

专 论

大豆的基因组研究

刘 峰 陈受宜

(中国科学院遗传研究所植物生物技术实验室, 北京 100101)

大豆起源于中国, 是具有重要经济价值的油料作物, 也是食物中植物蛋白质的主要来源。其种子约含 40% 蛋白和 20% 脂肪。在国际贸易市场上, 大豆的脂肪含量在主要油料作物中位居第一。目前, 大豆已成为一种国际性作物, 在美国、巴西、阿根廷、中国和印度等国家广为种植, 其中美国的大豆产量最高(Singh 等, 1999)。

大豆的基因组约为 100Mb, 其中 60% 左右是重复序列(Gurley 等, 1979)。其大小大约是拟南芥基因组的 715 倍, 水稻的 215 倍, 然而小于玉米基因组的一半, 为小麦的 1/14(Arumanagathan 和 Racle, 1991)。

遗传图谱的构建

十几年前我们对大豆基因组的了解还非常少。尽管已有 250 个形态和同功酶标记, 但大部分不能定位。在分子图谱建成前, 大豆中只有 17 个经典连锁群, 包含 57 个标记, 长度为 (420cM (Palmer 和 Keim, 1989)。大多数连锁群为两点或三点连锁。这与大豆的经济地位是不相称的。

现在, 大豆基因组的研究与以前的状况已大不相同, 甚至可与许多模式植物相媲美。大豆基因组的时代是从发展分子遗传图谱开始的。遗传连锁图在许多研究领域包括系统进化、分子标记辅助选择和基因的图位克隆等都发挥了巨大的作用。因此, 对于得到广泛研究的重要经济作物) 大豆来讲, 发展一张高密度的分子遗传图谱, 其上包括许多具有可分辨表型效应的经典标记, 同功酶标记和大量高信息度的 DNA 标记并且均匀分布在整个基因组中, 将具有重要意义(表 1, 2)。

在过去的十年中, 应用美国北方和南方的大豆种质资源, 分别发展了十多张遗传图谱, 所用的作图群体有种间和种内杂交得到的群体、F2 群体和重组自交系群体(Recombinant Inbred Lines, RIL)。

1988 年, Apuya 等应用大豆两个栽培品种 Min2 soy 和 Noir 1 杂交得到的 F2 群体构建了第一张大豆 RFLP 图谱, 共有 11 个 RFLP 标记分布在 4 个连锁群上。随后, Keim 等(1990)应用 G. Max 和 G. Soja 进行种间杂交, 构建了第二张 RFLP 图谱, 该图谱包括 130 个 RFLP 标记组成的 26 个连锁群, 覆盖

表 1 拟南芥、水稻和大豆的基因组研究进展

	拟南芥	水稻	大豆
染色体	n= 5	n= 12	n= 20
基因组大小	130Mb	430Mb	1100Mb
已定位的分子标记	1000	2275	1387
分子标记类型	RFLP, CAPS, SSR, AFLP, RAPD	RFLP, SSR, RFLP, SSR, AFLP, RAPDAFLP, RAPD	
EST	38, 000 (24, 000 non2redundant sequences)	35, 000(3200 个已定位)	约 20, 000 个
已完成的基因组序列	84Mb	/	/
重复序列比例	较少	50%	40-60%
物理图谱覆盖范围	几乎全部(YAC)	70%(YAC)/	
图位克隆的重要基因	12 个	5 个	无
转化方法	农杆菌	基因枪, 聚乙二醇(PEG), 农杆菌	农杆菌
数据库网址	http: M genome2www. stanford. edu/ Arabidopsis	http: Mwww. staff. or. jp	http: M129. 186. 26. 94

表 2 大豆遗传图谱的发展

时间	作者	作图群体	连锁群	总标记数	总长度(cM)
1988	Apuya et al	F2	4	11	
1990	Keim et al	F2	26	130	1200
1993	Shoemaker & Olson	F2	25	383	3000
1993	Lark et al	F2	31	132	1500
1995	Shoemaker & Specht	F2	26	138(104)	
1997	Keim et al	RIL	28	840	3441
1999	Cregan et al	F2 (2), RIL	20	1423	
1997	张德水等	F2	20	71	1446.8
1999	刘峰等	RIL	22	240	37131.5

