基于关系相似性的蛋白质交互作用识别

王宇伟,牛 耘
(南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

摘 要:针对目前蛋白质提取方法仅以单句信息为依据的不足,文中提出了以相似性为框架基于大规模文本的蛋白质交互关系识别方法。首先通过搜索医学文献数据库建立蛋白质对的签名档,然后提取签名档中的重要特征建立蛋白质对的向量空间模型,最后通过 K 近邻分类方法判断蛋白质对的交互关系。实验比较了向量空间模型下不同的距离度量策略对分类效果的影响,得出了比较合理的衡量相似性的函数。结果表明基于大规模文本采用基于余弦距离度量相似性的近邻方法识别蛋白质交互关系取得了较高且均衡的精确度和召回率,并且此方法直接利用了已有的交互信息,从而免除了额外的人工标注负担。

关键词:关系相似性;蛋白质交互;空间向量模型;K 近邻分类
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2015)02-0042-05
doi:10.3969/j.issn.1673-629X.2015.02.010

Identification of Protein-protein Interaction Based on Relational Similarity

WANG Yu-wei, NIU Yun
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: For the deficiencies of current approaches on Protein-Protein Interaction (PPI) identification which based on single sentences, propose a relational similarity method for automatic identification of protein-protein interactions by searching large scale text. The signature of a protein pair is obtained first by searching large scale biomedical text. Then, features are extracted from the signatures to build the vector space model of the protein pair. Finally, K nearest neighbor classifier is applied to identify PPIs. The influence of various distance measurement strategies under vector space model on classification effect are compared and the rational similar function is obtained. Results show that this approach achieves high and well balanced precision and recall when taking cosine as the similarity measurement. In addition, this approach makes use of known PPIs thus releases the burden of manual annotation.

Key words: relational similarity; protein-protein interaction; vector space model; K nearest neighbor classification

0 引 言

蛋白质是生物细胞最重要的成分,它们通过彼此间的作用完成细胞中的大部分过程,蛋白质之间的交互信息(Protein-Protein Interaction,PPI)对于生物学研究具有重要的意义。因而,生物医学领域专家手工地从医学文献中收集这些信息录入统一格式的数据库中,比如:HPRD[1],BIND[2],DIP[3],IntAct[4],MINT[5],等等。然而随着生物医学文献的急剧增长,现已很难从海量数据中抽取出有价值的信息。因此,自动地从医学文献中抽取 PPI 已成为一项重要的研究内容。

目前,从医学文本中自动识别蛋白质交互信息的技术主要包括:基于同现的方法[6]、基于规则(模式匹配)的方法[7]和基于机器学习的方法[8-10]。基于同现的方法通过统计两个蛋白质在句子中的共现次数来判断蛋白质间是否存在交互关系,这种方法简单,结果召回率较高但精确度低[11];基于规则或模板匹配的方法提高了精确度但却导致比较低的召回率,另外手动建立规则需要巨大开销且通常规则只适用于特定的部分数据[12],这些都导致建立的规则覆盖面十分有限。

近些年来,越来越多的 PPI 识别技术采用了基于机器学习的方法。这类方法从标注有交互关系的句子中抽取重要特征建立模型来判断蛋白质之间的交互关

收稿日期:2014-02-28 **修回日期**:2014-05-28 **网络出版时间**:2014-12-27
基金项目:国家自然科学基金资助项目(61202132)
作者简介:王宇伟(1989-),男,硕士研究生,研究方向为自然语言处理;牛 耘,副教授,研究方向为自然语言处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20141227.1341.014.html

系,或者通过设计核函数进一步利用句子结构表示(如字符串序列、句法依赖或句法分析)上的隐含特征。然而,目前基于机器学习的方法主要以单句为依据采用基于监督的方式进行蛋白质交互关系识别,这类方法主要有两方面的局限:

