根癌土壤杆菌 C58 Cereon 中分泌蛋白信号肽分析

范成明 李成云 赵明富 何月秋*

(云南农业大学植物保护学院 昆明 650201)

摘 要 利用 SignalP3.0、LipoP1.0、TMHMM2.0 和 TargetP1.01 4 种蛋白分析软件预测了 *Agrobacterium tumefaciens* C58 Cereon 菌株全部基因组的 4554 个 ORF 编码的蛋白信号肽,共发现 203 个信号肽,且它们的氨基酸残基相对保守。其中 158 条具分泌型信号肽,9 条具 RR-motif 型信号肽,28 条具信号肽酶 II 型信号肽,8 条具细菌素-信息素型信号肽,但只有分泌蛋白 AGR-C-1878p 和 AGR-C-1880p 的信号肽氨基酸残基完全相同,表明信号肽是高度变异的。

关键词 基因组,信号肽,蛋白分析

中图分类号:Q936 **文献标识码**:A **文章编号**:0001-6209(2005)04-0561-06

根癌土壤杆菌(*Agrobacterium tumefaciens*)作为一种载体,在基因工程中的作用是众所周知的。2001 年该菌株的基因组测序完成,并对基因组的组成和结构特点做了较为系统的分析[1],为我们在今后的试验中更好地利用该菌株进行基因工程的研究提供了更好的试验依据。但是对于根癌土壤杆菌基因组中信号肽蛋白的分析,还未见报道。

多数移位蛋白在多肽的 N 端具有信号肽。信号肽可以指导蛋白运送到蛋白的正确作用位点[2]。尽管不同的蛋白的信号肽存在差异,但信号肽的基本结构是相似的[3]。信号肽一般有 3 个明显的结构域[4,5],即 N-domain、H-domain、C-domain。随着人们对信号肽研究的深入,人们根据信号肽酶识别信号肽序列的不同将信号肽分成 4 类[6]:信号肽酶I型[7]、信号肽酶II型[8,9]、ComC 型[10]和细菌素-信息素型[11]。

随着生物信息学的迅速发展,对蛋白的结构和功能进行预测,如蛋白的信号肽预测[12,13]、作用位点的预测[13]和跨膜结构的预测[14]等,已成为生物信息学研究的重要组成部分,同时也为今后的试验提供更可靠的依据,使实验更有目的性。但每个软件都有自己的优点和不足,为了减少误差,在对核酸或蛋白序列进行分析时,一般都采用多个软件同时分析。当不同的软件分析的结果有较大的出入时,如何取舍常常是研究者很困惑的问题。本文以 4 种软件对根癌土壤杆菌染色体基因组中信号肽预测的结果为例,对其信号肽的特征及不同软件间产生差异的可能原因和解决办法进行了阐述。

1 材料和方法

1.1 材料

所利用的染色体基因组来源于 ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Agrobacterium_tumefaciens_C58_Cereon。染色体中共有 4554 个 ORF。

1.2 方法

利用 http://www.cbs.dtu.dk/services 提供的 SignalP3.0、LipoP1.0、TMHMM2.0 和 TargetP1.01 4 种蛋白分析软件,筛选 *Agrobacterium tumefaciens* C58 Cereon 菌株的全部染色体基因所编码的部分分泌型蛋白。

