大豆油脂储存蛋白的生物信息学分析

王巍杰,吴 丹,王金朋

(华北理工大学 生命科学学院,河北 唐山 063000)

摘 要:大豆种子含油量高低和油脂合成途径密切相关,油脂合成途径复杂,涉及诸多蛋白和酶,为此对大豆油脂储存蛋白进行生物信息学分析。大豆全基因组数据下载于 JGI 数据库,生物数据库查询结合 Perl 程序处理获取大豆油脂储存基因和蛋白,在大豆基因组中确定1 264个与油脂合成相关的基因,其中 23 个基因与油脂储存有密切的联系。利用 Protparam、SOPMA、ProtComp、SignalP 软件对 23 个基因的蛋白序列、蛋白基本理化性质及二级结构、亚细胞定位、信号肽等进行生物信息学分析。结果表明:23 个油脂储存基因不均匀分布在 12 条染色体上;23 个蛋白序列氨基酸数目为 165 ~ 1 012个;等电点为 5. 90 ~ 10. 03;外显子数目为 5 ~ 16 个;二级结构预测显示无规则卷曲和 α - 螺旋为主要构成成分;蛋白亚细胞定位主要位于内质网、质膜和胞外。用 MEGA6 软件内置的 Clustal W 程序对大豆中油脂储存基因的蛋白序列进行比对分析,采用邻接法(neighbor-joining,NJ)构建系统发育树,结果显示大豆油脂储存基因的亲缘关系和进化差异。

关键词:大豆;油脂储存;生物信息

中图分类号:S565. 1 **文献标识码**:A **DOI**:10. 11861/j. issn. 1000-9841. 2016. 02. 0234

Bioinformatics Analysis of Storage Oil Protein in Soybean

WANG Wei-jie, WU Dan, WANG Jin-peng

(College of Life Science, North China University of Science and Technology, Tangshan 063000, China)

Abstract: There are closed relationships between the quantity of oil storage and oil synthesis pathway in soybean seed, oil synthesis pathway is complicated process involved in many proteins and enzymes. Storage oil proteins, one of the important proteins, were analyzed by bioinformatics method in the paper. The genome of soybean was download from JGI database, the storage oil gene and related proteins were obtained by database searches and Perl program data mining, 23 genes closely related to oil storage were identified from 1 264 storage oil genes in soybean. By using software such as Protparam, SOPMA, ProtComp and SignalP, the features of protein sequences, physical and chemical properties, secondary structure, subcellular location, and signal peptide were analyzed. Bioinformatics analysis indicated that 23 oil storage genes maps unevenly on 12 soybean chromosomes. Twenty-three oil storage proteins vary in compositions of amino acid (165 and 1 012), isoelectric points (5. 90 and 10. 03), and exons (5 and 16). Secondary structure prediction results showed that alpha helix and random coil were common structures. Those proteins were located on endoplasmic reticulum, plasmalemma and extracellular. Phylogenetic tree were constructed by neighbor-joining methods basing on the multiple sequences alignments of Clustal W provided by MEGA6 software. Analysis of phylogenetic tree showed that the relationships of 23 storage oil genes and their differences in revolutionary history.

Keywords: Soybean; Storage oil; Bioinformatics

大豆(*Glycine max*)是我国重要的蛋白质和食用油来源,其用途非常广泛,某些成分具有药用功能,用于抑制肿瘤生长、调节高血压和其它心血管疾病[1-2]。大豆平均含油量为 20%,每年约有 86% 的大豆油用于人类食物和动物饲料,其余的 14% 用于如能源、肥皂、润滑油、化妆品、墨汁、粘合剂等行业[3]。随着植物油需求量的增加和消费者对膳食脂肪安全意识的提高,大豆种子的含油量和质量受到了遗传育种工作者的关注[4-5]。

