

· 研究报告 ·

根瘤感受样基因的进化：结构歧异与功能分化

朱新宇^{*}, 吕万胜, 余春梅, 汪保华

南通大学生命科学学院, 南通 226019

摘要 豆科植物百脉根(*Lotus japonicus*)的根瘤感受基因*Nin*与根瘤的早期发育有关。*Nin*的同源基因(*Nin-like*基因)功能上涉及氮代谢过程。从完成测序的豆科和非豆科植物基因组中获取*Nin-like*基因并进行系统发育分析。在此基础上, 追踪基因和蛋白质结构的歧异式样, 尝试建立结构歧异和功能分化的联系。通过比较, 新的*Nin-like*基因被鉴别。系统发育分析不仅重现了以前分辨的直系同源群(分支I、II和III), 且识别了它们之间的姐妹群关系。*Nin-like*基因的结构呈现多样性, 支持系统发育分析的结果。水稻*OsNLP5*基因缺乏内含子, 可能起源于基因返座事件。*NIN-like*蛋白结构域组织和功能位点在不同分支中存在差异, 提示它们的功能发生了分化。根瘤固氮植物*NIN-like*蛋白的GAF结构域中存在一个显著变异区, 三级建模分析显示这个变异区对应于百脉根非固氮*NIN-like*蛋白的一段保守构象, 这一变异可能使豆科植物具有根瘤固氮能力。研究结果为阐明*Nin-like*基因的功能提供了新的研究思路。

关键词 进化, 功能分化, *Nin-like*基因, 根瘤共生, 结构歧异

朱新宇, 吕万胜, 余春梅, 汪保华 (2013). 根瘤感受样基因的进化: 结构歧异与功能分化. 植物学报 48, 519–530.

最初, Schäuser等(1999)在豆科模式植物百脉根(*Lotus japonicus*)中鉴定了调控根瘤菌感染线形成和原基细胞启动的*Nin*(nodule inception)基因。随后, Borisov等(2003)在豌豆(*Pisum sativum*)中鉴定了*Nin*基因的直系同源基因*Sym35*, 以及 Schäuser等(2005)在拟南芥(*Arabidopsis thaliana*)和水稻(*Oryza sativa*)等非豆科植物中鉴别了*Nin-like*同源基因。由此证实*Nin*基因属于一个多基因家族。*NIN-like*蛋白包含1个保守的RWP-RK结构域(PF02042)、1个C末端的PB1结构域(SM00666)和1–2个N末端的GAF结构域(PF01590)。在功能上, RWP-RK结构域负责与DNA结合, 因此推测*NIN*蛋白作为一种转录因子行使功能(Schäuser et al., 1999, 2005); PB1结构域负责蛋白质-蛋白质间的相互作用, 以及与其它包含PB1结构域的蛋白质形成异聚体(Ito et al., 2001); GAF结构域一般存在于植物光敏色素和cGMP特定的磷酸二酯酶(cGMP-specific phosphodiesterases)中, 也存在于与固氮密切相关的根瘤菌转录调控蛋白*NifA*中(Aravind and Ponting, 1997)。研究结果(Schäuser et al., 2005)显示, 百脉根*Nin*基因在进化过程中丢失了一段位于GAF结构域区域的序列, 推测这一片段丢

失事件使得百脉根*Nin*基因招募为根瘤固氮基因。然而, 这一丢失事件尚未在更多豆科固氮植物中得到证实。研究显示, *Nin-like*基因尽管功能不同, 但都响应氮营养信号, 显示与氮源的利用有关, 功能上可能存在相互协同的关系(Scheible et al., 2004; Barbulova et al., 2007; Castaings et al., 2009)。目前, 这些功能不同但相关的*Nin-like*基因结构歧异式样与功能分化的关系尚未被充分认识。随着多个豆科模式植物核基因组测序的完成(Cannon et al., 2006; Sato et al., 2008; Schmutz et al., 2010), 探索*Nin*同源基因的进化歧异与功能分化的条件也逐渐成熟。本研究在基因和蛋白质水平上分析*Nin-like*基因的结构歧异式样, 探讨结构歧异与功能分化的联系, 旨在为*Nin-like*基因的功能研究提供新思路。

1 材料与方法

1.1 数据提取

以百脉根(*Lotus japonicus* (Regel) K. Larsen)*NIN*蛋白作为查询序列, 使用缺省参数, tBLASTn搜索4个非豆科模式植物和3个豆科模式植物的*Nin-like*基因(表1)。

收稿日期: 2012-07-27; 接受日期: 2012-12-27

基金项目: 国家自然科学基金(No.31000729)、江苏省高校自然科学基金(No.09KJB180006)和南通市应用研究计划(No.BK2012062)

* 通讯作者。E-mail: zhuxinyu@ntu.edu.cn

表1 本研究所涉及的物种和*Nin-like*基因的取样**Table 1** The plant species and nodule inception like genes surveyed in this study

