Hadoop 框架下节点重要性算法实现蛋白质功能预测①

林志兴[1], 郭金文[2], 林 劼[2]

[1](三明学院 现代教育技术中心, 三明 365004)
[2](福建师范大学 软件学院, 福州 350001)

摘 要: 论文从蛋白质序列数据的角度出发, 通过序列相似度循环匹配构造蛋白质网络, 并且通过网络节点重要性排序算法预测蛋白质功能. 以节点重要性重要性作为研究对象, 在蛋白质网络应用节点重要性算法 PageRank 计算网络中蛋白质节点 PR 值, 在 Hadoop 平台上进行开发实现功能预测的并行计算, 减小运行时间. 最后通过准确率, 召回率以及 F1-measure 三个指标来衡量结果, 并对比传统的功能预测方法, 验证结果的有效性.

关键词: 蛋白质序列; 功能预测; 循环匹配; 节点重要性; Hadoop 平台

Predicting Protein Function Method with Node Importance Algorithm Based on Hadoop

LIN Zhi-Xing[1], GUO Jin-Wen[2], LIN Jie[2]

[1](Modern Educational Technology Center, Sanming College, Sanming 365004, China)
[2](Department of Software, Fujian Normal University, Fuzhou 350001, China)

Abstract: This paper starts from the perspective of protein sequence data, and constructs the protein network by cyclic sequence similarity matching. Then a novel method based on ranking the importance of network nodes is proposed. Considering the importance of protein nodes in the network, the node importance algorithm PageRank (PR) is used to compute the nodes' PR value. The proposed method is also developed on the Hadoop Platform, which makes it more suitable for huge genome database with great efficiency and parallel computing. Finally, comparing the traditional method of function prediction by the Accurate rate, Recall rate and F1-measure measurements, our method has been validated and the result shows that the method is feasible and valuable for practical usage.

Key words: protein sequence; function prediction; circular matching; node importance; Hadoop platform

二十世纪初人类基因组计划宣告完成, 结合计算机进行生物实验的技术得到了迅速的发展, 大量的生物数据不断涌现. 如何分析这些数据, 解析这些数据中蕴含的信息, 将这些看似庞杂的数据转变成对我们有用的知识, 是生物信息学领域致力解决的问题, 蛋白质功能的预测属于生物信息学一个重要的研究内容[1]. 以往通过传统的生物实验方法虽然能够准确测定蛋白质功能, 但是要耗费大量的人力物力且通过人工进行的实验效率低下. 因此, 通过计算机采用数据挖掘和智能计算技术对蛋白质功能进行预测, 对指导蛋白质功能研究具有重要现实和理论意义[2].

蛋白质功能预测是利用计算机技术, 从已有的蛋白质数据信息, 包括结构、序列、相互作用等数据出发, 综合运用各种数学算法和统计方法预测未知的蛋白质功能. 从序列出发预测蛋白质功能是一个重要的方法, 序列相似性比对作为蛋白质功能预测的一种重要方法, 其主要步骤是将蛋白质的氨基酸映射为对应字符, 并根据蛋白质氨基酸的排列顺序表示为一条以氨基酸字符作为基本元素的序列, 通过对比序列之间的相似性, 将相似序列已知蛋白质工作作为基础, 预测新测序列的蛋白质功能, 这类方法有 BLAST[3], FASTA[4], 以及 PSI-BLAST[5]等. 不过, 以上方法仅仅从序列的相似性比对进行研究, 忽略蛋白质之间的关系的全局结构, 以及关键蛋白质的重要作用, 因此准

① 基金项目:福建省自然科学基金(2014J01220);三明学院科研基金(B201201/G);福建省教育厅科技基金(JB13187,JA15463)
收稿时间:2015-08-25;收到修改稿时间:2015-10-26

确率较难提升, 而且效率低下. 近年来, 出现了应用蛋白质序列、基因表达和蛋白质相互作用等组合数据对蛋白质进行功能预测的方法[6-8], 这些方法需要对未知功能蛋白质进行大量实验, 获取大量数据的基础上才能够进行预测, 增加了蛋白质功能预测的成本.

