基于序列功能注释的蛋白质相互作用预测方法研究

陈 霞[1], 陈浩文[2*]　(1. 长沙航空职业技术学院,湖南长沙 410124;2. 湖南大学信息科学与工程学院, 湖南长沙 410082)

摘要　蛋白质 - 蛋白质相互作用(PPI)是大多数细胞过程和生物功能的基础。该研究基于蛋白质功能注释方法(FNM)首次提出了结合蛋白的重要性的方法、结构域相互作用、基因本体论注释序列和注释;然后融合不同的策略,分别建立了 3 种方法结合的蛋白质序列特征与 FNM 功能注释功能。利用蛋白质相互作用预测构建蛋白质相互作用网络是进一步理解蛋白质功能的必要前提,也是理解细胞新陈代谢及复杂疾病形成发生的基础和关键。

关键词　蛋白质相互作用;多源信息融合;功能预测;本体注释
中图分类号　S126　　**文献标识码**　A　　**文章编号**　0517 - 6611(2015)28 - 352 - 02

Sequence and Functional Annotations-based Prediction of Protein-protein Interactions

CHEN Xia[1], CHEN Hao-wen[2*]　(1. Changsha Aeronautical Vocational and Technical College, Changsha, Hunan410124; 2. School of Information Science and Engineering, Hunan University, Changsha, Hunan 410082)

Abstract　Protein protein interaction (PPI) is the basis of most cellular processes and biological functions. This paper first proposed the method of binding proteins by protein functional annotation method (FNM) for the first time. And then, three methods of combining the characteristics of protein sequences and the function of FNM were established. Prediction of protein-protein interaction network is a necessary prerequisite for understanding the function of proteins, which is the basis and key to understand the formation of cell metabolism and complex diseases.

Key words　Protein interaction; Multi-source information fusion; Functional prediction; Ontology annotation

研究人员可以从实验检测方法或者计算生物学 2 个角度研究蛋白质的相互作用。尽管生物实验检测方法可以得到大量的 PPI 数据,但这些实验方法的成本昂贵,并且实验所导致的高假阳性等缺陷使得它们不能作为标准使用。而计算生物学方法具有低成本、效率高等优点,从而被研究人员广泛关注,该方法可以通过分析大规模数据来分析 PPI 网络随时间变化的特性。

2001 年 Bock 等[1]首先提出利用支持向量机预测蛋白质相互作用方法,该方法仅依靠蛋白质序列本身的数据即可以预测其相互作用,随后更多的研究者也提出了基于序列保守型的改进方法。但是随着数据的不断增加,新预测的蛋白质相互作用中存在大量的假阳性数据,因此,一些基于文本挖掘、蛋白质空间结构、基因功能注释等多源信息的方法相继被提出[2-3]。

1　基于蛋白质序列信息及本体注释信息融合的预测方法

单独利用某一种信息可能难以获得最优的效果,而将多种互补的信息融合能最大限度地预测蛋白质相互作用网络[4-6]。该研究综合利用蛋白质序列信息、结构信息、基因本体注释以及序列注释等,预测蛋白质相互作用。

图 1 显示了该研究融合策略方法的研究框架。从图 1 可看出,该方法融合了 4 种类型的先验知识如蛋白质重要性、域相互作用、基因及序列的本体注释。融合策略方法基于一个重要假设即蛋白质序列信息与其他信息(基因本体注释信息等)是互补的。根据不同的融合策略,该研究设计了多种蛋白质预测方法。该研究中蛋白质序列信息采用 CT 方法[9]获取,以下分别详细介绍 4 种功能注释方法及 3 种融合策略。

基金项目　湖南省教育厅资助科研项目(12C0921)。
作者简介　陈霞(1983 -),湖南邵阳人,讲师,硕士,从事生物信息学研究。*通讯作者,讲师,博士,从事生物信息学研究。
收稿日期　2015-08-21

1.1　蛋白质功能注释方法

为了从基因功能等角度获取蛋白质相关的先验信息以弥补蛋白质序列信息的局限性,以下采用了 4 种功能注释方法:

(1)重要程度。考虑到蛋白质对某个组织器官的作用,每个蛋白质可以被划分为重要的或者不重要的。利用公式(1)的编码方案可以将该信息描述为一个 1 维向量:

$$f_{EP}=\left[v_{EP}=\begin{cases}2,\text{当 2 个蛋白质都重要时}\\1,\text{只有 1 个蛋白质重要时}\\0,\text{其他}\end{cases}\right]_1^T \tag{1}$$

