生物信息方法分析大豆 Nodulin 家族蛋白

王巍杰[1],柴文静[1],杨文娇[2],侯瑛倩[2]

(1. 华北理工大学 生命科学学院,河北 唐山 063210; 2. 唐山拓普生物科技有限公司,河北 唐山 063002)

摘 要:为研究大豆 Nodulin 家族蛋白特性,在数据库 UniprotKB 中下载大豆 Nodulin 蛋白序列,利用 NCBI 的 BLAST 程序进行同源序列比对筛选和 Pfam 数据库综合分析,确定 14 个大豆 Nodulin 家族蛋白,通过生物信息学方法对 14 个家族蛋白的基本理化性质、跨膜区域、模体、系统发育和蛋白结构等进行分析。结果显示:14 条蛋白序列氨基酸数为 72 ~ 116、理论等电点为 9. 94 ~ 11. 59、分子质量为 7 408. 70 ~ 12 465. 55、脂肪系数为 85. 94 ~ 114. 12,除 K7KNK4 没有跨膜区外,其余 13 条 Nodulin 蛋白均有 1 ~ 2 跨膜区。二级结构主要由 α 螺旋和无规则卷曲组成,13 个模体中模体 1、模体 2、模体 3 和模体 4 保守性强,系统发育进化树由 6 个分枝组成,分枝Ⅵ的 A0A0B2P3C8、I1MWA0 三维结构预测结果也相似。

关键词:大豆;Nodulin 蛋白;模体;系统进化分析

中图分类号:S565. 1 **文献标识码**:A **DOI**:10. 11861/j. issn. 1000-9841. 2018. 01. 0039

Bioinformatics Analysis of Nodulin Family Proteins in Soybean

WANG Wei-jie[1], CHAI Wen-jing[1], YANG Wen-jiao[2], HOU Ying-qian[2]

(1. College of Life Science, North China University of Science and Technology, Tangshan 063210, China; 2. Tangshan Top Bio-Technology Co., Ltd, Tangshan 063002, China.)

Abstract: The nodulin protein sequence was downloaded from the database UniprotKB, and 14 nodulin family members were identified by BLAST search in the NCBI database. The basic physicochemical properties, gene structure, protein structure and phylogenetic tree were analyzed by bioinformatics tools. The bioinfoematics analysis indicated fourteen nodulin proteins vary in composition animo acid(72 and 116), isoionic point(94 and 11. 59), molecular weight(5 898. 87 and 12 465. 55), and the fat coefficient(85. 94 and 114. 12), transmembrane regions(1 and 2) except K7KNK4. The secondary protein structures were mainly composed of α helix and irregular curl. Motifs(1 to 4) were more conserved among 14 motifs, phylogenetic tree constructed by mega 8 were divided into 6 protein groups, A0A0B2P3C8 and I1MWA0 in group Ⅵ also shared the similar three demesional structures.

Keywords: Soybean; Nodulin protien; Motif; Phylogenetic analysis

大豆是重要的经济作物之一,蛋白含量高达 40%[1],大豆蛋白是一种含有人体所需的 9 种必须氨基酸且不含胆固醇的植物蛋白,具有很高的营养价值[2]。大豆氮素来源于根瘤固氮、土壤氮和肥料氮[3]。其中根瘤菌对大豆的作用明显,根瘤菌与大豆共生,产生 Nodulin 蛋白,在大豆根部形成根瘤,根瘤主要固定大气中的氮素供植物生长发育,根瘤菌的共生固氮可为宿主植物提供 50% 以上的氮元素,对大豆产量和蛋白含量有积极作用[4]。2010 年《Nature》杂志公布了大豆的基因组序列,全基因组由 1. 1 亿个碱基组成,其中46 430 个蛋白编码基因[5]。目前,有关大豆根瘤素家族蛋白的研究报道不多,Nodulin 基因最初因在豆科植物根瘤中的特异性表达而命名,在非结瘤植物如水稻、拟南芥中也分离鉴定出 Nodulin 同源基因[6]。

本文从 UniprotKB 和 Pfam[7] 数据库获取 14 个大豆 Nodulin 家族蛋白,采用生物信息方法进行理化、性质、结构、进化关系等生物信息学分析,揭示 Nodulin 家族蛋白特性,为深入研究 Nodulin 蛋白的作用机制和大豆及豆科植物生物固氮研究提供一定的理论基础,同时也为深入了解 Nodulin 家族蛋白在植物中的同源蛋白研究提供参考。

