Journal of Computer Applications
计算机应用,2015,35(6):1637 - 1642
ISSN 1001-9081
CODEN JYIIDU
2015-06-10
http://www.joca.cn

文章编号:1001-9081(2015)06-1637-06
doi:10.11772/j.issn.1001-9081.2015.06.1637

基于双重索引矩阵的蛋白质功能预测

孟　军*,张　信
(大连理工大学 计算机科学与技术学院,辽宁 大连 116024)
(*通信作者电子邮箱 mengjun@dlut.edu.cn)

摘　要:针对单一数据源预测蛋白质功能效果不佳以及蛋白质相互作用网络信息不完全等问题,提出一种多数据源融合和基于双重索引矩阵的随机游走的蛋白质功能预测(MSI-RWDIM)算法。该算法使用了蛋白质序列、基因表达和蛋白质相互作用数据预测蛋白质功能,并根据这些数据源特性构建相应的相互作用加权网络;然后融合各数据源加权网络并结合功能相关性网络构建双重索引矩阵,使用随机游走算法计算得分进而预测蛋白质功能。在酵母数据集的五折交叉验证中,MSI-RWDIM 算法具有较高的准确率和较低的覆盖率,还可降低功能标签损失率。研究结果表明,MSI-RWDIM 算法的总体性能优于常用的 k- 近邻、直推式多标签集成分类和快速同步加权方法。

关键词:多数据源融合;随机游走;双重索引矩阵;功能相关性网络;蛋白质功能预测
中图分类号:TP181　　**文献标志码**:A

Protein function prediction based on doubly indexed matrix

MENG Jun*, ZHANG Xin
(*School of Computer Science and Technology, Dalian University of Technology, Dalian Liaoning* 116024, *China*)

Abstract: The single data source cannot effectively predict the function of protein and the information of protein interaction network is incomplete. In order to solve the problem, A Multi-Source Integration and Random Walk with Doubly Indexed Matrix (MSI-RWDIM) algorithm was proposed. The proposed algorithm used protein sequence, gene expression and protein-protein interaction for the prediction of protein function. The weighting networks were constructed from the data sources with their characteristics. A network, which was fused by the weighting networks, integrated with function correlation network to construct a doubly indexed matrix. Random walk was used to calculate annotation scores and predict protein function. The cross-validation experiments on Yeast show that MSI-RWDIM can achieve higher prediction accuracy, lower coverage and lower loss rate of function labels. The research results show that the overall performance of MSI-RWDIM is much better than commonly used k-nearest neighbor, transductive multi-label ensemble classifier and fast simultaneous weighting method.

Key words: multiple data integration; random walk; doubly indexed matrix; function correlation network; protein function prediction

0　引言

随着高通量测序技术的发展,产生了大量基因组和蛋白质组学数据。近年来,利用蛋白质序列数据[1]、基因表达数据[2]和蛋白质相互作用数据[3]等单一数据源预测蛋白质功能成为研究热点。然而,单一数据源在表现功能时具有不完整性以及数据的高噪声等特点,导致预测效果不佳,因此利用计算方法有效地集成这些多源异构数据成为蛋白质功能预测研究的新挑战。多数据簇集成的基因注释方法(Multisource Association of Genes by Integration of Clusters, MAGIC)[4]使用贝叶斯网络模型集成多源异构高通量生物数据并推断蛋白质间的功能联系来预测蛋白质功能;基于特征融合的分类模型[5]通过拼接蛋白质各数据源对应蛋白的特征向量构建新的特征向量,训练支持向量机分类模型;多源 k- 近邻算法(Multi-Source k- Nearest Neighbor algorithm, MS-kNN)[6]是集成多数据源预测结果的功能预测方法,利用相似性度量求出蛋白质在各数据源中的 k 最近邻蛋白,根据在各数据源中的近邻蛋白求出注释得分,最后集成所有注释得分。考虑到蛋白质特征维度大以及相互作用网络的高维稀疏性的特点,这些方法在构建分类器模型时只简单地考虑特征融合,未能分析各数据特有功能信息,或者只是简单考虑邻居蛋白功能注释信息,并没有考虑网络拓扑结构。

标签传播方法不仅考虑近邻蛋白质,同时还考虑网络全局拓扑特性,充分利用蛋白质数据网络特性预测蛋白质功能。可以利用 Jaccard 系数衡量功能之间相关性,并将这种关系融入正规化的半监督学习框架[7];双关系图的功能预测(Protein function prediction based on Bi-relational Graph, PfunBG)算法[8]利用功能相关性扩展蛋白质相互作用,使用网络传播衡量蛋白质与功能邻近性;基于松弛标记和功能相关性结合的方法是通过功能相关性影响松弛标记迭代过程,从而预测蛋白质的功能[9]。上述方法利用了功能相关性网络的半监督学习框架,但都是基于单一数据源的预测模型,并且如何有效地结合相互作用网络和功能相关性方法都有待进一步研究。

