· 论著 ·

日本血吸虫基因组核糖体移码序列的预测分析

王海彬[1,2]　杨忠[3]　刘锋[3]　胡薇[2*]

【摘要】 **目的**　预测日本血吸虫基因组中核糖体移码的基因序列并进行鉴定。**方法**　挑选稳定并能够可靠预测 RNA 假结结构的软件，编写批量提交数据的程序并结合本地手段进行假结结构预测，计算序列最小自由能从而挑选稳定序列，进一步使用生物信息学软件 Fsfinder 分析序列中核糖体移码位点，进行开放性阅读框（open reading frame，ORF）分析，筛选出可能产生核糖体移码的日本血吸虫基因序列。利用日本血吸虫蛋白质组数据库中的肽段质谱数据进行比对，寻找对应的肽段信息。**结果**　从日本血吸虫的 8 452 条基因编码序列中预测出 26 条可能含有促使核糖体移码假结结构的序列。经过日本血吸虫蛋白质组数据库中的肽段质谱数据进行比对，发现日本血吸虫输入蛋白（*Sj*importin）移码之后产生的肽段。**结论**　整合已有的 RNA 假结预测软件以及核糖体移码预测软件，建立了日本血吸虫预测核糖体移码序列数据库，并成功获取日本血吸虫蛋白 *Sj*importin 核糖体移码表达证据。

【关键词】 日本血吸虫；核糖体移码；假结；生物信息学

Prediction and analysis of frameshifts in the genome of *Schistosoma japonicum*

WANG Hai-bin[1,2], *YANG Zhong*[3], *LIU Feng*[3], *HU Wei*[2,3*]. [1]*School of Biotechnology of East China University of Science and Technoloby, Shanghai 200237, China* [2]*National Institude of Parasitic Diseases, Chinese Center for Disease Control and Prevention, Key Laboratory of Parasite and Vector Biology, Ministry of Health, WHO Collaborating Center for Malaria, Schistosomiasis and Filariasis, Shanghai 200025, China* [3]*School of Life Sciences, Fudan University, Shanghai 200433, China*

[*] *Corresponding author: HU Wei, Email: huwwyz@163.com,*

Supported by National Key Program of Infectious Diseases (2009ZX1004-302)

【**Abstract**】 **Objective**　To predict and identify the frameshift sequences in the genome of *Schistosoma japonicum*. **Methods**　Perl software was used to predict candidate sequences with pseudoknots of *S. japonicum* automatically. The stable sequences were searched based on the calculation of the minimum free energy and the FSfinder software was used to predict frameshift sites of the sequences, which were compared with proteome database by BLAST. **Results**　From the 8 452 sequences in the *S. japonicum* database, 26 candidate sequences containing both pseudoknot and corresponding frameshift site were selected and the protein *Sj*importin bearing frameshift with the evidence in proteome database of *S. japonicum* was found. **Conclusion**　Combining the RNA pseudoknots prediction software and frameshift prediction software, the database of *S. japonicum* frameshift has been built and the evidence of frameshift for *Sj*importin has been got.

【**Key words**】 *Schistosoma japonicum*; Frameshift; Pseudoknot; Bioinformatics

核糖体移码是指在蛋白翻译过程中，核糖体（5′→3′）有时按一定的效率向 5′方向（－1 移码）或 3′方向（＋1 移码）移动一个碱基使阅读框发生变化的现象[1]。移码会产生一种新的由两个阅读框共同编码的融合蛋白。这种特殊位点上的非常规调控能够提供生物体的复杂多样性，并在蛋白表达层面上提供更多选择。已发现核糖体移码在病毒中广泛存在[2]，但是真核生物中报道较少，有待进一步研究。

核糖体移码通常在一些特征序列处发生。－1 核糖体移码的特征基因序列通常由一个“slippery”序列和一个名为假结（pseudoknot）的 RNA 二级结构共同组成[3]。移码发生于“slippery”序列上，而假结结构起到刺激作用[4]，能阻遏核糖体的移动并迫使其跳入另外一个开放性阅读框（open reading frame，ORF）。＋1 核糖体移码在多种生物中也得以

DOI：10.3760/cma.j.issn.1673-4122.2012.03.003

基金项目：艾滋病肝炎等传染病重大专项（2009ZX1004-302）

作者单位：[1]200237 上海，华东理工大学生物工程学院；[2]200025 上海，中国疾病预防控制中心寄生虫病预防控制所，卫生部寄生虫病原与媒介生物学重点实验室，世界卫生组织疟疾、血吸虫病和丝虫病合作中心；[3]200433 上海，复旦大学生命科学院