1 材料与方法

1. 1 数据材料

在数据库 UniprotKB 中下载大豆 Nodulin 蛋白序列,在 NCBI 数据库中利用 BLAST 进行同源序列比对筛选,结合 Pfam 数据库综合分析,确定 14 个 Nodulin 蛋白家族成员。

1. 2 大豆 Nodulin 蛋白家族理化性质分析

采用 ExPaXy 提供的在线分析工具 Protparam (http://web. expasy. org/protparam/)[8] 分析 Nodulin

收稿日期:2017-07-16

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。

家族蛋白的氨基酸数目、分子量、等电点等理化性质。

1.3 大豆 Nodulin 蛋白跨膜结构的预测和分析

采用 TMHMM Server V 2.0 在线软件(http://www.cbs.dtu.dk/services/TMHMM/)分析跨膜结构预测。

1.4 模体、多序列比对及系统进化分析

使用 MEGA 7.0[9] 软件内置的 Clustal W 对 Nodulin 蛋白序列进行多序列分析,采用相邻连接法(neighbor-joining,NJ)构建系统发生树,bootstrap 次数设为 1 000,其它参数为默认值。在线软件 MEME(http://meme-suite.org/tools/meme)寻找识别 Nodulin 蛋白的模体,最小模体长度设置为 6,其它参数为默认值。

1.5 大豆 Nodulin 蛋白二级结构预测分析

Nodulin 蛋白的二级结构预测用在线分析软件 GOR4(https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html)。

1.6 大豆 Nodulin 蛋白质三级结构预测

家族的三级结构分析用在线软件 SWISS-MODEL(https://swissmodel.expasy.org/interactive)[10-11]。

2 结果与分析

2.1 大豆 Nodulin 蛋白的理化性质分析

14 个大豆 Nodulin 家族蛋白的理化性质分析结果如表 1 所示:14 个蛋白的氨基酸数目为 56 ~ 116,其中有 3 个蛋白的氨基酸数目小于 100,A0A0B2P3C8 所含氨基酸数目最多,为 116。Nodulin 家族蛋白的氨基酸组成成分中,丙氨酸(Ala)含量最多,占总蛋白的 20% ~ 25% 左右,其次为丝氨酸(Ser),占蛋白总数的 6% ~ 11%。Nodulin 蛋白的分子量为 5 898.87 ~ 12 465.55。Nodulin 蛋白的等电点均大于 7.5,K7ML71 的等电点最大,为 11.59。

不稳定系数分析表明:A0A0B2P3C8 和 I1MWA0 两个蛋白的不稳定系数分别为 42.12 和 41.66 属于不稳定蛋白,其它 12 个蛋白为稳定蛋白。A0A0B2P3C8、I1MWA0 具有相似的疏水性区域,存在 3 个高分值峰,分别位于 20 ~ 25、45 ~ 55、80 ~ 90 区域,A0A0B2R5F3、A0A0B2P3C8、I1MWA0 蛋白总平均亲水性为负值,属于疏水蛋白(图 1)。

表 1 大豆 Nodulin 蛋白的理化性质分析

Table 1 Analysis of physicochemical properties of Nodulin protein

蛋白质登录号 Protein accession number	氨基酸数目 No. of AA	分子量 Molecular weight/kDa	理论等电点 Theoretical pI	不稳定系数 Instability index	亲水性平均系数 Hydropathicity	脂肪指数 Aliphatic index
A0A0B2R617	107	11285.05	10.97	20.87	0.132	89.63
A0A0B2Q0I7	107	11074.86	10.72	24.36	0.219	88.88
A0A0B2P3C8	116	12465.55	9.94	42.12	-0.037	96.98
A0A0B2Q195	105	11139.98	10.45	21.04	0.241	92.43
A0A0B2RN01	105	10890.54	10.13	20.30	0.168	94.19
A0A0B2R6H2	106	11091.80	10.68	23.52	0.121	85.94
A0A0B2R5F3	56	5898.87	11.00	5.89	-1.933	104.82
K7ML69	107	11043.84	10.22	21.68	0.301	92.43
K7ML70	107	11023.77	10.68	27.10	0.252	88.88
K7ML71	72	7408.70	11.59	9.77	0.475	102.08
I1MWA0	115	12336.43	10.09	41.66	-0.007	97.83
K7KNK3	107	11255.03	10.97	20.87	0.155	90.56
C6T443	105	10933.97	10.37	19.58	0.089	90.48
K7KNK4	99	10503.01	9.89	28.31	0.004	88.89