蛋白质的循环排列就是将该蛋白质氨基酸序列的首尾相连, 从中分离出新的 N-和 C-端的一种排列, 这样的排列可以与另一个蛋白质氨基酸序列进行近似匹配(允许不匹配数量小于某个给定的阀值), 也称为蛋白质的全局循环排列. 从 1979 年开始, 就陆续发现了很多蛋白质全局循环排列的例子, 并发现这些循环排列对蛋白质的功能、结构具有非常重要的作用[9-10].

现有大部分研究集中在循环模式在蛋白质功能上的关系, 其中一些研究发现蛋白质全局循环模式能够在蛋白质的功能中提供重要帮助[11-13]. Weiner 等人提供了一个蛋白质全局循环模式清单[14], 并说明基于循环模式的重复和删除机制在蛋白质中的重要性. 另外, 一些研究也表明在蛋白质折叠和二级结构中, 蛋白质的全局循环模式一样能起到重要的作用[15-16].

本论文采用将蛋白质序列进行循环匹配相似度计算, 根据两个序列相似度判别相似的蛋白质, 通过构建蛋白质序列循环匹配的关联网络, 应用图论作为计算基础, 采用节点重要性算法计算网络中蛋白质的 PR 值并且通过 Hadoop 平台实现并行计算, 对蛋白质的功能进行预测.

1　相关知识

1.1　蛋白质序列循环匹配

已有的研究已经揭示, 氨基酸序列的循环排列模式在蛋白质的功能表达方面具有相当强的关联性, 在蛋白质的功能推断中具有重要的意义, 其不仅包含了已有的一般模式所能够识别的联系, 而且还能够通过循环模式发现蛋白质之间新的关系[17]. 蛋白质的循环排列就是将蛋白质序列首尾相连, 构成一个环形, 并进行循环转换, 从中分离出新的 N-和 C-端, 与另一个蛋白质氨基酸序列进行相似度匹配计算, 如果匹配程度小于预定阀值(允许不匹配数量小于某个给定的阀值), 就说明这两个蛋白质序列全局循环排列匹配. 这种循环模式描述参考图 1, 从 C-端开始比对的两条不相匹配的蛋白质序列, 将其中一条蛋白质序列进行循环转换, 使之匹配上另一条蛋白质序列[18].

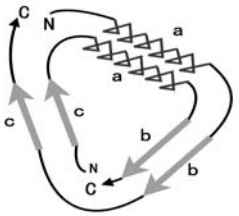


图 1　蛋白质循环匹配模式 (图改自文献[19])

论文中采用循环匹配的算法[20], 将蛋白质序列进行循环转换, 找出互相匹配的蛋白质, 并定义任意两条蛋白质序列 v_i 和 v_j, 其序列长度分别为 L_{v_i} 和 L_{v_j}, 其中 $L_{v_i} < L_{v_j}$, 则称两条蛋白质之间具有对应关系 M($v_i \to v_j$), 据此得到的一个有向的蛋白质匹配网络, 网络中节点的方向由蛋白质序列长的节点指向序列短的, 构建的网络示意图如图 2 所示, 圆圈越大的表示序列越长的蛋白质节点.

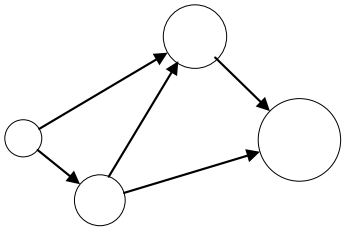


图 2　构建的蛋白质网络图

1.2　网络节点重要性算法

重要节点的定义是指存在网络中的那些能够在影响网络结构和功能的一些特殊的节点, 随着复杂网络知识这一研究领域的深入发展, 学者们根据所研究的具体问题提出了多种多样的节点重要性排序方法, 我们主要介绍 PageRank 算法.

PageRank 算法根据对网络节点的排序是基于两个假设: 1)数量假设: 如果连入目标节点的邻居节点数目越多则该节点越重要. 2)质量假设: 与目标节相连接的邻居节点质量越高, 则该节点越重要.

算法充分利用了这两个假设, 聚类开始时先赋予网络中每个网页节点相同的 PR 值一般赋值 1, 然后进行迭代, 每一步把每个节点当前的 PR 值平均分给它所指向的网络中的所有节点. 每个节点的新 PR 值为它所获得的 PR 值之和, 得到节点 vi 在 t 时刻的 PR 值用公式(1)计算.

$$PR_i(t) = \sum_{j \in B_j} \frac{PR_j(t-1)}{K_j^o} \quad (1)$$

其中, B_j 为连接到节点 j 的节点集合, K_j^o 为与连接到节点 v_i 的节点 v_j 的出度. 此公式一直迭代, 直到各个节点的 PR 值收敛.

