桃拉病毒基因组的生物信息学分析

李伟哲[1],贾阳阳[2],刘 露[2],肖 勤[2*]

(1.河北农业大学生命科学学院,河北保定 071001;2.河北农业大学海洋学院,河北秦皇岛 066003)

摘要 [目的]对桃拉病毒(*Taura syndrome* virus,TSV)的完整基因组进行生物信息学分析。[方法]通过生物信息学方法对基因序列组成、开放阅读框、蛋白质理化性质、二级结构预测分析、蛋白跨膜结构的存在与否、蛋白信号肽存在与否以及蛋白质三级结构进行了预测分析。[结果]登录 NCBI 网站下载 TSV(JX094350.1)10 128 bp 的基因片段,经生物信息学分析,编码氨基酸 3 286 个,理论等电点(pI)为 5.14,相对分子质量为 366 443.00 Da,不稳定系数(Ⅱ)为 37.76,属于稳定蛋白质;完整基因序列中包含 2 个开放阅读框(open reading frame,ORF);蛋白中存在跨膜结构;没有蛋白信号肽。[结论]对 TSV 的生物信息学分析有助于在分子水平上了解桃拉病毒的基因结构以及预测其感染机制,可为预防和治疗虾类的桃拉综合征提供有用的信息。

关键词 桃拉病毒;基因;蛋白质;生物信息学分析

中图分类号 S945.4 **文献标识码** A

文章编号 0517-6611(2019)08-0119-04

doi:10.3969/j.issn.0517-6611.2019.08.030

开放科学(资源服务)标识码(OSID):

Bioinformatics Analysis of *Taura syndrome* virus Genome

LI Wei-zhe[1], JIA Yang-yang[2], LIU Lu[2] et al (1.College of Life Sciences, Hebei Agricultural University, Baoding, Hebei 071001; 2.College of Ocean, Hebei Agricultural University, Qinhuangdao, Hebei 066003)

Abstract [Objective]To make bioinformatics analysis on *Taura syndrome* virus (TSV) gene in swine. [Method]The complete genes of TSV were analyzed by bioinformatics software, including its gene sequence analysis, open reading frame prediction (ORF) prediction, physicochemical properties of protein, secondary structure prediction, protein transmembrane and signal peptides prediction, and as well as protein tertiary structure prediction. [Result]The TSV gene (JX094350.1) with a length of 10 128 bp was successfully obtained from NCBI gene bank. The bioinformatics analysis showed that TSV gene was a total of 3 286 amino acids, a theoretical isoelectric point (pI) of 5.14, a theoretical molecular mass of 366 443 Da, and an instability coefficient (Ⅱ) 37.76, being a stable protein. The complete gene sequence contained two open reading frames (ORFs). There was a transmembrane structure in the protein, and there was not included protein signal peptide. [Conclusion]The bioinformatics analysis of TSV is helpful for understanding *Taura syndrome virus* on molecular level and the prediction of infection mechanism. It will provide useful informations for the prevention and treatment of *Taura syndrome*.

