论著·基础研究 doi:10.3969/j.issn.1671-8348.2015.17.003

果蝇 Tap 蛋白结构与功能的生物信息学分析*

刘洪超[1],胡 澍[1△],涂心明[2]
(河南科技大学医学院:1.生物医学实验教学中心/神经遗传实验室;
2.人体解剖与组织胚胎学教研室,河南洛阳 471003)

摘 要:目的 利用生物信息学方法分析果蝇 Tap 蛋白的结构和功能。**方法** 基于 NCBI 数据库中果蝇 Tap 蛋白的氨基酸序列,从蛋白质理化性质、跨膜区、信号肽、亚细胞定位、结构域、三维结构及物种间同源蛋白进化关系等方面进行分析。**结果** Tap 蛋白为不稳定亲水性蛋白,无跨膜区和信号肽,在细胞核中发挥生物学效应,具有碱性螺旋-环-螺旋(bHLH)结构域。Tap 蛋白与来源于 NeuroD1 的模板 2ql2.1.B 有 61.02%的氨基酸序列一致,Tap 蛋白与人类和啮齿动物的编码产物高度同源,在系统发生树中距离最近。**结论** 果蝇 Tap 蛋白具有 bHLH 蛋白家族的典型结构,可能在果蝇胚胎发育早期的神经发生、分化等过程中发挥作用。

[关键词] Tap 蛋白;螺旋-环-螺旋构型;分子结构;计算生物学;果蝇,黑腹
中图分类号:R394.1;Q963 **文献标识码**:A **文章编号**:1671-8348(2015)17-2311-04

Bioinformatics analysis of the structure and function of Drosophila Tap protein*

Liu Hongchao[1], *Hu Shu*[1△], *Tu Xinming*[2]
(1. *Laboratory of Neurogenetics/Biomedical Experimental Training Center*; 2. *Department of Human Anatomy, Histology and Embryology, Medical College, Henan University of Science and Technology, Luoyang, Henan* 471003, *China*)

[**Abstract**]**Objective** To analyze the structure and function of Drosophila Tap protein using bioinformatics methods. **Methods** Based on the amino acid sequences of Drosophila Tap protein from NCBI database, the bioinformatics analyses were performed to unravel the physicochemical property, the transmembrane region, the signal peptide, the subcellular localization, the domain, the tertiary structure, the phylogenetic tree of Tap protein and Tap related proteins from other species. **Results** Tap protein was an unstable hydrophilic protein, playing biological function in the nucleus. It contained a basic helix-loop-helix(bHLH) domain, but without transmembrane region and signal peptide. The amino acid sequence identity between Tap protein and NeuroD1-derived template 2ql2.1.B was 61.02%. Tap protein and its related proteins in human and rodents were most close in the phylogenetic tree, showing high homology. **Conclusion** Tap protein contains the typical bHLH structure and may play a role in the neurogenesis, neural precursor cell differentiation in early embryonic stage of Drosophila.

[**Key words**]Tap protein; basic helix-loop-helix motifs; molecular structure; computational biology; drosophila melanogaster

碱性螺旋-环-螺旋(basic helix-loop-helix,bHLH)蛋白家族广泛分布于生物界,存在较高的保守性。大部分 bHLH 蛋白以同源二聚体或异源二聚体发挥转录因子作用,主要调节各种干细胞向终末细胞的分化,在骨骼肌、胰腺发育、血液发生、果蝇神经干细胞分化以及脊椎动物脊髓、端脑皮层的发育过程中发挥着重要作用[1-4]。果蝇 *Tap* 基因,也称为 *biparous*,分别由两个研究小组针对 mec-3 和 delilah 的 bHLH 结构域通过低保真度杂交筛选和 PCR 筛选而鉴定出来,并被定位于 3 号染色体左臂(17359682-17361811)[5-6]。其序列全长约 2.13 kb,细胞遗传学定位于 74A5。基因序列比对结果表明果蝇 *Tap* 基因与脊椎动物的 *neurogenin*(*ngn*)和 *neuroD* 基因关系最为密切,是 *ngn* 基因的直系同源基因,然而 Tap 在果蝇发育过程中的功能至今未知[6-7]。对其进行功能与调控研究既具有原创性,又对哺乳动物 *ngn* 基因的相关研究具有重要的参考和借鉴意义。