从不同数据源可以得到蛋白质的不同表征特性,结合这

收稿日期:2015-01-13;**修回日期**:2015-04-03。　**基金项目**:国家自然科学基金资助项目(61472061)。
作者简介:孟军(1964 -),女,辽宁大连人,副教授,博士,CCF 会员,主要研究方向:机器学习、数据挖掘;　张信(1990 -),男,安徽安庆人,硕士研究生,主要研究方向:机器学习、数据挖掘。

些异构数据可以从多角度分析蛋白质功能[6]。蛋白质功能与蛋白质结构密切相关,蛋白质结构依赖于蛋白质序列,可从蛋白质序列信息预测蛋白质功能。共表达基因更可能是功能相关的,并影响生物功能,基因表达数据能反映共表达基因表达过程,可用于构建基于表达数据的相互作用网络预测功能。蛋白质并非单独完成生物功能,同时一个特定生物功能一般由一组蛋白质完成,基于这种特性的蛋白质相互作用可直接用来预测蛋白质功能。本文融合蛋白质序列、基因表达和蛋白质相互作用三类数据来构建相互作用加权网络,提出了多数据融合和基于双重索引矩阵的随机游走(Multi-Source Integration and Random Walk based on Doubly Indexed Matrix, MSI-RWDIM)的蛋白质功能预测算法。该方法根据蛋白质数据特性分别构建网络,将异构网络融合成单个加权网络,与功能相关性网络构建双重索引矩阵,使用随机游走算法得到蛋白质功能注释得分,从而预测蛋白质的功能。

1 相关工作

1.1 融合多数据源构建蛋白质网络

不同蛋白质数据在表征蛋白质功能方面具有不同形式,如何有效地构建相应蛋白质相互作用网络来表现蛋白质功能特征信息,对于准确地预测功能具有重要影响。本文根据各种蛋白质数据的本身特性,选用相应方法构建蛋白质加权网络。

1.1.1 蛋白质序列数据

蛋白质由 20 种氨基酸排列组成,氨基酸的排列信息与蛋白质的结构信息和功能信息密切相关。考虑到部分功能相关蛋白为低同源性或可能存在不与任何蛋白质具有同源性的孤立蛋白,采取不同于以往的基于同源性的序列匹配方法,本文使用特征提取策略计算蛋白质间相似性来构建相互作用网络。假如蛋白质集合 $P = \{P_1, P_2, \cdots, P_n\}$ 和相应的功能标签集合 $C = \{C_1, C_2, \cdots, C_m\}$,则对蛋白质 P_i 利用伪氨基酸组成(Pseudo Amino Acid Composition, PseAAC)[10] 方法提取 60 维蛋白质特征向量 $\boldsymbol{Pro}_i = (p_1, p_2, \cdots, p_{60})$,蛋白质间相似性度量如下:

$$d(P_u, P_v) = 1 - \frac{\boldsymbol{Pro}_u \cdot \boldsymbol{Pro}_v}{\| \boldsymbol{Pro}_u \| \| \boldsymbol{Pro}_v \|} \tag{1}$$

其中:$\boldsymbol{Pro}_u$ 和 $\boldsymbol{Pro}_v$ 分别为蛋白质 P_u 和 P_v 的特征向量。d 的取值范围为[0, 1],其值越小,蛋白质间序列相似性越大;值越大时,序列相似性越小。为保证矩阵的一致性,关联矩阵 $\boldsymbol{SEQ}$ 的计算方法如下:

$$\boldsymbol{SEQ}_{P_u, P_v} = \exp\left(\frac{-d(P_u, P_v)}{2\sigma^2}\right) \tag{2}$$

利用 $\boldsymbol{SEQ}^{\text{std}} = \boldsymbol{D}^{-1/2} \cdot \boldsymbol{SEQ} \cdot \boldsymbol{D}^{-1/2}$ 对构建的网络进行标准化处理,其中:$\boldsymbol{D}$ 为矩阵 $\boldsymbol{SEQ}$ 每行所有值的和构成的对角矩阵,$\boldsymbol{SEQ}^{\text{std}}$ 为标准化之后的矩阵。对于后面将要阐述的基因表达数据、蛋白质相互作用的有权网络以及融合后的网络都使用同样的标准化方法。

该方法构建的网络利用蛋白质序列间相似性程度来衡量近邻权重,但这种网络是稠密网络,其中很多相似边都具有很小的权值。蛋白质功能预测的效率很大程度上依赖于网络中权值非零边的数目,为此可将稠密网络调整为稀疏网络,即选取 k- 近邻蛋白质作为邻居蛋白构建网络。

1.1.2 基因表达数据

基因表达数据体现了基因在不同条件下的活动信息,通过其基因表达的改变来反映蛋白质当前的生命过程。共表达基因一般具有功能相关性,并在一定条件下影响生物过程。本文引入皮尔逊相关系数(Pearson Correlation Coefficient, PCC)将基因表达数据构建成共表达网络,并使用 $\boldsymbol{PCC}$ 来衡量共表达强弱程度,其定义如下:

$$\boldsymbol{PCC}_{P_u, P_v} = \frac{\sum_{i=1}^{m} (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{m} (u_i - \bar{u})^2 \sum_{i=1}^{m} (v_i - \bar{v})^2}} \tag{3}$$