Key words *Taura syndrome virus*; Gene; Protein; Bioinformatics analysis

1994 年,Lightner 等[1]在患有桃拉综合征(taura syndrome,TS)的凡纳滨对虾(*Litopenaeus vannamei*)中发现了桃拉病毒(*Taura syndrome* virus,TSV),之后被 Hasson 等[2]证实并命名。TSV 是一种直径为 31~32 nm 的非包覆二十面体颗粒,是单链正链 RNA,属于小 RNA 病毒粒子家族[3]。TSV 能够感染许多对虾种类,自然宿主如凡纳滨对虾和中国对虾(*Penaeus chinensis*)[4]。不同对虾品种对 TSV 的敏感度不同,其中凡纳滨对虾敏感度较高,野生型凡纳滨对虾仔虾对 TSV 的抵抗力比人工孵化仔虾的更高[5]。TSV 大多数情况下倾向危害体重较轻的幼虾,幼虾的累积死亡率高达 40%~90%[6]。TSV 感染共 3 个阶段,分别为急性期、过渡期和慢性期。在急性期,虾表皮上皮组织切片中可以看到典型的病理损伤,而在过渡期和慢性期则无。多数病虾属于急性期感染,急性期感染的大多数病虾胡须和尾巴的体表变红,且尾扇边缘会变成茶红色,外壳比较柔软;基本不进食,极少数可能会少量进食;常在水面缓慢游动。个别幸存病虾将进入到过渡期,过渡期仅有数天,但半数左右的病虾会于甲壳上留下不规则的黑斑[7]。随后进入长时间的慢性期,处于慢性期的病虾成为病毒携带者,可将病毒水平传播给其他易感虾群。我国的多数对虾养殖区由于养殖规模的不断扩大,已经出现了严重的桃拉综合征发病现象,因此了解 TSV 分子生物学信息对于防治此病可以提供信息帮助,而生物信息学相关分析可以满足这一现实需求。

生物信息学是继人类基因组计划之后的一门新兴学科,其将数学、计算机和生物学相关内容联合起来处理生物信息,对信息进行获取、加工、存储、分配、分析和解释。生物信息学主要研究内容包括核酸序列、蛋白质序列以及从其中获得的一些数据。生物学研究正在由传统实验观察阶段扩展到推理演算阶段,目前已有很多关于核酸和蛋白质的生物学数据库存在[8]。笔者基于生物信息学的方法,通过相关软件对 TSV 进行了生物信息学分析,旨在解析该病毒在分子生物学方面的一些信息,为研究预防和治疗桃拉综合征提供信息帮助。

1 材料与方法

1.1 材料

1.1.1 序列信息。在 NCBI(national center for biotechnology information, https://www.ncbi.nlm.nih.gov/)中的 Nucleotide 选项内可以查出 TSV 完整基因序列。

1.1.2 分析软件。TSV 生物信息学分析软件见表 1。

1.2 方法

1.2.1 TSV 基因序列信息。NCBI 中收录了世界各国科学家提交的基因序列、大多数期刊论文所研究过的基因序列以及

基金项目 留学人员科技活动择优资助项目(CN201602)。

作者简介 李伟哲(1992—),男,河北石家庄人,硕士研究生,研究方向:渔业环境及其调控。*通信作者,副教授,博士,从事水生动物病害防治及水产品安全检测研究。

收稿日期 2018-11-26

各专利中公布的基因序列,因此它的集成化程度是现有生物学数据库中最高的。各国家科学家提交注册的基因序列、各种期刊论文报道的基因序列和各种专利中公开的基因序列均被收录在 NCBI 中,并每 24 h 更新数据库内容[9]。

利用 NCBI 在线网站中的 Nucleotide 数据库,对 TSV 进行搜索,得到完整的基因序列。

表 1 TSV 生物信息学分析软件

Table 1 The analysis softwares of TSV bioinformatics

序号 No.	软件名称 Software name	分析用途 Analysis purposes
1	BioXM 本地软件	核酸序列的组成分析
2	NCBI 中 ORF Finder 软件	确定 TSV 基因开放阅读框位置
3	Translate Tool 软件	确定氨基酸序列
4	ProtParam 软件	分析理化性质包括相对分子质量、理论等电点、氨基酸组成等
5	SOPMA 软件	分析蛋白质二级结构如无规卷曲、β 转角、延伸链、α 螺旋
6	TMHMM 软件	预测 TSV 基因编码的蛋白质跨膜区
7	Signal P-4.1 Server 软件	预测蛋白是否含有信号肽
8	SWISS-MODEL 软件	构建蛋白质三级结构

