DOI:10.12113/202003003

文本分析技术在蛋白质生物信息学中应用的案例综述

苏绍玉[1],徐 婧[2],鄢仁祥[2*]
(1.福建省科学技术信息研究所, 福州 350003;2.福州大学 生物科学与工程学院,福州 350100)

摘 要:海量数据时代考察文本分析技术在生物信息学领域的应用具有重要的理论和现实价值。本文讨论了文本分析在蛋白质计算分析中的几个应用实例以及核心技术内容。文本分析技术应用于生物信息学领域可发挥引领和导向作用,在生物信息学中的应用又进一步促进了文本分析技术的发展。文本分析技术虽然广泛在生物信息学中应用,但是其发展仍然有需要尚待解决的几个问题,本文对此也进行了讨论。

关键词:文本分析;科技情报;生物信息学;蛋白质计算;大数据;人工智能
中图分类号:Q816;G202 **文献标志码**:A **文章编号**:1672-5565(2020)04-215-08

Systematic investigation of application of text analysis to protein bioinformatics

SU Shaoyu[1], XU Jing[2], YAN Renxiang[2*]
(1. *Fujian Institute of Scientific and Technological Information*, *Fuzhou* 350003, *China*;
2. *College of Biological Science and Engineering*, *Fuzhou University*, *Fuzhou* 350100, *China*)

Abstract: Investigating the application of text analysis technology to bioinformatics has important theoretical and practical value in the era of big data. Several application examples and core technical content of text analysis in protein computational analysis are discussed in this paper. Text analysis technology can play a leading and guiding role in the field of bioinformatics, and its application in bioinformatics further promotes the development of text analysis technology. Although text analysis technology has been widely used in bioinformatics, there are still some problems to be solved, which are also discussed in this paper.

Keywords: Text analysis; Sci-tech information; Bioinformatics; Protein computing; Big data; Artificial intelligence

文本一般是指由一定符号或者文字组成的信息。大部分的文本信息是可以直接人工读取其内容的。文本分析主要指对特定的文本数据的含义进行表示、提取其中特征、并对相应信息知识进行表示和分析的过程。文本分析是文本表示、信息挖掘、信息检索和深入探索的一个基本问题。通常来讲,文本分析是文本数据挖掘、信息检索和大数据分析中的核心技术之一。近些年来,随着大数据、人工智能和机器学习等一系列信息技术的发展,文本分析技术在众多领域得到了越来越多的应用。现今,各行各业包括科研和商业机构都处于信息爆炸产生的海量数据时代,在这种背景之下文本分析技术愈显得重要和亟待发展。

资料或者讯息主要分为视频、音频和文本资料三种类型。其中,文本是多数传统知识的载体形式。近年来随着机器学习中的深度学习(Deep learning)[1]算法的进展,音频和视频的识别率显著提高并得到越来越多的应用。这可以在最近几次的图像识别竞赛结果以及现实生活中各种视音频识别应用中得以体现。近年来在对资料进行分析中的另外一个进展就是文本分析技术。各种机器学习算法广泛地应用于文本的解析之中[2-5],特别是人工神经元网络[6]、随机森林[7]以及支持向量机[8]等主流方法。尤其明显的是在近几年,深度学习算法结合文本分析在不少领域中都有着重要的应用[9]。

文本的分析过程可以简单地分为设计规划、文

收稿日期:2020-03-13;**修回日期**:2020-04-11.
基金项目:国家自然科学基金项目(No.31500673);福建省属公益类科研院所基本科研专项基金(No. 2017R1008-10).
作者简介:苏绍玉,助理研究员,硕士,研究方向:蛋白质组学和生物信息学. E-mail:26639388@ qq.com.
* **通信作者**:鄢仁祥,副研究员,硕士生导师,研究方向:蛋白质组学和生物信息学. E-mail:yanrenxiang@ fzu.edu.cn.

本信息采集、文本信息解析和文本内涵的应用等环节。图 1 列出了文本分析的基本通用过程。这个基本过程主要就是文本数据的收集、统计分析建模、挖掘其中潜在知识、以及应用挖掘到的知识等过程。文本数据的收集一般可以从互联网(例如,使用爬虫程序或者直接下载)、各种数据库、现存的文件或者实际正在开展的调研项目中获取;之后采用统计模型或者机器学习算法等对数据进行深入分析以及建模;挖掘出来的知识一般以图表或者一些特定的模型形式表示;最终,经验证的可靠知识和规律则可以直接应用于实际之中。

