DOI:10.16016/j.1000-5404.2002.04.001
第24卷第4期 第三军医大学学报 Vol. 24, No. 4
2002年4月 ACTA ACADEMIAE MEDICINAE MILITARIS TERTIAE Apr. 2002

文章编号:1000-5404(2002)04-0375-03

专家论坛

微生物基因组研究进展及其对生命科学的影响

Progress in microbial genome research and its effect on life science

胡福泉 (第三军医大学基础医学部微生物学教研室、重庆市微生物工程重点实验室,重庆 400038)

1 微生物学在历史上造就的辉煌

微生物学是一个极具生命活力,发展十分活跃的学科,也是生命科学领域内重要的领头学科之一。在最近100年中,微生物领域内获得诺贝尔奖33项,有61名学者获此殊荣。生命科学领域内的许多重大发现,都是在微生物学领域或以微生物作为研究模型而获得的:①早在前基因组时代(1990年人类基因组计划启动之前),Tatum及Beadle对链孢霉菌赖氨酸营养缺陷型的研究带来了一个基因一个酶的学说;②Griffith通过肺炎球菌转化试验,证实了基因的载体是DNA;③Watson及Crick从研究病毒DNA的本质属性中,提出了DNA双螺旋结构模型;④Yanosfky通过研究大肠杆菌trpA基因,证明了蛋白质与基因的共线性关系;⑤Jacob及Monod根据大肠杆菌Lac系统的研究,提出了基因操纵子学说。微生物学在基因组时代对科学作出的贡献同样令人瞩目。比如:①以λ噬菌体为载体,利用部份酶切物理作图,构建了大肠杆菌基因组的Contig,这种基因作图方式是基因组计划中使用的基本技术;②利用微生物学中对复制子的深入研究和人工改造,构建了基因组克隆中使用的各种载体,如λ、Cosmid、YAC、BAC及PAC等;③在对大肠杆菌M13噬菌体深入研究的基础上建立了以M13单链DNA为基础的双脱氧核苷酸测序方法;④Sanger等人对λ噬菌体的全序列测序战术指导了基因组大片段测序工作;⑤1995年公布的由Fleischmann, RD等人建立的基因组随机测序与整合的测序战略是测序技术中的一场革命[1];⑥基因knock out技术在细菌和酵母中首获成功并被广泛应用;⑦利用酵母建立的two-yeast-hybidization系统成为了研究蛋白质与蛋白质相互作用的有力工具。综上所述,用著名的生物学家Bernand Dixon在PNSA中的评论最能说明微生物学在近代所作出的贡献,他说:"20世纪结束的时候,微生物学可以无愧地宣布,在生物学发展的3个关键的阶段上(DNA双螺旋结构与中心法则、遗传工程及人类基因组研究)它都站在了时代的最前列,发挥着不可替代的、独特的作用"。关于微生物学的未来,他说:"微生物学在本世纪和下个世纪都将继续是领头的学科"。

2 微生物基因组研究进展概况

在当代生命科学领域内,影响最大、处于最前沿的当数人类基因组学,故人们把我们所处这个时代称为"基因组时代"、"后基因组时代"。基因组时代的标志就是人类基因组计划(Human genome project, HGP),这是一项以测序和注释人类基因组为主要内容的、举世闻名的跨国科学研究计划。对微生物基因组的测序研究早在人类基因组计划之前就已有之,如Frank A等人(1980)就完成了花叶病毒基因组(8024 bp)的测序;Arrand JR等人(1981)完成了EB病毒的全基因组测序(172281 bp);Dunn JJ等人(1983)完成了T7噬菌体的基因组(39937 bp)测序;Peeters BP等人(1985)完成了大肠杆菌丝状噬菌体Ike的基因组(6883 bp)测序。Chee MS等人1990年完成了巨细胞病毒的基因组(229354 bp)测序[2]。在人类基因组计划中,有些模式微生物(如大肠杆菌和酵母菌)本身就是人类基因组计划的研究内容。由于微生物基因组相对较小,易于操作,故对微生物基因组的研究先行一步,起到"先行官"的作用。在微生物基因组研究方面所取得的理论和技术进展,为人类基因组计划提供了极有益的借鉴。而人类基因组计划的强大资金投入又极大地促进了微生物基因组计划的飞速发展。但作为微生物基因组计划((Microbial genome projects, MGP)的正式开始,是在1994年有美国能源部立项启动的。这项计划一经启动其发展就甚为神速,目前微生物基因组计划已经涉及到180多个细菌,测序工作量已达到约6亿碱基对。从表1可见微生物基因组计划发展的惊人速度。由于微生物种类的多样性,微生物基因组计划将扩展至千千万万不同的微生物基因组。可以预计,今后微生物基因组计划的总投入和测序工作量都将可能超越人类基因组计划。它对本学科、其它学科以及人类本身产生的深远影响将是不言而喻的。

