Cas9蛋白的结构预测及功能分析 (the structure prediction and functional analysis of CRISPR-associated Protein 9)

汇报：钱坤

第八组：李国亮，晋知文，王玲，钱坤

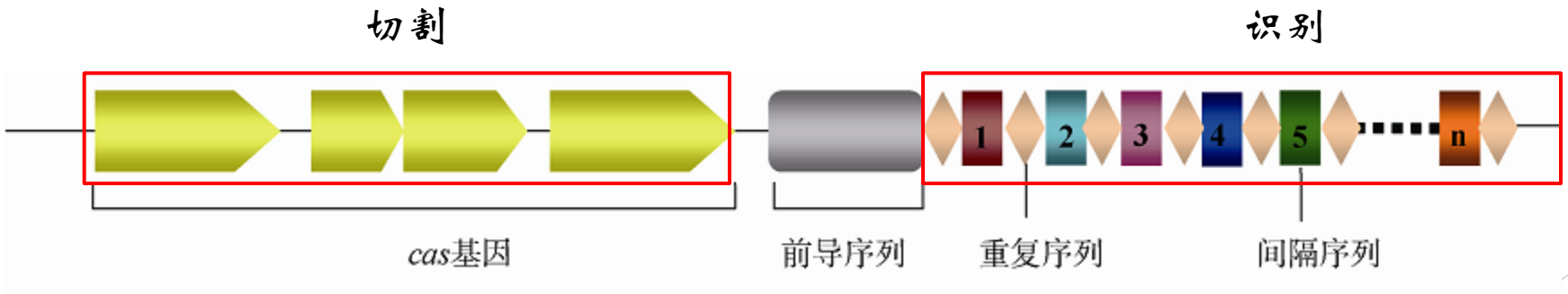
主要内容

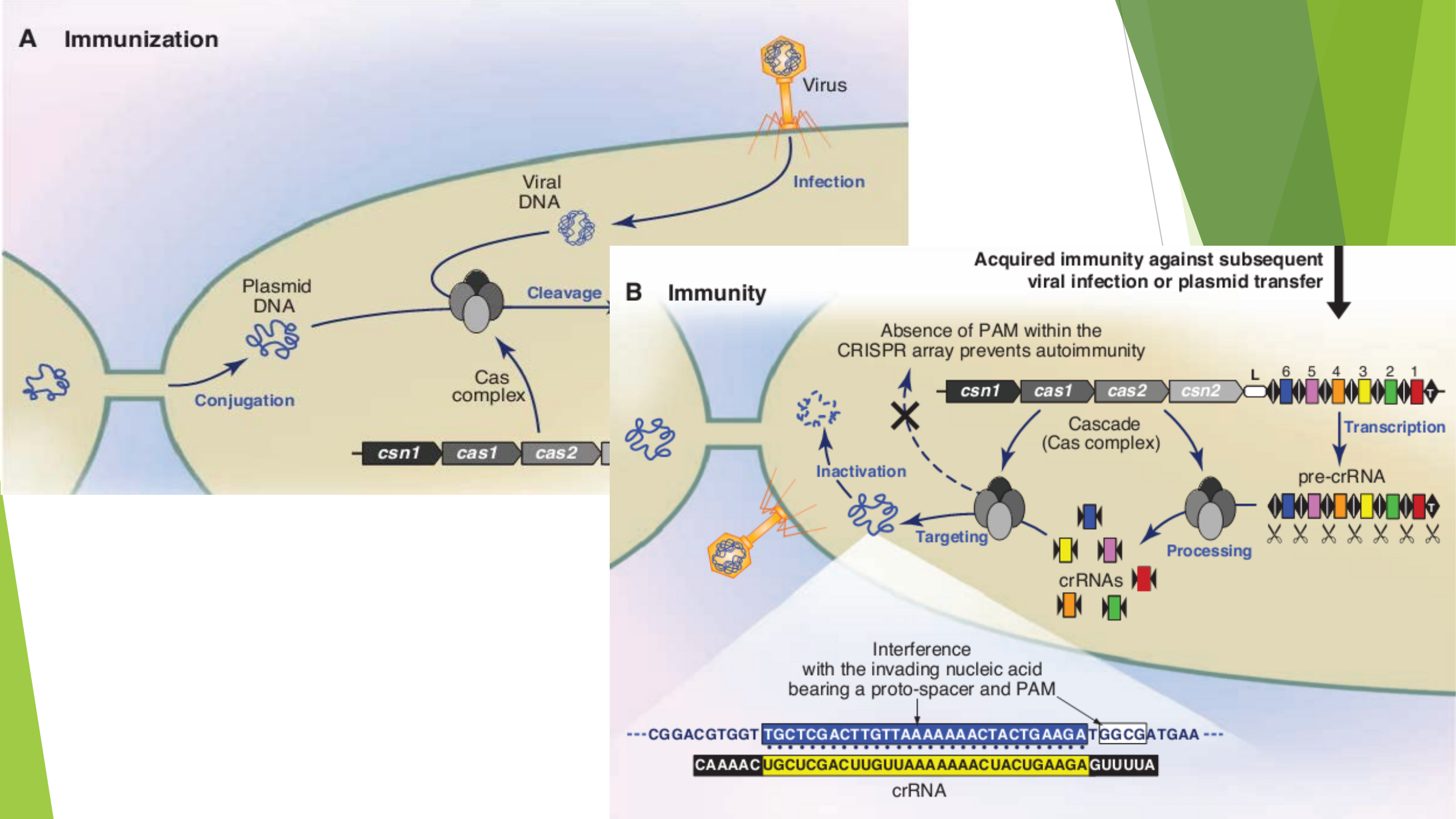
- CAS蛋白的背景
- 研究进展
- 基因分析
- 蛋白质分析
- 参考文献

背景

- CRISPR（clustered, regularly interspaced, short palindromic repeats）/CRISPR-associated(CAS):来源是细菌或古菌获得性免疫，由RNA指导Cas蛋白对靶向基因进行修饰。

CRISPR-Cas主要由两部分组成




A Immunization
Virus
Infection
Viral DNA
Plasmid DNA
Cleavage
Conjugation
Cas complex
csn1
cas1
cas2
Acquired immunity against subsequent viral infection or plasmid transfer
B Immunity
Absence of PAM within the CRISPR array prevents autoimmunity
csn1
cas1
cas2
csn2
L
6
5
4
3
2
1
Cascade (Cas complex)
Transcription
pre-crRNA
Inactivation
Targeting
Processing
crRNAs
Interference with the invading nucleic acid bearing a proto-spacer and PAM
---CGGACGTGGT TGCTCGACTTGTTAAAAAAACTACTGAAGA TGGCGATGAA ---
CAAAAC UGCUCGACUUGUUAAAAAAACUACUGAAGA GUUUUA
crRNA

Cas蛋白

- Many of the Cas proteins are members of the large superfamily of RAMP(receptor activity-modifying protein), which have features of RNA binding proteins。
- It contains two nuclease domains, a RuvC-like nuclease domain near the amino terminus and a HNH-like nuclease domain in the middle of the protein.
- The Cas proteins are expected to function in various aspects of maintenance of CRISPR gene loci (including addition of new invader-derived elements in response to infection) as well as Prokaryotic silencing RNA（psiRNA） biogenesis and psiRNA-mediated resistance to invaders.

研究进展

- CRISPR/Cas systems have been categorized into three main types(Ⅰ，II，Ⅲ), based on core elements content and sequences。
- The CRISPR3/Cas system of Streptococcus thermophiles DGCC7710 is a type II system that consists of four cas genes（cas9, cas1, cas2, and csn2）。
- cas9 (formerly named “cas5” or “csn1”)is the signature gene for type II systems 。
- cas9 protein is involved in CRISPR RNAs(crRNAs) processing and/or crRNA-mediated silencing of invasive DNA.
- DNA cleavage is executed by Cas9, which uses two distinct active sites, RuvC and HNH, to generate site-specific nicks on opposite DNA strands.