(2)蛋白质域相互作用。蛋白质域是蛋白质序列的一部分,是蛋白质结构的子单元及进化模块,它们一定程度上决定了蛋白质的功能。利用公式(2)将其编码为一个 1 维向量:

$$f_{DDI}=\left[v_{DDI}=\begin{cases}1,\text{如果存在 } DDI\\0,\text{否则}\end{cases}\right]_1^T \tag{2}$$

(3)基因本体注释。基因本体由 3 个部分组成,描述了生物过程、分子功能以及细胞组成等知识。该研究采用 Resnik 度量[7]。从公式(3)可以看出,这里需要使用一个 3 维向量用于描述 2 个蛋白质之间的基因注释相似性。

$$f_{GO}=[\text{gosim\_}bp(p1,p2),\text{gosim\_}mf(p1,p2),\text{gosim\_}cc(p1,p2)]_3^T \tag{3}$$

(4)序列注释。基于蛋白质序列自身的多个角度特征,如空间结构、功能性质等,39 个不同特征如激活位点、beta 折叠、结合位点等被用于注释蛋白质序列。利用方差分析进一步验证这些特征是否与蛋白质相互作用相关,最终选出 26 个特征用于分析:

$$f_{SN}=[v_1,\cdots,v_i,\cdots,v_{26}]_{26}^T \tag{4}$$

1.2　注释信息融合

为了验证该研究中 4 种注释的先验知识有效性,利用公式(5)将每个蛋白质表示为 31 维的向量,该注释信息的融合被称为 FNM:

$$f_{FNM}=[f_{EP},f_{DDI},f_{GO},f_{SN}]_{31}^T \tag{5}$$

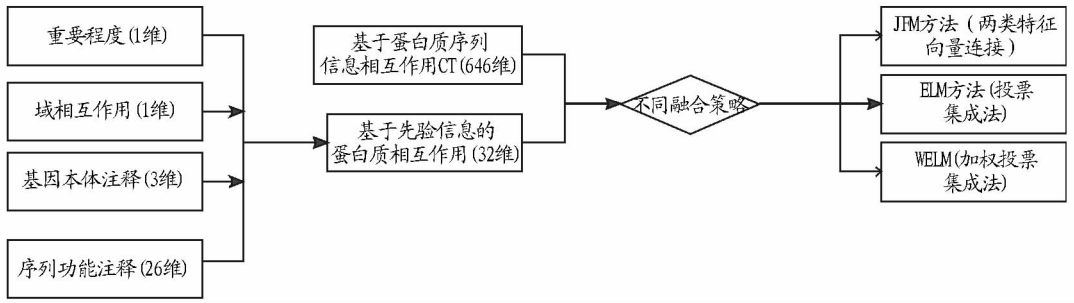

图 1　融合策略方法的基本框架

1.3　序列信息及注释信息融合　该研究将利用 3 种融合模型以集成序列信息及先验注释信息。第一种融合策略 JFM 利用公式(6)将 2 类特征融合为一个特征;而第二种策略 ELM 是利用 2 个基分类器构成集成学习模型,其中一个分类器利用 CT 方法提取的序列信息,一个分类器用于学习注释信息,如公式(7)所示;第三种融合策略 WELM 引入了一个权重因子,通过调节参数 p 以获得最大的预测准确度,如公式(8)所示。

$$f_{JFM} = [f_{CT}, f_{EP}, f_{DDI}, f_{GO}, f_{SN}]^T_{717} \tag{6}$$

$$dec\_ELM = 0.5 \times dec\_CT + 0.5 \times dec\_FNM \tag{7}$$

$$dec\_WELM = (\max p \times dec\_CT + (1 - \max p) \times dec - FNM)$$

$$\text{where } \max p = \arg \max_{p \in [0,1]} (accuracy) \tag{8}$$

2　数据集及评价指标

2.1　数据集　采用了 2 种规模的数据集用于评估该研究改进方法(融合策略方法)的性能。第一个数据集是 GUO 数据集[8],该数据集是已经存在的。该研究从多个数据集中收集蛋白质相互作用而构建了第二个数据集。前者仅仅包含酵母菌数据,而后者同时包含了酵母菌数据和人类数据,利用数据预处理方法最终得到 5 594 个蛋白质相互作用数据集。

2.2　评价指标　以下采用准确度(ACC),敏感度(SN)、阳性精确度(PE)以及 Matthew's 关联系数(MCC)评价该研究中的融合策略方法,如以公式(8)~(11)所示。同时,利用 ROC 曲线直观描述。

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$SN = \frac{TP}{TP + FN} \tag{9}$$

$$PE = \frac{TP}{TP + FP} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{11}$$