植物种子中的油脂合成过程非常复杂,涉及许多关键酶和基因。在大豆进化过程中,不断受环境变迁繁衍至今,形成了适应现代环境的内在机制,其中重要的油脂合成基因家族也形成了一个独有的进化模式。2010 年 1 月 14 日的《Nature》杂志公布了由美国农业部、美国能源部联合基因组研究所和普渡大学等多家科研机构联合完成的大豆完整基因组序列草图[6],对于认识其生物学机制,并在分子水平上改进大豆品质,提高产油量意义重大。目前,植物种子中油脂合成途径已经清楚,在拟南芥基因组中已经确定了 600 多个与油脂合成相关的基因,其中与油脂储存相关的基因有 19 个[7]。Jeremy 等[6]用拟南芥中油脂合成相关的基因与大豆基因组做同源比对分析,发现1 127个与油脂合成相关的基因,与油脂储存相关的基因有 22 个,这些基因及其相关通路对大豆油脂含量有重要的影响,通过对某些基因的修饰和调控,可增加大豆的油脂

收稿日期:2015-06-25

第一作者简介:王巍杰(1967-),女,教授,硕导,主要从事生物质应用和生物信息研究。E-mail: weijiewang@ yahoo. com。

产量[8]。

根据在细胞活动中功能的不同,与油脂合成相关的基因可以分为 9 类[6,9]。本文利用在线工具和生物信息相关方法对大豆种子中油脂储存相关基因进行初步生物信息学研究,以期为提高植物油脂合成品质提供数据基础。

1　材料与方法

1.1　数据材料

大豆全基因组数据下载于 JGI http://phytozome.jgi.doe.gov/pz/portal.html;拟南芥基因序列和蛋白序列下载于 TAIR ftp://ftp.arabidopsis.org/[10];拟南芥中与油脂合成相关的基因信息从数据库 *Arabidopsis* Lipid Gene(ALG) http://lipids.plantbiology.msu.edu/ 获得。

1.2　大豆油脂储存相关基因的鉴定

用 Perl 程序将拟南芥中与油脂储存相关的 19 个基因的蛋白序列和 CDs 序列从拟南芥全基因组中提取出来。以拟南芥中油脂储存基因的蛋白序列作为查询序列,执行 BLASTp(E-value 10^{-5})同源基因搜索,保存结果。对结果进行筛选,筛选条件为:E-value $< 10^{-20}$,score > 200,得到同源基因。用 Perl 程序分别将大豆中油脂储存基因的蛋白序列和 CDs 序列提取出来。

1.3　油脂储存基因的结构分析和基因定位

用 Perl 程序处理大豆基因组数据,获得大豆油脂储存相关基因在染色体上的位置信息,在 NCBI 搜索获得大豆 20 条染色体的长度,用 MapInsepect 软件对大豆油脂储存功能相关基因进行染色体物理定位。

从 Phytozome http://phytozome.jgi.doe.gov/jbrowse/index.html 得到大豆油脂储存蛋白的基因序列,用 Gene Structure Display Sever 2.0[11] http://gsds.cbi.pku.edu.cn/index.php 分析基因内含子、外显子组成。

1.4　大豆中油脂储存蛋白的生物信息学分析

大豆中油脂储存蛋白的基本性质采用 ExPaSy 提供的在线分析工具 Protparam[12] http://web.expasy.org/protparam/,分析其氨基酸数目、等电点、分子量等。二级结构的预测用在线分析工具 SOPMA[13] https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html。用 ProtComp http://linux1.softberry.com/berry.phtml?group=programs&subgroup=proloc&topic=protcompan 对大豆油脂储存蛋白进行亚细胞定位预测分析。用 CBS 提供的在线工具 SignalP http://www.cbs.dtu.dk/services/SignalP/ 对大豆油脂储存蛋白做信号肽分析,参数为默认值。

1.5　油脂储存功能相关基因系统发生分析

用 MEGA6 软件内置的 Clustal W 程序对大豆中油脂储存基因的蛋白序列进行比对分析,采用邻接法(neighbor-joining,NJ)构建系统发育树,Bootstrap 值设为1 000,其它参数为默认值。