Species (Abbr.)	Name	Gene model	GenBank accession	Length (aa)
<i>Arabidopsis thaliana</i> (At)	AtNLP1	AT2G17150	Q8H111	909
	AtNLP2	AT4G35270	D7MD23	963
	AtNLP3	AT4G38340	Q9SVF1	767
	AtNLP4	AT1G20640	D7KJ39	844
	AtNLP5	AT1G76350	D7KTC4	808
	AtNLP6	AT1G64530	D7KS88	841
	AtNLP7	AT4G24020	D7M8Y6	959
	AtNLP8	AT2G43500	D7LK24	947
	AtNLP9	AT3G59580	Q9M1B0	894
<i>Oryza sativa</i> (Os)	OsNLP1	Os03g03900	NP_001048860	942
	OsNLP2	Os04g41850	Q0JC27	936
	OsNLP3	Os01g13540	NP_001042530	938
	OsNLP4	Os11g16290	ABA92484	886
	OsNLP5	Os09g37710	EEE70166	865
<i>Glycine max</i> (Gm)	GmNLP1	Glyma06g00240.1	XP_003527589	682
	GmNLP2	Glyma04g00210.1	XP_003523488	744
	GmNLP3	Glyma16g30180.1	XP_003548181	963
	GmNLP4	Glyma09g25230.1	XP_003533182	933
	GmNLP5	Glyma11g13390.1	XP_003539038	965
	GmNLP6	Glyma12g05390.1	XP_003540690	946
	GmNLP7	Glyma15g03220.1	XP_003546980	973
	GmNLP8	Glyma20g29960.1	XP_003555384	897
	GmNLP9	Glyma13g42160.1	XP_003542064	974
	GmNLP10	Glyma10g37860.1	XP_003536463	883
<i>Medicago truncatula</i> (Mt)	MtNLP1	Medtr4g092610	XP_003606812	912
	MtNLP2	Medtr5g106690	XP_003618108	933
	MtNLP3	Medtr1g126970	XP_003592270	979
	MtNLP4	Medtr2g120530	XP_003597538	993
<i>Lotus japonicus</i> (Lj)	LjNIN	Chr2.CM0102.250.r2.m	CAB61243	878
	LjNLP1	Chr1.CM0178.280.r2.m	CAE30324	904
	LjNLP2	Chr3.CM0106.760.r2.m	CAE30325	972
	LjNLP3	Chr5.CM0148.170.r2.a	—	976
	LjNLP4	Chr3.CM0091.230.r2.m	—	985
<i>Pisum sativum</i> (Ps)	PsNIN		CAD37948	921
<i>Selaginella moellendorffii</i> (Sm)	SmNLP1	Gw1.3._1005.1	XP_002963404	606
	SmNLP2	EstExt_Genewise1Plus.C_180498	XP_002971920	777
<i>Physcomitrella patens</i> (Pp)	PpNLP1	EstExt_Genewise1.C_260189	XP_001757159	697
	PpNLP2	EstExt_gwp_gw1.C_1280044	XP_001770664	799
	PpNLP3	EstExt_Genewise1.C_2500013	XP_001779081	697
	PpNLP4	Gw1.109.16.1	XP_001769027	661

*Lotus japonicus*的数据来自<http://www.kazusa.or.jp/lotus/>, 其余均来自phytozome(<http://www.phytozome.net/search.php>)或JGI数据库(<http://genome.jgi-psf.org/>)。—: GenBank蛋白数据库中缺乏相应序列。

The data of *Lotus japonicus* are retrieved from <http://www.kazusa.or.jp/lotus/>, the rest from phytozome (<http://www.phytozome.net/search.php>) or JGI (<http://genome.jgi-psf.org/>)。—: Denotes the lack of corresponding sequences.

其中, 百脉根序列从百脉根基因组数据库(<http://www.kazusa.or.jp/lotus/>)中提取(Sato et al., 2008),

其余从phytozome或JGI数据库中提取(Goodstein et al., 2012)。此外, 为了更好地理解豆科固氮植物*Nin-*

*like*基因的进化史, 我们还加入了豆科非模式植物豌豆的*Nin*基因(Borisov et al., 2003)。在几种候选序列中, 如果同一基因存在选择性剪切, 仅选择最长的序列。如果存在截断序列, 则重新预测基因模型。最终选择的同源序列要求与百脉根NIN蛋白具有同样的结构域组织, 即同时包含GAF、RWP-RK和PB1结构域。蛋白结构域组织信息通过SMART平台获取(Schultz et al., 1998)。对于真正的*Nin-like*基因, 我们预期它们在候选序列中排名靠前。对每条候选序列, 用手工检查它们的结构域组织, 且对最终确定的候选者进行EST搜索以获取这些基因的表达信息(Lee et al., 2005)。

1.2 序列比对和系统发育分析

系统发育分析使用蛋白质序列。截取每条蛋白质序列中GAF、RWP-RK和PB1结构域区域的序列, 通过Clustal X进行序列比对(Thompson et al., 1997)。首先建立3个独立的结构域数据集, 手工校对比对结果, 然后合并这3个比对为单一的数据矩阵。为了进行基因结构分析, 使用PAL2NAL程序(Suyama et al., 2011), 以全长蛋白质序列比对为基础构建基于密码子的比对。用最大似然法(ML)(Felsenstein, 1981)和邻接法(NJ)(Saitou and Nei, 1987)构建系统树时分别用PHYML(Guindon and Gascuel, 2003)和MEGA程序(Tamura et al., 2011)。ML分析中, 使用ProTest程序(Abascal et al., 2005)测试进化模型和速率异质性参数。NJ分析中选择成对删除空位方式。支持度通过non-parametric bootstrap评估; 分别使用1 000和500次重复抽样分析进行NJ和ML系统树牢固度评估。支持强度分为低(50%–75%)、中(76%–85%)和高(86%–100%)。

1.3 基因结构分析

使用spidey程序(Wheelan et al., 2001)比较全长蛋白质比对和相应的cDNA比对, 以获得基因结构, 然后手工分析内含子-外显子边界和相位信息。相位0指内含子插入在密码子之间; 相位1指内含子插入密码子第1个核苷酸和第2个核苷酸之间; 相位2指定内含子插入密码子第2个核苷酸和第3个核苷酸之间。内含子的位置信息依据密码子比对和内含子-外显子边界信息获得。本研究中内含子位置即使被1个碱基对分开

也被认为是不相同的, 尽管不能排除它们有共同起源的可能(Rogozin et al., 2000)。

1.4 蛋白质结构分析

利用ScanProsite程序(<http://prosite.expasy.org/scanprosite/>)(Gasteiger et al., 2005)分析蛋白质功能位点的类型和数量。利用WoLF PSORT程序(<http://wolfpsort.org/>)(Horton et al., 2007)分析蛋白质的亚细胞定位。通过Phyre(<http://www.sbg.bio.ic.ac.uk/~phyre/>)(Kelley and Sternberg, 2009)远程同源建模方法对NIN-like蛋白的N末端GAF结构域进行三级结构建模。以人类磷酸二酯酶(phosphodiesterase)C末端GAF结构域晶体结构(PDB代码: 2ZMF; SCOP代码:c2e4sA)为模板。采用Swiss-Pdbviewer程序(Guex and Peitsch, 1997)比较和显示蛋白质三级结构模型。

2 结果与讨论

2.1 数据抽提

我们从7个模式植物和1个非模式植物中共获得40条NIN-like蛋白质序列和相应的cDNA(表1)。比较以往的研究(Schauser et al., 2005), 我们在百脉根基因组数据中鉴别了2个新成员, 即*LjNLP3*(chr5.CM0-148.170.r2.a)和*LjNLP4*(chr3.CM0091.230.r2.m); 在水稻基因组数据库中鉴别了2个新成员, 即*OsNLP4*(Os11g16290)和*OsNLP5*(Os09g37710)。本研究保持以前研究中使用的*Nin-like*基因名称(Schauser et al., 2005)以便比较。

