

羊布鲁氏菌 16M 基因组分泌蛋白的生物信息学分析

杨羽, 吴清民*

(中国农业大学动物医学院 农业部人畜共患病重点开放实验室

动物布病防控研究室, 北京 100193)

摘要: 对羊布鲁氏菌全基因组 16M 中全部 3 197 个氨基酸序列进行生物信息学分析。利用 SignalP 和 TatP 软件分析 N-端信号肽, 结果显示具有 N-端信号肽的序列有 288 个; 继而用 TMHMM 和 Phobius 软件对这 288 个序列进行跨膜区预测, 得到不含跨膜区的蛋白 208 个; 最后利用 LipoP 对这 208 个蛋白进行分类, 得到具有信号肽的蛋白 191 个。使用 SecretomeP 软件对被预测为无信号肽的蛋白质进行分析, 结果显示有 391 个可能通过非经典途径分泌。由于分泌蛋白在细菌的致病过程中起着重要作用, 而布鲁氏菌基因组编码的大多数蛋白的功能尚未确定, 因此分析和预测布鲁氏菌的分泌蛋白可为更完整地、系统地研究布鲁氏菌的分子致病机理提供非常重要的信息。

关键词: 布鲁氏菌; 信号肽; 分泌蛋白

中图分类号: S852.614

文献标识码: A

文章编号: 0366-6964(2009)07-1059-04

Analysis of the Secreted Protein Encoding Genes in Genome of *Brucella melitensis* 16M

YANG Yu, WU Qing-Min*

(College of Veterinary Medicine, China Agricultural University, Beijing 100193, China)

Abstract: The complete 3 197 open reading frames (ORFs) of the *Brucella melitensis*'s whole genome sequence has been analyzed. Among them, there are 191 proteins with N-terminal signal peptides based on the combination of prediction algorithms SignalP v3.0 and TatP, transmembrane domains prediction algorithms TMHMM v2.0 and Phobius. And there are also 391 non-signal peptide secretion proteins based on the analyzed by SecretomeP. The functions of most of these proteins have not been identified clearly. For the important effect of secreted proteins have in bacteria pathology, the research of secreted proteins of brucella would be very useful to understand the pathogenic mechanism of brucella and to develop new methods for diagnosis.

Key words: brucella; signal peptide; secreted protein

布鲁氏菌病(布病)是一种由布鲁氏菌引起的世界范围内的人兽共患传染病, 该病病原体最早于 1905 年在马耳他从山羊奶中被分离出来^[1]。主要在亚洲、非洲和拉丁美洲流行^[2], 我国多见于内蒙、东北, 西北等牧区。布鲁氏菌感染动物可以引起流产、死胎、睾丸炎和关节炎等; 人布鲁氏菌病又称马耳他热或波状热, 病人以发热、肌肉关节和神经疼痛等为特征。多年来, 针对布鲁氏菌病致病机理的研究从未间断。随着 2002 年 *B. melitensis* 16M 基因

组测序工作的完成, 为研究者在基因组水平上揭示布鲁氏菌的分子致病机理奠定了基础^[3]。

B. melitensis 16M 基因组共有 2 个染色体, 总长度约为 3.3 Mb, 预测有 3 197 个 ORFs^[3], 其中近年来鉴定出多种编码基因及其蛋白在细菌致病过程中起着非常重要的作用, 但分泌蛋白在布鲁氏菌致病中作用的研究少见报道。分泌蛋白通常是由 N 端的信号肽引导通过细胞膜被运输到胞外^[4]。信号肽一般由 10~40 个氨基酸残基组成。尽管不同蛋白的信号肽

收稿日期: 2008-11-03

基金项目: 国家自然科学基金项目(30871882); 教育部长江学者和创新团队发展计划项目(IR T0866)

作者简介: 杨羽(1982-), 男, 天津人, 硕士生, 主要从事布鲁氏菌病的研究, E-mail: pcdazui@yahoo.com.cn; Tel: 010-62733901

* 通讯作者: 吴清民, E-mail: wuqin@cau.edu.cn

存在差异,但信号肽的基本结构是相似的。信号肽一般有 3 个明显的结构域,即 N 端结构域、H 结构域和 C 端结构域^[5]。N 端结构域为带正电荷的碱性氨基酸残基组成,C 端结构域是富含丙氨酸的片段,疏水中心区包含有 6~15 个氨基酸残基,对肽段的穿膜和正确定位具有重要的作用^[6]。有些蛋白虽然缺乏 N 端信号肽也能达到胞外区域^[7],但大多数可溶性蛋白主要还是通过正常的分泌途径进行转运。