2　结果与分析

2.1　大豆油脂储存相关基因信息

BLASTp 比对结果显示:从大豆基因组中找到 23 个与油脂储存功能相关的基因,这 23 个基因的蛋白序列存在着较大的差异,氨基酸数目从 165 ~ 1 012个不等,对应分子量19 197.0 ~ 113 732.9 Da,8 个蛋白属于稳定蛋白;蛋白不均匀分布在 12 条染色体上(图 1)。1、3、7、10、12、19、20 号染色体上各有 1 个蛋白,11、13、17 号染色体上各有 2 个蛋白,16 号染色体上有 3 个蛋白,9 号染色体上有 7 个蛋白(数量最多),且基因在染色体上的位置相对集中。

基因结构分析结果显示:23 个基因中 *Glyma*09*g*07520.2、*Glyma*13*g*16560.1、*Glyma*17*g*06120.1 所含外显子数目最多,为 16 个,*Glyma*09*g*25350.1、*Glyma*10*g*33350.2 含有 5 个外显子,数量最少。

2.2　大豆油脂储存蛋白一级结构分析

利用在线分析工具 Protparam 对大豆油脂储存蛋白的理化特性进行分析,由表 1 可见 23 个蛋白的氨基酸数目有所差距,在 165 ~ 1 012个氨基酸不等,其中 2 个蛋白的氨基酸数目小于 200,7 个蛋白的氨基酸数目在 200 ~ 300 个,5 个蛋白的氨基酸数目在 300 ~ 400 个,8 个蛋白的氨基酸数目在 400 ~ 700 个,*Glyma*09*g*25350.1 编码蛋白所含氨基酸数目为 165 个(数目最少),*Glyma*12*g*08915.1 含有的氨基酸数目1 012个(数目最多)。23 个蛋白对应的分子量在19 197.0 ~ 113 732.9 Da。

等电点分析表明:14 个蛋白的等电点大于 7.5,显碱性,其中 *Glyma*11*g*09411.1 编码蛋白的等电点最大,为 10.03;6 个蛋白的等电点小于 6.5,显酸性,其中 *Glyma*20*g*34300.1 编码蛋白的等电点最小,为 5.9。

脂溶性指数分析表明:19 个蛋白的脂溶性指数小于 100,4 个蛋白的脂溶性指数大于 100。23 条氨基酸的疏水性大部分为负值,表明实验中分析的大豆油脂储存蛋白绝大部分属于亲水蛋白。

不稳定指数分析表明:*Glyma*03*g*41030.1、*Glyma*09*g*22330.1、*Glyma*09*g*22455.1、*Glyma*09*g*25350.1、*Glyma*12*g*08915.1、*Glyma*13*g*16790.1、*Glyma*16*g*21970.1、*Glyma*17*g*05910.2 编码蛋白的不稳定指数小于 40.00,这 8 个蛋白为稳定蛋白,其它 15 个蛋白的不稳定指数均大于 40.00,为不稳定蛋白。