2.2 系统发育

排除3个结构域各自比对中的模糊区域和自行征插入后, 组合获得长度为293个氨基酸残基位点的矩阵。在AIC标准(Kullback and Leibler, 1951)下的最优进化模型为JTT(Jones et al., 1992), 数据异质性参数为 $G=1.258$, $I=0.063$ 。以小立碗藓(*Physcomitrella patens*)NIN-like蛋白作为外类群。采用ML分析产生最优树(图1)的lnL分值为-10 573.337 444。NJ分析产生几乎与ML法分析相同的拓扑结构, 差别主要存在于少数极短的分支上(图1)。我们构建的系统树显示(图1), 所有参与分析的被子植物NIN-like蛋白成员可以分为3个高度支持的分支, 分别定名为分支I、II和

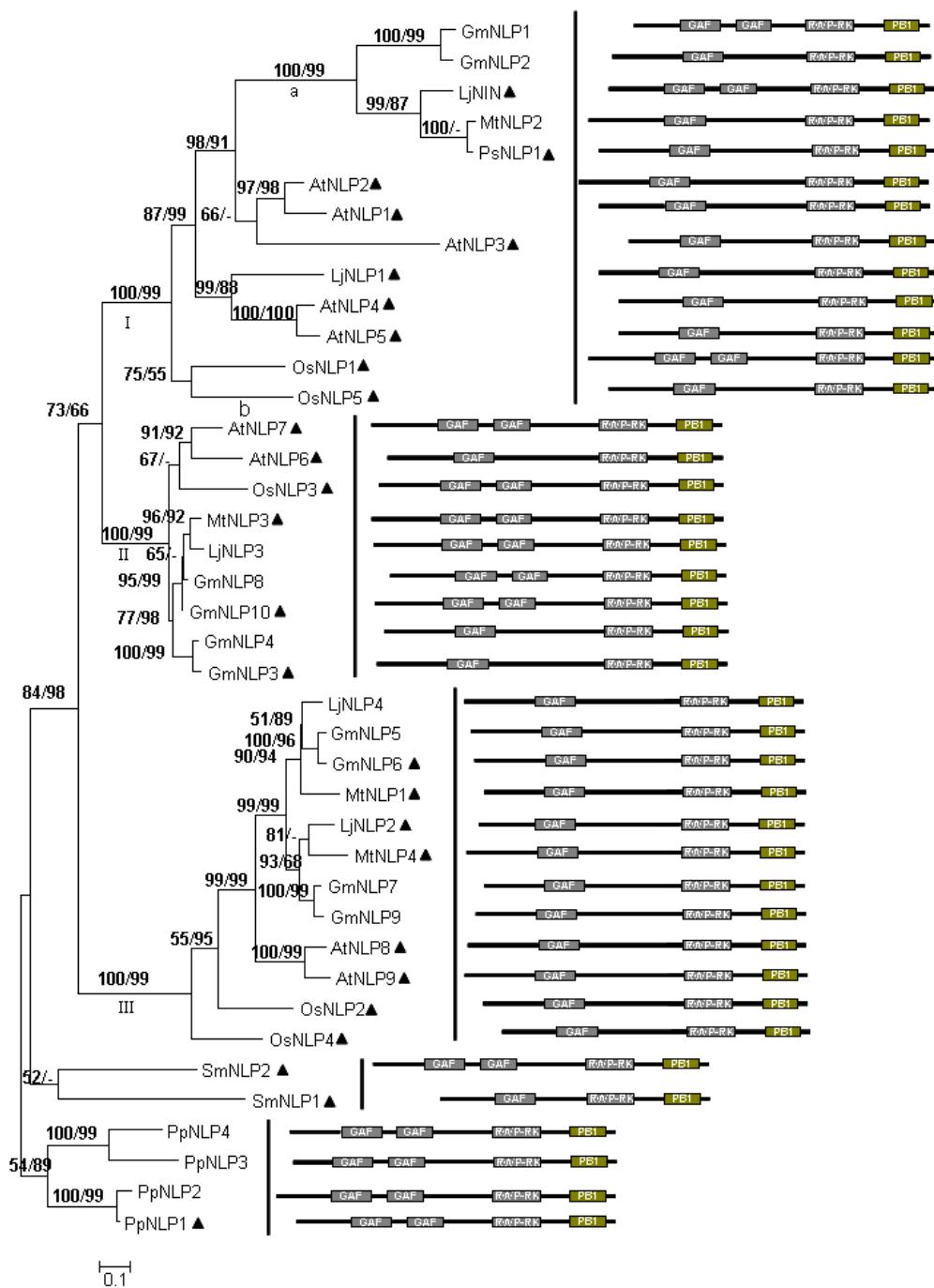


图1 NIN-like蛋白的ML系统发育树

分支上方分别为大于50的ML和NJ法bootstrap百分值; 分支下方为分支名称。- 表示在NJ分析中没有可分辨的分支关系; 三角指示该基因具有EST支持; 星号表示推测的早期基因重复事件。a代表来自豆科植物的NIN-like蛋白固氮分支。右侧为每个蛋白对应的结构域组织。

Figure 1 ML phylogenetic tree of NIN-like proteins

The numbers above branches are ML and NJ bootstrap percentages >50 , respectively; Those below branches are the clade name. “-” denote the relationships not resolved in NJ analysis; The genes with available EST sequences are indicated by triangles; The stars denote the inferred early duplication events. The lowercase letter “a” represents the nitrogen fixation subclade from legume plants. The domain organizations within each clade are depicted on the right.