[*] 通信作者：胡薇，Email：huwwyz@163.com

发现,如原核生物中蛋白质链释放因子 B(polypeptide release factor, prfB)编码的释放因子 2(release factor, RF2)[5],以及真核生物中鸟氨酸脱羧酶抗酶(ornithine decarboxylase antizyme, ODC)。在鸟氨酸脱羧酶抗酶核糖体框移过程中,RNA 上的假结结构也是必需的[6]。因此假结结构的发现和预测是研究核糖体移码突变的重要切入点。

假结是一种稳定的 RNA 二级结构,首次发现于烟草花叶病毒[7]。其最简形式由两个螺旋组成,中间由单链或环形连接,单链的环区域经常和邻近的茎区域发生反应(loop1-stem2 或者 loop2-stem1)来形成疏水键并参与到整个分子构象[8]。由于茎和环作用不同,假结可以形成一个结构不同的群体,并在选择性基因表达中起到至关重要的作用[9-10]。RNA 二级结构预测方法通常有基于热力学最小自由能方法以及比较序列分析方法两种,其中随机文法方法将 RNA 序列看为一定语法规则的语句,通过配对关系来得到二级结构。该方法计算复杂度较高,需要相关的先验知识。最小自由能方法假设 RNA 最稳定的二级结构就是正常二级结构,按照热力学理论,在稳定状态时其自由能最小,而通过对 RNA 序列给定含有自由能参数的热力学模型,就可以求得自由能最小的二级结构[11]。目前常采用 Pknots-RG、GeneBee 和 Dotknot 软件。

Pknots-RG 采取三种算法来预测假结结构,第一为传统的最小自由能(minimum free energy, MFE)算法;第二为增强折叠型算法,在该算法中,程序会报告一个最佳的含有最少一个假结的折叠结构,这在怀疑含有假结但并未用 MFE 算法检测出来时相当重要;第三种算法为本地折叠算法,该算法计算出的假结不是含有最低能量的结构,而是为最佳能量与长度比,该算法有助于预测错误折叠结构中含有的假结结构。GeneBee 是俄罗斯研发的一个程序,它是通过序列比对的方法来预测 RNA 二级结构的,通过在 GenBank 中匹配相似的区域特征来达成对于 RNA 二级结构的预测。通过采用补偿替代算法,它可以来连配序列从而达到预测二级结构的目的。Dotknot 软件在预测长序列上有着很大的优势。

由于在预测假结结构的最小自由能时会遇到 NP-problem 问题,实际上对包括假结结构的 RNA 结构预测的算法很难解决预测时间长、精确性差的问题。因此在基因序列中寻找可能的假结时需要启发性的方法来进行预测,然后获取其他数据进行支持。

1 材料和方法

1.1 日本血吸虫编码基因序列及蛋白质组数据来源

本研究中 8 452 条日本血吸虫基因编码序列来自于中国国家人类基因组计划南方研究中心。原下载地址为:http://function.chgc.sh.cn/sj-proteome/search/protein_search_all.php,现更新后下载地址为:http://www.chgc.sh.cn/japonicum/Resources.html。日本血吸虫蛋白质组数据库包括日本血吸虫蛋白质的质谱数据,质谱方法为 QSTAR Pulser I[12],获取地址为:http://function.chgc.sh.cn/sj-proteome/download.htm。

1.2 预测软件及数据库来源

Pknots-RG 软件在线地址:http://bibiserv.techfak.uni-bielefeld.de/pknotsrg,Dotknot 软件在线地址:http://dotknot.csse.uwa.edu.au, Genebee 软件在线地址:http://www.genebee.msu.su/genebee.html, Pseudobase 数据库在线地址:http://ekevanbatenburg.nl/PKBASE/PKBGETCLS.HTML,FSfinder 软件在线地址:http://wilab.inha.ac.kr/FSFinder/。

1.3 批量提交系统环境

硬件环境为 Dell 台式电脑,Linux 服务器[taihu(Dell 2Xeon2.4GHTcpu 8G 内存), xihu (Dell 2Xeon2.4GHTcpu 4G 内存)],RedHat Linux 操作系统,perl3.0 程序语言。