基因及其表达物的结构与功能的解析和预测对生物科学、医药科学等领域的研究与开发有着极为重要的指导作用。生物信息学运用信息科学的原理和技术处理大量分散和复杂的生物学和医学科学数据,使其更易解读,更富有具体指向,由此成为生命科学领域研究所必需的工具。本研究基于 NCBI 数据库中果蝇 Tap 蛋白的氨基酸序列,应用生物信息学方法对 Tap 进行理化性质、跨膜区、信号肽、亚细胞定位、结构域、三维结构等方面的预测,分析其可能的生物学功能,并通过与其他物种同源蛋白氨基酸序列的比对和分析,为系统研究果蝇 *Tap* 基因的功能与调控提供参考。

1 材料与方法

1.1 研究基因及蛋白 果蝇 *Tap* 基因 mRNA 序列:NM_079400.3;果蝇 Tap 蛋白氨基酸序列:NP_524124.1。

1.2 方法

1.2.1 *Tap* 基因开放阅读框(ORF)分析 含有开放阅读框是一个 DNA 序列编码特定蛋白质的部分或全部的先决条件。本研究采用 ORF Finder 在线工具分析和界定果蝇 *Tap* 基因的开放阅读框。

1.2.2 Tap 蛋白理化性质分析 蛋白质的理化性质是蛋白质

* **基金项目**:国家自然科学基金面上项目(31171256);自然科学基金河南科技大学配套经费(09001504);河南科技大学博士启动基金(13530057)。 **作者简介**:刘洪超(1990-),在读硕士,主要从事神经发育和遗传研究。 △ **通讯作者**,Tel:(0379)64810765;E-mail:hushu51@sina.com。

研究的基础，本研究采用 ProtParam 工具预测 Tap 蛋白的相对分子质量、氨基酸组成、理论等电点 PI、半衰期、不稳定系数、总平均亲水性等。

1.2.3 Tap 蛋白二级结构和疏水性分析 蛋白质的二级结构主要指多肽链中主链原子在各局部空间的排列分布情况，基本的二级结构包括 α-螺旋，β-折叠，β-转角，无规卷曲等结构组件，对其预测和分析有助于认识蛋白质的空间结构。Predict-Protein 工具可以对蛋白质的二级结构、模体、二硫键结构、跨膜区等许多结构信息进行预测分析，平均准确率超过 72%，最佳残基预测准确率达 90%以上，被视为蛋白质二级结构预测的标准。本研究采用该工具中的 PROFSec 和 PROFAcc 分别对 Tap 蛋白的二级结构和疏水性进行预测。

1.2.4 Tap 蛋白跨膜区预测 目前常用的蛋白质跨膜区分析工具是依靠跨膜蛋白数据库来预测跨膜区位置和跨膜方向。本研究中，TMHMM 软件包和 TMpred 程序被用来对 Tap 蛋白的跨膜区进行预测。

1.2.5 Tap 蛋白信号肽预测 信号肽是决定新生肽在细胞中定位或决定某些氨基酸残基修饰的肽段，通常由 15～30 个氨基酸残基组成，位于新生肽的 N 端。本研究中 SignalP 4.1 Server 自动利用神经网络模型对 Tap 进行信号肽预测，得到 3 种 C、Y、S-score 计算结果[8]，同时，NLStradamus 在线工具被用来预测 Tap 可能的核定位信号(nuclear localization signal，NLS)。

