绵羊 *NRCAM* 基因的生物信息学分析

靳泽希[1]，冯 芬[1]，邓晓银[1]，王维民[1,2]

（1. 甘肃农业大学，甘肃 兰州 730070；2. 甘肃省肉羊繁育生物技术工程实验室，甘肃 民勤 733300）

摘要：利用生物基因组学数据库，对绵羊神经原相关的细胞粘附分子基因（neuro-related celladhesion molecule，NRCAM）进行生物信息学分析，以初步了解其结构并对其编码的蛋白质的功能进行预测。结果表明，绵羊 *NRCAM* 基因含有 1 个最大长度为 3 648 bp 的开放阅读框，编码 1215 个氨基酸残基。*NRCAM* 基因编码产物的分子质量为 134 367.13 KDa，理论等电点为 5.49。亚细胞定位主要位于细胞质（26.1%），属于分泌蛋白，且存在信号肽序列。存在 5 段 IGc2 区、1 段 IG 区，1 段低复杂性区域以及 4 段 FN3 区，且有 1 段跨膜结构；二级结构以无规卷曲和 β 折叠为主，三级结构主要由无规卷曲和 β 折叠缠绕形成。

关键词：绵羊；*NRCAM*基因；生物信息学分析

中图分类号：S826 **文献标志码**：A **文章编号**：1001-1463(2020)12-0019-06

doi:10.3969/j.issn.1001-1463.2020.12.006

Bioinformatics Analysis of Sheep *NRCAM* Gene

JIN Zexi[1], FENG Fen[1], DENG Xiaoyin[1], WANG Weimin[1,2]

(1. College of Animal Science and Technology, Gansu Agricultural University, Lanzhou Gansu 730070, China; 2. Engineering Laboratory of Mutton Sheep Breeding and Reproduction Biotechnology in Gansu Province, Minqin Gansu 733300, China)

Abstract: The bioinformatics analysis of sheep *NRCAM* gene was carried out by using bioinformatics database and software, and its structure was preliminarily understood and predicted. The results showed that sheep *NRCAM* gene contained an open reading frame of 3 648 bp, encoding 1215 amino acid residues. The molecular mass of *NRCAM* protein was 134 367.13 KDa, and the theoretical isoelectric point PI was 5.49. The subcells were mainly located in cytoplasmic (26.1%) and belong to secretory protein. There are signal peptide sequences. There are five segments of IGc2 regions, a segments of IG region, a segments of low complexity region,four segments of FN3 region and a segments of transmembrane structure. The secondary structure is mainly random crimp and β-pleated sheet, and the tertiary structure is mainly formed by random crimp and β-pleated sheet,winding and folding.

Key words: Sheep; *NRCAM* gene; Bioinformatics

神经原相关的细胞粘附分子（neuro-related celladhesion molecule，NRCAM）是一种跨膜的细胞粘附分子，它有多种亚型，目前已经鉴定出的有 20 多种。NRCAM 属单基因

收稿日期：2020-07-03

基金项目：甘肃农业大学学生科研训练计划项目（202004013）。

作者简介：靳泽希（1999—），男，甘肃张掖人，本科在读，研究方向为动物科学（畜牧兽医方向）。联系电话：（0）17797691813。Email：1178144923@qq.com。

通信作者：王维民（1984—），男，湖北广水人，副教授，主要从事绵羊遗传育种与繁殖工作。联系电话：（0931）7631225。Email： wangwm@gsau.edu.cn。