1.4 评价三种程序可信度的方法

从已知假结数据库 Pseudobase(http://ekevanbatenburg.nl/PKBASE/PKBABOUT.HTML#S1)中一共随机抽取共 50 条序列进行验证,测试 Pknots-RG、Dotknot 以及 Genebee 三种软件的准确度。

1.5 批量在线提交程序的编写及软件的本地化

1.5.1 GeneBee 软件在线批量提交程序的编写

通过 perl 语言中的循环语句,使数据逐条提交至网址 http://www.genebee.msu.su/genebee.html,并判断返回值中是否含有“pseudoknot”字样,编写 perl 程序时,采取 POST 方法获取网页返回值。

1.5.2 Dotknot 软件批量提交程序的编写

通过编程语言中的循环语句,使数据逐条提交至网址 http://dotknot.csse.uwa.edu.au,并判断返

回值中是否含有( )和[ ]相互嵌套的部分字样,编写 perl 程序时,采取 POST 方法获取网页返回值。

1.5.3 Pknots-RG 软件的本地运行

采用现有 Pknots-RG 本地软件进行分析预测。将 fasta 格式数据提交到本地软件中,返回值由软件自动生成。

1.6 核糖体移码位点预测及数据处理

使用 FSfinder 手动进行核糖体移码位点预测。首先,使用 Fsfinder 预测序列中可能存在的移码位点并记录。然后按照从 3′向 5′为正方向,只有在某个 ORF 中间内部存在移码位点,并且移码位点和假结结构距离 4 ~ 11 nt 时,才认为其有可能发生移码的现象。然后将整个序列按照移码序列的情况来进行翻译,并将其数据记录下来。

1.7 利用蛋白质组数据验证假结结构及对应的程序性核糖体移码现象

为了检验生物信息学假结结构预测方法的可靠性并借此确认日本血吸虫中实际存在核糖体移码现象,使用 MASCOT 软件对日本血吸虫蛋白质组数据库中肽段数据进行比对,通过将基因序列和日本血吸虫的蛋白质谱数据进行比对,发现日本血吸虫体内实际发生程序性核糖体移码的蛋白,从而在蛋白质组层面上确认核糖体移码现象的发生。用日本血吸虫蛋白质组质谱数据对筛选出的基因核酸序列的正向库和反向库(DNA 序列)进行分别搜索。利用本地 BLAST 手段,将前步返回的肽段序列与挑选预测含有核糖体移码现象的日本血吸虫基因的正向 DNA 序列进行 BLAST 比对,获得具体的匹配基因和位置,从而发现日本血吸虫蛋白质组中存在的核糖体移码现象。

2 结果

2.1 软件可靠性分析

利用假结数据库 Pseudobase 中 50 条已经确定的假结序列评价这三个预测软件 Pknots-RG、Dotknot 以及 Genebee 的实际预测情况(表 1),检验结果显示:Pknots-RG 预测可靠性为 64% (32/50),Dotknot 预测可靠性为 88% (44/50),GeneBee 预测可靠性为 62% (32/50),而三种软件预测结果交集的可靠性为 44% (22/50)。

表 1 Pseudobase 数据库中挑选数据测试结果

Table 1 The result of testing the data from Psudobase

序列名 Sequence name	物种来源 Species	Pknots-RG	Dotknot	Genebee
BChV	beet chlorosis virus	+	+	−
BEV	*Berne virus*	+	+	+
BLV	Bovine leukemia virus	+	+	+
BWYV	Beet western yellows virus	+	+	−
BYDV – NY – RPV	barley yellow dwarf virus	+	+	+
CABYV	cucubit aphid – borne yellows virus	−	−	+
EAV	equine arteritis virus	+	+	+
EIAV	Equine infectious anemia virus	−	+	−
FIV	Feline immunodeficiency virus	+	+	+
HCV 229E	human coronavirus 229E	−	+	+
Hs Ma3	*Homo sapiens*	+	+	+
IBV	infectious bronchitis virus	−	+	−
LDV – C	lactate dehydrogenase – elevating virus, strain C	+	+	+
MMTV_gag/pro	mouse mammary tumor virus	+	+	−
Mm_Edr	*Mus musculus* (mouse)	+	+	+
PEMV	pea enation mosaic virus	+	+	−
PLRV – S	potato leafroll virus	+	+	−
PLRV – W	potato leafroll virus	−	−	−
PRRSV – 16244B	Porcine reproductive respiratory syndrome virus, North American isolate (16244B)	−	+	−
PRRSV – LV	Porcine reproductive respiratory syndrome virus, European Lelystad strain (LV)	−	+	−