其中:u_i 和 v_i 表示蛋白质 P_u 和 P_v 在某一条件下的表达值,$\bar{u}$ 和 $\bar{v}$ 表示蛋白质 P_u 和 P_v 的所有表达值的均值。m 为表达数据的条件维度。$\boldsymbol{PCC}$ 的取值范围为[-1, 1],其值小于 0 时,说明蛋白质之间为负相关;大于 0 时,表现为正相关;等于 0 时,说明不存在相关性。构建的网络同样采用 k- 近邻方法进行稀疏化。标准化后得到基因表达数据网络 $\boldsymbol{PCC}^{\text{std}}$。

1.1.3 蛋白质相互作用数据

为体现无向的无权蛋白质相互作用网络中相互作用的程度,需要衡量网络中边的权重。可以用边聚集系数(Edge Clustering Coefficient, ECC)[11] 来描述一个节点在网络中的属性,其已广泛应用在分析蛋白质相互作用网络的拓扑特性等领域。$\boldsymbol{ECC}$ 是一个局部变量,能有效地表述某条边上蛋白质之间的亲疏程度和重要性,可用来评估蛋白质相互作用网络边的权重。给定网络中边(P_u, P_v),其边聚类系数 $\boldsymbol{ECC}$ 定义如下:

$$\boldsymbol{ECC}_{P_u, P_v} = \frac{z_{u,v}}{\min(dim_u - 1, dim_v - 1)} \tag{4}$$

其中:$Z_{u,v}$ 为网络中边(P_u, P_v) 构成的三角形数量,dim_u 和 dim_v 分别为点 P_u 和 P_v 的在网络中的度。其值越大,表明两个节点之间相关性越大。标准化后得到蛋白质相互作用加权网络 $\boldsymbol{ECC}^{\text{std}}$。

1.1.4 多数据源融合

利用不同源蛋白质数据构建多种异构相互作用网络,由于这些网络基于不同性质和来源,各数据源间相互独立,经过标准化处理,可使用朴素贝叶斯模型[12] 将这些异构网络融合成单个加权网络。当某条边由多个网络共同拥有时,表明此边在网络中具有更高的可信性,合并后权值一般大于单个网络对应的权值。朴素贝叶斯模型公式如下所示:

$$\boldsymbol{W}_{i,j}^{\text{com}} = 1 - \prod_{r} (1 - \boldsymbol{W}_{i,j}^{r}) \tag{5}$$

其中:$\boldsymbol{W}_{i,j}^{r}$ 代表在三个数据源网络 $\boldsymbol{SEQ}^{\text{std}}$、$\boldsymbol{PCC}^{\text{std}}$ 和 $\boldsymbol{ECC}^{\text{std}}$ 中边(i, j) 的权值,$\boldsymbol{W}_{i,j}^{\text{com}}$ 为合并后边(i, j) 的权值。本文使用基于不同性质和来源的蛋白质数据,符合此方法要求的独立性假设。通过此方法可将三个数据源网络合并为相互作用网络 $\boldsymbol{G}^{\text{com}}$。当网络数目增多时,寻找一个广泛适用的方法能快速可扩展地融合这些网络仍然是一个挑战;并且多个网络合并可能会覆盖某个网络对特定类别功能预测优势。

1.2 功能相关性网络构建

功能相关性网络可以分为层次相关性和非层次相关性两

类[7]。层次相关性根据基因本体(Gene Ontology,GO)结构图中父子功能和祖孙功能结构关系来衡量功能间亲疏程度,这种方法的局限性在于它只考虑了 GO 术语的层次结构。观察发现不同 GO 术语经常共同注释一些常见蛋白质[7],可根据功能间共同注释蛋白信息计算相关性,即非层次相关性。为此,本文使用基于共同注释蛋白质的功能相似性度量来衡量功能相关程度。针对标签 f_1 和 f_2,其功能相关性定义如下:

$$\boldsymbol{Corr}(f_1, f_2) = |Q_{f_1} \cap Q_{f_2}| / |Q_{f_1} \cup Q_{f_2}| \tag{6}$$

其中:Q_{f_1} 和 Q_{f_2} 为训练蛋白质中注释了功能标签 f_1 和 f_2 蛋白质集合。$\boldsymbol{Corr}(f_1, f_2)$ 的取值范围为[0, 1],其值越接近 1,说明功能标签共同注释某些蛋白可能性越大。利用此方法构建蛋白质功能相关性网络 $\boldsymbol{G}^{\text{corr}}$。