家族，其不同亚型的形成是由单个 *NRCAM* 基因通过不同的转录、转录后加工、翻译、翻译后加工形成的。NRCAM 属于免疫球蛋白超基因家族[1-2]，它是一种能介导细胞之间及细胞与细胞外基质间相互作用的糖蛋白，在细胞的识别及转移、肿瘤的浸润与生长、神经再生、跨膜信号的传导、学习和记忆等方面均发挥一定的作用。神经细胞粘附分子在组织形成和细胞迁移以及神经突长出中起着重要作用，它还可以通过胞内区与细胞骨架蛋白或第二信使的结合参与信号传导过程。例如，传统的钙粘素通过参与细胞极性建立、细胞增殖、轴突延长和聚集等基本过程，在动物细胞的形态中发挥着重要作用[3]。Zhou WB 等[4]发现，将周围神经植入脑中几天后，丘脑、纹状体部位神经元直接朝向周围神经移植物的方向生长，进入雪旺细胞柱中，而在雪旺细胞和神经元表面均有 NRCAM 的表达，表明 NRCAM 在神经的再生过程中担当着重要角色。Doherty P 等[5]对鸡 *NRCAM* 基因的分析发现，*NRCAM* 基因由内含子和 26 个外显子组成，这 26 个外显子的结构在不同的物种和属之间是相当恒定的，但内含子是不同的。目前，人、家鼠、牛、狗、猪、绵羊、鸡、兔子等动物的 *NRCAM* 基因序列均已经公布，但对其结构和功能的研究有待进一步研究。我们以生物基因组数据库调取的绵羊 NRCAM 的序列为基础，利用生物信息学方法对不同物种 *NRCAM* 基因及其编码蛋白的理化性质、二级结构及多参数预测、蛋白质跨膜结构、信号肽预测、亚细胞定位和三级结构等进行了分析，以期为深入研究 *NRCAM* 基因及其编码蛋白基本结构和生物学功能提供理论基础。

1 材料与方法

1.1 序列来源

数据来源于 NCBI 网站的 GenBank 数据库[6]，包括绵羊（XM_027968593.1）、牛（NM_ 001206562.1）、人（NM_001193583.1）、家鼠（XM_017594291.1）、猪（XM_021063526.1）、狗（XM_014120801.2）、兔子（XM_008258357.2）和鸡（XM_015280741.2）等 8 个物种的 mRNA 序列。括号内为 GenBank 登录号。

1.2 方法

绵羊 *NRCAM* 基因开放阅读框（Open reading frame，ORF）采用 NCBI 的 ORF Finder 程序分析，参照 Kozak 法则；NRCAM 编码产物的理化性质采用 Bioedit 及 ExPASy 分析软件预测[7]；亚细胞定位采用 PSORT Ⅱ 预测[8-9]；蛋白潜在信号肽剪切位点预测采用 Signalp 3.0 软件；跨膜螺旋区域的预测采用 TMHMM 程序；蛋白保守结构域分析采用 Smart 软件。采用 ProtScale 进行蛋白亲疏水性分析。二级结构采用 Jpred 分析预测。采用 Swiss-model 软件分析蛋白三级结构多序列比对，同源性分析采用 DNAMAN 软件。

2 结果与分析

2.1 绵羊 *NRCAM* 基因开放阅读框分析

通过图 1 可以看出，绵羊 *NRCAM* 基因序列中有 1 条最大长度为 3 648 bp 的 ORF，起始密码子位于 405 bp 处，终止密码子位于 4 052 bp 处，推测编码 1 215 个氨基酸残基。

2.2 绵羊 *NRCAM* 基因编码产物的理化性质分析

蛋白质的基本性质包括其相对分子质量、氨基酸组成和等电点等[10]。对绵羊 *NRCAM* 基因编码产物理化性质的分析表明，绵羊 *NRCAM* 基因编码 1 215 个氨基酸残基，其分子式为 $C_{5995}H_{9389}N_{1613}O_{1830}S_{32}$，分子质量为 134 367.13 KDa，理论等电点 pI 为 5.49。其氨基酸组成如图 2 所示，其中含量最多的氨基酸是 Leu（亮氨酸），所占比例为 8.3%；含量最少的氨基酸是 Cys（半胱氨酸），所占比例 1.2%。负电荷残基总数（Asp + Glu）为 148，正电荷残基总数（Arg + Lys）为 121。