(未完待续)

（续表 1）

序列名 Sequence name	物种来源 Species	Pknots-RG	Dotknot	Genebee
RSV	Rous sarcoma virus	–	+	–
SARS – CoV	SARS coronavirus	+	+	+
SRV1_gag/pro	simian retrovirus – 1	+	+	+
ScYLV	sugarcane yellow leaf virus	+	+	+
VMV	Visna – Maedi virus	–	+	+
WBV	White Bream Virus	+	+	+
AKV – MuLV	AKV murine leukemia virus	+	+	+
BaEV	baboon endogenous virus	–	+	+
Cas – Br – E – MuLV	Cas – Br – E murine leukemia virus	+	+	+
FeLV	Feline leukemia virus	+	+	–
GaLV	gibbon ape leukemia virus	+	+	–
Mo – MuLV	Moloney murine leukemia virus	+	+	+
SNV	spleen necrosis virus	–	+	+
AMV3	alfalfa mosaic virus	–	–	–
APLV	andean potato latent virus	+	+	+
BBMV3	broad bean mottle virus	+	–	+
BMV3	brome mosaic virus	–	+	+
BSBV1	beet soil – borne virus	+	+	+
BSBV2	beet soil – borne virus	+	+	+
BSBV3	beet soil – borne virus	+	+	+
BSMVbeta	barley stripe mosaic virus	–	+	–
BVQ1	beet virus Q	–	+	–
BVQ2）	beet virus Q	+	+	+
BVQ3	beet virus Q	+	+	+
CaYMV	cacao yellow mosaic virus	+	+	+
CaYVV	calopogonium yellow vein virus	+	+	–
CcTMV	tobacco mosaic virus	–	+	+
CCMV3	cowpea chlorotic mottle virus	–	–	+
CGMMV	cucumber green mottle mosaic virus	+	+	–
CGMMV_PKbulge	cucumber green mottle mosaic virus	–	–	–

– ：表示没有预测出包含假结的序列，+ ：表示预测出来包含假结的序列

– ;indicates the sequence which was not predicted to contain pseudoknot, + : indicates the sequence which was predicted to contain pseudoknot

2.2　数据批量在线提交及本地化

利用三种软件分别分析日本血吸虫数据库中的 8 452 条序列，通过对序列进行自由能分析，将自由能小于 4 kkcal/mol 的序列看作稳定的假结结构，获得 151 条预测含有稳定假结结构的序列。

2.3　核糖体移码位点预测及数据处理

使用 FSfinder 软件预测 151 条序列中的核糖体移码位点，并判断该移码位点和假结以及 ORF 在序列上是否相互对应，分析获得既含有核糖体移码位点也含有对应假结结构以及 ORF 的序列为 91 条，这 91 条序列即为日本血吸虫预测核糖体移码基因序列。

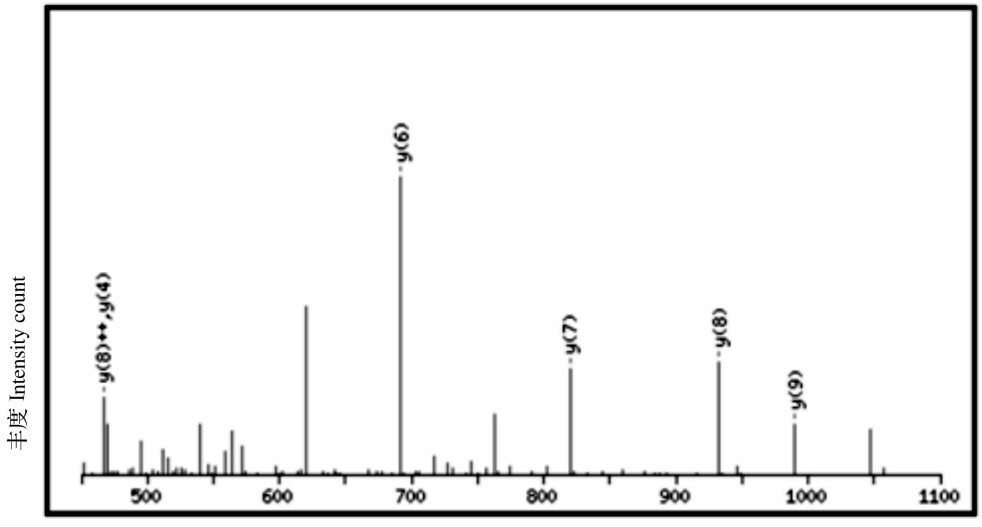
2.4　利用蛋白质组数据验证假结结构及对应的程序性核糖体移码现象

2.4.1　质谱数据比对

用日本血吸虫蛋白质组质谱数据对筛选出的 91 个基因核酸序列的正向库和反向库（都是 DNA 序列）进行分别搜索。在质谱数据里，有1 395个肽段在这个 DNA 库里得到了匹配。删除重复序列数据后剩余 103 条。经过 BLAST 比对，在这 91 条基因序列中，排除假阳性，能够被质谱检测到的序列剩余 7 条。

