生物信息学以及植物新基因的发现研究

魏小春,郑　群

(石河子大学 园艺系,新疆 石河子 832003)

摘　要:新基因的发现对植物生命科学的研究及发展起着不可估量的作用,可以增强植物抗逆性、提高作物产量、改善经济植物品质等。然而,在发现植物新基因的方法中,生物信息学始终扮演着重要的角色。

关键词:新基因;生物信息学;抗逆性;产量;品质

中图分类号:Q 943.2　文献标识码:A　文章编号:1001-0009(2009)05-0118-04

生物学中一个普遍的问题就是寻找新基因。发现新基因是当前国际上基因组研究的热点,使用生物信息学的方法是发现新基因的重要手段[1]。传统上,基因和蛋白质的发现需要用分子生物学和生物化学技术。互补DNA可以用克隆的方法从基因文库中克隆得到,而蛋白质纯化后可以用酶活性等生物化学指标测序。生物信息学方法也能够为新基因的存在提供一些证据。从目的来看,一个新基因就是指在数据库中发现的一些还没有被注释的DNA序列。

生物信息学是20世纪80年代末随着人类基因组计划的启动而兴起的一门新的交叉学科,由数据库、计算机网络和应用软件三大部分构成[2-3]。利用生物信息学方法是发现新基因的重要手段。比如酵母基因组包含了5 932个基因,其中60%的基因是通过信息分析得到的。使用EST序列信息寻找新基因是当前国际上基因争夺战的热点。利用生物信息学方法预测基因的功能也是强有力的武器。如上述酵母基因组包含的5 932个基因中,已知其功能的占6 500个,其中30%来自试验,3 500个(约54%)来自信息技术。

第一作者简介:魏小春(1983-),男,在读硕士,主要研究方向为植物生理生化。E-mail:jweixiaochun@126.com。

通讯作者:郑群(1968-),男,博士,副教授,现从事蔬菜生理生态研究工作。E-mail:zq1508@sina.com。

基金项目:石河子大学基金资助项目(ZRKX200706)。

收稿日期:2009-01-10

1　新基因的丰富内涵

基因的发现和基因工程的长足进展,使人类突破了传统观念的束缚,开始大胆尝试干预生命、改变生命。20世纪80年代,在加拿大的一片苹果园里,就通过转基因获得了别有滋味,既甜又脆,十分可口的苹果"芭蕾皇后"。用基因疗法克服植物病害当英国正在紧锣密鼓地进行超级苹果的研究时,在澳大利亚和马来西亚,科学家开始了超级菠萝的研究工作。澳大利亚和马来西亚是两个菠萝生产大国,近年来,这两个国家面临着一个棘手的问题-菠萝的黑心病和菠萝树冠的退化。为了挽回黑心病给菠萝产业带来的损失,科学家将用基因技术找到引起菠萝黑心病的酶,并用基因技术和克隆技术攻克它。对于树冠的退化现象,科学家也将通过改变它的基因来解决[4]。

中国超级稻计划项目组已在某种水稻上发现2个位点,2个位点的基因共拥有使水稻增产36%的潜能。世界杂交水稻之父袁隆平日前在世界农业科技大会公布了项目组的新发现。袁隆平说,生物技术可以加速中

Study on the Genes Related to the Fruits' Dehiscence in Higher Plant

YANG Ai-zhen[1,2], ZHANG Zhi-yi[1], ZHAO Fu-kuan[2], WANG You-nian[2]

(1. College of Biology, Beijing Forestry University, Beijing 100089, China; 2. Department of Biotechnology, Beijing University of Agriculture, Beijing 102206, China)

Abstract: This paper took *Arabidopsis thaliana* as an example, outlined the genes and downstream genes relavant to the formation of disassociation layer after fruits ripen, explored the interaction between genes, lay the theoretical foundation of the fruits dehiscence regulation, lead ways to solve agricultural practice.