1.2.6 Tap 蛋白亚细胞定位预测 蛋白质的亚细胞定位是指某种蛋白在细胞内的具体存在部位，Tap 蛋白的亚细胞定位采用 PSORT 和 LOCtree 来进行分析[9]。

1.2.7 Tap 蛋白结构域和功能预测 结构域是蛋白质序列的结构、功能和进化单元，对于蛋白质结构和功能的研究具有重要意义。Tap 蛋白的结构域和功能分类采用 InterPro 预测。该工具整合了 UniProt、CATH-Gene3D、PROSITE 等 16 个成员数据库，充分利用各成员库的优势，整合蛋白质家族、结构域和功能位点等资源于一体，通过蛋白质家族分类、结构域和重要功能位点预测等途径对蛋白质进行功能注释[10]。

1.2.8 Tap 蛋白三维结构预测 蛋白质三维结构的预测方法通常包括同源建模法、串线法/折叠识别法和从头预测法，而同源建模法是基于序列同源比对，对于序列相似度大于 30%的序列模拟比较有效，是目前最常用的方法。本研究运用 SWISS-MODEL 蛋白质结构服务器对 Tap 蛋白进行同源性建模以分析其三维结构。

1.2.9 Tap 蛋白氨基酸序列同源性分析 将果蝇 Tap 蛋白的氨基酸序列输入 BLASTP，采用 BLASTp 程序检索 Swiss-Prot 数据库进行同源蛋白的搜索与比较。

2 结 果

2.1 果蝇 *Tap* 基因 ORF 分析 ORF Finder 工具在 *Tap* 基因 mRNA 序列中寻找出一个长 1 197 bp 的 ORF，起始密码子位于 432 bp，终止密码子位于 1 628 bp，推测编码 398 个氨基酸残基，见图 1。

2.2 果蝇 Tap 蛋白基本理化性质分析 ProtParam 分析工具对 Tap 蛋白的理化性质进行了预测，其氨基酸组成见表 1。果蝇 *Tap* 基因编码 398 个氨基酸，含量最多的两种氨基酸是脯氨酸(Pro)和丝氨酸(Ser)，所占比例分别为 10.1%和 9.0%。负电荷残基总数[天冬氨酸(Asp)＋谷氨酸(Glu)]为 43，正电荷残基总数[精氨酸(Arg)＋赖氨酸(Lys)]为 39。其分子式为 $C_{1987}H_{3007}N_{569}O_{606}S_9$，理论分子量约 44.85×10^3，理论等电点为 6.49。不稳定系数为 66.79，属于不稳定蛋白；脂肪系数为 54.22，总平均亲水指数为－0.849，预测 Tap 蛋白为水溶性蛋白。

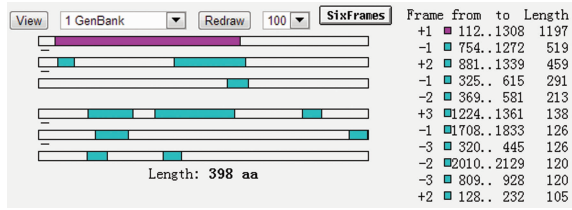

图 1 果蝇 *Tap* 基因序列的开放阅读框分析

表 1 果蝇 Tap 蛋白的氨基酸组成

氨基酸	*n*	含量(%)	氨基酸	*n*	含量(%)
Ala (A)	28	7.0	Leu (L)	29	7.3
Arg (R)	23	5.8	Lys (K)	16	4.0
Asn (N)	18	4.5	Met (M)	8	2.0
Asp (D)	22	5.5	Phe (F)	24	6.0
Cys (C)	1	0.3	Pro (P)	40	10.1
Gln (Q)	32	8.0	Ser (S)	36	9.0
Glu (E)	21	5.3	Thr (T)	21	5.3
Gly (G)	23	5.8	Trp (W)	2	0.5
His (H)	17	4.3	Tyr (Y)	14	3.5
Ile (I)	8	2.0	Val (V)	15	3.8