1.2.1 SignalP3.0 计算分析:利用 SignalP-NN 中的 Dscore 和 Smean 及 SignalP-HMM 中的 Cmax3 个参数对土壤杆菌的基因组进行初步的筛选,将其分成具有信号肽的和不具有信号肽的两大类群。

1.2.2 L 值的计算:L 值为 $-198.235-123.455\times$(S mean score)$+1983.44\times$(HMM Cmax score)进行计算筛选。除去 L 值小于 0 的蛋白序列[12]。

1.2.3 LipoP1.0 计算分析:它可以大致将蛋白分成四大类[16]:细胞质蛋白、跨膜蛋白、被 I 型信号肽酶识别的蛋白和被 II 型信号肽酶识别的蛋白中被 I 型信号肽酶识别的蛋白。

1.2.4 TMHMM2.0 计算分析:TMHMM2.0 可以计算出蛋白中是否具有跨膜结构。由于部分跨膜结构的结构域也能被 I 型信号肽酶识别,可能被错误地预测为分泌型蛋白[17],为了提高结果的可靠性,本文

* 通讯作者:Tel/Fax:86-871-5228532;E-mail:heyueqiu@163.net

作者简介:范成明(1977-),男,山东省邹平县人,博士研究生。E-mail:fanchengming@126.com

收稿日期:2005-01-04,修回日期:2005-05-05

只选取不具有跨膜结构域的蛋白。

1.2.5 TargetP1.01 计算分析：该软件可以将蛋白按其亚细胞作用位点分成五大类：叶绿体型、线粒体型、分泌途径型、任何位点型、不能预测位点型。

通过 1.2.1～1.2.5 的计算分析，对结果进行复合筛选，即选出 5 种分析结果的重叠部分，即是具有分泌特征的蛋白。最后根据信号肽的特点[5]将各种信号肽归类。

2 结果

2.1 土壤杆菌中分泌蛋白信号肽的筛选

本文通过复合筛选得到 203 条具有信号肽的且有分泌功能的蛋白，其中有 167 条Ⅰ型信号肽酶识别的信号肽[158 条分泌型信号肽(表 1)和 9 条 RR-motif 型信号肽(表 2)]；有 8 条细菌素-信息素型信号肽(表 3)；28 种脂蛋白信号肽(表 4)。

表 1 分泌型信号肽长度及对应蛋白

Table 1 Length of the Sec-type signal peptides and name of protein

Protein	Length	Protein	Length	Protein	Length	Protein	Length
GR_C_2067p	19	AGR_L_315p	22	AGR_L_1777p	24	AGR_L_982p	27
AGR_C_4118p	19	AGR_L_3185p	22	AGR_L_198p	24	AGR_C_3924p	28
AGR_L_1726p	19	AGR_L_3413p	22	AGR_L_2241p	24	AGR_C_4360p	28
AGR_L_1p	19	AGR_L_373p	22	AGR_L_2247glp	24	AGR_L_129GMp	28
AGR_L_3547p	19	AGR_L_407p	22	AGR_L_2419p	24	AGR_L_2811p	28
AGR_L_632p	19	AGR_L_425p	22	AGR_L_2862p	24	AGR_L_3087p	28
AGR_C_2347p	20	AGR_C_1196p	23	AGR_L_3125p	24	AGR_C_2502p	29
AGR_C_3637p	20	AGR_C_2092p	23	AGR_L_3165p	24	AGR_C_2586p	29
AGR_C_4189p	20	AGR_C_2707p	23	AGR_C_1451p	25	AGR_C_3918p	29
AGR_L_1197p	20	AGR_C_3018p	23	AGR_C_2142p	25	AGR_L_3070p	29
AGR_L_2575p	20	AGR_C_3235p	23	AGR_C_3448p	25	AGR_L_3248p	29
AGR_L_2900p	20	AGR_C_337p	23	AGR_C_3849p	25	AGR_L_573p	29
AGR_L_3270p	20	AGR_C_3603p	23	AGR_C_489p	25	AGR_C_1156p	30
AGR_L_572p	20	AGR_C_3652p	23	AGR_L_1020p	25	AGR_C_1426p	30
AGR_C_1045p	21	AGR_C_4134p	23	AGR_L_228p	25	AGR_L_2992p	30
AGR_C_2741p	21	AGR_C_4336p	23	AGR_L_3560p	25	AGR_L_873p	30
AGR_C_2812p	21	AGR_C_4976p	23	AGR_L_930p	25	AGR_C_2458p	31
AGR_C_288p	21	AGR_C_738p	23	AGR_C_1036p	26	AGR_C_314p	31
AGR_C_3306p	21	AGR_L_1160p	23	AGR_C_1474p	26	AGR_L_2894p	31
AGR_C_529p	21	AGR_L_1296p	23	AGR_C_1631p	26	AGR_L_612p	31
AGR_L_1124p	21	AGR_L_2505p	23	AGR_C_1881p	26	AGR_C_2742p	32
AGR_L_1462p	21	AGR_L_2863p	23	AGR_C_2594p	26	AGR_C_3080p	32
AGR_L_297p	21	AGR_L_2941p	23	AGR_C_2805p	26	AGR_L_39p	32
AGR_L_3137p	21	AGR_L_3061p	23	AGR_C_384p	26	AGR_L_702p	32
AGR_L_3319p	21	AGR_L_325p	23	AGR_C_4053p	26	AGR_C_2334p	33
AGR_L_335p	21	AGR_L_608p	23	AGR_C_4215p	26	AGR_C_4604p	33
AGR_L_761p	21	AGR_C_1084p	24	AGR_C_433p	26	AGR_L_2413p	33
AGR_L_890p	21	AGR_C_113p	24	AGR_C_970p	26	AGR_L_3153p	34
AGR_L_993p	21	AGR_C_140p	24	AGR_L_1009p	26	AGR_C_4209p	35
AGR_C_1878p	22	AGR_C_1639p	24	AGR_L_1021p	26	AGR_C_2380p	36
AGR_C_1880p	22	AGR_C_3316p	24	AGR_L_2651p	26	AGR_C_2705p	36
AGR_C_2482p	22	AGR_C_3508p	24	AGR_L_3262p	26	AGR_L_295p	36
AGR_C_2695p	22	AGR_C_3762p	24	AGR_C_1003p	27	AGR_L_514p	36
AGR_C_2949p	22	AGR_C_3922p	24	AGR_C_1396p	27	AGR_L_3023glp	37
AGR_C_4294p	22	AGR_C_4267p	24	AGR_C_200p	27	AGR_C_1201p	38
AGR_L_1237p	22	AGR_C_4582p	24	AGR_C_316p	27	AGR_C_3292p	38
AGR_L_2211p	22	AGR_C_75p	24	AGR_C_498p	27	AGR_C_1197p	40
AGR_L_2612p	22	AGR_C_94p	24	AGR_L_146p	27	AGR_C_3730p	48
AGR_L_2662p	22	AGR_L_1481p	24	AGR_L_798p	27		
AGR_L_271p	22	AGR_L_1638p	24	AGR_L_976p	27		

