微生物基因组的生物信息学研究平台的建立*

赵贵军　何智良　卢　阳　叶兰汀　徐安龙**

（中山大学生物化学系　广州生物信息中心　广州　510275）

摘要：随着人类基因组计划及其它测序工作顺利进行，人们已经得到了大量的基因序列。如何阐明这些序列的功能和意义，是功能基因组学的主要任务。生物信息学和比较基因组学为加速这一进程提供了有利的工具。该研究建立了对已经完成全基因组测序和部分测序的25种细菌的基因组的生物信息学研究平台，提供了WEB形式的服务（http：//202．116．74．108）。25种细菌的全基因组蛋白质序列可以在NCBI的ftp：//ftp．ncbi．nlm．nih．gov/genbank/genomes/bacteria下载。该系统可以按照基因序列号、功能和种属名查询基因序列，根据美国国家信息中心（NCBI）的功能代码表对每个基因进行了自动和手工分类，并可查询分类情况，在此基础上建立了几种亲缘关系相近的种属的同源基因相互注释功能的应用。

关键词：生物信息学，细菌基因组，功能注释

中图分类号：Q93　　**文献标识码**：A　　**文章编号**：0253-2654（2002）04-0022-06

ESTABLISHMENT OF A NOVEL WEB-BASED BIOINFORMATICS PLATFORM FOR THE STUDY OF MICROBES GENOMES

ZHAO Gui-Jun，HE Zhi-Liang，LU Yang，YE Lang-Ting，XU An-Long

（*Dept．of Biochemistry　Zhongshan University，Guangzhou Center for Biotechnology Information，Guangzhou*　510275）

Abstract：With the advancement in complete sequencing of various genome，preliminary projection of gene function of

* 国家自然科学基金重点项目（No．69935020）

Major Project of Granted by Chinese National Natural Science Found（No．69935020）

** 联系人　E-mail：ls36@zsu．edu．cn

收稿日期：2001-02-23，**修回日期**：2001-06-18

novel sequences using bioinformatics methods was key to the further study on functional genomes. In this report, a web based program has been used to the gene functional classification, which was carried out according to NCBI COG's code table, and provided gene query interface and gene annotation tool according to homology. Gene functional classification for 24 completed and uncompleted sequencing bacteria genomes have been undertaken and the functional distribution among these genomes was discussed. The database and results could be accessed at http://202.116.74.108/.
Key words: Bioinformatics, Gene annotation, Bacteria genomes

大规模测序技术和生物信息学的飞速发展，导致了越来越多的生物基因组测序的完成。由于细菌基因组的规模较小，因此它们完成测序的数量也是最多，在 NCBI 有许多种细菌的全基因组序列和部分测序基因组可供全球免费下载。每种细菌的遗传背景和生活环境不同，其基因组的大小和表达的基因产物的种类、数量也将不同。本文将这些基因组的表达产物按照 NCBI 的标准来进行分类，对分类情况进行比较和统计，来研究基因组里基因按照功能分类的分布情况。

随着人类基因组计划的初步完成及其他模式生物体和细菌基因组测序的大量完成，分析基因编码规律，阐明基因功能就成了一个非常紧迫和重要的工作。各亲源关系相近的种属之间的直向同系物（orthologs）由于起源相同，在功能上也保持相近。利用已知基因功能和序列上的同源性，来初步预测未知基因功能是一种行之有效和快速的基因组注释方法。运用 BLAST 方法[1]，可以快速找到和已知基因序列同源的相应基因。本文利用这一思路，初步构建了几个亲缘关系相近的细菌种属的 BLAST 结果数据库，并应用 JAVA 和 CGI（通用网关接口）等工具做成了一个可以通过 WEB 查询的系统。