2.4.2　读码位置搜索

通过在 7 条序列中进行位置的匹配，发现序列编号为 SJCHGC09414 的基因序列按照 +2 及 +3 读码顺序翻译产生的蛋白质和日本血吸虫蛋白质组数据库中存在数据相吻合的情况，从而在蛋白质组层面上发现了日本血吸虫中核糖体移码现象的产生（图 1、2）。其中，+2 读码顺序时的匹配蛋白为 AVGLQHSALTY，+3 读码顺序时的匹配蛋白为 KALGGF。AVGLQHSALTY 位于 +2 读码的 375 ~ 385 氨基酸处，KALGGF 位于 +3 读码的 398 ~ 403 氨基酸处，其一致性均为 100%，说明序列编号为 SJCHGC09414 的基因序列同时按照 +2 及 +3 读码顺序翻译时在日本血吸虫体内产生蛋白，与预测的结果相符。



MS/MS Fragmentation of **AVGLQHSALTY**
Found in **gi|56757933**, Schistosoma japonicum SJCHGC09414 protein mRNA, complete cds
Translated in frame 2 (nucleic acid sequence)
Match to Query 202: 1158.537168 from(580.275860,2+)
Title: File: sj-egg-shell-scx-4 (recalibrated).wiff, Sample: Sample001 (sample number 1), Elution: 54.2 min, Period: 1, Cycle(s): 328 (Experiment 4)
Data file \\tsclient\H\SJ-projects\Sj-MSMS-data\Ion-peaks\Cen-peaks\unzip\sj-egg-shell-scx-4.mgf
Monoisotopic mass of neutral peptide Mr(calc): 1158.6033 **Ions Score:** 39 **Expect:** 0.00023 **Matches:** 6/86 fragment ions using 7 most intense peaks

图 1　*Sj*importin 蛋白 +2 读码与日本血吸虫蛋白质组数据库比对结果

Fig. 1　*Sj*importin protein ORF +2 BLAST with proteomic database of *Schistosoma japonicum*

2.5　日本血吸虫输入蛋白生物信息学功能分析

将 SJCHGC09414 序列在 NCBI 的 nr 数据库进行 BLAST 分析，发现该蛋白包括三个结构域，均属于 armadillo/beta-catenin repeat（ARM）超家族，通过和其他生物进行多序列联配发现三个结构域相对保守。该蛋白在曼氏血吸虫中也有同源蛋白表达，两者的序列一致性为 86%。通过和其他生物的比对，将其命名为日本血吸虫输入蛋白（*Sj*importin）。*Sj*importin 和其余物种的多序列联配图见图 3。输入蛋白广泛分布于各生物中，经过 BLAST 比对以及进化树分析后发现 SJCHGC09414 蛋白为输入蛋白 alpha-2 亚基家族，和曼氏血吸虫输入蛋白一致性为 86%，和人类的一致性为 45%（图 4）。

3　讨论

核糖体移码在蛋白表达以及生物过程调控中起到相当重要的作用，假结结构的预测是发现核糖体移码的重要手段。然而，假结结构是 RNA 二级结构预测的难点。目前，预测假结结构的方法还不完善，精确度和灵敏度均有待提高。本研究通过对目前流行的 RNA 二级结构预测软件进行测试发现：无论从实用性还是从可靠性来说，利用单一软件进行分析是不够的，因此结合多种预测软件是获得较为精确的 RNA 假结结构的必要条件。

