生物物理学报 2011, 27(6) 545-554 DOI: 10.3724/SP.J.1260.2011.00545 ISSN: 1000-6737 CN: 11-1992/Q

研究论文

基于Z曲线方法重新注释脑膜炎奈瑟菌MC58株基因组

魏闻, 郭锋彪, 吴映辉

电子科技大学生命科学与技术学院，成都 610054

摘要:

已测序的微生物基因组中包含的注释开放阅读框(open reading frames, ORFs)可以分为两大类：第一类对应于功能已知的蛋白质编码基因；第二类则为功能未知的假设ORFs，其中通常有一部分实际上不编码蛋白质。采用基于Z曲线的方法从属于第一类的功能已知基因出发训练参数，进而确定第二类ORFs中非编码的部分。通过支持向量机的学习及分类，结果显示十重交叉检验平均正确率为98.45%，说明Z曲线联合支持向量机是一种高度准确的基因识别方法。最终，确定216个假设ORFs实际上不编码蛋白质。通过采用Blastp进行序列比对，保留的假设ORFs中有341个在高可靠性的条件下获得了功能信息。根据蛋白质直系同源簇方法进行功能分类，分别有30、53、59和159个新注释的假设ORFs属于信息储存和加工类、细胞加工和信号传递类、新陈代谢类和特征不明显类。另外还有70个不属于其中的任何一类。注释结果比RefSeq及GenBank提供的原注释更加准确，更加完整。

关键词： 脑膜炎奈瑟菌MC58株 基因组重注释 21变量Z曲线 支持向量机

Reannotation of ORFs in *Neisseria meningitidis* MC58 Based on Z Curve Method

WEI Wen, GUO Fengbiao, WU Yinghui

School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

Abstract:

Annotated ORFs in microbial genomes could be usually categorized into two groups: the first group corresponds to known genes; whereas the second one includes unknown-function ORFs. Because the annotation is not always accurate, it is necessary and important to confirm which ORF of the latter group is genuine gene and which is not. Starting from known genes in the former group, the authors used the combination of 21 Z curve variables and SVM to re-predict coding potentials of ORFs contained in the latter group. Ten-fold cross-validation result showed that the average accuracy of the method was greater than 98.45% for recognizing the known genes and the non-gene sequences in *Neisseria meningitides* genome. In other words, very high accuracy of recognition can be obtained by combining SVM and Z curve method. When applying the model to 810 hypothetical ORFs, 216 ones were consistently recognized as non-coding ORFs. Furthermore, functions had been assigned to 341 hypothetical ORFs with high reliability by using Blastp search. According to the COG functional categories, 30, 53, 59 and 159 newly annotated hypothetical genes belong to the "information storage and processing", "cellular processes and signaling", "metabolism" and "Poorly characterized", respectively. Consequently, it provided a more comprehensive and precise annotation for *Neisseria meningitidis* MC58 than the original GenBank and RefSeq annotations.

Keywords: *Neisseria meningitides* MC58 Genome reannotation 21 Z curve parameters Support vector machine

收稿日期 2010-09-20 **修回日期** 2010-12-29 **网络版发布日期**

DOI: 10.3724/SP.J.1260.2011.00545

基金项目:

国家自然科学基金项目(31071109，60801058)，电子科技大学“中央高校基本科研业务费”资助(ZYGX2009J082)

通讯作者: 郭锋彪，电话/传真：(028)83208232，E-mail：fbguo@uestc.edu.cn

作者简介:

