人类、小鼠、大鼠和斑马鱼中 Neurogenin家族的分析

梁巍 夏桔丹 徐琳杰 杨勇

2008-06-16

概要

- 背景
- 系统进化树的构建
- 外显子-内含子结构的比较分析
- NGN结构域（motif）分析
- bHLH结构域的序列分析
- bHLH结构域的结构分析
- 斑马鱼与小鼠NGN3中bHLH结构域的比较
- 对斑马鱼中NGN两个亚家族的探索
- 结论

背景

- 神经原质蛋白（神经元素）Neurogenin是一类在神经细胞前体中表达，控制神经细胞前体发育成神经细胞的转录因子。并且，Neurogenin还是bHLH蛋白中的一类。

- 目前一般认为早表达的bHLH蛋白控制细胞命运的决定，晚表达的bHLH蛋白则控制细胞分化。Neurogenin是一类早表达的bHLH蛋白，而Neurod则是晚表达的bHLH蛋白。

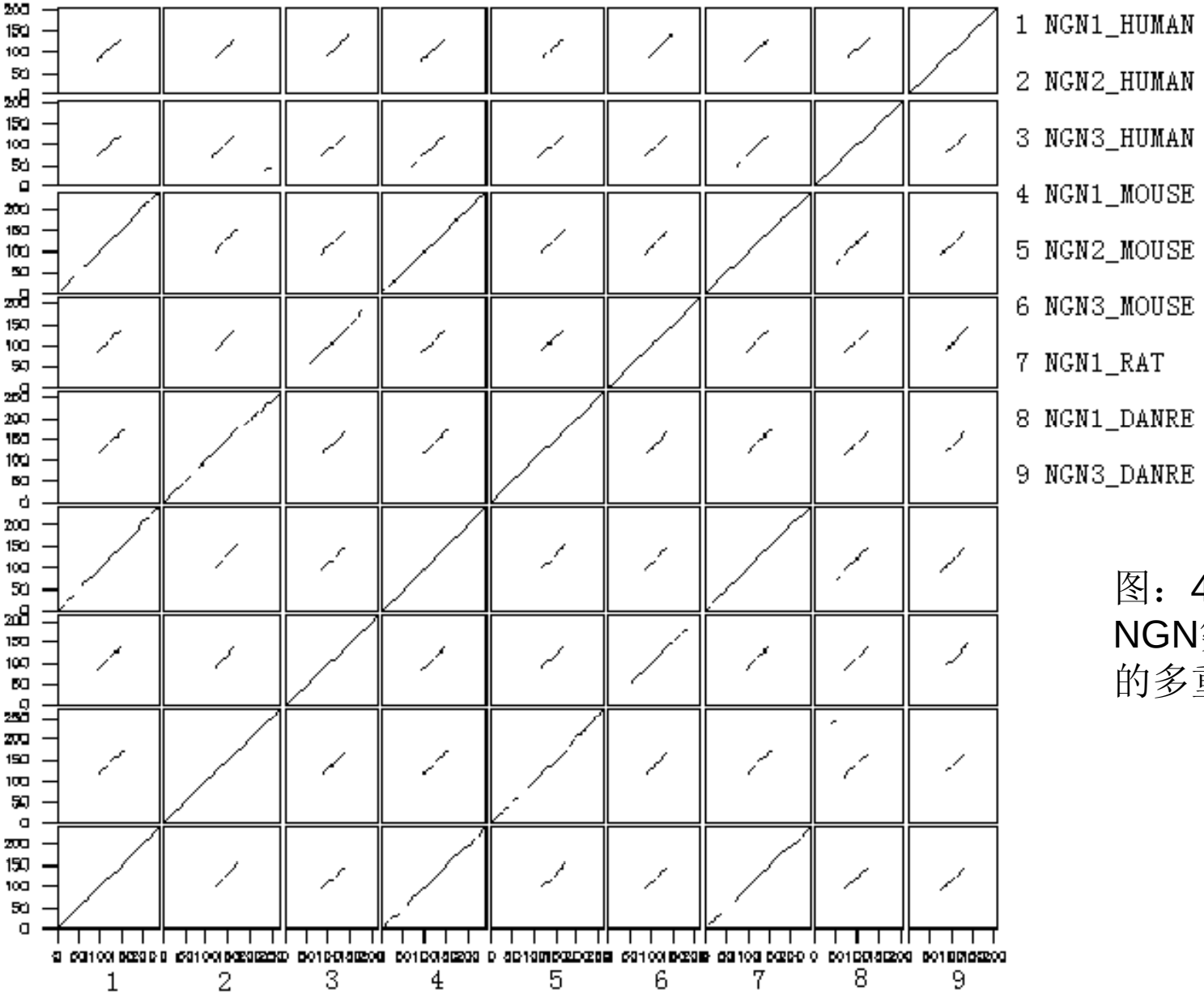
- 人类、小鼠和斑马鱼中的NGN家族不完全一致。在斑马鱼中没有人和小鼠共有的亚家族NGN2。同时，亚家族NGN3的表达图式有所不同。在人和小鼠中，NGN3不仅在神经中表达，同时也作为胰腺内分泌腺的标记基因，在胰腺细胞的分化和命运决定中起重要作用；而斑马鱼的NGN3仅在神经中表达。由于斑马鱼基因组中有重复现象，我们猜测在斑马鱼中存在另一个NGN3的拷贝，它和目前已知的NGN3分别在胰腺和神经中行使功能。

数据的取得

Organism	Gene Name	Ensembl Gene ID	Protein Name	Uniprot Protein ID	Protein Seq Long	DNA binding site	HLH motif
zebrafish	neurog1	ENSDARG00000056130	Neurogenin-1	O42606	208	71-82	83-123
	neurog3	ENSDARG00000016951	Neurogenin-3	Q9DG56	208		
human	NEUROG1	ENSG00000181965	Neurogenin-1	Q92886	237	93-104	105-145
	NEUROG2	ENSG00000178403	Neurogenin-2	Q9H2A3	272	110-165	122-133
	NEUROG3	ENSG00000122859	Neurogenin-3	Q9Y4Z2	214	84-95	96-136
mouse	Neurog1	ENSMUSG00000048904	Neurogenin-1	P70660	244	94-105	106-146
	Neurog2	ENSMUSG00000027967	Neurogenin-2	P70447	263	113-124	125-165
	Neurog3	ENSMUSG00000044312	Neurogenin-3	P70661	214	84-95	96-136
rat	Neurod1	ENSRNOG00000022405	Neurogenin-1	P70595	244	106-146	94-105