式中,TP 表示真阳性;FP 表示假阳性;TN 表示真阴性;FN 表示假阴性。

3　结果与分析

表 2 中列出了 6 个基于序列信息的方法以及该研究提出的 4 个策略(FNM、JFM、ELM 以及 WELM)分别在 GUO 数据集上的试验结果。前 6 个方法中 CT 方法具有最好的性能。因此,该研究后续的融合过程中选择采用该方法的信息作为序列特征。FNM 由于只包含了 31 维的特征,从而导致信息不足,但是它的计算复杂度较小。而同时融合了注释信息以及序列信息的 JFM、ELM 以及 WELM 大大提高了性能。图 2 是所有方法的 ROC 曲线性能图。如图 2 所示,同时融合序列信息以及注释信息的 3 个方法的性能最优,并且这种优势贯穿所有 FP 上。这现象表明该研究改进的基于序列功能注释的蛋白质相互作用预测方法的假阳性和假阴性都得以降低,从而提高了真阳性和真阴性。

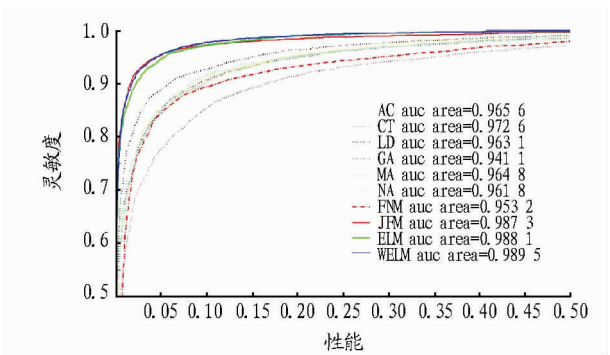

图 2　ROC 曲线比较结果

表 2　GUO 数据集上 5 折交叉验证结果比较

方法	SN//%	PE//%	ACC//%	MCC//%	AUC//%
AC	88.13	92.69	90.59	81.27	96.56
CT	89.99	93.85	92.05	84.16	97.26
LD	88.31	91.95	90.43	80.91	96.31
GA	85.66	89.29	87.69	75.45	94.11
MA	88.31	92.75	90.70	81.50	96.48
NA	87.79	92.24	90.20	80.50	96.18
FNM	88.06	91.77	90.08	80.22	95.32
JFM	93.62	96.59	95.16	90.35	98.73
ELM	93.08	96.11	94.65	89.35	98.81
WELM	93.60	96.46	95.08	90.21	98.95

4　结语

该研究详细阐述了一种基于融合信息的蛋白质相互作用预测方法,该方法利用序列信息和功能注释信息的互补性,设计不同的融合方案,然后在不同数据集上进行了试验比较。试验结果从多个准则上验证了该研究改进的融合策略方法具有较好的泛化能力,且假阳性率较低。

参考文献

[1] BOCK J R, GOUGH D A. Predicting protein – protein interactions from primary structure[J]. Bioinformatics, 2001, 17(5): 455 – 460.

[2] BUI Q C, KATRENKO S, SLOOT P M A. A hybrid approach to extract protein – protein interactions[J]. Bioinformatics, 2011, 27(2): 259 – 265.

(下转第 392 页)

种多品种的混养模式容易造成动物之间疾病的交叉感染,放大了本来就不低的养殖业产业风险。因此,每户农户饲养 2 种动物为宜,控制饲养品种,提高饲养动物的专业性和技术程度,提高质量,扩大规模。

3.2 提高生产专业化水平,建设支持体系 倡议中国扶贫项目,当地政府和大学开展相应的养殖技术培训项目或者培训班,传播现代养殖技术给当地农民。同时,养殖业的相关支持体系也应该健全。中国项目可以针对从事养殖业的农户挑选设立示范户,给予一定的援助,以达到示范带头、辐射其他农户的目的。当地政府向村中已有的兽医提供更多的补助和培训机会,如果政府缺乏资金,农民自有的合作社可以尝试集资给村中兽医提供培训,比较兽医防止动物因病死去所带来的成本止损和培训所需成本二者的差异,再由村民集体决定。继续引进和推动国际(或区域性)组织在当地开发、组织项目,支持农户养殖业发展。

3.3 激活农民合作社,促进互惠合作 Peapea 村本已有牲畜协会这种农民专业合作社,因此应激活农民合作社,促使其发挥作用。以前的研究表明合作社在提供农户专业技术、激活农民积极性、提供农户收入以及赋权方面发挥着积极的作用。农民合作社的运作提高了农户生计的可持续性。