表1 最近1年微生物基因组计划进展情况

Tab 1 Progress of microbial genome projects in last year

	Oct. 20, 2000	July 25, 2001	Oct. 31, 2001
Completed	37	55	67
Annotation in progress	14	17	18
Sequencing in progress	76	93	105
In total	127	165	190

表2 截止2001年已完成并正式发表的微生物基因组

Tab 2 Microbial genomes completed in public databases before 2001

Microorganism	Base pair	Coding protin
Agrobacterium tumefaciens circular chr.	2,841,581 bp	2721 proteins
Agrobacterium tumefaciens linear chr.	2,074,782 bp	1833 proteins
Agrobacterium tumefaciens circular chr.	2,841,490 bp	2785 proteins
Agrobacterium tumefaciens linear chr.	2,075,560 bp	1876 proteins
Aeropyrum pernix	1,669,695 bp	2694 proteins
Aquifex aeolicus	1,551,335 bp	1522 proteins
Archaeoglobus fulgidus	2,178,400 bp	2420 proteins
Bacillus halodurans C-125	4,202,352 bp	4066 proteins
Bacillus subtilis	4,214,814 bp	4100 proteins
Borrelia burgdorferi	910,724 bp	850 proteins
Brucella melitensis	3,294,931 bp	3198 proteins
Buchnera sp. APS	640,681 bp	564 proteins
Campylobacter jejuni	1,641,481 bp	1654 proteins

基金项目:国家自然科学基金资助项目(30070037);重庆市攻关课题资助项目(2000)

作者简介:胡福泉(1953-),男,四川省珙县人,博士,教授,博士生导师,主要从事微生物学方面的研究,发表论文41篇。电话:(023)68752240,E-mail:hufuquan@yahoo.com

收稿日期:2002-02-11;修回日期:2002-03-05

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

续表

Microorganism	Base pair	coding protin
Caulobacter crescentus	4, 016, 947 bp	3737 proteins
Chlamydia muridarum	1, 069, 412 bp	909 proteins
Chlamydia trachomatis D/UW-3/CX	1, 042, 519 bp	894 proteins
Chlamydophila pneumoniae CWL029	1, 230, 230 bp	1052 proteins
Chlamydophila pneumoniae AR39	1, 229, 853 bp	997 proteins
Chlamydophila pneumoniae J138	1, 228, 267 bp	1070 proteins
Clostridium acetobutylicum	3, 940, 880 bp	3672 proteins
Deinococcus radiodurans	2, 648, 638 bp	2580 proteins
Escherichia coli K12	4, 639, 221 bp	4289 proteins
Escherichia coli O157: H7 EDI933	5, 528, 970 bp	5349 proteins
Escherichia coli O157: H7	5, 498, 450 bp	5361 proteins
Haemophilus influenzae	1, 830, 138 bp	1709 proteins
Halobacterium sp. NRC1	2, 014, 239 bp	2058 proteins
Helicobacter pylori 26695	1, 667, 867 bp	1566 proteins
Helicobacter pylori J99	1, 643, 831 bp	1491 proteins
Lactococcus lactis	2, 365, 589 bp	2266 proteins
Listeria monocytogenes	2, 944, 528 bp	2846 proteins
Listeria innocua	3, 011, 208 bp	2968 proteins
Methanobacterium thermoautotrophicum	1, 751, 377 bp	1869 proteins
Methanococcus jannaschii	1, 664, 970 bp	1715 proteins
Mesorhizobium loti	7, 036, 074 bp	6752 proteins
Mycobacterium leprae	3, 268, 203 bp	1605 proteins
Mycobacterium tuberculosis	4, 411, 529 bp	3918 proteins
Mycobacterium tuberculosis CDC1551	4, 403, 836 bp	4187 proteins
Mycoplasma genitalium	580, 073 bp	484 proteins
Mycoplasma pneumoniae	816, 394 bp	677 proteins
Mycoplasma pulmonis	963, 879 bp	782 proteins
Neisseria meningitidis MC58	2, 272, 325 bp	2025 proteins
Neisseria meningitidis Z2491	2, 184, 406 bp	2121 proteins
Pasteurella multocida	2, 257, 487 bp	2014 proteins
Pseudomonas aeruginosa	6, 264, 403 bp	5565 proteins
Pyrococcus alyssi	1, 765, 118 bp	1765 proteins
Pyrococcus horikoshii	1, 738, 505 bp	1979 proteins
Rickettsia conorii	1, 268, 755 bp	1374 proteins
Rickettsia prowazekii	1, 111, 529 bp	834 proteins
Ralstonia solanacearum	5, 810, 922 bp	5116 proteins
Salmonella typhimurium LT2	4, 857, 432 bp	4451 proteins
Salmonella typhi	4, 809, 037 bp	4395 proteins
Sinorhizobium meliloti	3, 654, 135 bp	3341 proteins
Staphylococcus aureus N315	2, 813, 641 bp	2595 proteins
Staphylococcus aureus Mu50	2, 878, 134 bp	2697 proteins
Streptococcus pneumoniae R6	2, 038, 615 bp	2043 proteins
Streptococcus pneumoniae TIGR4	2, 160, 837 bp	2094 proteins
Streptococcus pyogenes	1, 852, 451 bp	1696 proteins
Sulfolobus solfataricus	2, 992, 245 bp	2977 proteins
Sulfolobus tokodaii	2, 694, 765 bp	5670 proteins
Synechocystis PCC 6803	3, 573, 470 bp	3169 proteins
Thermoplasma acidophilum	1, 564, 906 bp	1509 proteins
Thermoplasma volcanium GSS1	1, 585, 104 bp	1499 proteins
Thermotoga maritima	1, 860, 725 bp	1846 proteins
Treponema pallidum	1, 138, 011 bp	1031 proteins
Ureaplasma urealyticum	751, 719 bp	613 proteins
Vibrio cholerae	4, 033, 464 bp	3827 proteins
Xylella fastidiosa	2, 679, 306 bp	2766 proteins
Yersinia pestis	4, 653, 728 bp	4008 proteins