2　基于双重索引矩阵的随机游走

通过前面的分析可得到基于 N 种蛋白质间的相互作用网络 $\boldsymbol{G}^{\text{com}}(V_1, E_1)$ 和 M 种标签间的蛋白质功能相关性网络 $\boldsymbol{G}^{\text{com}}(V_2, E_2)$,其中:$V_1$ 代表蛋白质节点集合,V_2 表示功能节点集合,E_1 代表相互作用网络中边的集合,E_2 代表功能相关性网络中边的集合。不同于以往随机游走算法应用于相互作用网络,本文使用相互作用网络与功能相关性网络结合的方法。首先构建一个双重索引矩阵 $\boldsymbol{A}$,其权值来自网络 $\boldsymbol{G}^{\text{com}}$ 和 $\boldsymbol{G}^{\text{corr}}$。$\boldsymbol{A}$ 是 $|N| \times |M| \times |N| \times |M|$ 维矩阵,其元素定义如下:

$$\boldsymbol{A}_{ij,uv} = \begin{cases} \boldsymbol{G}^{\text{com}}_{i,u}\boldsymbol{G}^{\text{corr}}_{j,v}, & u \in Nei(i),\ v \in Nei(j) \\ 0, & \text{其他} \end{cases} \tag{7}$$

其中:i 和 j 分别为蛋白质节点和功能节点;$Nei(i)$ 为节点 i 在相互作用网络中的邻居集合;$Nei(j)$ 为节点 j 在功能相关性网络中的邻居集合;$\boldsymbol{G}^{\text{com}}_{i,u}$ 和 $\boldsymbol{G}^{\text{corr}}_{j,v}$ 分别来自网络 $\boldsymbol{G}^{\text{com}}$ 和 $\boldsymbol{G}^{\text{corr}}$ 对应节点间的权值。

基于矩阵 $\boldsymbol{A}$,应用随机游走算法,迭代收敛后得到蛋白质功能注释得分。注释得分 $\boldsymbol{S}_{i,j}$ 计算公式如下:

$$\boldsymbol{S}_{i,j} = \alpha \sum_{u \in Nei(i)} \sum_{v \in Nei(j)} \boldsymbol{G}^{\text{com}}_{i,u}\boldsymbol{G}^{\text{corr}}_{j,v}\boldsymbol{S}_{u,v} + (1-\alpha)\boldsymbol{Y}_{i,j} \tag{8}$$

其中:节点 i 来自相互作用网络 $\boldsymbol{G}^{\text{com}}$,节点 j 来自功能相关性网络 $\boldsymbol{G}^{\text{corr}}$;$\boldsymbol{Y}_{i,j}$ 为节点 i 和 j 先验知识得分;α 为控制随机游走的过程参数。考虑整个网络以及迭代过程[13],其公式重写如下:

$$\boldsymbol{S}^t = \alpha\boldsymbol{A}\boldsymbol{S}^{t-1} + (1-\alpha)\boldsymbol{Y} \tag{9}$$

MSI-RWDIM 算法具体实现步骤为:

输入　蛋白质序列、基因表达和蛋白质相互作用数据;蛋白质功能注释集合 C,参数 α 和 k。

输出　蛋白质预测注释得分矩阵 $\boldsymbol{S}$。

步骤 1　采用三种数据源所共有的 N 种蛋白质,构建基于三种数据源的异构有权网络,进行标准化处理($\boldsymbol{SEQ}^{\text{std}}$,$\boldsymbol{PCC}^{\text{std}}$ 和 $\boldsymbol{ECC}^{\text{std}}$),提取并初始化 M 种标签的标签矩阵 $\boldsymbol{Y}$;

步骤 2　根据式(5)合并多源异构网络得到网络 $\boldsymbol{G}^{\text{com}}$;

步骤 3　根据式(6)计算功能相关性得到网络 $\boldsymbol{G}^{\text{corr}}$;

步骤 4　由矩阵标准化后的 $\boldsymbol{G}^{\text{com}}$ 和 $\boldsymbol{G}^{\text{corr}}$ 得到双重索引矩阵 $\boldsymbol{A}$;

步骤 5　将矩阵 $\boldsymbol{A}$、α 代入式(9),迭代直到收敛;

步骤 6　最后得到注释矩阵 $\boldsymbol{S}$ 为 $\boldsymbol{S}^t$ 收敛结果,将 top n 得分值对应的功能标签赋予未注释蛋白质。

3　实验结果分析

3.1　实验数据集

本文使用酵母的蛋白质序列、基因表达和蛋白质相互作用三种数据。其中,蛋白质相互作用数据来自于 BioGRID(http://thebiogrid.org/)[14],提取了包含 5 579 种蛋白质的 81 031 条蛋白质相互作用对。基因表达数据来自于 SGD(http://www.yeastgenome.org/)[15] 数据库,此数据集包含 6 307 种蛋白质在 215 种条件下的表达值。蛋白质序列从 MIPS(http://mips.helmholtz-muenchen.de/)[16] 下载,包含 6 717 条蛋白质序列信息。通过分析三种数据,提取所有数据源共同拥有的 4 526 种蛋白质,构建其相互作用网络。本实验使用 GO 术语注释蛋白质功能,并从 SGD 数据库中提取到生物过程(Biological Process,BP)术语和分子功能(Molecular Function,MF)术语。考虑到标签集中包含一些未被证实的功能注释,剔除了 GO 中有电子注释推断(Inferred from Electronic Annotation,IEA)的标签,同时选取蛋白质标签数量在 3 到 300 之间的标签,得到 1 034 个 BP 术语和 428 个 MF 术语。

