整合序列与蛋白相互作用特征的亚细胞定位预测

王明会，龚 艺，王 强，冯焕清，李 骜

(中国科学与技术大学信息科学技术学院 合肥 230027)

【摘要】提出了一种基于序列和PPI特征的距离公式，可综合序列氨基酸组成和PPI对象、强弱等信息对两个蛋白质的相似性进行表征，并在此基础上提出了一种用于蛋白质亚细胞定位预测的K近邻算法。利用留一法对性能进行了评估，结果显示，在序列基础上加入PPI特征，可明显有助于亚细胞定位的预测；同时基于上述距离的K近邻算法也优于使用相同特征的SVM算法，表明该算法可以对蛋白质的亚细胞定位信息进行准确有效的预测。

关 键 词 生物信息学； K近邻算法； 蛋白质相互作用； 亚细胞定位

中图分类号 TP391； Q71 **文献标志码** A doi:10.3969/j.issn.1001-0548.2015.03.026

Prediction of Protein Subcellular Localization by Incorporating Sequence and Protein-Protein Interaction Features

WANG Ming-hui, GONG Yi, WANG Qiang, FENG Huan-qing, and LI Ao

(School of Information Science and Technology, University of Science and Technology of China Hefei 230027)

Abstract Information of protein subcellular localization is indispensable to study protein function, as a protein can perform its function only after it is correctly transported to a specific subcellular compartment. Thus it is very important to provide accurate prediction of protein subcellular localization in biological studies. In contrast to sequence features (e.g. amino acids composition) that are widely used in subcellular localization prediction, features extracting protein-protein interaction (PPI) are largely ignored, although they reflect the co-localization information of different proteins. In this study, we propose a novel distance formula based on both protein sequence and PPI features, which precisely measures the similarity of proteins by incorporating protein information including amino acid composition, PPI and the corresponding interaction scores. Based on this distance formula, we further introduce a k-nearest neighbor (KNN) algorithm for predicting subcellular localization. The results of leave-one-out test on a benchmark dataset show that PPI features significantly improve the performance of protein subcellular localization. Meanwhile, this KNN algorithm also outperformes SVM algorithm adopting the same features, suggesting the efficiency of the proposed algorithm for predicting protein subcellular localization.

Key words bioinformatics; K-nearest neighbor algorithm; protein-protein interaction; subcellular localization

生物体细胞内存在许多细胞区域和细胞器，蛋白质合成后只有转运到正确的细胞器或区域中才能发挥作用，参与各种生命活动。因此蛋白质的亚细胞定位(subcellular localization)信息对于揭示蛋白质的功能及其生命活动中发挥的作用是必不可少的[1-3]。同时，蛋白质亚细胞定位在药物设计、药物靶点的辨别和优化等方面也发挥着重要的作用。

目前可确定蛋白质亚定位的传统实验技术主要有绿色荧光蛋白标记[1]等，但由于实验效率较低，已经无法满足当前蛋白质组学快速发展的需求。为解决上述问题，利用生物信息学方法进行蛋白质亚细胞定位的研究现已取得了相当多的成果[4-8]。这些方法首先提取反映蛋白质亚细胞定位的相关特征信息，并将其转化成输入特征向量，在此基础上选择合适的机器学习和统计学方法加以预测。现有研究表明，以氨基酸组成(amino acid composition，AAC)为主的蛋白质序列信息对预测其亚定位有很大的帮助，蛋白质的序列相似程度越高，则其越趋向于存在于相同的细胞区域或细胞器内，因此是目前蛋白质亚细胞定位中的常用特征[4-8]。但是，仅通过序列

收稿日期：2013－12－18；修回日期：2014－10－27

基金项目：国家自然科学基金(61101061, 31100955)；中央高校基本科研业务费专项资金(WK2100230011)；高等学校博士学科点专项科研基金(20113402120028)

作者简介：王明会(1982－)，女，博士，副教授，主要从事生物信息学和生物统计方面的研究.

