doi:10.3969/j.issn.1672-5565.2013.01.09

木质素生物合成酶 CCR 基因的生物信息学分析

陈 刚[1,2],徐秉良[1,2*],白江平[1,2]

(1. 甘肃农业大学草业学院、草业生态系统教育部重点实验室、中-美草地畜牧业可持续发展研究中心、甘肃省草业工程实验室,甘肃 兰州 730070;2. 甘肃省干旱生境作物学重点实验室-甘肃省作物遗传改良与种质创新重点实验室,甘肃 兰州 730070)

摘 要:肉桂酰辅酶 A 还原酶(Cinnamoyl-CoA reductase,CCR)是催化木质素特异途径的第一个关键酶,是调节碳素流向木质素潜在的控制关节点,对木质素单体的生物合成起着重要作用。通过 NCBI 数据库收集来自裸子植物、单子叶植物及双子叶植物的35条 CCR 基因的完整信息,对35条 CCR 基因的 cDNA 及其编码的氨基酸序列的进化规律、理化性质、结构域、导肽、信号肽、跨膜结构域、亲/疏水性以及蛋白质结构等性状进行了生物信息学分析与预测,构建了 CCR 基因的系统发育树。分析结果表明,单子叶植物 CCR 基因中 GC 的含量明显高于双子叶植物;CCR 基因编码的氨基酸序列存在9个保守区域;所编码氨基酸的理化性质基本一致,但单子叶、双子叶及裸子植物的 CCR 基因编码主要氨基酸的种类和含量存在着差异;CCR 蛋白的 N-端存在一个脱氢酶/差向异构酶/辅酶Ⅰ结合蛋白的结构域,无导肽、信号肽及跨膜结构域,属亲水性蛋白;进化树绘制以及同源建模结果表明,CCR 基因的进化和植物的进化基本一致,CCR 蛋白三级结构模型的空间结构稳定,建模结果可靠。分析结果对于深入研究 CCR 蛋白在木质素合成中的作用具有一定的理论指导意义。

关键词:木质素,CCR 基因,生物信息学,进化树,同源建模

中图分类号:Q518.2 **文献标识码**:C **文章编号**:1672-5565(2013)-01-050-08

Bioinformatics analysis of CCR—one of the key enzymes mediated lignin biosynthesis in plants

CHEN Gang[1,2], XU Bing-liang[1,2*], BAI Jiang-ping[1,2]

(1. *College of Grassland Science of Gansu Agricultural University/Key Laboratory of Grassland Ecosystem of MOE/ Sino-U. S. Centers for Grazing Land Ecosystem Sustainability/Pratacultural Engineering Laboratory of Gansu Province, Lanzhou* 730070, *China*;
2. *Gansu Key Lab of Crop Improvement & Germplasm Enhancement, Lanzhou* 730070, *China*)

Abstract: The Cinnamoyl-CoA reductase (CCR) is the first key enzyme that catalyzed the specific pathway in lignin biosynthesis, it is also the speed limit enzymne that mediated the flow of the Carbon to the lignin. It thus plays an important role in the biosynthesis of lignin monomer. In order to fully understand the characteristics of this enzyme, a total of 35 CCR genes belonging to monocotyledon, dicotyledon and gymnosperm were selected from NCBI database, the bioinformational methods were employed to analyze the amino acid composition, conserved functional domains, physical and chemical characteristics, as well as the leader peptides, signal protein, transmembrane domain and Hydrophobicity/hydrophilicity of peptides coding by those genes. The phylogenetic tree of CCR genes was constructed and three-dimensional structure of CCR proteins were predicted and analyzed. The results showed that GC content in gymnosperm is higher than monocotyledon. Nine conserved domains were existed in all tested CCR geness and those domains shared great similarity among genes. The physical and chemical characteristics of CCR peptides are similar, but the primary structure of proteins are significantly different from each other. the data showed that one dehydrogenase/epimerase/NAD binding domain was located at the N-terminus of CCR. The three-dimensional structure of CCR revealed that this protein is stable. No obvious leader or signal peptide and trans-

收稿日期:2012-08-08;**修回日期**:2012-08-30.

基金项目:国家自然科学基金项目(31060261,30671267)和甘肃省自然科学基金项目(1107RGZA152),共同资助。

作者简介:陈刚,男,甘肃兰州人,在读硕士,主要从事植物病理学的研究。E-mail:chengang246810@163.com.

*通讯作者:徐秉良,E-mail:xubl@gsau.edu.cn.

membrane domain were detected suggested that these enzymes are hydropholicity protein and mainly functioning in the cytosol of the cell. The results also suggested that the evolution of CCR genes might corresponding to the development of the kindom of 1plants. Therefore, the results provided great information on the CCR and be for the further research of CCR protein in lignin biosythesis.