- 已知信息：

① 基因全长-7655bp

② mRNA序列-4229bp

③ 蛋白序列-1409aa

④ 蛋白功能：内切酶活性；与核酸共价结合

- 未知信息：

蛋白3D结构

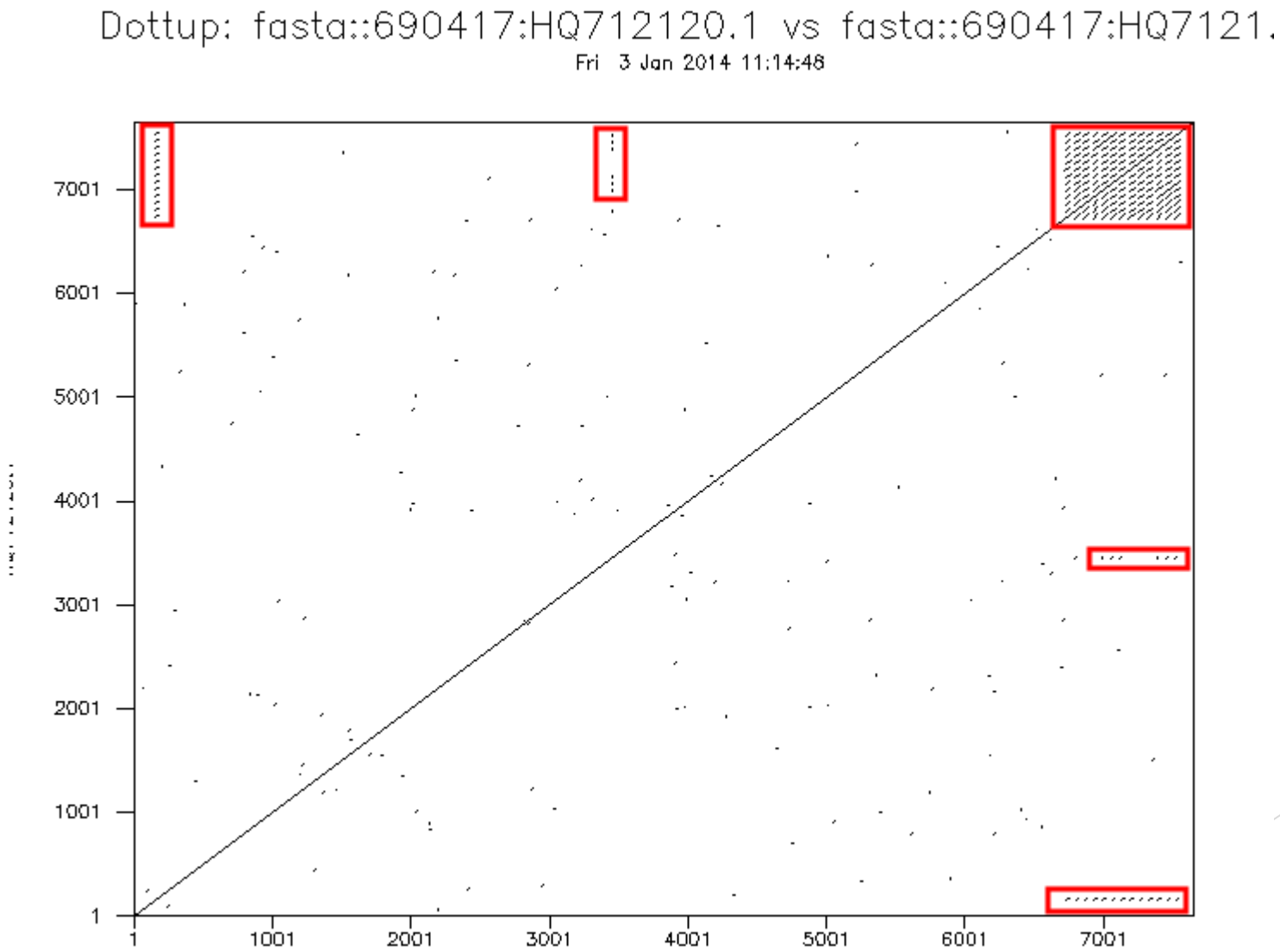
- 分析目的：