特征并不能反映蛋白质亚细胞定位的全部信息，相应的预测方法性能不够理想。另一方面，蛋白-蛋白相互作用(protein-protein Interaction，PPI)是反映蛋白相互作用和功能特性关系的重要特征[9-11]，蛋白质存在相互作用的前提是共处于细胞的同一位置，因此如果两个蛋白质存在较明显的相互作用，则其很可能存在共同的亚细胞定位。因此，如能合理使用PPI信息，将有效地提高蛋白质亚细胞定位的预测性能。

蛋白质亚细胞定位的常用预测算法有支持向量机(support vector machine，SVM)、K近邻(K-nearest neighbor，KNN)等[1]。SVM是一种基于统计学习理论的机器学习方法，该方法在结构风险最小化的原则下，保证最小的分类错误率，其缺点是在输入特征维数很高时算法复杂度大，同时性能不够理想。K近邻是一种简单有效的有监督分类方法，但是需预先定义数据之间的距离，目前大多方法是根据氨基酸组成等序列信息计算两个蛋白质的欧式距离[1-2,8]，但这种距离计算方法无法有效地整合蛋白质PPI信息。

针对上述问题，本文提出了一种结合PPI和氨基酸组成信息的距离公式，用以综合评估两个蛋白质在序列和内在功能特性上的相似性，在此基础上利用K近邻算法对数据进行了训练和测试，取得了令人满意的效果。

1 数据与算法

1.1 数据

本文从现有的Uniprot、Organelle和LOCATE3个蛋白质数据库中获得相关的蛋白质亚定位信息，从中提取出有亚定位标注的人类蛋白质，并对其进行BLAST去冗余和去除序列过短的蛋白质，最终提取胞外区、细胞核、细胞质、细胞骨架、细胞膜共5个具有代表性的亚细胞定位，具体信息如表1所示。此外，为获得相关蛋白质的PPI信息，从生物信息学数据库STRING中下载了全部共80 138条PPI记录，每条记录中都包括一对相互作用的蛋白质和相互作用强弱的数值，采用1～1 000之内的整数表示。

表1 亚细胞定位数据集

亚定位	蛋白质数
胞外区	273
细胞核	1 044
细胞质	1 734
细胞骨架	229
细胞膜	566

1.2 评价方法

为了检验算法的有效性，在评估算法性能的过程中采用以下4个评价指标：敏感性(S_n)、特异性(S_p)、准确率(ACC)和马氏相关系数(MCC)，分别定义为：

$$S_n = \frac{TP}{TP + FN} \tag{1}$$

$$S_p = \frac{TN}{TN+FP} \tag{2}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \tag{4}$$

式中，TN、TP、FN、FP分别表示用该模型测试得到的真阴性、真阳性、假阴性和假阳性数据的数目；S_n反映模型对阳性数据的预测水平；S_p反映模型对阴性数据的预测水平；ACC反映整体数据的正确预测率；MCC反映了模型对整体数据的预测水平。

1.3 算法

K近邻算法的基本思想是：对于一个分类标签的测试样本，通过找到训练数据集中距离它最近的k个近邻，再通过这k个近邻的分类标签来确定该测试样本的标签，因此确定测试样本的近邻是决定该算法性能的重要因素。在蛋白质亚细胞定位的预测研究中，对蛋白质P可使用氨基酸组成特征向量$\boldsymbol{P}_{AAC}$表征其序列信息，有：

$$\boldsymbol{P}_{AAC} = [f_1 \quad f_2 \quad \cdots \quad f_{20}] \tag{5}$$

式中，$f_i (i = 1,2,\cdots,20)$表示第i种氨基酸在蛋白质序列中出现的频率。在此基础上，可以定义任意两个蛋白质P、P'之间的距离，实际中通常采用欧氏距离进行计算，如表2所示。