Key words: Lignin, CCR Gene, Bioinformatics, Phylogenetic Tree, Homology Modeling

木质素是一种具有芳香族特性的三维高分子化合物。作为地球上含量仅次于纤维素的天然有机物,木质素具有重要的生理功能,特别是在植物抗倒伏、抗病和抗逆境方面发挥着重要的作用[1]。木质素在植物体内的生物合成途径尚不完全清楚,但普遍认为大致包括莽草酸途径、苯丙氨酸途径及木质素特异合成途径 3 个主要阶段,目前对于木质素生物合成途径的研究越来越多的集中在木质素的特异合成途径上[2]。

肉桂酰辅酶 A 还原酶(CCR)作为催化木质素特异途径的第一个关键酶,催化 3 种羟基肉桂酸的 CoA 酯还原生成相应的肉桂醛,可能对木质素合成途径的碳流具有潜在的调控作用,是调节碳素流向木质素潜在的控制关节点,对木质素单体的生物合成起着重要作用[3]。因此对 CCR 的研究将有助于对植物木质素生物合成途径的进一步研究。

到目前为止,已从拟南芥、大麦、小麦、番茄等多种植物中克隆得到了 CCR 基因的全长或部分编码序列[4],但对 CCR 基因缺乏系统的生物信息学分析和研究报道,特别是 CCR 基因编码氨基酸序列的保守区域、CCR 蛋白导肽、信号肽、亚细胞定位、跨膜结构域、功能位点及三级结构的研究尚未见报道。为此,本研究拟采用生物信息学工具与分析方法,对 NCBI 数据库中分别来自裸子植物、单子叶植物及双子叶植物的 35 条 CCR 基因完整 cDNA 及其编码的氨基酸序列进行数据挖掘,旨在为对 CCR 基因的进一步研究和利用提供一定的理论依据。

1　材料与方法

1.1　材料

数据资料来源于 NCBI 数据库中已注册的,分别来自裸子植物、单子叶植物及双子叶植物共计 35 条 CCR 基因的核酸及其编码的氨基酸序列(表 1)。

1.2　方法

利用 NCBI 中的 ORF Finder 和 BioXM 2.6 软件对 CCR 基因完整 cDNA 序列的 GC 含量进行分析;采用 ClustalX 和 Mega4 软件构建 CCR 基因的系统发生树;通过 NCBI 的 Conserved Domains 数据库,对 CCR 基因编码的氨基酸序列进行保守区分析;采用 ExPASy、SMART、Post Prediction、TargetP 1.1 Server、SignalP 4.0 Server、TMHMM Server v. 2.0、ProtScale 及 Cn3D 对 CCR 基因编码的主要氨基酸的平均含量、理化性质、CCR 蛋白结构域、亚细胞定位、导肽、信号肽、跨膜结构域、亲/疏水性以及 CCR 基因编码的氨基酸的活性位点、NADP 结合位点及底物结合位点进行预测和分析;最后采用 Swiss - Model 对 CCR 基因编码蛋白质的三级结构进行同源建模,并用 PyMOL 对建模结果进行处理。

2　结果与分析

2.1　CCR 基因完整 cDNA 序列的分析

2.1.1　CCR 基因 GC 含量的分析

采用 NCBI 中 ORF Finder 和 BioXM 2.6 对 35 条 CCR 基因完整 cDNA 序列进行 GC 含量分析[5](表 2),结果表明,单子叶植物 CCR 基因的 GC 含量,尤其是编码区 GC 含量远高于双子叶植物。单子叶植物中甘蔗 CCR 的 GC 含量最高,达 69.89%;黑麦草 CCR2 的 GC 含量最低,为 61.34%,平均为 66.92%;而双子叶植物中 GC 含量最高为大叶相思,达 53.71%,苜蓿的最低,为 41.27%,平均 48.84%。

2.1.2　CCR 基因系统发育树的构建

采用 ClustalX 程序对 35 条 CCR 基因的完整 cDNA 序列进行多重比对(采用默认的 IUB 记分矩阵),采用 Mega4 程序对产生的多重比对结果构建系统发育树(采取最大简约法),并采用随机逐步比较的方式搜索最佳系统树,对生成的系统发育树进行 Bootstrap 校正,最终生成系统发育树[6](图 1)。

表 1 不同植物 CCR 基因 cDNA 及其编码氨基酸的序列号

Table1 The sequence number of cDNA and the corresponding amino acid of CCR in different plants 物种