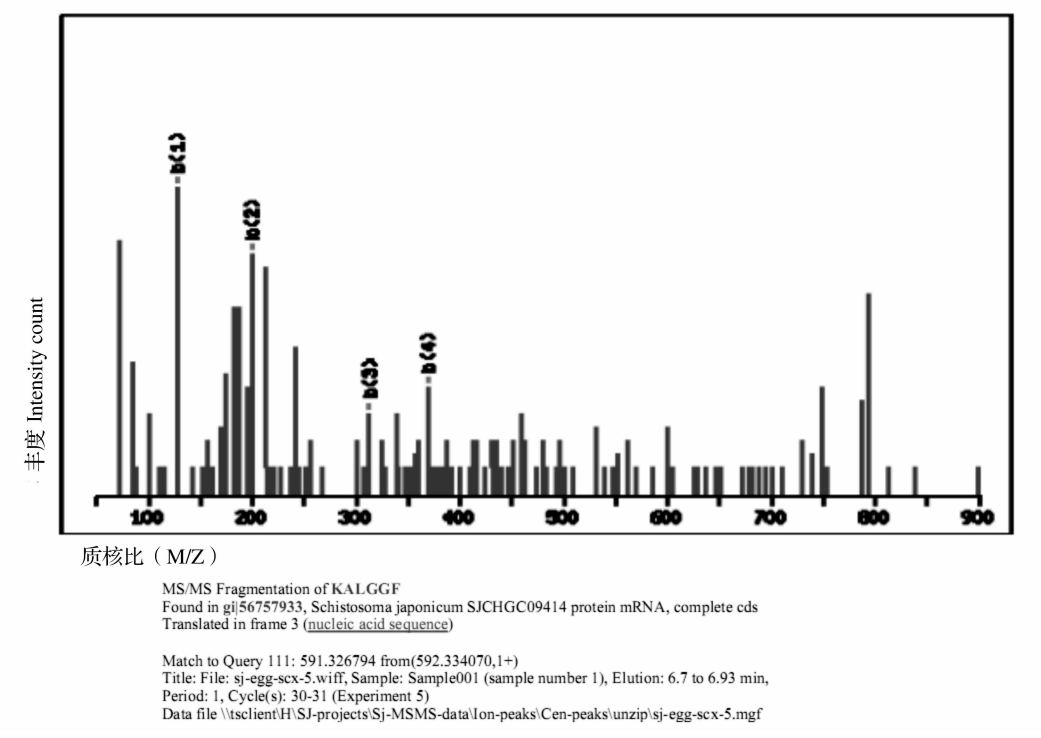


图 2　*Sj*importin 蛋白 +3 读码与日本血吸虫蛋白质组数据库比对结果

Fig. 2　*Sj*importin protein ORF +3 BLAST with proteomic database of *Schistosoma japonicum*

此外，数据库的质量对于生物信息学的预测起到了至关重要的作用。数据库中序列的每一个错配都可能导致错误的结果。即便采用 3 种预测假结结构的软件共同计算，一同缩小样品数量，提高精确度，也不能避免其中错误预测的产生。因此要提高预测的成功率，必须改进序列数据库的质量，提高原本数据的可靠性。

本研究通过整合现有假结结构预测软件以及核糖体移码预测软件，利用程序语言，编写出自动在线提交的程序，优化了生物信息学分析的工作流，对日本血吸虫基因编码数据库进行分析，经过预测及鉴定，在日本血吸虫蛋白编码数据库中发现 *Sj*importin 蛋白在日本血吸虫体内存在移码表达的现象。输入蛋白在真核生物体内起到将分子从细胞质（cytoplasm）转运到细胞核内的作用。通常来说，输入蛋白介导的转运发生在细胞核孔。由于绝大多数的蛋白都不能够将自身运输穿过核孔，因此输入蛋白无疑在蛋白质运输过程中发挥着重要作用。*Sj*importin 蛋白在日本血吸虫的功能仍有待进一步研究，特别是其核糖体移码的翻译调控方式的潜在作用值得关注。

在以上的研究中既预测了日本血吸虫中的假结结构并建立数据库，也预测了核糖体移码现象，这些研究拓展了核糖体移码对于蛋白表达调控的认识，为今后药物靶点筛选的进一步研究打下了基础。