大豆基因组 1200cM。Shoemaker 和 Olson(1993)同样使用 G. Max @G. Soja 杂交的 F2 群体, 发展了含有 25 个连锁群的遗传图谱, 包括 365 个 RFLP 标记, 11 个 RAPD 标记, 3 个经典性状标记和 4 个同功酶标记。大豆经典图谱共有 68 个位点分布在 20 个连锁群上, 每个连锁群上只有几个位点(Palmer 和 Shoemaker, 1998)。1995 年, Shoemaker 和 Specht 将来源于不同群体的各种标记整合到一张图谱上。他们用近等基因系 Clark @Harosoy 杂交得到的 F2 群体构建了一张含有 13 个经典标记, 7 个同功酶标记, 110 个 RFLP 标记和 8 个 RAPD 标记的遗传图谱。通过使用一套锚定的 RFLP 探针, 根据其在上述群体和 G. Max @G. Soja 两个群体中的分离, 确定了分子和经典图谱的同源连锁群。Keim 等(1997)应用 BSR- 101 @PI437.654 构建的含有 300 个单株的 RIL 群体, 构建了 RFLP 框架图, 然后选取其中 42 个 RIL 株系进行 AFLP 标记的遗传作图, 构建了含有 28 个连锁群的遗传图谱, 长度为 3441cM。共有 165 个 RFLP, 25 个 RAPD, 650 个 AFLP 标记。尽管 AFLP 经常成簇排列, 相对于其它分子标记, AFLP 更均匀地分布在所有连锁群中。

1999 年, Cregan 等应用 606 个 SSR 标记, 同时对三个作图群体, USDA/ IOWA G. max @G. soja 的 F2 群体, Univ. Of Utah Minsoy @Noir 1 的重组自交系和 Univ. Of Nebraska Clark @Harosoy 的 F2 群体, 进行作图。大多数 SSR 标记在两个或三个群体的亲本中都具有多态性。这些图谱上总标记数(分子标记, 同功酶和经典标记) 分别为 703(G. max @G. soja), 539(Minsoy @Noi 1) 和 460(Clark @Harsoy)。由于 SSR 位点的特异性, 因此可以将三个群体分别

构建的遗传图谱整合成一张遗传图谱, 含有 20 个同源连锁群, 对应于大豆单倍体基因组中的 20 条染色体。大豆中报道的全部经典连锁群除了 CLG06 以外都整合进这张图谱中。综合这三张图谱, 特异位点的总数为 1423 个, 其中有 606 个 SSR, 689 个 RFLP, 79RAPD, 11 个 AFLP, 10 个同工酶和 26 个经典标记。

发展这些遗传图谱的同时, 也发展了各种类型的分子标记。现已定位了几百个 RFLP 探针和 RAPD 标记(Lark 等, 1993; Rafalski 和 Tingey, 1993; Shoemaker 等, 1995), 600 多个 AFLP 标记(Keim 等, 1997) 和大约 600 多个微卫星标记(Cregan 等, 1999)。

基因组的比较遗传作图研究发现禾谷类具有高度保守的基因组结构, 据此提出禾谷类具有一个共同的祖先基因组假说。这在基因组信息从一个禾谷类品种到另一个品种的交流方面具有重要意义(Bennetzen 和 Freeling, 1993)。然而豆科植物中的比较作图则较为复杂。大豆在二倍化过程中发生的大量 DNA 重排事件使其很难鉴定其与相关豆科植物间具有共线性的染色体区段(Boutin 等, 1995)。绿豆(2n= 11)和菜豆(2n= 11)都属于 Phaseolus 属, 表现出高度的保守性和标记间顺序的相似性, 绿豆的大多数连锁群含有一个或几个来自菜豆的连锁区段, 但与大豆(2n= 20)比较时则大不相同, 前二者的连锁群上通常只有较短的大豆连锁区镶嵌其中。