3.2　算法的性能分析

为了评估 MSI-RWDIM 算法的有效性和可行性,采用平均查准率(Average precision)、1 - 排名损失率(1 - RankingLoss)和覆盖度(Coverage)指标[17],同时引入反映分类敏感性和特异性的接受者操作特征(Receiver Operating Characteristic,ROC)曲线下面积值(Area Under of the roc Curve,AUC)指标来衡量其预测性能。

Average precision 表示平均查准率,评估了排序得分向量中排在真实标签前面的标签中也是真实标签的概率。

$$Avegprec = \frac{1}{q}\sum_{i=1}^{q}\frac{1}{|L_i|}\sum_{\lambda \in L_i}\frac{|\lambda' \in L_i \,|\, r(i,\lambda') \leqslant r(i,\lambda)|}{r(i,\lambda)} \tag{10}$$

其中:q 为测试集数量;L_i 为待预测蛋白质 P_i 标签集合;$r(i, \lambda)$ 为标签 λ 在蛋白质预测标签排序序列中的位置。Average precision 取值范围为[0, 1],其值越大,性能越好。

RankingLoss 表示预测到标签排序错误的类标签对的数目,其值为 0 时,性能最好。

$$RL = \frac{1}{q}\sum_{i=1}^{q}\frac{|\{(\lambda_a, \lambda_b) \in L_i \times \overline{L}_i \,|\, r(i,\lambda_a) > r(i,\lambda_b)\}|}{|L_i|\,|\overline{L}_i|} \tag{11}$$

其中 $\overline{L}_i$ 为 L_i 的补集,表示蛋白质没有的标签集。为了与前面指标保持一致性,此处使用 1 - RankingLoss,其值越大,性能越好。

Coverage 表示对于预测到的标签排序结果,排在最前面的多少个标签可以覆盖实例所有真实标签,其值越小,性能越好。

$$Cov = \frac{1}{q}\sum_{i=1}^{q}\max_{\gamma \in L_i} r(i,\lambda) - 1 \tag{12}$$

AUC 是用来度量分类模型好坏的指标,通过 ROC 曲线分析其值大小,而 ROC 曲线通过真阳性率(True Positive Rate,TPR)和假阳性率(False Positive Rate,FPR)值对绘制而成。如果一个实例被正确预测为正类,即为真正类(True Positive,

TP)；如果实例是被错误预测为负类的正类，即为假负类(False Negative，FN)；如果实例是被错误预测为正类的负类，称之为假正类(False Positive，FP)；负类被预测为负类，称之为真负类(True Negative，TN)。AUC 取值范围为[0，1]，取值越大，效果越好。为适应多标签分类问题，使用适应多标签学习的 AUC[18]。

$$\begin{cases} TPR = TP/(TP + FN) \\ FPR = FP/(TN + FP) \end{cases} \tag{13}$$

在酵母数据集上采用五折交叉验证(5-Fold Cross Validation，5-CV)方法分析性能。MSI-RWDIM 与三种常用的集成预测方法 MS-kNN、直推式多标签集成分类(Transductive Multi-label Ensemble Classifier，TMEC)[18]和快速同步加权方法(fast Simultaneous Weight method，SW)[19]进行对比实验。其中：MS-kNN 分别对蛋白质序列数据、基因表达数据和蛋白质相互作用三种数据源采用 k-近邻方法计算功能注释得分，然后融合所有功能注释得分；TMEC 扩展相互作用网络为蛋白质和功能相关的有向双关系图，将其应用于随机游走算法；SW 通过在融合网络方面进行改进，提出基于单约束线性回归问题的权重优化方案。同时，为了说明双重索引矩阵的高效性，与基于相互作用矩阵的随机游走(Multi-Source Integration and Random Walk based on interaction network，MSI-RW)方法进行比较。

针对 BP 术语，表 1 显示 MSI-RWDIM 可以得到较高的 Average precision，相对于其他三种方法分别提高了将近 6%、9% 和 15%。对于 1 - RankingLoss，除与 TMEC 有轻微差距，比其他方法都有一定程度提高。同时，MSI-RWDIM 在 Coverage 和 AUC 上都优于其他数据集成方法。MSI-RWDIM 的各项指标比 MSI-RW 有明显提高，因此基于双重索引矩阵的方法效果更好。各种方法的 ROC 曲线如图 1 所示。针对 MF 术语，从表 2 可知，Average precision 和 Coverage 上相对于其他方法都有很大提高，而在 1 - RankingLoss(1 - RL)和 AUC 上，除与 TMEC 基本持平外，都有一定程度的优势。

表 1 酵母数据集在 BP 术语上各项指标性能

方法	性能指标			
	Avegprec	1 - *RL*	*Cov*	*AUC*
MSI-RWDIM	0.4108	0.9020	119.8067	0.9041
MSI-RW	0.3903	0.8820	151.1644	0.8851
TMEC	0.3487	0.9088	159.5262	0.8801
SW	0.3295	0.8490	181.9639	0.8553
MS-kNN	0.2629	0.6895	200.2804	0.8095