但此公式存在一个问题, 如果网络中存在一个 0 出度的节点(通常称之为悬挂节点)时, 会不断吸收 PR 值, 使得 PR 值一直停留在此节点而无法传递出来. 因此, PageRank 算法在此基础上增加了一个阻尼系数, 即随机跳转概率 d, 在每一步迭代过程中, 网络的节点 PR 值都将以 d 的概率随机分给网络中的所有节点, 而以 $1-d$ 的概率均分给它所指向的节点. 通过这个步骤此有向网络将形成一个强连通的网络, 其邻接矩阵是一个不可约阵, 矩阵存在特征值 1, 修改后的公式如公式(2)表示.

$$PR_i(t)=\frac{d}{n}+(1-d)\sum_{j\in B_j}\frac{PR_j(t-1)}{K_j^o} \tag{2}$$

其中 n 为网络规模, 即节点个数, d 为阻尼系数, 其余参数均与公式 5.2 一致. 阻尼系数 d 通常设置为 0.15. Larry Page 和 Sergey Brin 两人从理论上证明了, 此算法是一个绝对可收敛的计算过程, 并且最终计算的节点 PR 值与初始值无关.

在实际计算过程中, PageRank 算法都是以矩阵运算的实现, 考虑到算法所需要迭代次数较多, 计算过程时间较长. 并且, 随着蛋白质数据库中的蛋白质数量的增加, 以及新增蛋白质数量的快速增长趋势, 传统的单机计算将面临计算资源不足的困境. 因此, 我们将此算法在Hadoop平台上实现, 通过并行运算的方式, 减少运行时间, 达到实际应用的目的.

1.3 Hadoop 平台

Hadoop 是一个设计为对大量数据进行分布式处理的分布式文件系统的数据处理框架. MapReduce 是 Hadoop 框架一个最核心的设计, 它为海量的数据提供了计算过程实现的可行性. MapReduce 程序包含 3 个主要部分: Map 函数部分、Reduce 函数部分和 Main 函数部分. Main 函数是分布式应用程序的入口, 主要处理作业控制和文件输入/输出. Map 函数、Reduce 函数是 MapReduce 程序的灵魂, 其中 Map 函数是一个作业归集并分配数据的过程, 接受一组数据并将其转换为一个键/值对列表并将此列表输出, 传给 Reduce 函数作为其输入. Reduce 函数根据 Map 函数生成的列表键值, 缩小键/值对列表, 进一步处理计算. 此过程如图 3 所示.

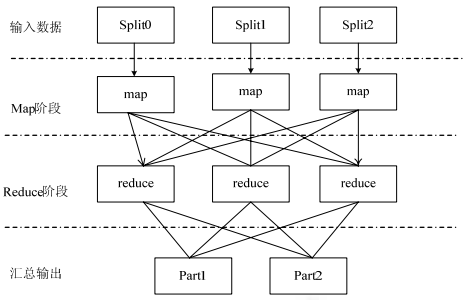


图 3 MapReduce 框架原理

2 基于MapReduce并行计算的功能预测

我们在 Eclipse 软件平台进行开发, 在 Hadoop 系统框架下实现并行运算 PageRank 算法, 预测蛋白质功能. 算法流程如图 4 所示.

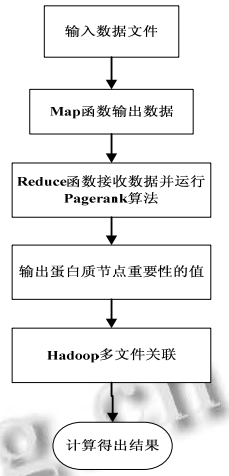
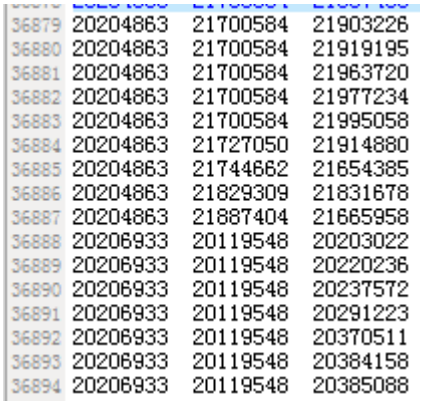