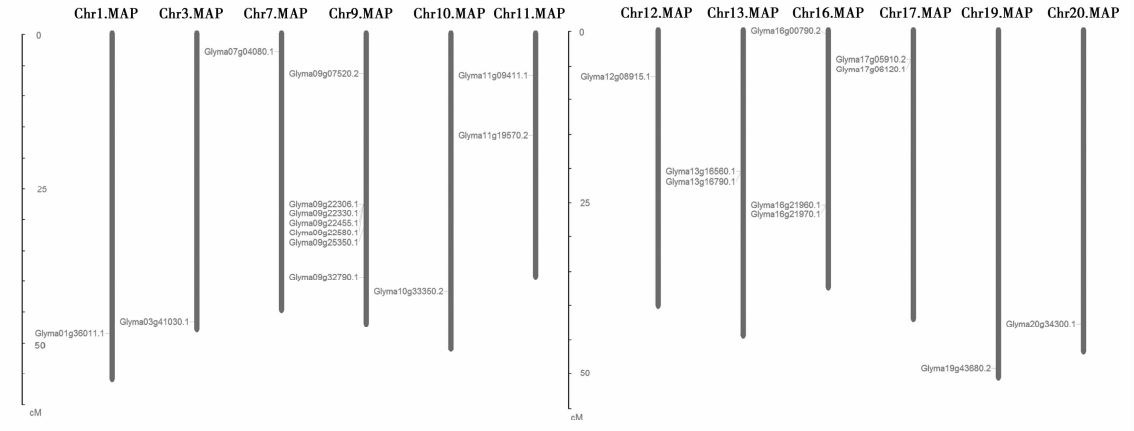

图 1 大豆油脂储存基因染色体定位

Fig. 1 Chromosome mapping of soybean storage oil gene

表 1 大豆油脂储存蛋白一级结构和外显子数

Table 1 The first structure and exon analysis of soybean storage oil proteins

名称 Name	氨基酸数目 No. of AA	分子量 Molecular weight /Da	等电点 pI	脂肪指数 Aliphatic index	疏水性 Hydrophobicity	不稳定指数 Instability index	外显子数 No . of exon
*Glyma*01*g*36011. 1	329	36884. 6	9. 70	99. 00	0. 312	42. 81	9
*Glyma*03*g*41030. 1	240	27167. 9	7. 10	74. 00	-0. 317	38. 03	6
*Glyma*07*g*04080. 1	676	75571. 0	6. 28	71. 97	-0. 360	41. 96	6
*Glyma*09*g*07520. 2	517	59286. 5	8. 93	100. 48	0. 216	46. 20	16
*Glyma*09*g*22306. 1	204	22592. 7	8. 92	79. 80	-0. 469	41. 51	6
*Glyma*09*g*22330. 1	201	22515. 7	9. 24	83. 43	-0. 492	30. 02	6
*Glyma*09*g*22455. 1	200	22568. 7	7. 07	84. 35	-0. 476	35. 12	6
*Glyma*09*g*22580. 1	202	22828. 8	7. 86	77. 23	-0. 531	42. 29	6
*Glyma*09*g*25350. 1	165	19197. 0	6. 15	82. 06	-0. 062	37. 22	5
*Glyma*09*g*32790. 1	337	38202. 0	9. 62	98. 37	0. 164	40. 48	9
*Glyma*10*g*33350. 2	189	21275. 0	6. 04	84. 66	-0. 310	53. 90	5
*Glyma*11*g*09411. 1	270	30762. 6	10. 03	94. 93	0. 261	41. 31	9
*Glyma*11*g*19570. 2	582	65783. 5	5. 91	87. 75	-0. 112	44. 95	7
*Glyma*12*g*08915. 1	1012	113732. 9	7. 44	89. 13	-0. 058	35. 84	12
*Glyma*13*g*16560. 1	498	57332. 4	8. 89	102. 75	0. 279	46. 75	16
*Glyma*13*g*16790. 1	668	74776. 8	8. 59	79. 43	-0. 256	36. 87	6
*Glyma*16*g*00790. 2	668	74557. 0	6. 28	71. 81	-0. 335	42. 50	6
*Glyma*16*g*21960. 1	350	39591. 7	9. 68	95. 83	0. 187	44. 08	9
*Glyma*16*g*21970. 1	317	36002. 6	9. 26	102. 68	0. 357	38. 33	9
*Glyma*17*g*05910. 2	668	74833. 8	8. 59	78. 98	-0. 270	36. 12	6
*Glyma*17*g*06120. 1	504	57999. 1	8. 89	102. 68	0. 250	45. 04	16
*Glyma*19*g*43680. 2	315	35959. 3	8. 95	78. 95	-0. 204	45. 11	6
*Glyma*20*g*34300. 1	227	25886. 5	5. 90	85. 51	-0. 158	61. 28	7