III。这一结果重现了以前基于拟南芥、水稻和百脉根3个物种的分析(Schauser et al., 2005)。所不同的是, 我们分辨的3个分支均获得高度支持。在3个分支的关系上, 分析结果显示, 分支I和分支II为姐妹群, 并同时获得ML和NJ方法的低度支持, 提示它们可能有最近的共同起源; 而分支III显示为较早分化的一支。在分支I的内部, 由豆科成员组成的固氮亚支a获得高度支持(图1), 提示在分支I谱系进化中豆科植物*NIN-like*基因的一次祖先基因重复事件形成了这个亚支。

2.3 结构域组织的进化

为了追踪*NIN-like*蛋白结构域组织的歧异式样, 我们在SMATR平台上获取它们的结构域组织信息。所有参与分析的*NIN-like*蛋白在其C侧具有1个RWP-RK和1个PB1结构域, 而N侧的GAF结构域存在变化(图1)。在分支I和分支II的成员中, GAF结构域的数量为1–2个; 如果存在2个GAF结构域, 位于C侧的GAF结构域是退化(degeneration)的, 即发生显著变异或丢失, 而位于N侧的GAF结构域其长度和氨基酸组成具有高度保守性。在分支III中, 所有成员只保留了N侧的GAF结构域, 而C侧的GAF结构域则完全消失。由于在完成测序的基部陆生植物小立碗藓中所有*NIN-like*蛋白保持2个GAF结构域(图1), 我们认为具有2个GAF结构域的组织形式为祖先状态。在进化过程中, 位于C侧的GAF结构域逐渐丢失。

2.4 基因结构的进化

本研究对3个被子植物分支的模式植物*Nin-like*基因的结构进行了分析和比较, 表2显示了内含子数目和相位分析结果。我们发现, 在33个被分析的*Nin-like*基因中, 分支I的OsNLP5缺乏内含子, 其余基因的内

含子数目变异范围为3–9; 在32个具有内含子的基因中共有142个内含子, 平均每个基因4.44个; 每个基因的平均内含子数量在不同群之间也是不同的, 分支I、分支II和分支III分别为3.09、5.44和4.92个; 在142个内含子中, 75个(53%)相位是0, 56个(39%)相位是1, 11个(8%)相位是2。表3显示了内含子插入位置分析结果。由表3可知, 在所有被分析的*Nin-like*基因中存在3个保守的内含子插入位置, 即位置11、19和31。依据内含子插入位置和/或数量的变化, *Nin-like*基因可以划分为3类基因结构, 基本对应于系统发育分析的3个分支(图1)。在分支I中, 一般仅存在3个保守的内含子插入位置, 无其它内含子插入, 保守位置19在不同成员之间的变化是由于1个核苷酸对或一个密码子位置变化产生的(如位置16、17和18)。在分支II中, 位置11的上游和位置31的下游存在数目不等且随机的内含子插入。在分支III中, 与分支I中位置19的情况类似, 位置11也由于1个核苷酸对或1个密码子位置变化导致不同成员之间存在变异(如位置12和13), 在位置19和31之间以及位置31下游存在数目不等的随机内含子插入。

2.5 蛋白质功能位点和三维结构的歧异

所有40个*NIN-like*蛋白具有细胞核定位信号, 符合转录因子细胞区室定位特征。对3个分支的*NIN-like*蛋白的功能位点进行分析, 共鉴别了8个功能位点。其中, N-豆蔻酰化(N-myristylation)、N-糖基化(N-glycosylation)、蛋白激酶CK2磷酸化(casein kinase II phosphorylation)和蛋白激酶C磷酸化位点(protein kinase C phosphorylation)存在于所有被分析的蛋白序列中, 数目多且位置随机。另外4类功能位点选择性分布在不同的蛋白序列中, 位点数目为1–2个(表4)。cAMP和cGMP依赖的蛋白激酶磷酸化位点