1 文本分析技术在生物信息学中的应用

文本分析的方法已经广泛应用于蛋白质的研究之中。蛋白质相关的计算研究是生物信息学中的一个重要方向。从结构层次的角度来看,蛋白质的一级序列、二级结构、三级结构以及功能等信息都可以用文本进行表示(本文将会对蛋白质的四个结构层次进行详细论述见图 1)。在生物学中,蛋白质分子广泛地参与细胞中的各项生理和生活过程,包括生长、发育、生殖、代谢和免疫调节等各个环节,并与各种重要疾病的发生、发展和治疗密切相关。同时,蛋白质是一种重要的药物靶标,其相应机理一直受到科学界的极大关注。蛋白质的结构,特别是三维空间构象,对其生物学功能有重要作用。然而,与蛋白质在生物学中的重要性形成强烈反差的是科学界对于其结构、功能以及相应机理的了解非常贫乏。主要原因是通过实验手段来获得相应信息的过程极其困难,同时需要消耗一定的科研经费。在学术界中,各个专业的科学家通常主要从计算模拟和实验验证的两个角度同时进行来系统性地研究蛋白的作用机制。蛋白质结构与功能的研究通常被认为是学术价值和应用价值兼备的研究课题之一,其研究不仅可增强学术界对蛋白质折叠规律的了解,而且开发出的预测方法还可直接用于指导蛋白质工程、功能基因组学和新药研发等实验研究过程。因此,近年来生物信息学的研究中越来越偏重于往蛋白质相关的方向发展。与此同时,在“互联网+”和大数据快速发展的背景下,文本分析技术越来越多的应用于蛋白质分析的研究中,这对生物学领域以及情报分析行业都有重要意义。文本分析研究中有效融合了数据分析和信息处理以挖掘蛋白质相关的有价值的信息,可在生物学研究中发挥引领和导向作用。

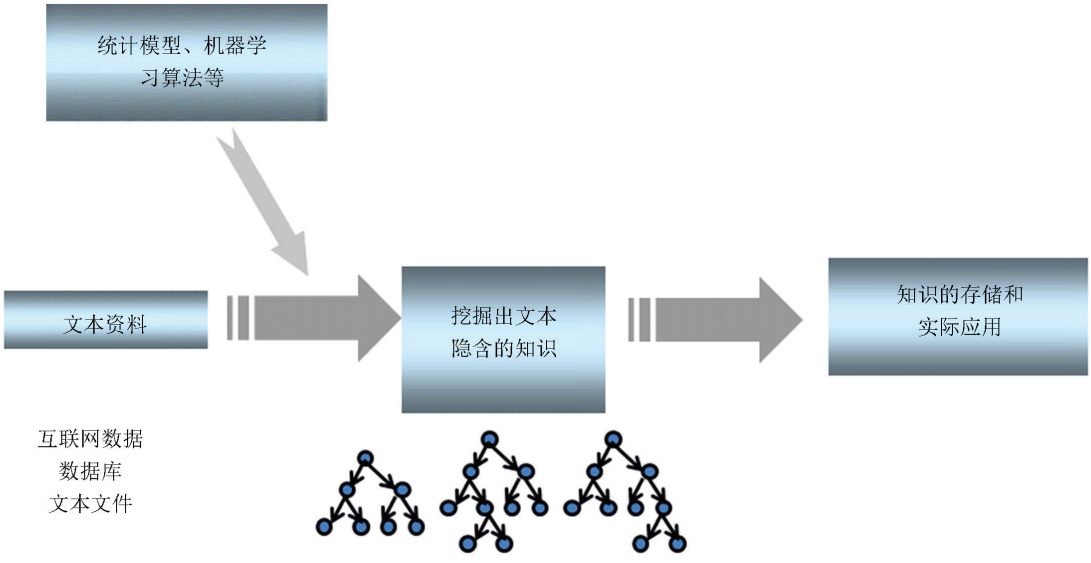

图 1 文本分析的基本过程

Fig.1 Basic procedure for text analysis

传统的信息情报[10]研究工作一般按照任务和项目进行相应的研究。信息情报技术目前已经介入包括经济、社会、科技预测研究和重大智库决策支持研究之中。同时,也包括生物科学的研究,信息情报技术特别是文本分析技术目前也已经得到了较大的应用。情报分析技术的进展需要全面客观理性发现、分析和研究特定领域的需求。科技情报机构一般不仅仅为政府部门提供科技情报服务,也应为科学技术的进步,例如生物学的方法,提供一定的技术储备和应用理论基础。通常来讲,情报信息技术的