图 4 Hadoop 平台下实现 Pagerank 算法功能预测流程

预测过程分为如下几个步骤:

(1) 输入数据文件. 这里的输入数据文件包含 3 个文件, 第一个是蛋白质相互匹配信息文件, 是将蛋白质的序列通过循环匹配算法得出的匹配结果. 数据格式如图 5 所示, 这是文件随机截取的一部分(36879 行-36894 行), 主要包含三列, 第一列为待预测的未知功能蛋白质, 第二列和第三列均为已知的蛋白质, 同行的数据表示相互匹配. 第二个文件为蛋白质的 mapping 文件, 第三个为 unigo 文件, 其中 Mapping 文件的设计是为了使得在各大生物信息数据库中交互检索更方便. 因为蛋白质功能信息文件中注释使用的是蛋白质的 Name, 因此我们要先进行数据预处理, 通过蛋白质 AC 号和蛋白质 Name 关联文件, 将序列与功能

信息对应. Unigo 文件是蛋白质功能信息, 其中保存的是蛋白质以及其相关已知功能.



36879	20204863	21700584	21903226
36880	20204863	21700584	21919195
36881	20204863	21700584	21963720
36882	20204863	21700584	21977234
36883	20204863	21700584	21995058
36884	20204863	21727050	21914880
36885	20204863	21744662	21654385
36886	20204863	21829309	21831678
36887	20204863	21887404	21665958
36888	20206933	20119548	20203022
36889	20206933	20119548	20220236
36890	20206933	20119548	20237572
36891	20206933	20119548	20291223
36892	20206933	20119548	20370511
36893	20206933	20119548	20384158
36894	20206933	20119548	20385088

图 5　蛋白质相互匹配信息文件数据格式

(2) Map 函数处理. 通过 Mapping 读取蛋白质相互匹配信息文件, 以第一列待预测的蛋白质作为键值 Key, 其余两列作为 value 值, 将其输出, 作为 Reduce 的输入, 这样相同的待预测蛋白质的节点数据将被输入到同一个 Hadoop 计算机节点的 Reduce 中. 通过这样的处理, 该算法可以并行处理多个蛋白质功能预测任务.

(3) Reduce 函数处理. 这个步骤中, 我们将 Pagerank 算法在 Reduce 函数中实现, 在主函数设置实现函数的节点个数, 并设置合适的 hash 函数, 将对应的关联蛋白质通过 hash 函数进行分布, 相同的待预测蛋白质节点的 PCR 网络数据就可以被分配到不同的 Reduce 中, 实现 Hadoop 集群不同计算机节点分别并行计算不同的待预测蛋白质节点的 PCR 中蛋白质节点重要性值.

(4) 输出数据并进行文件关联. 通过第三步, 我们获得待预测的蛋白质 PCR 网络的已知蛋白质节点的 PR 值. 然后, 通过 Hadoop 编写的不同文件关联程序, 将已知的蛋白质这些 AC 号与 unigo 中的功能信息对应起来, 将蛋白质的 PR 值赋给其具有的功能.

(5) 计算获取结果. 将相同 PCR 网络中同个功能的 PR 值求和, 获得各个功能的最终 PR 值, 通过 z 值计算, 取不同 z 值情况下的满足 z 值设定的 PR, 作为预测结果, 并进行最后的比较得出结果, z 值采用公式(3)进行计算.

$$z = \frac{X - \bar{x}}{\sigma} \tag{3}$$

其中步骤 2,3 是实现并行运算的关键步骤, 也是 Hadoop 集群运行的主要步骤, 定义 MapReduce 框架的在 Map 和 Reduce 阶段输入和输出的 Key 和 Value 形成邻接矩阵并运行 PageRank 算法, 我们在表 1 中描述此过程.