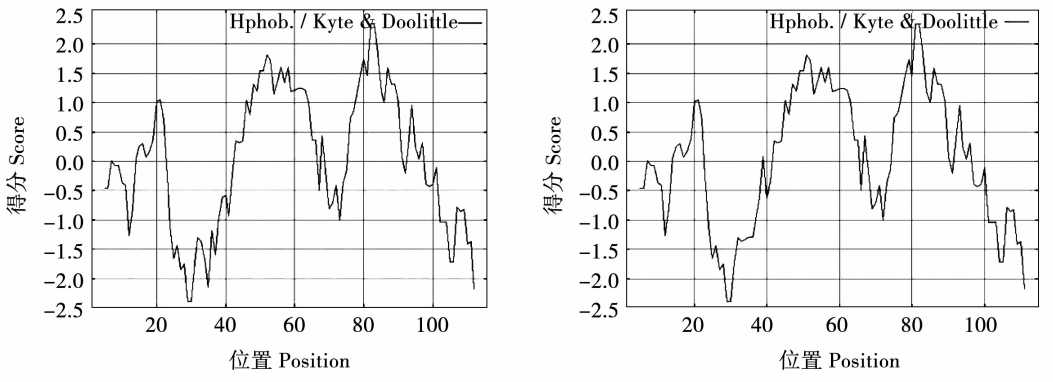

图 1　部分 Nodulin 蛋白的疏水性

Fig. 1　Partial of Nodulin protein hydrophobicity

2.2　大豆 Nodulin 蛋白质跨膜区域分析

14 条蛋白序列预测分析结果如图 2 所示:除 K7KNK4 蛋白没有预测到跨膜区域,A0A0B2R5F3 蛋白在 20 ~ 37 氨基酸处有一个跨膜区。A0A0B2R617、A0A0B2RN01、A0A0B2Q0I7、A0A0B2P3C8、A0A0B2Q195、A0A0B2R6H2、K7ML69、K7ML70、K7ML71、I1MWA0、K7KNK3、C6T443 这 12 条蛋白均有由内向外,再由外向内的两个跨膜区域,其中蛋白 A0A0B2R617、A0A0B2Q195、K7ML69、K7ML70、K7KNK3、A0A0B2Q0I7 的跨膜区位于 34 ~ 56 和 71 ~ 88 氨基酸处,A0A0B2P3C8、A0A0B2R6H2、K7ML71、I1MWA0 的跨膜区域分布没有明显的特征。

