本加权[26]等文本处理方法也被应用

表 1 常用的蛋白质结构功能预测网址

问题	名称	网址	参考文献
蛋白质无序区预测	IUPred	http://iupred.enzim.hu/	文献[30]
	DISpro	http://scratch.proteomics.ics.uci.edu/	文献[31]
	Spritz	http://distill.ucd.ie/spritz/	文献[32]
	RONN	http://www.strubi.ox.ac.uk/RONN/	文献[33]
蛋白质相互作用位点预测	Cons-PPIS	http://pipe.scs.fsu.edu/ppisp.html	文献[34]
	SPPIDER	http://sppider.cchmc.org	文献[35]
	PP-Pred	http://bioinformatics.leeds.ac.uk/ppi-pred	文献[36]
	PINUP	http://sparks.informatics.iupui.edu/PINUP	文献[37]
蛋白质相互作用预测	PIE	http://pie.snu.ac.kr/	文献[38]
	Pred-PPI	http://cic.scu.edu.cn/bioinformatics/predict_ppi/default.html	文献[39]
	InterPreTS	http://www.russell.embl.de/cgi-bin/tools/interprets.pl	文献[40]

于特征提取与训练数据处理之中。在预测相互作用的同时，研究者也想分析出蛋白序列中发生相互作用的区域，即蛋白质相互作用位点的预测。条件随机域[27]、隐马尔可夫模型[28]等语法分析中常用到的信息方法也被用于分析相互作用位点，并取得了较好的效果。除此之外，随机森林、支持向量机、人工神经网络、贝叶斯网路、线性回归等机器学习方法在蛋白质相互作用位点预测中均有应用。然而，也有研究者提出怀疑：单纯依靠蛋白序列无法为预测相互作用提供足够信息[29]。因此，需要更多的文本挖掘和机器学习的研究者开发新的特征与分类方法来解决这一问题。目前常用的蛋白质相互作用与相互作用位点预测软件网址如表 1 所示。

3 结束语

随着自然语言与文本挖掘方法研究的日趋成熟，寻找应用领域将是未来研究的中心。以生物信息学为代表的交叉学科正在吸引着越来越多的信息科学研究人员，将文本挖掘的技术和方法应用于生物信息学研究会引起越来越多文本挖掘研究人员的注意。同时，生物信息学研究者也需要逐渐深入学习文本挖掘方法，以解决具体的生物信息学问题。

在生物文献检索领域中，除了文中提到的蛋白质相互作用预测、基因疾病关系预测以外，还有许多问题需要利用文本挖掘的方法在文献数据库中搜索相关知识，尤其是需要随时更新文献检索结果的问题，如药物不良反应与分子成分的关系、单核苷酸多态性(single nucleotide polymorphisms，SNP)位点与疾病、药物不良反应的关系等。

在生物信息学领域中，与蛋白质组学相关的、需要根据氨基酸序列进行预测的结构和功能相关研究，几乎都可以利用文本挖掘技术进行处理。词频统计、条件随机域、隐马尔科夫模型、上下文无关文法等成熟的文本挖掘技术已经成功的应用于二级结构预测、不规则区域预测、相互作用及相互作用位点预测之中，但文本挖掘的最新研究成果还有待研究者应用于蛋白质、DNA 语言之中。蛋白质的三四级结构、远距离家族同源检测、无序区检测与相互作用网络构建、药物靶点预测等问题尚没有有效的计算处理方法，还需要更多的信息科学研究人员提供更为有效的算法。除此之外，半监督学习、主动学习等新的机器学习与文本挖掘方法被提出，也必将在生物文献检索与生物信息学中得到应用。目前，基于反馈的推荐系统将成为生物文献检索的新热点问题。

生物信息学的发展为信息科学，尤其是文本挖掘研究者提供了广阔的应用空间，同时也需要文本挖掘算法研究人员针对生物数据的特点，提出更有效的智能算法。本文总结了生物信息学中应用的文本挖掘方法及问题的同时，也给出了相关的较为成功预测软件网址，可供新接触生物信息学的文本挖掘研究人员测试和对比。作者期望更多的文本挖掘研究人员能够将自己的方法应用在生物信息学领域，进而推动生物信息学乃至分子遗传学研究的发展。

参考文献:

[1] Seonho Kim, Juntae Yoon, Yang Jihoon. Kernel approaches for genic interaction extraction[J].Bioinformatics,2008,24(1):118-126.

[2] Katrin Fundel,Robert Kuffner,Ralf Zimmer.RelEx-relation extraction using dependency parse trees[J].Bioinformatics,2007,23(3):365-371.

[3] Joshua M Temkin,Mark R Gilder.Extraction of protein interaction information from unstructured text using a context-free grammar[J].Bioinformatics,2003,19(16):2046-2053.

[4] Andre Skusa,Alexander Ruegg,Jacob Kahler.Extraction of biological interaction networks from scientific literature [J]. Briefings in Bioinformatics,2005,6(3):263-276.

[5] Rainer Malik, Lude Franke, Arno Siebes. Combination of text-mining algorithms increases the performance[J].Bioinformatics, 2006,22(17):2151-2157.

[6] Rajesh Chowdhary,Zhang Jinfeng,Liu Jun S.Bayesian inference of protein-protein interactions from biological literature[J].Bioinformatics,2009,25(12):1536-1542.

[7] Jiang Qinghua,Wang Yadong,Hao Yangyang,et al.Mir2disease:a manually curated database for MicroRNA deregulation in human disease[J].Nucleic Acids Research,2009,37(s1):D98-D104.

[8] Cheng Dean,Craig Knox,Nelson Young,et al.PolySearch: a web-based text mining system extracting relationships between human diseases,genes,mutations,drugs and metabolites[J].Nucleic Acids Research,2008,36(s2):W399-W405.