	物种 Species	CCR 基因 CCR Gene	cDNA 序列号 Number of cDNA	氨基酸的序列号 Number of amino acid
单子叶植物 (monocotyledon)	大麦 Hordeum vulgare	HvCCR	AY149607	AAN71760
	小麦 Triticum aestivum	TaCCR	AY771357	AAX08107
	水稻 Oryza sativa	OsCCR	GQ848067	ADM86880
	甘蔗 Saccharum officinarum	SoCCR	AJ231134	CAA13176
	狗尾草 Pennisetum purpureum	PpCCR	HQ889311	ADY39751
	黑麦草 Lolium perenne	LpCCR1	AY061888	AAL47182
		LpCCR2	AF278698	AAG09817
	玉米 Zea mays	ZmCCR1	NM001112018	NP001105488
		ZmCCR2	NM001112245	NP001105715
双子叶植物 (dicotyledon)	板蓝根 Isatis tinctoria	ItCCR	GQ872418	ADC40029
	草莓 Fragaria ananassa	FaCCR	AY285922	AAP46143
	大叶相思 Acacia auriculiformis	AaCCR	DQ001168	AAY86360
	丹参 Salvia miltiorrhiza	SmCCR	JF784010	AEB69789
	党参 Codonopsis lanceolata	CiCCR	AB243011	BAE48787
	番茄 Lycopersicon esculentum	LeCCR1	DQ019125	AAY41879
		LeCCR2	NM001247368	NP001234297
	光皮桦 Betula luminifera	BlCCR	FJ410450	ACJ38670
	红麻 Hibiscus cannabinus	HcCCR	HM151381	ADK24219
	胡杨 Populus tremuloides	PtCCR	AF217958	AAF43141
	辣椒 Capsicum annuum	CaCCR	EU616555	ACF17647
	蓝莓 Vaccinium corymbosum	VcCCR	FJ197338	ACI14382
	马铃薯 Solanum tuberosum	StCCR	AY149608	AAN71761
	毛白杨 Populus tomentosa	PtoCCR	AY479973	AAR83344
	苜蓿 Medicago truncatula	MtCCR	XM003604190	XP003604238
	拟南芥 Arabidopsis thaliana	AtCCR1	NM101463	NP173047
		AtCCR2	NM106730	NP178197
	泡桐 Paulownia sp.	PsCCR	EU338487	ACD13265
	沙梨 Pyrus pyrifolia	PpyCCR	GU138672	ADK62523
	橡胶树 Hevea brasiliensis	HbCCR	HQ229954	ADU64758
	银合欢 Leucaena leucocephala	LlCCR	DQ986907	ABL01801
	油茶 Camellia oleifera	CoCCR	FJ883579	ACQ41893
	麻风树 Jatropha curcas	JcCCR	GQ149700	ACS32301
	甘蓝 Brassica oleracea	BoCCR	HM805092	AEK27175
裸子植物 (gymnosperm)	火炬松 Pinus taeda	PtaCCR	AY064169	AAL47684
	银杏 Ginkgo biloba	GbCCR	HQ901358	AEO13438

表 2　CCR 基因 ORF 长度及 GC 含量

Table 2　The ORF length and GC content of CCR geneCCR 基因

	CCR 基因 CCR Gene	ORF 长度(bp) Length of ORF	ORF 中 GC 含量(%) GC content of ORF(%)	CCR 基因 CCR Gene	ORF 长度(bp) Length of ORF	ORF 中 GC 含量(%) GC content of ORF(%)
单子叶植物 (monocotyledon)	HvCCR	1047	67.72	LpCCR1	1089	68.51
	TaCCR	1074	64.71	LpCCR2	103561.34	
	OsCCR	1120	64.41	ZmCCR1	1116	69.81
	SoCCR	1119	69.89	ZmCCR2	1041	65.8
	PpCCR	1110	70.1			
双子叶植物 (dicotyledon)	ItCCR	1026	52.3	JcCCR	963	43.13
	FaCCR	1020	51.82	StCCR	999	46.69
	AaCCR	960	53.71	PtoCCR	1017	48.62
	SmCCR	966	46.52	MtCCR	1011	41.27
	CiCCR	1011	49.31	AtCCR1	1035	51.74
	LeCCR1	999	43.67	AtCCR2	999	50.6
	LeCCR2	999	46.59	PsCCR	999	50.1
	BlCCR	1011	51.29	PpyCCR	1020	49.36
	HcCCR	1017	51.28	HbCCR	1017	49.51
	PtCCR	1014	47.68	LlCCR	1011	49.01
	CaCCR	1005	47.6	CoCCR	990	49.14
	VcCCR	1044	51.01	BoCCR	999	50.1
裸子植物 (gymnosperm)	PtaCCR	975	48.77	GbCCR	972	47.16