2.3 大豆油脂储存蛋白二级结构预测分析

对 23 个与油脂储存功能相关基因编码的蛋白进行二级结构预测分析(表 2)。二级结构预测结果表明:大部分蛋白的二级结构以无规则卷曲为主要构成元件,以 α - 螺旋为次要构成元件,β - 转角和延伸链的百分比最少。

ProtComp 亚细胞定位结果表明:蛋白主要定位于内质网、胞外和质膜。用 SignalP 对大豆油脂储存蛋白做信号肽分析,未发现信号肽位置。

表 2　大豆油脂储存蛋白二级结构预测和亚细胞定位

Table2　The secondary structure prediction and subcellular location

名称 Name	α－螺旋 α-Helix	β－转角 β-Turn	延伸链 Extended strand	无规则卷曲 Random coil	亚细胞定位 Subcellular location
*Glyma*01*g*36011. 1	97	35	93	104	内质网
*Glyma*03*g*41030. 1	59	35	48	98	胞外
*Glyma*07*g*04080. 1	201	71	125	279	质膜
*Glyma*09*g*07520. 2	208	45	92	172	内质网
*Glyma*09*g*22306. 1	71	26	35	72	胞外
*Glyma*09*g*22330. 1	72	23	30	76	胞外
*Glyma*09*g*22455. 1	69	23	39	69	胞外
*Glyma*09*g*22580. 1	74	27	30	71	胞外
*Glyma*09*g*25350. 1	48	20	39	58	胞外
*Glyma*09*g*32790. 1	104	33	87	113	内质网
*Glyma*10*g*33350. 2	56	18	32	83	胞外
*Glyma*11*g*09411. 1	54	38	84	94	内质网
*Glyma*11*g*19570. 2	209	61	126	186	质膜
*Glyma*12*g*08915. 1	368	79	219	346	质膜
*Glyma*13*g*16560. 1	202	37	83	176	内质网
*Glyma*13*g*16790. 1	208	64	154	242	质膜
*Glyma*16*g*00790. 2	219	66	126	257	质膜
*Glyma*16*g*21960. 1	113	32	89	116	内质网
*Glyma*16*g*21970. 1	122	31	78	86	内质网
*Glyma*17*g*05910. 2	206	69	152	241	质膜
*Glyma*17*g*06120. 1	203	36	94	171	内质网
*Glyma*19*g*43680. 2	84	36	70	125	胞外
*Glyma*20*g*34300. 1	106	27	28	66	胞外