表 2 具有 RR-motif 的被信号肽酶Ⅰ型信号肽

Table 2 Signal peptides recognized by SPase Ⅰ with twin-arginine

Protein	Signal sequence
AGR_C_1468p	MQFTRRHTLKFAGISCAAT ALAGAL
AGR_L_1912p	MRRHLMTTTAAMLLAMTGS AFAGME
AGR_C_3890p	MISHCRRLLATTTALVIAST AIAAAE
AGR_C_2981p	MICRRSLLLGGALLATVMKAPH LLADGD
AGR_L_1217p	MRRTIVKIAMTGFLVTSGLVSP ALSQEL
AGR_L_2772p	MTFSRRQFHKIALAAGAFVALPGG SFAAAE
AGR_C_1694p	MSNAREERRAIAAVLVAGLGFVPAT AHAQDP
AGR_L_456p	MEGYMRRATLFAGLVAGFSTFAFNA AQAVEI

The conserved residues of the RR-motif are indicated in bold letters. The predicted SPase Ⅰ cleavage site, residues from position －3 to －1 relative to the SPase Ⅰ cleavage site are underlined.

表 3 细菌素-信息素型信号肽

Table 3 Predicted bacteriocin and pheromone signal peptides

Protein	Signal peptide Spase Ⅰ		
AGR_C_971p	MCLASAFSLGALAP	ALA	QAP
AGR_C_1570p	MNIVSTSVALLLAA	ASA	EVE
AGR_C_996p	MSWVDPMTSILTNAA	AMA	ALQ
AGR_L_1576p	MMLAAAMLAINISGV	AAA	VPV
AGR_C_1692p	MEVSSMSSSISSALSI MLAAVLPVLVFPHGASS	ASS	QVA
AGR_C_1007p	MALSALAVLLLTMASTS	ENA	SAS
AGR_C_3641p	MPAHSLFSLSVSAALLLSAVPLP	AQS	AGW
AGR_C_2989p	MSTAHFTFCTLFSLTVAAMPVA	APA	ADG

The predicted SPase Ⅰ cleavage site, residues from position －3 to －1 relative to the SPase Ⅰ cleavage site are underlined.