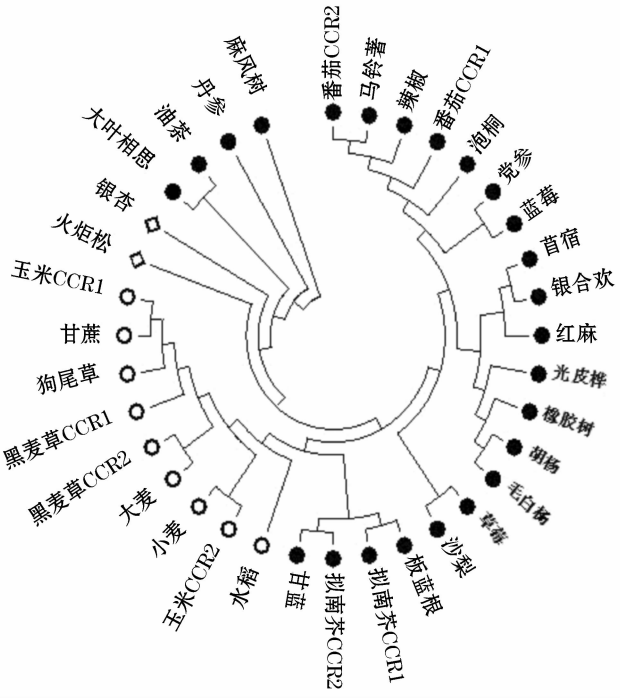

图 1　CCR 基因系统发育树

Fig. 1　Phylogenetic tree of CCR gene

由图 1 可以看出,植物 CCR 基因在进化树中大致可分为二大类、三小类,这与植物在进化的过程中分化为被子植物和裸子植物,被子植物又进一步分化成单子叶植物和双子叶植物的进化方式相一致,但双子叶植物中的拟南芥、板蓝根和甘蔗的 CCR 基因与单子叶植物的 CCR 基因聚在一类,麻风树、丹参、油茶及大叶相思的 CCR 基因与裸子植物银杏,火炬松的 CCR 基因聚在一类,其余双子叶植物的 CCR 基因全部聚在一类中。这表明在植物分化之前,CCR 基因已经存在于植物中,而且在植物的进化时间上超前于植物的分化时间。在分类地位上,虽然拟南芥、板蓝根、甘蔗与单子叶植物差异较大,但其 CCR 基因与单子叶植物的 CCR 基因聚在同一类中,麻风树、丹参、油茶及大叶相思的 CCR 基因和裸子植物的 CCR 基因关系较近,这种基因的聚类和植物分类存在冲突的现象在植物中已被广泛发现[7]。通过构建 CCR 基因系统发育树,CCR 基因的聚类与植物的分类大体一致,表明 CCR 基因的进化和植物的进化基本是一致的,CCR 基因和植物的进化过程密切相关。

2.2 CCR 基因编码氨基酸保守区域及理化性质

2.2.1 CCR 基因编码氨基酸序列保守区的分析

采用 ClustalX(v1.83)软件对 CCR 基因编码氨基酸序列的保守区域进行分析[6],结果表明,从蛋白质的 N 端到 C 端,依次发现了以下 9 个氨基酸保守区(图略):① VCVTGAGGFIASWLVKLLL;② GYTVKGTVRNP;③ GVFHTASP;④ VTDDPEQMVEPAV;⑤ VRRVVFTSSIGAV;⑥ TKNWYCYGKAVAE;⑦ GVDLVVVNPVLVIGPLLQ;⑧ ASGRYLCAE;⑨ TVKSLQEKGHL。在 NCBI 的 Conserved Domains 数据库中,对上述 9 个保守区域进行分析[6],结果表明 9 个保守区域共同构成了 NADB_Rossmann Superfamily 氨基酸保守区,功能注释为 Rossmann - fold NAD(P)(+) - binding proteins,在反应中起催化还原的作用。

2.2.2 CCR 基因编码主要氨基酸平均含量的分析

利用 ExPASy ProtParam 对 35 条 CCR 基因编码的含量较高的氨基酸进行统计[8],发现不论单子叶植物、双子叶植物、还是裸子植物,CCR 基因编码的含量最高的氨基酸均为 Ala、Val 及 Leu(麻风树中包括 Ala 和 Val),但单子叶、双子叶植物中以 Val 的平均含量最高,裸子植物中中以 Leu 的平均含量最高。植物 CCR 基因编码的含量较高的氨基酸在单子叶植物中为:Val(11.83%) > Ala(11.79%) > Leu(8.58%) > Asp(6.56%) > Gly(6.52%);双子叶植物中为:Val(10.20%) > Leu(9.40%) > Ala(8.45%) > Lys(7.47%) > Glu(6.55%);裸子植物中为:Leu(10.25%) > Val(10.05%) > Ala(8.65%) > Lys(7.90%) > Glu(6.20%) = Gly(6.20%),其中 Val、Ala、Leu,Gly 均为非极性氨基酸,Glu、Asp 均为酸性氨基酸,Lys 为碱性氨基酸。