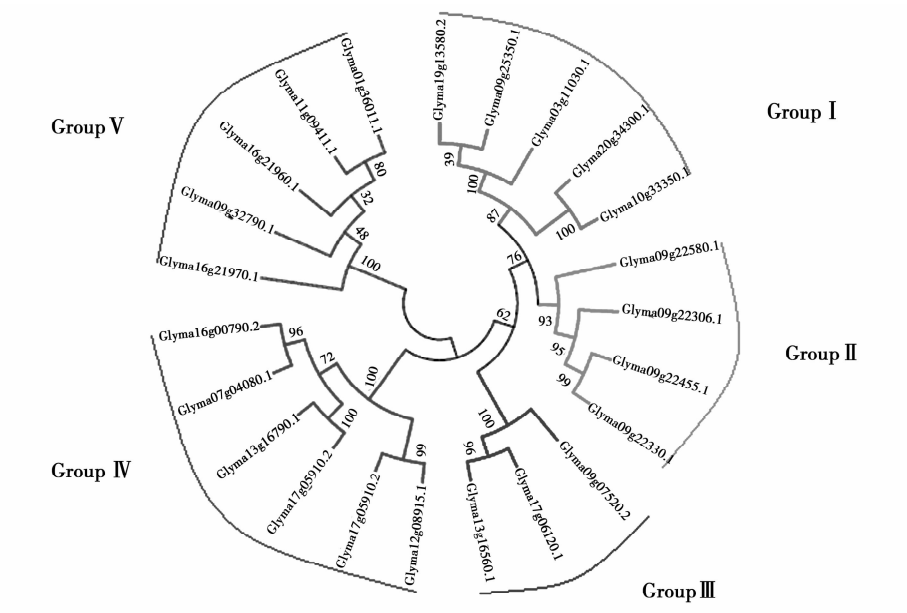

图 2　大豆油脂储存基因系统发育树

Fig. 2　The phylogenetic analysis of soybean storage oil proteins

2. 4　大豆油脂储存基因系统发育分析

对大豆中油脂储存基因单独构建系统发育树(图 2),由进化分析可以看出,整个进化树被分为 5 个分枝: Group Ⅰ 包括 5 个成员 *Glyma*03*g*41030. 1、*Glyma*09*g*25350. 1、*Glyma*10*g*33350. 2、*Glyma*19*g*43680. 2、*Glyma*20*g*34300. 1; Group Ⅱ 包括 4 个成员 *Glyma*09*g*22306. 1、*Glyma*09*g*22330. 1、*Glyma*09*g*22455. 1、*Glyma*09*g*22580. 1; Group Ⅲ 3 个成员 *Glyma*09*g*07520. 2、*Glyma*13*g*16560. 1、

*Glyma*17*g*06120.1; Group Ⅳ 包括 6 个成员 *Glyma*07*g*04080.1、*Glyma*11*g*19570.2、*Glyma*12*g*08915.1、*Glyma*13*g*16790.1、*Glyma*16*g*00790.2、*Glyma*17*g*05910.2;Group Ⅴ 包括 5 个成员 *Glyma*01*g*36011.1、*Glyma*09*g*32790.1、*Glyma*11*g*09411.1、*Glyma*16*g*21960.1、*Glyma*16*g*21970.1。Group Ⅰ、Ⅱ 的枝长较短,说明这些基因出现历史较短,发生功能变化的可能性也较小。Group Ⅲ、Ⅳ、Ⅴ 的枝长较长,说明这些基因出生较早,可能经历了比较复杂的进化过程。进一步分析发现,每个分枝包含的基因具有相似的基因结构,例如,Group Ⅱ 中的四个基因均含有 6 个外显子,Group Ⅲ 中的 3 个基因均含有 16 个外显子。

3 结 论

随着测序技术的不断发展,生物数据呈爆炸性增长,利用生物信息学的方法对基因进行分析已成为生物学研究的趋势之一。大豆在进化过程中经历过 2 次主要的加倍化事件,一次发生在5 900万年前,这次加倍化事件造成了大豆与苜蓿、百脉根的分化,然后大豆自身又发生了一次二倍化,约在 1 300万年前[14],所以大豆基因组中很多基因都是以多拷贝形式存在的。系统发育分析表明,23 个油脂储存蛋白被分为 5 个分枝,Group Ⅰ、Ⅴ 中均有 5 个成员,Group Ⅱ 中有 4 个成员,Group Ⅲ 的成员最少,为 3 个,Group Ⅳ 的成员最多,为 6 个。Group Ⅰ、Ⅱ 的枝长较短,说明这些基因出现历史较短,发生功能变化的可能性也较小。Group Ⅲ、Ⅳ、Ⅴ 的枝长较长,说明这些基因出生较早,可能经历了比较复杂的进化过程。每个分枝中所包含的基因具有相似的基因结构,外显子数目一致,同一分枝上的基因亚细胞定位结果也一样。