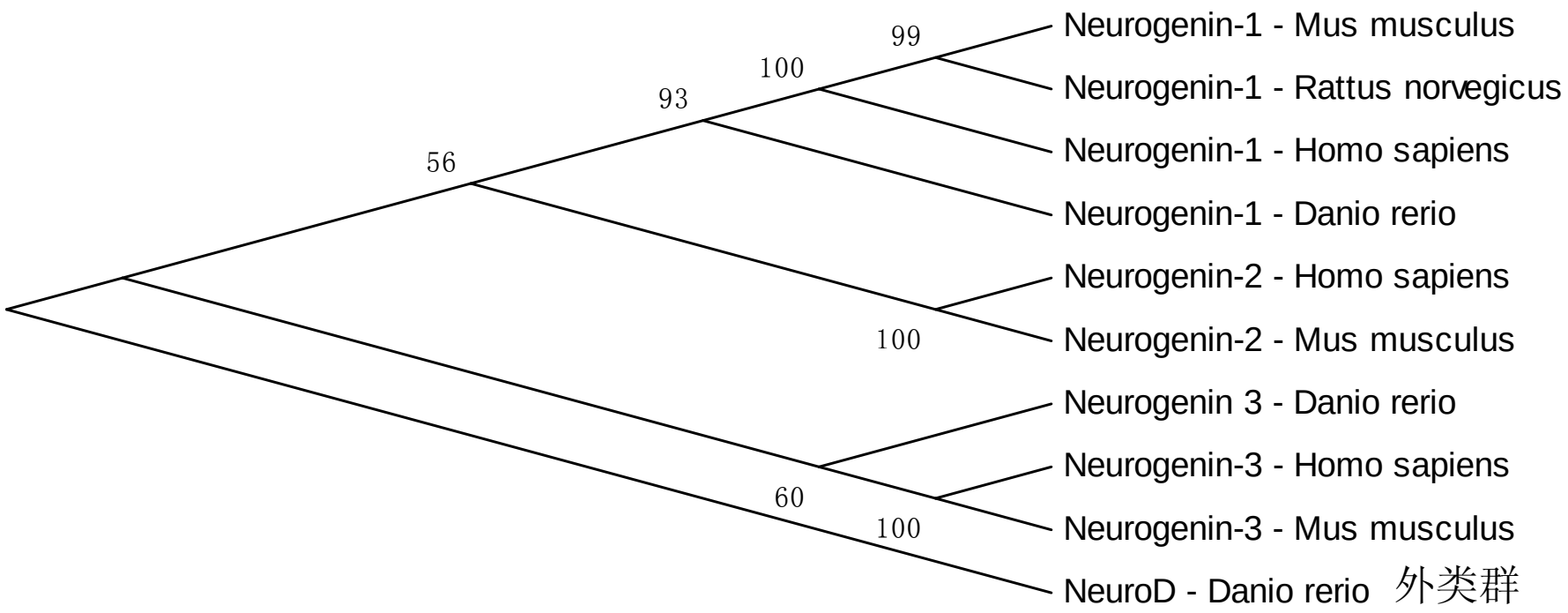
表：4个物种的9个Neurogenin（NGN）的序列相关信息

NGN蛋白序列的初步分析



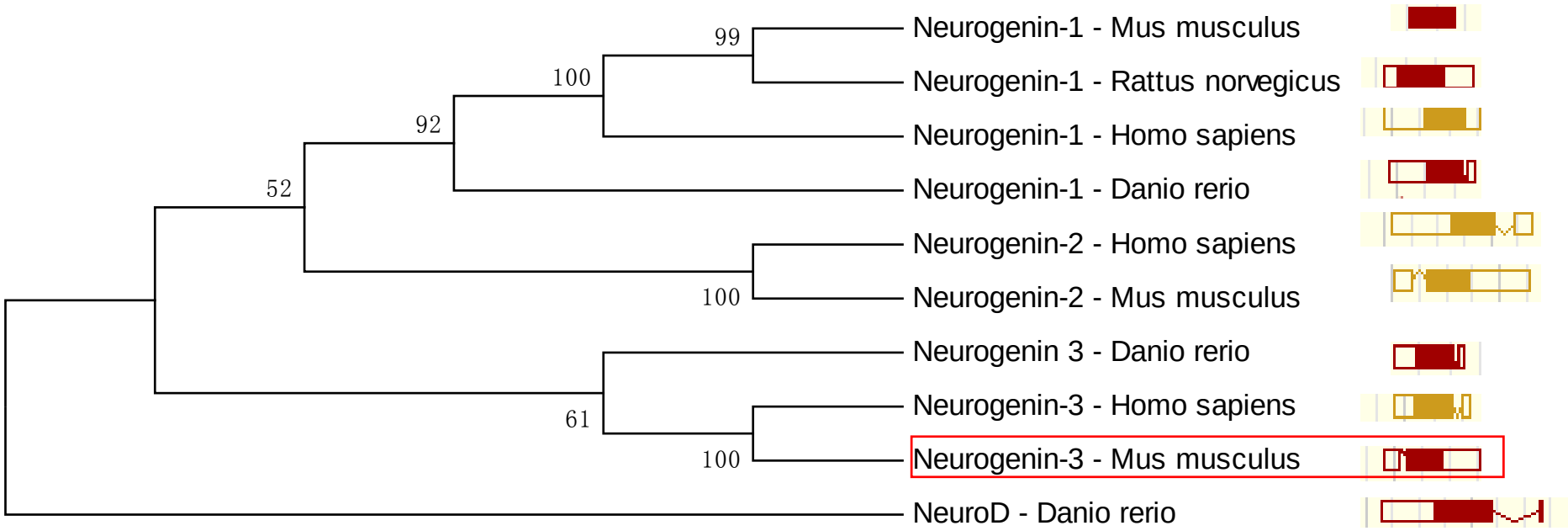
图：4个物种的9个NGN氨基酸序列的多重点阵图

系统进化树的构建



图：Mega 4.0软件N-J算法构建的系统进化树（锥状）

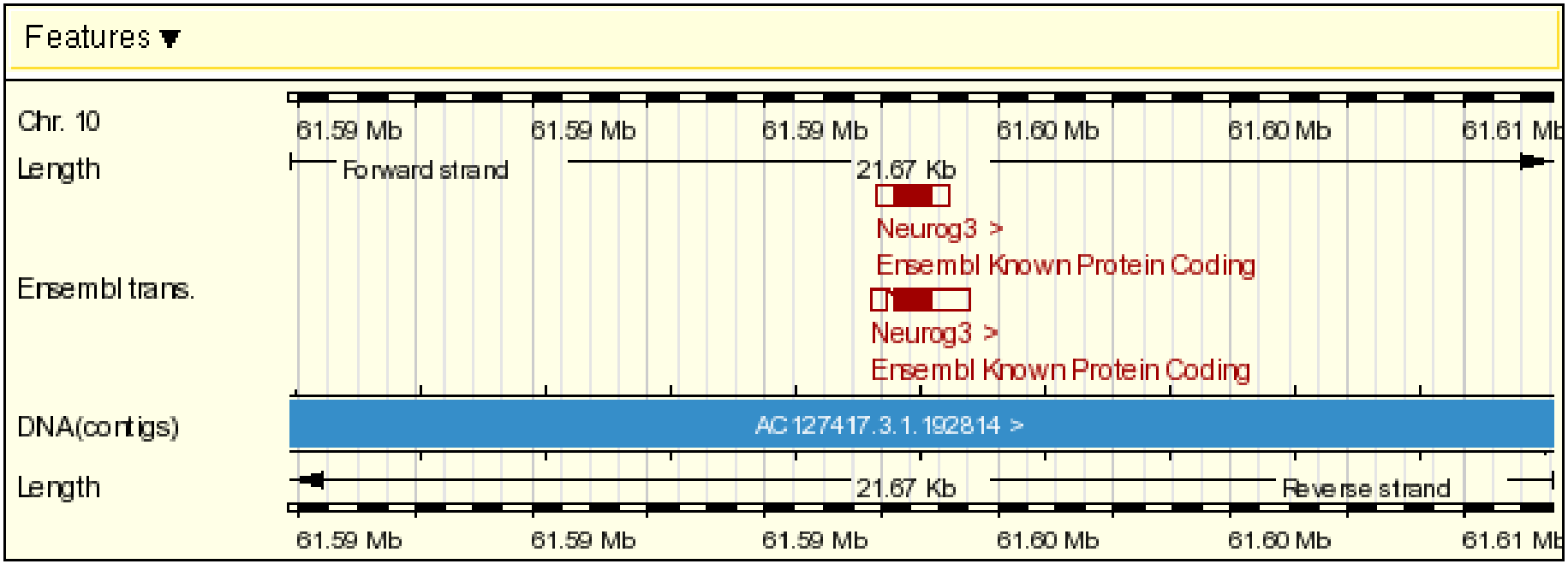
外显子-内含子结构的比较分析



图：9个*ngn*的外显子-内含子结构图

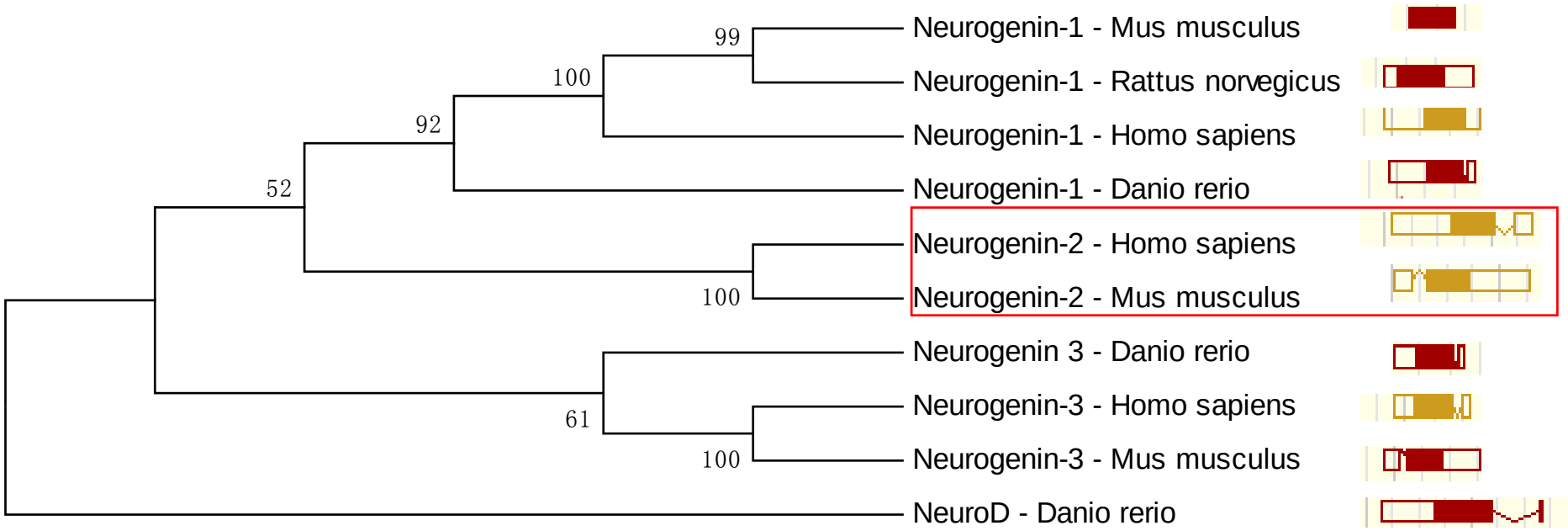