1.2.2 TSV 基因序列的组成分析。BioXM 本地软件的编制和运行对于计算机软硬件所处的环境要求不高,基本 Windows 系统均可以运行,进行基因序列组成分析时采取的算法是通读全部序列[10],因此可直接将序列放入分析框。

利用 BioXM 本地软件,对所得到的完整基因序列进行组成成分分析。

1.2.3 TSV 开放阅读框架分析。ORF 是可编码蛋白质的一段碱基序列,其代表蛋白结构数目[11]。ORF Finder 是生物信息服务平台中的一种,可在数据库中寻找编码框,查询可能存在的蛋白质编码区域[7]。

利用 ORF Finder 在线软件,对 TSV 完整基因序列的 ORF 出现位置进行检索,需满足以下条件:最小 ORF 长度(NT)为 600,遗传密码使用起始密码子“ATG”,忽略嵌套 ORFs。

1.2.4 TSV 蛋白质理化性质分析。蛋白质一级结构指多肽链内氨基酸残基由 N 末端到 C 末端的顺序排列,也称之为基本结构。根据 ORF Finder 确定的 TSV 蛋白质一级结构预测分析,进行理化性质分析。

将获取的 ORF 结果和 Translate Tool 软件得到氨基酸序列通过 ProtParam 在线软件(http://au.Expasy.org/)进行理化性质分析,主要包括蛋白质理论分子量、氨基酸组成、理论等电点、理论不稳定系数以及疏水性等参数[12]。

1.2.5 TSV 蛋白质二级结构预测与分析。多肽主链在空间中盘绕、折叠可构成一种立体结构形态,将其称之为蛋白质二级结构。此结构包括无规卷曲、β 转角、延伸链和 α 螺旋等,它不仅是一级结构与三级结构之间的连接,而且是预测三维空间结构的重要环节。通过 SOPMA 在线软件,采用 5 种方法(Levin 同源预测方法、CNRS 方法、GOR 方法、PHD 方法和双重预测方法)对蛋白质二级结构进行分析和预测,将预测结果汇集整理[13-14]。

1.2.6 TSV 蛋白质序列的跨膜结构。膜蛋白拥有独特的结构,并担负着许多细胞生物功能,如细胞之间信号传导,物质运输以及免疫等[15]。因此,预测蛋白质跨膜结构是否存在十分重要。利用 TMHMM 在线软件(http://www.cbs.dtu.dk/services/TMHMM/)对 TSV 基因编码的蛋白质是否存在跨膜结构进行预测[16]。

1.2.7 TSV 蛋白信号肽分析。信号肽由氨基酸组成,通常处于分泌蛋白的 N 端。它负责把蛋白质引导至细胞含不同膜结构的亚细胞器内,作用不可替代,可以用来分析蛋白质的细胞定位[17]。通过 Signal P-4.1 Server 隐马尔可夫模型(HMM)算法在线对 TSV 基因编码的蛋白质中是否存在信号肽进行分析预测[18]。

1.2.8 TSV 蛋白质三级结构预测与分析。蛋白质三级结构是一种特定的立体构象,其是多肽链利用侧链基团之间相互作用发生卷曲折叠,并依靠次级键维系而形成。同源建模法、折叠识别法(串线法)和从头预测法是蛋白质三维结构普遍的预测方法,其中同源建模法是最常用的方法,可通过生物信息学软件 Expasy 中的 SWISS-MODEL 软件对 TSV 蛋白质三级结构的立体构象进行预测[19]。

2 结果与分析

2.1 TSV 基因序列信息 利用 NCBI 网站获得 TSV 基因序列,该基因序列号为 JX094350.1,总长度为 10 128 bp,并将其以 FASTA 格式下载到本地文件夹内。

2.2 TSV 基因序列的组成分析 BioXM 软件结果显示,序列长度为 10 128 bp;腺嘌呤核苷酸(A)共 2 869 个,占总核苷酸的 28.33%;鸟嘌呤核苷酸(G)共 2 311 个,占总数的22.82%;胞嘧啶核苷酸(C)共 2 061 个,占总数的 20.34%;尿嘧啶核苷酸(U)共 2 887 个,占总数的 28.51%;A+U 的含量(56.84%)高于G+C 的含量(43.16%);分子量为 3 121 404 Da。