作者Email: fbguo@uestc.edu.cn

参考文献：

1. Gao N, Chen LL, Ji HF, Wang W, Chang JW, Gao B, Zhang L, Zhang SC, Zhang HY. DIGAP —— A database of improved gene annotation for phytopathogens. BMC Genomics, 2010, 11: 54
2. Chen LL, Zhang CT. Gene recognition from questionable ORFs in bacterial and archaeal genomes. J Biomol Struct Dyn, 2003, 21(1): 99~109
3. Luo C, Hu GQ, Zhu H. Genome reannotation of Escherichia coli CFT073 with new insights into virulence. BMC Genomics, 2009, 10: 552
4. Yu JF, Sun X. Reannotation of protein-coding genes based on an improved graphical representation of DNA sequence. J Comput Chem, 2010, 31(11): 2126~2135
5. Cortes C, Vapnik V. Support-vector networks. Mach Learn, 1995, 20(3): 273~297
6. Chang CC, Lin CJ. LIBSVM: A library for support vector machines, 2001. [2010-09-06]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/
7. Zhang CT, Zhang R. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. Nucl Acids Res, 1991, 19(22): 6313~6317
8. Guo FB, Ou HY, Zhang CT. ZCURVE: A new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucl Acids Res, 2003, 31(6): 1780~1789
9. Fickett JW, Tung CS. Assessment of protein coding measures. Nucl Acids Res, 1992, 20(24): 6441~6450
10. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: An updated version includes eukaryotes. BMC Bioinformactics, 2003, 4:41
11. Nielsen P, Krogh A. Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics, 2005, 21(24): 4322~4329
12. Trifonov EN. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequence. J Mol Biol, 1987, 194(4): 643~652
13. Zhang CT, Chou KC. A graphic approach to analyzing codon usage in 1562 Escherichia coli protein coding sequences. J Mol Biol, 1994, 238(1): 1~8
14. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics, 2001, 17(9): 763~774

本刊中的类似文章

1．闻芳,卢欣,孙之荣,李衍达.基于支持向量机(SVM)的剪接位点识别[J]. 生物物理学报, 1999,15(4): 733-739
2．胡磊,乔立安,公衍道,赵南明.利用支持向量机预测II类MHC分子结合多肽[J]. 生物物理学报, 2001,17(4): 669-675
3．王娴,李骜,王明会,冯焕清.基于支持向量机方法的蛋白可溶性预测[J]. 生物物理学报, 2005,21(1): 60-64
4．邱天爽,郑效来,鲍海平,赵庚申.一种基于支持向量机技术的癫痫脑电棘尖波识别方法[J]. 生物物理学报, 2005,21(4): 317-321
5．晏春,杜耀华,高青斌,王正志.基于支持向量机的人类5’非翻译区剪接位点识别[J]. 生物物理学报, 2005,21(4): 284-288
6．施建宇,潘泉,张绍武,程咏梅.基于氨基酸组成分布的蛋白质同源寡聚体分类研究[J]. 生物物理学报, 2006,22(1): 49-56
7．胡秀珍,李前忠.用支持向量机识别 β-发夹模体[J]. 生物物理学报, 2007,23(6): 463-469
8．姜小莹 朱俊东 李晓波 张同亮.使用伪氨基酸组成和模糊支持向量机预测蛋白质结构类[J]. 生物物理学报, 2008,24(1): 43-48
9．杨会芳 程咏梅 张绍武 潘泉.基于分段伪氨基酸组成成分特征提取方法预测蛋白质亚细胞定位[J]. 生物物理学报, 2008,24(3): 232-238
10．施建宇 张艳宁.使用图像特征构建快速有效的蛋白质折叠识别方法[J]. 生物物理学报, 2009,25(2): 106-116
11．罗丽, 张绍武, 陈伟, 潘泉.基于分段氨基酸组成成分的蛋白质相互作用预测[J]. 生物物理学报, 2009,25(4): 282-286
12．王立贵, 应晓敏, 曹源, 查磊, 李伍举.sRNASVM——基于SVM方法构建大肠杆菌sRNA预测模型[J]. 生物物理学报, 2009,25(4): 287-293
13．李启鹏, 张绍武, 潘泉, 陈伟.基于多策略滑动伸缩窗特征提取方法预测蛋白质同源寡聚体[J]. 生物物理学报, 2009,25(5): 335-342
14．吴建盛 马昕 周童 汤丽华 胡栋.G 蛋白偶联受体及其类型的预测[J]. 生物物理学报, 2010,26(2): 138-148
15．赵秀娟 裴智勇 刘佳 蔡禄.多样性增量结合支持向量机方法预测酵母核小体定位[J]. 生物物理学报, 2010,26(5): 421-428
16．刘雷, 胡秀珍.基于添加模体信息和功率谱密度的组合向量预测27类蛋白质折叠子[J]. 生物物理学报, 2010,26(9): 823-832
17．蔡冬梅, 周卫东, 李淑芳, 王纪文, 贾桂娟, 刘学伍.基于去趋势波动分析和支持向量机的癫痫脑电分类[J]. 生物物理学报, 2011,27(2): 175-182

文章评论

反馈人		邮箱地址	
反馈标题		验证码	0828