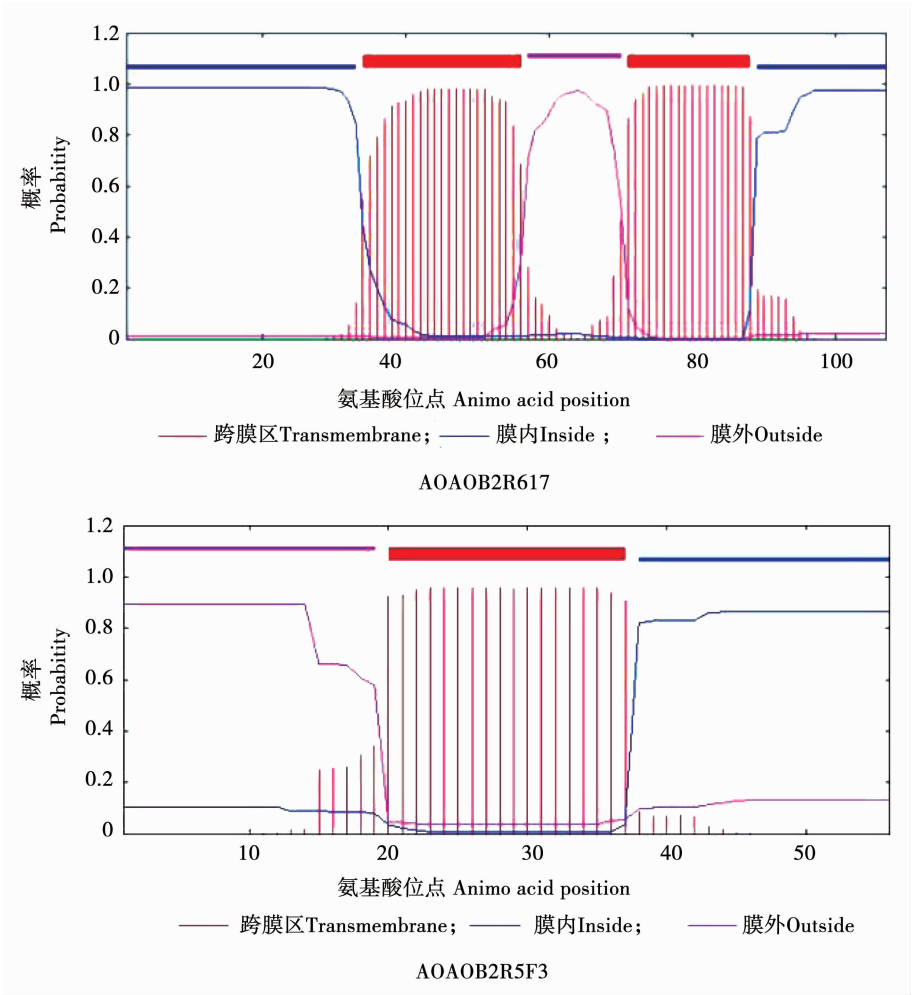

图 2　Nodulin 蛋白家族部分蛋白成员的跨膜结构预测

Fig. 2　Transmembrane structure prediction of partial members of Nodulin protein

2.3 大豆 Nodulin 家族蛋白的模体和多序列比对分析

14 条大豆 Nodulin 蛋白的模体(图 3)分析发现大于 6 个氨基酸残基的模体有 5 个,模体 1 为 50 个氨基酸、模体 2 为 28 个氨基酸、模体 3 为 15 个氨基酸、模体 5 为 29 个氨基酸、模体 6 为 8 个氨基酸,模体 1 出现在除 K7KNK4 以外的 13 条 Nodulin 蛋白中,模体 2 出现在除 A0A0B2P3C8、A0A0B2R5F3、1MWA0 以外的 12 条 Nodulin 蛋白中,模体 3 出现在除 A0A0B2R6H2、A0A0B2R5F3、K7ML71 外的 11 条 Nodulin 蛋白中,模体 5 出现在 A0A0B2P3C8、I1MWA0 中,模体 6 出现在 A0A0B2P3C8、K7KNK3 的羧基端。6 个氨基酸残基的模体有 8 个,其中模体 4 出现在 A0A0B2P3C8、K7KNK3 的氨基酸氨基端和除 K7KNK4 以外其余 11 条 Nodulin 蛋白的羧基端;模体 8、模体 9、模体 10、模体 11、模体 12、模体 13 出现在 K7KNK4 中;其余模体分布没有明显的规律。模体分析显示模体 1、模体 2、模体 3、模体 4 在 Nodulin 家族蛋白中是十分保守的。

14 条 Nodulin 蛋白的多序列比对结果(图 4)显示,A0A0B2R5F3 和 K7ML71 氨基酸数目相对于其余 12 条蛋白数目较少,其余 12 条蛋白保守性较好,蛋白的保守区域主要集中在除 A0A0B2R5F3 外的蛋白质中部和靠近氨基端的序列中,位于模体 1 中,22 个保守位点氨基酸中非极性氨基酸含量较高,可能和 Nodulin 蛋白的疏水性有关。

