

## 微生物基因组注释系统 MGAP

禹 胄 李 涛 蔡 涛 赵进东 罗静初\*

(北京大学生命科学学院北京大学蛋白质工程和植物基因工程重点实验室 北京大学生物信息中心 北京 100871)

**摘 要** 利用生物信息学方法和工具开发了微生物基因组注释系统(Microbial genome annotation package, MGAP),并用于蓝细菌 PCC7002 的基因组注释。该系统由基因组注释系统和基于 Web 的用户接口程序两部分组成。基因组注释系统整合多个基因识别、功能预测和序列分析软件,以及蛋白质序列数据库、蛋白质资源信息系统和直系同源蛋白质家族数据库等。用户接口程序包括基因组环状图展示、基因和开放读码框在染色体上的分布图,以及注释信息检索工具。该系统基于 PC 微机和 Linux 操作系统,用 MySQL 作数据库管理系统、用 Apache 作 Web 服务器程序,用 Perl 脚本语言编写应用程序接口,上述软件均可免费获得。

**关键词** 微生物基因组 基因组注释 生物信息学 蓝细菌 数据库

中图分类号:Q332 文献标识码:A 文章编号:1001-6209(2003)06-0805-04

基因组注释(Genome annotation)是利用生物信息学方法和工具,对基因组所有基因的生物学功能进行高通量注释,是当前功能基因组学研究的一个热点。基因组注释的研究内容包括基因识别和基因功能注释两个方面。基因识别的核心是确定全基因组序列中所有基因的确切位置。从基因组序列预测新基因,现阶段主要是 3 种方法的结合(1)分析 mRNA 和 EST 数据以直接得到结果(2)通过相似性比对从已知基因和蛋白质序列得到间接证据<sup>[1]</sup>(3)基于各种统计模型和算法从头预测。对预测出的基因进行高通量功能注释可以借助于以下方法,利用已知功能基因的注释信息为新基因注释(1)序列数据库相似性搜索(2)序列模体(Motif)搜索(3)直系同源序列聚类分析(Cluster of orthologous group, COG)<sup>[2]</sup>。

随着微生物全基因组序列测定速率的加快,开发有 Web 接口的高效、综合基因组注释系统十分必要。近年来,国际上已有一些这样的工具,如基于 Java 的微生物基因组数据库接口。尽管 JMGD 提供了一个很好的图形化接口程序,却并不具有基因组自动注释功能。德国国家环境和健康研究中心开发的蛋白质摘录、描述和分析工具(Protein extraction, description, and analysis tool, PEDANT)是大型基因组分析系统,整合了大量基因组功能信息和结构信息。PEDANT 注释功能强大<sup>[3]</sup>,适用范围广,但没有便于操作的图形界面,而且需要较强的硬件系统支持。目前,微生物基因组全序列测定通常由中小实验室独立完成,有必要开发和集成基于 PC/Linux 系统并以免费数据库管理系统、免费软件和公

基金项目:国家'863 计划'(2001AA2310011, 2002AA231061);国家自然科学基金(G1998010100);北京市科委基金(H010210350112)

\* 通讯作者。Tel: 86-10-62757281; Fax: 86-10-62759001; E-mail: luojc@pku.edu.cn

作者简介:禹 胄(1977-)男,内蒙古人,生物信息学硕士。E-mail: yuzhou@hkucc.hku.hk

北京大学生命科学学院王竹参与了部分工作

收稿日期:2003-01-08,修回日期:2003-07-21

共数据库资源为主的基因组信息注释系统。

## 1 系统和方法

### 1.1 开发环境

本系统基于 PC 微机,操作系统为 Linux。测试系统为 PIII 550 双 CPU 微机,内存 1GB,运行 RedHat 7.0 Linux 系统。数据库管理系统使用 MySQL,Web 服务器程序使用 Apache,应用程序接口用 Perl 脚本语言编写。本系统也可在单 CPU 微机上运行,内存不小于 512MB。所有系统软件和应用软件均可以从 Internet 网上免费获得。

### 1.2 测试数据

本系统用蓝细菌(*Synechococcus* sp.)PCC7002 基因组初步拼接所得最大重叠连续群(Contig)<sup>[4]</sup>作测试数据,共 303247bp。