2.3 Tap 蛋白的二级结构和疏水性分析 PROFSec 预测 Tap 蛋白的二级结构主要为环(Loop)，占 87.19%，α-螺旋为 12.81%，无 β-折叠结构。PROFAcc 预测 Tap 蛋白的亲水性高，表面暴露超过 16%的氨基酸残基占 54.02%，与 ProtParam 预测结果一致，Tap 蛋白可能为水溶性蛋白。

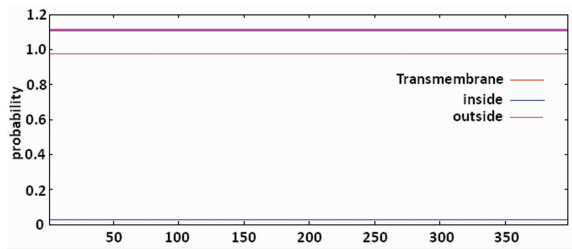

图 2 TMHMM 2.0 对 Tap 蛋白跨膜区的预测

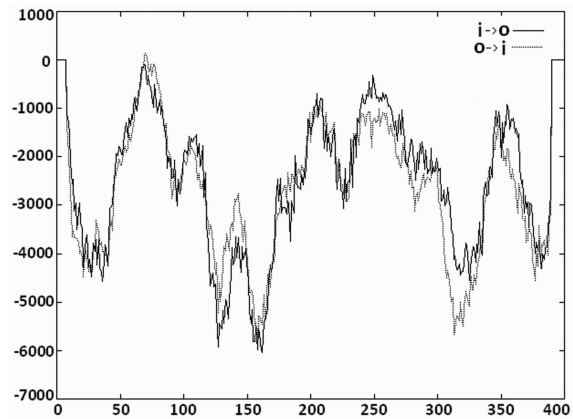

图 3 TMpred 对 Tap 蛋白跨膜区的预测

2.4 Tap 蛋白跨膜区预测 TMHMM Server 和 TMpred 对

Tap 跨膜区预测的结果为：跨膜区中的氨基酸期望值为 0.001 45，该值大于 18 时才有跨膜区，故该结果表明 Tap 蛋白无跨膜区，见图 2～3。

2.5 Tap 蛋白信号肽预测　对于一个典型的信号肽，C-score 和 Y-score 趋向于+1，S-score 在剪切位点之前高，而在剪切位点之后变低。采用神经网络模型预测的信号肽结果如图 4 所示，表明 Tap 蛋白不存在信号肽。NLS 是另一种形式的信号肽，一般含 4～8 个氨基酸残基，富含精氨酸、赖氨酸等碱性氨基酸，可位于多肽序列的任何部分，其介导蛋白质通过核孔复合体入核。NLStradamus 工具分析结果表明 Tap 蛋白氨基酸序列 126～160 位间存在 NLS，Tap 蛋白可能依此定位于细胞核，见图 5。

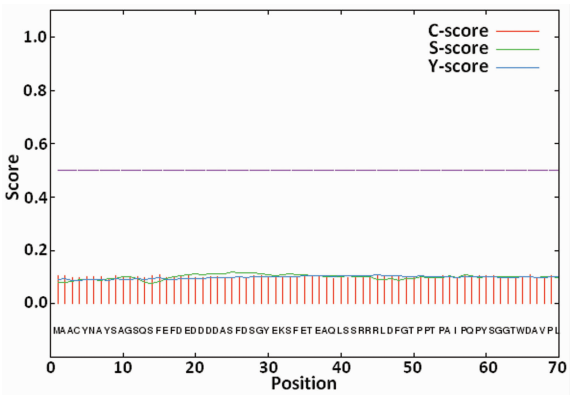

图 4　Tap 蛋白信号肽预测

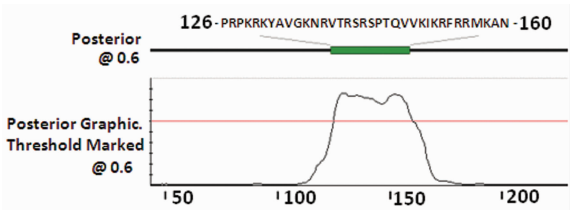

图 5　Tap 蛋白 NLS 预测

2.6 Tap 蛋白亚细胞定位分析　PSORT Ⅱ工具预测结果表明 Tap 蛋白分布于细胞核、细胞骨架、细胞质、分泌系统囊泡的可能性分别为 69.6%、17.4%、8.7%和 4.3%。LOCtree 分析结果表明 Tap 蛋白定位于细胞核，属于 DNA 结合蛋白，与 NLStradamus 预测结果一致。因此 Tap 蛋白可能位于细胞核发挥其生物学功能。