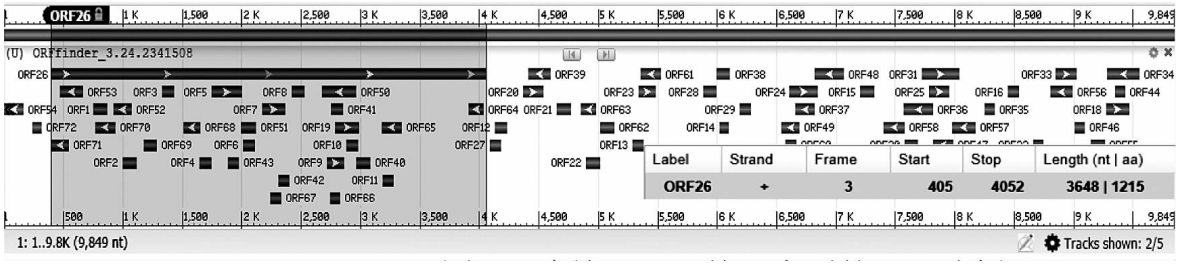


图 1 绵羊 *NRCAM* 基因序列的 ORF 分析

基因编码产物半衰期为 30 h，不稳定指数为 40.22，不稳定指数为 40.22>40.00，可确定该基因编码产物属不稳定蛋白。

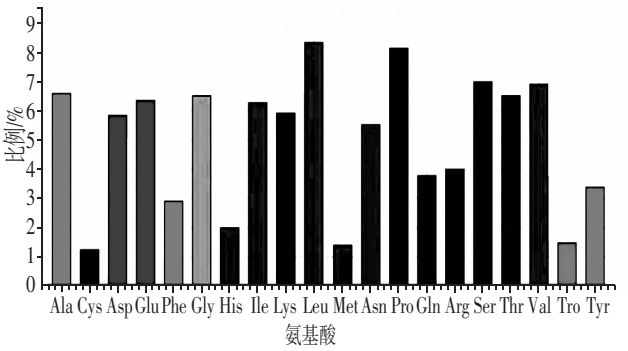


图 2 绵羊 *NRCAM* 基因编码产物氨基酸组成

2.3 绵羊 *NRCAM* 基因编码产物亚细胞定位分析

绵羊 *NRCAM* 基因对蛋白亚细胞的定位结果见表 1。可以看出，绵羊 NRCAM 蛋白的亚细胞分布于细胞质的可能性为 26.1%，分布于细胞核的可能性为 17.4%，分布于囊泡分泌系统、线粒体的可能性均为 13.0%，分布于高尔基体、内质网的可能性均为 8.7%，分布于细胞骨架、细胞外及细胞壁、质膜的可能性均为 4.3%。由此推断，绵羊 *NRCAM* 基因的编码产物主要在细胞质中发挥生物学作用。

表 1 NRCAM 编码产物的亚细胞定位预测结果

亚细胞定位	概率/%
细胞质	26.1
细胞核	17.4
线粒体	13.0
囊泡的分泌系统	13.0
高尔基体	8.7
内质网	8.7
细胞骨架	4.3
细胞外及细胞壁	4.3
质膜	4.3

2.4 绵羊 *NRCAM* 基因编码产物的同源性分析

从图 3、图4可以看出，NRCAM 在很多物种中都有表达，且绵羊与牛的 NRCAM 氨基酸序列同源性较高，这也说明它们在进化过程具有较近的亲缘关系。*NRCAM* 基因编码产物同源树证明，该基因的编码产物在绵羊和牛上的同源性最高，达 99%。