从ensmebl中得到9个*ngn*的外显子-内含子结构图，其中红色是Ensembl已知基因，黄色标记的表明Ensembl和havana录入此基因的转录本一致，基本上是非常确定的基因。有色实体部分为cds序列，白色空白部分为UTR区域，中间的折线部分为内含子。

外显子-内含子结构的比较分析



图：Ensembl数据库中小鼠的2个Neurog3基因差异

外显子-内含子结构的比较分析



图：9个*ngn*的外显子-内含子结构图

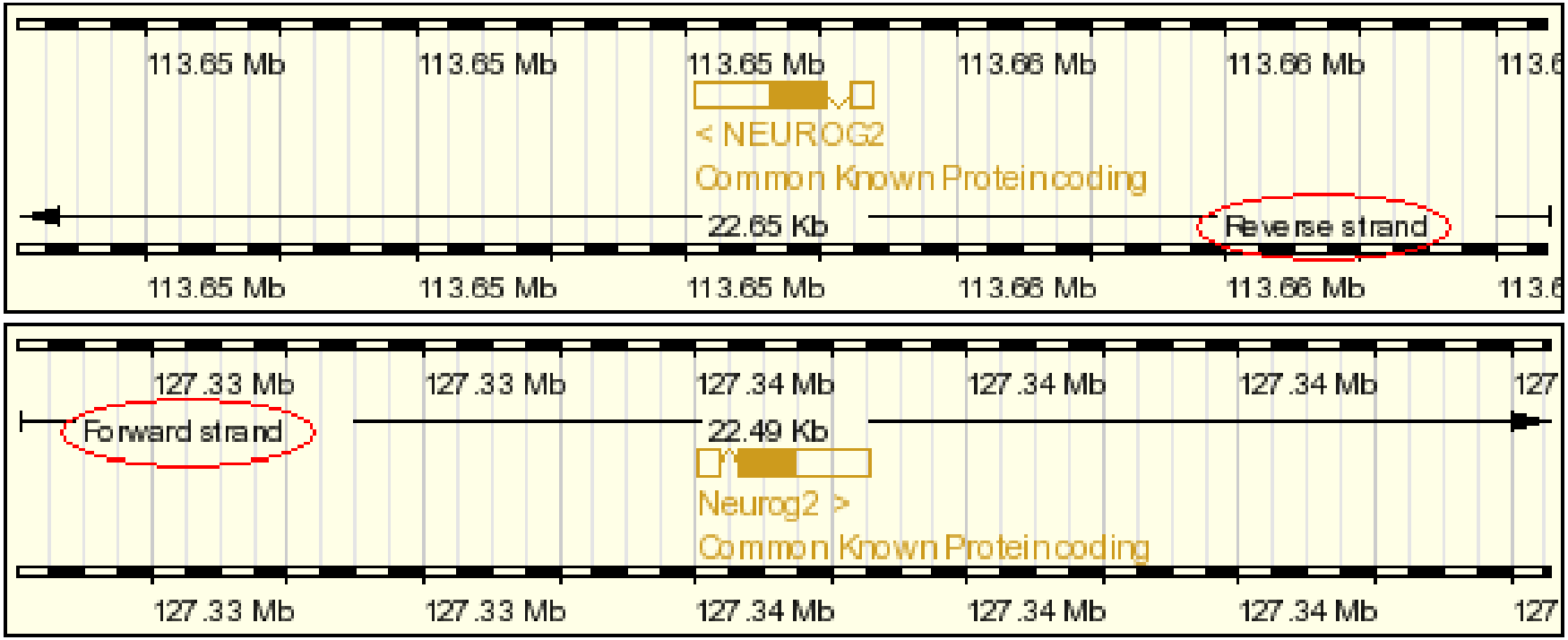
从ensmebl中得到9个*ngn*的外显子-内含子结构图，其中红色是Ensembl已知基因，黄色标记的表明Ensembl和havana录入此基因的转录本一致，基本上是非常确定的基因。有色实体部分为cds序列，白色空白部分为UTR区域，中间的折线部分为内含子。