2.7 Tap 蛋白结构域与功能分析　InterPro 对果蝇 Tap 蛋白的结构域和功能分析结果如图 6，Tap 蛋白包含有 Myc-type，bHLH 结构域（IPR011598），具有蛋白质二聚化功能活性（GO:0046983）。

2.8 Tap 蛋白的同源性建模和三维结构预测　同源建模结果表明，Tap 蛋白（155～213）和蛋白质结构数据库中的模板 2ql2.1.B 有 61.02%的氨基酸序列一致[11]，该模板来源于 NeuroD1。以 2ql2.1.B 为模板构建出 Tap 蛋白的三维结构如图 7 所示，模型质量评价结果 GMQE 为 0.07，QMEAN4 为－1.43。

2.9 果蝇 Tap 蛋白氨基酸序列同源性分析　同源性分析结果显示果蝇 Tap 蛋白序列与 NGN1、NeuroD4、NeuroD1、NGN3、NeuroD2、NGN2 序列有 70%、61%、58%、58%、57%、57%的相似性，与人源和鼠源的 NGN1 蛋白同源性最高，据此推测果蝇 Tap 蛋白可能与哺乳动物的 NGN 蛋白为同源蛋白，这也说明了它们在进化过程中的亲缘关系。通过从相关数据库中收集下载序列相似性大于 50%的各物种编码蛋白的资料来构建系统发生树，如图 8 所示，果蝇 Tap 蛋白与人类和啮齿动物的编码产物具有高度同源性，在系统发生树中距离最近，它们可能发挥着相同的生物学功能。