发展一般涵盖不同时空领域的政治、经济、社会、生物学、人文文化等各方面的信息。特别是近些年来,生物学和信息技术结合得越来越紧密,情报分析中的文本分析技术在这个特定领域中也得到了越来越广泛的应用。我们将在下面段落中对文本分析在生物信息学中蛋白质的计算分析具体应用进行举例。

2　文本分析技术应用于蛋白质折叠类型的识别

蛋白质折叠识别,即从蛋白质序列识别出其三维结构类型,是生物信息学中的一个重要研究方向。蛋白质一级序列结构中一般使用 20 个字母表示 20 种标准氨基酸,相应的序列信息以文本文件的形式进行储存。目前学术界可以获得的最大的蛋白质序列数据库为美国国立生物技术信息中心(National Center for Biotechnology Information)[11]的 NR(Non-Redundant Protein Sequence Database)数据库。NR 数据库不同于一般数据库(例如 MySQL 和微软 SQL server 数据库),NR 数据库就是采用最简单的文本文件的格式(也称为 fasta 格式)来储存蛋白质序列。所以不少文本分析的算法可以直接应用于蛋白质的序列分析。

蛋白质折叠识别涉及蛋白质一级序列到三维空间结构类型(即蛋白质三级结构)的映射(见图 2),主要使用方法为穿线法(Threading)或者也称为逆折叠法(Inverse folding)[12],意为把目标序列穿过蛋白质的三维空间结构以判断序列与结构之间的匹配程度。匹配程度越高,则相应的打分分数也越高。折叠识别方法的理论基础是蛋白质在序列不太相似的情况下,蛋白质的空间结构仍然可能相似,即蛋白质之间存在弱同源性。这是因为蛋白质在千百万年甚至更长时间的进化历程中,为了保持功能的相对稳定结构也会保持相对稳定,但是序列由于各种突变因素的存在而不断变化。这个特性可以总结为蛋白质的结构比序列保守。所以自然界中存在序列很不相似但是结构相似的蛋白质结构。文本分析方法在蛋白质折叠识别中也有重要应用。其中代表性的方法之一为 Cheng 和 Baldi 提出的一种基于文本分析和机器学习的方法。在这种方法中,作者主要采用文本分析中信息检索(Information retrieval)[13]的方法。给定一个查询蛋白质的一级结构序列,折叠的目标识别是将所有可能的模板按照它们序列与结构相关性排序,这个过程类似于谷歌和其他搜索引擎对与用户查询关联的网页进行排名。在这样的分析算法中,越相似的蛋白质理论上也就越容易排名比较靠前。在这个案例中,文本分析技术是一种总体的策略,即采用类似于搜索引擎算法的策略把蛋白质的折叠类型给检索出来。在具体的参数构建中(即机器学习中输入的特征向量),这个研究具体采用可以表征蛋白质序列和结构相似性的各种指标。