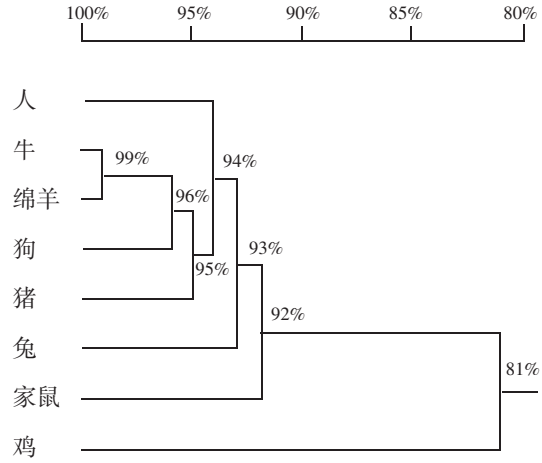
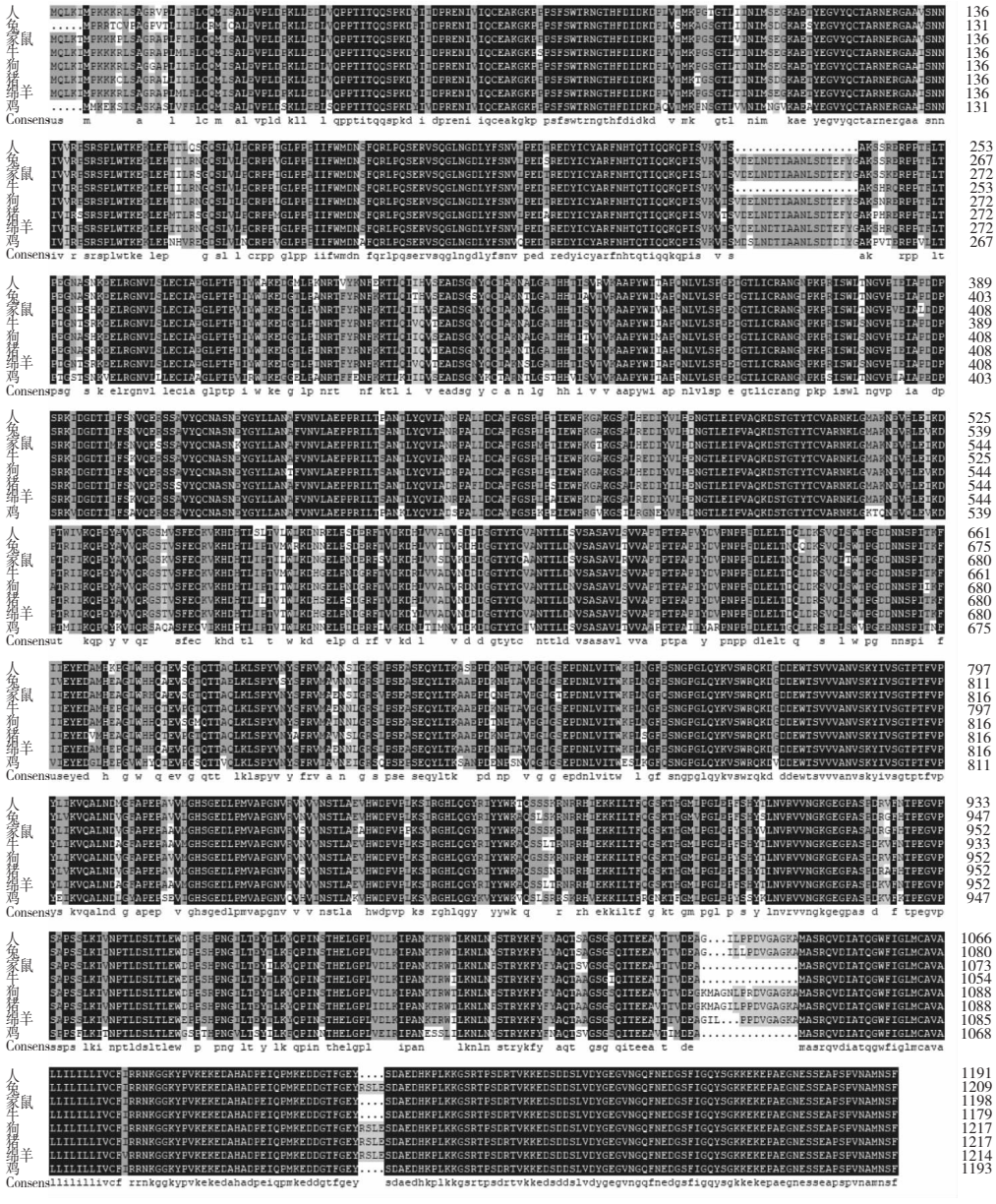