外显子-内含子结构的比较分析

Pairwise Alignment Result

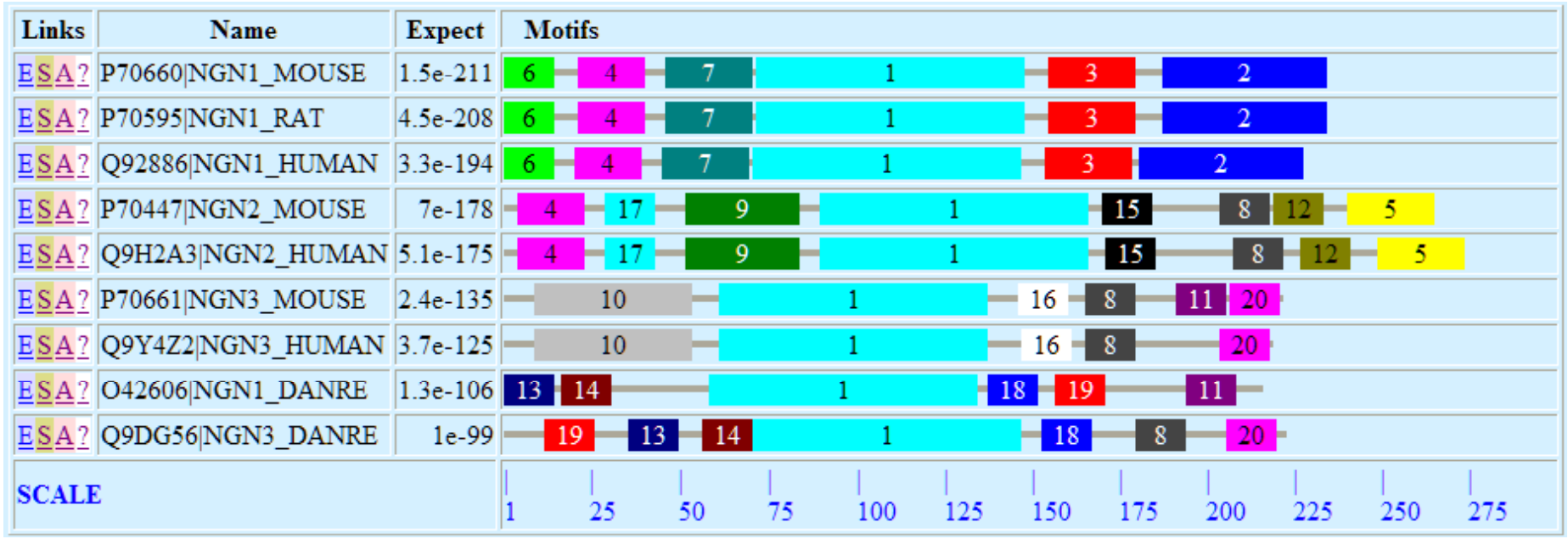
LENGTH	SCORE	IDENTITY	SIMILARITY	GAPS
2749	8174.0	2075/2749 (75.5%)	2075/2749 (75.5%)	355/2749 (12.9%)

图：小鼠*ngn2*基因组序列倒置后与人类的*ngn2*基因组序列进行序列比对的结果



图：人和小鼠*ngn2*基因的转录方向

NGN结构域（motif）分析



图：多重期望序列模体识别系统分析4个物种的9个NGN蛋白质序列保守结构域图形输出

Maximum width：100， minimum width：10，maximum number of motifs：20

NGN结构域（motif）分析

```
BL   MOTIF 1 width=80 seqs=9
P70661|NGN3_MOUSE (65) RGGRNRPKSELALSKQRRSRRKKANDRERNRMHNLNSALDALRGVLPTFPDDAKLTKIETLRFAHNYIWALTQTLRIADH
P70595|NGN1_RAT   (75) RRGRARVRSEALLHSLRRSRRVKANDRERNRMHNLNAALDALRSVLPSFPDDTKLTKIETLRFAYNYIWALAETLRLADQ
P70660|NGN1_MOUSE (75) RRGRARVRSEALLHSLRRSRRVKANDRERNRMHNLNAALDALRSVLPSFPDDTKLTKIETLRFAYNYIWALAETLRLADQ
Q92886|NGN1_HUMAN (74) RRGRTRVRSEALLHSLRRSRRVKANDRERNRMHNLNAALDALRSVLPSFPDDTKLTKIETLRFAYNYIWALAETLRLADQ
Q9Y4Z2|NGN3_HUMAN (65) RGGRSRPKSELALSKQRRSRRKKANDRERNRMHDLNSALDALRGVLPTFPDDAKLTKIETLRFAHNYIWALTQTLRIADH
O42606|NGN1_DANRE (52) KRRRGRARNETTVHVVKKNRRLKANDRERNRMHNLNDALDALRSVLPAFPDDTKLTKIETLRFAHNYIWALSETIRIADQ
P70447|NGN2_MOUSE (94) RAVSRGAKTAETVQRIKKTRRLKANNRERNRMHNLNAALDALREVLPTFPEDAKLTKIETLRFAHNYIWALTETLRLADH
Q9H2A3|NGN2_HUMAN (94) RAVSRGAKTAETVQRIKKTRRLKANNRERNRMHNLNAALDALREVLPTFPEDAKLTKIETLRFAHNYIWALTETLRLADH
Q9DG56|NGN3_DANRE (60) KTSNGKLKKLMSTSRQRGNRRVKANDRGRHRMHNLNSALDNLRSVLPTFPDDAKLTKIETLRFARNYIWALSETLRIADH
```

图：多重期望序列模体识别系统分析4个物种的9个NGN蛋白质序列的第一个保守结构域（mofit1）序列信息

Organism	Protein name	DNA binding: Basic motif	Domain: HLH motif	bHLH	bHLH length
zebrafish	Neurogenin-1	71-82	83-123	71-123	53
	Neurogenin-3			79-131	53
human	Neurogenin-1	93-104	105-145	93-145	53
	Neurogenin-2	110-165	122-133	110-165	56
	Neurogenin-3	84-95	96-136	84-136	53
mouse	Neurogenin-1	94-105	106-146	94-146	53
	Neurogenin-2	113-124	125-165	113-165	53
	Neurogenin-3	84-95	96-136	84-136	53
rat	Neurogenin-1	106-146	94-105	94-146	53

表：Uniprot数据库中的bHLH相关数据

NGN结构域（motif）分析

- 保守结构域1（浅蓝色）长为80aa，其结构域在9个NGN的位置和长度都与bHLH类似，估计为MEME预测的bHLH结构域。但是Pfam中的bHLH结构域长度大约都为53aa，位置也略有差异。原因应该是MEME程序本身问题。

- 只有保守结构域1为9个NGN共有，说明NGN都含有一个bHLH。从其他的保守结构域可以看出9个NGN的进化关系：人、小鼠、大鼠的NGN1有6个相同的保守结构域；人和小鼠的NGN2也共享了8个相同的保守结构域；人和小鼠的NGN3也有5个保守结构域相同；人、小鼠和大鼠都是羊膜动物，而斑马鱼是非羊膜动物，二者进化关系较远，其NGN1和NGN3物种间差异较明显。但根据预测，斑马鱼NGN1和NGN3有5个相同的保守结构域，说明斑马鱼NGN1和NGN3进化关系相比于其他羊膜动物的NGN1、NGN3更近一些，这与系统进化树的结果不符。原因可能是由于最大预测的保守结构域的数目为20，长度最小为10，程序灵敏度较低，使得一些不是保守结构域的序列被预测出来，斑马鱼NGN1和NGN3的那些除1外的被预测的保守结构域才会相同，这不能完全反映进化关系。