① 利用生物信息学工具预测蛋白质结构

② 验证蛋白功能

基因分析

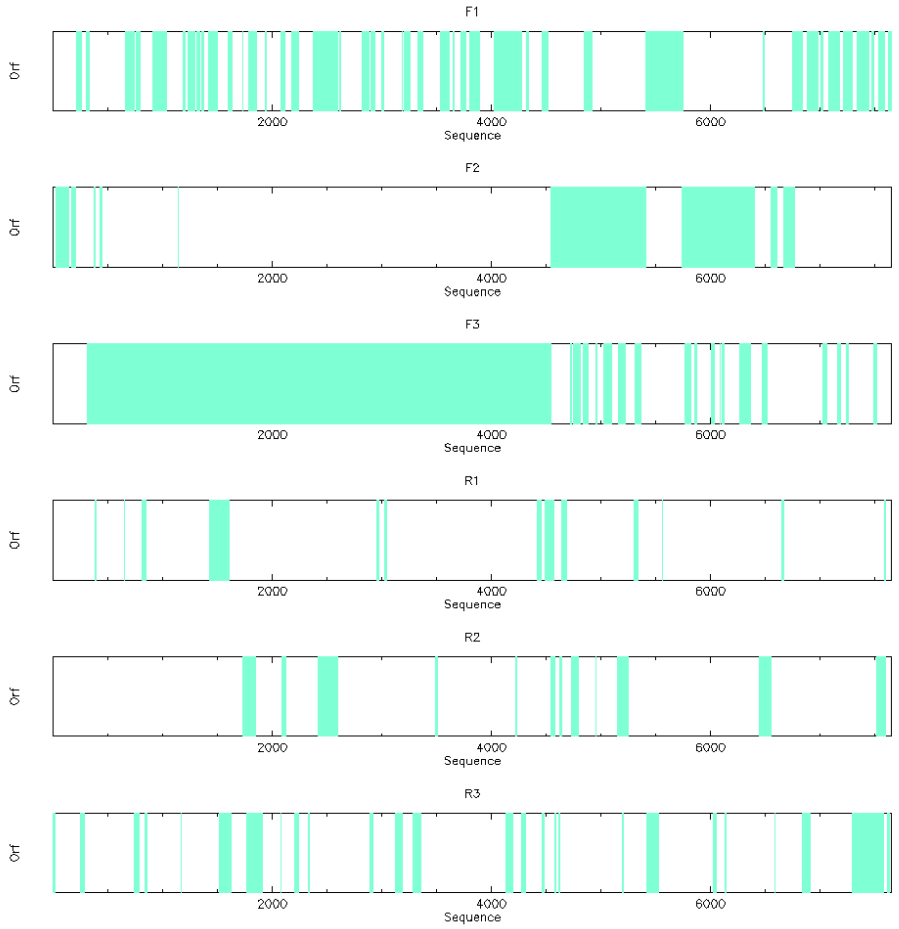
- Dottup查看基因结构



有重复序列

多种方法预测ORF

plotorf



getorf (pep_minisize=200)