(1)在进行交互关系判定时仅依赖于一句话的信息。由于蛋白质交互信息描述语言的多样性和句法结构的复杂性,以单句为依据的方法难以对交互特征进行全面的把握。

(2)一句话中往往包含三个以上蛋白质。这种方法需要对句子中的每对蛋白质对都进行标注,这势必需要耗费大量的人力资源。

针对以上这些问题,文中提出了一种以关系相似性为框架基于大规模文本的蛋白质交互识别方法,该方法直接以现有的 PPI 数据库作为训练数据从而避免了额外的人工标注负担,并且充分利用大规模文本库资源,依据文本中丰富的上下文信息,更全面地获取交互关系特征。实验分析了不同的相似性度量策略对 PPI 识别效果的影响。结果表明基于余弦距离度量的相似性识别蛋白质交互的方法取得了较高且均衡的精确度及召回率。

1　基于大规模文本的 PPI 识别方法

基于大规模文本的 PPI 识别方法可以看作是一个文本分类问题[13-14]。首先,在医学文献数据中搜索包含目标蛋白质对的句子集合作为此目标蛋白质对的签名档;然后从众多蛋白质对的签名档中提取特征,并且把每个蛋白质对实例映射为一个 n 维的特征向量;最后通过计算向量之间的相似性来判断目标蛋白质对之间是否存在交互关系。

具体方法步骤如下所述。

1.1　关系抽取

PubMed 数据库收录了超过一千八百万篇生物医学文献摘要,是建立 PPI 网络重要的信息来源。对于每一个目标蛋白质对(Protein1,Protein2),在 PubMed 数据库中通过搜索同时包含 Protein1 和 Protein2 这两个蛋白质的句子集合作为目标蛋白质对的签名档。由于 PubMed 没有提供直接检索句子的接口,于是分以下两步来完成:

(1)检索包含目标蛋白质对的摘要。

调用 PubMed 数据库提供的接口搜索包含目标蛋白质对的摘要。

· 首先,以 Protein1 和 Protein2 作为参数调用检索命令 esearch 搜索那些摘要的 ID;

· 然后,实际的摘要文本再以检索到的 ID 作为输入调用获取命令 efetch 来获得。

(2)搜索包含目标蛋白质对的句子。

接下来,通过处理上一步检索到的摘要集合得到包含目标蛋白质的句子集合。

· 首先,使用伊利诺州大学 urbana-champaign 分校认知计算研究组开发的句子识别工具[15]来识别摘要集合中的句子。

· 然后,只保留那些同时含有 Protein1 和 Protein2 这两个蛋白质的句子作为目标蛋白质对的签名档内容。

· 最后,每一个蛋白质对都会有一个集合与之对应,也就是它的签名档,集合中包含一个或多个句子。建好蛋白质对的签名档之后,即可以利用这些上下文信息对目标蛋白质对的交互关系做出判断。

1.2　关系表示

文中采用向量空间模型来表示蛋白质 Protein1 和 Protein2 之间的关系 R。向量的维是刻画这一关系的单词特征,因目标蛋白质对的签名档中包含了关系 R 的完整描述,把所有签名档中的单词去除那些停止词,单字符单词以及无意义的数字之后剩余的单词作为特征,并且排除了那些低频词(出现此单词的签名档数少于 25 个),最终留有 4 867 个单词特征。每一个关系 R 用一个 4 867 维的特征向量来刻画,这个向量的特征权重对应于该特征单词是否在目标蛋白质对对应的签名档中出现,出现时特征值为 1,否则就为 0。

1.3　相似性计算

这一步通过比较目标蛋白质对与已知交互关系的蛋白质对的相似性来判定目标蛋白质对是否具有交互关系。文中比较了几种不同的相似性计算策略对于 PPI 识别效果的影响。目标蛋白质对以及已知交互关系的蛋白质对的相互作用关系分别用两个向量 $\vec{v_1}$,$\vec{v_2}$ 表示,它们之间的相似性可以用以下几种距离度量方法(公式(1) ~ (4))来度量,距离值越小代表相似度越大。

其中,$\vec{v_1} = (v_{1,1},\cdots,v_{1,i},\cdots,v_{1,n})$,$\vec{v_2} = (v_{2,1},\cdots,v_{2,i},\cdots,v_{2,n})$。