随着生物信息学的迅速发展,对蛋白质的结构和功能进行预测,如蛋白质的信号肽预测、作用位点的预测和跨膜结构的预测等已成为生物信息学研究的重要组成部分。通过结构和功能预测可为进一步试验提供更可靠的依据,使试验更有目的性。作者通过 6 种蛋白分析软件对布鲁氏菌基因组的分泌性蛋白进行了分析预测。

1 材料与方法

1.1 材料

所用材料是已公布的羊布鲁氏菌 16M 基因组中共 3 197 个蛋白质的氨基酸序列^[8]。

1.2 方法

利用 SignalP v3.0(<http://www.cbs.dtu.dk/services/SignalP/>)软件^[9-10]对其氨基酸序列的信号肽进行分析;利用 TatP v1.0(<http://www.cbs.dtu.dk/services/TatP/>)软件^[11]对 Tat 信号肽进行分析;利用 TMHMM v2.0(<http://www.cbs.dtu.dk/services/TMHMM/>)软件^[12]对该序列的跨膜区进行分析;利用 Phobius(<http://phobius.cgb.ki.se/>)软件^[13-14]对只有 1 个跨膜区并和预测信号肽区域重合的蛋白进行再次分析;利用 LipoP v1.0(<http://www.cbs.dtu.dk/services/LipoP/>)软件^[15]对脂蛋白信号肽进行分析;利用 SecretomeP(<http://www.cbs.dtu.dk/services/SecretomeP/>)软件^[16]对 SignalP 判定无信号肽的序列进行分析。

2 结果

2.1 SignalP 分析结果

SignalP v3.0 是基于神经网络(Neural network, NN)和隐马可夫模型(Hidden Markov, HMM)^[17] 2 种算法预测蛋白质信号肽的软件。前者能够对信号肽的分泌途径以及切割位点进行预测,预测值标示为 S-mean;后者主要是对该氨基酸序列是否具有信号肽进行分析和预测,同时也对信

号肽切割位点及分布进行预测,预测值表示为 Sprob。分析结果显示,神经网络算法(NN)分析具有信号肽的蛋白质有 370 个;隐马可夫模型(HMM)分析具有信号肽的蛋白质有 281 个;同时满足 2 种算法的有 228 个。采用分泌信号肽判别式^[18]: $L = -918.235 - 123.455x(\text{S-mean}) + 1983.44x(\text{HMM 值})$,所计算出的 L 值对上述预测值进行判定,当 L 值大于零时即表示该蛋白质具有分泌信号肽。经过分析,布鲁氏菌基因组中具有信号肽的蛋白有 288 个。

2.2 TatP 分析结果

细菌有一种途径被称作双精氨酸转运系统(Twin-arginine Translocation,简称 Tat),也可以进行蛋白质的分泌。通过此类系统转运的蛋白质信号肽中具有双精氨酸(RR)保守序列的核心。这种转运系统有一个很特别的功能,就是可以转运已经完全折叠的蛋白质,而传统的 sec 分泌途径只能分泌未折叠的蛋白。经过 TatP v1.0 分析,基因组中共有 162 个蛋白具有 Tat 信号肽,其中有 44 个与 SignalP 的预测结果重合,且信号肽切割位点相同,但是所有这些蛋白都没有 RR 保守序列,因此作者认为没有具有 TatP 信号肽的分泌蛋白。

2.3 TMHMM 和 Phobius 分析结果

分泌蛋白与跨膜蛋白都含有信号肽,但分泌性蛋白一般不含有跨膜区,因此本试验中使用 TMHMM v2.0 软件对含有信号肽的蛋白进行了跨膜区分析。分析结果显示,在上述 288 个蛋白中没有跨膜区的有 126 个,有 1 个跨膜区的有 91 个,有 2 个或更多跨膜区的蛋白有 71 个。蛋白质含有跨膜区提示该蛋白可能作为膜受体起作用,也可能是定于与膜上的锚定蛋白或者离子通道蛋白,具有 2 个以上的跨膜结构域的蛋白被认为不具有分泌信号肽。但是,当具有 1 个跨膜区的时候,TMHMM v2.0 软件不能将信号肽与跨膜结构域区分,容易造成假阴性的结果,在这种情况下信号肽长度小于 60 个氨基酸残基,且位于 N 端时仍有可能属于分泌信号肽。因此采用 Phobius 软件对这 91 条具有 1 个跨膜区的序列进行再次分析,Phobius 软件可以区别信号肽和跨膜区。结果 91 个蛋白中有 9 个确实有跨膜区,其余的 82 个都没有跨膜区。