1　材料与方法

1.1　基因组序列来源

本文所用的蛋白质序列除霍乱弧菌（*Vibrio cholerae*，来自 TIGR）外其余全部来自美国国家生物信息中心（NCBI），可以从 ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/bacteria 下载，它们分别是：*Aeropyrum pernix* K1（*Aero*）、*Archaeoglobus fulgidus*（*Aful*）、*Aquifex aeolicus*[2]（*Aquae*）、*Borrelia burgdorferi*（*Bbur.*）、*Bacillus subtilis*（*Bsub*）、*Campylobacter jejuni*（*Cjej*）、*Chlamydia pneumoniae*[3]（*Cpneu*）、*Chlamydia trachomatis*[4]（*Ctra*）、*Escherichia coli*（*E.coli*）、*Haemophilus influenzae* Rd（*Hinf*）、*Helicobacter pylori* 26695（*Hpyl*）、*Mycoplasma genitalium*（*Mgen*）、*Methanococcus jannaschii*（*Mjan*）、*Mycoplasma pneumoniae*（*Mpneu*）、*Methanobacterium thermoautotrophicum*（*Mthe*）、*Mycobacterium tuberculosis* H37Rv[5]（*Mtub*）、*Neisseria meningitides* MC58（*Nmen*）、*Pyrococcus abyssi* GE5（*Pabyssi*）、*Pyrococcus horikoshii*[6]（*Pyro*）、*Richettsia prowazekii strain* Madrid E[7]（*Rpxx*）、*Synechocystis* PCC6803（*Synecho*）、*Thermotoga maritima* MSB8（*Tmar*）、*Treponema pallidum*[8]（*Tpal*）、*Ureaplasma urealyticum*（*Uure*）、*Vibrio cholerae*（*Vcholerae*）。目前基因组序列还在不断扩充当中。

1.2　编程

将各基因组的 ACCESSION NUMBER，GI NUMBER 和基因名字，简短的功能注释等都做成一个数据库，该数据库可以通过 INTERNET 查询，见图 1 所示。