· 欧几里得(欧氏)距离:

$$D(\vec{v_1},\vec{v_2}) = \sqrt{\sum_{k=1}^{n}(v_{1,k} - v_{2,k})^2} \tag{1}$$

· 曼哈顿距离:

$$D(\vec{v_1},\vec{v_2}) = \sum_{k=1}^{n}|v_{1,k} - v_{2,k}| \tag{2}$$

· 杰卡德距离:这种度量方法是将向量 $\vec{v_1}$,$\vec{v_2}$ 当作两个集合 X,Y 来处理,集合的元素分别是两向量代表的蛋白质对的签名档中出现的特征。

$$D(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} \tag{3}$$

· 余弦距离：

$$D(\vec{v_1},\vec{v_2}) = \frac{\sum_{k=1}^{n} v_{1,k} \bullet v_{2,k}}{\sqrt{\sum_{k=1}^{n} (v_{1,k})^2} \bullet \sqrt{\sum_{k=1}^{n} (v_{2,k})^2}} \quad (4)$$

1.4 *K* 近邻分类

得到了实例的相似性之后，基于它们的相似性采用 *K* 近邻分类方法识别目标蛋白质之间的交互关系。*K* 近邻方法是基于统计的分类方法，也是文本分类中比较常用的方法[16-17]，其基本思想对应此分类任务是根据得到的实例间的距离考察目标蛋白质对的（关系相似性）最相近的 k 个蛋白质对实例，这 k 个最相似的实例中哪种类别的实例最多，就将目标蛋白质对分为哪一类。

令 $C=\{+1,-1\}$，+1 表示目标蛋白质之间有交互，-1 表示无交互。$f(p)$ 表示目标蛋白质的类别号，*K* 近邻算法如图 1 所示。

训练算法：

对于每个训练样例<x,$f(x)$>，把这个样例加入训练列表 training_examples

分类算法：

给定一个待识别的目标蛋白质对实例 p

在 training_examples 中选出距离最靠近 p 的 k 个实例，并用 x_1，x_2，…，x_k 表示

返回 $f(p) \leftarrow \underset{v\in C}{\operatorname{argmax}} \sum_{i=1}^{k} \delta(v, f(x_i))$

其中，如果 $a=b$ 那么 $\delta(a,b)=1$，否则 $\delta(a,b)=0$

图 1 *K* 近邻分类算法

在此算法中，如果存在多个距离目标蛋白质对一样近的实例，把相同距离的这些实例当作一个实例来处理，其类别号由这些实例的多数决定，哪种类别实例最多，则类别号归为哪一类。这样算法中的 x_1，x_2，…，x_k 与 p 之间的相似度互不相同。

2 实验及结果分析

2.1 实验数据

实验全部的训练数据来自于现有的 PPI 数据库而不需要额外的人工标注。把有交互的蛋白质对看作正样例，无交互的看作负样例。正样例来源于专家手工收集的人类 PPI 数据库 HPRD[1]，抽出其中那些包含在 PubMed 数据库一篇以上摘要中的蛋白质对作为有交互的蛋白质对训练集，共 1 420 对。而对于负样例，采用生物信息学领域常用的方法，首先对 HPRD 中的蛋白质随机组合成蛋白质对，并且去除包含在 HPRD 数据库中的组合，最后只保留那些包含在 PubMed 数据库一篇以上摘要中的组合，作为最后无交互的蛋白质对训练集，总共 1 353 对。因此，实验数据集中共包含了 2 773 个蛋白质对。

2.2 实验设置

实验采用的结果性能评价指标是当前 PPI 抽取系统主要使用的三个指标：精确度（Precision = TP/(TP+FP)）、召回率（Recall = TP/(TP+FN)）和 *F* 值（*F*-Score = $2P\times R/(P+R)$）。对于数据集中的每个蛋白质对都基于它们的签名档建立对应的特征向量，并采用留一交叉验证法（leave-one-out）进行测试，即将每个蛋白质对作为测试样例，其余的作为训练样例，这样共测试 2 773 次。最后采用图 1 描述的分类算法完成识别过程。