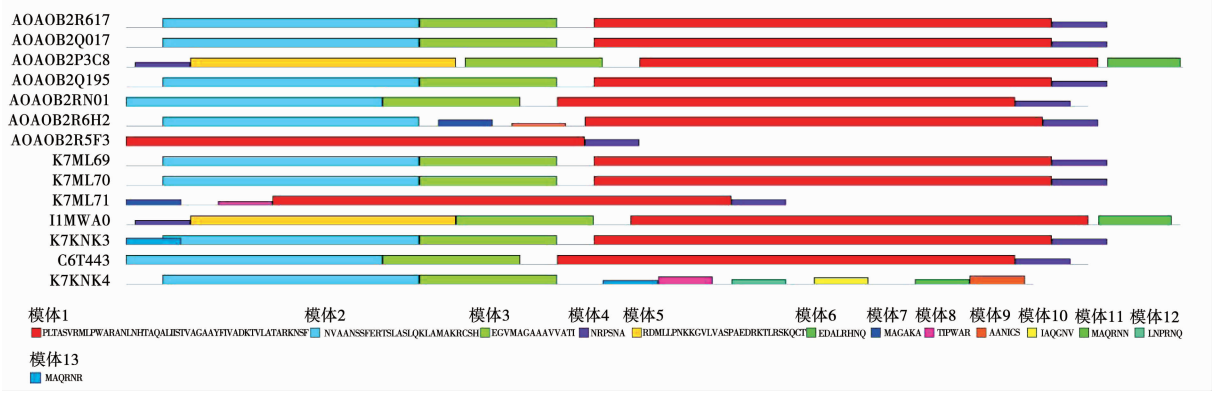

图 3 Nodulin 蛋白的模体

Fig. 3 The conserved motif of Nodulin protein

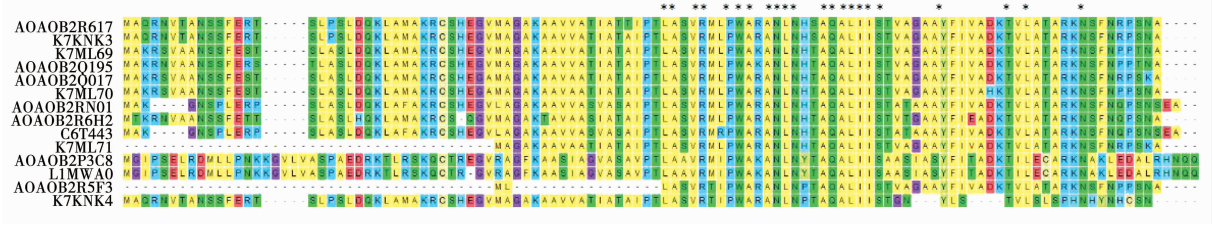

图 4 Nodulin 蛋白的多序列比对结果

Fig. 4 Multiple sequence alignment results of Nodulin protein

2.4 Nodulin 家族蛋白系统发育分析

对 14 条 Nodulin 蛋白构建系统发育树(图 5),由进化分析看出,进化树被分为 6 个分枝:分枝Ⅰ包括 4 个成员 A0A0B2R617、K7KNK3、A0A0B2Q0I7、K7ML71;分枝Ⅱ包括 3 个成员 K7ML70、A0A0B2Q195、K7ML69;分枝Ⅲ包括 3 个成员 A0A0B2R6H2、A0A0B2RN01、C6T443;分枝Ⅵ有 1 个成员 A0A0B2R5F3;分枝Ⅴ有 1 个成员 K7KNK4;分枝Ⅵ包括 2 个成员 A0A0B2P3C8、I1MWA0。分枝Ⅰ、分枝Ⅱ的 Nodulin 蛋白枝较短,说明出现的历史较短,功能变化不大;分枝Ⅵ的 Nodulin 蛋白枝较长,出现的时间较早,经历了复杂的进化过程;分枝Ⅵ和分枝Ⅴ均有 1 个成员组成,蛋白枝较长。结合模体和多序列进一步分析发现,每个分枝包含的 Nodulin 蛋白相似,例如,分枝Ⅵ的 A0A0B2P3C8、I1MWA0 均有相同的 5 个模体;组成分枝Ⅵ的 K7KNK4 是唯一一条不含模体 1 的蛋白,有 8 个模体,其中模体 8、模体 9、模体 10、模体 11、模体 12 在其它 13 个 nodulin 蛋白中未出现,K7KNK4 和其它蛋白序列差异较大,进化速度较慢。