1.3　基因组注释

依照 NCBI 的基因功能分类（见表 1），对每种细菌的每个基因进行归类，并作出统

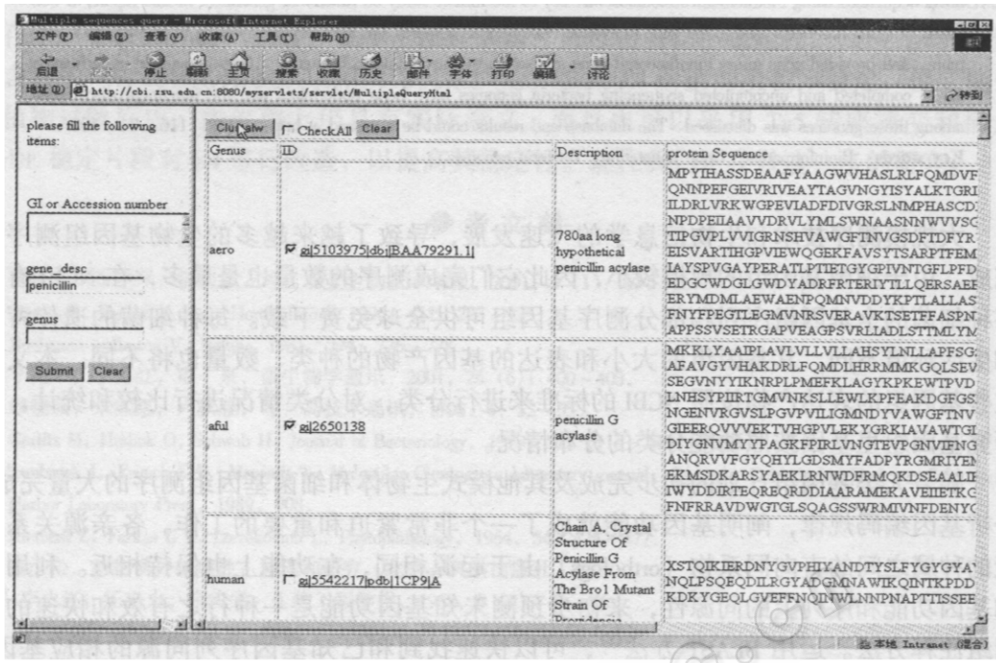

图1 基因组序列查询界面

表1 基因功能分类代码表

	Code	Description
Information storage and processing	J	Translation, ribosomal structure and biogenesis
	K	Transcription
	L	DNA replication, recombination and repair
Cellular processes	D	Cell division and chromosome partitioning
	O	Post translational modification, protein turnover, chaperones
	M	Cell envelope biogenesis, outer membrane
	N	Cell motility and secretion
	P	Inorganicion transport and metabolism
	T	Signal transduction mechanisms
Metabolism	C	Energy production and conversion
	G	Carbohydrate transport and metabolism
	E	Amino acid transport and metabolism
	F	Nucleotide transport and metabolism
	H	Coenzyme metabolism
	I	Lipid metabolism
Poorly characterized	R	General function prediction only
	S	Function unknown

计和图表，见图2。在NCBI的COG系统（http://www.ncbi.nlm.nih.gov/COG/），他们对19种完全测序的细菌基因组进行了自动功能代码分类，这19种基因组是：*Aful*、*Aquae*、*Bsub*、*Bbur*、*Ctra*、*Cpneu*、*E.coli*、*Mjan*、*Mthe*、*Mtub*、*Rpxx*、*Pyro*、*Tmar*、*Tpal*、*Mgen*、*Hinf*、*Hpyl*、*Mpneu*、*Synecho*。编制程序找出这19种基因组在我们的数据库中功能未知（即功能代码为R和S）的每个基因，然后从COG取回它们的功能代码和注释，更新我们自己的数据库。

1.4 BLAST比较分析

BLAST程序为2.0版，从NCBI下载。参数为缺省参数。把选定种属中的所有蛋白质序列与和它亲源关系较近的种属中的所有蛋白质序列作一对一的BLAST分析。由于这一步骤计算机运算量比较大，目前只完成了*Uure*、*Mgen*和*Mpneu*之间，*Aful*、*Mthe*、*Pabyssi*、*Pyro*和*Mjan*之间，*Pyro*和*Pabyssi*之间的BLAST分析。

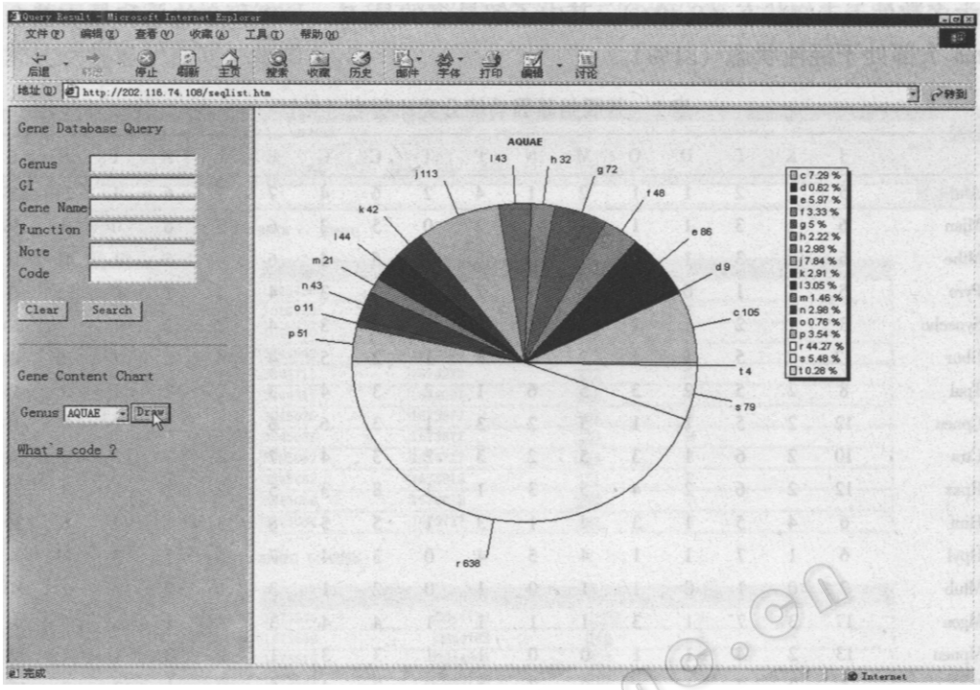

图 2　注释了的基因组查询和功能基因分布图

(http: //202. 116. 74. 108/seqlist. htm)