大豆是一个古四倍体

尽管使用了不同的群体和各种类型的分子标记发展大豆遗传图谱, 但在进行不同群体的图谱比较时, 仍存在一些模糊的问题。这主要是由基因组本身的复制和 RFLP 探针通常只能检测众多复制位点中的一个引起的。对大豆基因组的内在结构进行分析, 揭示了一些关于基因组进化和基因表达的有趣现象。

很明显, 远祖时期, 大豆发生了某种类型的多倍体化。大豆二倍体的染色体数目为 2n= 40, 然而 Phaseolae 亚科的大多数属为 2n= 22。Lackey (1980)据此推断, Glycine 可能来源于一个二倍体的祖先(n= 11), 随后产生非整倍体的丢失导致 n= 10, 进而多倍体化成为现在的 2n= 40。如下所示:

2n= 22 $\xrightarrow{\text{染色体丢失}}$ 2n= 20 $\xrightarrow{\text{多倍化}}$ 4n= 40 $\xrightarrow{\text{二倍体化}}$ 2n= 40

然而, 尽管其为多倍体, 大豆基因组的绝大部分

在行为上仍表现为二倍体。多倍体的/二倍化0是一个典型进化事件,通常是由于增加、缺失、突变和重排等快速抑制了连锁群的非同源配对引起的(Ohno, 1970)。有许多证据证明大豆实际上是一个古四倍体。

在大豆种质库中,发现了许多重复基因,即两个独立的基因控制相同的性状(Palmer 和 Kilen, 1987)。例如: pal/pa2(软毛的直立和平贴生长), pd1/pd2(正常/密集的软毛), d1/d2(种胚色)和 y7/y8(绿/黄簇叶)。根据它们在 F2 群体中呈现的 15B 1 分离比,确定它们为重复基因。另外,对基因组的分子遗传分析提供了基因组复制的最直接的证据。Shoemaker 等(1996)对 280 个随机挑选的 Pst I 基因克隆与完全酶切的大豆基因组 DNA 进行 Southern 分析得到的平均杂交条带数目的分析显示了基因组中大量的复制现象。有 90% 以上的探针检测到 2 条以上的条带,近 60% 检测到 3 条或以上的条带。这说明基因组中仅为单拷贝序列的部分不到 10%。基余大部分可能不仅四倍化,还经历了基因组复制。对基因组中复制片段线性结构的全面分析和不同来源数据的整合也支持这一观点。

对来自 9 个不同群体的 800 多个标记综合分析得到一张整合图谱,通过 RFLP 杂交技术检测到广泛的同源关系(Shoemaker 等, 1996)。复制片段平均长度为 45cM,个别的超过 100cM。复制区段平均复制了 2155 次,有的达到 6 次。另外,还发现/巢式0复制,说明其中大豆的一个原始基因组可能在进化过程中又进行了额外的四倍体化。

大豆基因组中序列的复制,不能简单地解释为四倍体化事件。例如,对于任一连锁群,其上定位的标记也可以在其它多个连锁群中检测到。一个中等大小的连锁群上包含能在另外 8 个连锁群上同时检测到的标记。这说明,四倍体化和片段复制可能不是产生基因组复制的唯一机制。

如上所述,由于基因组序列的复制,在整合 RFLP 数据时产生了困难。因为一个 RFLP 探针在基因组中可检测到多个位点。在一个群体中用某一探针检测到的多态性片段代表了一个位点,而相同探针在另一群体中检测到的多态性片段可能代表了另一个完全不同且不连锁的位点。这与不同群体中哪一个位点具有多态有关。虽然通过使用/锚定0的探针/酶组合(特定的探针/酶组合确定了相同的多态性位点,从而在图谱上定位在相同的位置),可以克服这一困难,然而最终解决此问题,只有发展信息

