

丹参柯巴基焦磷酸合酶的生物信息学分析

李卿¹, 邸鹏¹, 陆文铨¹, 张磊^{2*}, 陈万生^{1*}

1. 中国人民解放军第二军医大学附属长征医院 药学部, 上海 200003

2. 中国人民解放军第二军医大学药学院 药用植物学教研室, 上海 200433

摘要: **目的** 分析丹参柯巴基焦磷酸合酶(CPS)的氨基酸序列特征, 并与其他植物进行比较, 为丹参CPS蛋白的后续研究提供有利参考。**方法** 利用生物信息学方法, 以丹参为主, 对13科16种植物的18条CPS蛋白序列进行了序列组成、生化特性、导肽、信号肽、跨膜结构域、疏水性/亲水性、蛋白质二级、三级结构及功能域等预测分析, 并进行了丹参CPS蛋白序列与其他植物的同源比对以及系统进化树构建。**结果** CPS蛋白氨基酸残基数集中在730以上; 相对分子量91 610左右; 平均理论等电点5.87, 提示CPS为酸性蛋白。序列结构预测结果显示, CPS蛋白具有明显的疏水区和亲水区, 不存在信号肽, 不具有跨膜结构域, 可能有叶绿体转运肽。蛋白质二级结构中最主要的结构元件是 α -螺旋和无规则卷曲。蛋白同源性比对结果显示, 丹参CPS与同科植物迷迭香和玄参科植物野甘草的同源性最高。**结论** 对丹参等植物的CPS蛋白进行了详细的生物信息学分析, 可为今后深入研究该酶的结构特征和功能提供参考。

关键词: 丹参; 柯巴基焦磷酸合酶(CPS); 生物信息学; 一级结构; 二级结构; 三级结构

中图分类号: R282.12 **文献标志码:** A **文章编号:** 0253-2670(2015)06-0887-08

DOI: 10.7501/j.issn.0253-2670.2015.06.020

Bioinformatics analysis of copalyl diphosphate synthase in *Salviae Miltiorrhizae Radix et Rhizoma*

LI Qing¹, DI Peng¹, LU Wen-quan¹, ZHANG Lei², CHEN Wan-sheng¹

1. Department of Pharmacy, Shanghai Changzheng Hospital, Second Military Medical University, Shanghai 200003, China

2. Department of Pharmaceutical Botany, School of Pharmacy, Second Military Medical University, Shanghai 200433, China

Abstract: Objective To provide the useful information for the further study of copalyl diphosphate synthase (CPS) of *Salviae Miltiorrhizae Radix et Rhizoma* through amino acid (aa) sequences comparison in *S. miltiorrhiza* and other plants. **Methods** Bioinformatics analysis methods were used to perform the prediction of composition and physicochemical character, leader peptide, signal peptide, transmembrane domain, hydrophobicity/hydrophilicity, secondary structure, tertiary structure, and functional domain of 18 CPS amino acid sequences in *S. miltiorrhiza* and other different plants which belong to 16 species and 13 families. Homology analysis and phylogenetic trees of CPS were also carried out. **Results** Most CPS proteins contain more than 730 aa. Their molecular weight is approximately 91 610 and the isoelectric point of them is about 5.87, which suggests that CPS is acidic protein. The results of the sequence structure prediction show that CPS has apparent hydrophobic region and hydrophilicity region, no signal peptide, no transmembrane domain, and may probably have chloroplast transit peptide. The main secondary structures of the proteins are α -helix and random coil. Homology analysis reveals that *S. miltiorrhiza* has the highest homology compared with *Rosmarinus officinalis* and *Scoparia dulcis*. **Conclusion** This work provides a systemic sequence analysis of CPS in *S. miltiorrhiza* and other plants. It will provide the useful information for CPS studies in the future.

Key words: *Salviae Miltiorrhizae Radix et Rhizoma*; copalyl diphosphate synthase; bioinformatics; primary structure; secondary structure; tertiary structure

丹参 *Salviae Miltiorrhizae Radix et Rhizoma* 是 *Salvia miltiorrhiza* Bge. 的干燥根及根茎, 味苦, 性微寒, 归心、肝经, 具有祛瘀止痛、活血通经、清心除

收稿日期: 2014-10-18

基金项目: 国家自然科学基金资助项目(81325024, 81303160)

作者简介: 李卿(1982—), 女, 博士, 主管药师, 研究方向为中药资源学。E-mail: qli@smmu.edu.cn

*通信作者 陈万生 E-mail: chenwanshengsmmu@aliyun.com

张磊 E-mail: zhanglei@smmu.edu.cn

烦的功效^[1]。丹参酮类化合物，是公认的丹参有效成分之一。现代药理研究表明，丹参酮具有清除氧自由基，改善缺氧后引起的心肌代谢紊乱及心功能障碍；修复血管内皮细胞，抗动脉粥样硬化形成；降低心肌耗氧量，缩小心肌梗死面积等作用；还具有抗菌、抗炎、钙拮抗、抗心肌肥厚以及性激素样作用^[2]。