表 4　脂蛋白信号肽

Table 4　Lipoprotein signal peptides

Protein	Signal peptide SPase Ⅱ		
AGR_C_862p	MTDTSLQTLTRRGFVLISAGSV	LSA	CV
AGR_C_967p	MSTRRLPALLLPLAL	LAG	CQ
AGR_C_1378p	MVTVIAKSNSRTKSLLSSVAEVSAVVSMLVV	LSG	CV
AGR_C_1443p	MSGNFLRLSAAMSLLAV	VAG	CN
AGR_C_1473p	MRVSIVLGSLAALVA	LTA	CQ
AGR_C_1491p	MLFRSVTLAAFAIA	LSA	CT
AGR_C_2283p	MPLAGSRHVTALTLAVLTA	LSG	CA
AGR_C_3026p	MKSVLIIAAVFGFSASAA	LAE	CA
AGR_C_3031p	MRLSINGQKRRALFLAPLFAAL	MAG	CA
AGR_C_3166p	MKTLLSSAAITLSLA	LSG	CT
AGR_C_3386p	MSLAFRLNAYKATGLILAAAA	LAA	CQ
AGR_C_3722p	MRVWKQVGVSLAVVG	AGL	CL
AGR_C_3839p	MRNSGKFRGRSALLLSSTVGLALA	LSA	CT
AGR_C_4244p	MSISRRGVLFGLPLF	LAG	CA
AGR_C_4452p	MVSNLRIAERKGSVMRGVFAVVFLMLV	LAG	CA
AGR_C_4477p	MRQAAYSLRRRLLGWLLISTA	IIG	CV
AGR_C_4619p	MNATDTQGRTMSRRILPLFASLCVTAV	LAG	CS
AGR_C_4740p	MGRLPSRRAHCNVGESFMKKAIIFALVGLS	LAS	CT
AGR_C_4789p	MQFRYVVTGLAVVMS	LAG	CQ
AGR_C_4934p	MIKKIAIVALCGTY	LSA	CT
AGR_C_4986p	MLSDVFSRWSRVAAAISVVMMAGL	LAG	CQ
AGR_L_1719p	MNPIHIVVAAPLL	LGG	CT
AGR_L_1720p	MMTRPFTLGSKLAAILSLPLF	LSG	CV
AGR_L_2071p	MKTILIAAALGGIFA	LAS	CS
AGR_L_2074p	MFRINRSLFLLIPAIA	LSS	CQ
AGR_L_2135p	MNFKTSAALLSAAIA	VSS	CV
AGR_L_2246p	MSRTNISALSPMQKLARNPAVIAMTLALA	LAG	CA
AGR_L_2979p	MNETTTTRRGFLLGAGGLALAG	LAG	CN

The predicted SPase Ⅱ cleavage site，residues from position －3 to －1 relative to the SPase Ⅱ cleavage site are underlined.

2.2　土壤杆菌中各类信号肽的特征

2.2.1　分泌型信号肽的特征：土壤杆菌中信号肽长度变化如图 1 所示。在 158 条被信号肽酶 Ⅰ 型识别的分泌蛋白中，其信号肽的长度变化范围为 19～48aa，平均为 25.4aa。同时，从信号肽相似性的比较来看，在所筛选出的 203 条信号肽中只有两条蛋白即 AGR-C-1878p 和 AGR-C-1880p 的信号肽的氨基酸残基完全相同，同为 MNIKSLLIGSAAALAAVSGAHA，并且通过 blast 比对这两个前蛋白的成熟蛋白，结果表明这两个蛋白的全序列和成熟蛋白也具有较高的相似性。同时信号肽长度的变化说明信号肽具有高度的变异性，这可能与其成熟蛋白的靶标位点精确识别和功能的精确分工密切相关。

在土壤杆菌信号肽的氨基酸组成中（图 2），非极性氨基酸残基出现频率为 0.4%～24.0%，平均为 8.2%，其中残基 W 的出现频率最低为 0.4%，而残

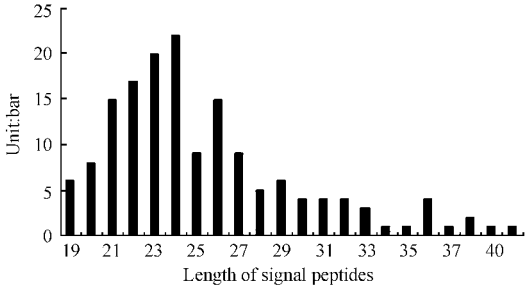

图 1　土壤杆菌基因组中分泌型信号肽长度分布

Fig.1　Length distribution of predicted secretory (Sec-typy) signal peptides in *Agrobacterium tumefaciens* C58 cereon