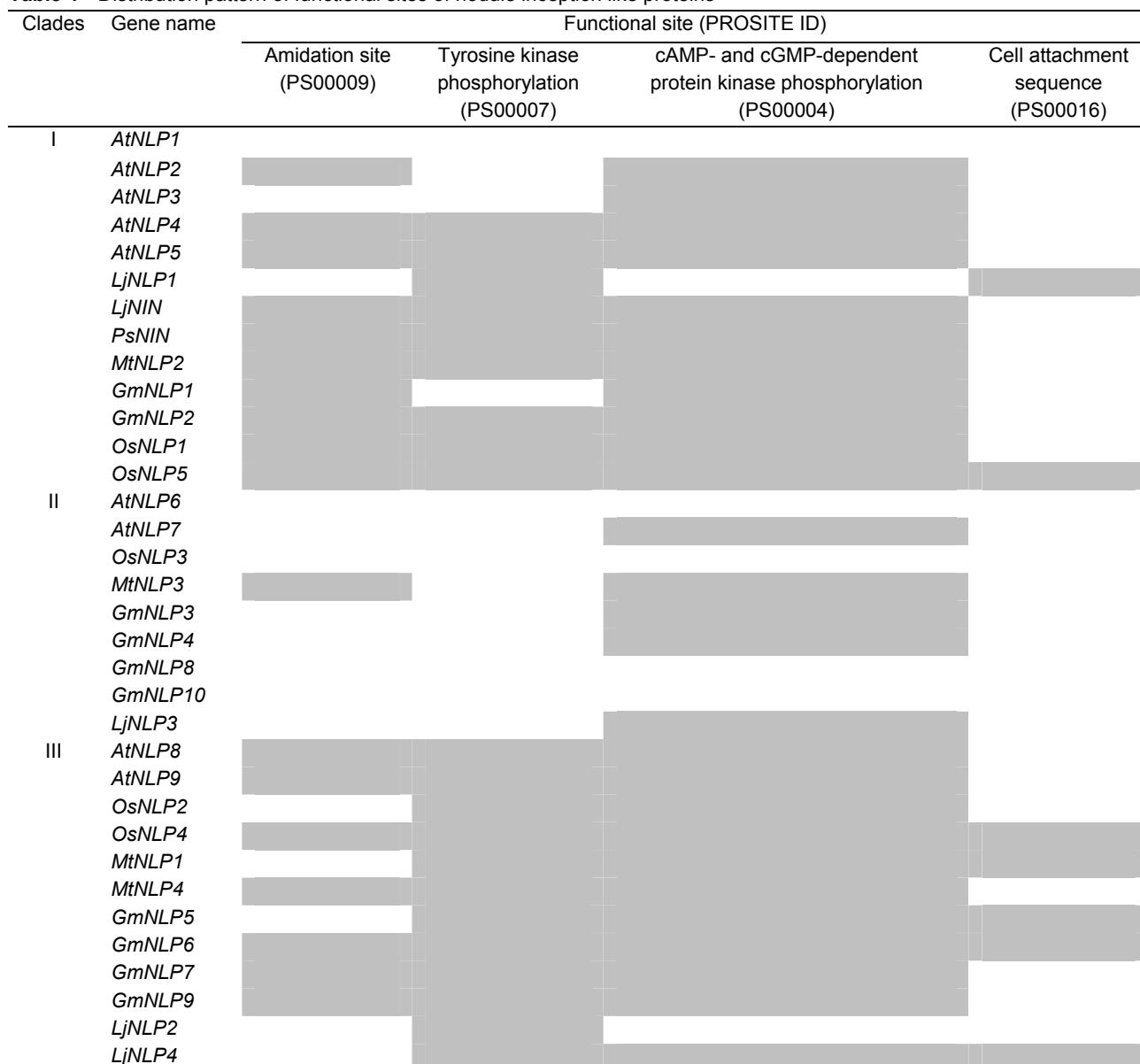
表2 根瘤感受样(*Nin-like*)基因的内含子数量和相位

Table 2 Phase and number of introns in nodule inception like genes

Clade (No. of genes)	No. of introns in each phase (%)			Total No. of introns	Mean No. of introns per gene
	0	1	2		
I (11)	17	14	3	34	3.09
II (9)	22	24	3	49	5.44
III (12)	36	18	5	59	4.92
Total (32)	75(53)	56(39)	11(8)	142	4.44

表 3 本研究取样的根瘤感受样(*Nin-like*)基因的内含子插入位置信息

Table 3 Position information of intron insertion in the nodule inception like genes sampled in this study

表4 根瘤感受样(NIN-like)蛋白功能位点的分布式样**Table 4** Distribution pattern of functional sites of nodule inception like proteins

表中阴影部分表示存在相应的功能位点。

The shaded part in this table represents the presence of corresponding functional sites.

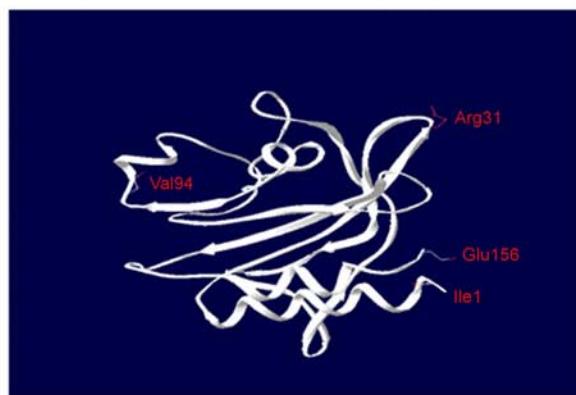
(cAMP- and cGMP-dependent protein kinase phosphorylation)广泛分布于3个分支的蛋白序列中,仅少数序列缺乏。酰胺化位点(amidation site)和酪氨酸激酶磷酸化位点(tyrosine kinase phosphorylation)一般分布在分支I和分支III中,在分支II中几乎完全缺乏。细胞附着序列(cell attachment sequence)零星分布在分支I和分支III的少数序列中,在分支II中则完全缺乏。不同分支的NIN-like蛋白功能位点分布的差别

意味着它们接受不同的信号因子,最终可能影响不同靶基因的表达。有趣的是,细胞附着序列,即Arg-Gly-Asp(RGD)三肽,以往只发现于胞外与细胞表面受体相互作用的蛋白质中(Ruoslahti and Pierschbacher, 1986)。

在N侧的GAF结构域序列中,豆科固氮亚支(图1的亚支a)的成员存在一个变异区域(图2A),研究推测这个变异区域可能导致百脉根的*Nin*基因被招募为根

A

AtNLP1	VKEPLLQAIIS GLNEAVQ-D- KDFLVQIIVVP IQ-QECKSFL TTMAQPHLFN QEQYSS--LAE YHVSETYNF
AtNLP2	VKEPLVQALE GLNNEEVQ-D- KDFLQIQIILP IQ-QECKHNL TTSEQPHFFN PKYSS--LKR YRDVSAYNF
OsNLP5	VPERFDQALAA YIRETQS-D- ADVLVQLWVP VKCNGDQLVL TTSCQPFILD QRSNS--LIQ PREVSTKYQF
OsNLP1	-GD----- GELLVQWVVP TR-IGD RQVL TTCCQPPFLD RRNQR--LAN YRVSMKQYF
AtNLP4	VTEPLVQAVKE HIKDYTT-A- RGSLIQLWVP VN-RGCKWVL TTKEQPFSHD PLCQR--LAN YREISVNHF
AtNLP5	--ERLVQAVT HIKDFTS-E- RGSLIQLWVP VD-RGCKWVL TTKEQPFSHD PMCQR--LAH YREISENYQF
LjNLP1	IMEKLPLALK WIRQFNU-N- KDMLIQIIVVP VP-RGD RPIL SANNLPPSL SCSEN--LAR YREISECFQF
AtNLP3	LKEPVACAMG HLQEVCM-E- RELLTLQWVP VETRSCR-VL STEROFPYSIN TESOSOSLAL YEDASAGYSE
MtNLP2	VKEPLVVAVG YLKKEYTKNSS NWVLIQIIVVP MRQ-----RSALIHTQN HYLQQESSSA PWSVN-----
PsNIN	-----
LjNIN	VKEPLVIAVG YLKKEYTKNSS -NVLIQIIVVP LRPQGILHDHD YHTHNYLLSNN PP PQP EAAAD HESVSLGFPFM
GmNLP2	VKEPLVIAVG YLKKEYAKNS- -NLPIQWVVP ERQ-----SARAQPQDN IYPYAAALIN GDAAAAFQIQE
GmNLP3	VKEPLVIAVG YLKKEYAKNS- -NLLIQLWWP ERQ-----SARAQPQDN -YPY-AALIN -TTSAFQFQE
AtNLP1	PAD EGHMD--- FVGLPGR VF LQKFPEWT PDVRFFRSD E YPRIKEAQRC DWRGS LALPV FERGSCTCLG
AtNLP2	LAD EDHSKE--- SVGLPGR VF LKKLPEWT PDVRFFRSHE YPRIKRAEQC DWRGS LALPV FERGSCTCLG
OsNLP5	SADVASGS--- SPGLPGR VF FIGRLPWEWT PDVKYFTSYE YPRINHAQYL DVHCTMGLPV FERGNVYSCLG
OsNLP1	SAD ESRAPA--- DLGLPGR VF VFGVRVPEWT PDVKYFTSYE YPRVQHQAQYF DIRGSVALPV FEP RPSA CLG
AtNLP4	SAR QDD S--- KALAGLPGK VF LGKLPPEWT PDVRFFKSEE YPRVQHQAQDC DVRGTLALPV FEQGSKICLGL
AtNLP5	STE QED DS--- SIDLVGLPGR VF LGKLPPEWT PDVRFFKSEE YPRVQHQAQDC DVRGTLALPV FEQGSQICLGL
LjNLP1	SAR EDSK--- ELVPGLPGR VF RDOKVPEWT PDVRFFRSD E YPRVEHAREP DICGT LALPV FEQGSRTCLG
AtNLP3	AAEVGS EQ--- LVGGLPGR VF LRRMPPEWT PDVRFFRSD E YPRIGYARRY QVRAT LALPL FQQTSGNCVVA
MtNLP2	---- PN ----- MN VHVRFFRSHD YPR-HQQQQQ -YGSLLALPV FERGSCTCLG
PsNIN	----- RFFRSHE YPR-HQQQQQ QYGSLLALPV FERGSCTCLG
LjNIN	PAA PNSNLY--- SSVHVRFFRSHE YPR-UQAAQY ---GS LALPV FERGTCTCLG
GmNLP2	DWVHUNDQY--- WT PNVVRFFRSHE YPR-HLRTPG ---SLALPV FERTGTA MCLG
GmNLP3	DWVH---DQ--- WT PNVVRFFRSHE YPR-HLRPPG ---SLALPV FESGSAMCLV
AtNLP1	VVEIVTTTQK -MNYPQEELEK MCKALAEAVDL RSS-
AtNLP2	VVEIVTTTQK -MNYPQEELEK ICKALES---
OsNLP5	VIELIMTQK -LNFTSELNT ICSALQAVNL TSTE
OsNLP1	VVELVMTTQK -VNYSABIAEN ICNALKEVDL RSSD
AtNLP4	VIEVVMTTQK -VKLRPPELES IC RALQAVDL RSTE
AtNLP5	VIEVVMTTQK -VKLSPDLES IC RALQAVDL RSTE
LjNLP1	VIEVVMTTQK -INVVPQLES VCKALEVVDL ---
AtNLP3	VMEMVTTTQK -LEYASQSLST ICHALEAEPDL RTSQ
MtNLP2	VIEFVISONQT LINYRPQLDH LSNAL-EAVD FRSS
PsNIN	VIEFVIANQN LINYRPQLDH LSNAL-EAVD FRSS
LjNIN	VLEIVITNQQT TINYN---- VSNALDQAVD FRSS
GmNLP2	----- VVEIL-----
GmNLP3	PDRECTYRT----- IVEVLT CVCVK A---