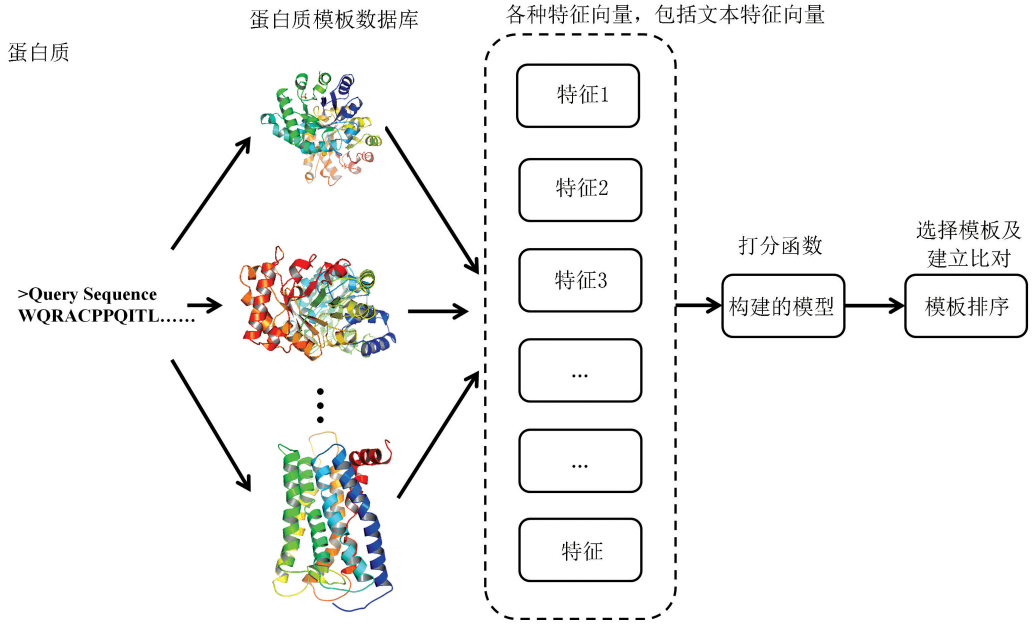

图 2　蛋白质折叠识别的基本流程

Fig.2　Basic flow for protein fold recognition

文本与蛋白质的三维空间有着紧密关系。虽然蛋白质的结构可以用三维图形的形式显示,但是存储蛋白质结构的文件却是文本文件。这个文件一般是以 PDB 的格式进行存储的。目前 PDB 数据库(http://www.rcsb.org/)中存储的生物大分子结构数据量有 16 万多个,其绝大多数是蛋白质的结构文件。因为蛋白质的结构文件采用文本形式存储,所以文本分析技术也广泛应用于蛋白质结构分析中,包括信息的提取、分析以及结构优化等。

3 蛋白质二级结构的预测

在生物学中,蛋白质的二级结构除了用图形结构表示外,同时也可以采用本文的形式进行储存。在蛋白质二级结构中,学术界一般用字母 H 代表螺旋、E 代表折叠以及 C 代表无规则卷曲。蛋白质二级结构预测的经典算法是 Psipred[14]。Psipred 算法由 Jones 于 20 世纪 90 年代开发,现今仍然有广泛的应用。序列一般通过 NCBI PSI-BLAST[15]程序搜索 NCBI 的 NR 数据库生成序列谱(Profile)信息,序列谱一般可以反应 20 种氨基酸出现的频率。表 1 是使用 PSI-BLAST 程序对蛋白质进行迭代搜索得到的 PSSM 矩阵,即蛋白质每个位置 20 种氨基酸出现的频率。研究人员可以进步一地使用人工神经元网络(Neural network)算法来对序列谱信息进行建模,并准确预测出相应蛋白质每个氨基酸的二级结构。有时也会测试主流的深度学习算法,例如 Tensorflow[16]等在相应特征相应上的预测性能。在实际研究中,科研人员一般构建一系列有效的特征向量,用于准确地预测未知蛋白的二级结构。这样面对一个新的蛋白质,其二级结构信息将准确地展示在科研人员面前,这为进一步的科研提供了非常有用的拓扑结构信息。目前主流的蛋白质二级结构测预测精度一般可以达到 80%以上。图 3 列出了常见的二级结构程序所使用的算法框架。

表 1 4ay7[17]蛋白质前 9 个氨基酸的 PSSM 矩阵

Table 1 PSSM matrix of the top 9 amino acids of protein 4ay7[17]

AA	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
E	−5	−3	−4	2	−7	4	6	−6	−4	−7	−7	4	−6	−7	−5	−4	−5	−7	−6	−6
F	−9	−8	−11	−12	−10	−9	−11	−11	−7	−4	−1	−8	12	−4	−11	−10	−8	−5	−3	−8
T	−6	−7	4	−1	−8	−6	−6	−6	−5	−3	−5	−4	−5	−9	−5	−1	7	−10	−9	−8
L	−1	−1	−3	−5	−7	−3	−3	−1	2	−6	−2	3	3	−3	4	4	−3	−8	−2	−5
K	−4	5	4	−5	−9	−2	−3	−6	−3	−2	−2	5	1	−5	−5	−4	−2	−3	−4	−5
T	−5	−4	−2	4	−10	2	7	−9	−5	−9	−9	−2	−6	−10	−7	−4	−3	−10	−7	−8
R	−6	8	−1	−7	−5	−4	−5	−8	−8	−4	1	−3	−8	−9	−6	−7	−1	2	−7	−4
L	−3	−7	−6	−8	−4	−9	−8	−7	−9	3	3	−6	3	5	−9	−6	−4	−3	−1	4
L	−2	0	−3	−4	−5	−1	−1	−6	−3	2	4	1	2	0	−8	−5	−3	−4	−1	0

根据形态以及功能特点,生物体内的蛋白可以分成纤维状蛋白(Fibrous protein)、球状蛋白(Globular protein)和膜蛋白(Membrane protein)三大类。有一些特定的蛋白质,例如膜蛋白,其二级结构预测准确率相对较低。这可能是因为膜蛋白处在生物膜之中,膜蛋白的跨膜区与非跨膜区的区别以及功能作用差异是比较大的。这使得膜蛋白有着与球蛋白不同的生物化学特性。准确地获得膜蛋白跨膜区与非跨膜区的信息将有时对判断膜蛋白的生物学功能起到关键的决定性作用[18]。因此,开发针对膜蛋白的二级结构预测算法也是未来的一个重要方向。