$$d_{AAC}(P,P') = \sqrt{\sum_{i=1}^{20}(f_i - f_i')^2} \tag{6}$$

由于PPI强弱关系的数值与上述欧式距离在分布上具有明显的差异，因此为将两者相结合，采用了加权混合的方式计算两个存在相互作用的蛋白间的距离，有：

$$d(P,P') = (1-c)\frac{1}{d_{PPI}(P,P')} + c d_{AAC}(P,P') \tag{7}$$

式中，$d_{PPI}(P,P')$表示蛋白质P、P'之间相互作用的强弱数值，若两个蛋白之间的PPI作用越明显，则其之间的距离越近；c为预先指定的权重系数。

表2　不同蛋白质亚细胞定位预测方法的性能比较

亚定位	KNN(PPI+AAC)				KNN(AAC)				SVM(PPI+AAC)			
	MCC	S_n	S_p	ACC	MCC	S_n	S_p	ACC	MCC	S_n	S_p	ACC
胞外区	0.41	0.47	0.95	0.91	0.38	0.44	0.95	0.90	0.11	0.15	0.95	0.88
细胞核	0.44	0.70	0.76	0.74	0.36	0.62	0.76	0.72	0.08	0.31	0.76	0.63
细胞质	0.17	0.61	0.56	0.59	0.17	0.62	0.56	0.59	0.12	0.33	0.56	0.45
细胞骨架	0.50	0.65	0.94	0.92	0.48	0.63	0.93	0.91	0.29	0.37	0.94	0.90
细胞膜	0.47	0.63	0.88	0.84	0.42	0.56	0.88	0.83	0.16	0.26	0.88	0.78

在使用SVM算法进行性能比较时，所使用的PPI特征向量为：

$$\boldsymbol{P}_{\mathrm{PPI}}=[p_1 \quad p_2 \quad \cdots \quad p_M] \tag{8}$$

式中，M为PPI数据集中出现的蛋白质总数；$p_i(i=1,2,\cdots,M)$表示该蛋白质P与第i个蛋白质相互作用的强弱数值，如果没有相互作用即为0。由此将氨基酸组成和PPI特征结合得到输入SVM的最终特征向量为：

$$\boldsymbol{P}_{\mathrm{AAC+PPI}}=[f_1 \quad f_2 \quad \cdots \quad f_{20} \quad p_1 \quad p_2 \quad \cdots \quad p_M] \tag{9}$$

2　结果与讨论

为检验蛋白质亚细胞定位与蛋白之间相互作用的联系，首先利用获得的PPI信息构建了PPI的网络，同时将网络节点的蛋白质亚细胞定位信息用不同颜色标示出来，如图1所示。由图可以看出，该网络由多个聚类构成，每种聚类分别对应于具有相同定位的蛋白质，它们之间具有密切的相互作用关系。而处于不同定位的蛋白质之间尽管也存在一定程度的联系，但相对共定位的蛋白而言其PPI作用明显降低。因此，蛋白质PPI信息可以反映出蛋白质之间在亚细胞定位方面的内在联系。

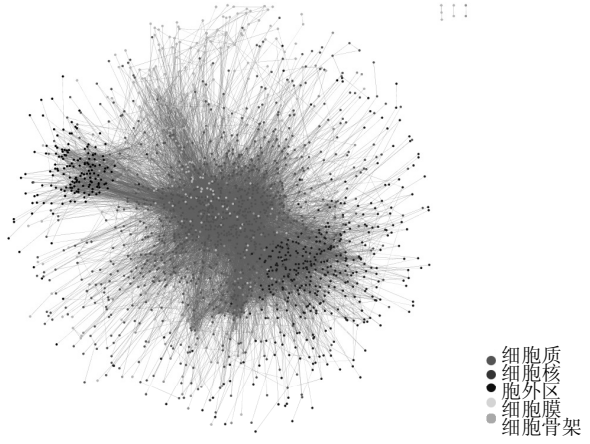
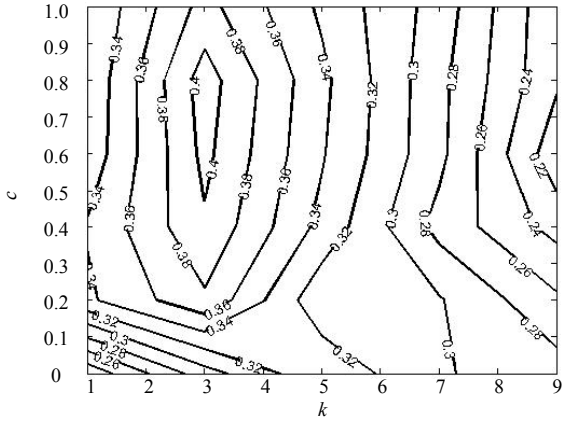