[9] Ivan Iossifov, Raul Rodriguez-Esteban, Ilya Mayzus, et al. Looking at cerebellar malformations through text-mined Interactomes of mice and humans [J]. PLoS Computational Biology, 2009,5(11):e1000559.

[10] Lars Juhl Jensen,Jasmin Saric,Peer Bork.Literature mining for the biologist:from information retrieval to biological discovery [J].Nature Genetics,2006,7(2):119-129.

[11] Wang Jiannan,Inci Cetindil,Ji Shengyue,et al.Interactive and fuzzy search:a dynamic way to explore MEDLINE[J].Bioinformatics,2010,26(18):2313-2320.

[12] Kay L.A book of life? How the genome became an information system and DNA a language[J].Perspect Biol Med,1998,41(4): 504-528.

[13] Gamier J,Osguthorpe D J,Robson B.Analysis of the accuracy and implications of simple method for predicting the secondary structure of globular proteins[J].Journal of Molecular Biology, 1978,120(1):97-120.

[14] Rost B,Sander C.Prediction of protein secondary structure at better than 70%accuracy[J].Journal of Molecular Biology,1993,232 (2):584-599.

[15] Hua S J,Sun Z R.A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach [J]. Journal of Molecular Biology, 2001,308 (2):397-407.

[16] Liu Y,Carbonell J,Klein-Seetharaman J,et al.Comparison of probabilistic combination methods for protein secondary structure prediction[J].Bioinformatics,2004,20(17):3099-3107.

[17] Romero P,Obradovic Z,Li X,et al.Sequence complexity of disordered protein[J].Proteins,2001,42(1):38-48.

[18] Hirose S,Shimizu K,Kanai S,et al.Poodle-L:a two-level SVM prediction system for reliably predicting long disordered regions [J].Bioinfromatics,2007,23(16):2046-2053.

[19] Wang L, Sauer U H. Ond-Crf: predicting order and disorder in proteins using conditional random fields [J]. Bioinformatics, 2008,24(11):1401-1402.

[20] Bock J R, Gough D A. Predicting protein-protein interactions from primary structure[J].Bioinformatics,2001,17(5):455-460.

[21] Ben-Hur A,Noble W S.Kernel methods for predicting protein-protein interactions[J].Bioinformatics,2005,21(s1):i38-i46.

[22] Darnell S J,Page D,Mitchell J C.An automated decision-tree approach to predicting protein interaction hot spots [J]. Proteins: Structure,Function,and Bioinformatics,2007,68(4):813-823.

[23] Chen X W,Liu M.Prediction of protein-protein interactions using random decision forest framework [J]. Bioinformatics, 2005,21 (24):4394-4400.

[24] Jansen R, Yu H, Greenbaum D, et al. A bayesian networks approach for predicting protein-protein interactions from genomic data[J].Science,2003,302(5644):449-453.

[25] Gomez S M,Noble W S,Rzhetsky A.Learning to predict protein-protein interactions from protein sequences [J]. Bioinformatics, 2003,19(15):1875-1881.

[26] Li Minghui, Wang Xiaolong, Lin Lei, et al. Effect of example weights on prediction of protein-protein interactions[J].Computational Biology and Chemistry,2006,30(5):386-392.

[27] Li Minghui,Lin Lei,Wang Xiaolong,et al.Protein-protein interaction site prediction based on conditional random fields[J].Bioinformatics,2007,23(5):597-604.

[28] Friedrich T,Pils B,Dandekar T,et al.Modelling interaction sites in protein Domains with interaction profile hidden Markov models [J].Bioinformatics,2006,22(23):2851-2857.

[29] Yu Jiantao,Guo Maozu,Chris Needham,et al.Westhead.Simple sequence-based kernels do not predict protein-protein interactions[J].Bioinformatics,2010,26(20):2610-2614.

[30] Dosztanyi Z,Csizmok V,Tompa P,et al.The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins [J]. J Mol Biol,2005,347(4):827-839.

[31] Cheng J,Sweredoski M J,Baldi P.Accurate prediction of protein disordered regions by mining protein structure data[J].Data Mining and Knowledge Discovery,2005,11(3):213-222.

[32] Vullo A,Bortolami O,Pollastri G,et al.Spritz:a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines[J].Nucleic Acids Res,2006,34(s2):W164-W168.

[33] Yang Z R,Thomson R,McNeil P,et al.Ronn:the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins [J]. Bioinfromatics, 2005,21 (16): 3369-3376.

[34] Chen H,Zhou H X.Prediction of interface residues in protein-protein complexes by a consensus neural network method:test against NMR data[J].Proteins,2005,61(1):21-35.

[35] Porollo A,Meller J.Prediction-based fingerprints of protein-protein interactions[J].Proteins,2007,66(3):630-645.

[36] Bradford J R,Needham C J,Bulpitt A J,et al.Insights into protein-protein interfaces using a Bayesian network prediction method [J].J Mol Biol,2006,362(2):365-386.

[37] Liang S,Zhang C,Liu S,et al.Protein binding site prediction using an empirical scoring function [J]. Nucleic Acids Res, 2006,34 (13):3698-3707.

[38] Kim S,Shin S-Y,Lee I-H,et al.PIE:an online prediction system for protein-protein interactions from text[J].Nucleic Acids Research,2008,36(s2):W411-W415.

[39] Guo Yanzhi,Yu Lezheng,Wen Zhining,et al.Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences[J].Nucleic Acids Research,2008,36(9):3025-3030.

[40] Aloy P, Russell RB. InterPreTS: protein interaction prediction through tertiary structure[J].Bioinformatics,2003,19(1):161-162.