- 同时，我们也用SMART对9个NGN的结构域进行了预测。结果表明，SMART预测准确度远高于MEME。但是，SMART只能对单个蛋白的结构域进行预测，而MEME可以预测多个蛋白的保守结构域，两个软件各有用处。

NGN结构域（motif）分析

Organism	protein name	motif name	motif begin to end	motif length	E-value
zebrafish	Neurogenin-1	bHLH	76-128	53	2.44E-18
	Neurogenin-3	bHLH	84-136	53	1.15E-14
human	Neurogenin-1	bHLH	98-150	53	4.49E-17
	Neurogenin-2	bHLH	118-170	53	1.38E-17
	Neurogenin-3	bHLH	89-141	53	1.38E-17
mouse	Neurogenin-1	bHLH	99-151	53	4.49E-17
	Neurogenin-2	bHLH	118-170	53	1.38E-17
	Neurogenin-3	bHLH	89-141	53	7.94E-18
rat	Neurogenin-1	bHLH	99-151	53	4.49E-17

表：用SMART分析9个NGN的结构域

bHLH结构域的序列分析

```
CLUSTAL FORMAT for T-COFFEE Version_5.31 [http://www.tcoffee.org],
                          CPU=0.91 sec, SCORE=82, Nseq=9, Len=56

NGN1_bHLH_DANRE  ---RRLKANDRERNRMHNLNDALDALRSVLPAFPDDTKLTKIETLRFAHNYIWALS
NGN1_bHLH_HUMAN  ---RRVKANDRERNRMHNLNAALDALRSVLPSFPDDTKLTKIETLRFAYNYIWALA
NGN1_bHLH_MOUSE  ---RRVKANDRERNRMHNLNAALDALRSVLPSFPDDTKLTKIETLRFAYNYIWALA
NGN1_bHLH_RAT    ---RRVKANDRERNRMHNLNAALDALRSVLPSFPDDTKLTKIETLRFAYNYIWALA
NGN2_bHLH_HUMAN  KKTRRLKANNRERNRMHNLNAALDALREVLPTFPEDAKLTKIETLRFAHNYIWALT
NGN2_bHLH_MOUSE  ---RRLKANNRERNRMHNLNAALDALREVLPTFPEDAKLTKIETLRFAHNYIWALT
NGN3_bHLH_MOUSE  ---RRKKANDRERNRMHNLNSALDALRGVLPTFPDDAKLTKIETLRFAHNYIWALT
NGN3_bHLH_HUMAN  ---RRKKANDRERNRMHDLNSALDALRGVLPTFPDDAKLTKIETLRFAHNYIWALT
NGN3_bHLH_DANRE  ---RRVKANDRGRHRMHNLNSALDNLRSVLPTFPDDAKLTKIETLRFARNYIWALS
                    ** ***:* *:***:** *** ** ***:**:*:*********** ******:

 CLUSTAL FORMAT for T-COFFEE Version_5.31 [http://www.tcoffee.org],
                            CPU=0.92 sec, SCORE=83, Nseq=9, Len=53
NGN1_bHLH_DANRE  RRLKANDRERNRMHNLNDALDALRSVLPAFPDDTKLTKIETLRFAHNYIWALS
NGN1_bHLH_HUMAN  RRVKANDRERNRMHNLNAALDALRSVLPSFPDDTKLTKIETLRFAYNYIWALA
NGN1_bHLH_MOUSE  RRVKANDRERNRMHNLNAALDALRSVLPSFPDDTKLTKIETLRFAYNYIWALA
NGN1_bHLH_RAT    RRVKANDRERNRMHNLNAALDALRSVLPSFPDDTKLTKIETLRFAYNYIWALA
NGN2_bHLH_HUMAN  RRLKANNRERNRMHNLNAALDALREVLPTFPEDAKLTKIETLRFAHNYIWALT
NGN2_bHLH_MOUSE  RRLKANNRERNRMHNLNAALDALREVLPTFPEDAKLTKIETLRFAHNYIWALT
NGN3_bHLH_MOUSE  RRKKANDRERNRMHNLNSALDALRGVLPTFPDDAKLTKIETLRFAHNYIWALT
NGN3_bHLH_HUMAN  RRKKANDRERNRMHDLNSALDALRGVLPTFPDDAKLTKIETLRFAHNYIWALT
NGN3_bHLH_DANRE  RRVKANDRGRHRMHNLNSALDNLRSVLPTFPDDAKLTKIETLRFARNYIWALS
                  ** ***:* *:***:** *** ** ***:**:*:*********** ******:
```

图：T-COFFEE对9个NGN蛋白中的bHLH结构域进行多序列比对结果

bHLH结构域的序列分析

图：9个NGN蛋白中的bHLH结构域的序列图标

从序列图标中可以很直观的看出9个NGN蛋白中的bHLH结构域在人、小鼠、大鼠和斑马鱼中十分保守。这也从侧面说明其功能的重要性。

bHLH结构域的结构分析

- bHLH结构域在NGN蛋白中有重要的功能。作为NGN蛋白的DNA结合位点，它的结构很有特点。下面我们对NGN蛋白中的bHLH结构域的三维结构作一些分析。

- PDB数据库中没有NGN的结构数据，无法直接做出分析。通过搜索，我们发现PDB数据库中有一个小鼠中MyoD的bHLH的结构数据（1mdy），这是数据库中仅有的一个bHLH的结构数据。因此我们想通过小鼠中MyoD的bHLH的结构来预测斑马鱼NGN1和NGN3中bHLH结构域的结构。

bHLH结构域的结构分析

Pairwise Alignment Result

LENGTH	SCORE	IDENTITY	SIMILARITY	GAPS
54	78.0	20/54 (37.0%)	29/54 (53.7%)	2/54 (3.7%)

```
MYOD1_Bhlh_MO  1 DRRKAATMRERRRLSKVNEAFETLKRCTSSNPNQ-RLPKVEILRNAIRYIEGLQ 53
                 ||..|..|||.|:..:|:|.:.|:.....:.|:. :|.|:|.||.|..||..|.
NGN1_bHLH_DAN  1 -RRLKANDRERNRMHNLNDALDALRSVLPAFPDDTKLTKIETLRFAHNYIWALS 53
```

图：小鼠MyoD的bHLH序列与斑马鱼NGN1的bHLH序列比对结果

Pairwise Alignment Result

LENGTH	SCORE	IDENTITY	SIMILARITY	GAPS
54	69.0	19/54 (35.2%)	27/54 (50.0%)	2/54 (3.7%)