表 1　MapReduce 过程

阶段	各个执行步骤信息

Map 阶段	输入: 蛋白质相互匹配信息文件, 三列数据为蛋白质 AC 号
	map 方法:
	String[]lineSplit=value.toString().split("　　"); //逐行读入数据
	word.set(lineSplit[0]); //设置 key 值
	v.set(lineSplit[1]+"　"+lineSplit[2]); //设置 Value 值
	context.write(word,v); //写入 context
	……
	输出<key, value>: (第一列待预测的蛋白质 AC 号, 第二列和第三列已知蛋白质数据 AC 号)

Reduce 阶段	输入为 Map 阶段的输出<key, value>
	reduce 方法:
	for (Text val : values) {
	//循环读取每个节点根据读取分配过来的 key 和对应的 Value 值
	}
	//value 值放入链表 mapList,存放的是已知蛋白的对应关系
	//初始化邻接矩阵 links[a][b]
	for　links[a][b]=0;
	//有对应关系的边在将邻接矩阵赋 1
	for　links[a][b] = 1;
	/*执行 PageRank 算法*/
	输出三列: (待预测功能的蛋白质, 与其相匹配的已知蛋白质, PR 值)

3　实验结果及分析

本文采用的功能数据来源于 UniProtKB-GOA 数据库, 数据可通过 http://www.ebi.ac.uk/GOA/进行下载, 最新的版本是 2015 年 9 月 16 日, 其中共有 1,500,000 条蛋白质序列及功能信息.

实验结果的衡量指标为: 准确率(Precision), 召回率(Recall)和 F1-measure. 准确率(Precision), 召回率

(Recall)和 F1-measure. 准确率主要是衡量被测量真值与预测结果之间一致的程度, 召回率用来衡量预测的结果包含了全部正确结果的程度, 准确率和召回率指标看似没有必然的关系, 甚至有的时候是矛盾的, 与信息检索领域中使用的评估指标一样, 采用一个综合度量指标 F1-Measure 来调和平均准确率和召回率. 使用三个指标之前, 我们首先要定义 3 个参数 TP, FP, TN, 其中 TP 为蛋白质数据库中指定蛋白质功能与在试验中预测出功能相符的功能个数, FP 为试验中预测出蛋白质功能但实际并不是指定蛋白质功能的个数, 这就是误判数量, TN 为蛋白质标准库中指定蛋白质功能没有被预测到的功能个数, 则 3 个评价指标定义如下:

$$\Pr ecesion = \frac{TP}{TP + FP}$$

$$\mathrm{Re}call = \frac{TP}{TP + TN}$$

$$F1 - measure = \frac{2 \times \Pr ecision \times \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call}$$

实验采用 10 倍交叉验证方法, 运行在 Hadoop 集群为 10 台 PC 机上, 我们设置的 Map 和 Reduce 的节点个数为 10 个, 对比交叉验证的 10 组数据, 在集群下与单机下分别实现交叉验证的 10 组数据节点重要性排序算法, 计算运行的时间, 结果如表 2 所示.

表 2 单机和集群下运行时间对照

组别编号	PC 单机(min)	Hadoop 集群(min)
1	1586	160
2	1549	158
3	1680	172
4	1662	169
5	1535	157
6	1640	168
7	1534	156
8	1678	170
9	1602	163
10	1642	168

理论上分析, 我们设置 Hadoop 的计算节点个数为 10 个, 则集群的运行时间理论上应该是单机的十分之一, 但是由于集群中的计算机节点在数据处理过程中涉及到数据的交互传递耗时, 因此运行的时间为单机状态下运行时间的十分之一还多点, 但是这已经明显提高了计算运行的效率. 预测结果如图 6 所示.

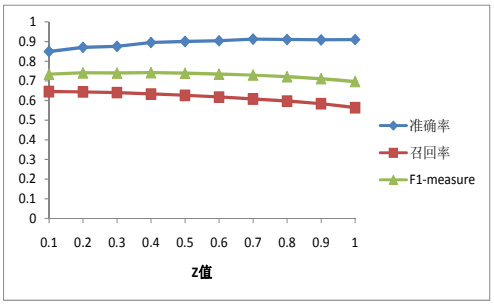


图 6 PageRank 算法预测功能结果

从结果我们可以看出通过 PageRank 计算后进行的功能预测, 能够取得较好的预测结果, 整体上准确率在 0.85 以上, F1-measure 维持在 0.7 左右, 在 z 值取 0.4 时, F1-measure 取得最好的 0.7417. 我们取 z=0.4 时得到的结果对比其他方法, 如表 3 所示, 其中 PPI 网络的聚类 DCS 预测结果来自文献[21], BLAST 直接注释方法取自文献[22].