1.5　数据库建立

该数据库在 IRIX6.5 操作系统上建立，数据库操作系统采用 MySQL，采用 JAVA 语言编写程序，将 BLAST 分析结果全部载入数据库，这样得到一个通过 SCORE 值关联的两个不同种属基因组序列相对应的同源数据库。

1.6　编写 CGI

写一个 CGI，其输入参数有 3 个：用于比较的 2 个属的名称，以及一个 BLAST 分值。CGI 根据这 3 个参数以及上述两个数据库，生成一个关于这 2 个属中功能未知（代码为 R 和 S）的蛋白质的功能预测报告。

2　结果与讨论

如图 1 所示，在左边 3 个栏目中，用户可以根据 GI 或 ACCESSION NUMBER 基因的功能或性质、基因种属名来选择查询自己所需要的序列，系统支持模糊查询。在“Gi or Accession number”中，用户可以输入任意多个需要查询序列号码，中间用空格隔开。本例在”gene-desc”中查询“penicillin”，在右边返回数据库中所有包含“penicillin”的序列记录。

数据库中除了 18 种完全测序细菌基因组蛋白质序列外，还补充了很多部分测序细菌基因组序列和人类、老鼠、拟南芥、线虫和果蝇等模式生物体序列。通过选择、组合，用户可以点击“Clustalw”来做多重比对，从而分析青霉素相关基因在种属中的分布、变异、进化以及和环境的关系。

统计各基因组的功能代码分布情况并列表由表 2 可知，所有种属基因组的基因功

能大多数处于未知状态（R和S），其中了解最多的是 *Rpxx*，R和S的总和是41%，而 *Mtub* 大部处于混沌状态（81%）。

表2 基因组基因功能分类比例表（%）

	J	K	L	D	O	M	N	P	T	C	G	E	F	H	I	R	S
Aful	5	1	2	1	1	2	1	4	2	6	4	7	7	6	1	24	26
Mjan	6	1	3	1	1	1	1	2	0	5	2	6	2	6	0	4	56
Mthe	6	2	3	1	0	3	0	5	1	5	4	6	4	5	1	41	11
Pyro	5	1	1	0	1	0	1	2	0	2	3	4	2	1	0	75	0
Synecho	3	1	2	1	1	2	1	3	0	3	3	4	2	2	1	1	70
Bbur	12	2	5	1	1	3	7	8	1	2	5	4	4	1	3	30	13
Tpal	8	2	5	2	3	5	6	1	2	3	4	3	2	2	1	7	43
Cpneu	12	2	5	1	1	5	2	3	1	3	6	6	5	3	2	45	0
Ctra	10	2	6	1	3	5	2	3	2	3	4	7	2	3	3	5	38
Rpxx	12	2	6	2	4	5	3	1	1	8	3	5	2	2	3	4	37
Hinf	6	4	5	1	3	4	1	3	1	5	5	8	3	4	2	6	39
Hpyl	6	1	7	1	1	4	5	4	0	3	4	7	5	3	3	42	4
Mtub	2	0	1	0	1	1	0	1	0	2	1	3	1	2	2	1	80
Mgen	17	3	7	1	3	1	1	1	1	4	4	3	4	1	1	6	41
Mpneu	13	2	4	1	1	0	0	1	1	3	3	1	2	0	1	1	65
Bsub	4	2	3	1	0	2	2	1	0	2	4	5	2	1	2	0	70
Ecoli	3	1	4	1	0	2	2	2	0	1	9	6	3	3	1	26	36
Aqaue	8	3	3	1	1	2	3	4	0	8	5	6	4	2	3	47	0