对木质素生物合成途径中另外两种基因 4CL 和 C3H 所编码的主要氨基酸进行统计分析,结果表明,不论在单子叶植物、双子叶植物还是裸子植物中,4CL 和 C3H 所编码的最主要的氨基酸均为 Val、Ala 与 Gly[9-10],均属非极性氨基酸,与 CCR 编码的主要氨基酸的种类一致,但 4CL、C3H 及 CCR 所编码的含量最高的氨基酸的种类并不相同,CCR 编码的含量最高的氨基酸为 Val,4CL 编码的含量最高的氨基酸为 Ala,而 C3H 编码的含量最高的氨基酸为 Leu。

2.2.3 CCR 基因编码氨基酸理化性质的分析

采用 ExPASy ProtParam 对 35 条 CCR 基因编码氨基酸的理化性质进行分析[8],结果表明,不同植物 CCR 基因编码的氨基酸残基数、分子质量、酸性/碱性氨基酸比例、蛋白质不稳定性指数基本一致。在蛋白质的稳定性中,除水稻的 CCR 蛋白为不稳定蛋白外,其他植物的 CCR 蛋白均为稳定性蛋白。

2.3 CCR 蛋白结构和功能的预测及分析

2.3.1 CCR 蛋白导肽、信号肽的预测和分析

导肽是一段新合成的肽链携带的通过细胞膜进入细胞器的所必须的识别序列。采用 TargetP 1.1 Server 对 35 条 CCR 基因编码的氨基酸序列进行预测和分析[11],结果表明其序列含叶绿体转运肽及线粒体目标肽的分值均较低,无氨基酸残基裂解位点,可靠性Ⅳ级,说明 CCR 蛋白不存在上述 2 种导肽。

信号肽位于蛋白质的 N 端,指导分泌性蛋白到内质网膜上合成,在蛋白质合成结束之前被切除,一般有 16 ~ 26 个氨基酸残基,其中包括疏水核心区、信号肽的 C 端和 N 端。采用 SignalP 4.0 Server 对 CCR 蛋白的信号肽进行分析[12],结果表明,植物 CCR 蛋白无信号肽。

2.3.2 CCR 蛋白跨膜结构域、亚细胞定位的预测和分析

跨膜结构域通常由 20 个左右的疏水性氨基酸残基组成,主要形成 α 螺旋。采用 TMHMM Server v.2.0 软件对 35 条 CCR 基因编码氨基酸的跨膜结构域进行预测和分析[13]。结果表明,CCR 蛋白不存在跨膜蛋白,大部分蛋白质位于膜内,但有一部分氨基酸残基的肽段嵌入膜内。

对 35 条 CCR 基因编码的氨基酸采用 Post Prediction 进行亚细胞定位[14],结果表明,65.7% 的 CCR 蛋白定位于质膜上、22.9% 的定位于细胞质中、8.6% 的定位于内质网上,另有 2.9% 的定位于细胞核上。由此推断,在不同的植物体中,虽然 CCR 蛋白的定位有所不同,但绝大多数定位于质膜上,少数定位于细胞质中。

2.3.3 CCR 蛋白亲/疏水性的预测及分析

采用 ExPASy ProtScale 对 35 条 CCR 基因编码的氨基酸序列分析[11],结果表明,其氨基酸序列中亲水性氨基酸、疏水性氨基酸均匀的分布于整个肽链中,亲水性氨基酸多于疏水性氨基酸,因此可认为 CCR 蛋白属于亲水性蛋白。

2.3.4 CCR 基因编码蛋白结构域的预测及分析

在 NCBI 的 Conserved Domains 数据库中,对 35 条 CCR 基因进行分析,结果表明,与提交的序列最匹配的保守结构域模型为 FR_SDR_e,其在相同的蛋白质序列中生成的重叠为 NADB_Rossmann Superfamily(功能注释为 Rossmann - fold NAD(P)(+) - binding proteins)。

同时,采用 SMART 对 CCR 蛋白氨基酸序列的功能结构域进行分析[15],结果表明,在蛋白的 N 端

存在一个脱氢酶/差向异构酶/NAD 结合蛋白的结构域,即与 3Beta_HSD/Epimerase/NAD_binding_4 等保守域具有很高的同源性(图 2)。