表 2 酵母数据集在 MF 术语上各项指标性能

方法	性能指标			
	Avegprec	1 - *RL*	*Cov*	*AUC*
MSI-RWDIM	0.4486	0.9079	35.2348	0.9175
MSI-RW	0.4209	0.8929	40.5727	0.8997
TMEC	0.4046	0.9188	44.6894	0.9191
SW	0.2457	0.8414	57.7789	0.8505
MS-kNN	0.3609	0.7258	60.2510	0.8173

综上所述，MSI-RWDIM 方法能有效地利用各数据源的互补性和功能相关性信息预测蛋白质功能，在预测效果上较其他算法有了一定提高。

本文方法性能提高的主要原因：1)随机游走算法通过模拟粒子随机游走过程，能够利用网络全局拓扑特性，而在考虑局部特性时，算法每一次随机游走都会以一定概率返回出发点，最后，经过多次游走过程达到稳定状态；2)考虑到一些功能标签同时注释相同蛋白质，引进双重索引矩阵，将功能相关性网络与相互作用网络融合，能有效提高预测效果；3)三种数据集的有效融合，避免了单数据源在表现功能方面的片面性。

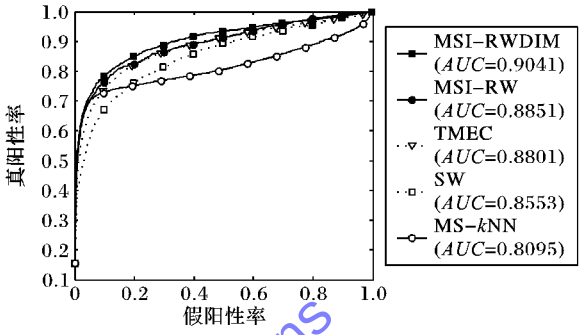

图 1 五种方法在 BP 术语上的 ROC 曲线

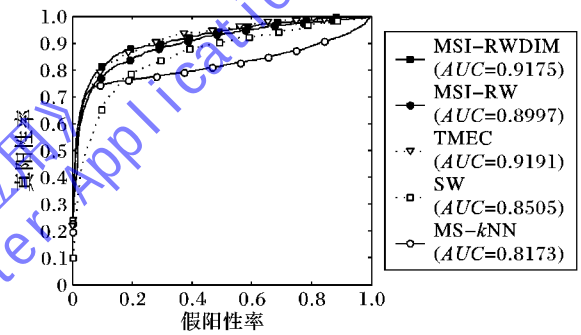

图 2 五种方法在 MF 术语上的 ROC 曲线

3.3 参数选择

MSI-RWDIM 算法主要有 k 和 α 两个参数。α 控制随机游走过程中返回起始点的概率，它限制了粒子离开起始点的距离，权衡了局部和全局拓扑特性。为了有效选取 α 值，测试了其在 10 个取值点(0.05，0.15，…，0.95)上的 *AUC* 值。由图 3 可知，当 $\alpha = 0.15$ 时，两种标签集在 *AUC* 上都能取得最好的预测结果。从 α 取值可以看出，网络的局部拓扑对于功能预测具有更高的贡献度，这也符合近邻蛋白质具有较强的功能相关性特征。k 为构建稀疏网络时近邻蛋白数目。如果 k 值太大，很多相似度小的无关蛋白也作为近邻蛋白，会使预测结果产生偏差；如果 k 值太小，又会忽略可能存在的近邻蛋白。根据以往经验和实验取 k 为 100 时稀疏矩阵性能不会降低[19]。

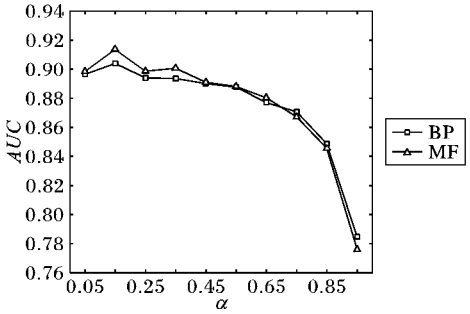

图 3 参数 α 在不同取值时的 *AUC* 值

3.4 预测结果分析

用酵母数据集对预测得到的蛋白质 GO 术语概率得分排

序,将得分值排在前面的功能标签与文献报道以及 SGD 数据库功能注释信息进行比较分析。表 3 显示了 9 种蛋白质在 BP 术语和 MF 术语下的预测功能、文献报道功能以及预测到该标签得分的排名情况。

最新报道显示,YGR283C 和 YMR310C 具有甲基转移酶活性[20];而从 MSI-RWDIM 方法的预测结果显示得分最高位置就预测到转移酶活性和甲基化作用。YAL053W 编码的蛋白质参与到细胞壁组织活动以及 FAD 导入活性[21];而从 SGD 数据库查询到信息,发现其具有 FAD 跨膜运输活性,与实际预测结果相符。YBR280C 参与目标蛋白前体白酶体降解过程以及具有泛素蛋白转移酶活性[22-23];本文方法预测结果显示,该蛋白质具有 GO:0031146 和 GO:0004824 功能标签,其预测结果与 GO 术语含义实际功能基本吻合。同时,根据文献[24]报道,YNR038W 具有 60S 核糖体亚单位组装和 RNA 聚合酶亚基 Rpc19 需求的 RNA 解旋酶活性。本文方法同样在排序第二的标签预测到此功能。在其他蛋白质上的预测与实际结果同样具有很高的吻合度,这充分说明了方法的可行性和可信性,希望以后对预测结果进行更完善的验证。