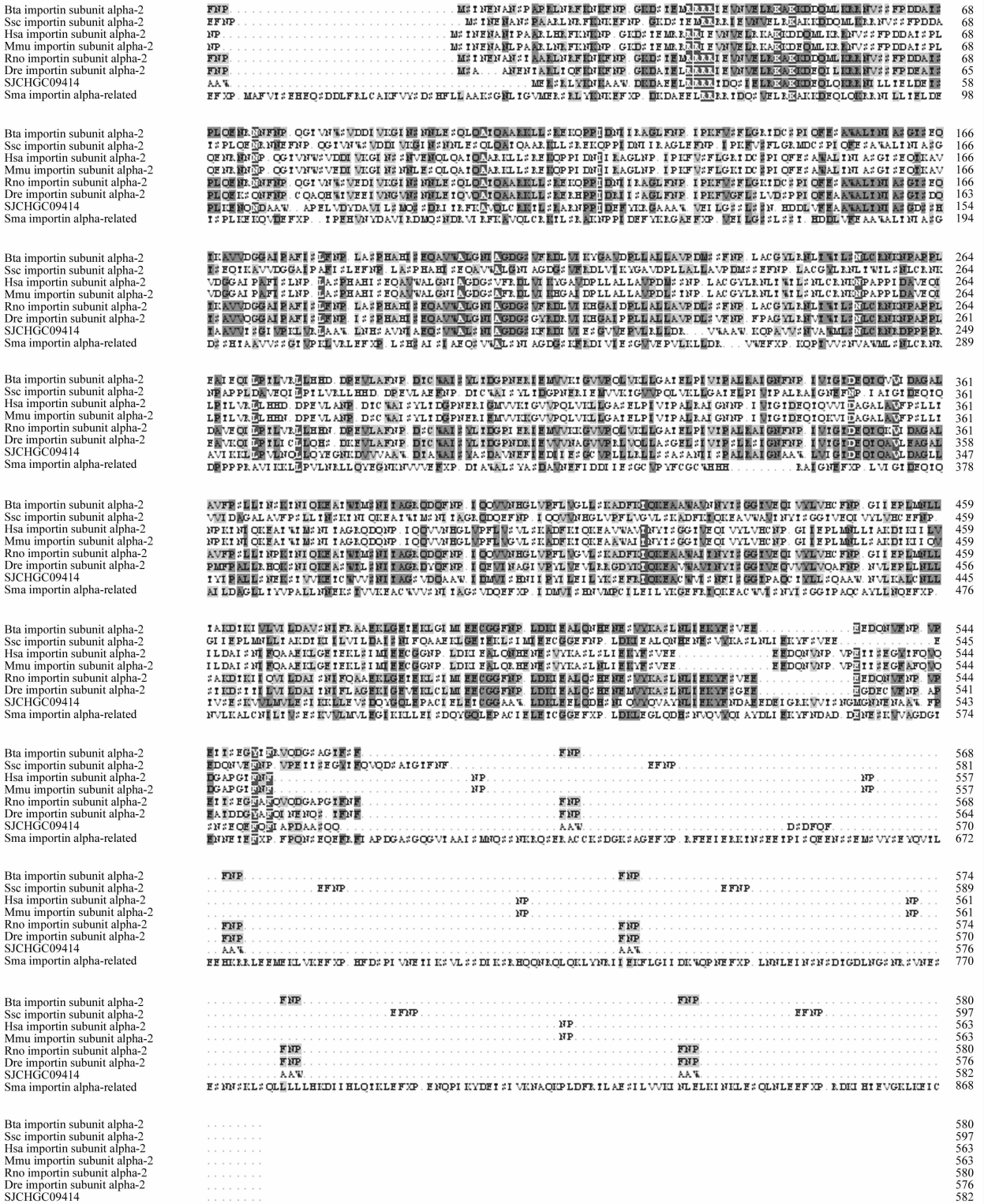


图 3　*Sj*importin 与其他物种序列的同源比对分析

Sma：曼氏血吸虫，Dre：斑马鱼，Rno：褐家鼠，Mmu：小家鼠，Hsa：人类，Ssc：野猪，Bta：牛

Fig. 3　Homologous alignment of *Sj*importin sequence with other species

Sma：*S. mansoni*，Dre：*Danio rerio*，Rno：*Rattus norvegicus*，Mmu：*Mus musculus*，

Hsa：*Homo sapiens*，Ssc：*Sus scrofa*，Bta：*Bos taurus*

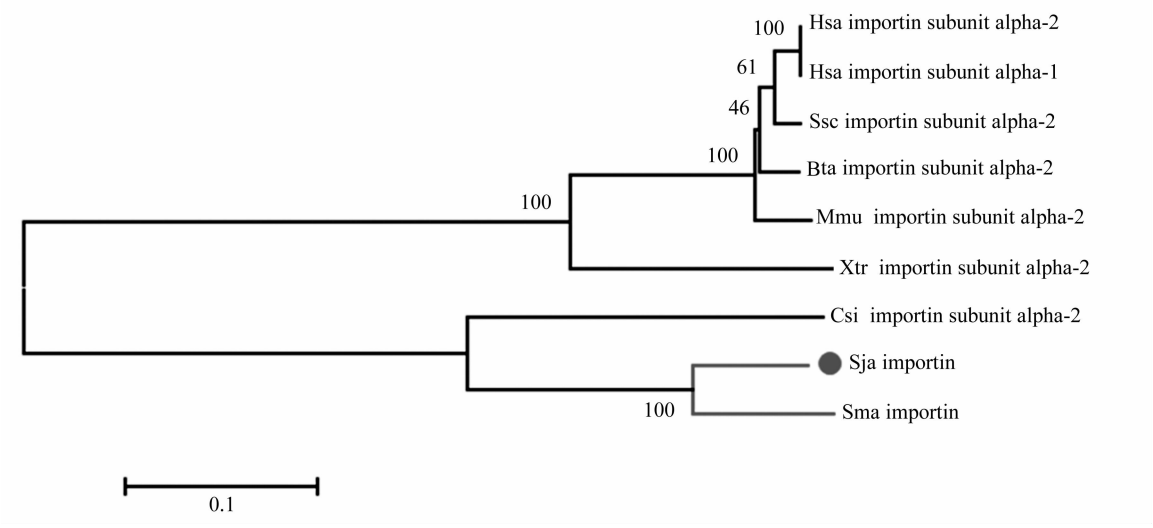


图 4　*Sj*importin 进化树

Sja:日本血吸虫,Sma:曼氏血吸虫,Csi:中华支睾吸虫,Xtr:热带爪蟾,Mmu:小鼠,Bta:牛,Ssc:野猪,Hsa:人

Fig. 4　Phylogenetic tree of the *Sj*importin

Sja:*S. japonicum*, Sma:*S. mansoni*, Csi:*Clonorchis sinensis*, Xtr:*Xenopus tropicalis*, Mmu:*Mus musculus*, Bta:*Bos taurus*, Ssc:*Sus scrofa*, Hsa:*Homo sapiens*

参 考 文 献

[1] 文宏津,龚炳永. 程序性核糖体移码与抗病毒药物筛选[J]. 国外医药抗生素分册, 1999, 20(4):148-152.

[2] Nixon PL, Rangan A, Kim YG, et al. Solution structure of a luteoviral P1-P2 frameshifting mRNA pseudoknot[J]. J Mol Biol, 2002, 322(3): 621-633.

[3] Moon S, Byun Y, Kim HJ, et al. Predicting genes expressed via +1 and -1 frameshifts[J]. Nucleic Acids Res, 2004, 32(16): 4884-4892.

[4] Atkins JF, Gesteland RF. Intricacies of ribosomal frameshifting [J]. Nat Struct Biol,1999, 6(3): 206-207.