B**图2** LjNLP1蛋白GAF结构域的三级结构模型

(A) GAF结构域的多序列比对(阴影部分示发生在固氮亚支a(图1)中的显著变异区段); (B) LjNLP1蛋白GAF结构域的三级结构模型标注的氨基酸位置代表LjNLP1的GAF结构域的起点(Ile1)和终点(Glu156)以及对应于豆科植物变异区的起点(Arg31)和终点(Val94)。

Figure 2 Three-dimensional structural model of GAF domain in LjNLP1 protein

(A) Multiple sequence alignment of GAF domain regions; The shaded part indicates the significantly variable region occurred in NIN-like proteins from legume plants (see the clade a in Figure 1); (B) Three-dimensional structural model of GAF domain in LjNLP1 protein; The labelled amino acids, Ile1 and Glu156, represent the starting and ending points of GAF domain in LjNLP1, respectively, and in which, the Arg31 and Val94 represent the starting and ending points corresponding to the variable part in legume plants.

表5 已知功能的根瘤感受样(*Nin-like*)基因**Table 5** Known functions of nodule inception like genes

Clade	Gene name	Function	Reference
I	<i>LjNIN</i> , <i>PsNIN</i> , <i>OsNLP1</i> , <i>AtNLP3</i>	<i>LjNIN</i> and <i>PsNIN</i> involved in early stages of root nodule formation; <i>OsNLP1</i> did not rescue abortion of infection in <i>nin</i> mutant plants, whereas <i>LjNIN</i> did; <i>AtNLP3</i> responded quickly to N nutrition	Schauser et al., 1999; Borisov et al., 2003; Scheible et al., 2004; Yokota et al., 2010
	<i>AtNLP7</i>	Response to N nutrition; may play a role in stomatal movements and drought resistance	Castaings et al., 2009
	<i>AtNLP8</i>	Response to N nutrition	Scheible et al., 2004

瘤固氮基因(Schauser et al., 2005)。本研究显示在取样的3个豆科模式植物(大豆、蒺藜苜蓿(*Medicago truncatula*)和百脉根)中均存在这样1个变异区。对分支I的*LjNLP1*的GAF结构域进行远程同源建模(图2B)发现, 对应于豆科固氮亚支N侧GAF结构域变异区域的是一段保守的β-折叠+转角+α-螺旋构象, 范围从Arg(R)31到Val(V)94(图2B)。其中Glu(E)73到Glu(E)89片段对应于*LjNIN*中的一段丢失区域, 这一片段包含一个α-螺旋区。由于固氮亚支在进化过程中发生N侧GAF结构域的变异或片段丢失, 可能改变了这些基因的功能, 分化为根瘤固氮基因。

2.6 讨论

基因家族是真核生物的特征之一(Horan et al., 2005)。家族内的基因在经历进化选择过程后, 最终形成不同的直系同源群以执行不同的功能(Lespinet et al., 2002)。通过扩大豆科固氮植物物种的取样, 有助于加深理解根瘤感受样(*Nin-like*)多基因家族的结构歧异式样与功能分化的关系。我们的分析不仅重现了以前划分的3个直系同源群, 且鉴别了它们之间的系统发育关系, 即分支I和分支II的姐妹群关系(图1)。在分支I中, 固氮亚支a显示起源于分支内的一次祖先*NIN-like*基因的重复事件, 其后裔基因由于GAF结构域的变异和/或其它未知的机制获得根瘤固氮能力。分支III的成员一致地丢失1个GAF结构域。由于基部陆地植物*NIN-like*蛋白包含2个GAF结构域, 按照进化的简约规则, 我们认为具有2个GAF结构域的*NIN-like*蛋白为祖先结构形式。