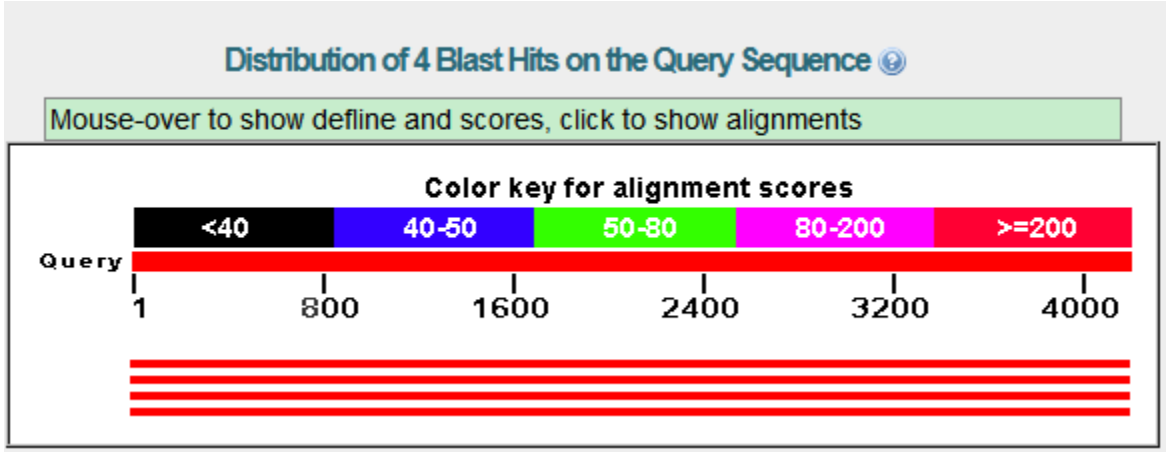
>HQ712120.1_1 [315 - 4541] Streptococcus thermophilus strain DGCC 7710 SerB (serB) gene, partial cds; CRISPR3 gene locus, complete sequence; and predicted protein gene, partial cds
MLFNKCIIISINLDFSNKEKCMTKPYSIGLDIGTNSVGWAVITDNYKVPSKKMKVLGNTS
KKYIKKNLLGVLLFDSGITAEGRRLKRTARRRYTRRRNRILYLQEIFSTEMATLDDAFFQ
RLDDSFLVFDDKRDSKYPIFGNLVEEKVYHDEFPTIYHLRKYLADSTKKADLRLVYLALA
HMIKYRGHFLIEGEFNSKNNDIQKNFQDFLDTYNAIFESDLSLENSKQLEEIVKDKISKL
EKKDRILKLFPGEKNSGIFSEFLKLIVGNQADFRKCFNLDEKASLHFSKESYDEDLETLL
GYIGDDYSDVFLKAKKLYDAILLSGFLTVTDNETEAPLSSAMIKRYNEHKEDLALLKEYI
RNISLKTYNEVFKDDTKNGYAGYIDGKTNQEDFYVYLKNLLAEFEGADYFLEKIDREDFL
RKQRTFDNGSIPYQIHLQEMRAILDKQAKFYPFLAKNKERIEKILTFRIPYYVGPLARGN
SDFAWSIRKRNEKITPWNFEDVIDKESSAEAFINRMTSFDLYLPEEKVLPKHSLLYETFN
VYNELTKVRFIAESMRDYQFLDSKQKKDIVRLYFKDKRKVTDKDIIEYLHAIYGYDGIEL
KGIEKQFNSSLSTYHDLLNIINDKEFLDDSSNEAIIEEIIHTLTIFEDREMIKQRLSKFE
NIFDKSVLKKLSRRHYTGWGKLSAKLINGIRDEKSGNTILDYLIDDGISNRNFMQLIHDD
ALSFKKKIQKAQIIGDEDKGNIKEVVKSLPGSPAIKKGILQSIKIVDELVKVMGGRKPES
IVVEMARENQYTNQGKSNSQQRLKRLEKSLKELGSKILKENIPAKLSKIDNNALQNDRLY
LYYLQNGKDMYTGDDLDIDRLSNYDIDHIIPQAFLKDNSIDNKVLVSSASNRGKSDDFPS
LEVVKKRKTFWYQLLKSKLISQRKFDNLTKAERGGLLPEDKAGFIQRQLVETRQITKHVA
RLLDEKFNNKKDENNRAVRTVKIITLKSTLVSQFRKDFELYKVREINDFHHAHDAYLNAV
IASALLKKYPKLEPEFVYGDYPKYNSFRERKSATEKVYFYSNIMNIFKKSISLADGRVIE
RPLIEVNEETGESVWNKESDLATVRRVLSYPQVNVVKKVEEQNHGLDRGKPKGLFNANLS
SKPKPNSNENLVGAKEYLDPKKYGGYAGISNSFAVLVKGTIEKGAKKKITNVLEFQGISI
LDRINYRKDKLNFLLEKGYKDIELIIELPKYSLFELSDGSRRMLASILSTNNKRGEIHKG
NQIFLSQKFVKLLYHAKRISNTINENHRKYVENHKKEFEELFYYILEFNENYVGAKKNGK