[5] Weiss RB, Dunn DM, Atkins JF, et al. Slippery runs, shifty stops, backward steps, forward hots: -2, -1, +1, +2, +5, and +6 ribosomal frameshifting[J]. Cold Spring Harb Symp Quant Biol, 1987, 52: 687-693.

[6] Ivanlov IP, Gesteland RF, Atkins JF. Atizyme expression: a subversion of triplet decoding, which is remarkably conserved by evolution, is a sensor for an autoregulatory circuit[J]. Nucleic Acids Res, 2000, 28(17): 3185-3196.

[7] Rietveld K, Van Poelgeest R, Pleij CW, et al. The tRNA-like structure at the 30 terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA[J]. Nucleic Acids Res, 1982, 10(6): 1929-1946.

[8] Staple DW, Butcher SE. Pseudoknots: RNA structures with diverse functions[J]. PLoS Biol, 2005, 3(6): e213.

[9] Theimer CA, Blois CA, Feigon J. Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function[J]. Mol Cell, 2005, 17 (5): 671-682.

[10] Brierley I, Digard P, Inglis SC. Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot[J]. Cell, 1989, 57(4): 537-547.

[11] Eddy SR, Durbin R. RNA sequence analysis using coariance models[J]. Nucleic Acids Res, 1994, 22(11): 2079-2088.

[12] Liu F, Lu J, Hu W, et al. New perspectives on host-parasite interplay by comparative transcriptomic and proteomic analyses of *Schistosoma japonicum*[J]. PLoS Pathog, 2006, 2(4): 268-281.

（收稿日期:2012-02-17）
（本文编辑:陈勤）