大豆基因组中油脂合成相关基因是一个重要的比较大的家族,涉及 9 种功能。同源比对分析确定了1 267个与油脂合成相关的基因,其中与油脂储存相关的基因有 23 个,虽然所占的比例很小,却也发挥着非常重要的作用。对得到的 23 个基因的蛋白序列和 CDs 序列,利用在线分析软件对油脂储存蛋白结构进行了分析。这 23 个基因的蛋白序列存在着较大的差异,氨基酸数目 165 ~ 1 012个不等,对应的分子量在19 197.0 ~ 113 732.9 Da,8 个蛋白属于稳定蛋白;二级结构预测结果显示,大部分蛋白的二级结构以无规则卷曲为主要构成元件,α - 螺旋为次要构成元件;亚细胞定位分析结果显示蛋白定位于内质网、质膜和胞外;信号肽分析未发现有信号位点;23 个油脂储存基因的外显子数目为5 ~ 16 个。植物种子油脂合成的过程之复杂,涉及的基因之多,给研究工作带了很大的困难,育种工作者从分子水平入手,通过调控一些关键酶的表达量,或者通过转基因的方法可以控制种子中的含油量和质量[15],但是这些方法受到了很多因素的制约。利用生物信息学方法,从基因水平上研究油脂合成基因的序列特征、进化规律,可以从本质上改善种子中油脂的含量和质量。本研究对大豆油脂储存蛋白进行初步分析,为深入了解该家族蛋白的合成调控、结构和功能等提供了参考数据。

参考文献

[1] He J,Gu D, Wu X, et al. Effect of soybean protein on blood pressure: A randomized controlled trial[J]. Annals of internal medicine, 2005, 143: 1-9.

[2] Soprani T, Uliana V K, Ribeiro R F Jr, et al. Cardiac protein changes in rats after soybean oil treatment:A proteomic study[J]. Lipids in Health and Disease, 2015(14):26.

[3] Li F, Hanson M V, Larock R C. Soybean oil-divinylbenzene thermosetting polymers: structure, properties and their relationships [J]. Polymer, 2001, 42: 1567-1579.

[4] Cahoon E B. Genetic enhancement of soybean oil for industrial uses: Prospects and challenges [J]. AgBioForum, 2003, 6: 11-131.

[5] Clemente T E, Cahoon E B. Soybean oil: Genetic approaches for modification of functionality and total content1[J]. Plant Physiology, 2009, 9 (151): 1030-1040.

[6] Schmutz J, Cannon S B, Schlueter J, et al. Genome sequence of the palaeopolyploid soybean [J]. Nature, 2010, 463 (7278): 178-83.

[7] Beisson F, Abraham J K Koo, Ruuska S, et al. *Arabidopsis* genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database[J]. Plant Physiology, 2003, 132: 681-697.

[8] Fehr W R. Reeding for modified fatty acid composition in soybean [J]. Crop Science(Suppl 3), 2007, 47: 72-87.

[9] Xu X P, Liu H, Tian L, et al. Integrated and comparative proteomics of high-oil and high-protein soybean seeds [J]. Food chemistry, 2015, 172:105-16.

[10] Poole R L. The TAIR database[J]. Methods in Molecular Biology, 2007, 406: 179-212.

[11] Guo A Y, Zhu Q H, Chen X, et al. GSDS: A gene structure display server[J]. Hereditas, 2007, 29 (8): 1023-1026.

[12] Wilkins M R, Gasteiger E, Bairoch A, et al. Protein identification and analysis tools on the ExPASy server[J]. Methods in Molecular Biology, 1999, 112: 571-607.

[13] Geourjon C, Deléage G. SOPMA: Significant improvement in protein secondary structure prediction by prediction from alignments and joint prediction[J]. Computer Applications in the Biosciences, 1995, 11(6): 681-684.

[14] Gary S, Lila V, Wayne A, et al. National science foundation-sponsored workshop report. Draft plan for soybean genomics[J]. Plant Physiology, 2004, 135 (1): 59-70.

[15] Haun W, Coffman A, Clasen B M, et al. Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family[J]. Plant Biochemistry and Biotechnology, 2014,12(7): 934-940.