4 基于互联网数据库的蛋白质序列功能信息解析

从新测序的蛋白质组和基因组序列中准确识别出目标蛋白是现代组数据功能注释的重要步骤之一。自从各种基因组和蛋白质组测序计划成功开展起,各种测序技术得到极大的发展。因此,大量的物种随后相继被测序。现在几乎每天都会产生大量的基因和蛋白质序列[19]。而新测序的基因和蛋白质序列中一部分数据一般是没有注释功能的。如何从一个新测序物种筛选出潜在的目标基因和蛋白质是科学家们亟待面对的重要的工作之一,这同时也是研究人员判断一个后续研究项目的重要性和是否开展一个研究项目的关键科学依据。现代蛋白质组数据的特点是序列多样性复杂、以及数据量大等特点。首先,科研人员通常研发一系列创新性和使用的算法[20]。新算法可以从全基因组规模的数据中识别出蛋白质相互作用网络,以及预测蛋白质三维结构和药物配体结合模式。同时,一系列创新性的编码及模型算法方式将用于构建针对特定目标蛋白的生物信息学模型[21]。得到计算模型后,研究人员通常将进行相应的蛋白质组学实验对模型进行验证和进一步研究。

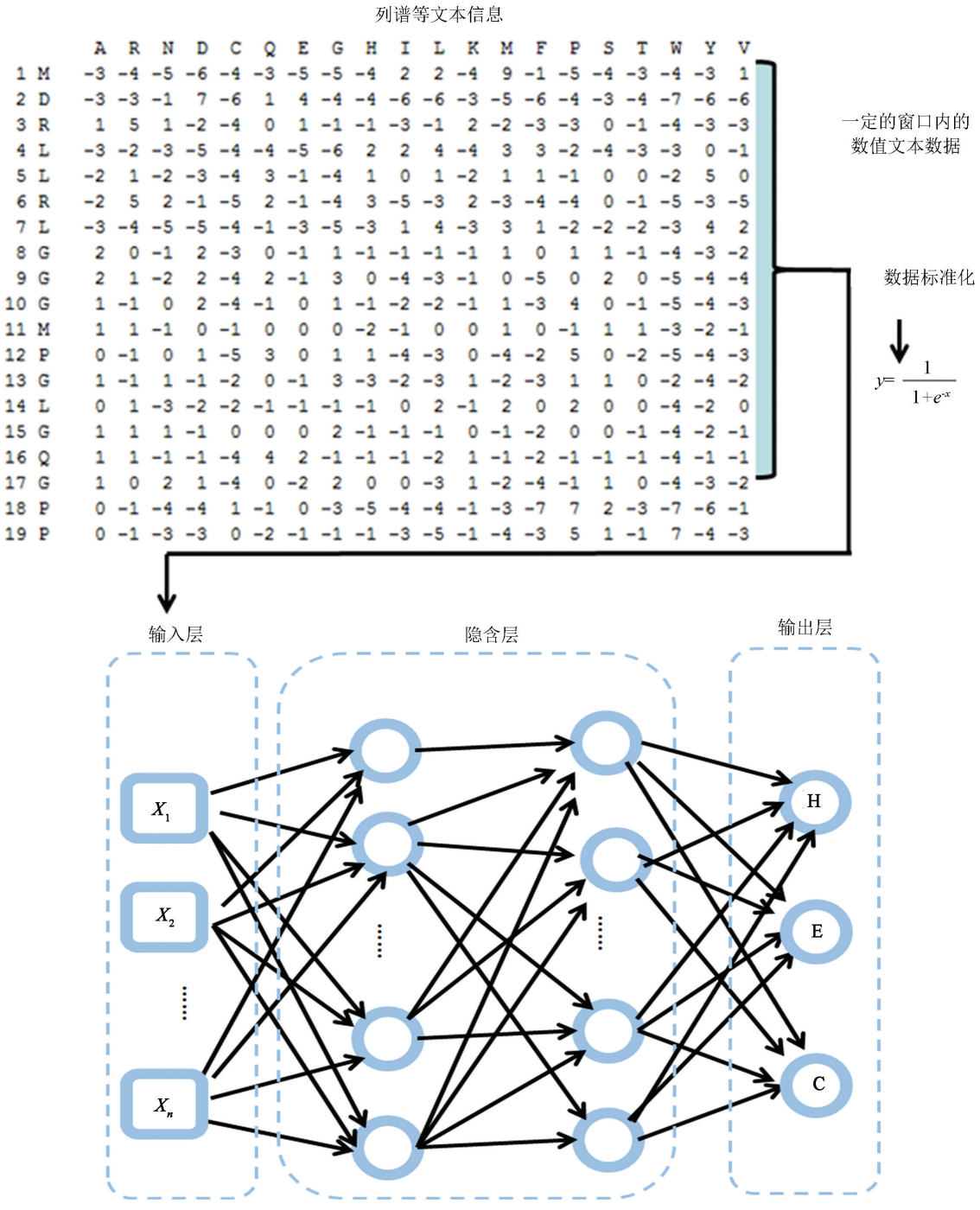

图 3　基于文本分析和机器学习的蛋白质二级结构预测

Fig.3　Prediction of protein secondary structure using text analysis and machine learning