图 3 8 个物种的 *NRCAM* 基因编码产物的同源树

2.5 绵羊 *NRCAM* 基因编码产物潜在信号肽剪切位点预测

信号肽序列是存在于分泌蛋白基因编码序列中、在起始密码子之后的 1 段富含疏水氨基酸多肽的序列。通过检测绵羊 NRCAM 蛋白潜在信号肽的存在情况可判断该基因编码的产物是否为分泌蛋白和跨膜蛋白以及跨膜蛋白的基本信息。从图 5 看出，绵羊



（8 个物种包括：绵羊、人、兔、家鼠、牛、狗、猪和鸡）

图 4 8 个物种的 *NRCAM* 基因编码产物序列的同源性分析

NRCAM 基因编码产物的 *C* 值、*Y* 值和 *S* 值分别为 0.474、0.580 和 0.929。推断 *NRCAM* 基因的编码产物包含信号肽，剪切位点位于 29、30 残基处，属于分泌蛋白。

2.6 绵羊 *NRCAM* 基因编码产物跨膜螺旋结构预测

用 TMHMM2.0 软件分析的结果显示，该基因编码的蛋白有 1 段跨膜结构（图6），其中 1 ~ 1 074 位氨基酸在细胞膜外，其余氨基酸在细胞质内。

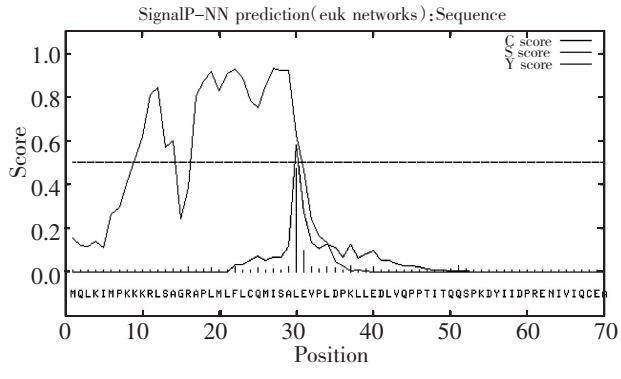


图 5 绵羊 *NRCAM* 基因蛋白潜在信号肽剪切位点分析结果

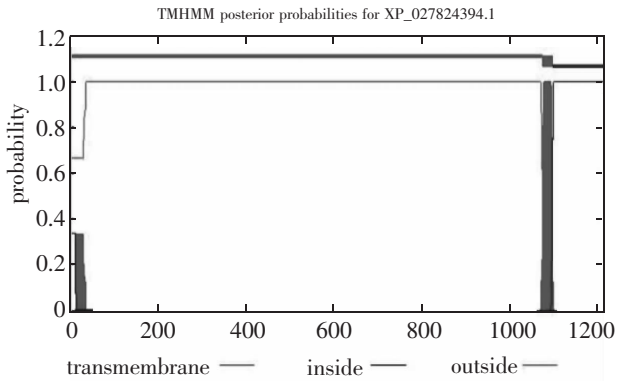


图 6 绵羊 *NRCAM* 基因蛋白跨膜螺旋结构分析结果

2.7 绵羊 *NRCAM* 基因编码产物保守结构域分析

由 Smart 软件分析可知，绵羊 NRCAM 第 59～130 位、第 283～347 位、第 373～439 位、第 467～532 位和第 558～623 位氨基酸残基存在 IGc2 区，第 152～239 位氨基酸残基存在于 IG 区，第 625～635 位氨基酸残基均为低复杂性区域，第 647～730 位、第 747～830 位、第 846～937 位和第 952～1 037 位氨基酸残基存在于 FN3 区，第1075～1 097 位氨基酸残基存在于跨膜区(图 7、表2)。

表 2 绵羊 NRCAM 蛋白保守结构域分析数据

名称	起始位点	终止位点	E值
IGc2	59	130	0.000 144
IG	152	239	0.000 015 9
IGc2	283	347	1.95e−15
IGc2	373	439	3.76e−8
IGc2	467	532	1.26e−9
IGc2	558	623	0.000 001 2
low complexity	625	635	N/A
FN3	647	730	1.07e−10
FN3	747	830	0.00474
FN3	846	937	3.37e−8
FN3	952	1 037	5.56e−9
transmembrane region	1 075	1 097	N/A