度高且只有单一位点的分子标记))) 微卫星标记(刘峰等, 1998)。

如上所述, Cregan 等(1999)根据 600 多个单一位点的微卫星标记在 2 到 3 个群体的分离数据,将连锁群整合成一张公共图谱,含有 20 条连锁群,覆盖长度为 2750cM。包含预期的 20 个连锁群的遗传图谱的建成将成为大豆基因组研究中重要的里程碑。

农艺性状基因的定位

分子遗传图谱只有能准确定位基因或 QTL,才能得到充分利用。通常一次试验只能分析一个性状。然而 Shoemaker 等(1995)在一次实验中定位了 18 个基因,这是通过在用以作图的亲本中积累突变完成的。由于大部分基因已定位在经典连锁群上,通常多数经典连锁群在一次试验中就能整合在相应的分子遗传图谱上。目前,已定位了 60 多个质量性状位点。

除了许多抗病基因或/质量性状0基因外,大多数重要农艺性状是由几个或多个基因共同控制的(数量性状)。控制这种性状的基因座位称为数量性状基因位点(quantitative trait loci, QTL)。由于大多数数量性状存在遗传与环境的互作,数量性状的育种需要在多个地点进行两年或更长时间的田间重复实验。显然这个过程耗时费力,而分子标记对主效 QTL 的定位可使我们根据标记的分离数据推测 QTL 的存在。这对于育种来说,既省时间,又节约经费。基因/发现0(discovery)和 QTL 定位是这种类型的标记辅助选择(MAS)的前提。

1990 年发表了大豆中第一篇 QTL 定位的报道(Keim 等, 1990)。此后,相继报道了与众多数量性状相关 QTL 的定位。这些性状包括生殖性状、种子特征、种子成分、突变性状、生长形态性状、抗病和营养效率等。

对这些文献的分析有一些有趣的发现。QTL 并不是随机分布于基因组中,它们经常是位于特定的连锁群和特定的区域。例如, L 连锁群包含 18 个性状的 QTL 位点,而其它连锁群上则没有这些性状的定位报道。另外,相关的 QTL,如种子性状或抗病性等,可能成簇定位于某一特定区域。Imsande 等(1998)报道了 8 个不同的种子特征 QTL 位点位于一个连锁群,而另一个连锁群则有 9 个与大豆不同抗病性相关的 QTL。

大豆 BAC 库的构建和物理图谱

高密度图谱的构建、目的基因的发现和定位使得图位克隆法成为可能。该法不仅需要相邻标记之间的精细图谱,还要提供大片段 DNA 文库用以连接相邻的分子标记。已经建成几个覆盖大豆整个基因组的 BAC 文库 (Marek 和 Shoemaker, 1997; Danesh 等, 1998)。应用这些文库已构建了目的基因附近详细的物理重叠群(contig)(Marek 和 Shoemaker, 1997)。并且这些文库来自不同的基因型,使用不同的限制性内切酶,可方便地进行科研交流。

Marek 和 Shoemaker (1997) 以栽培品种 Williams82 为材料,纯化的提取的高分子量 DNA (HMW)经 HindⅢ 部分酶切后,与 PBI0BAC11 连接后进行转化,构建了一个含有 40000 个克隆的 BAC 文库(相当于 4~5 倍基因组);对 224 个随机克隆进行分析,平均插入片段长度为 150kb,覆盖任一特定序列的可能性为 98%。用叶绿体基因 psbA 进行检测,表明叶绿体 DNA 所占的比例小于 1%,用抗病基因类似物(RGA)克隆筛选该文库,构建了相应的 BAC 克隆重叠群(contig)。并将其中第一类 contig 定位在 J 连锁群含有几个已知抗病基因所在的区域。该重叠群长度为 400kb,共有 9 个克隆。