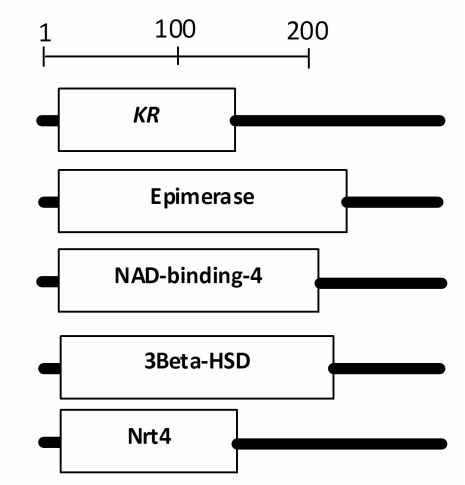

图 2　CCR 蛋白结构域的预测

Fig. 2　Predicted domain sites of CCR

2.3.5　CCR 蛋白活性位点、NADP 结合位点及底物结合位点的预测和分析

酶的特殊催化能力只局限在大分子的一定区域,只有少数特殊的氨基酸残基参与底物结合及催化作用,这些特异的氨基酸残基比较集中的区域,即与酶活力直接相关的区域称为酶的活性部位。酶的活性部位通常又分为结合部位和催化部位[5]。为了进一步了解 CCR 蛋白活性位点、NADP 结合位点及底物结合位点在 CCR 中的分布,利用 Cn3D 对 FR_SDR_e 模型进行分析[16](图 3),结果表明拟南芥 CCR1 的活性位点分别位于第 98 位(A)、第 122 位(S)、第 156 位(Y)和第 160 位(K);NADP 结合位点分别位于第 12 位(G)、第 14 位(G)、第 15 位(G)、第 16 位(F)、第 17 位(I)、第 36 位(V)、第 37 位(R)、第 62 位(A)、第 63 位(D)、第 64 位(L)、第 83 位(T)、第 84 位(A)、第 85 位(S)、第 86 位(P)、第 87 位(M)、第 120 位(T)、第 121 位(S)、第 156 位(Y)、第 160 位(K)、第 183 位(P)、第 184 位(V)、第 185 位(L)、第 186 位(V)和第 198 位(S);底物结合位点分别位于第 87 位(M)、第 89 位(D)、第 122 位(S)、第 123 位(I)、第 124 位(G)、第 127 位(Y)、第 156 位(Y)、第 183 位(P)、第 184 位(V)、第 185 位(L)、第 198 位(S)、第 201 位(H)、第 215 位(N)、第 219 位(V)和第 285 位(F)。

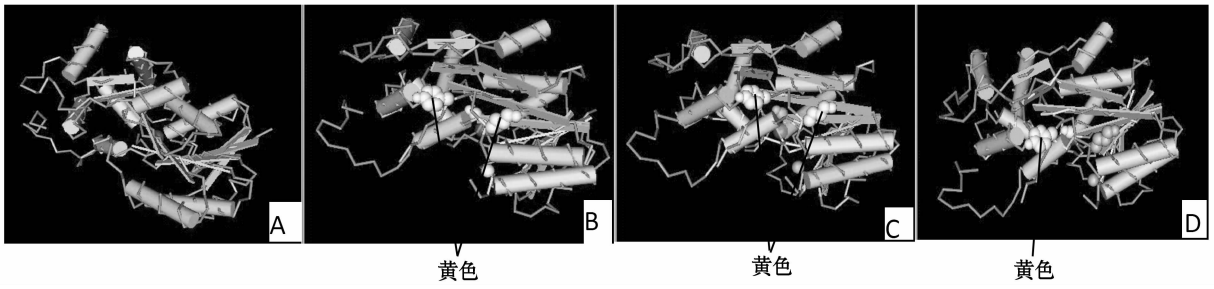

图 3　CCR 蛋白活性位点、NADP 结合位点及底物结合位点的预测

A:FR_SDR_e 模型;B:黄色表示活性位点;C:黄色表示 NADP 结合位点;D:黄色表示底物结合位点。

Fig. 3　The actice sites, NADP – building sites and substrate – building sites of CCR protein

A:The model of FR_SDR_e;B:Yeoolow show active sites. C:Yellow shows NADP binding sites.

D:Yellow shows substrate binding sites.

2.3.6　CCR 蛋白三级结构的预测和分析

蛋白质要实现其催化等活性首先要正确的完成折叠,因此对蛋白多肽构成以及一级结构的分析远远不能满足对蛋白酶功能的了解。蛋白质三级结构的分析,对理解蛋白质结构和功能之间的关系起了至关重要的作用。目前的 X – ray 和 NMR 等实验技术预测蛋白质的结构代价相当高,随着生物信息学的发展,用生物软件预测蛋白质的结构变成现实[17]。采用 Swiss – Model 对植物 CCR 基因编码蛋白质的三级结构进行同源建模,并用 PyMOL 对建模结果进行处理[18]。结果表明 CCR 蛋白的三维结构以 α – 螺旋和无规卷曲为主要的结构元件,延伸链分布于整个肽链之中(图 4)。