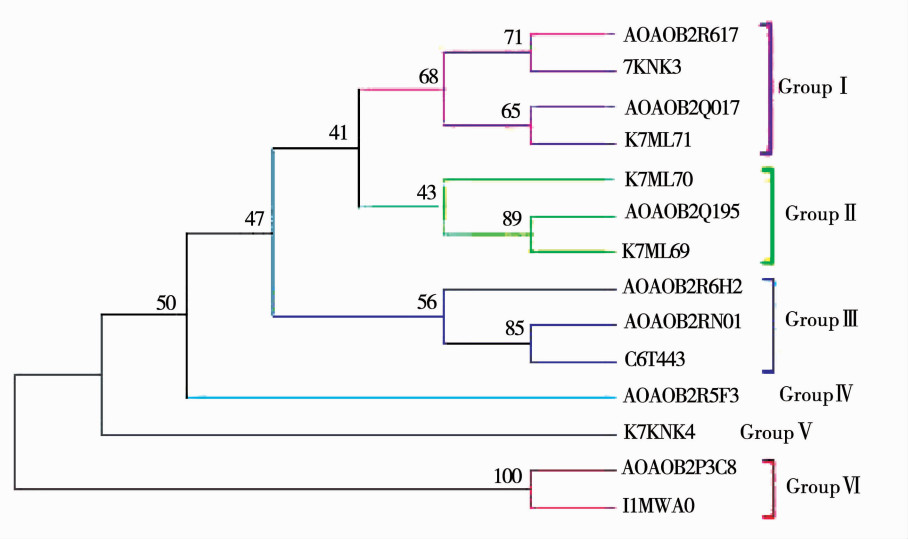

图 5　Nodulin 蛋白的系统发育树

Fig. 5　Phylogenetic tree of Nodulin protein

2.5　蛋白质二级结构预测

采用在线分析工具 GOR4 进行二级结构预测(表 2),除 K7KNK4 蛋白无规则卷曲占 50% 以上,其它 13 条蛋白二级结构相似。13 条蛋白序列的二级结构 α 螺旋占较高比例,占 50% ~71%;无规则卷曲为次要构成元件,占 25% ~46%;延伸链所占比例最少,占 1.8% ~5.2%。

表 2　Nodulin 蛋白二级结构组成

Table 2　The secondary structure of Nodulin protein

蛋白质登录号 Protein accession number	α 螺旋 Alpha helix(Hh)/%	延伸链 Extended strand(Ee)/%	无规则卷曲 Random coil(Cc)/%
A0A0B2R617	62.62	1.87	35.51
A0A0B2Q0I7	71.03	3.74	25.23
A0A0B2P3C8	69.83	5.17	25.00
A0A0B2Q195	64.49	3.74	31.78
A0A0B2RN01	68.57	1.90	29.52
A0A0B2R6H2	66.04	3.77	30.19
A0A0B2R5F3	50.00	3.57	46.43
K7ML69	63.55	3.74	32.71
K7ML70	71.03	3.74	25.23
K7ML71	68.06	2.78	29.17
I1MWA0	63.48	5.22	31.30
K7KNK3	64.49	1.87	33.64
C6T443	68.57	1.90	29.52
K7KNK4	38.38	7.07	54.55

2.6　Nodulin 蛋白三维结构预测

14 个 Nodulin 蛋白中 A0A0B2P3C8、I1MWA0、K7ML69 的三维结构的预测结果如图 6 所示,3 种蛋白以 α 螺旋结构占主导,和二级结构分析结果一致。其中 A0A0B2P3C8 和 I1MWA0 的三维结构相似,疏水性也相似,和蛋白系统进化分析结果一致,位于发育树的同一分枝。