2.8 绵羊 *NRCAM* 基因编码产物亲疏水性分析

该基因编码蛋白疏水性最大值为 4.078(1 090位)，最小值为 −3.022(790～791位)，图形的高峰值(正值)区域表示疏水的区域，而负值的“低谷”区域是亲水区域。整条链中亲水性氨基酸残基多于疏水性氨基酸残基。因此可推测该基因编码的蛋白是亲水性蛋白(图8)。

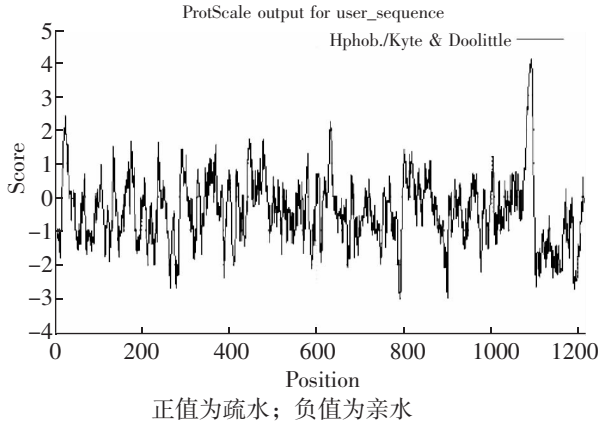


图 8 绵羊 *NRCAM* 基因编码蛋白质疏水性/亲水性预测分析

2.9 绵羊 *NRCAM* 基因编码产物二级结构的预测

通过 Jpred 软件分析可知（图9），绵羊 NRCAM 蛋白二级结构如下：α 螺旋(Hh)、β 折叠(Ee)、无规卷曲(Cc)分别占 2.96%、42.13%、54.89%。可以看出 NRCAM 基因编码的二级结构中无规卷曲占主导地位，其次是 β 折叠。

2.10 绵羊 *NRCAM* 基因编码产物三级结构预测与分析

通过在线工具 Swiss-model 对绵羊 NRCAM 蛋白三级结构的预测和分析（图10)可知，*NRCAM* 基因编码蛋白的三级结构与二级结构预测的结果一致，主要由无规卷曲和

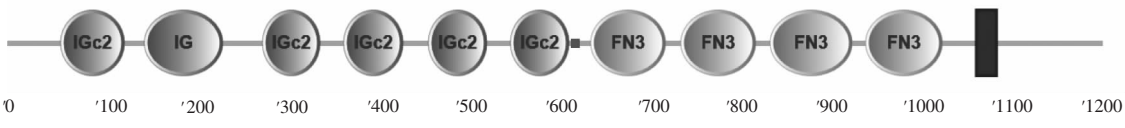
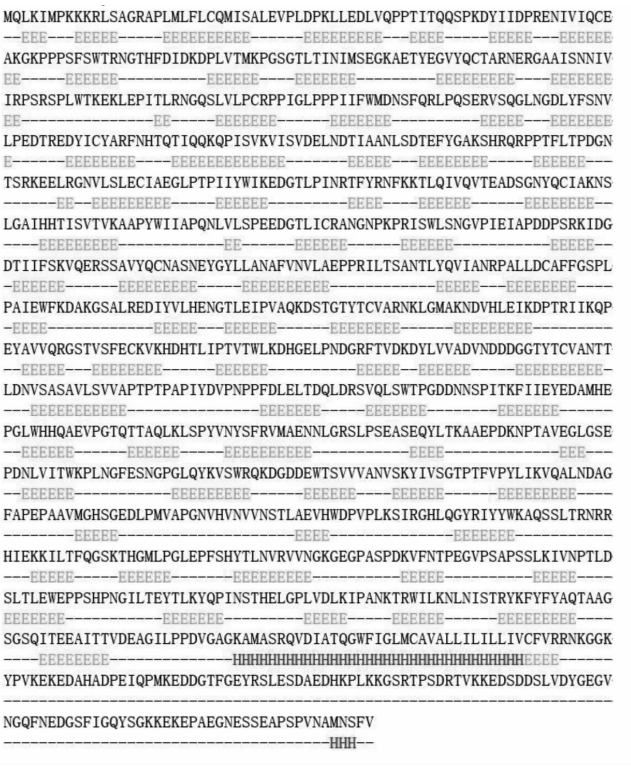