Key words: Higher plant; Fruit; Gene; Hormone

国的超级稻研究。如果转入玉米的某种基因,超级水稻的单产还会有大幅增长的潜力[5]。中科院上海生科院植物生理生态所、植物分子遗传国家重点试验室林鸿宣研究员领导的研究组,在水稻产量相关功能基因研究上取得突破性进展,成功克隆了控制水稻粒重的数量性状基因 *GW2*,并深入阐明了该基因的生物学功能和作用机理,显示该基因在高产分子育种中具有应用前景。该研究成果为作物高产育种提供了具有自主知识产权和重要应用前景的新基因,为阐明作物产量和种子发育的分子遗传调控机理提出了新见解[6]。

2 新基因起源的分子机制

人们对新基因起源这一问题的兴趣可以追溯到 20 世纪 30 年代,尽管当时对遗传物质的本质还没有清晰的认识,Haldane[7] 和 Muller[8] 就已提出通过基因重复可以产生新的基因。此后,得益于分子生物学试验手段的进步和遗传学的发展,人们进一步认识了基因的本质,观察到大量的试验现象,如染色体重复、基因家族和断裂基因等,并在此基础上提出了一些新基因产生的假说[9-10]。20 世纪 80 年代中期以后,大规模基因组序列信息的获得以及分子进化和群体遗传学理论的成熟,更使得在基因组水平的理论预测成为可能[11]。1993 年,华裔学者龙漫远(Long)等人[12] 发现了第 1 个年轻基因——精卫基因(*jingwei*),从此新基因起源的研究进入了一个新的时期。此后,又有司芬克斯(*sphinx*)基因和猴王基因(*monkey king*)等大约 20 多个年轻基因被报道。一个新基因的产生是一个复杂的过程,常常综合了多种机制。有关新基因起源的分子机制,Long 等人[13] 已作过系统的介绍,其主要有基因重复(gene duplication)、外显子重排(exon shuffling)、逆转座(retro-transposition)、可移动元件(mobile elements)、基因水平转移(gene lateral transfer)和基因分裂与融合(gene fission and fusion)等。

3 利用生物信息学技术发现和克隆新基因

随着人类基因组计划(HGP)的开展和各种模式生物(如拟南芥、水稻等)的基因组预测工作相继完成,重要的是要从大量不连贯的信息中发现其中隐藏的重要信息。基因组信息学的首要任务之一就是发现新的基因和新的功能,使用基因组信息学的方法是发现新基因的重要手段。在植物中发现一个新基因,就可以利用该基因进行作物品种的改良和新品种的培育。

3.1 利用 EST 数据库发现新基因(电脑克隆)

由于 GENBANK 中收集到的 ESTs 数目呈现指数增长的趋势,利用 EST 数据库信息数据进行功能基因的电子克隆是目前最常用的手段。首先获得感兴趣并可能有潜在功能位点的 EST 作为种子序列,用 Blast 对 dbEST 库进行同源性搜索,找到重叠部分且同源性高的 ESTs 进行拼接,获得序列重叠群,然后再以拼接好的重叠群为新的种子序列,重复上述拼接步骤,直到不能继续获得延伸为止,即为电子克隆的最终序列。然后可利用相关的网络和软件工具对其进行 cDNA 完整性分析及结构和功能的分析预测[14]。

EST 作为遗传标记最初应用于人类基因组遗传作图。EST 技术使基因克隆发生革命性变革,以 EST 为重要来源的染色体物理图谱进一步方便了对候选基因的连锁分析,可将其确定在更狭小的染色体区段内,缩小了基因的筛选范围。EST 本身为表达基因的产物,用其取代。DNA 全长的筛选、基因组的鉴定等繁琐的试验操作,可大大提高分离基因的效率。将所获 EST 用生物信息学的方法与各公共数据库中已知序列进行比较,可迅速而准确地确定基因功能。由于在构建 cDNA 文库时要尽可能地选用全长 cDNA,所以一旦发现有价值的 EST,就可以找到对应的克隆,获得的全长 cDNA 可以直接用于如转基因等的研究。利用 EST 方法进行发现、分离新基因的研究,不仅是人类基因组研究的热点,而且是植物基因组研究的重要内容。但是由于植物基因数据库相对较小,EST 研究中有相当一部分为未知基因,如拟南芥 1993 年所测的 1 152 个 EST 中,68%为新基因[15];1997 年约 1.5 万个不重复 EST 中 60%为未知的[16]。拟南芥种子中约 40%的 EST 与 dbEST 数据库中的序列无同源性,推测可能是种子特异表达基因。水稻 EST 研究,1997 年所测 EST 中约有 75%为新基因[17],1998 年未知部分仍为 75%[18],2003 年中国杭州华大基因研发中心所测水稻的 86 136 条 EST 中有 52%为新基因[19]。随着植物基因功能鉴定工作的不断进行,利用同源比较分析基因功能将成为未来规模化鉴定基因功能的主要方式之一。