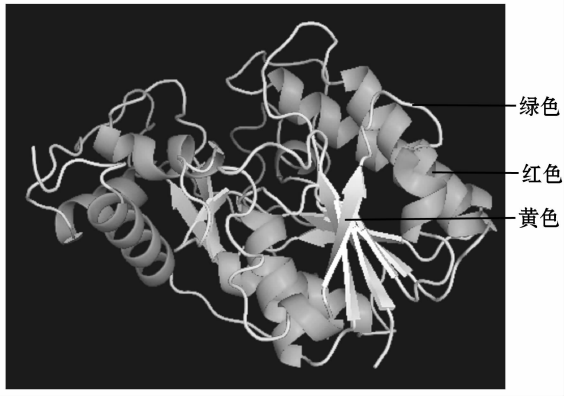

图 4　CCR 蛋白三维结构模型的预测

红色:α – 螺旋;黄色:β – 折叠延伸链;绿色:无视卷曲

Fig. 4　Three – dimensional structure prediction of CCR protein

Red:Alpha helix;Yelllow:Beta sheet extended strand;Green:Random coil

采用 Swiss – Pdb Viewer 分析拟南芥 CCR1 的同源建模结果,结果表明,预测的蛋白质残基的二面角(ψ 和 φ)位于黄色核心区域(图 5),表明其空间结构稳定,所以用同源建模的方法对植物 CCR 基因编码的氨基酸序列进行上述建模的结果非常可靠。

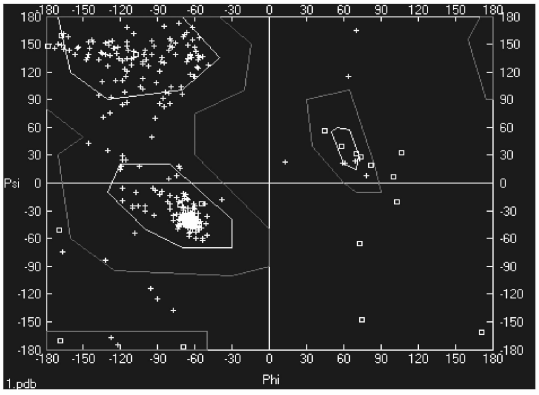

图 5 拟南芥 CCR1 蛋白模型的拉氏构象图

Fig. 5 Ramachandran plot prediction of CCR1 protein in Arabidopsis thaliana

3 结论与讨论

本研究应用生物信息学手段,对 NCBI 数据库中分别来自裸子植物、单子叶植物及双子叶植物的 35 条 CCR 基因完整 cDNA 及其编码氨基酸序列的组成成分、理化性质、保守序列、导肽、信号肽、跨膜结构域、亲/疏水性,结构域及 CCR 蛋白的三级结构进行了预测和分析,构建了 CCR 基因的系统进化树和 CCR 蛋白三级结构的模型。

分析结果表明,单子叶植物 CCR 基因 GC 含量,尤其是编码区的 GC 含量较双子叶植物的普遍偏高,这种现象不仅在 CCR 基因中如此,在对其他基因的研究中亦有发现[9-10],因此推测高的 GC 含量可能是单子叶植物基因区别于双子叶植物基因的一个典型特征。CCR 基因 GC 含量在两大类植物中的明显差异可能与植物的进化过程和生存环境的差异有一定联系[9];CCR 基因的进化与植物的进化基本一致,但少数 CCR 基因的聚类和植物分类存在冲突,有研究表明基因的倍增和重组、水平的基因转移等都是这种差异存在的原因[10];CCR 基因编码的氨基酸从 N 端到 C 端依次发现了 9 个保守区域,在反应中共同起催化还原的作用,但目前大多数的文献中对 KNWYCYGK 这一保守区域报道较多,且认为这一区域在超二级结构上可形成 βαβ 结构,并推测它可能是 CCR 的催化位点,也可能是其辅因子 NADPH 的结合区域,尤其是其上的两个赖氨酸残基(K)可能直接与底物结合,但本文通过对已报道的 35 条 CCR 蛋白二级结构以及上述功能位点的分析发现 KNWYCYGK 在超二级结构上并不能形成 βαβ 结构,仅能形成一段 α – 螺旋和部分无规卷曲,其中仅有 Y(N – 端)、K(C – 端)与 CCR 蛋白的催化位点、NADPH 结合位点及底物结合位点 K(C – 端)有关。