### 1.3 MGAP 的基因组注释系统

基因组注释系统是 MGAP 的核心,整合了许多常用的基因识别和蛋白质功能预测软件,包括 GeneMarks<sup>[5]</sup>、IPRsearch、BLASTPGP 和 FASTA3 等,以及多个数据库,如非冗余蛋白质序列数据库(Non redundant, NR)、已知三维空间结构的蛋白质序列数据库(PDBSeq)、国际蛋白质资源信息系统(InterPro)<sup>[6]</sup>和直系同源蛋白质家族数据库(Cluster of orthologous groups, COG)等,编写了相应的模块进行自动操作,并把每一步注释结果导入数据库中。MGAP 整合的一般模块,可以被其他任何一种微生物基因组直接使用。不同实验室可根据实际研究需要,增加相应模块或数据,如蓝细菌 *Anabaena* sp. strain PCC 7120 的蛋白质序列库等<sup>[7]</sup>。

基因识别是 MGAP 的第一步,本系统采用微生物基因组基因识别最为权威的 GeneMarks 软件进行基因预测,通过 <http://opal.biology.gatech.edu/GeneMark/genemarks.cgi> 网站提交重叠连续群测试序列(303247bp),使用 GeneMarks 缺省参数,预测得到 279 个基因。然后用 MGAP 的数据加载模块(Loaddata)将预测结果导入 ORF 表中。

### 1.4 MGAP 的用户接口

用户接口用于展示注释结果,提供易于操作和分析平台。本系统用户接口基于 Web 设计开发,用户可通过浏览器访问基因组注释系统,包括基因组环状图展示、基因和 ORF 在染色体上分布图,并对注释信息进行检索。基因组环状基因分布图构建基于如下信息:预测所得基因的起始位置、长度、编码基因的正负链信息,以及预测的基因功能分类。

## 2 结果

图 1 是 MGAP 系统对 PCC7002 基因组重叠连续群测试序列注释结果。A 为基因展示图,B 为 ORF 显示页面。A 中由外向内依次为(1)正链编码基因(2)负链编码基因(3)GC 含量统计(4)GC 偏离量统计。该系统构建的环状基因组,可显示正负链上的编码基因,用相应颜色表示功能类别。本系统沿用经典蛋白质功能分类方法<sup>[8]</sup>,即把微生物基因组所有基因按功能分为 16 大类,进而细分为 113 个子类。此外,还增加了统计 GC 含量和 GC 偏离量(GC Bias)功能。计算 GC 含量时以 200bp 为滑动窗口,计算 GC 偏离量时以 13kb 为滑动窗口。GC 偏离量表示 G 和 C 含量的差别,定义为  $(G-C)/(G+C)$ <sup>[9]</sup>。点击 A

图中环状基因组展示图 ,则可得到 B 图基因组局部 ORF 显示页面。点击图中某个 ORF ,即可调出其所有注释信息 ,包括该 ORF 在基因组中的位置、长度、正负链信息、核酸和蛋白序列 ,以及对 NR 蛋白库、COG 数据库、InterPro、PDBseq 数据库的搜索结果。所有结果都有相对应的连接可以直接连到原始数据库。

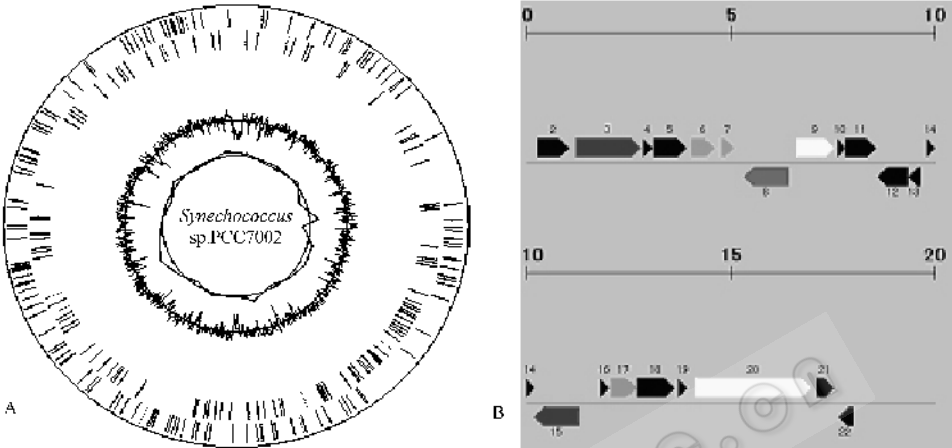


图 1 MGAP 系统对 PCC7002 基因组重叠连续群测试序列注释结果

Fig. 1 MGAP annotation results of the PCC7002 test contig

3 讨论

新基因组功能注释是基因组研究的重要方面 ,MGAP 把注释所用软件和公共数据库进行有机集成 ,使注释过程自动进行并把结果存储到数据库系统中 ,最终提供友好的界面 ,可为中小实验室提供方便实用的微生物基因组注释系统 ,减少人工参与 ,提高注释效率。该系统考虑到国内一般中小实验室的实际情况 ,基于廉价的 PC 微机和免费 Linux、MySQL、Apache 和 Perl 等软件系统开发。