2.4 LipoP v1.0 分析结果

一个具有信号肽的蛋白即使没有跨膜区,也有可能被共价的脂基团锚定在膜上,这种蛋白的信号

肽被称为脂蛋白信号肽,即被 II 型信号肽酶识别的蛋白。SignalP v3.0 软件虽然可以预测信号肽序列,但不能区分被 I 型信号肽酶识别的蛋白和被 II 型信号肽酶识别的蛋白。因此作者采用 LipoP v1.0 软件对这 208 个蛋白进行进一步的分析,Lipo 软件可以将蛋白分为 4 大类: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。分析结果表明其中被 I 型信号肽酶识别的蛋白有 170 个,被 II 型信号肽酶识别的蛋白有 21 个,其余为细胞质蛋白。

2.5 SecretomeP 软件分析结果

由于并非所有的分泌蛋白都具有信号肽,某些非经典途径分泌的蛋白并不带有信号肽。因此,为了找出布鲁氏菌基因组中不带有信号肽的分泌蛋白,作者使用 SecretomeP 软件对 3 197 个 ORF 中

的其余 2 909 个氨基酸序列进行分析。通过 SecretomeP 软件将其与除去信号肽的经典分泌蛋白序列进行比对,从而对其是否编码分泌蛋白进行预测。结果显示,2 909 个氨基酸序列中有 391 个可能编码非经典分泌蛋白。

2.6 综合分析

本研究中作者共分析羊布鲁氏菌 16M 基因组中蛋白氨基酸序列 3 197 个,通过以上 6 种不同算法软件对布鲁氏菌基因组的分析,结果表明在羊布鲁氏菌所有蛋白质序列中的分泌蛋白一共有 561 个,其中具有可溶性分泌信号肽的蛋白有 170 个,占总比例的 5.3%;通过非经典途径分泌的蛋白有 391 个,占总比例的 12.2%。在这 561 个蛋白中最长的为 2 555 个氨基酸,最小的为 42 个氨基酸。分布最多的集中在 50~350 个氨基酸,有 428 个序列(图 1)。

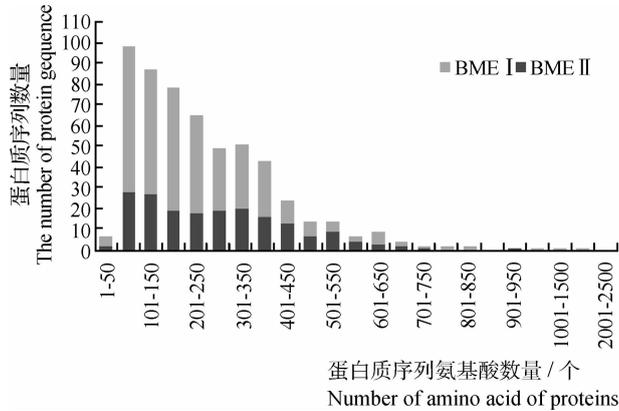


图 1 561 个分泌蛋白序列大小(氨基酸个数)分布

Fig. 1 Sequences length (aa) distribution in the 561 secreted proteins

信号肽长度介于 16~60 个氨基酸,其中具有 23 个氨基酸的信号肽数量最多,有 14 个,其次是 28 个氨基酸的信号肽,数量为 13 个(图 2)。从统计图

中可以看出,这种细菌的信号肽长度在 23~28 个氨基酸之间所占比例最高,为 36%。

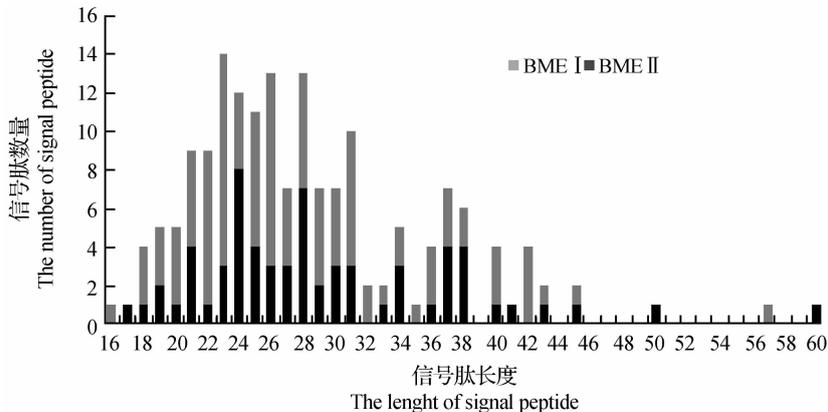


图 2 561 个分泌蛋白信号肽长度(氨基酸个数)分布

Fig. 2 The signal peptides length (AA) distribution of 561 secreted proteins

3 讨论

由于使用的算法不同,不同的分析软件,对同一序列分析的结果也可能不同。作者通过使用不同软件对同一基因组进行分析,目的就是通过不同算法的分析,可以有效地提高预测结果的准确性。