LLNSAFQSWQNHSIDELCSSFIGPTGSERKGLFELTSRGSAADFEFLGVKIPRYRDYTPS
SLLKDATLIHQSVTGLYETRIDLAKLGEG
>HQ712120.1_2 [4544 - 5410] Streptococcus thermophilus strain DGCC 7710 SerB (serB) gene, partial cds; CRISPR3 gene locus, complete sequence; and predicted protein gene, partial cds
MAGWRTVVVNIHSKLSYKNNHLIFRNSYKTEMIHLSEIDILLLETTDIVLTTMLVKRLVD
ENILVIFCDDKRLPTAFLTPYYARHDSSLQIARQIAWKENVKCEVWTAIIAQKILNQSYY
LGECSFFEKSQSIMELYHGLERFDPSNREGHSARIYFNTLFGNDFTRESDNDINAALDYG
YTLLLSMFAREVVVCGCMTQIGLKHANQFNQFNLASDIMEFFRPIIDRIVYQNRHNNFVK
IKKELFSIFSETYLYNGKEMYLSNIVSDYTKKVIKALNQLGEEIPEFRI
>HQ712120.1_3 [5410 - 5751] Streptococcus thermophilus strain DGCC 7710 SerB (serB) gene, partial cds; CRISPR3 gene locus, complete sequence; and predicted protein gene, partial cds
MSYRYMRMILMFDMPTDTAEERKAYRKFRKFLLSEGFIMHQFSVYSKLLLNHTANTAMVG
RLKANNPKKGNITILTVTEKQFARMIYLYGDKNTSIANSEERLVFLGDNYCDED
>HQ712120.1_4 [5744 - 6400] Streptococcus thermophilus strain DGCC 7710 SerB (serB) gene, partial cds; CRISPR3 gene locus, complete sequence; and predicted protein gene, partial cds
MKINFSLLDEPMEVNLGTVLVIEDVSVFAQLVKEFYQYDEQSNLTIFDSKIRSIRSSELL
LITDILGYDINTSQVLKLLHTDIVSQLNDKPEVRSEIDSLVSLITDIIMAECIENELDIE
YDEITLLELIKALGVRIETKSCTVFEKIFEILQIFKYLVKKRILVFVNSLSYFSKDEIYQ
ILEYTKLSQADVLFLEPRQIEGIQQGFILDKDYILMFYNN
>HQ712120.1_5 [7580 - 7296] (REVERSE SENSE) Streptococcus thermophilus strain DGCC 7710 SerB (serB) gene, partial cds; CRISPR3 gene locus, complete sequence; and predicted protein gene, partial cds
MSTENCDNKSFGTIRNNTALKLSATVNLTIHSFGTIRNNTALKLYNCSSILWDSFGTIRN
NTALKLCNLFGSIVIGFGTIRNNTALKQYRLHYIC