Vande Loo 等[20] 第 1 次成功地应用 EST 方法分离到蓖麻油酸生物合成的关键酶基因。黄骥等[21] 以来源于水稻盐胁迫 cDNA 文库的 1 个 500 bp 的 EST S121 为信息探针,搜索位于 GenBank 的水稻 EST 库,发现有 2 个 EST 与 S121 部分序列一致,经过拼接组装获得了 1 个 886 bp 的全长 cDNA 序列,同源性比较结果表明,其可能编码一个新的水稻锌指蛋白基因。Meyers 等[22] 经基因序列结构的生物信息学分析发现了多数水稻的抗病基因具有 NBS 结构,这为发现水稻的抗病基因奠定了基础。通过序列结构分析发现,抗性基因在基因组演变中是比较保守的[23]。Rushton[24] 等利用含 1 500 个 EST 的基因芯片在小麦上鉴定了 117 个抗 Cochliobolus cae-boum 的基因。

3.2 利用基因芯片发现新基因

应用基因芯片技术,对比每一个测定基因和已知基因,即可进行新基因的寻找和表达的监测。Aharoni 等[25] 从草莓中分离了 1 701 个 cDNA 片段,与 480 个矮

牵牛植物克隆做对照,构建成微阵列芯片来研究草莓果的不同成熟时期果色与成熟度的关系。以此方法发现并证实了有两个基因是新基因,其中之一是草莓乙酰基转移酶基因(SAAT)。在成熟的草莓中,此酶在特殊化合物的合成过程中有很重要作用,同时多种水果中都有这种酶的存在。Aharoni 还发现红色果实比白色果实 SAAT 的表达活性高。Guteraman 等[26]也利用 DNA 芯片从玫瑰花有气味和没有气味的四倍体栽培植株中鉴定了新的与香味相关的基因。全长的 cDNA 对植物基因的功能分析是必需的。Seki 等[27]利用生物素化的 CAP 捕捉方法构建了不同条件下拟南芥植物的全长 cDNA 文库,例如在干旱、寒冷、失重的植物从萌发到成熟种子的不同发育阶段。利用拟南芥 1 300 个全长的 cDNA 制备了 cDNA 微阵列以鉴定干旱和寒冷诱导的基因和 DREB1A/CBF3(转录因子,控制胁迫环境下的基因表达)的靶基因。分别分离了 44 个干旱诱导基因和 19 个寒冷诱导基因的 cDNA,其中 30 个干旱诱导基因和 10 个寒冷诱导基因是以前没有报道过的。12 个压力诱导基因被鉴定为 DREB1A 的靶基因,它们中的 6 个是新基因。在 RNA 点杂交和微阵列分析的基础上,6 个基因被鉴定为受 DREB1A 控制的新的干旱和寒冷诱导基因。

植物细胞的分裂周期是受众多基因调控的复杂过程,其中由 M 期进入 S 期受 E2F 转录因子的调控。Vandepoele 等[28]用基因芯片技术对 E2F 转录因子所调控的基因进行了研究,结合生物信息学方法,共发现了 70 个与之相关的新基因。Misson 等[29]用基因组芯片研究了拟南芥在磷元素缺乏条件下的基因表达谱变化发现对磷缺乏起协同诱导表达的基因有 612 个,对磷缺乏敏感而被抑制的基因有 254 个,经进一步分类比较后得知这些基因都是新基因。用该技术在植物中还发现了许多与抗旱、抗寒、抗病、高产等相关的新基因,为人们充分认识和利用植物资源提供了方便、快捷的工具。在水稻基因芯片的研究方面上已经完成了水稻 4 号染色体特异 DNA 芯片的研制,系统开展了水稻 4 号染色体基因特异表达谱分析,从而发现了一批新的表达基因。

3.3　从基因组 DNA 序列中预测新基因

由于几个重要模式植物基因组测序的完成以及重要经济作物测序的进行,使得植物生命科学的研究出现了繁荣的景象。从基因组序列预测新基因,本质上是把基因组上编码蛋白质的区域和非编码蛋白质的区域区分开来。对于理论方法来讲就是要找到在编码区和非编码区哪些数学、物理学特征是不一样的。将这些序列与已知基因数据库进行比较,就可以发现新的基因了。发现了新基因就会对生命活动的认识加深一步。