2.3 TSV 开放阅读框架(ORF)分析 ORF Finder 软件在线分析结果见图 1,在满足最小 ORF 长度(NT)为 600、遗传密码使用起始密码子“ATG”并忽略嵌套 ORFs 条件下,TSV 基因潜在的编码框共 2 个,其中 ORF1 由第 6 878~9 913 位之间的 1 011 个氨基酸组成,ORF2 由第 312~6 671 位之间的 2 119个氨基酸组成。

2.4 TSV 蛋白质理化性质分析 TSV 基因共编码 3 286 个氨基酸,将氨基酸序列导入分析软件,结果见表 2。由表 2 可知,疏水性氨基酸包括丙氨酸(A)、异亮氨酸(I)、亮氨酸(L)、苯丙氨酸(F)、色氨酸(W)、缬氨酸(V)共 1 077 个,占氨基酸总数的32.8%;极性氨基酸包括天冬酰胺(N)、半胱氨酸(C)、谷氨酰胺(Q)、丝氨酸(S)、苏氨酸(T)、酪氨酸(Y)共 1 006 个,占氨基酸总数的 30.6%;强碱性氨基酸包括赖氨酸(K)和精氨酸(R)共 305 个,占 9.3%;强酸性氨基酸包括天冬氨酸(D)和谷氨酸(E)共 430 个,占氨基酸总数的 13.1%;稀有氨基酸中只含有吡咯赖氨酸(Pyl)2 个,占氨基酸总数的 0.1%,不含有硒半胱氨酸(Sec)。同时可得知,理论等电点(pI)为 5.14;相对分子质量为 366 443 Da;原子组成

为 $C_{16157}H_{25346}N_{4360}O_{5098}S_{131}$；不稳定系数(Ⅱ)为 37.76，属于稳定蛋白类；脂肪系数为 82.49；平均亲水性为-0.284。

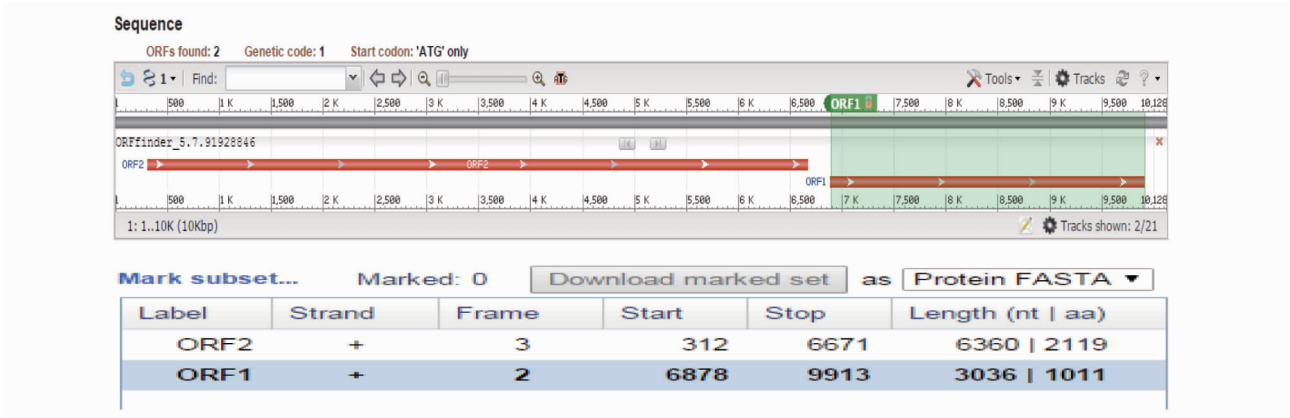

图 1　开放阅读框结果

Fig.1　The results of open reading frame (ORF)