2.3 实验结果及讨论

实验比较了不同的相似性度量策略对识别结果的影响，表 1 至表 3 分别列出了采用欧几里得和曼哈顿距离、杰卡德距离以及余弦距离度量相似性并采用 *K* 近邻分类所得到的精确度、召回率及 *F* 值，*K* 赋值为奇数从 1 增长至 15，数据以% 为单位。图 2 和图 3 对比了采用这四种距离度量相似性后识别结果的 *F* 值的变化趋势。其中欧几里得距离与曼哈顿距离度量相似性在建立的向量空间模型下得到了相同的分类结果（如表 1）。

表 1 欧几里得及曼哈顿距离度量相似性的 *K* 近邻分类结果

K	Positive			Negative		
	Precision	Recall	*F*-Score	Precision	Recall	*F*-Score
1	86.990 8	46.619 717	60.706 093	62.326 042	92.682 93	74.531 944
3	93.963 78	32.887 325	48.721 966	58.128 296	97.782 71	72.912 65
5	95.238 1	25.352 112	40.044 49	55.741 127	98.669 624	71.237 99
7	97.222 22	22.183 098	36.123 85	54.879 543	99.334 81	70.699 63
9	97.689 766	20.845 07	34.358 677	54.493 927	99.482 63	70.415 9
11	98.168 495	18.873 24	31.659 777	53.92	99.630 45	69.971 45
13	98.113 205	18.309 858	30.860 53	53.748 005	99.630 45	69.826 47
15	98.418 976	17.535 212	29.766 888	53.531 746	99.704 36	69.661 76

表 2　杰卡德距离度量相似性的 K 近邻分类结果

K	Positive			Negative		
	Precision	Recall	F-Score	Precision	Recall	F-Score
1	75.859 6	74.577 46	75.213 066	73.783 585	75.092 384	74.432 23
3	77.027 02	72.253 525	74.563 96	72.657 875	77.383 59	74.946 31
5	79.049 84	71.478 874	75.073 96	72.800 54	80.118 256	76.284 3
7	79.091 62	71.126 76	74.898 03	72.593 58	80.266 075	76.237 274
9	79.903 534	70.0	74.624 626	72.138 65	81.522 545	76.544 07
11	79.935 27	69.577 46	74.397 59	71.893 295	81.670 364	76.470 59
13	80.241 936	70.070 42	74.812 03	72.276 58	81.892 09	76.784 48
15	79.887 82	70.211 266	74.737 625	72.262 3	81.448 63	76.580 96

表 3　余弦距离度量相似性的 K 近邻分类结果

K	Positive			Negative		
	Precision	Recall	F-Score	Precision	Recall	F-Score
1	75.571 43	74.507 04	75.035 46	73.634 38	74.722 84	74.174 62
3	77.646 19	73.873 24	75.712 74	73.909 99	77.679 23	75.747 75
5	77.777 78	72.957 75	75.290 695	73.351 84	78.122 69	75.662 14
7	77.638 374	74.084 51	75.819 82	74.047 96	77.605 32	75.784 91
9	77.282 85	73.309 86	75.243 95	73.422 16	77.383 59	75.350 845
11	76.406 136	73.661 97	75.008 96	73.361 824	76.127 13	74.718 9
13	76.444 77	73.591 55	74.991 035	73.328 59	76.201 035	74.737 22
15	76.283 44	74.295 78	75.276 49	73.741 005	75.757 576	74.735 695