目前，我国正在进行之中的微生物基因组计划有：① Thermoanaerobacter tengcongensis；② Shigella；③ Staphylococcus epidermidis，④ Leptospira interrogans serovar icterohaemorrhagiae Lai。

3 微生物基因组计划的深远影响

对一个基因组，单纯测序工作的完成，还仅仅是解决了从字母到字母（From letter to letter）的排序问题，得到的仅仅是一部由 A，T，G，C 四个字母组成的没有“词”（Word）、“词组”（Phase）、“句子”（Sentence）的“天书”。要读懂这部天书就必须对“天书”进行注释，即要弄清构成这部天书的“词”、“词组”、“句子”、“段落”乃至“全文”的意义，也就是说要解读基因组序列中各功能元件的意义。一个基因组测序完成后，必须基于人类对该基因组所积累的认识，进行尽可能的注释。目前，已经发表的那些有关微生物基因组测序的文章，都还只是对基因组的“部分注释”或“相对注释”。对一个细菌基因组的完全注释，即读懂“全文”，决不是一代人所能完成的工作，这将是“后基因组时代”的艰巨任务。

目前，大部分常见医学微生物都已有代表株完成测序或正在被测序之中。尽管目前对已完成测序的微生物基因组都还只是“部分注释”，但所获得的信息已经使得既是微生物学专业工作者也应接不暇。这些信息正在全面更新微生物学的现有知识，对微生物学乃至整个生命科学产生着极为深远的影响。