必须指出 ,目前所有计算机注释信息 均不能保证完全准确。MGAP 在一定程度上依赖于现有数据库中的注释信息。由于各种原因 ,这些注释信息必然有一些错误。显然 ,这些错误信息将不可避免地引入新的注释系统。为此 ,MGAP 综合了多种注释方法 ,并互为补充。例如 ,一个 ORF 既有 BLASTP 从 NR 数据库搜索到的相似序列 ,又在 InterPro 蛋白质模体库中找到相应功能位点 ,也可找到高分匹配的 COGs ,那么该注释结果就比较可靠。此外 ,必要的人工注释 ,可以避免或纠正自动注释的错误。例如 ,由于测序错误产生的读码框移位或是缺失 ,会导致一个基因被拆分成两段 ,这种错误目前只能由手工纠正。基因组注释是一个复杂、繁琐的过程 ,需要大量的生物学知识。详尽、准确的注释需要经过严格的生物学实验才能获得。

本系统对测试序列的注释结果仍有许多未知功能基因 ,需不断扩充新数据而逐步更新。MGAP 的新版本将增加交互式用户注释模块 ,进一步扩充和增强该系统注释功能。

## 参 考 文 献

- [ 1 ] Bailey L C Jr , Searls D B , Overton G C . Analysis of EST-driven gene annotation in human genomic sequence . *Genome Res* , 1998 , **8** ( 4 ) : 362 ~ 376 .
- [ 2 ] Tatusov R L , Koonin E V , Lipman D J . A genomic perspective on protein families . *Science* , 1997 , **278** ( 5338 ) : 631 ~ 637 .
- [ 3 ] Frishman D . Functional and structural genomics using PEDANT . *Bioinformatics* , 2001 , **17** : 44 ~ 57 .
- [ 4 ] Yu Z , Li T , Zhao J . PGAAS : a prokaryotic genome assembly assistant system . *Bioinformatics* , 2002 , **18** ( 5 ) : 661 ~ 665 .
- [ 5 ] Besemer J , Lomsadze A , Borodovsky M . GeneMarkS : a self-training method for prediction of gene starts in microbial genomes . Implications for finding sequence motifs in regulatory regions . *Nucleic Acids Res* , 2001 , **29** ( 12 ) : 2607 ~ 2618 .
- [ 6 ] Apweiler R . The Interpro database , an integrated documentation resource for protein families , domains and functional sites . *Nucleic Acids Res* , 2001 , **29** : 37 ~ 40 .
- [ 7 ] Kaneko T , Nakamura Y . Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120 . *DNA Res* , 2001 , **8** ( 5 ) : 205 ~ 213 .
- [ 8 ] Riley M . Functions of the gene products of *Escherichia coli* . *Microbiol Rev* , 1993 , **57** ( 4 ) : 862 ~ 952 .
- [ 9 ] Grigoriev A . Analyzing genomes with cumulative skew diagrams . *Nucleic Acids Res* , 1998 , **26** ( 10 ) : 2286 ~ 2290 .

## MGAP-A Microbe Genome Annotation Platform

Yu Zhou Li Tao Cai Tao Zhao Jindong Luo Jingchu \*

( College of Life Sciences , National Laboratory of Protein Engineering and Plant Genetic Engineering ,  
Centre of Bioinformatics , Peking University , Beijing 100871 , China )

**Abstract :** A Microbe Genome Annotation Platform ( MGAP ) was developed and applied to the cyanobacterium PCC7002 genome annotation. Various bioinformatics software tools from sequence analysis to gene identification and function prediction were implemented in MGAP. Protein sequence databases SWISSPROT and PDBseq , protein information resource InterPro and COG were also integrated in the platform. The web interface of MGAP has the functionality to display a circular map of gene distribution and GC contents throughout the genome. Detailed information such as the DNA and protein sequence , the location of genes on chromosomes can be viewed by clicking the corresponding object within the map. MGAP is based on a PC/Linux system affordable for small biological laboratories and has the advantage of using free software tools including MySQL , Apache and Perl.

**Key words :** Microbe genome , Genome annotation , Bioinformatics , *Synechococcus* sp. strain PCC7002 , Database

Foundation item :The Natural Science Foundation of China( G1998010100 ) ;Chinese National Programs for High Technology Research and Development ( 2001AA231011 , 2002AA231061 )

\* Corresponding author . Tel : 86-10-62757281 ; Fax : 86-10-62759001 ; E-mail : luojc@pku.edu.cn

Received date : 01-08-2003