丹参酮属于脂溶性二萜醌类化合物，主要来源于 MEP 途径（2-C-甲基-D-赤藓糖醇-4-磷酸途径）^[3]，研究发现，由柯巴基焦磷酸（copalyl diphosphate, CPP）形成的松香二烯可能是丹参中松香烷型二萜醌类化合物生物合成途径中的前体物质^[4-5]。而柯巴基焦磷酸合酶（copalyl diphosphate synthase, CPS）的作用是催化线性结构的牻牛儿基牻牛儿基焦磷酸（geranylgeranyl pyrophosphate, GGPP）环化形成 4 种可能的 CPP 异构体。高伟等^[6]通过基因芯片等技术发现，丹参酮 II_A 量高的样品中 CPS 基因表达也升高，表明该基因可能参与了丹参酮类化合物的生物合成。Zhou 等^[7]在酿酒酵母中融合表达 GGPP 合酶、法尼基焦磷酸（FPP）合酶、

CPS 和类贝壳杉烯合酶（KSL）后，产生了丹参酮的前体物质次丹参酮二烯，再次证实了 CPS 在丹参酮类化合物生物合成过程中的作用。

CPS 是植物三环二萜类化合物生物合成过程中的重要环化酶，正在受到越来越多的关注，但是关于其功能的系统研究还未见报道，利用生物信息学方法，对 CPS 进行全面预测，能为后续研究指明方向。通过生物信息学方法对丹参中已克隆的 3 个 CPS 基因编码蛋白进行了一级结构、二级结构和三级结构的预测分析，并与其他 15 种植物进行比较，同时进行了这 16 种植物 CPS 蛋白序列同源比对及系统进化树构建，旨在探讨丹参 CPS 蛋白功能，为丹参 CPS 蛋白的后续研究提供有利参考。

1 数据下载

以“Copalyl diphosphate synthase”为搜索关键词，从美国国立生物技术信息中心（National Center of Biotechnology Information, NCBI）下载完整的丹参 CPS 蛋白序列，并下载其他 15 种植物的 CPS 蛋白序列以供比较。药材信息见表 1。

表 1 丹参及其他植物 CPS 基因的基本信息

Table 1 Basic information of CPS in *Salviae Miltiorrhizae Radix et Rhizoma* and other plants

名称	来源	基因 ID 号	名称	来源	基因 ID 号
SmiCPS1 丹参 <i>Salvia miltiorrhiza</i>		ABV57835.1	MdCPS	苹果新品种 <i>Malus domestica</i> × <i>Malus honanensis</i>	AGN92853.1
SmiCPS2 丹参		AEZ55684.1	MtCPS	蒺藜苜蓿 <i>Medicago truncatula</i>	AES60489.1
SmiCPS3 丹参		AGF69544.1	PcCPS	西洋梨 <i>Pyrus communis</i>	AGF25267.1
AgCPS 巨冷杉 <i>Abies grandis</i>		Q38710.1	RoCPS	迷迭香 <i>Rosmarinus officinalis</i>	AHL67261.1
CaCPS 小粒咖啡 <i>Coffea arabica</i>		ACQ99373.1	SdCPS	野甘草 <i>Scoparia dulcis</i>	BAD91286.1
CmaCPS 笋瓜 <i>Cucurbita maxima</i>		AAD04292.1	SmoCPS	江南卷柏 <i>Selaginella moellendorffii</i>	J9QS23.1
CmoCPS 栗 <i>Castanea mollissima</i>		AEF32082.1	SlCPS	番茄 <i>Solanum lycopersicum</i>	AEP82766.1
HaCPS 向日葵 <i>Helianthus annuus</i>		CBL42915.1	TcaCPS	可可 <i>Theobroma cacao</i>	EOX94746.1
IeCPS 毛萼香茶菜 <i>Isodon eriocalyx</i>		AEP03175.1	TcrCPS	台湾杉 <i>Taiwania cryptomerioides</i>	AFE61356.1

2 方法

利用各网站提供的生物信息学工具对丹参及其他植物 CPS 蛋白序列进行在线分析。用 ProtParam (<http://web.expasy.org/protparam/>) 预测 CPS 蛋白的组成成分和理化性质；用 ProtScale (<http://web.expasy.org/protscale/>) 对 CPS 蛋白的疏水/亲水性进行预测；CPS 蛋白的导肽和信号肽的预测分别用 TargetP 1.1 Server (<http://www.cbs.dtu.dk/services/TargetP/>)^[8]和 SignalP 4.1 Server(<http://www.cbs.dtu.dk/services/SignalP/>)^[9]完成；跨膜结构用 TMHMM Server v. 2.0 (<http://www.cbs.dtu.dk/services/TMHMM->