表 3　蛋白质功能预测结果分析

蛋白质	BP 术语			MF 术语		
	预测功能	文献报道	排名	预测功能	文献报道	排名
YGR283C	Methylation(GO:0032259)	Methylation(SGD)	1	transferase activity(GO:0016740)	methyltransferase activity[20]	5
YAL053W	fungal-type cell wall organization and biogenesis(GO:0031505); transmembrane transport(GO:0055085); transport(GO:0006810)	fungal-type cell wall biogenesis [21]; transmembrane transport; FAD transport(SGD)	1, 2, 6	transmembrane signaling receptor activity(GO:0004888); transporter activity(GO:0005215)	FAD transmembrane transporter activity (SGD)	1, 4
YBR280C	SCF-dependent proteasomal ubiquitin-dependent protein catabolic process(GO:0031146)	SCF-dependent proteasomal ubiquitin-dependent protein catabolic process[22]	1	ubiquitin-protein transferase activity (GO:0004824)	ubiquitin-protein transferase activity[23]	1
YNR038W	maturation of LSU-rRNA from tricistronic rRNA transcript (GO:0000463)	maturation of LSU-rRNA from tricistronic rRNA transcript(SGD)	1	ATP-dependent RNA helicase activity (GO:0004004)	ATP-dependent DEAD-box RNA helicase activity[24]	2
YPL230W	positive regulation of transcription from RNA polymerase II promoter (GO:0045944); response to stress (GO:0006950)	regulation of transcription from RNA polymerase II promoter in response to salt stress[25]	2, 3	sequence-specific DNA binding(GO:0043565)	DNA binding[25]	2
YOR141C	chromatin remodeling(GO:0006338); DNA repair(GO:0006281)	chromatin remodeling; cellular response to DNA damage stimulus[26]	2, 5	ATP-dependent 3′-5′ DNA helicase activity(GO:0043140); chromatin binding(GO:0003682)	ATP-dependent 3′-5′ DNA helicase activity; mRNA binding(SGD)	2, 4
YNR024W	nuclear polyadenylation-dependent tRNA catabolic process(GO:0071038); nuclear polyadenylation-dependent mRNA catabolic process (GO:0071042)	nuclear polyadenylation-dependent CUT catabolic process; nuclear polyadenylation-dependent rRNA catabolic process[27]	1, 2	RNA binding (GO:0003723)	poly(U) RNA binding [27]	1
YMR310C	Methylation(GO:0032259)	Methylation(SGD)	1	methyltransferase activity (GO:0008168)	methyltransferase activity[20]	1
YJR082C	histone acetylation (GO:0016573); DNA repair(GO:0006281)	histone acetylation; DNA repair(SGD)	1, 2	histone acetyltransferase activity(GO:0004402)	histone acetyltransferase activity[28]	1

4　结语

本文提出一种基于双重索引矩阵的随机游走算法,并采用多数据源融合方法进行蛋白质功能的预测。该算法结合功能相关性网络构建双重索引矩阵,基于蛋白质网络和功能相关性网络同时进行随机游走迭代,将达到收敛后的待预测蛋白质对应所有功能注释得分排序在 top n 的功能赋予该蛋白质。算法主要焦点在于如何在多源蛋白质网络和功能相关性网络中应用随机游走来提高预测精度。在酵母蛋白质序列、基因表达和蛋白质相互作用三种数据源上的五折交叉验证实验结果表明,提出的方法能够提高功能的预测准确率,并具有较小覆盖度。但针对注释蛋白质少的功能标签,算法预测到的概率还是偏低。今后研究的重点在于设计更有效的功能权重度量策略构建功能相关性网络,结合并优化层次相关性和

非层次相关性度量方法。

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 [J]. Proteins: Structure, Function, and Bioinformatics, 2009, 74(3): 566 – 582.

[2] PUELMA T, GUTIERREZ R A, SOTO A. Discriminative local subspaces in gene expression data for effective gene function prediction [J]. Bioinformatics, 2012, 28(17): 2256 – 2264.

[3] MOOSAVI S, RAHGOZAR M, RAHIMI A. Protein function prediction using neighbor relativity in protein-protein interaction network [J]. Computational Biology and Chemistry, 2013, 43: 11 – 16.

[4] TROYANSKAYA O G, DOLINSKI K, OWEN A B, *et al*. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae) [J]. Proceedings of the National Academy of Sciences of the United States of America, 2003, 100(14): 8348 – 8353.

[5] BARUTCUOGLU Z, SCHAPIRE R E, TROYANSKAYA O G. Hierarchical multi-label prediction of gene function [J]. Bioinformatics, 2006, 22(7): 830 – 836.