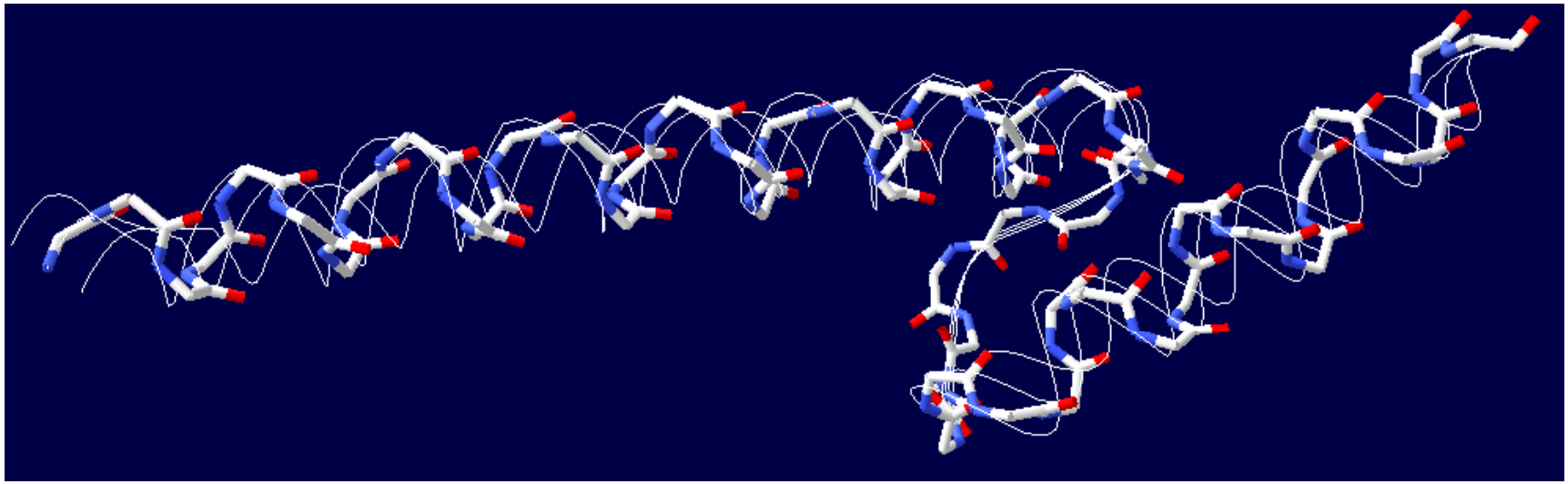
```
MYOD1_Bhlh_MO  1 DRRKAATMRERRRLSKVNEAFETLKRCTSSNPNQ-RLPKVEILRNAIRYIEGLQ 53
                 ||..|..|.|.|:..:|.|.:.|:.....:.|:. :|.|:|.||.|..||..|.
NGN3_bHLH_DAN  1 -RRVKANDRGRHRMHNLNSALDNLRSVLPTFPDDAKLTKIETLRFARNYIWALS 53
```

图：小鼠MyoD的bHLH序列与斑马鱼NGN3的bHLH序列比对结果

bHLH结构域的结构分析

	DNA binding site	helix	loop	helix
MyoD1_mouse	1-12	2-28	29-37	38-53
NGN1_danre	1-13	1-28	29-37	38-53
NGN3_danre	1-13	1-28	29-37	38-53

表：斑马鱼中NGN1、NGN3和小鼠MyoD中bHLH的结构预测



图：通过Swiss-PDBviewer展示的小鼠MyoD中bHLH的基本结构

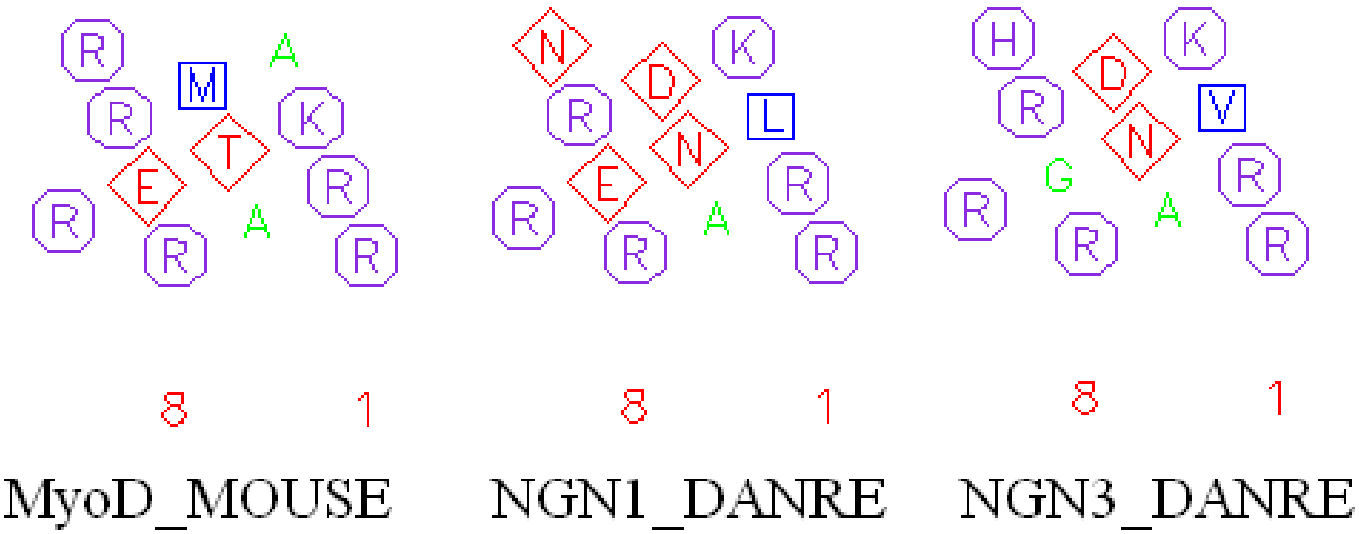
bHLH结构域的结构分析

- 下面对小鼠MyoD和斑马鱼NGN1、NGN3中bHLH的DNA结合位点进行分析。

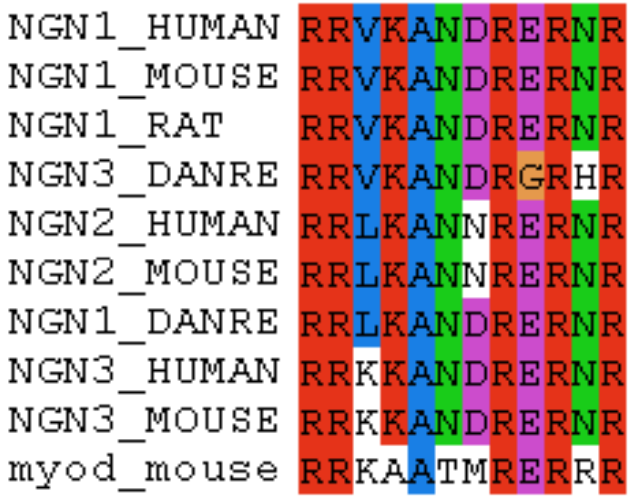
- 我们发现带正电的碱性氨基酸残基非常多（R、K），尤其是精氨酸（R），每个序列至少有5个。由于DNA分子带负电，因此bHLH中DNA结合位点有许多带正电的碱性氨基酸残基，这些以精氨酸为主的带正电的碱性氨基酸残基在bHLH与DNA结合过程中起到很重要的作用。这也从另一个方面说明了斑马鱼NGN1和NGN3的bHLH的结构预测还是比较可信的。