2.0/) 进行预测；CPS 蛋白二级结构的预测用 SOPMA (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html) 完成，功能域的预测用 Pfam 27.0(<http://pfam.janelia.org/>)和 SMART (<http://smart.embl-heidelberg.de/>)^[10]进行；用 Phyre2 (<http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index>)^[11]和 PDBsum Generate (<http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html>) 完成 CPS 蛋白高级结构同源建模和建模结果的检验。

CPS 蛋白序列同源比对采用 blastall 2.2.26 本地

版完成; CPS 蛋白家族系统进化树的构建采用邻位连接法 (neighbor-joining, NJ), 用 Bootstrap method 检验进化树的拓扑结构, 重复抽样次数 (No. of bootstrap replication) 设为 1 000, Gaps/Missing Data Treatment 选择 Pairwise deletion, 由 MEGA 5.0 软件完成。

3 结果与分析

3.1 CPS 蛋白序列的组成成分及理化特性分析

利用 ProtParam 在线工具对丹参及其他植物 CPS 蛋白序列进行理化性质分析。16 种植物 CPS 蛋白的氨基酸残基 (amino acids, aa) 数除蒺藜苜蓿为 158 aa 外, 其他都在 730 aa 以上; 各蛋白序列的相对分子质量为 17 780~99 540, 中位值为 91 610; 理论等电点均在 6 左右, 平均 5.87, 提示 CPS 蛋白为酸性蛋白。从 CPS 氨基酸组成中可以看到, 16 种植物的 CPS 蛋白所含酸性氨基酸残基比例均高于所含碱性氨基酸残基比例, 进一步提示 CPS 蛋白为酸性蛋白。各种植物 CPS 蛋白中, 含量最丰富的氨基酸残基主要集中在亮氨酸 (Leu)、谷氨酸 (Glu)、丝氨酸 (Ser)、丙氨酸 (Ala)、天冬氨酸 (Asp) 和赖氨酸 (Lys)。在这几种植物氨基酸残基中, Leu 是各种植物中量最高的氨基酸 (蒺藜苜蓿除外)。蒺藜苜蓿中, 量最高的氨基酸是 Glu, 占 10.1%, 其次是 Leu 和天冬酰胺 (Asn), 均为 9.5%。通过 ProtParam 分析, 除丹参 *SmiCPS2*、笋瓜、向日葵和毛萼香茶菜的 CPS 蛋白

不稳定系数小于 40, 为稳定蛋白, 其他几种植物的 CPS 蛋白均为不稳定蛋白。

3.2 CPS 蛋白导肽的预测和分析

导肽 (leader peptide) 是一段引导新合成的肽链进入细胞器的识别序列^[12], 导肽的预测与分析对蛋白质的功能分析、作用机制和作用途径等具有重要意义^[13]。通过在线预测工具 TargetP 1.1 Server, 将置信区间设为 0.95, 对丹参及其他多种植物的 CPS 蛋白序列进行了预测 (表 2)。结果显示, *SmiCPS1*、*SmiCPS2* 和 *SmiCPS3* 3 条丹参 CPS 蛋白的导肽预测可靠性分别为 V、IV 和 II 级, *SmiCPS1* 和 *SmiCPS3* 可能具有叶绿体转运肽 (chloroplast transit peptide), 且在各自蛋白序列第 35 位和第 21 位分别存在一个导肽分裂位点。无法确定 *SmiCPS2* 是否具有导肽, 也未发现其导肽分裂位点。其他 15 种植物 CPS 蛋白的导肽预测分析结果显示, 迷迭香的可靠性为 V 级, 其余植物都在 III 级以上。与 *SmiCPS1* 和 *SmiCPS3* 相似, 巨冷杉、小粒咖啡、笋瓜、野甘草、番茄、可可和台湾杉都具有导肽分裂位点, 具有导肽性, 且它们的导肽很可能都是叶绿体转运肽, 提示这些植物的 CPS 蛋白合成后, 可能转运到叶绿体中发挥作用。剩下植物与 *SmiCPS2* 相似, 都不存在导肽分裂位点, 不能确定具有何种导肽。

表 2 丹参及其他植物 CPS 蛋白的导肽预测

Table 2 Leader peptide prediction of CPS protein in *Salviae Miltiorrhizae Radix* et *Rhizoma* and other plants