表 2　ProtParam 在线分析氨基酸组成

Table 2　Amino acid composition by on-line ProtParam analysis

名称 Name	数量 Number//个	比重 Proportion//%
Ala (A)	218	6.6
Arg (R)	157	4.8
Asn (N)	158	4.8
Asp (D)	185	5.6
Cys(C)	57	1.7
Gln(Q)	103	3.1
Glu(E)	245	7.5
Gly(G)	169	5.1
His(H)	79	2.4
Ile(I)	186	5.7
Leu(L)	262	8.0
Lys(K)	148	4.5
Met(M)	74	2.3
Phe(F)	122	3.7
Pro(P)	144	4.4
Ser(S)	291	8.9
Thr(T)	281	8.6
Trp(W)	32	1.0
Tyr(Y)	116	3.5
Val(V)	257	7.8
Pyl(O)	2	0.1
Sec(U)	0	0.0

2.5　TSV 蛋白二级结构预测与分析　通过 SOPMA 对 TSV 蛋白在线分析，结果见图 2。由图 2 可知，其中 α 螺旋占 37.70%，延伸链占 18.43%，β 转角占 7.54%，无规则卷曲占 36.33%，以 α 螺旋和无规则卷曲结构为主。

2.6　TSV 蛋白质序列的跨膜结构　通过跨膜结构分析程序 Expasy 的 HMHMM 在线对 TSV 进行跨膜结构预测分析。结果如图 3 所示：横坐标代表氨基酸顺序位置，纵坐标代表该区域是跨膜区的概率，大于 0.5 表示该区域具有跨膜螺旋的可能性大，小于 0.5 则可能性小；红线和蓝线分别代表膜外和膜内区域，两者交互位置表示出现跨膜[20]。由图 3 可知，TSV 基因编码的蛋白质存在跨膜区域。

2.7　TSV 蛋白信号肽　通过 SignalP-4.1 在线软件对 TSV 基因编码的蛋白信号肽存在与否进行预测，结果如图 4 所示，C-score 代表剪切位点的值，此值与氨基酸一一对应，C 值最高处通常是剪切位点；S-score 代表每个氨基酸对应一个值并连接成曲线表明变化趋势，值较高的区域可能为信号肽区域；Y-score 同时考虑 S 值和 C 值，比单独的 C 值或 S 值更准确[21]。因为典型信号肽的结果图中 C-score 和 Y-score 均向 +1 靠近，S-score 曲线则在切点前高，在切点之后变低[22]，数据显示 TSV 基因编码的蛋白质存在信号肽的可能性为 0.112，因此预测不存在信号肽。

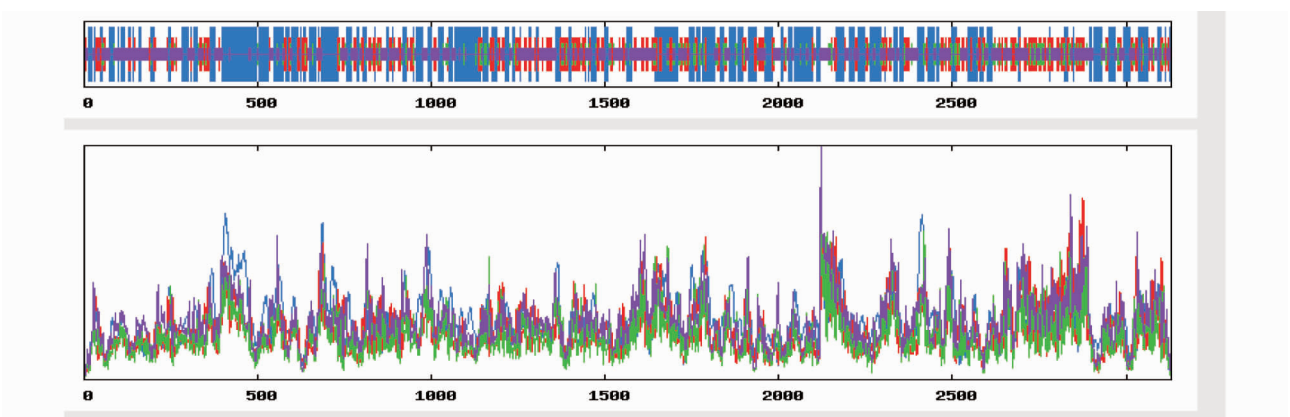

图 2　SOPMA 在线分析蛋白质二级结构预测结果

Fig 2　The forecast results of protein secondary structure by on-line SOPMA analysis