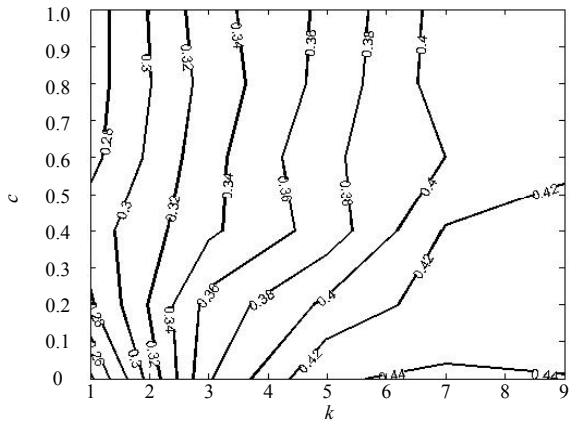
图1　亚细胞定位与蛋白质相互作用网络

本文提出的K近邻算法中有两个重要参数：近邻数k和计算蛋白距离公式中的系数c。在数据的训练和性能评估时，需要对上述参数进行选择以保证最优的分类性能。本文采用常见的网格搜索策略在整个参数空间进行寻优，由于不同亚细胞定位的数据之间数目差别很大，因此使用了对有偏数据鲁棒的马氏相关系数(MCC)作为评估指标，如图2所示。对于所有的亚细胞定位数据，通过参数寻优均可显著提高预测性能。如对于胞外区数据选择k=1、c=0时，预测结果的MCC仅为0.22；而通过网格搜索确定最优参数k=3、c=0.5后，K近邻算法的预测性能获得明显提升，其MCC达到了0.41。

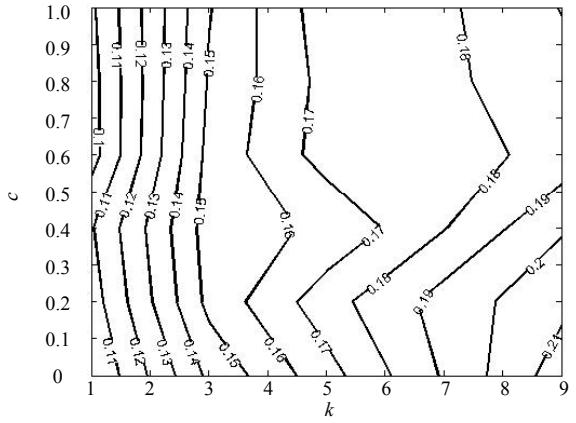
为客观评估亚细胞定位的预测性能，进一步使用留一法对本文的方法与仅使用氨基酸组成的K近邻算法进行了比较，如表2所示。除了对细胞质定位的灵敏度略低(1%)以外，本文算法的性能指标均具较明显的优势，如对于细胞核数据本文算法的马氏相关系数和灵敏度分别达到了0.44和0.70，而使用氨基酸组成的K近邻算法的相关指标仅为0.36和0.62。上述结果表明，引入PPI信息有助于定位蛋白质所属的细胞区域并提升亚细胞定位的预测精度。此外，对相关研究中广泛使用的SVM算法也进行了性能比较。由于SVM的性能同样也受参数影响，因此在实验中使用了LibSVM工具包[12]中提供的网格搜索函数对其进行了参数优化。表2的结果显示，本文算法在所有测试中均好于使用相同特征的SVM算法，这可能是由于输入SVM的PPI特征维数过高造成的。因此，在使用氨基酸组成和PPI信息时，K近邻算法能更好地对不同亚细胞区域进行区分。