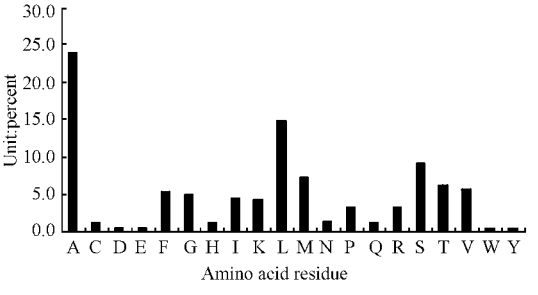

图 2　20 种氨基酸残基在分泌型蛋白信号肽中出现频率

Fig.2　20 charged residues of predicted signal peptides of the secreted proteins

基 A 的出现频率最高为 24.0%,其次为氨基酸残基 L 为 14.8% ;带正电的氨基酸残基 R、K 和 H 分别为 3.3%,4.4%和 1.2%,平均为 3.0% ;带负电的氨基酸残基 D 和 E 都为 0.5%,平均为 0.5% ;不带电的极性氨基酸残基出现频率为 0.4% ~ 9.1%,平均为 2.8%,残基 S 的出现频率最高为 9.1%,残基 Y 的最低为 0.4%。同时,发现在信号肽组成中使用率大于 5%的氨基酸残基,多数属于脂肪族氨基酸,主要是中性和含有羟基或硫氨基的氨基酸。这可能与分泌蛋白的属性相关,使信号肽更易穿过质膜,从而行使信号指导功能。

从土壤杆菌分泌型信号肽的 N 结构域、H 结构域 C 结构域的变化来看,N 结构域的长度变化为 2 ~ 19aa,平均为 5.8aa。H 结构域的长度变化为 11 ~ 39aa,平均为 17.4aa。它们的变化与 Harold[5] 所报道 *Bacillus subtilis* 的结构有差异。可能在不同的生物中信号肽的 N 结构域和 H 结构域的变化有所不同,这还需要在今后的研究中进行探讨。C 结构域(表 5)在-3 位置(相对于酶切位点)、+1 位置和 +1 位置上氨基酸残基 A 使用的频率最高,分别为 81.0%、93.7%和 25.9% ;可以看出在-3 和-1 的位置上,氨基酸残基的组成非常保守,在-3 位置上,出 A 之外,L、S 和 V 的使用频率分别为 1.9%、6.3%和 8.2%,而 D、E、F、H、K、M、N、P、R、T、W 和 Y 使用频率为 0,在-1 位点除了氨基酸残基 A(93.7%)外,只有氨基酸残基 G(使用频率为 1.9%)和 S(使用频率为 4.4%),其他氨基酸残基在该位点上的使用频率均为 0。 -3 和 -1 位是 Ⅰ 型信号肽酶识别比较关键的位点,因而在氨基酸残基的使用上是较保守的。

表 5 分泌型信号肽中 20 种氨基酸在酶切位点(-3 ~ 3)使用频率

Table 5 20 charged residues at the cleavage position from -3 to +3 of the Sec-type signal peptides

aa	-3	-2	-1	+1	+2	+3	aa	-3	-2	-1	+1	+2	+3
A	81.0	7.6	93.7	25.9	3.2	10.8	M	0.0	8.9	0.0	0.6	0.0	0.6
C	0.6	0.0	0.0	0.6	0.0	1.9	N	0.0	4.4	0.0	1.9	0.6	1.9
D	0.0	0.0	0.0	7.6	28.5	3.2	P	0.0	1.3	0.0	0.6	7.6	6.3
E	0.0	3.2	0.0	15.8	14.6	5.7	Q	0.6	10.8	0.0	22.2	5.1	4.4
F	0.0	12.7	0.0	1.3	0.6	2.5	R	0.0	1.9	0.0	1.3	0.0	6.3
G	0.6	2.5	1.9	7.0	3.2	3.2	S	6.3	8.9	4.4	3.8	11.4	6.3
H	0.0	10.1	0.0	1.9	1.3	0.6	T	0.0	1.9	0.0	2.5	16.5	6.3
I	0.6	0.6	0.0	0.6	1.9	8.2	V	8.2	2.5	0.0	0.6	3.2	10.8
K	0.0	1.9	0.0	3.2	1.9	8.2	W	0.0	0.6	0.0	0.0	0.0	0.6
L	1.9	17.7	0.0	2.5	0.0	11.4	Y	0.0	2.5	0.0	0.0	0.6	0.6