分泌蛋白的跨膜结构除了 α 螺旋之外,还有 β 折叠结构,由于条件限制,作者并没有分析分泌蛋白的 β 折叠结构。因此,可能会有一些拥有 β 折叠结构的蛋白也被列入分泌蛋白的结果中。

通过试验验证是找到分泌蛋白最可靠的办法,因为少数分泌蛋白并不遵循普遍的情况,有些分泌蛋白可能拥有不止一个跨膜区^[7]。虽然利用生物信息学方法分析的结果会和实际分泌到菌体外的蛋白在数量和种类上有一定的差距,但其中大多数属于分泌蛋白的可靠性已经在一些生物体中被证实。布鲁氏菌作为一种重要的人畜共患病病原菌,基因组中保留如此众多的潜在分泌蛋白,其表达量大小、在何种情况下表达以及生物学功能等问题都值得研究。通过生物信息学方法分析并预测这些蛋白的功能,依据预测结果设计试验验证,既可节省时间,还将为完整、系统地研究该菌的分子致病机理奠定基础。

参考文献:

- [1] GODFROID J, CLOECKAERT A, LIAUTARD J P, et al. From the discovery of the Malta fever's agent to the discovery of a marine mammal reservoir, brucellosis has continuously been a re-emerging zoonosis[J]. *Vet Res*, 2005, 36(3): 313-326.
- [2] CORBEL M J. Brucellosis; an overview[J]. *Emerg Infect Dis*, 1997, 3(2): 213-221.
- [3] DELVECCHIO V G, KAPATRAL V, REDKAR R J, et al. The genome sequence of the facultative intracellular pathogen *Brucella melitensis* [J]. *Proc Natl Acad Sci U S A*, 2002, 99(1): 443-448.
- [4] VON H G. The signal peptide[J]. *J Membr Biol*, 1990, 115(3): 195-201.
- [5] MARTOGLIO B, DOBBERSTEIN B. Signal sequences: more than just greasy peptides[J]. *Trends Cell Biol*, 1998, 8(10): 410-415.
- [6] 周晓翌, 李成云, 赵之伟, 等. 粗糙脉孢菌基因组分泌蛋白的初步分析[J]. *遗传*, 2006, 28(02): 200-207.
- [7] MARCHESINI M I, UGALDE J E, CZIBENER C,

et al. N-terminal-capturing screening system for the isolation of *Brucella abortus* genes encoding surface exposed and secreted proteins[J]. *Microb Pathog*, 2004, 37(2): 95-105.

- [8] *Brucella melitensis* 16M[OL]//Organism Details. http://img.jgi.doe.gov/cgi-bin/pub/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=637000040.
- [9] BENDTSEN J D, NIELSEN H, VON H G, et al. Improved prediction of signal peptides: SignalP 3.0 [J]. *J Mol Biol*, 2004, 340(4): 783-795.
- [10] NIELSEN H, KROGH A. Prediction of signal peptides and signal anchors by a hidden Markov model [J]. *Proc Int Conf Intell Syst Mol Biol*, 1998, 6: 122-130.
- [11] BENDTSEN J D, NIELSEN H, WIDDICK D, et al. Prediction of twin-arginine signal peptides[J]. *BMC Bioinformatics*, 2005, 6: 167.
- [12] KROGH A, LARSSON B, VON H G, et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes[J]. *J Mol Biol*, 2001, 305(3): 567-580.
- [13] KALL L, KROGH A, SONNHAMMER E L. A combined transmembrane topology and signal peptide prediction method[J]. *J Mol Biol*, 2004, 338(5): 1027-1036.
- [14] KALL L, KROGH A, SONNHAMMER E L. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server [J]. *Nucleic Acids Res*, 2007, 35(Web Server issue): 429-432.
- [15] JUNCKER A S, WILLENBROCK H, VON H G, et al. Prediction of lipoprotein signal peptides in Gram-negative bacteria [J]. *Protein Sci*, 2003, 12(8): 1652-1662.
- [16] BENDTSEN J D, KIEMER L, FAUSBOLL A, et al. Non-classical protein secretion in bacteria[J]. *BMC Microbiol*, 2005, 5: 58.
- [17] BARASH S, WANG W, SHI Y. Human secretory signal peptide description by hidden Markov model and generation of a strong artificial signal peptide for secreted protein expression [J]. *Biochem Biophys Res Commun*, 2002, 294(4): 835-842.
- [18] LEE S A, WORMSLEY S, KAMOUN S, et al. An analysis of the *Candida albicans* genome database for soluble secreted proteins using computer-based prediction algorithms[J]. *Yeast*, 2003, 20(7): 595-610.