名称	叶绿体转运肽	线粒体目标肽	分泌途径信号肽	其他	定位	可靠性	导肽氨基酸残基数
<i>SmiCPS1</i>	0.751	0.638	0.010	0.013	C	V	35
<i>SmiCPS2</i>	0.030	0.641	0.096	0.310	*	IV	0
<i>SmiCPS3</i>	0.882	0.096	0.031	0.052	C	II	21
<i>AgCPS</i>	0.931	0.017	0.043	0.112	C	I	70
<i>CaCPS</i>	0.949	0.098	0.005	0.028	C	I	53
<i>CmaCPS</i>	0.950	0.071	0.019	0.055	C	I	46
<i>CmoCPS</i>	0.668	0.035	0.163	0.113	*	III	0
<i>HaCPS</i>	0.702	0.157	0.062	0.130	*	III	0
<i>IeCPS</i>	0.640	0.149	0.048	0.143	*	III	0
<i>MdCPS</i>	0.146	0.559	0.042	0.063	*	III	0
<i>MtCPS</i>	0.037	0.182	0.040	0.899	-	II	0
<i>PcCPS</i>	0.175	0.651	0.035	0.027	*	III	0
<i>RoCPS</i>	0.548	0.385	0.010	0.016	*	V	0
<i>SdCPS</i>	0.758	0.184	0.018	0.022	C	III	46
<i>SlCPS</i>	0.921	0.093	0.038	0.044	C	I	46
<i>SmoCPS</i>	0.091	0.178	0.068	0.835	*	II	0
<i>TcaCPS</i>	0.934	0.024	0.119	0.177	C	II	52
<i>TcrCPS</i>	0.973	0.034	0.010	0.047	C	I	66

“C”-叶绿体转运肽; “*”-不能确定具有何种导肽; “-”-其他位置导肽; 可靠性-数值越小, 越可靠

“C”-chloroplast transit peptide; “*”-don't know; “-”-any other location; reliability class-from 1 to 5, where 1 indicates the strongest prediction.

3.3 CPS 蛋白信号肽的预测和分析

信号肽位于新合成肽链 N 端, 是引导新合成肽链转移到内质网上的一段多肽^[14]。从 CPS 蛋白导肽的预测结果中发现, 所有蛋白都不存在信号肽, 为了验证这个结果, 利用更专业的信号肽在线分析工具 SignalP 4.1 Server, 对丹参 CPS 蛋白信号肽的存在位置进行了分析, 结果表明丹参 CPS 蛋白中不存在信号肽。对其他 15 种植物 CPS 蛋白序列进行相同预测, 也得到类似结果。

3.4 CPS 蛋白疏水性/亲水性的预测和分析

蛋白质亲疏水性氨基酸组成是蛋白质折叠的主要驱动力, ProtScale 程序能用来绘制蛋白质亲疏水性序列谱, 反映蛋白质的折叠情况^[15]。利用 ProtScale 对丹参 CPS 蛋白序列进行疏水性/亲水性预测, 结果显示, *SmiCPS1* 多肽链第 600 位氨基酸具有最低分值-2.733, 亲水性最强; 第 160 位氨基酸具有最高分值 2.333, 疏水性最强; *SmiCPS2* 的最低分值-2.878 位于第 8 位, 最高分值 2.9 位于第 17 位; *SmiCPS3* 第 685 位亲水性最强, 为-3.244, 第 163 位疏水性最强, 为 2.367。从总体上看, 3 条序列的疏水区域明显小于亲水区域, 均表现为亲水性。用相同方法对其他植物 CPS 蛋白序列进行亲疏

水性分析, 结果与丹参相似, 提示 CPS 可能是亲水性蛋白。

3.5 CPS 蛋白跨膜结构域的预测和分析

跨膜结构域一般由 20 个左右的疏水性氨基酸残基组成, 主要形成 α -螺旋, 常由跨膜蛋白的效应区域所展现^[12]。利用在线工具 TMHMM Server v. 2.0 对丹参 CPS 蛋白进行跨膜结构分析, 结果显示, 丹参 CPS 蛋白整条肽链都位于细胞膜之外, 不存在跨膜结构。其他植物的 CPS 蛋白跨膜结构域分析结果与丹参一致, 提示本实验中的 CPS 蛋白均不具跨膜结构域。

3.6 CPS 蛋白二级结构的预测和分析

蛋白质二级结构是指蛋白质多肽链氨基酸残基借助氢键折叠和盘绕形成的 α -螺旋、 β -折叠、无规则卷曲以及模体等组件, 其中, α -螺旋和 β -折叠是最常见的蛋白质二级结构^[13]。利用 SOPMA 对丹参 CPS 蛋白序列进行二级结构预测(表 3), 结果显示, α -螺旋和无规则卷曲是丹参 CPS 蛋白二级结构的主要结构元件, 其次是延伸链和 β -折叠。其他植物 CPS 蛋白序列的二级结构预测结果显示, 除江南卷柏和台湾杉外, 均与丹参一致, 都是 α -螺旋所占比例最多, 其次是无规则卷曲和延伸链, 所占比例最

表 3 丹参等植物 CPS 蛋白二级结构主要构成组件比例

Table 3 Main component proportion of secondary structure of CPS protein in *Salviae Miltiorrhizae Radix et Rhizoma* and other plants