这种方法实质上是把基因组中编码蛋白质和非编码蛋白质的区域区分开来,将这些序列与已知基因数据库进行比较,就可以发现新的基因。廖问陶等[30]根据拟南芥 Toc33 基因的编码序列,设计一对 PCR 引物,扩增诸葛菜叶体外膜蛋白转运器构件蛋白基因 Toc33,最后得到 2 个片段,这 2 个片段与拟南芥 Toc33 基因有较高的同源性,分别命名为 OvToc33-1,OvToc33-2,进而得到了诸葛菜 Toc33 基因的全编码区序列。

3.4　通过保守区来发现和克隆基因

蛋白质家族和超家族最初是根据序列类似性大小(主要指残基因相同率)来划分的。虽然,它对同源蛋白质的一级结构、功能和进化研究曾起到一定作用,但是,对嵌合蛋白质和多结构域蛋白质等复杂蛋白质的序列分析表明,蛋白质之间的相似性可能只局限于某个序列区域或结构域,而全序列的"平均"类似性可能很小;一个序列可能被指定这一个以上的蛋白质家族。因此,现在蛋白质超家族的概念是具有某种共同结构域的所有分子组成的分子集合。例如所有具有免疫球蛋白样结构域的分子组成免疫球蛋白超家族,而类似补体成分 C1r、C2 和补体因子 B、H 等 60 个氨基酸的内部同源单位(结构域)的所有分子组成补体超家族。这种定义似乎较好地反映蛋白质结构、功能和进化的关系。

一旦来源于某一个种属的一个特定的蛋白质与某一蛋白质家族达到统计学相关的标准,则来源于其他物种的这种蛋白质就自动地被认为是该蛋白质家族的成员,除非有别的证据能否定这个推论。一个序列可能被指定为一个以上的蛋白质家族,而蛋白质之间的相似性可能只局限于某个序列区域或结构域。可以通过这个区或结构域在遗传和进化上的保守性来发现或克隆同一家族的其他成员(包括不同物种的)。有以下 2 种方法:①可以通过经同源蛋白质的保守区来设计引物,从另一物种中克隆出同一家族的蛋白基因。基于结构相似暗喻功能相近的假设,就可以分离到已知功能线索的基因。②对某些与人类或其他动物(尤其是灵长类和其他哺乳动物)的某个基因(或基因家族中的保守区段)有高度同源性的基因,可将同源基因或保守的区段进行电子筛库,得到人类的 EST 信息,进一步进行拼接、延伸而有可能得到全长的 cDNA 序列信息[31]。

4　展望

生物信息学的特点是投资少、见效快、效益大,适合于我国的现实条件。从英特网上源源不断地采集数据,进行分析、归类与重组,发现新线索、新现象和新规律,用以指导试验工作的设计,即直接从现有公共数据库中的 EST 出发,用生物信息学的方法寻找可能有研究价值的新基因,并用试验方法来研究证实。这是一条既快又省的科研路线,可避免不必要的重复,少走弯路,尽快提高我国生命科学的研究水平。

对农业方面来说,不可能也不应该照搬国内外医药

行业克隆新基因所用的方法,而应该走生物信息学和定位克隆相结合的道路。具体说就是一方面进行各种遗传疾病家系、种质资源的采集,从家系分析入手寻找致病基因在染色体上的位置,然后对这个区域进行测序,再利用生物信息学的手段预测候选基因和它的功能,并用试验加以证实;另一方面直接从现有有关农业动植物公共数据库中的EST出发,用生物信息学的方法寻找可能有价值的新基因并用试验方法来研究证实,这种双管齐下的克隆新基因的方法可能更适合我国农业院校从事农业生物信息学的客观条件。

参考文献

[1] 陈润生.生物信息学及其研究进展[J].医学研究通讯,2002,31(12):2-5,26.

[2] Baldi P, Brunak S. Bioinformatics: The Machine Learning Approach[J]. Cambridge Mass: MIT Press, 2001.

[3] Goodman N. Biological data becomes computer literate: new advances inbioinformatics[J]. Curr Opinin Biotechnol, 2002, 13(1): 68-71.

[4] 新基因丰富植物生命内涵[EB/OL]. 中国农业网,2006-2-17.

[5] 新基因增产潜能巨大-级稻研究突破在望.央视国际,2001-1-8.

[6] 我国水稻产量相关功能基因研究取得重大进展[EB/OL]. http://www.edu.cn.2007-04-10.

[7] Haldane J B S. The cause of evolution[M]. London: Longmans and Green, 1932.

[8] Muller H J. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere[J]. Genetics, 1935, 17: 237-252.