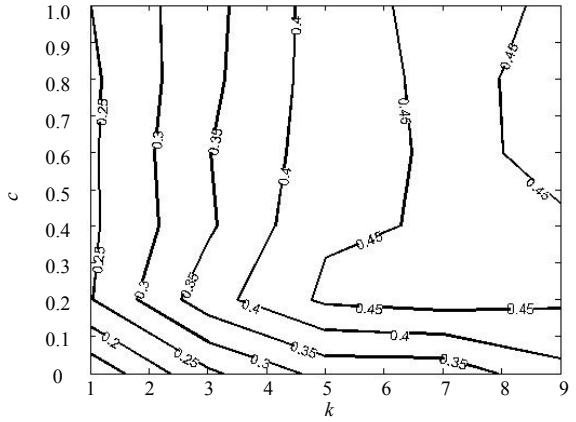
a. 跑外区



b. 细胞核



c. 细胞质



d. 细胞膜

图2 算法参数的网格搜索寻优

3 总 结

本文探讨了蛋白质相互作用信息对蛋白质亚细胞器定位预测的影响。通过网络聚类分析的结果表明，存在密切作用关系的蛋白质具有相同亚细胞定位的趋势，因此上述信息可以用于蛋白质的亚细胞定位的预测工作。为有效地整合蛋白质序列和PPI信息，本文进一步提出了一种表征蛋白质在序列和功能上相似性的距离公式，在此基础上使用K近邻算法获得了明显的性能提升。本文的工作为蛋白质亚细胞定位提供了一种新的思路，对相关预测方法的研究具有积极的意义。

参 考 文 献

[1] KENICHIRO I, KENTA N. Prediction of subcellular locations of proteins: Where to proceed?[J]. Proteomics, 2010(10): 3970-3983.

[2] CHOU Kuo-chen, WU Zhi-cheng, XIAO Xuan. iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites[J] . Mol BioSyst, 2012(8): 629-641.

[3] DU Pu-feng, YU Yuan. SubMito-PSPCP: Predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions[J]. Biomed Res Int, 2013: 263829.

[4] PIERLEONI A, MARTELLI P L, CASADIO R. MemLoci: Predicting subcellular localization of membrane proteins in eukaryotes[J]. Bioinformatics, 2011, 27(9): 1224-1230.

[5] XIE Dan, LI Ao, WANG Ming-hui, et al. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST[J]. Nucleic Acids Research, 2005, 33(suppl 2): 105-110.

[6] LI Li-qi, ZHANG Yuan, ZOU Ling-yun, et al. An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity[J]. PLoS ONE, 2012, 7(1): e31057.

[7] MARCIN M, MARCIN P, JANUSZ B M. MetaLocGramN: a meta-predictor of protein subcellular localization for Gram-negative bacteria[J]. Biochimica ET Biophysica Acta (BBA)-Proteins and Proteomics, 2012, 1824(12): 1425-1433.

[8] CHOU Kuo-chen, SHEN Hong-bin. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0[J]. PLoS ONE, 2010, 5(4): e9931.

[9] LIU Han-qing, BECK T N, GOLEMIS E A, et al. Integrating in silico resources to map a signaling network[M]. Methods Mol Biol, 2014, 1101: 197-245.

[10] LI Bi-qing, YOU Jin, CHEN Lei, et al. Identification of lung-cancer-related genes with the shortest path approach in a protein-protein interaction network[J]. BioMed Research International, 2013: 267375.

[11] PIETSCH J, RIWALDT S, BAUER J, et al. Interaction of proteins identified in human thyroid cells[J]. International Journal of Molecular Sciences, 2013, 14(1): 1164-1178.

[12] CHANG Chih-chung, LIN Chih-Jen. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27.

编 辑 黄 莘