名称	α -螺旋/%	β -折叠/%	无规则卷曲/%	延伸链/%
<i>SmiCPS1</i>	58.01	4.04	31.53	6.43
<i>SmiCPS2</i>	60.77	4.62	30.12	4.49
<i>SmiCPS3</i>	56.75	5.04	31.40	6.81
<i>AgCPS</i>	60.02	3.00	31.11	5.88
<i>CaCPS</i>	58.47	4.75	29.72	7.06
<i>CmaCPS</i>	57.96	4.74	31.59	5.71
<i>CmoCPS</i>	59.60	4.46	29.62	6.32
<i>HaCPS</i>	59.15	4.39	29.95	6.52
<i>IeCPS</i>	56.42	4.41	33.12	6.05
<i>MdCPS</i>	58.57	4.88	30.41	6.13
<i>MtCPS</i>	60.76	4.43	28.48	6.33
<i>PcCPS</i>	58.70	5.13	30.41	5.76
<i>RoCPS</i>	58.32	3.50	34.04	4.13
<i>SdCPS</i>	56.73	4.24	32.00	7.03
<i>SlCPS</i>	59.50	4.12	29.25	7.12
<i>SmoCPS</i>	66.44	3.79	25.98	3.79
<i>TcaCPS</i>	57.93	4.39	31.83	5.85
<i>TcrCPS</i>	61.65	5.45	27.58	5.33

少的是 β -折叠。在江南卷柏中，延伸链和 β -折叠所占比例一致，都是 3.79%；而在台湾杉中， β -折叠 (5.45%) 的比例略高于延伸链 (5.33%)。

3.7 CPS 蛋白功能域的预测和分析

功能域 (functional domain) 又称结构域，是蛋白质分子中介于二级与三级结构之间的一种独立结构和功能单位，具有特定的生物学功能^[16-17]。利用 Pfam 27.0 程序对丹参 CPS 蛋白序列进行功能结构域预测，结果显示，丹参 CPS 蛋白具有一个 terpene synthase N terminal domain 和一个 terpene synthase C terminal domain，行使植物次级代谢过程中萜类物质合成作用。为确保预测结果的准确性，通过在线工具 SMART 进行验证，结果与 Pfam 27.0 一致。对其他植物进行同样的功能域预测和分析后发现，除蒺藜苜蓿只具有一个 terpene synthase C terminal domain，其他植物的 CPS 蛋白均存在同丹参一致的 2 个结构域。蒺藜苜蓿的 CPS 蛋白序列明显短于其他植物，这可能是导致它缺少 terpene synthase N terminal domain 的原因。

3.8 CPS 蛋白三级结构的预测和分析

蛋白质的功能与其三级结构密切相关，对蛋白质高级结构的预测和分析，有助于理解蛋白质结构与功能之间的相关性^[18-19]。利用 Phyre2 线串法 (从

头计算模式) 对丹参 CPS 蛋白进行三级结构预测，结果如图 1 所示。

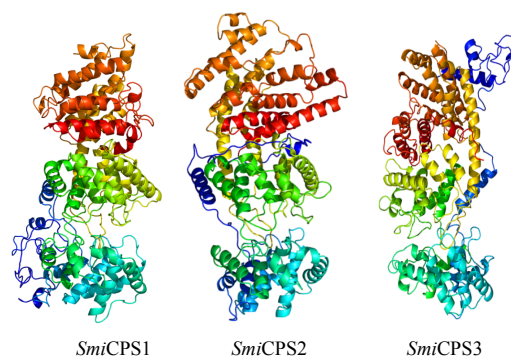


图 1 丹参 CPS 蛋白三维结构预测

Fig. 1 3D structures prediction of *SmiCPS* in *Salviae Miltiorrhizae Radix et Rhizoma*

利用基于 PROCHECK 程序的蛋白质结构检验工具 PDBsum Generate 对预测结果进行检测，计算得出 Ramachandran 图，结果如表 4 所示，丹参 *SmiCPS2* 落在不允许区的比例刚好为 1%，而 *SmiCPS1* 和 *SmiCPS3* 都大于 1%，另外，三者氨基酸残基位于最佳区的比例都小于 90%，不过分布在最佳区和次允许区的比例超过了 90%，表明丹参 CPS 蛋白的空间构象虽然没有落在最有利区，但其构象是合理的，不过稳定性不足。另外，*SmiCPS1*、

表 4 丹参及其他植物 CPS 蛋白三维结构稳定性预测

Table 4 Prediction of stability about 3D structures of CPS protein in *Salviae Miltiorrhizae Radix et Rhizoma* and other plants