在微生物基因组完成测序和注释之后，人们对该微生物的每一个基本生物学特性、致病性、免疫性等重要问题的理解都将更全面、更精细和更准确。从目前已经完成测序的为数有限的微生物基因组学研究来看，已经向人们展示了大量前所未知的全新事实：如发现了某些细菌其实不止一个染色体，嗜盐古细菌 sp. NRC-1 基因组有一个大染色体和两个小染色体（Mini-chromosome），大染色体含有 2, 014, 239 个碱基对，两个小染色体分别含 191, 346 个碱基对和 365, 425 个碱基对[3]；霍乱弧菌也含有 2 条环状染色体，较大的染色体含 2961146 bp，G+C 含量为 46.9%，较小染色体含 1072314 bp，G+C 含量为 47.7%[4]；细菌基因组携带前噬菌体（Prophage）基因组是一种普遍现象，其携带率达百分九十几，有的甚至高达 100%；这些溶原性前噬菌体在不同细菌之间的转移，形成了一种“基因流”；这种基因流在细菌致病性、耐药性的传递中具有重要意义[5~7]；细菌基因组大约 90%左右的序列都是编码序列，相比之下，人类基因组的编码序列只占 3%~5%，可以下结论认为：前核生物对碱基序列的使用是高利用率的，这反映出细菌遗传序列的“储备量”（即未利用序列）较人类少得多[3~8]；麻风杆菌基因组研究揭示，该菌与结核杆菌有共同的祖先，在进化的长河中，麻风杆菌发生了退化，产生了众多假基因，推定有 4.4 Mb 的结核杆菌基因组编码了约 4000 种蛋白，按比例，3.3 Mb 的麻风杆菌基因组应编码约 3000 种蛋白。但实际上，麻风杆菌只检测到 391 种可溶性蛋白质，提示麻风杆菌中存在大量的假基因没有活性[8,9]。这种退化现象的原因令人感兴趣。基因组测序研究显示，胞内寄生物都曾有退化现象，如立克次体、衣原体及一些胞内共生生物，一旦在高度特化的生境中不再需要某些功能，决定这些功能的基因就会退化而失去活性。这种现象使一些胞内寄生微生物只需拥有最小限度的基因，这体现出了生物在生存与进化中的“经济性”原则[8~13]。仅此几例，足可以说明微生物基因组研究向我们揭示的崭新画面。

在研究生物蛋白质进化关系方面，目前有一种非常有用的

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

“垂直同源蛋白质组的聚类(Cluster of orthologous groups of proteins, COG)分析法, 它是通过计算机对世界大型生命数据库进行检索和同源性比对, 从而对基因或蛋白质的序列进行分析, 进而确定它们的种系发生关系的聚类方法[14,15]。该方法是分析蛋白质进化关系的最强有力方法, 分析的结果将每一个蛋白质进化系称为一个 COG 类。COGs 的聚类是通过“全基因组对全基因组比较”(All-against-all sequence comparison)来实现的, 将来自某一基因组的蛋白, 分别一个一个的对另一个基因组来源的每一个蛋白进行比较, 找出“最佳命中”(Best hit)的蛋白。如此反复循环, 在各个已完成全基因组测序的微生物基因组中进行轮番比对, 每一个蛋白都需找到“彼此最佳命中”(Reciprocal best hit)蛋白。那些彼此最佳命中蛋白就构成一个 COG 类。构成一个 COG 中的任何一个蛋白, 都必须出现在至少 3 个微生物种中, 才能构成一个 COG 类。同一个 COG 类中的蛋白质是垂直进化的关系(Verical evolutionary descent)。值得注意的是, 这种垂直进化关系不仅仅指“一对一”(One-to-one)的传递关系, 当进化产生出“谱系特异性基因复本”(Lineage-specific gene duplications)时, 这种关系也可表现为“一对多”(One-to-many)及“多对多”(Many-to-many)的关系。

COGs 分类中使用的是“COGNITOR”软件(网址: http://www.ncbi.nlm.nih.gov/COG)。COGs 数据库及其配置的“COGNITOR”分析软件是作为新测序微生物基因组注释的工作平台和研究基因组进化的平台。该工作平台目前基本上只使用于微生物基因组分析。在 COG 数据库中已经收集的基因组达 44 个, 包括细菌、古细菌和酿酒酵母。最近又加入了一个古细菌 Aeropyrum pernix[16] 及两个多细胞真核基因组, 即线虫 Caenorhabditis elegants 和果蝇 Drosophila melanogaster[17]。这些数据的加入是很有意义的, 这向数据库中提供了一些原来缺乏的进化谱系的数据。由于这些信息的加入, 使得基因组注释中, “有把握的功能预测”(Confident functional predication)提高到约 50%[15]。

显然, 在微生物基因组研究中使用的 COGs 分类, 对于在整个生物学界鉴定基因功能、蛋白质进化和生物进化都将作出巨大的贡献。

人类认识自然的目的在于适应自然、利用自然、甚至改造自然, 使之为人类服务。人类从来没有放弃过这种努力, 例如培育出的各种生物“良种”, 预防传染病的疫苗, 就是成功的例子。基因组研究的结果, 使得生物的遗传学背景被精确认识, 为人类改造生物创造了前所未有的良好前景。由于微生物基因组相对较小, 如病毒(包括噬菌体)基因组只有数千个碱基, 在利用现代基因工程技术对生物进行改造的进程中, 显然在微生物这类微小生物中最富可操作性, 最有可能取得成功。在人们普遍认为信息技术和生物技术将成为未来社会的两大支柱产业的时候, 我们有理由预测微生物学将为未来社会带来巨大的社会效益和经济效益, 造就更大的辉煌。