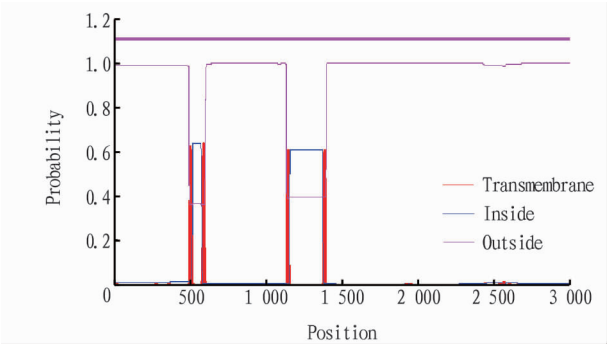

图 3 HMHMM 在线软件分析蛋白质跨膜区域结果

Fig.3 The analysis results of protein transmembrane region by on-line HMHMM software

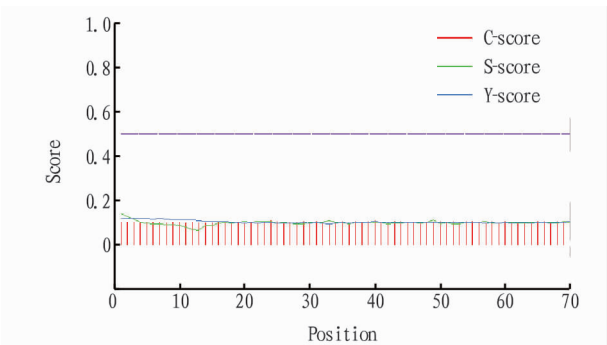

图 4 Signal P-4.1 分析蛋白信号肽结果

Fig.4 The analysis results of protein signal peptide by Signal P-4.1 software

2.8 TSV 蛋白质三级结构预测与分析 通过 SWISS-MODEL 模型软件对 TSV 蛋白质的三维结构进行预测,结果如图 5 所示。

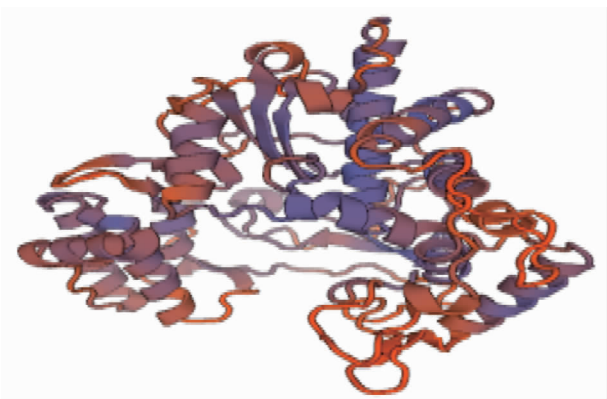

图 5 SWISS-MODEL 在线分析蛋白三级结构结果

Fig.5 The analysis results of protein tertiary structure by online SWISS-MODEL software

3 结论

在 NCBI 中现有的 TSV 基因中总长虽有所不同,但均表明 TSV 基因有 2 个 ORF,与该研究的结果一致。目前对于桃拉病毒的研究主要集中在 TSV 的分离、鉴定和检测等方面[23],没有针对其全基因组在蛋白质结构预测方面的研究。通过对 TSV 基因(JX094350.1)生物信息学分析,得到了在分子水平上 TSV 基因组更多的信息,为进一步研究提供了便利和经验,同时也为预防和治疗桃拉综合征提供重要信息。