*Nin-like*基因表现出结构多样性(表3), 意味着在其进化过程中发生了频繁的内含子插入和丢失(Park et al., 2008)。由于内含子可能有调控功能, 内含子的

插入和丢失可能促成基因的功能分化, 它们通过促进外显子重排或直接导致调控元件的差异(Lynch and Conery, 2000); 另一方面, 3个分支内部在内含子位置11、19和31是高度保守的(表3), 处于强选择性压力下, 意味着这些基因具有相似的功能。我们认为, 无内含子的*OsNLP5*基因可能起源于基因返座(Roy et al., 2003)而不是内含子丢失事件, 原因为其潜在的返座源基因*OsNLP1*具有3个内含子(表3), 而发生3次基因丢失事件的可能性很低(Zhu et al., 2012)。一般认为, 返座基因由于缺乏调控元件而不能表达(Graur and Li, 2000), 但有研究显示返座基因不仅能够表达且能够承担非冗余的功能(Kong et al., 2004)。*OsNLP5*和它潜在的返座源基因*OsNLP1*都具有作为活跃转录基因的EST证据(图1中的分支b), 它们在表达式样上的差异值得进一步关注(Sakai et al., 2011)。

在3个直系同源群中, 一些成员的功能已经确定(表5)。分支I的固氮亚支a的*LjNIN*和*PsNIN*, 功能上与固氮根瘤的早期形成有关(Schauser et al., 1999, 2005; Borisov et al., 2003), 这2个基因都响应根瘤细菌的微共生条件而形成过度的根毛卷曲。分支I的*OsNLP1*不能替代百脉根*LjNIN*的功能(Yokota et al., 2010)。尽管*OsNLP1*与*LjNIN*同属一个进化分支, 但它们的功能已经分化。进一步证明豆科植物根瘤固氮能力是一种后发特征, 起源于不具有根瘤固氮能力的同源基因。分支I的*AtNLP3*能够响应硝酸盐的N源信号, 表明尽管*AtNLP3*不具有根瘤固氮能力, 但与N源的利用有关, 处于N源利用的基因网络中(Scheible et al., 2004)。在分支II中, *AtNLP7*的功能已被初步确定。*AtNLP7*突变显示典型的氮饥饿表型, 并影响硝酸盐利用相关基因(如硝酸盐吸收和诱导基因)的功能, 因

此, *AtNLP7*与*AtNLP3*功能类似, 都涉及硝酸盐利用和代谢调控(Castaings et al., 2009)。有趣的是, 尽管2个基因处于不同的进化分支, 两者功能位点分布却是相同的(表4), 因此它们在功能上的分化需要进一步研究。在分支III中, 拟南芥*AtNLP8*同样响应硝酸盐的N源信号(Scheible et al., 2004)。值得注意的是, *AtNLP8*与*AtNLP7*和*AtNLP3*的功能位点分布不同(表4), 表明尽管它们都涉及响应硝酸盐的N源信号, 但可能接受不同的调控信号。以上研究提示, 不同分支的*Nin-like*基因均涉及N源利用过程, 发挥不同但相互关联的功能。由于不同的进化背景, GAF的功能可能不同, 但其中有一个确定的功能是cGMP-binding(Aravind and Ponting, 1997)。研究发现, 与其它非豆科植物相比较, 豆科植物百脉根的NIN-like蛋白的GAF结构域存在一段显著变异区域, 且正是这个变异区域导致百脉根获得固氮能力。本研究中蛋白质三级结构比较分析显示, 在这个变异区域内, 处于同一个进化分支的百脉根NIN-like蛋白(LjNIN1)具有保守的 β -折叠+转角+ α -螺旋三级结构构象(图2B), 而固氮植物可能发生了变异(如LjNIN)或丢失(如PsNIN)(图2A)。由于蛋白质功能常常与它的构象密切相关(Hou et al., 2007), 在核心结构区域的突变能急剧改变结合口袋(binding pocket)的形状。我们的建模分析提示, 固氮植物可能通过改变GAF结构域对信号分子的结合潜能或性质影响靶基因的转录, 并最终获得固氮能力。

参考文献

- Abascal F, Zardoya R, Posada D** (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105.
- Aravind L, Ponting CP** (1997). The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends Biochem Sci* **22**, 458–459.
- Barbulova A, Rogato A, D'Apuzzo E, Omrane S, Chiurazzi M** (2007). Differential effects of combined N sources on early steps of the Nod factor-dependent transduction pathway in *Lotus japonicus*. *Mol Plant Microbe Interact* **20**, 994–1003.
- Borisov AY, Madsen LH, Tsyganov VE, Umehara Y, Voroshilova VA, Batagov AO, Sandal N, Mortensen A, Schauser L, Ellis N, Tikhonovich IA, Stougaard J** (2003). The *Sym35* gene required for root nodule development in pea is an ortholog of *Nin* from *Lotus japonicus*. *Plant Physiol* **131**, 1009–1017.
- Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang XH, Mudge J, Vasdevani J, Schiex T, Spannagi M, Monaghan E, Nicholson C, Humphray SJ, Schoof H, Mayer KFX, Rogers J, Quétier F, Oldroyd GE, Debelle F, Cook DR, Retzel EF, Roe BA, Town CD, Tabata S, Van de Peer Y, Young ND** (2006). Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc Natl Acad Sci USA* **103**, 14959–14964.
- Castaings L, Camargo A, Pocholle D, Gaudon V, Texier Y, Boutet-Mercey S, Taconnat L, Renou JP, Daniel-Vedele F, Fernandez E, Meyer C, Krapp A** (2009). The nodule inception-like protein 7 modulates nitrate sensing and metabolism in *Arabidopsis*. *Plant J* **57**, 426–435.
- Felsenstein J** (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368–376.
- Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A** (2005). Protein identification and analysis tools on the ExPASy server. In: Walker JM, ed. *The Proteomics Protocols Handbook*. Totowa: Humana Press. pp. 571–607.
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS** (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40** (Database issue), D1178–D1186.
- Graur D, Li WH** (2000). *Fundamentals of Molecular Evolution*, 2nd edn. Sunderland (MA): Sinauer Associates. pp. 336–337.
- Guex N, Peitsch MC** (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723.
- Guindon S, Gascuel O** (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696–704.
- Horan K, Laurica J, Bailey-Serres J, Raikhel N, Girke T** (2005). Genome cluster database. A sequence family analysis platform for *Arabidopsis* and rice. *Plant Physiol* **138**, 47–54.
- Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K** (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res* **35** (Web Server issue), W585–W587.