2.2.2 RR-motif 型信号肽 :能被信号肽酶 Ⅰ 识别的信号肽除了分泌型信号肽之外,还有一类 RR-motif 型信号肽。它除了具有分泌型信号肽所具有的结构特征之外,其最明显的特征是具有 R-R-X- # - # (X 表示任意氨基酸残基, # 表示疏水氨基酸残基)[19] 这样的模式。在该菌株中得到了如表 2 所示的含有 RR-motif 型信号肽的分泌蛋白。根据本文的试验结果可知,SignalP 可以预测出细菌素-信息素型信号肽,这与 Harold[5] 认为该软件不能预测细菌素-信息素型信号肽相悖,这可能是不同版本的软件所致。

2.2.3 细菌素-信息素型信号肽 :通过 SignalP 算法,得到 8 条不含 H 结构域的细菌素-信息素型信号肽(表 3)。含有该类信号肽的蛋白是通过 ABC(ATP-binding cassette)转运子分泌到胞外。

2.2.4 信号肽酶 Ⅱ 型信号肽(或脂蛋白信号肽) :从 4554 个 ORF 中,共筛选得到 28 条含有脂蛋白信号肽的蛋白(表 4)。其信号肽的平均长度为 21.7aa(图 3),16 ~ 34aa。且 H 结构域的平均长度为 12.1aa,N 结构域的平均长度为 6.6aa。同时,在该类信号肽中,也发现具有 RR-motif 模式的信号肽,说明有些脂蛋白也可能通过 Tat 分泌途径进行移位。

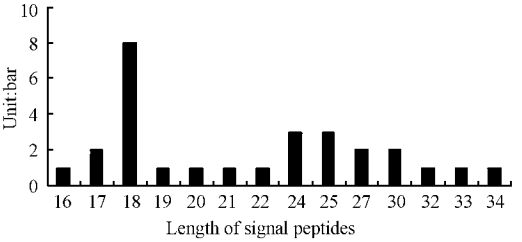

图 3 土壤杆菌中脂蛋白信号肽的长度分布

Fig.3 Length distribution of predicted lipoprotein signal peptides in *Agrobacterium tumefaciens* C58 Cereon

该类信号肽在氨基酸残基的使用频率(图 4)与分泌型信号肽相似。非极性氨基酸的平均使用频率最高为 7.6%,带正电荷的氨基酸平均为 3.2%,带负电荷的氨基酸平均为 0.7%,不带电极性氨基酸为 4.0%。可以看出信号肽的氨基酸的组成相对是保守的。

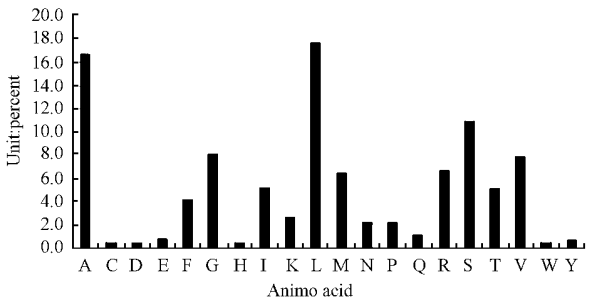

图 4　20 种氨基酸残基在脂蛋白信号肽中出现频率

Fig.4　20 charged residues of predicted signal peptides of the lipoproteins

2.2.5　SignalP-NN 中判断参数 Smean 和 Dscore：SignalP-NN 的参数 Dscore 是 SignalP3.0 中新增加的，且 Dscore 比 Smean 对信号肽有更高的鉴别能力[15]。在以往的版本中主要以 Smean 作为一个重要的判断参数。为了对比两个参数的鉴别力，本文在数据的处理中，分别用 SignalP-NN 参数 Smean 和 Dscore，与 SignalP-HMM 参数 Cmax 组合即 Smean 和 Cmax 组合（简称：S-C）和 Dscore 和 Cmax（简称：D-C）来判断蛋白是否具有信号肽，并对所得结果进行比较。单由 S-C 筛选出具有信号肽的蛋白有 461 条；同样，由 D-C得到 453 条具有信号肽的蛋白；其中两种方法共有蛋白有 419 条。