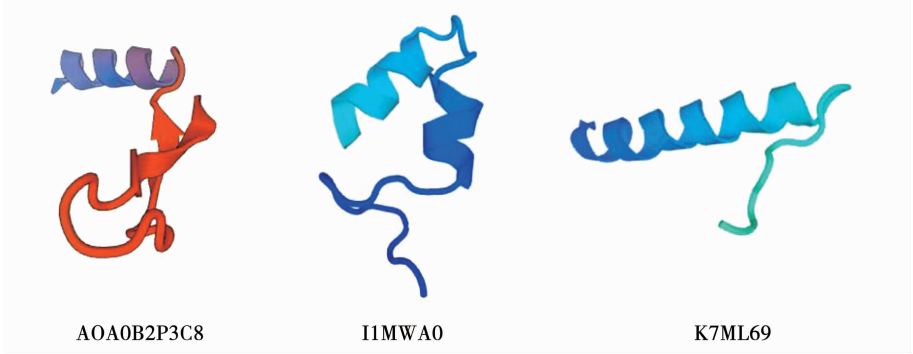

图 6 Nodulin 蛋白三级结构

Fig. 6 Nodulin protein tertiary structure prediction results

3 结论与讨论

Nodulin 家族蛋白的氨基酸数目为 56 ~ 116，分子量为 7 408. 70 ~ 12 465. 55，氨基酸组成以丙氨酸和丝氨酸含量较多；等电点大于 7. 5，推测 Nodulin 蛋白为碱性蛋白；脂肪系数为 85. 94 ~ 114. 12；除 A0A0B2P3C8、I1MWA0 外的 12 条蛋白为稳定性较好的蛋白；A0A0B2P3C8、I1MWA0 蛋白属于疏水蛋白，两个蛋白存在极相似的 3 个疏水区域，且三级结构相似，揭示蛋白的三维结构可能与蛋白的亲疏水性有一定的关系；其余 12 条蛋白为亲水蛋白，亲水程度不同；二级结构中 K7KNK4 蛋白无规则卷曲占 50% 以上，其它 13 条蛋白 α 螺旋占较高比例；除 K7KNK4 蛋白外，其余 13 条蛋白为跨膜蛋白，存在 1 ~ 2 跨膜区域；14 个 Nodulin 家族蛋白序列较保守，蛋白的保守区域主要集中在除 A0A0B2R5F3 外的蛋白质中部和靠近氨基端的序列中，位于模体 1 中。系统发育分析表明 Nodulin 蛋白被分为 6 个分枝，分枝Ⅰ有 4 个蛋白，分枝Ⅱ、分枝Ⅲ均有 3 个蛋白，分枝Ⅵ、分枝Ⅴ均有 1 个蛋白，分枝Ⅵ有 2 个蛋白，分枝Ⅰ、Ⅱ、Ⅲ的 Nodulin 蛋白枝较短，说明出现的历史较短，功能变化不大；分枝Ⅵ和分枝Ⅴ均有 1 个成员组成，蛋白枝较长；分枝Ⅵ包括 A0A0B2P3C8、I1MWA0，蛋白枝也较长，进化速度较慢，二者的三维结构预测结果也相似。

参考文献

[1] 刘新旗，涂丛慧，张连慧，等. 大豆蛋白的营养保健功能研究现状[J]. 北京工商大学学报(自然科学版)，2012，30(2)：1-6. (Liu X Q，Tu C H，Zhang L H，et al. Research status of nutrition and health function of soybean protein[J]. Journal of Beijing Technology and Business University(Natural Science Edition)，2012，30(2)：1-6.

[2] 田琨. 大豆蛋白的结构表征及应用研究[D]. 上海：复旦大学，2010. (Tian K. Structural characterization and application of soybean protein[D]. Shanghai：Fudan University，2010.)

[3] 严君. 大豆结瘤固氮及生长发育对土壤环境无机氮含量的响应[D]. 哈尔滨：东北农业大学，2011. (Yan J. Responses of nitrogen fixation and growth and development to soybean nitrogen in inorganic nitrogen[D]. Harbin：Northeast Agricultural University，2011.)

[4] 李俊，沈德龙，林先贵. 农业微生物研究与产业化进展[C]. 北京：科学出版社，2011：289. (Li J，Shen D L，Lin X G. Advances in research and industrialization of agricultural microorganism[C]. BeiJing：Science Press，2011：289).

[5] Jeremy S，Steven B，Cannon J，et al. Genome sequence of the palaeontology[J]. Soybean Nature，2010，463(7278)：178-183.

[6] Fedorova M，Vance C P. Genome-wide identification of nodule-specific transcripts in the model legume Medicago truncatula [J]. Plant Physiology，2002，130(2)：519.

[7] 姚玉波. 大豆根瘤固氮特性与影响因素的研究[D]. 哈尔滨：东北农业大学，2012. (Yao Y B. Study on nitrogen fixation characteristics and influencing factors of soybean root[D]. Harbin：Northeast Agricultural University，2012.).

[8] Finn R D，Mistry J，Schuster-Bockler B，et al. Pfam：clans，web tools and services [J]. Nucleic Acids Research，2006，34：247-251.

[9] Wilkns M R，Gasteiger E，Bairoch A，et al. Protein identification and analysis tools on the ExPASy server[J]. Methods in Molecular Biology，1992，112：571-607.

[10] Kumar S，Stecher G，Tamura K. MEGA7：Molecular evolutionary genetics analysis Version 7.0 for bigger datasets[J]. Molecular biology and evolution，2016，33(7)：1870-1874.

[11] Bordoli L，Schwede T. Automated protein structure modeling with SWISS-MODEL workspace and the protein model portal [J]. Methods Molecular Biology，2012，857：107-136.