名称	最佳区比例/%	次允许区比例/%	一般允许区比例/%	不允许区比例/%	G-factors 值
<i>SmiCPS1</i>	80.50	15.30	1.80	2.40	-1.83
<i>SmiCPS2</i>	88.20	9.00	1.70	1.00	-0.09
<i>SmiCPS3</i>	81.80	14.50	1.70	2.00	-1.56
<i>AgCPS</i>	89.70	8.90	0.80	0.60	0.02
<i>CaCPS</i>	90.20	8.20	1.20	0.40	-0.01
<i>CmaCPS</i>	80.80	14.90	2.30	2.00	-1.50
<i>CmoCPS</i>	80.90	14.20	1.60	3.30	-1.54
<i>HaCPS</i>	82.10	13.90	1.80	2.20	-1.49
<i>IeCPS</i>	81.50	13.20	2.50	2.80	-1.76
<i>MdCPS</i>	81.00	15.00	1.90	2.10	-1.51
<i>MiCPS</i>	83.30	13.80	1.40	1.40	-0.11
<i>PcCPS</i>	83.40	12.80	2.20	1.50	-1.55
<i>RoCPS</i>	88.00	8.80	1.50	1.70	-0.16
<i>SdCPS</i>	80.90	14.20	2.60	2.30	-1.61
<i>SlCPS</i>	79.90	14.10	3.60	2.30	-1.52
<i>SmoCPS</i>	92.30	6.70	0.60	0.40	0.04
<i>TcaCPS</i>	81.90	13.30	2.20	2.70	-1.57
<i>TcrCPS</i>	88.70	9.60	0.90	0.90	0

G-Factors < -0.5 为罕见结构；G-factors < -1.0 为十分罕见结构

G-Factors < -0.5 unusual; G-factors < -1.0 highly unusual

SmiCPS2 和 *SmiCPS3* 的 G-Factors 值分别为-1.83、-0.09 和-1.56, 提示 *SmiCPS2* 的结构属于正常范围, *SmiCPS1* 和 *SmiCPS3* 的结构十分罕见。

对其余植物 CPS 蛋白进行同样的预测和检验 (表 4), 结果显示, 所有植物 CPS 落在最佳区和次允许区的比例均超过了 90%, 其空间构像都是合理的; 笋瓜和江南卷柏的 CPS 蛋白有超过 90% 的氨基酸残基落在最有利区; 笋瓜、江南卷柏、巨冷杉和台湾杉的 CPS 蛋白落在不允许区的比例不到 1%, 空间结构稳定性最好。与 *SmiCPS2* 相似, 迷迭香、蒺藜苜蓿、台湾杉、巨冷杉、江南卷柏和笋瓜的

G-Factors 值均大于-0.5, 它们的空间结构都为正常结构, 而其他植物 CPS 蛋白可能具有罕见的空间结构。笋瓜和江南卷柏每个预测值都较理想, 提示这 2 个蛋白的空间结构真实性较高。

3.9 CPS 蛋白序列同源比对分析及系统进化树构建

利用 blastall 2.2.26 本地版软件将丹参 CPS 蛋白序列与其他植物进行同源比对 (表 5), 结果显示, 丹参 *SmiCPS1* 蛋白与 *SmiCPS2*、*RoCPS* 和 *HaCPS* 具有较高相似性, *SmiCPS2* 与 *RoCPS*、*TcaCPS* 和 *HaCPS* 具有较高相似性, 而 *SmiCPS3* 与 *SdCPS*、*CaCPS* 和 *CmoCPS* 具有较高相似性。

表 5 丹参与其他植物 CPS 蛋白同源性比较

Table 5 Homology analysis of CPS protein in *Salviae Miltiorrhizae Radix et Rhizoma* and other plants

名称	一致性/%			e-value 值		
	<i>SmiCPS1</i>	<i>SmiCPS2</i>	<i>SmiCPS3</i>	<i>SmiCPS1</i>	<i>SmiCPS2</i>	<i>SmiCPS3</i>
<i>SmiCPS1</i>	—	63.36	45.26	0	0	0
<i>SmiCPS2</i>	63.36	—	45.34	0	0	0
<i>SmiCPS3</i>	45.26	45.34	—	0	0	0
<i>AgCPS</i>	40.07	43.82	46.93	7×10^{-154}	1×10^{-149}	1×10^{-168}
<i>CaCPS</i>	48.69	44.72	64.29	0	0	0
<i>CmaCPS</i>	46.90	43.15	57.09	0	0	0
<i>CmoCPS</i>	48.89	44.82	61.91	0	0	0
<i>HaCPS</i>	49.58	46.04	56.38	0	0	0
<i>IeCPS</i>	46.57	42.04	57.82	0	0	0
<i>MdCPS</i>	47.14	43.91	61.55	0	0	0
<i>MtCPS</i>	38.16	28.57	46.05	3×10^{-10}	2×10^{-11}	6×10^{-15}
<i>PcCPS</i>	46.87	43.77	61.29	0	0	0
<i>RoCPS</i>	62.66	86.45	46.53	0	0	0
<i>SdCPS</i>	48.55	43.97	68.64	0	0	0
<i>SlCPS</i>	47.64	45.73	61.62	0	0	0
<i>SmoCPS</i>	42.74	45.36	48.72	1×10^{-172}	5×10^{-157}	0
<i>TcaCPS</i>	47.51	46.16	61.58	0	0	0
<i>TcrCPS</i>	40.88	40.19	44.34	2×10^{-153}	6×10^{-139}	6×10^{-165}