通过分别统计 S-C 和 D-C 两种方法所得全部的 SignalP-NN 的 Smean 和 SignalP-HMM 参数 Cmax，发现：S-C 的 SignalP-NN 参数 Smean 的平均值为 0.747，而 D-C 的平均值为 0.737；S-C 的 SignalP-HMM 参数 Cmax 的平均值为 0.859，而 D-C 的平均值为 0.875。通过它的比较难以说明那种方法更好。

通过两种方法所得的非共有蛋白相关参数的得分来看：方法 D-C 得出的 SignalP-NN 的参数 Cmax 和 Ymax 的数值都比方法 S-C 得出的相对应的参数值高。对比两种方法所得的 SignalP-HMM 的参数 Cmax，可以看出利用 S-C 所得的 SignalP-HMM 中的 Cmax 得分比 D-C 的普遍要低。

且在两种方法共有的蛋白中进行统计分析，发现：SignalP-NN 的 Smean 最小值为 0.499，平均值为 0.766；SignalP-HMM 最小值为 0.503，平均值为 0.873。这两个参数的得分都较高。

众所周知，SignalP-NN 的 Smean 和 SignalP-HMM 的 Cmax 这两个重要参数的值越高，蛋白含有信号肽的可能性就越大。因而笔者在进行预测时，为了使预测结果更加准确，选用了 SignalP-NN 中的 Smean 和 D 和 SignalP-HMM 参数 Cmax 这 3 个参数对 4554 个 ORF 进行分泌型蛋白的筛选。

3　讨论

对于这几个软件预测的准确性有好多学者[15,20]已对其进行了探讨，在目前可用的网络软件中，本文所使用的软件都是属于准确性较高的软件，如 Karsten 等[20]报道在革兰氏阴性菌中利用 SignalP-NN 和 SignalP-HMM 对其中信号肽进行预测时，这两种方法的正确性分别为 91.4% 和 93.1%。但是在实际的应用过程中，本文发现在 SignalP3.0 中 SignalP-NN 的 D-score 和 Smean 和 SignalP-HMM 的 Cmax 这几个重要的参数在信号肽的判断上存在着一定的差异，并不是所有的被这 3 个判断为含有信号肽的蛋白的 3 个参数的得分都比较高。这不免给我们的判断带来了一定的困惑。同时本文为了保证结果的可靠性，选用 3 个参数同时对土壤杆菌的 ORF 进行预测，以便提高预测结果的可靠性。

利用不同的软件对同一基因组进行分析，其结果会有所差异，这主要是由于软件所使用的算法不同。同时每个软件都不可能保证其预测结果完全正确。因而，在进行信号肽的分析时首先应该选择权威的软件，并采用多种策略利用多个软件对目标基因组进行筛选，一般采用 2～4 个软件，以保证预测结果的真实性。再者，由于信号肽在不同生物中的存在这一定的差异，利用于软件对其进行分析时，一定要注意研究对象的生物类型。

通过对土壤杆菌的分析，本文发现在构成信号肽的氨基酸的使用上是保守的，如主要以非极性的为主且在酶切位点的氨基酸残基的组成中几乎没有酸性氨基酸和碱性氨基酸，这可能与信号肽的与质膜识别的属性有关。但是信号肽却是高度进化的，在 203 条具有信号肽的蛋白中，只有两条蛋白的信号肽是相同的。这种情况可能与信号肽功能的精密分工是密切相关的。因为蛋白的要行使其正确地功能，首先必须移位到其正确的亚细胞作用位点。亚细胞位点的识别正是由信号肽决定的，不同的作用位点，就应该有不同的信号肽。但是每一类蛋白的信号肽的保守和进化程度如何，还需要更多的信号肽的信息。

参　考　文　献

[1] Derek W W, Joao C S, Rajinder K, *et al*. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science*, 2001, **294**(14): 2317－2323.

[2] Emanuelsson O, Nielsen H, Brunak S, *et al*. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, 2000, **300**: 1005－1016.

[3] von Heijne G. Life and death of a signal peptide. *Nature*, 1998, **396**: 111 – 113.

[4] Akita M, Sasaki S, Matsuyama S, *et al*. SecA interacts with secretory proteins by recognizing the positive charge at the amino terminus of the signal peptide in *Escherichia coli*. *Journal of Biological Chemistry*, 1990, **265**: 8162 – 8169.

[5] Paetzel M, Dalbey R E, Strynadka N C. Crystal structure of a bacterial signal peptidase in complex with a beta-lactam inhibitor. *Nature*, 1998, **396**: 186 – 190.