- 使用Softberry预测ORF

N	Tu/Op	Conserved pairs(N/Pv)	S		Start		End	Score
1	1 Op 1	.	+	CDS	315	-	4544	1041
2	1 Op 2	.	+	CDS	4544	-	5413	311
3	1 Op 3	.	+	CDS	5425	-	5754	234
4	1 Op 4	.	+	CDS	5744	-	6403	431
5	2 Tu 1	.	-	CDS	7293	-	7580	113

多种orf预测方法显示315-4544bp为编码区间，CDS共4229bp，编码氨基酸1409个

- mRNA序列比对



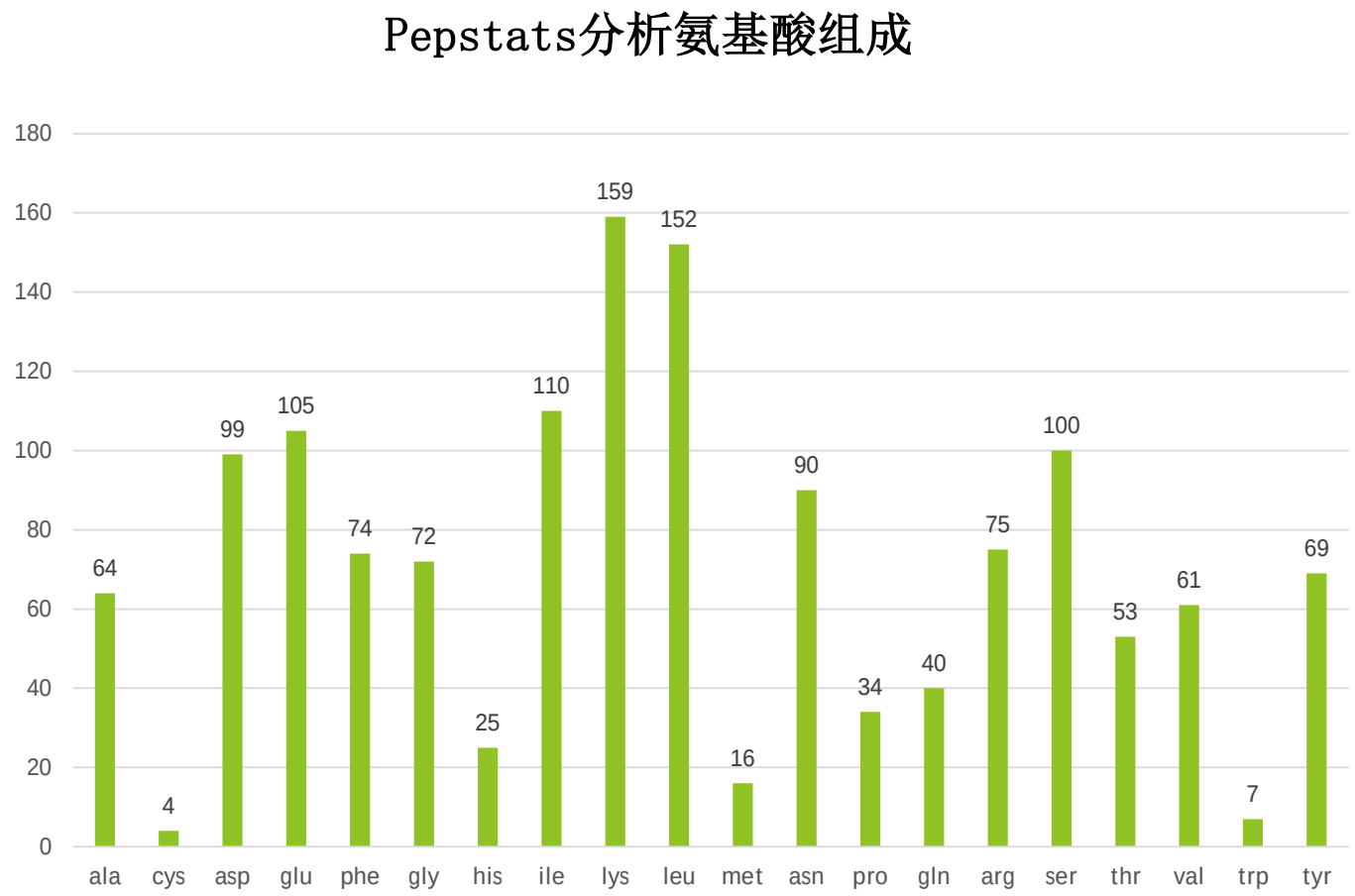
Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