[6] LAN L, DJURIC N, GUO Y, *et al*. MS-*k*NN: protein function prediction by integrating multiple data sources [J]. BMC Bioinformatics, 2013, 14(Suppl 3): S8.

[7] ZHANG X-F, DAI D-Q. A framework for incorporating functional interrelationships into protein function prediction algorithms [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(3): 740 – 753.

[8] JIANG J Q . Learning protein functions from bi - relational graph of proteins and function annotations [C]// WABI 2011: Proceedings of the 11th International Workshop, LNCS 6833. Berlin: Springer-Verlag, 2011: 128 – 138.

[9] HU P, JIANG H, EMILI A. Predicting protein functions by relaxation labelling protein interaction network [J]. BMC Bioinformatics, 2010, 11(Suppl 1): S64.

[10] CHOU K-C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology [J]. Current Proteomics, 2009, 6(4): 262 – 274.

[11] WANG J, LI M, WANG H, *et al*. Identification of essential proteins based on edge clustering coefficient [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(4): 1070 – 1080.

[12] GUO X, GAO L, LIAO Q, *et al*. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks [J]. Nucleic Acids Research, 2013, 41(2): e35.

[13] ZHOU D, BOUSQUET O, LAL T N, *et al*. Learning with local and global consistency [EB/OL]. [2014 – 12 – 03]. http://131.107.65.14/en-us/um/people/denzho/papers/LLGC.pdf.

[14] CHATR-ARYAMONTRI A, BREIKREUTZ B J, HEINICKE S, *et al*. The BioGRID interaction database: 2013 update [J]. Nucleic Acids Research, 2013, 41(D1): D816 – D823.

[15] DWIGHT S S, HARRIS M A, DOLINSKI K, *et al*. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO) [J]. Nucleic Acids Research, 2002, 30(1): 69 – 72.

[16] MEWES H W, RUEPP A, THEIS F, *et al*. MIPS: curated databases and comprehensive secondary data resources in 2010 [J]. Nucleic Acids Research, 2010, 39(Suppl 1): D220 – D224.

[17] REN Z, WANG L, FU Z, *et al*. Ensemble learning algorithm of muti-label classification based on Ranking Loss [J]. Journal of Computer Applications, 2013, 33(S1): 40 – 42. (任志博, 王莉莉, 付忠良, 等. 基于 Ranking Loss 的多标签分类集成学习算法 [J]. 计算机应用, 2013, 33(S1): 40 – 42.)

[18] YU G, RANGWALA H, DOMENICONI C, *et al*. Protein function prediction using multi-label ensemble classification [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013, 10(4): 1045 – 1057.

[19] MOSTAFAVI S, MORRIS Q. Fast integration of heterogeneous data sources for predicting gene function with limited annotation [J]. Bioinformatics, 2010, 26(14): 1759 – 1765.

[20] YOUNG B D, WEISS D I, ZURITA-LOPEZ C I, *et al*. Identification of methylated proteins in the Yeast small ribosomal subunit: A Role for SPOUT methyltransferases in protein arginine methylation [J]. Biochemistry, 2012, 51(25): 5091 – 5104.

[21] PROTCHENKO O, RODRIGUEZ-SUAREZ R, ANDROPHY R, *et al*. A screen for genes of heme uptake identifies the FLC family required for import of FAD into the endoplasmic reticulum [J]. Journal of Biological Chemistry, 2006, 281(30): 21445 – 21457.

[22] MARK K G, SIMONETTA M, MAIOLICA A, *et al*. Ubiquitin ligase trapping identifies an SCFSaf1 pathway targeting unprocessed vacuolar/lysosomal proteins [J]. Molecular Cell, 2014, 53(1): 148 – 161.

[23] KATO M, KITO K, OTA K, *et al*. Remodeling of the SCF complex-mediated ubiquitination system by compositional alteration of incorporated F-box proteins [J]. Proteomics, 2010, 10(1): 115 – 123.

[24] FOURATI Z, ROY B, MILLAN C, *et al*. A highly conserved region essential for NMD in the Upf2 N-terminal domain [J]. Journal of Molecular Biology, 2014, 426(22): 3689 – 3702.

[25] HLYNIALUK C, SCHIERHOLTZ R, VERNOOY A, *et al*. Nsf1/Ypl230w participates in transcriptional activation during non-fermentative growth and in response to salt stress in Saccharomyces cerevisiae [J]. Microbiology, 2008, 154(Pt 8): 2482 – 2491.

[26] SARAVANAN M, WUERGES J, BOSE D, *et al*. Interactions between the nucleosome histone core and Arp8 in the INO80 chromatin remodeling complex [J]. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(51): 20883 – 20888.

[27] MILLIGAN L, DECOURTY, L, SAVEANU C, *et al*. A yeast exosome cofactor, Mpp6, functions in RNA surveillance and in the degradation of noncoding RNA transcripts [J]. Molecular and Cellular Biology, 2008, 28(17): 5446 – 5457.

[28] ZHONG P P, MELCHER K. Identification and characterization of the activation domain of Ifh1, an activator of model TATA-less genes [J]. Biochemical and Biophysical Research Communications, 2010, 392(1): 77 – 82.