在后基因组时代, 微生物学将起着什么样的作用呢? 可以肯定地说, 微生物学将在未来生命科学中继续起着领头作用。这是因为, 微生物是最小的生命体, 其结构与功能的关系较之高等生物简单得多, 相应的其基因组也就小得多。因此, 微生物将是功能基因组研究的良好对象。微生物的许多生物学性状都是单基因控制的, 要阐明这些基因的功能显然要容易得多。在微生物系统中阐明了某一基因的功能, 通过同源性比较, 人们可以揭示在其它生物中的相应基因的功能。对于许多生命现象奥秘的揭示, 将首先在微生物中实现。在微生物学领域内取得的理论和技术成就, 将继续成为其它生物研究工作的借鉴。

关键词: 微生物; 基因组; 注释; 蛋白进化
中图法分类号: R37　　文献标识码: A

参考文献:

[1] Fleischmann R D, Adams M D, White O, *et al*. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd[J]. Science, 1995, 269(5 223): 496—512.

[2] Chee M S, Bankier A T, Beck S. Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD 169[J]. Curr Top Microbiol Immunol, 1990, 154: 125—169.

[3] Ng W V, Kennedy S P, Mahairas G G, *et al*. Genome sequence of Halobacterium species NRC-1[J]. Proc Natl Acad Sci, USA, 2000, 97(22): 12 176—12 181.

[4] Heidelberg J F, Eisen J A, Nelson W C, *et al*. DNA Sequence of both chromosomes of the cholera pathogen Vibrio cholerae[J]. Nature, 2000, 406(6 795): 477—483.

[5] Kuroda M, Ohta T, Uchiyama I, *et al*. Whole genome sequencing of meticillin-resistant Staphylococcus aureus[J]. Lancet, 2001, 357(9 264): 1 225—1 240.

[6] Stover C K, Pham X-QT, Erwin A L *et al*. Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen [J]. Nature, 2000, 406(6 795): 959—964.

[7] Blattner F R, Plunkett G III, Bloch C A, *et al*. The complete genome sequence of Escherichia coli K-12[J] Science, 1997, 277(5 331): 1 453—1 474.

[8] Cole S T, Eiglmeier K, Parkhill J, *et al*. Massive gene decay in the leprosy bacillus[J]. Nature, 2001, 409(6 823): 1 007—1 011.

[9] Cole S T, Brosch R, Parkhill J, *et al*. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence[J]. Nature, 1998, 393(6 685): 537—544.

[10] DelVecchio V G, Kapatral V, Redkar R J, *et al*. The genome sequence of the facultative intracellular pathogen Brucella melitensis[J]. Proc Natl Acad Sci USA, 2002, 99: 443—448.

[11] Shirai M, Hirakawa H, Kimoto M, *et al*. Comparison of whole genome sequences of Chlamydia pneumoniae J138 from Japan and CWL029 from USA[J]. Nucleic Acids Res, 2000, 28(12): 2 311—2 314.

[12] Himmelreich R, Hilbert H, Plagens H, *et al*. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae[J]. Nucleic Acids Res, 1996, 24(22): 4 420—4 449.

[13] Ogata H, Audic S, Barbe V, *et al*. Selfish DNA in protein-coding genes of Rickettsia[J]. Science, 2000, 290(5 490): 347—3 502.

[14] Tatusov R L, Koonin E V, Lipman D J, *et al*. A genomic perspective on protein families[J]. Science, 1997, 278(5 338): 631—637.

[15] Tatusov R L, Natale D A, Garkavtsev I V, *et al*. The COG database: new developments in phylogenetic classification of proteins from complete genomes[J]. Nucleic Acids Res, 2001, 29(1): 22—28.

[16] Kawarabayasi Y, Hino Y, Horikawa H. *et al*. Complete Genome Sequence of an Aerobic Hyper-thermophilic Crenarchaeon, Aeropyrum pernix K1[J]. DNA Res, 1999, 6(2): 83—101.

[17] Adams M D, Celniker S E, Holt R A, *et al*: The genome sequence of Drosophila melanogaster[J]. Science, 2000, 287(5 461): 2 185—2 195.

(编辑　冷怀明)

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net