CCR 基因编码氨基酸的理化性质基本一致,但不同植物中 CCR 基因编码的主要氨基酸的种类和含量存在着差异;CCR 基因与木质素合成过程中 C3H、4CL 基因与所编码的主要氨基酸种类相同,均为 Val、Ala 及 Gly,因此推测这 3 种氨基酸可能与木质素合成过程中相关的酶有重要联系;不同植物 CCR 基因编码的氨基酸残基数、分子质量、酸性/碱性氨基酸比例、蛋白质不稳定性指数基本一致;CCR 蛋白无导肽、信号肽及跨膜结构域,属亲水性蛋白;主要定位于质膜上,少数定位于细胞质中,且通过对 CCR 蛋白跨膜结构的预测和分析结果可知,定位于质膜上的蛋白质主要以外在蛋白或脂锚定蛋白的形式存在,少数以整合蛋白的形式部分嵌入质膜中,另外,CCR 蛋白质在核糖体上合成后,可能并不进行蛋白转运,而是直接与质膜结合,或保留在细胞质基质中起催化还原的作用;CCR 蛋白的 N 端存在一个脱氢酶/差向异构酶/NAD 结合蛋白结构域,是其进行催化还原反应的主要部位;CCR 蛋白三级结构模型的空间结构稳定,建模结果可靠。分析结果对于深入研究 CCR 蛋白在木质素合成中的作用具有一定的指导意义。

参考文献(References)

[1] Hano C, Addi M, Bensaddek L. Differential accumulation of monolignol – derived compounds in elicited flax(Linum usitatissimum) cell suspension cultures [J]. Planta, 2006, 223(5): 975 – 989.

[2] 耿 飒,徐存拴,李玉昌. 木质素的生物合成及其调控研究进展[J]. 西北植物学报,2003,23(1):171 – 181.

[3] Lacombe E, Hawkins S. Cinnamoyl CoA reductase, the first committed enzyme of the lignin branch biosynthetic pathway: cloning, expression and phylogenetic relationships [J]. Plant Journal, 1997, 11(3): 429 – 441.

[4] 李金花,张绮纹,牛正田,卢孟柱,Carl J Douglas. 木质素生物合成及其基因调控的研究进展[J]. 世界林业研究,2007,20(1): 29 – 37.

[5] 薛庆中主编,DNA 和蛋白质序列数据分析工具[M]. 第二版. 北京:科学出版社,2009,72 – 100.

[6] Kumar S, Tamura K, Nei M. Integrated software for molecular evolutionary genetics analysis and sequence alignment [J]. Briefings in Bioinformatics, 2004, 5: 150 – 163.

[7] Doyle J J. Trees within trees: genes and species, molecules and morphology [J]. Syst Biol, 1997, 46: 537 – 553.

[8] Kyce J, and Doolittle RF. A simple method for displaying the hydropathic character of a protein [J]. Mol Biol, 1982, 157(6): 105

-132.

[9] 黄胜雄,胡尚连,孙 霞,曹 颖,卢学琴,蒋 瑶. 木质素生物合成酶 4CL 基因的遗传进化分析[J]. 西北农林科技大学学报(自然科学版),2008,36(10):199-206.

[10] 薛永常,聂会忠,刘长斌. 木质素合成酶 C3H 基因的生物信息学分析[J]. 生物信息学,2009,7(1):13-17.

[11] Emanuelsson O,Nielsen H,and Brunak S. Predicting subcellular localization of proteins based on their N-terminsl smino acid sequence [J]. Mol Biol,2000,300(4):1005-1016.

[12] Bendtsen J D,Nielsen H and Von Heijne G. Improved prediction of signal peptides:SingalP 4.0 [J]. Mol Biol.,2004,340(4):783-795.

[13] Iked A M,Arai M,and Lao D M. Transmembrane topology prediction methods:a reassessment and improvement by a consensus method using a dataset of experimentally characterized transmembrane topologies [J]. In Silico Biol,2002,2(1):19-33.

[14] 刘旭光,张 杰编著,分子生物学软件应用[M],第一版. 北京:北京大学医学出版社,2007:178.

[15] Page R D M,Charleston M A. From gene to organismal phylogeny:reconciled trees and the gene tree/species tree problem [J]. Mol Phylogenet Evol,1997,7:231-240.

[16] 付海辉,辛培尧,许玉兰,刘 岩,韦援教,董 娇,曹有龙,周,军. 几种经济植物 UFGT 基因的生物信息学分析[J]. 基因组学与应用生物学,2010,30(1):92-102.

[17] Arnold K,Bordoli L,Kopp J,Schwede T. The SWISS-MODEL workspace:A web-based environment for protein structure homology modeling[J]. Bioinformatics,2006,22(2):195-201.

[18] Laskowski R A,Macarthur M W,Moss D S,Thornton J M. PROCHECK:A program to check the stereo chemical quality of protein structures[J]. Appl. Cryst.,1993,26(2):283-291.