阐明这部分基因功能，对实验方法学和生物信息学提出了挑战，这也是功能基因组学的主要任务。当然通过同源注释和比较基因组学，可以预测一些基因的功能。随着实验手段和其它间接方法的不断证实，大量功能未知和不确定（S和R）的基因将减少，有的基因的功能将明确改变，该表反映的基因功能分类比例也将不断变化。但这并不妨碍我们在这里先讨论一些明显的有趣的结果。在已知基因功能中与翻译RNA结构有关（J）的基因最多，这可能和该种类基因较保守，数目较多，容易辨认出来有关。与DNA复制、重组和修复有关（L）的基因在整个基因组中的比例也比较高，其中古菌（*Aful*、*Mjan*、*Mthe*、*Pyro* 等）的比例（3%左右）较其它细菌（5%左右）低，提示由于生存环境的不同（古菌大多生活在高温等极端环境下），对于DNA严紧复制的要求也不同。古菌为了适应极端环境，必须使有限的基因组附带更多的特化的遗传信息，而对于稳定遗传的要求较少，即古菌比较容易突变，以适应极端生存环境的变化需要。病原菌 *Tpal*、*Cpneu*、*Ctra*、*Rpxx*、*Hinf*、*Hpyl* 与外膜有关的基因比例较大，这可能和他们致病以及逃避宿主免疫机制有关，而结核杆菌的比例较少，可能和它的生长缓慢有关。螺旋体 *Bbur*、*Tpal* 和幽门螺杆菌 *Hpyl* 的运动能力较强，因此它们与细胞运动和分泌有关的基因（L）比例较其它属细菌高，为6%左右。

为了更加直观地显示基因组功能分类结果，本文结合数据库Access和CGI（通用网关接口）技术，做成了一个可以通过WEB查询的界面，使用户可以方便地查询结果。基因组的WEB查询界面如图2所示，可以根据基因组的种属名、基因的GI号码、基因名字、产物注释和我们所给予的功能分类代码来查询。查询速度快，结果能够分页显示。在Gene Content Chart中，用户可以选择基因组的属名，程序将作出各种分类代码在