图 4 列出了常见以蛋白质序列为起始的功能注释流程。在蛋白质功能研究中,科研人员通常开发能适应整个蛋白质组规模的目标蛋白识别的算法。将在保证预测准确定性的前提下,优化程序让程序的运行速度更快。相应算法可以直接应用于蛋白质组注释,以及筛选出重要的目标蛋白,这为基因组科学家们相应的后续研究提供了强有力的计算支持。在这个过程中,各种算法包括文本分析方法得到广泛的应用。

5　文本分析技术在蛋白质生物信息学中的其他应用

在以上提到的几个例子之外,文本分析技术在蛋白质生物信息学的其他方面也有着重要应用。Ross 等人提出,在传统构建进化树的方法上,也可以结合文本分析和同源性方法进行构建[22]。Ross 等人提出的方法扩充了文本的表示以及应用领域,

其方法也有助于生物科研界对文本数据和生物进化之间的理解。论文文献本身就是一种文本形式。Hassani 等人使用文本分析技术,系统性搜索现有医学文献数据。基于此种方法,Hassani 等人据创建与植物胁迫反应有关的拟南芥蛋白质的知识库。在构建的知识库的基础上,提出多种评分指标来识别关键的蛋白质-胁迫关联[23]。Pavlopoulos 等人开发出来一种名为 OnTheFly 的在线工具,可以实现对 Office、PDF 以及普通文本文件的信息提取,构建知识图形和网络结构,这个具体可以运用于生物学数据的分析和模型构建(例如蛋白质相互作用网络的构建)。

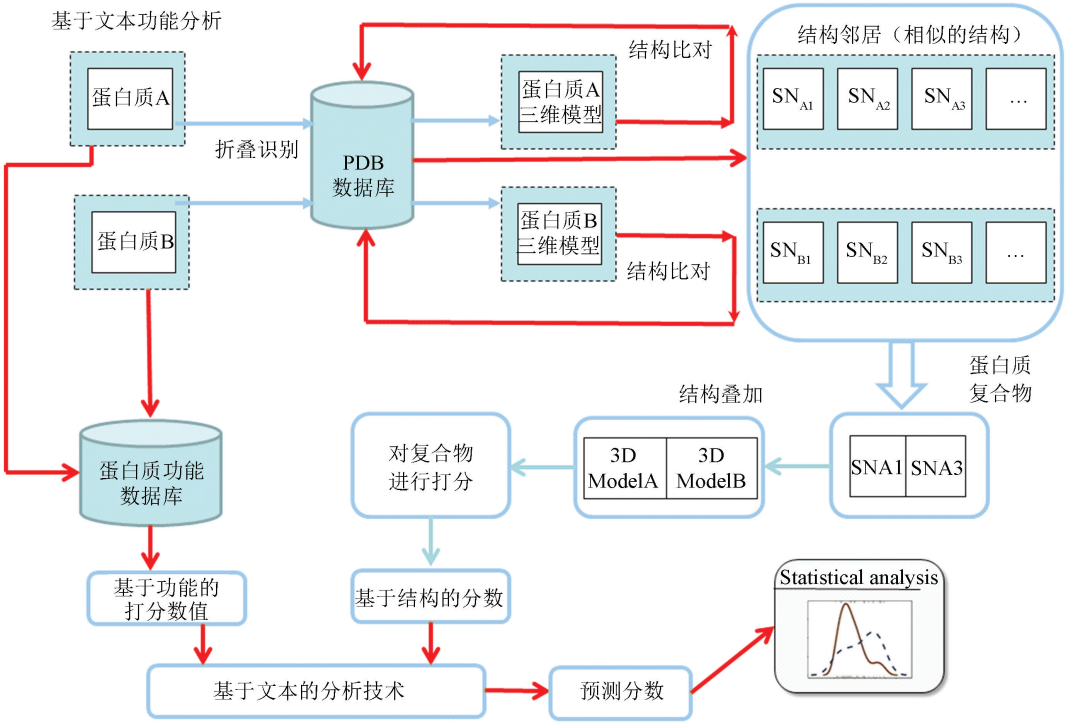

图 4 基于蛋白质结构与功能数据库的蛋白质功能预测模型

Fig.4 Protein function prediction model based on protein structure and function databases