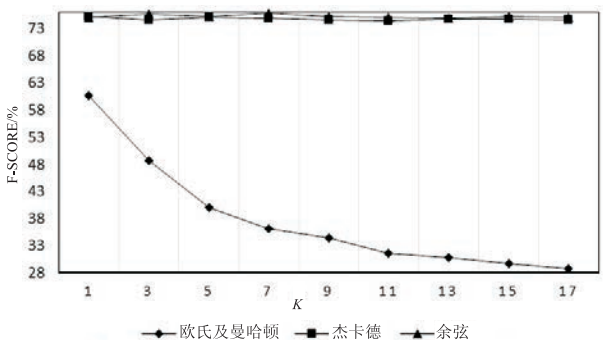

图 2　有交互关系的蛋白质识别结果的 F 值变化趋势对比

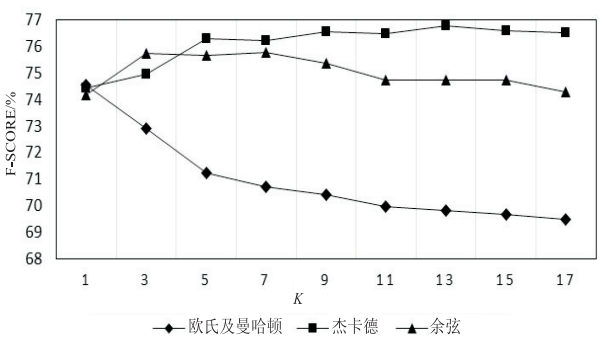

图 3　无交互关系的蛋白质识别结果的 F 值变化趋势对比

从以上结果的比较可以看出,采用欧氏距离及曼哈顿距离在以上建立的向量空间模型下来度量关系的相似性,识别结果(见表 1)的精确度与召回率相差很大,对于有交互的蛋白质对精确度很高而召回率很低,随着 K 值的增大,精确度在不断增加而召回率不断下降;而对于无交互关系的蛋白质对与此相反,精确度很低,召回率很高,随着 K 的增大,召回率增加而精确度下降,但是有交互的蛋白质组和无交互蛋白质组的精确度和召回率的差距都是随 K 的增加越来越大。基于杰卡德度量方法的识别结果(见表 2)一开始精确度和召回率较均衡,但随着 K 的增大,精确度和召回率与欧氏方法有着同样的变化趋势,差距也是逐渐变大。而采用余弦距离作为相似性度量策略(见表 3),不管是有交互的蛋白质组还是无交互的蛋白质组,精确度及召回率都维持在一个较高的值,且两者较为均衡,随着 K 的增加,精确度、召回率的变化幅度不大,两者之间的差距先是有所变大,后又逐渐缩小到和最初一样。当 K 上升至 7 时,精确度和召回率的 F 值都达到了最大值,分别为 75.82% 和 75.78%。

对于 F 值,由图中可以看到,采用欧氏距离与曼哈顿距离度量相似性后的识别结果是随着 K 的上升降到了一个很低的值,而杰卡德距离与余弦距离的度量相似性后的识别结果则相近,并且随着 K 的增长而变化较小。

综合来看,采用余弦距离作为此向量空间模型下的相似性计算函数较为合理。从基于余弦距离度量的识别结果来看,描述蛋白质交互关系的文本存在着共性,即描述交互关系的不同句子中,对目标蛋白质对的表达是相似的,而从大规模文本中提取的这些单词能有效地表示这些共性,从而可以建立起好的相似性计算模型对关系做出正确的判断。

3 结束语

文中提出了一种关系相似性框架下基于大规模文本数据识别蛋白质交互的方法,与广泛采用的基于单句的机器学习方法不同,该方法直接以蛋白质对为研究对象,以大规模文本为依据提取特征建立相似性计算模型,并利用 K 近邻分类器识别给定的两个蛋白质是否存在相互作用关系。识别结果可直接用于 PPI 网络的构建,此方法还能充分利用已有的 PPI 数据而无需额外的人工标注。文中比较了多种相似性度量策略对识别效果的影响,结果可以看出基于余弦距离的相似性计算函数在建立的向量空间模型下较为合理,根据余弦相似性采用近邻的方法自动识别 PPI 取得了较高且较均衡的精确度和召回率。

参考文献:

[1] Prasad T SK, Goel R, Kandasamy K, et al. Human protein reference database – 2009 update[J]. Nucleic Acids Research, 2009, 37: 767–772.

[2] Bader G D, Betel D, Hogue C W V. BIND: the biomolecular interaction network database[J]. Nucleic Acids Research, 2003, 31(1): 248–250.

[3] Salwinski L, Miller C S, Smith A J, et al. The database of interacting proteins: 2004 update[J]. Nucleic Acids Research, 2004, 32: 449–451.

[4] Kerrien S, Alam – Faruque Y, Aranda B, et al. IntAct – open source resource for molecular interaction data[J]. Nucleic Acids Research, 2007, 35: 561–565.