[6] Harold T, Albert B, Jan D H, *et al*. Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. *Microbiology and Molecular Biology Reviews*, 2000, **9**: 515 – 547.

[7] Tjalsma H, Bolhuis A, van Roosmalen M L, *et al*. Functional analysis of the secretory precursor processing machinery of *Bacillus subtilis*: identification of a eubacterial homolog of archaeal and eukaryotic signal peptidases. *Genes*, 1998, **12**: 2318 – 2331.

[8] Tjalsma H, Kontinen V P, Pra gai Z, *et al*. The role of lipoprotein processing by signal peptidase Ⅱ in the Gram-positive eubacterium *Bacillus subtilis*: signal peptidase Ⅱ is required for the efficient secretion of a-amylase, a non-lipoprotein. *Journal of Biology Chemistry*, 1999, **274**: 1698 – 1707.

[9] Tjalsma H, Zanen G, Venema G, *et al*. The potential active site of the lipoprotein-specific (type Ⅱ) signal peptidase of *Bacillus subtilis*. *Journal of Biology Chemistry*, 1999, **275**: 25102 – 25108.

[10] Chung Y S, Dubnau D. ComC is required for the processing and translocation of ComGC, a pilin-like competence protein of *Bacillus subtilis*. *Molecular Microbiology*, 1995, **15**: 543 – 551.

[11] Paik S H, Chakicherla A, Hansen J N. Identification and characterization of the structural and transporter genes for, and the chemical and biological properties of, sublancin 168, a novel lantibiotic produced by *Bacillus subtilis* 168. *Journal Biology Chemistry*, 1998, **273**: 23134 – 23142.

[12] Samuel A L, Steven W, Sophien K, *et al*. An analysis of the *Candida albicans* genome database for soluble secreted proteins using computer-based prediction algorithms. *Yeast*, 2003, **20**: 595 – 610.

[13] Henrik N, Jacob E, S ren B, *et al*. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 1997, **10**: 1 – 6.

[14] Anders K, Larsson B, Gunnar V H, *et al*. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, 2001, **305**(3): 567 – 580.

[15] Jannick D B, Henrik N. Improved prediction of signal peptides: SignalP 3.0. *Journal Molecular Biology*, 2004, **340**: 783 – 795.

[16] Juncker A S, Willenbrock H, Gunnar V H, *et al*. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Science*, 2003, **12** (8): 1652 – 1662.

[17] Rahfeld J U, Rucknagel K P, Schelbert B, *et al*. Confirmation of the existence of a third family among peptidyl-prolyl cis/trans isomerases: Amino acid sequence and recombinant production of parvulin. *FEBS Letter*, 1994, **352**: 180 – 184.

[18] Cristo bal S, de Gier J W, Nielsen H, *et al*, Competition between Sec-and Tat-dependent protein translocation in *Escherichia coli*. *EMBO Journal*, 1999, **18**: 2982 – 2990.

[19] Karsten H, Andreas G, Maurice S, *et al*. PrediSi: Prediction of signal peptides and their cleavage positions. *Nucleic Acids Research*, 2004, **32** (Web Server issue): W375 – W379.

Analysis of signal peptides of the secreted proteins in *Agrobacterium tumefaciens* C58

FAN Cheng-ming LI Cheng-yun ZHAO Ming-fu HE Yue-qiu*

(*Plant Protection College*, *Yunnan Agricultural University*, *Yunnan Agricultural University*, *Kunming* 650201, *China*)

Abstract: The 4554 ORFs of *Agrobacterium tumefaciens* C58 Cereon were used for the prediction of signal peptides by the network tools, such as SignalP3.0, LipoP1.0, TMHMM2.0 and TargetP1.01. Total 203 signal peptides with conserved amino residues are found, among them, 158 are secretary types, 9 are RR-motif types, 28 are SignalPase Ⅱ types and 8 are bacteriocin-pheromone types. However, only two signal peptides from the secreted proteins, AGR-C-1878p and AGR-C-1880p have the same amino sequences, showing the signal peptides of the strain are highly variable.

Key words: Genome, Signal peptides, Protein prediction, Secretome

* Corresponding author. Tel/Fax: 86-871-5228532; E-mail: heyueqiu@163.net

Received date: 01-04-2005