另外,深度学习在蛋白质生物信息学中也有广泛的应用。例如深度学习算法应用于蛋白质的无规则卷曲模型构建[24]。在蛋白质功能位点预测方面(例如磷酸化位点),深度学习算法也有应用实例[25]。近些年来,随着深度学习算法的发展,之前使用传统文本分析方法建立的模型,研究者也尝试对一部分案例通过深度学习算法进行进一步改进。

6 结 语

文本分析技术是开展科技情报研究工作以及其他相关科学研究的技术支撑之一,在新时期决策难度与风险不断加大的情境下,国内各个相关部门和研究机构有必要审时度势,积极研究这种技术并把之应用于政府资料的收集、分析以及科学研究的各个方面。令人欣慰,文本分析技术已经广泛应用于生物信息学领域中,这也体现在国内不少高校的信息技术相关专业科研教学人员也广泛从事生物信息学研究。本文主要探讨了文本技术在蛋白质计算研究中的几个应用实例。在不同领域中的广泛应用,这也进一步促进了文本分析技术的进展。进展的文本分析技术又可以进一步促进其在科技情报分析中的应用。

本文讨论了文本分析在蛋白质计算分析中的几个应用实例以及核心技术内容。在生物信息学中的应用,进一步促进了文本分析技术的进展。特别是在生物序列的编码,表示序列以及结构相似性的方面,进一步拓宽了文本分析的面,同时也增加了文本分析的内涵。科学研究方面,特别是生物学研究方面,通过理论计算模拟和实验过程有助于对相应分子机制的理解。例如,对蛋白质特有的结构特征进行深入研究,进而阐述与蛋白质相关的生物学功能的分子机制,同时也有助于深入理解膜蛋白序列-结构-功能关系的相应分子机理。相应的研究成果能够应用于基因组和蛋白质组规模的数据应用,也可为新型药物靶标的筛选和药物的开发提供帮助。

文本分析技术在生物信息学中的应用远不止在蛋白质方面,在生物基因组信息学方面也有重要应用。一般而言,能把生物学数据转化为文本数据的方面,都有应用文本分析技术的可能,具体是否可以使用可以阅读相应文献或者进行实际研究测试。另外,本文虽然讨论的是文本分析技术在蛋白质生物信息学领域的应用。但是,一些数据是用文本分析方法是不能直接处理的,例如图片、声音以及视频数据等。图像声音以及视频需要使用其他方法进行数据的基础处理。随着计算机技术的发展,有一些文本采用二进制的格式进行存储,这种格式人工无法直接读取,这可能也是文本技术需要面对的一个问题。文本分析技术仍然存在许多尚待改进的方面。例如,如果是动态地获得网络的上的数据? 以及如何降低建模预测出案例的假阳性率? 以及得到的数据和模型如何实现实时吸取新的文本数据中的知识? 构建模型的文本分析方法是否可以进一步优化等? 这些也许是科研人员需要思考以及不断提出新方法改进的方向。

参考文献(References)

[1]LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444. DOI: 10.1038/nature14539.

[2]GAJENDRAN V K, LIN J R, FYHRIE D P. An application of bioinformatics and text mining to the discovery of novel genes related to bone biology[J]. Bone, 2007, 40(5): 1378-1388. DOI:10.1016/j.bone.2006.12.067.

[3]CHOU K C, ZHANG C T, KEZDY F J, et al. A vector projection method for predicting the specificity of GalNAc-transferase[J]. Proteins, 1995, 21(2): 118-126. DOI: 10.1002/prot.340210205.

[4]ELHAMMER A P, POORMAN R A, BROWN E, et al. The specificity of UDP-GalNAc: Polypeptide N-acetylgalactosaminyltransferase as inferred from a database of in vivo substrates and from the in vitro glycosylation of proteins and peptides[J]. The Journal of Biological Chemistry, 1993, 268(14): 10029-10038. DOI:10.1023/A:1026465232149.