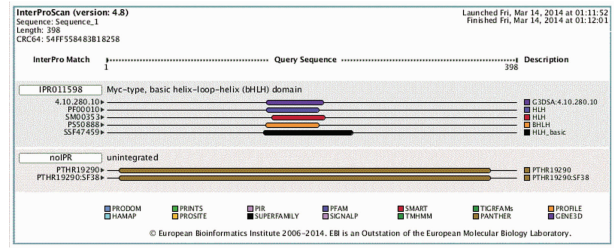

图 6　Tap 蛋白结构域与功能分析

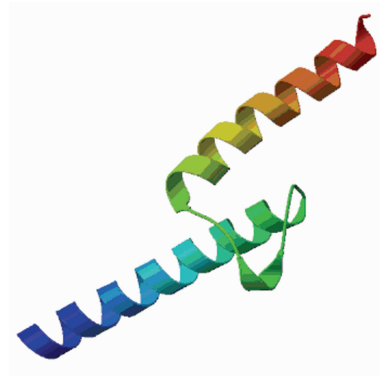

图 7　Tap 蛋白的三维结构预测

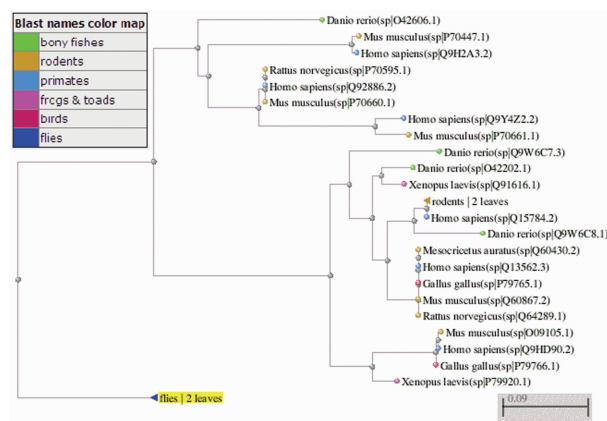

图 8　果蝇 Tap 蛋白氨基酸序列与其他物种间的系统发生树

3　讨　论

基因序列比对表明果蝇 *Tap* 基因与脊椎动物的 *ngn* 和 *NeuroD* 关系最为密切。*ngn* 基因表现出原神经基因的特征，在脊椎动物中启动了神经发生过程，而 *NeuroD* 通常作为原神经蛋白的下游基因调节神经分化过程[12-13]。虽然 *Tap* 基因序列与原神经基因相似，但实际上其在基因激活次序中晚于原神经基因，因此 *Tap* 基因的功能至今未知[6-7]。本研究综合利用多种生物信息学资源和工具，如 ORF Finder、ProtParam、PredictProtein、InterPro、BLASTp 等，对果蝇 Tap 蛋白的性质、结构、功能以及不同物种间氨基酸序列的同源性及进化关系进行了分析。从分析结果可知，果蝇 *Tap* 基因的 ORF 长 1 197 bp，编码 398 个氨基酸，富含 Pro 和 Ser，理论分子量约 44.85×10^3。脂肪系数为 54.22，总平均亲水指数为－0.849，整条多

肽链表现为亲水性，预测 Tap 蛋白为水溶性蛋白。Tap 蛋白的二级结构主要为 Loop，其二级结构容易接近水分子。Loop 作为蛋白质二级结构的一种，与其他二级结构 α-螺旋和 β-折叠不同的是其柔性很大，不能很好地限定于某种特定形式的结构。由于 Loop 结构经常出现在活跃点和对接点，可用于分子识别，所以在蛋白质的特征和功能中起着关键作用。Tap 蛋白无 N 端信号肽，不形成跨膜区，氨基酸序列 126～160 位间存在 NLS，可能定位于细胞核而发挥其生物学功能。结构域和功能分析表明 Tap 蛋白含有 bHLH 结构域，具有蛋白质二聚化活性。bHLH 蛋白含有一段近 60 个氨基酸残基的特征性序列模体，即一个 HLH 结构域和其上游富含碱性氨基酸残基的碱性结构域，HLH 区域参与蛋白质二聚化，而碱性区域与 DNA 序列特异性结合，大部分 bHLH 蛋白以同源或异源二聚体形式与 DNA 结合而发挥生物学功能。蛋白质三维结构同源建模分析发现，果蝇 Tap 蛋白氨基酸序列与来源于 NeuroD1 的模板 2ql2.1.B 有 61.02% 的相似性，表明 Tap 蛋白可能与 NeuroD1 具有相同的结构与功能：NeuroD 通过 HLH 结构域与 TCF3/E47 形成异源二聚体后与 E-box 序列结合。NeuroD 作为转录激活因子，通过结合于启动子 E-box 核心序列活化转录；与神经发生转录调节因子编码基因的增强子调控元件密切相关；参与调节细胞分化与迁移，如促进早期视网膜神经节细胞和内耳感觉神经元的形成，胰腺的胰岛细胞和小肠的肠内分泌细胞形成，在小脑皮质参与树突的发生与维持等[14-15]。氨基酸序列同源性分析结果表明，果蝇 Tap 蛋白与人类和啮齿动物进化距离最近，编码产物具有高度同源性，与人源和鼠源的 NGN1 蛋白同源性最高。果蝇 Tap 蛋白可能与哺乳动物的 NGN 蛋白为同源蛋白，它们可能发挥着相同的生物学功能。NGN 家族是脊椎动物神经系统发育过程中转录调控网络的重要部分，作为转录因子，广泛参与神经发生、细胞分化和细胞谱系决定等过程。在哺乳动物大脑皮质发育过程中，NGN1 和 NGN2 仅表达于神经发生时期的皮质脑室带，与 E12 和 E47 等 bHLH 蛋白形成异源二聚体，启动组织特异性基因表达，促使干细胞向神经元方向分化。