图 7 绵羊 NRCAM 蛋白保守结构域



（–表示无规卷曲，H 表示 a 螺旋，E 表示 β 折叠）

图 9 绵羊 *NRCAM* 基因编码蛋白质二级结构预测

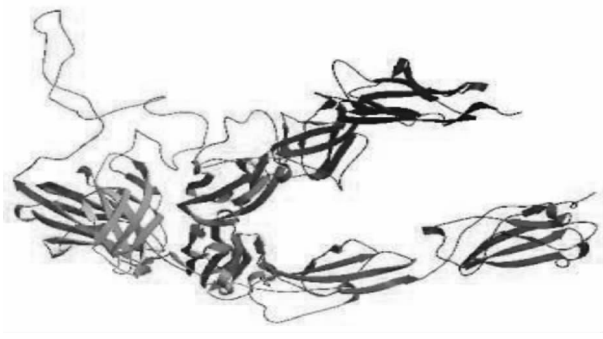


图 10 绵羊 NRCAM 蛋白三级结构的分析结果

β 折叠缠绕形成。

3 结论

绵羊 *NRCAM* 基因含有 1 个最大长度为 3 648 bp 的 ORF，编码 1 215 个氨基酸残基；亮氨酸所占比例最多，为 8.3%，分子质量为 134 367.13 KDa，理论等电点 pI 为 5.49。NRCAM 编码的产物为不稳定性蛋白。NRCAM 蛋白的亚细胞定位在细胞质的可能性最大，为 26.1%。*NRCAM* 基因在很多物种中都有表达，绵羊和牛在同源树中同源性达到 99%。NRCAM 基因的编码产物中包含信号肽，该蛋白是分泌蛋白。该基因编码的蛋白有 1 段跨膜结构。*NRCAM* 基因编码的蛋白为亲水性蛋白，亲水性氨基酸残基多于疏水性氨基酸残基。绵羊 *NRCAM* 基因编码产物的二级结构主要以无规卷曲和 β 折叠为主，三级结构主要由无规卷曲和 β 折叠缠绕形成。

参考文献：

[1] 史金阳，KRISSANSEN G W，夏 畅，等. 融合蛋白 NrCAM–Fc 的重组体的构建[J]. 中国医科大学学报，2002，31(1)：15–16.

[2] BRUMMENDORF T，RATHJEN F G. Cell adhesion molecules.1. Immunoglobulin superfamily[J]. Protein Profile，1995(2)：963–1108.

[3] 胡晓燕，周 严，袁建刚，等. 神经细胞粘附分子的研究进展[J]. 生命科学，2001(5)：200–204.

[4] ZHOU W B，ZHOU C F. Role of neural cell adhesion molecule and polysialic acid on the neuronal development and regeneration [J]. Progress in Physiological Sciences，1996，27(2)：118–122.

[5] DOHERTY P，FAZCLI M S，WALSE F S. The neural cell adhesion molecule and synaptic plasticity[J]. J. Neurobiology，1995；26：437–446.

[6] 张小雪，潘香羽，李发弟，等. 绵羊 *ESR* 基因生物信息学分析[J]. 甘肃农业科技，2014(9)：30–33.

[7] 罗 轶. 鸡 *FATP1* 基因 cDNA 的克隆、组织表达及其生物信息学分析[D]. 雅安：四川农业大学，2008.

[8] 张小雪，李发弟，王维民. 绵羊 *ANXA10* 基因生物信息学分析[J]. 甘肃农业科技，2016(6)：1–4.

[9] NAKAI K，HORTON P. PSORT：a program for detecting sorting signals in proteins and predicting their subcellular localization[J]. Trends Biochem. Sci.，1999(24)：34–36.

[10] 张小雪，李发弟，王维民. 绵羊 *STMN2* 基因生物信息学分析[J]. 甘肃农业科技，2016(7)：58–61.

（本文责编：陈 伟）