- Hou LM, Honaker MT, Shireman LM, Balogh LM, Roberts AG, Ng KC, Nath A, Atkins WM** (2007). Functional promiscuity correlates with conformational heterogeneity in A-class glutathione S-transferases. *J Biol Chem* **282**, 23264–23274.
- Ito T, Matsui Y, Ago T, Ota K, Sumimoto H** (2001). Novel modular domain PB1 recognizes PC motif to mediate functional protein-protein interactions. *EMBO J* **20**, 3938–3946.
- Jones DT, Taylor WR, Thornton JM** (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275–282.
- Kelley LA, Sternberg MJ** (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* **4**, 363–371.
- Kong H, Leebens-Mack J, Ni W, dePamphilis CW, Ma H** (2004). Highly heterogeneous rates of evolution in the SKP1 gene family in plants and animals: functional and evolutionary implications. *Mol Bio Evol* **21**, 117–128.
- Kullback S, Leibler RA** (1951). On information and sufficiency. *Ann Math Stat* **22**, 79–86.
- Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J** (2005). The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res* **33** (Database issue), D71–D74.
- Lespinet O, Wolf YI, Koonin EV, Aravind L** (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* **12**, 1048–1059.
- Lynch M, Conery JS** (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155.
- Park KC, Kwon SJ, Kim PH, Bureau T, Kim NS** (2008). Gene structure dynamics and divergence of the polygalacturonase gene family of plants and fungus. *Genome* **51**, 30–40.
- Rogozin IB, Lyons-Weiler J, Koonin EV** (2000). Intron sliding in conserved gene families. *Trends Genet* **16**, 430–432.
- Roy SW, Fedorov A, Gilbert W** (2003). Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci USA* **100**, 7158–7162.
- Ruosahti E, Pierschbacher MD** (1986). Arg-Gly-Asp: a versatile cell recognition signal. *Cell* **44**, 517–518.
- Saitou N, Nei M** (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.
- Sakai H, Mizuno H, Kawahara Y, Wakimoto H, Ikawa H, Kawahigashi H, Kanamori H, Matsumoto T, Itoh T, Gaut BS** (2011). Retrogenes in rice (*Oryza sativa* L. ssp. *japonica*) exhibit correlated expression with their source genes. *Genome Biol Evol* **3**, 1357–1368.
- Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, Fujishiro T, Katoh M, Kohara M, Kishida Y, Minami C, Nakayama S, Nakazaki N, Shimizu Y, Shinpo S, Takahashi C, Wada T, Yamada M, Ohmido N, Hayashi M, Fukui K, Baba T, Nakamichi T, Mori H, Tabata S** (2008). Genome structure of the legume, *Lotus japonicus*. *DNA Res* **15**, 227–239.
- Schauser L, Roussis A, Stiller J, Stougaard J** (1999). A plant regulator controlling development of symbiotic root nodules. *Nature* **402**, 191–195.
- Schauser L, Wieloch W, Stougaard J** (2005). Evolution of NIN-like proteins in Arabidopsis, rice, and *Lotus japonicus*. *J Mol Evol* **60**, 229–237.
- Scheible WR, Morcuende R, Czechowski T, Fritz C, Osuna D, Palacios-Rojas N, Schindelasch D, Thimm O, Udvardi MK, Stitt M** (2004). Genome-wide reprogramming of primary and secondary metabolism, protein synthesis, cellular growth processes, and the regulatory infrastructure of Arabidopsis in response to nitrogen. *Plant Physiol* **136**, 2483–2499.
- Schmutz J, Cannon SB, Schlueter J, Ma JX, Mitros T, Nelson W, Hyten DL, Song QJ, Thelen JJ, Cheng JL, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du JC, Tian ZX, Zhu LC, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA** (2010). Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183.
- Schultz J, Milpetz F, Bork P, Ponting CP** (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci USA* **95**, 5857–5864.
- Suyama M, Torrents D, Bork P** (2011). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34** (Web Server issue), W609–W612.

- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S** (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731–2739.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG** (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882.
- Wheelan SJ, Church DM, Ostell JM** (2001). Spidey: a tool for mRNA-to-genomic alignments. *Genome Res* **11**, 1952–1957.
- Yokota K, Soyano T, Kouchi H, Hayashi M** (2010). Function of GRAS proteins in root nodule symbiosis is retained in homologs of a non-legume rice. *Plant Cell Physiol* **51**, 1436–1442.
- Zhu XY, Chen CY, Wang BH** (2012). Phylogenetics and evolution of *Trx SET* genes in fully sequenced land plants. *Genome* **55**, 269–280.

Evolution of the Nodule Inception-like Genes: Structural Divergence and Functional Differentiation

Xinyu Zhu*, Wansheng Lü, Chunmei Yu, Baohua Wang

School of Life Sciences, Nantong University, Nantong 226019, China

Abstract The nodule inception gene *Nin* is related to early stages of root nodule formation in the model plant *Lotus japonicus*. Functionally, its homologs, *Nin-like* genes, are also involved in N metabolism. In this study, *Nin-like* gene data were retrieved from completely sequencing genomes of legume and non-legume plants for investigating phylogeny to trace the divergence of genes and proteins and build relationships between structural divergence and functional differentiation. We revealed novel *Nin-like* genes; previously resolved orthologous groups were recovered and their sister relationships resolved. *Nin-like* genes showed a diversity of structures, which supports the results of phylogenetic and functional studies; *OsNLP5* from *Oryza sativa* was found to be intronless, which probably originated from a gene retroposition event. We found discrepancies in domain organizations and functional sites among *NIN*-like proteins, which indicates possible functional differentiation. 3-D structural analysis revealed the occurrence of significant conformational changes in GAF domains of *NIN*-like proteins from nodule legume plants, which could become the basis of recruitment as nodule inception genes. Our findings will contribute to the design of further experimental studies.

Key words evolution, functional differentiation, *Nin-like* genes, rhizobia symbiosis, structural divergence

Zhu XY, Lü WS, Yu CM, Wang BH (2013). Evolution of the nodule inception-like genes: structural divergence and functional differentiation. *Chin Bull Bot* **48**, 519–530.

* Author for correspondence. E-mail: zhuxinyu@ntu.edu.cn

(责任编辑: 白羽红)