来源于同一祖先的不同植物在进化过程中的关系可以通过进化树来描述, 通过构建植物进化树, 可以了解一种植物在进化过程中的地位。本研究利用 MEGA 5.0 软件对丹参等 16 种植物的 CPS 蛋白序列进行系统进化树构建 (图 2), 结果显示, 丹参 *SmiCPS1* 和 *SmiCPS2* 与同科植物迷迭香聚为一类, *SmiCPS3* 先与玄参科植物野甘草聚为一类, 又和唇形科植物毛萼香茶菜位于同一枝上。玄参科和唇形科都属于双子叶植物纲合瓣花亚纲管状花目, 它们

的亲缘关系较近。另外, 西洋梨和同科植物一个苹果的新品种聚在一起; 两个裸子植物巨冷杉和台湾杉聚在一起, 又和蕨类植物江南卷柏聚为一类; 蒺藜苜蓿单独成为一支, 这可能是由于其序列过短, 变异较大引起的。

4 讨论

CPS 是植物三环二萜类化合物生物合成过程中的起始环化酶, 具有特殊作用, 特别是 *SmiCPS*, 与丹参酮合成密切相关, 正在受到越来越多的关注。

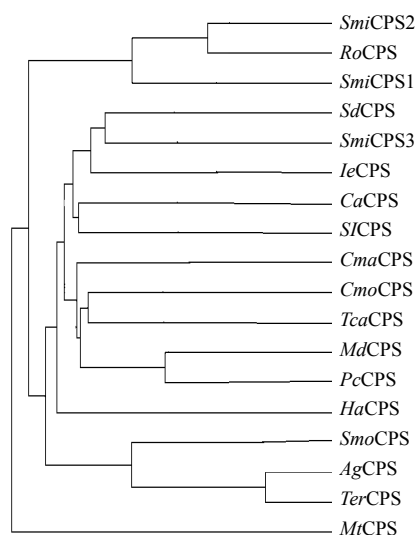


图2 丹参及其他植物 CPS 蛋白序列的 NJ 树

Fig. 2 NJ phylogenetic tree of CPS protein in *Salviae Miltiorrhizae Radix et Rhizoma* and other plants

生物信息学是一门包括生物学、计算机科学、数学、统计学的交叉学科，被广泛应用于基因组学、蛋白质组学、系统生物学和比较基因组学，是一门备受关注的学科，是当今生命科学和自然科学的重大前沿领域之一，也是 21 世纪自然科学的核心领域之一^[20-21]。采用各种分子生物学数据库和分析软件对已知的核酸和蛋白序列进行分析，从而推断及预测其结构和功能，已成为生物信息学研究的一种趋势^[22]。通过生物信息学方法预测，能为实验验证指明方向，减少实验误差，有利于科学研究的成功率。

本实验利用生物信息学方法，将丹参作为主要分析对象，对包括蕨类植物、裸子植物和被子植物的 13 科 16 属 16 种植物的 18 条 CPS 蛋白进行了序列组成、生化特性、结构特点和功能等预测分析。蛋白序列理化性质分析表明，CPS 蛋白为酸性蛋白，其中丹参 *SmiCPS2* 和笋瓜、向日葵、毛萼香茶菜的 CPS 蛋白为稳定类蛋白。导肽预测结果显示，丹参 *SmiCPS1* 和 *SmiCPS3* 以及巨冷杉、小粒咖啡、笋瓜、野甘草、番茄、可可和台湾杉的 CPS 蛋白可能具有叶绿体转运肽。结合信号肽预测结果，可推知 CPS 蛋白在游离核糖体上合成后，可能通过两种途径发挥作用，一是通过导肽进入叶绿体发挥作用；二是不进行蛋白转运，保留在细胞质基质中产生催化作用。这与萜类化合物的生物合成途径相符，文献报道，萜类化合物

的生物合成途径有两条，位于细胞质基质中的甲羟戊酸途径(MVA)和位于质体中的 MEP 途径^[23]，叶绿体即是质体的一种。

蛋白质折叠时能形成亲水表面和疏水内核，并于潜在跨膜区出现高疏水值区域。本实验对丹参等植物 CPS 蛋白进行疏水性/亲水性预测，结果显示 CPS 蛋白不存在高疏水值区域，推测 CPS 蛋白可能不存在跨膜结构域。之后的跨膜结构域预测结果证实了这一推测。

目前，主要通过核磁共振(NMR)和 X-射线晶体衍射(X-ray crystallography)等实验方法获得蛋白质的三维结构，但所需实验技术和设备费用要求都很高。借助计算机软件进行蛋白质高级结构预测可从一定程度上缓解这一难题。本文通过同源建模方法获得了丹参等植物 CPS 蛋白高级结构模型，并通过 PROCHECK 对所构建模型进行了检验，得到 Ramachandran 图。Ramachandran 图用来评估模拟结构与自然结构的相同程度，能够反映蛋白质立体化学质量(stereochemical quality)^[12]。根据构象的稳定性，Ramachandran 图分为最佳区(most favoured regions)、次允许区(additional allowed regions)、一般允许区(generously allowed regions)和不允许区(disallowed regions)^[13]。通常情况下，所预测模型氨基酸分布在最佳区的比例超过整个蛋白质的 90%，可认为该蛋白氨基酸位于构象最有利区；超过 90% 的模型氨基酸落在最佳区和允许区，则表明所构建的空间构像是合理的；不允许区低于 1% 的模型结构空间稳定性较好^[13,15]。从 Ramachandran 图上可以看出，丹参等植物的 CPS 蛋白模型理论上都符合立体化学规则，丹参 *SmiCPS2*、迷迭香、蒺藜苜蓿、台湾杉、巨冷杉、江南卷柏和笋瓜的 CPS 蛋白具有正常的空间结构，笋瓜、江南卷柏、巨冷杉和台湾杉的 CPS 蛋白空间结构稳定性最好。