[9] Ohno S. Evolution by Gene Duplication[M]. German: Springer-Verlag, 1970.

[10] Gilbert W. Why genes in pieces[J]. Nature, 1978, 271: 501.

[11] Gilbert W. The exon theory of genes[C]// Cold Spring Harbor Symposia on Quantitative Bilology, 1987, LII: 901-905.

[12] Long M, Langley C H. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila[J]. Science, 1993, 260: 91-95.

[13] Long M, Betran E, Thornton K, et al. The origin of new genes: glimpses from the young and old[J]. Nat Rev Genet, 2003, 4(11): 865-875.

[14] Kent W J. Blat-Blast-like alignment tool[J]. Genome Res, 2002, 12(4): 656-664.

[15] Hoog C. Isolation of a large novel mammalian genes by a differential cDNA library screening strategy[J]. Nucleic Acids Research, 1991, 19(22): 6123-6127.

[16] Delseny M, Cooke R, Raynal M, et al. The Arabidopsis thaliana cDNA sequencing projects[J]. FEBS Lett, 1997, 403(3): 221-224.

[17] Yamanoto K, Sasaki T. Large scale EST sequencing in rice[J]. Plant Molecular Biology, 1997, 35: 135-144.

[18] Femandez P, Heinz R, Paniego N, et al. Characterization of sunflower ESTs from leaf and developing flower[C]// Plant Animal Genome IX Conference, 2001: 67.

[19] Zhou Y, Tang J B, Walker M G, et al. Gene ideniifieation and expression analysis of 86136 expressed sequence tags(ESTs)from the rice genome. Geno[J]. Prot. Bioinfo, 2003(1): 26-42.

[20] Vande Loo F, Broun P, Turner S, et al. An oleate12-hydroxylase frome Ricinus communisL. is a fatty acyllesat-uraschomolog[J]. Proceeding of Natural Acdaemic Science of USA, 1995, 38, 45-59.

[21] 黄骥,张红生,曹雅君,等.水稻功能基因的电子克隆策略[J].中国水稻科学,2002,16(4):295-298.

[22] Meyers B C, Chin D B, Shen K A, et al. The major resistance gene cluster in lettuce is highly duplicated and spans several megabases[J]. Plant Cell, 1998, 10(11): 1817-1832.

[23] 朱小源,杨健源,曾列先,等.基因组学与植物抗病性研究进展[J].广东农业科学,2004(1):37-39.

[24] Rushton P J, Somssich I E. Transcriptional control of plant genes responsive to pathogens[J]. Curr Opin Plant Bio J, 1998, 1(4): 311-315.

[25] Aharoni A, Keizer L C P, Bouwmeester H J, et al. Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays[J]. Plant Cell, 2000(12): 647-661.

[26] Guterman I, Shslit M, Menda N, et al. Rose scent: genomics approachto discovering movel floral fragrance-related genes[J]. The Plant Cell, 2002, 14: 2325-2338.

[27] Seki M, Narusaka M, Abe H, et al. Monitoring the expression pattern of 1300 Arabidopsis genes under draught and cold stresses by using a full-length cDNA microarray[J]. The Plant Cell, 2001, 13: 61-72.

[28] Vandepoele K, Vlieghe K, Florquin K, et al. Genome-wide identification of potential plant E2F target genes[J]. Plant Physiol, 2005, 139: 316.

[29] Misson J, Raghothama K G, Jain A, et al. A genome-wide transcriptional analysis using Arabidopsis thaliana affymetrix gene chips determined plant responses to phosphate deprivation[J]. Plant Biol, 2005, 102(33): 119-134.

[30] 廖问陶.诸葛菜(*Orychophragmus violaceus*)及甘蓝型油菜(*Brassica napus*)Toc33基因编码区的克隆与分析[D].成都:四川大学硕士学位论文,2003.

[31] 王关林,方宏筠.植物基因工程[M].北京:科学出版社,2004.

The Application of the Bioinformatics in Finding New Genes

WEI Xiao-chun, ZHENG Qun

(Horticultural Department, Shihezi University, Shihezi, Xinjiang 832000, China)

Abstract: The finding of the new genes is of great important to the research and development of the plant science. For new genes could be applied to enhance the resistance, improve the output, better the quality of the plant etc. Thus, bioinformatics plays a large role all the time in the methods of finding new genes.

Key words: New gene; Bioinformatics; Resistance; Output; Quality