随后, Danesh 等(1998)以栽培大豆 Faribault 为材料,纯化的 HMW DNA 经 EcoRⅠ 部分酶切后与 PECSBAC4 (Frijters 等, 1997) 连接后,转化 DH10B 感受态细胞,构建了含有 30,000 个克隆的 BAC 库(相当于 3 个基因组大小),平均插入片段为 120kb。其中 55%的克隆含有高拷贝序列,25%含有中拷贝序列,20%的克隆只含有低拷贝序列。应用与大豆抗孢囊线虫主效基因紧密连锁的两个分子标记, RFLP 标记 B053T (117cM) SSR 标记 BARC Satt309 (0.4cM)成功地筛选到了几个相应的 BAC 克隆,并定位到图谱中预期的位置。这些克隆组成三个不连续的重叠群,覆盖了目的基因附近 365kb 的区域,相当于 319cM (Concibido 等, 1996)。目前他们正在填补克隆之间的缺口,并扩大作图群体的规模,期望能找到与该基因连锁更紧密的分子标记。对四个克隆的端部和中部进行 DNA 亚克隆后, RFLP 分析证明它们共分离,间距小于 1cM,说明没有嵌合现象。另外也发现该区域存在高比例的重组,因为在这些克隆的端部和中部亚克隆之间,应用两个含有 218 个单株的作图群体可以检测到发生了交换。

由于这两个 BAC 库是用不同的酶切构建的,因

此可以互补,二者共代表了大豆 7~8 个基因组大小,适合于进行任何一种图位克隆目的基因的尝试。

另外, Kasuga 等(1997)应用 AFLP 标记找到了与大豆抗根腐病基因 RPS12k 紧密连锁的两个标记, TC1 0107cM RPS12k 0.06cM CG1。HMW DNA 经多种酶切后与这两个探针杂交发现,二者有一个共同的条带,小于 125kb,说明这两个标记之间的物理距离小于 125kb。已构建了一个含有该抗性基因的 BAC 库,准备用这两个标记筛选含有 RPS12K 的 BAC 克隆,直接用于染色体登录(Chromosome Landing),从而最终克隆该抗性基因。

大豆的 EST 计划及其进展

大豆 EST 计划的目标是克隆和测定 300,000 个序列。由于植物不同部位和不同发育阶段包含不同的基因信息,因此必须从不同发育阶段和不同器官来源中进行克隆和测序。由于将产生大量的信息,该计划也将同时发展相应的软件项目以帮助整理有价值的信息。该计划的具体目标如下:

- 1 组织科学家利用大豆遗传资源共同研究以促进大豆工业的发展。
- 2 发展 cDNA 文库并进行测序,最终完成 300,000 个序列测定。
- 3 应用 cDNA 序列扩大和提高现有大豆基因组计划的范围和有效性,促进发现与鉴定基因。
- 4 找出更新的策略用以鉴定/标准 0 的 EST 计划中不能鉴定的基因。
- 5 发展基于计算机的方法,以有效查询、分类和正常使用该项目产生的大量原始数据
- 6 以最快的速度提供有关该项目的研究结果(如序列信息,克隆等)

现已构建完成了 16 个高质量的大豆 cDNA 文库,以产生高质量的序列。到目前为止,已测定了 2 万多个 cDNA 序列。计划每年完成 100,000 个序列测定,最终完成 300,000 个序列。

另外,公共图谱上定位的约 250 个 RFLP 标记(基因组克隆)也已测序完成。发现一半以上与其它材料中已克隆的基因相关,并且发现一些有趣的基因,如抗病基因,一个编码磷酸转移酶的基因和一个编码硝酸盐转移酶的基因。此外,还发展了生物信息学工具以帮助分析 EST 数据,结合大豆不同部位和不同环境条件下表达的基因信息,整合大豆 DNA 序列和遗传图谱与其它材料的相关数据。

大豆细胞有丝分裂时染色体的压缩,使其很难

进行细胞遗传学分析。虽然, 通过对粗线期染色体分析, 已建成完整的细胞遗传图谱, 即核型(Singh 和 Hymowitz, 1988)。但是仍难以将分子连锁群定位于染色体上。初级三体($2n=41$) 可用于快速把目的基因定位到特定的染色体上, 并且完成连锁群与染色体的整合。大豆全套的初级三体已构建完成(Xu 等, 1998)。这将有利于经典和分子遗传图谱的整合。