- 用clustal w对9个NGN和MyoD中bHLH的DNA结合位点作多序列比对。可以更清楚地看出精氨酸（R，红色）的保守性。这进一步说明精氨酸在bHLH与DNA结合中的作用。精氨酸是带正电的碱性氨基酸，而DNA是带负电的，又有磷酸基团，因此偏酸性。这样一来，bHLH与DNA就可以通过正负电的相互作用结合在一起。

bHLH结构域的结构分析



图：pepnet软件展示的小鼠MyoD和斑马鱼NGN1、NGN3中bHLH的DNA结合位点螺旋结构图



图：9个NGN和MyoD中bHLH的DNA结合位点作多序列比对

bHLH结构域的结构分析

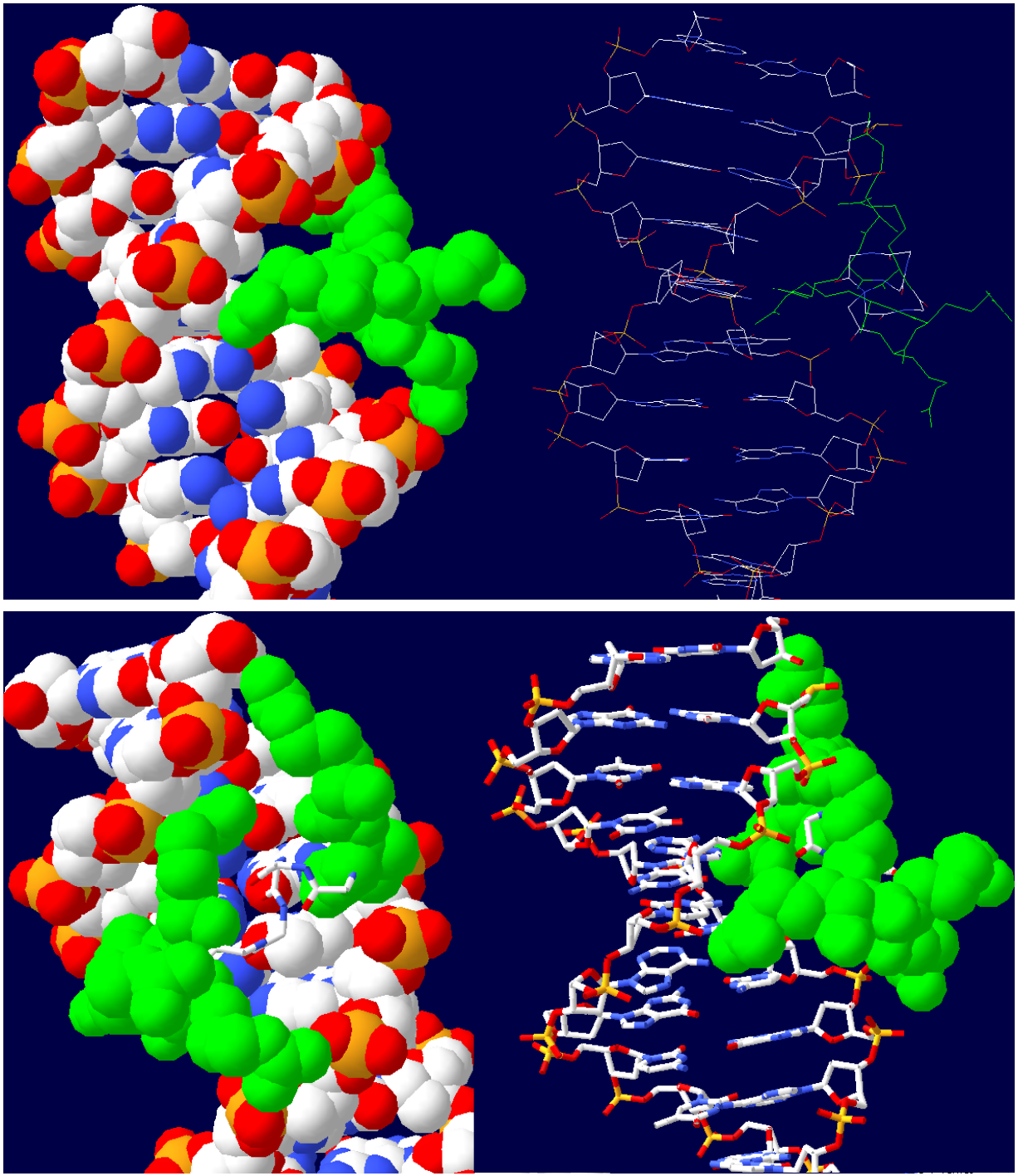
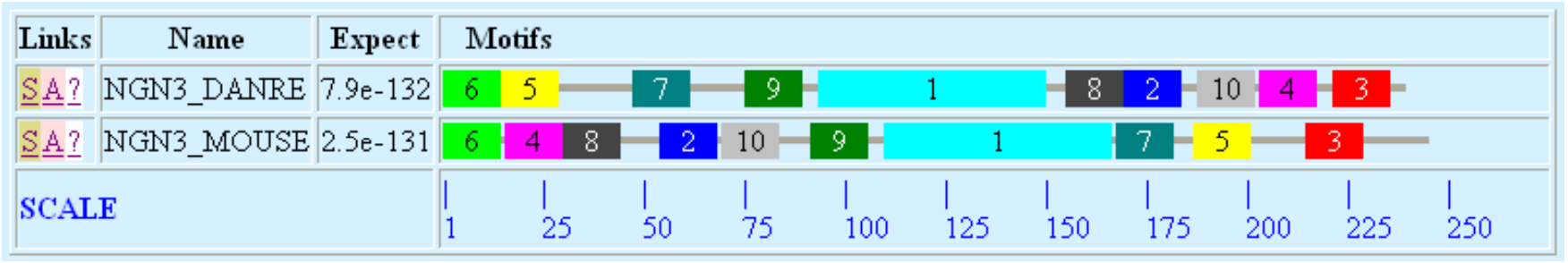


图20：Swiss-PDBviewer展示bHLH与DNA结合结构图

其中绿色的为精氨酸残基

斑马鱼与小鼠NGN3中bHLH结构域的比较



图：MEME对斑马鱼与小鼠NGN3的保守结构域的预测结果

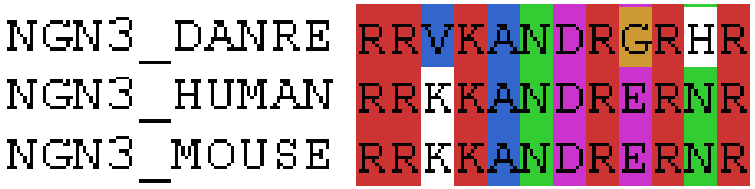
Maximum width：60， minimum width：10，maximum number of motifs：10

Pairwise Alignment Result

LENGTH	SCORE	IDENTITY	SIMILARITY	GAPS
53	235.0	46/53 (86.8%)	48/53 (90.6%)	0/53 (0.0%)

```
NGN3_bHLH_DAN  1 RRVKANDRGRHRMHNLNSALDNLRSVLPTFPDDAKLTKIETLRFARNYIWALS 53
                 ||.|||||.|:||||||||||.||.||||||||||||||||||.||||||:
NGN3_bHLH_MOU  1 RRKKANDRERNRMHNLNSALDALRGVLPTFPDDAKLTKIETLRFAHNYIWALT 53
```