本研究为了提高预测的准确性，对大多数类型的预测均选用了多种不同工具。由于不同预测程序分析时所依据的原理和采用的算法有所不同，它们所预测结果中的一致性部分具有较高的可信度。生物信息学分析表明果蝇 Tap 蛋白具有 bHLH 结构域，与哺乳动物 NGN 在序列上同源性最高，与 NeuroD 具有相同的高级结构，推测果蝇 Tap 蛋白可能作为转录因子参与果蝇神经发生与细胞分化等过程。本文对果蝇 Tap 蛋白的生物信息学分析结果，对深入研究 Tap 蛋白的功能注释与调控网络和进一步了解 *Tap* 基因对果蝇发育的影响提供参考。

参考文献：

[1] Havis E, Coumailleau P, Bonnet A, et al. Sim2 prevents entry into the myogenic program by repressing MyoD transcription during limb embryonic myogenesis[J]. Development, 2012, 139(11): 1910-1920.

[2] Liu XD, Chen X, Zhong B, et al. Transcription factor achaete-scute homologue 2 initiates follicular T-helper-cell development[J]. Nature, 2014, 507(7493): 513.

[3] Bhattacharya A, Baker NE. A network of broadly expressed HLH genes regulates tissue-specific cell fates [J]. Cell, 2011, 147(4): 881-892.

[4] Wilkinson G, Dennis D, Schuurmans C. Proneural genes in neocortical development[J]. Neuroscience, 2013, 253: 256-273.

[5] Bush A, Hiromi Y, Cole M. Biparous: a novel bHLH gene expressed in neuronal and glial precursors in Drosophila [J]. Dev Biol, 1996, 180(2): 759-772.

[6] Gautier P, Ledent V, Massaer M, et al. Tap, a drosophila bHLH gene expressed in chemosensory organs[J]. Gene, 1997, 191(1): 15-21.

[7] Ledent V, Gaillard F, Gautier P, et al. Expression and function of Tap in the gustatory and olfactory organs of Drosophila[J]. Int J Dev Biol, 1998, 42(2): 163-170.

[8] Petersen TN, Brunak S, Von Heijne G, et al. SignalP 4.0: discriminating signal peptides from transmembrane regions[J]. Nat Methods, 2011, 8(10): 785-786.

[9] Goldberg T, Hamp T, Rost B. LocTree2 predicts localization for all domains of life[J]. Bioinformatics, 2012, 28 (18): 458-465.

[10] Hunter S, Jones P, Mitchell A, et al. InterPro in 2011: new developments in the family and domain prediction database[J]. Nucleic Acids Res, 2012, 40(Database issue): D306-312.

[11] Longo A, Guanga GP, Rose RB. Crystal structure of E47 neuroD1/Beta2 bHLH domain DNA complex: Heterodimer selectivity and DNA recognition[J]. Biochemistry, 2008, 47(1): 218-229.

[12] Lacomme M, Liaubet L, Pituello F, et al. NEUROG2 drives cell cycle exit of neuronal precursors by specifically repressing a subset of cyclins acting at the G1 and S phases of the cell cycle[J]. Mol Cell Biol, 2012, 32(13): 2596-2607.

[13] D'amico LA, Boujard D, Coumailleau P. The neurogenic factor NeuroD1 is expressed in post-mitotic cells during juvenile and adult Xenopus neurogenesis and not in progenitor or radial glial cells[J]. PLoS One, 2013, 8(6): e66487.

[14] Cherry TJ, Wang S, Bormuth I, et al. NeuroD factors regulate cell fate and neurite stratification in the developing retina[J]. J Neurosci, 2011, 31(20): 7365-7379.

[15] Kim WY. NeuroD regulates neuronal migration[J]. Mol Cells, 2013, 35(5): 444-449.

（收稿日期：2014-09-15 修回日期：2015-03-20）