大豆基因组的内在结构仍然是一个谜, 但现在正渐渐得到阐明。基因组研究将告诉我们更多数量性状的进化、多倍体进化和复制遗传因子表达等多方面的知识。

参考文献

[1] Singh RJ, and Hymowitz T. 1999. *Genome* 42: 605- 616

[2] Gurley WB, Hepburn AG and Key JL, 1979. *Biochimica et Biophysica Acta* 561: 167- 183

[3] Arumanagathan K. and Racle ED, 1991. *Plant Mol. Biol. Rep.* 9: 229- 241.

[4] Palmer RG and Kilen TC 1987. Qualitative genetics and cytogenetics. In: Wilcox JR(ed.) *Soybeans: Improvement, Production, and Uses*. 2nd edn. No. 16. American Society of Agronomy. Inc. , Crop Science Society of America, Inc. , and Soil Science Society of America, Inc. , Madison, Wisconsin, pp. 135- 209.

[5] Apuya NR, Frazier, Keim P, Roth EJ and Lark KG. 1988. *Theor Appl Genet*, 75: 889- 901.

[6] Keim P, Diers BW, Olson TC and Shoemaker RC 1990. *Genetics* 126: 735- 742.

[7] Shoemaker, R. C. and Olson, T. C. , 1993. Molecular linkage

map of soybean [*Glycine max* (L.) Merr.]. In: O'Brien SJ (ed.) *Genetic maps*. Cold Spring Harbor Laboratory Publisher. Cold Spring Harbor, New York, pp. 6. 131- 6. 138.

[8] Palmer RG and Shoemaker RC. 1998. *Soybean genetics*. p. 45- 82. In: M. Hrustic, M. Vidic and D. Jockovic(ed.) *Soybean Institute of Field and Vegetable Crops*. Novi Sad, Yugoslavia.

[9] Shoemaker RC, and Specht JE(1995). *Crop Sci* 35: 436- 446

[10] Keim P, 1998. *Crop Sci.* 37: 537- 543

[11] Cregan PB, 1999. *Crop Sci.* In press

[12] Lark KG, *Theoretical and Applied Genetics* 86: 901- 906.

[13] Rafalski JA. *Trends Genet.* 9: 275- 279.

[14] Bennetzen J, Freeling M. 1993. *Trends in Genetics* 9: 259- 261

[15] Boutin SR, 1995. *Genome* 38: 928- 937

[16] Lackey J. 1980. *American Journal of Botany* 67: 595- 602

[17] Ohno S. 1970. *Evolution by Gene Duplication*. Springer verlag, New York.

[18] Shoemaker RC, (1996) *Genetics* 144: 329- 338

[19] 刘峰, 陈受宜. 1978. *大豆科学* 17(3): 256- 261

[20] Imsande M, 1998. *Soybean Genetics Newsletter* 25: 146- 148

[21] Marek LF, *Genome* 40: 420- 427

[22] Danesh D, *Theor. Appl. Genet.* 96: 196- 206

[23] Frijters ACJ, 1997. *Theor. Appl. Genet.* 94: 390- 399

[24] Conclbido VC, 1996. *Crop Sci.* 36: 1643- 1650

[25] Concibido VC, *Theor. Appl. Genet.* 93: 234- 241

[26] Kasuga T, Salimath SS, Shi J, Gijzen M, Buzzell RI, and Bhat2 tacharyya MK. 1997. High resolution genetic and physical mapping of molecular markers linked to the *Phytophthora* resistance gene *Rsp12k* in soybean. *MPMI* 10B1035= 1044

[27] Singh RJ *Theoretical and Applied Genetics* 76: 705- 711.

[28] Xu SJ, Singh RJ, *Soybean Genet. Newslet.* 24: 121- 122