整个基因组中所占的比例图。各种不同的颜色代表不同的分类，在整个基因组中的数目和比例都显示在同一张图上。

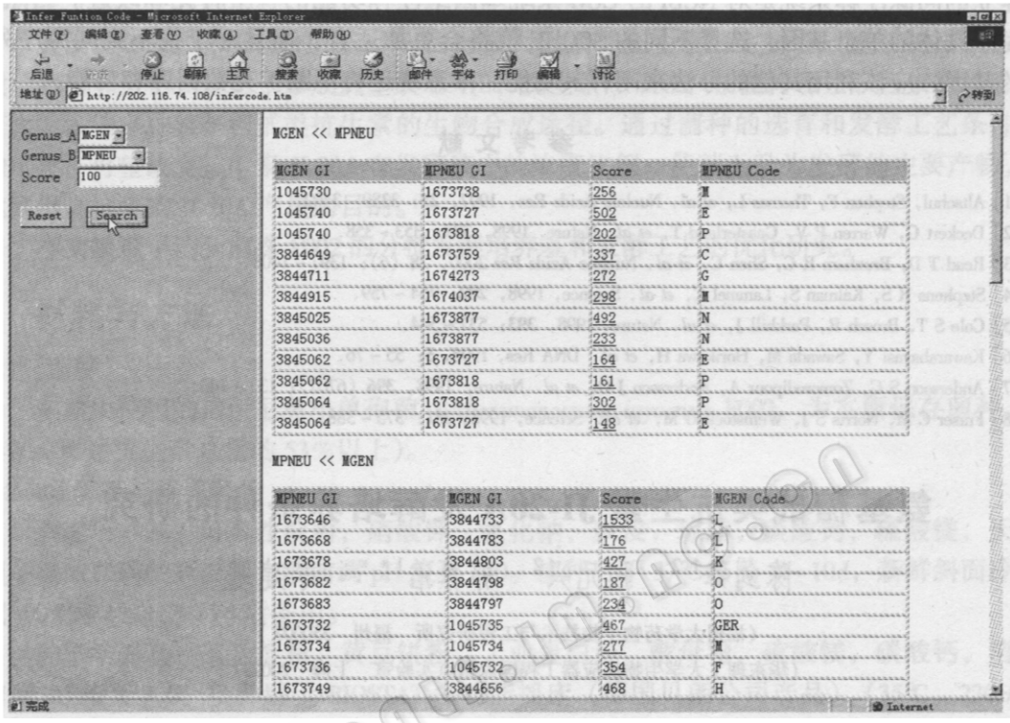

图 3　同源序列相互注释的查询界面（http：//202．116．74．108/infercode．htm）

该数据库将根据国际上的数据更新而不断更新，从而更具有实用价值。目前数据库中有 *Aful*，*Aquae*，*Aero*，*Bsub*，*Bbur*，*Cjej*，*Cpneu*，*Ctra*，*Hinf*，*Ecoli*，*Hpyl*，*Mjan*，*Mgen*，*Mpneu*，*Mthe*，*Mtub*，*Nmen*，*Pabyssi*，*Pyro*，*Rpxx*，*Synecho*，*Tpal*，*Tmal*，*Uure* 等 24 种细菌的基因组信息，并全部作了初步的功能分类。一些不易确定的基因参照了 NCBI 的 COG 分类，对于仍旧不能确定功能的大量基因（R 和 S 代码），可以利用图 3 所显示的 WEB 查询界面来查询。该程序基于这样的事实，有的种属不能确定的基因，在别的种属中功能比较确定，当他们的蛋白质同源程度很高，而且种属亲源关系比较接近的时候，通过那一种属的基因就可以来注释该种属不能确定的基因功能。反过来也如此。蛋白质之间相似关系是通过 BLAST 方法编程来确定的，缺省参数，选择 SCORE 最高的配对。由于 BLAST 需要很好的计算机速度和性能，目前该数据库只有支原体 *Uure*、*Mgen* 和 *Mpneu* 之间，古菌 *Aful*、*Mthe*、*Pabyssi*、*Pyro* 和 *Mjan* 之间，两种热球菌 *Pyro* 和 *Pabyssi* 之间的 BLAST 分析。

利用该查询系统时，先选择 Genus-B，即被 BLAST 的种属，有 *Mpneu*，*Mjan*，*Pabyssi*3 种可选，当他们分别被选定的时候，Genus-A 也随之变化，与数据库相适应。SCORE 值当然是越高，蛋白质序列相似程度越大，经过分析，发现 SCORE 在 100 时，相似性已经相当高的了，而且使用者可以直接点击 SCORE 值，查看两个序列比对结果。

这几种基因组在美国 NCBI 的 COGs 系统都有功能注释和代码分类，但是它们数据库中的数据是在至少 3 个基因组中都存在同源性的序列，因此不是完整基因组的功能

注释；对于两个基因组的情况，该系统就忽略了，然而如果这两个基因组亲缘关系很近，序列的相似性又很高，那么它们相互功能注释的可靠性也是很高的。因此本文所建立的同源注释系统是对 NCBI 的 COG 系统基因组注释功能的一个补充和改进。当然，对于具体的单个基因，选择不同的 SCORE 值将会更加适合，而且对于找不到同源序列的基因，应该利用其他的办法来阐释其功能。本系统还将不断扩充数据库和功能。

参 考 文 献

[1] Altschul, Stephen F, Thomas L, *et al*. Nucleic Acids Res, 1997, **25**: 3389 ~ 3402.
[2] Deckert G, Warren P V, Gaasterland T, *et al*. Nature, 1998, **392**: 353 ~ 358.
[3] Read T D, Brunham R C, Shen C, *et al*. Nucleic Acids Res 2000, **28** (6): 1397 ~ 1406.
[4] Stephens R S, Kalman S, Lammel C, *et al*. Science, 1998, **282**: 754 ~ 759.
[5] Cole S T, Brosch R, Parkhill J, *et al*. Nature, 1998, **393**: 537 ~ 544.
[6] Kawarabayasi Y, Sawada M, Horikawa H, *et al*. DNA Res, 1998, **5**: 55 ~ 76.
[7] Andersson S G, Zomorodipour A, Andersson J O, *et al*. Nature, 1998, **396** (6707): 133 ~ 140.
[8] Fraser C M, Norris S J, Weinstock G M, *et al*. Science, 1998, **281**: 375 ~ 388.