表 3 功能预测结果对比

方法	准确率	召回率	F1-measure
BLAST 直接注释	0.54	0.62	0.58
PPI 网络的聚类 DCS	0.71	0.62	0.66
本文的方法	0.88	0.64	0.74

3 结语

本文从蛋白质序列研究入手, 将蛋白质作为网络节点, 通过序列循环匹配算法寻找循环相似度高的蛋白质作为边来构建蛋白质循环匹配网络, 然后应用节点重要性排序算法, 并将其运行在 Hadoop 集群下实现蛋白质功能预测. 对比以往传统的蛋白质功能预测方法几乎没有涉及到结合采用蛋白质节点关键性作为功能预测的手段, 本文的方法在预测结果具有一定的提高, 并且结合 Hadoop 集群环境运行实现并行计算, 大大减小了运行的时间. 论文提出的采用节点重要性排序算法进行功能预测的方法, 所使用的算法是 Pagerank 原始算法, 还未对此算法进行优化, 因此我们可以看出其在运行效率还是比较低的, 因此可以考虑将 PageRank 算法改进或利用一些其他更可靠的算法, 这是今后有待继续深入研究的内容.

参考文献

1 Hawkins T, Chitale M, Luban S, et al. PFP: Automated prediction of gene ontology functional annotations with confidence scores

using protein sequence data. Proteins-structure Function & Bioinformatics, 2009, 74(3): 566−582.

2 孙啸,陆祖宏,谢建明.生物信息学基础.北京:清华大学出版社,2005:15−53.

3 Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. Journal of Molecular Biology, 1990, 215(3): 403−410.

4 Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods in Enzymology, 1990, 183(1): 63−98.

5 Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 1997, 25(8): 3389−3402.

6 陈义明,贺细平,乔波.一种基于树的蛋白质功能预测算法:KDE–CSSA.湖南农业大学学报(自然科学版),2015,1: 62−66.

7 孟军,张信.基于双重索引矩阵的蛋白质功能预测.计算机应用,2015,6:1637−1642.

8 罗纪文.基于二阶马尔可夫随机场的蛋白质功能预测.科技信息, 2014.

9 Bujnicki J. Sequence permutations in the molecular evolution of dna methyl transferases. BMC Evolutionary Biology, 2: 3, 2002.

10 Cunningham BA, Hemperly JJ, Hopp TP, Edelman GM. Favin versus concanavalin A: Circularly permuted amino acid sequences. Proc. Natl. Acad. Sci. USA, 1979, 76(7): 3218–3222.

11 Daly NL, Craik DJ. Acyclic permutants of naturally occurring cyclic proteins. The Journal of Biological Chemistry, 2000, 275(25): 3218–22.

12 Craik DJ. Seamless proteins tie up their loose ends. Science, 2006, 311: 1563–1564.

13 Craik DJ. Circling the enemy: Cyclic proteins in plant defence. Trends Plant Sci., 2009, 14(6): 328–335.

14 Weiner J, Bornberg-Bauer E. Evolution of circular permutations in multidomain proteins. Mol. Biol. Evol, 2006, 23(4): 734–743.

15 Lindberg M, Tangrot J, Oliveberg M. Complete change of the protein folding transition state upon circular permutation. Nature Struct. Biol., 2002, 9: 818–822.

16 Haglund E, Lindberg MO, Oliveberg M. Changes of protein folding pathways by circular permutation. overlapping nuclei promote global cooperativity. J Biol Chem., 2008, 283(41): 27904–27915.

17 Cunningham BA, Hemperly JJ, Hopp TP, et al. Favin versus concanavalin A: Circularly permuted amino acid sequences. Proc. of the National Academy of Sciences of the United States of America, 1979, 76(7): 3218−3222.

18 Jeltsch A. Circular permutations in the molecular evolution of DNA methyl transferases. Journal of Molecular Evolution, 1999, 49(1): 161−164.

19 Bliven S, Prlic A. Circular permutation in proteins. Plos Computational Biology, 2012, 8: e1002445.

20 Lin J, Adjeroh D. All-against-all circular pattern matching. Computer Journal, 2012, 55(7): 897−906.

21 Wei P, Wang J, Cai J, et al. Improving protein function prediction using domain and protein complexes in PPI networks. Bmc Systems Biology, 2014, 8(3): 1260−1260.

22 Tatusova TA, Madden TL. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. Fems Microbiology Letters, 1999, 174(2): 247−250.