参考文献

[1] LIGHTNER D V,REDMAN R M,HASSON K W,et al.Taura syndrome in *Penaeus vannamei* (Crustacea:Decapoda):Gross signs,histopathology and ultrastructure[J].Diseases of aquatic organisms,1995,21(1):53-59.
[2] HASSON K W,LIGHTNER D V,POULOS B T,et al.Taura syndrome in *Penaeus vannamei*:Demonstration of a viral etilolgy[J].Diseases of aquatic organisms,1995,23(2):115-126.
[3] 战文斌.水产动物病害学[M].北京:中国农业出版社,2011:239-240.
[4] LIGHTNER D V,REDMAN R M.Stategies for the control of viral disease of shrimp in the Americas[J].Fish Pathol,1998,33(4):165-180.
[5] 刘棠.凡纳滨对虾桃拉综合征病毒主要结构蛋白基因的克隆及原核表达[D].厦门:厦门大学,2008.
[6] BONAMI J R,HASSON K W,MARI J,et al.Taura syndrome of marine penaeid shrimp:Characterization of the viral agent[J].Journal of general virology,1997,78(Pt 2):313-319.
[7] 陈颜峰.如何减轻南美白对虾桃拉综合征的危害[J].科学种养,2012(7):50.
[8] 司源,郭亦琦,孔航辉.基于 ORF Finder 方法的植物 ITS 片段结构特点分析[J].华北农学报,2005,20(5):54-56.
[9] 张见影,伦志军,李正红.NCBI 基因序列数据库使用和检索方法[J].现代情报,2003(12): 224-225.
[10] 黄骥,张红生.基于 Windows 的核酸序列分析软件的开发[J].生物信息学,2004,2(1):13-17.
[11] ARNOLD K,BORDOLI L,KOPP J,et al.The SWISS-MODEL workspace:A web-based environment for protein structure homology modelling[J].Bioinformatics,2006,22(2):195-201.
[12] 钟静,吴小明,胡颖.大豆 FLAs 蛋白理化性质和结构特征的生物信息学分析[J].河南农业科学,2017,46(3):34-40.
[13] 刘祥.绿脓杆菌外膜蛋白 OprF 的生物信息学分析[J].生物技术,2015,25(4): 343-348.
[14] BAXEVANIS A D,FRANCIS OUELLETTE B F.Bioinformatics:A practical guide to the analysis of genes and proteins[M].New York:Wiley Interscience,2001.
[15] 裔东亮.蛋白质跨膜结构与二硫键连接模式研究[D].上海:上海交通大学,2009.
[16] 姚清国.运用 TMHMM 软件对水稻水通道蛋白 OsPIP2;6 跨膜结构的分析[J].河南农业,2017(29):59.
[17] GARDY J L,SPENCER C,WANG K,et al.PSORT-B:Improving protein subcellular localization prediction for Gram-negative bacteria[J].Nucleic acids research,2003,31(13):3613-3617.
[18] KARPLUS K,KARCHIN R,BARRETT C,et al.What is the value added by human intervention in protein structure prediction? [J].Proteins:Structure,function,and bioinformatics,2001,45(S5):86-91.
[19] 张德峰,付玉荣,伊正君.结核分枝杆菌 CarD 蛋白结构与功能的生物信息学分析[J].中国病原生物学杂志,2017(7):605-608.
[20] ZHANG M Q.Large-scale gene expression data analysis:A new challenge to computational biologists[J]. Genome research,1999,9(8):681-688.
[21] 陈尤莺.分类算法在生物信息学中的应用[D].福州:福建师范大学,2013.
[22] 刘洪超,胡澍,涂心明.果蝇 Tap 蛋白结构与功能的生物信息学分析[J].重庆医学,2015,44(17):2311-2314.
[23] 黎铭,陈晓汉.对虾桃拉综合征病毒(TSV)的分子生物学研究进展[J].广西农业科学,2008,39(6):834-837.