首先,农民合作社可以减少农户在生产过程中生产要素的投入,合作社统一购买牲畜所需要的治病药物或者营养补充剂、笼舍搭建所需建材、简易的投食装置等,会比农户单独在市场上购买更有议价权,因此有更大可能以较低的价格购买到这些生产要素。其次,合作社能减低农户出售产品时在市场中的脆弱地位。合作社可以统一集中各农户所饲养动物出售,拓宽产品的销售渠道,同时农户间组成利益共同体进行市场交易活动,也避免农户单独被动地参与到市场竞争中的情况。通过加入合作社,农户在产品销售上有更多选择,合作社能够为农户提供一定的保障,增强农户的社会资本,提高收入的安全性,促进农户生计的可持续发展。第三,合作社中的农民能够互相交流养殖技术,互惠合作,实现共同进步与共同发展。

4 小结

通过对 Peapea 村的调查研究,笔者对该村家庭养殖有了较完整的认识。Peapea 村养殖业发展十分落后,养殖业养殖方式落后、基础设施匮乏、相应的补贴和支持体系不健全等因素都制约着养殖业的发展。值得欣喜的是,当地农民对于发展养殖业的意愿都很高,也已经有一些国际性和区域性组织机构在当地开展促进养殖业发展的工作。就当地现有的养殖业发展水平而言,中国有相对丰富的“中国经验”可以借鉴给当地学习。通过改善基础设施,提高农民饲养的专业化水平,激活当地的农民合作社等方式,都可以有效地促进农民增收致富,达到减贫的目标。

参考文献

[1] ELLIS F. Rural livelihoods and diversity in developing countries[M]. Oxford:Oxford University Press,2000.
[2] ELLIS F,BAHIIGWA G. Livelihoods and rural poverty reduction in Uganda[J]. World development,2003,31(6):997 - 1013.
[3] ELLIS F,BAHIIGWA G. Livelihoods and rural poverty reduction in Tanzania[J]. World development,2003,31(8):1367 - 1384.
[4] 吴莹莹. 农户生计多样化和土地利用变化[D]. 重庆:西南大学,2009:33.
[5] 阎建忠,吴莹莹,张镱锂,等. 青藏高原东部样带农牧民生计的多样化[J]. 地理学报,2009,64(2):221 - 233.
[6] 我国家庭农场规模及发展情况分析[EB/OL]. (2013 - 06 - 05)[2015 - 08 - 20]. http://www.51report.com/free/3018126.html.
[7] 齐想. 欧洲和非洲家庭式家禽生产掠影[J]. 中国家禽,2011,33(5):61 - 62.
[8] 程军波. 中国家禽业走向非洲[J]. 中国禽业导刊,2008,25(19):2 - 9.
[9] 郭占锋,李小云. 对当前非洲农业研究的若干思考[J]. 农业经济,2012(3):17 - 19.
[10] 钟玲,王妍蕾,齐顾波. 坦桑尼亚的减贫历程及挑战[J]. 中国农业大学学报(社会科学版),2013,30(2):147 - 156.
[11] 胡国勇,路卓铭. 坦桑尼亚的贫困状况、减贫策略及其对我国的启示[J]. 社会科学家,2007,9(5):56 - 60.
[12] 王钰,施正香,黄仕伟. 发展家庭式奶牛养殖牧场的探讨[J]. 中国畜牧杂志,2015,51(12):38 - 43.

(上接第 353 页)

[3] ZHANG Y,LIN H,YANG Z,et al. Hash subgraph pairwise kernel for protein - protein interaction extraction[J]. IEEE/ACM transactions on computational biology and bioinformatics (TCBB),2012,9(4):1190 - 1202.
[4] SCHAEFER M H,LOPES T J S,MAH N,et al. Adding protein context to the human protein - protein interaction network to reveal meaningful interactions[J]. PLoS Comput Biol,2013,9(1):1002860.
[5] CHEN G,LI J,WANG J. Evaluation of gene ontology semantic similarities on protein interaction datasets[J]. International journal of bioinformatics research and applications,2013,9(2):173 - 183.
[6] XENARIOS I,SALW NSKI L,DUAN X J,et al. The data-base of interacting proteins:A research tool for studying cellular networks of protein interactions[J]. Nucleic acids research,2002,30(1):303 - 305.
[7] RESNIK P. Semantic similarity in a taxonomy:An information-based measure and its application to problems of ambiguity in natural language[J]. Journal of artificial intelligence research,1999,11(7):95 - 130.
[8] GUO Y,YU L,WEN Z,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.
[9] SHEN J,ZHANG J,LUO X,et al. Predicting protein-protein interactions based only on sequences information[J]. Proceedings of the national academy of sciences,2007,104(11):4337 - 4341.