[5] Ceol A, Aryamontri A C, Licata L, et al. MINT, the molecular interaction database: 2009 update[J]. Nucleic Acids Research, 2010, 38: 532–539.

[6] Bunescu R, Mooney R, Ramani A, et al. Integrating co–occurrence statistics with information extraction for robust retrieval of protein interactions from Medline[C]//Proceedings of the workshop on linking natural language processing and biology: towards deeper biological literature analysis. [s. l.]: Association for Computational Linguistics, 2006: 49–56.

[7] Koike A, Kobayashi Y, Takagi T. Kinase pathway database: an integrated protein–Kinase and NLP–based protein–interaction resource[J]. Genome Research, 2003, 13(6A): 1231–1243.

[8] 杨志豪,洪 莉,林鸿飞,等. 基于支持向量机的生物医学文献蛋白质关系抽取[J]. 智能系统学报,2008,3(4):361–369.

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

[10] 崔宝今,林鸿飞,张 霄. 基于半监督学习的蛋白质关系抽取研究[J]. 山东大学学报:工学版,2009,39(3):16–21.

[11] Grimes G R, Wen T Q, Mewissen M, et al. PDQ Wizard: automated prioritization and characterization of gene and protein lists using biomedical literature[J]. Bioinformatics, 2006, 22(16): 2055–2057.

[12] Ananiadou S, Kell D B, Tsujii J. Text mining and its potential applications in systems biology[J]. Trends in Biotechnology, 2006, 24(12): 571–579.

[13] 陈治纲,何丕廉,孙越恒,等. 基于向量空间模型的文本分类方法的研究与实现[J]. 计算机应用,2004,24(06Z):277–279.

[14] 饶文碧,柯慧燕. Web 文本分类技术研究及其实现[J]. 计算机技术与发展,2006,16(3):116–118.

[15] University of Illinois at Urbana–champaign. Sentence segmentation tool[EB/OL]. [2011–09–23]. http://cogcomp.cs.illinois.edu/page/tools_view/2.

[16] 许 幸,张启蕊. 基于 KNN 算法的医药信息文本分类系统的研究[J]. 计算机技术与发展,2009,19(4):206–209.

[17] 王 煜,白 石,王正欧. 用于 Web 文本分类的快速 KNN 算法[J]. 情报学报,2007,26(1):60–64.

(上接第 24 页)

2010: 21–30.

[5] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks[C]//Proc of the 19th ACM international conference on information and knowledge management. New York: ACM, 2010: 1633–1636.

[6] Romero D M, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics: idioms, plitical hashtags, and complex contagion on Twitter[C]//Proc of the 20th international conference on World Wide Web. New York: ACM, 2011: 695–704.

[7] 吴雨蓉. 微博信息传播模式分析[J]. 渤海大学学报:哲学社会科学版,2012(2):140–143.

[8] Weng J, Lim E P, Jiang J, et al. TwitterRank: finding topic sensitive influential Twitters[C]//Proc of the 3rd ACM international conference on web search and data mining. New York: ACM, 2010: 261–270.

[9] 李 军,陈 震,黄霁崴. 微博影响力评价研究[J]. 信息网络安全,2012(3):10–13.

[10] 韩朝阳,张仁军. 面向旅游应用的微博信息信度和效度评价[J]. 重庆理工大学学报:社会科学版,2011,25(10):37–40.

[11] 齐 娜,宋立荣. 医疗健康领域微博信息传播中的信息质量问题[J]. 科技导报,2012,30(17):60–65.

[12] 张豪锋,杨绪辉. 教育微博社群中首帖质量的分析与对策[J]. 远程教育杂志,2012(2):98–103.

[13] 莫祖英,马费成,罗 毅. 微博信息质量评价模型构建研究[J]. 信息资源管理学报,2013(2):12–18.

[14] 周庆山,梁兴堃,曹雨佳. 微博中意见领袖甄别与内容特征的实证研究[J]. 山东图书馆学刊,2012(1):22–27.

基于关系相似性的蛋白质交互作用识别

作者： 王宇伟， 牛耘， WANG Yu-wei， NIU Yun
作者单位： 南京航空航天大学 计算机科学与技术学院,江苏 南京,210016
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(2)