利用生物信息学方法对丹参 CPS 蛋白序列的生理生化特性进行预测和分析，可以为 CPS 蛋白及其编码基因的克隆提供可靠的依据；对其序列结构的预测和分析，可为其蛋白表达与修饰提供指导；对其二级及高级结构的预测和分析有利于深入探讨该酶结构和功能之间的关系、作用机制和代谢过程。本研究的结论虽属推断和预测，但能为进一步深入研究丹参 CPS 蛋白的功能提供参考。

参考文献

- [1] 中国药典 [S]. 一部. 2010.
- [2] 张建虎, 叶 芊. 丹参酮药理作用 [J]. 中华医学研究杂志, 2005, 5(6): 151-152.
- [3] Wang J W, Wu J Y. Tanshinone biosynthesis in *Salvia miltiorrhiza* and production in plant tissue cultures [J]. *Appl Microbiol Biotechnol*, 2010, 88(2): 437-449.
- [4] Cyr A, Wilderman P R, Determan M, *et al.* A modular approach for facile biosynthesis of labdane-related diterpenes [J]. *J Am Chem Soc*, 2007, 129(21): 6684-6685.
- [5] 秦双双, 陈 新. 丹参次生代谢产物丹参酮的调控研究 [J]. 武汉工业学院学报, 2009, 28(4): 34-37.
- [6] 高 伟, 崔光红, 孔建强, 等. 丹参柯巴基焦磷酸合酶基因的优化表达、纯化及抗体制备 [J]. 药学学报, 2008, 43(7): 766-772.
- [7] Zhou Y J, Gao W, Rong Q X, *et al.* Modular pathway engineering of diterpenoid synthases and the mevalonic acid pathway for miltiradiene production [J]. *J Am Chem Soc*, 2012, 134(6): 3234-3241.
- [8] Emanuelsson O, Nielsen H, Brunak S, *et al.* Predicting subcellular localization of proteins based on their n-terminal amino acid sequence [J]. *J Mol Biol*, 2000, 300(4): 1005-1016.
- [9] Petersen T N, Brunak S, von Heijne G, *et al.* SignalP 4. 0: discriminating signal peptides from transmembrane regions [J]. *Nat Methods*, 2011, 8(10): 785-786.
- [10] Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource [J]. *Nucl Acids Res*, 2012, 40(Database issue): 302-305.
- [11] Kelley L A, Sternberg M J. Protein structure prediction on the web: a case study using the Phyreserver [J]. *Nat Protoc*, 2009, 4(3): 363-371.
- [12] 董 娇, 周 军, 辛培尧, 等. 不同植物LDOX/ANS基因的生物信息学分析 [J]. 基因组学与应用生物学, 2010, 29(5): 815-822.
- [13] 龙 芳, 李绍鹏, 李茂富. 7种植物ALAD基因的生物信息学分析 [J]. 基因组学与应用生物学, 2013, 32(6): 802-814.
- [14] 韦雪芳, 王冬梅, 刘 思, 等. 信号肽及其在蛋白质表达中的应用 [J]. 生物技术通报, 2006(6): 38-42.
- [15] 薛庆中. DNA 和蛋白质序列数据分析工具 [M]. 北京: 科学出版社, 2012.
- [16] 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.
- [17] 王镜岩, 朱圣庚, 徐长法. 生物化学 [M]. 北京: 高等教育出版社, 2002.
- [18] 蔡娜娜, 陈月辉, 李 伟. 基于神经网络的蛋白质三级结构预测 [J]. 计算机工程, 2010, 36(9): 176-177.
- [19] 陈克克, 武 雪. 植物查耳酮异构酶生物信息学分析 [J]. 生物信息学, 2009, 7(3): 163-167.
- [20] Mount D W. 生物信息学 [M]. 曹志伟, 译. 北京: 科学出版社, 2006.
- [21] 许忠能. 生物信息学 [M]. 北京: 清华大学出版社, 2008.
- [22] 王汉屏. 不同植物防御素的生物信息学分析 [J]. 植物生理学通讯, 2008, 44(1): 25-32.
- [23] Osbourn A E, Lanzotti V. *Plant-derived Natural Products* [M]. New York: Springer US, 2009.