图：斑马鱼与小鼠NGN3中bHLH结构域序列比对结果



图：斑马鱼与小鼠、人NGN3中bHLH结构域DNA结合位点序列比对结果

斑马鱼与小鼠NGN3中bHLH结构域的比较

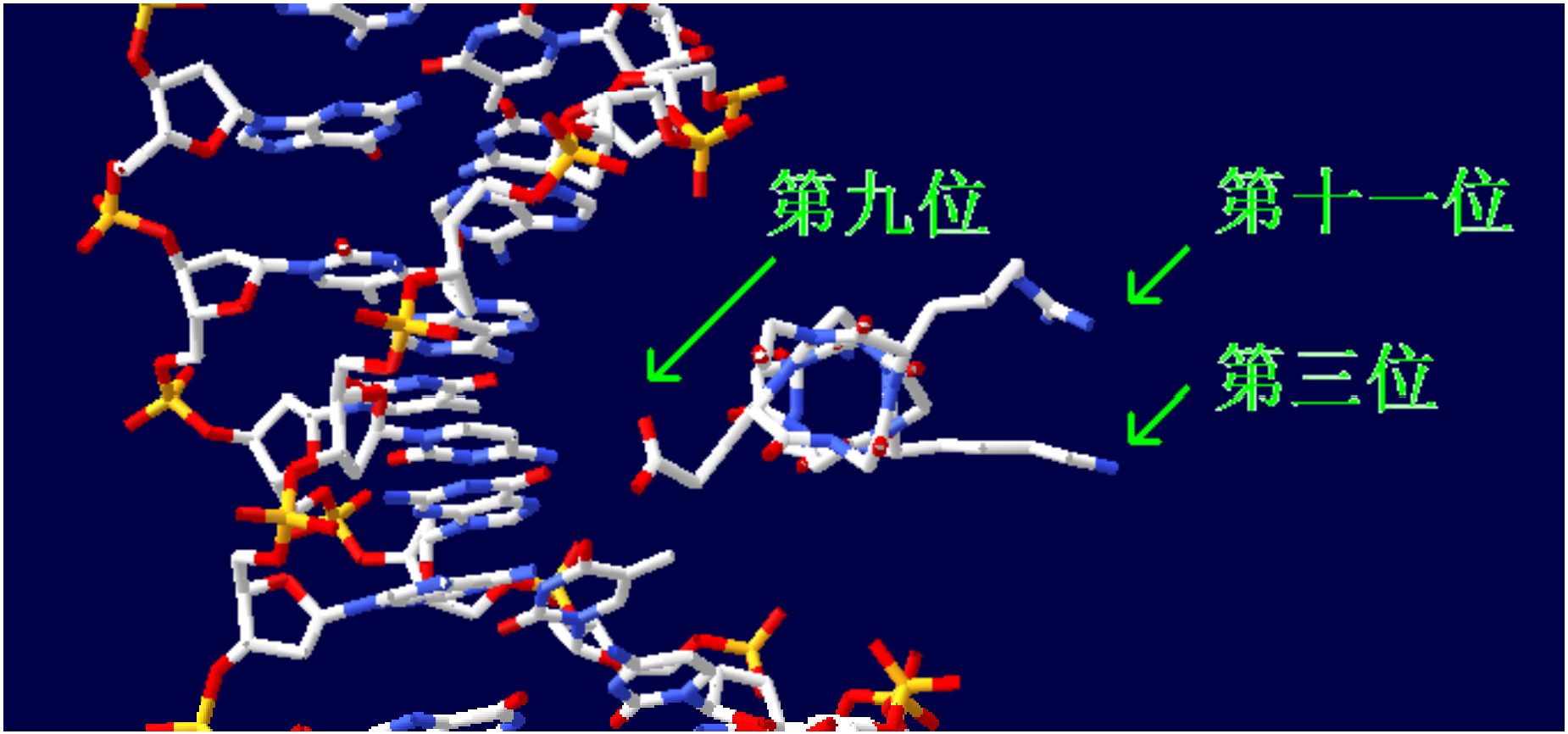
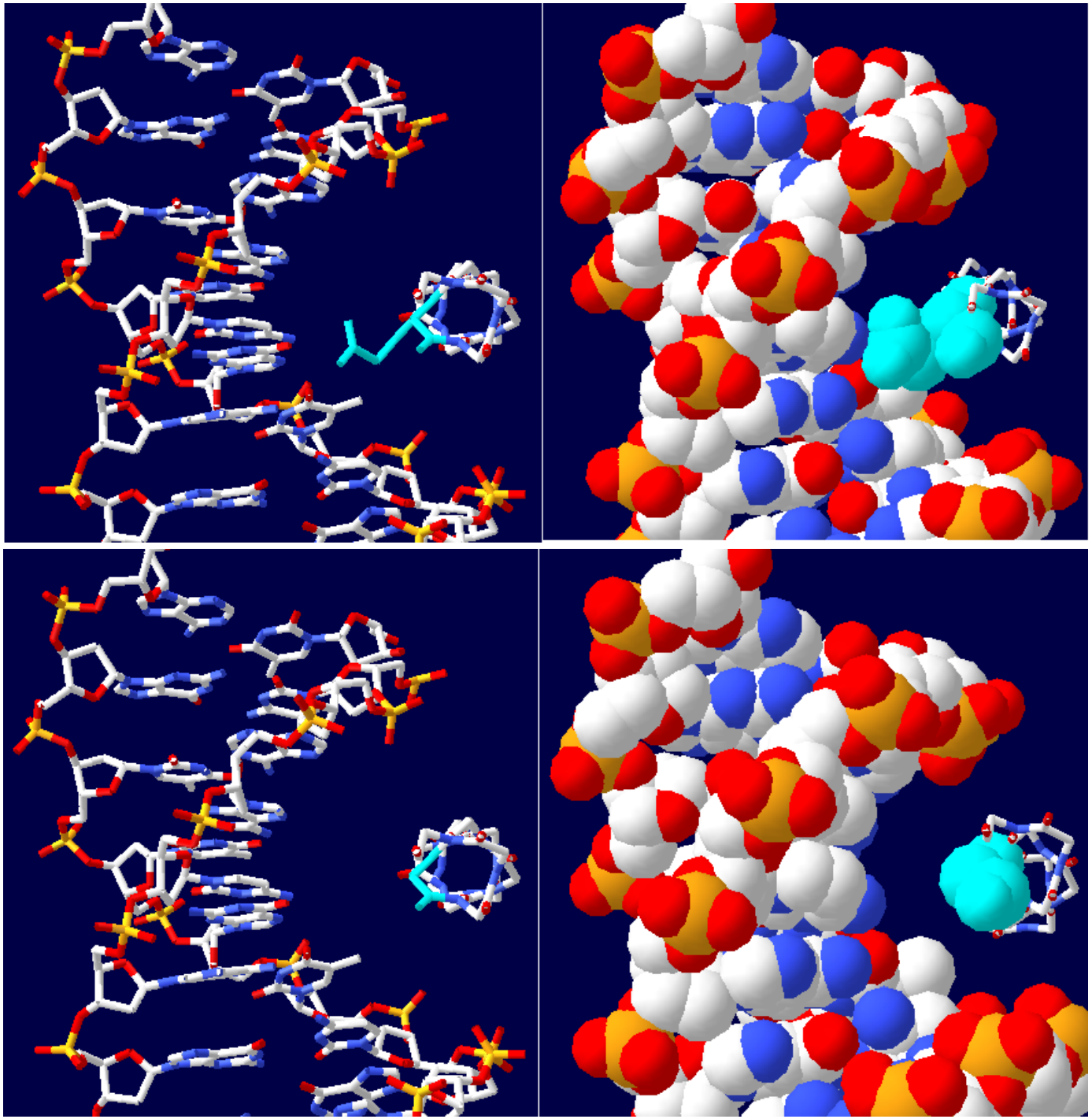


图24：Swiss-PDBviewer展示3个不同氨基酸的位置

斑马鱼与小鼠NGN3中bHLH结构域的比较



图：Swiss-PDBviewer展示第九位氨基酸结构图

显示为天蓝色