[5]JULENIUS K, MOLGAARD A, GUPTA R, et al. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites[J]. Glycobiology, 2005, 15(2): 153-164. DOI: 10.1093/glycob/cwh151.

[6]RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536. DOI:10.1038/323533a0.

[7]BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32. DOI:10.1023/A:1010933404324.

[8]VERT J P. Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings[J]. Pacific Symposium on Biocomputing, 2002, 7: 649-660. DOI:10.1142/9789812799623_0060.

[9]RAZAGHI A, VILLACRES C, JUNG V, et al. Improved therapeutic efficacy of mammalian expressed-recombinant interferon gamma against ovarian cancer cells[J]. Experimental Cell Research, 2017, 359(1): 20-29. DOI:10.1016/j.yexcr.2017.08.014.

[10]HAMBY S E, HIRST J D. Prediction of glycosylation sites using random forests[J]. BMC Bioinformatics, 2008, 9: 500. DOI:10.1186/1471-2105-9-500.

[11]PRUITT K D, TATUSOVA T, KLIMKE W, et al. NCBI Reference Sequences: Current status, policy and new initiatives[J]. Nucleic Acids Research, 2009, 37(Database issue): D32-D36. DOI:10.1093/nar/gkn721.

[12]DAMOULAS T, GIROLAMI M A. Probabilistic multi-class multi-kernel learning: On protein fold recognition and remote homology detection[J]. Bioinformatics, 2008, 24(10): 1264-1270. DOI:10.1093/bioinformatics/btn112.

[13]CHENG J, BALDI P. A machine learning information retrieval approach to protein fold recognition[J]. Bioinformatics, 2006, 22(12): 1456-1463. DOI:10.1093/bioinformatics/btl102.

[14]JONES D T. Protein secondary structure prediction based on position-specific scoring matrices[J]. Journal of Molecular Biology, 1999, 292(2): 195-202. DOI:10.1006/jmbi.1999.3091.

[15]ALTSCHUL S F, MADDEN T L, SCHAFFER A A, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25(17): 3389-3402. DOI:10.1093/nar/25.17.3389.

[16]GOEL R, HARSHA H C, PANDEY A, et al. Human protein reference database and human proteinpedia as resources for phosphoproteome analysis[J]. Molecular BioSystems, 2012, 8(2): 453-463. DOI: 10.1039/c1mb05340j.

[17]MAVERAKIS E, KIM K, SHIMODA M, et al. Glycans in the immune system and the altered glycan theory of autoimmunity: A critical review[J]. Journal of Autoimmunity, 2015, 57: 1-13. DOI:10.1016/j.jaut.2014.12.002.

[18]WISTRAND M, KÄLL L, SONNHAMMER E L. A general model of G protein-coupled receptor sequences and its application to detect remote homologs[J]. Protein Science, 2006, 15(3): 509-521. DOI:10.1110/ps.051745906.

[19]WIEDEMANN N, KOZJAK V, CHACINSKA A, et al. Machinery for protein sorting and assembly in the mitochondrial outer membrane[J]. Nature, 2003, 424(6948): 565-571. DOI:10.1038/nature01753.

[20]BIRNEY E, BATEMAN A, CLAMP M E, et al. Mining

the draft human genome[J]. Nature, 2001, 409(6822): 827-828. DOI:10.1038/35057004.

[21] PENG J, SCHWARTZ D, ELIAS J E, et al. A proteomics approach to understanding protein ubiquitination [J]. Nature Biotechnology, 2003, 21(8): 921-926. DOI:10.1038/nbt849.

[22] SEITZ O. Glycopeptide synthesis and the effects of glycosylation on protein structure and activity[J]. ChemBioChem, 2000, 1(4): 214-246. DOI:10.1002/1439-7633.

[23] BOHNE-LANG A, WILHELM C, LIETH V D. GlyProt: in silico glycosylation of proteins[J]. Nucleic Acids Research, 2005, 33(Web Server issue): W214-W219. DOI: 10.1093/nar/gki385.

[24] BRODEHL A, STANASIUK C, ANSELMETTI D, et al. Incorporation of desmocollin-2 into the plasma membrane requires N-glycosylation at multiple sites[J]. FEBS Open Bio, 2019, 9(5): 996-1007. DOI: 10.1002/2211-5463.12631.

[25] RIDER P J F, NADERI M, BERGERON S, et al. Cysteines and N-glycosylation sites conserved among all alphaherpesviruses regulate membrane fusion in herpes simplex virus 1 infection[J]. Journal of Virology, 2017, 91(21): e00873-17. DOI:10.1128/JVI.00873-17.

[责任编辑:吴永英]