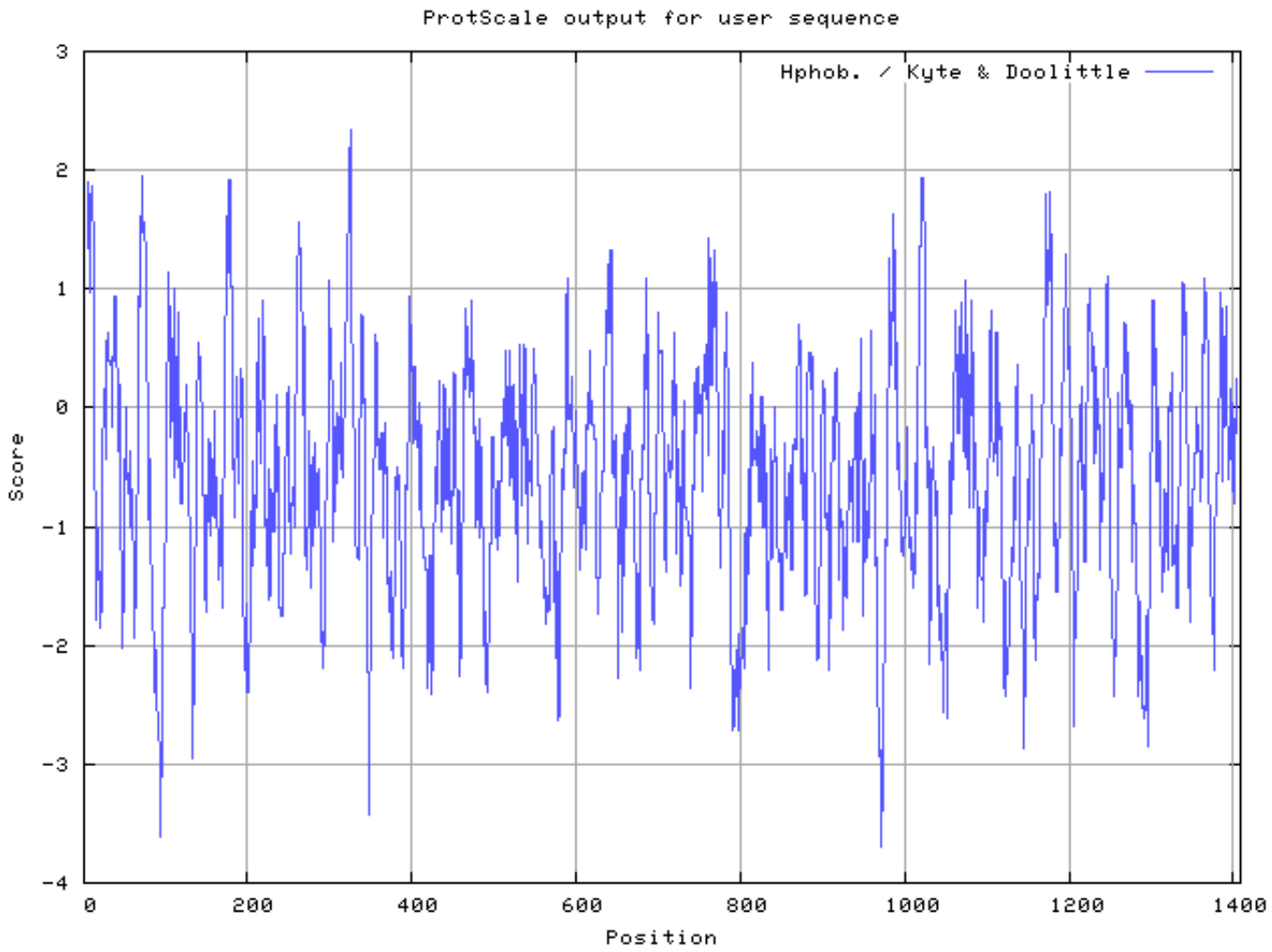
	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Streptococcus thermophilus strain DGCC 7710 SerB (serB) gene, partial cds; CRISPR3 gene locus, complete sequence; and predicted p	7696	7696	100%	0.0	100%	HQ712120.1
☐	Streptococcus thermophilus MN-ZLW-002, complete genome	7685	7685	100%	0.0	99%	CP003499.1
☐	Streptococcus thermophilus ND03, complete genome	7679	7679	100%	0.0	99%	CP002340.1
☐	Streptococcus thermophilus LMD-9, complete genome	7640	7640	100%	0.0	99%	CP000419.1

Database：nucleotide collection
E-value：0.001

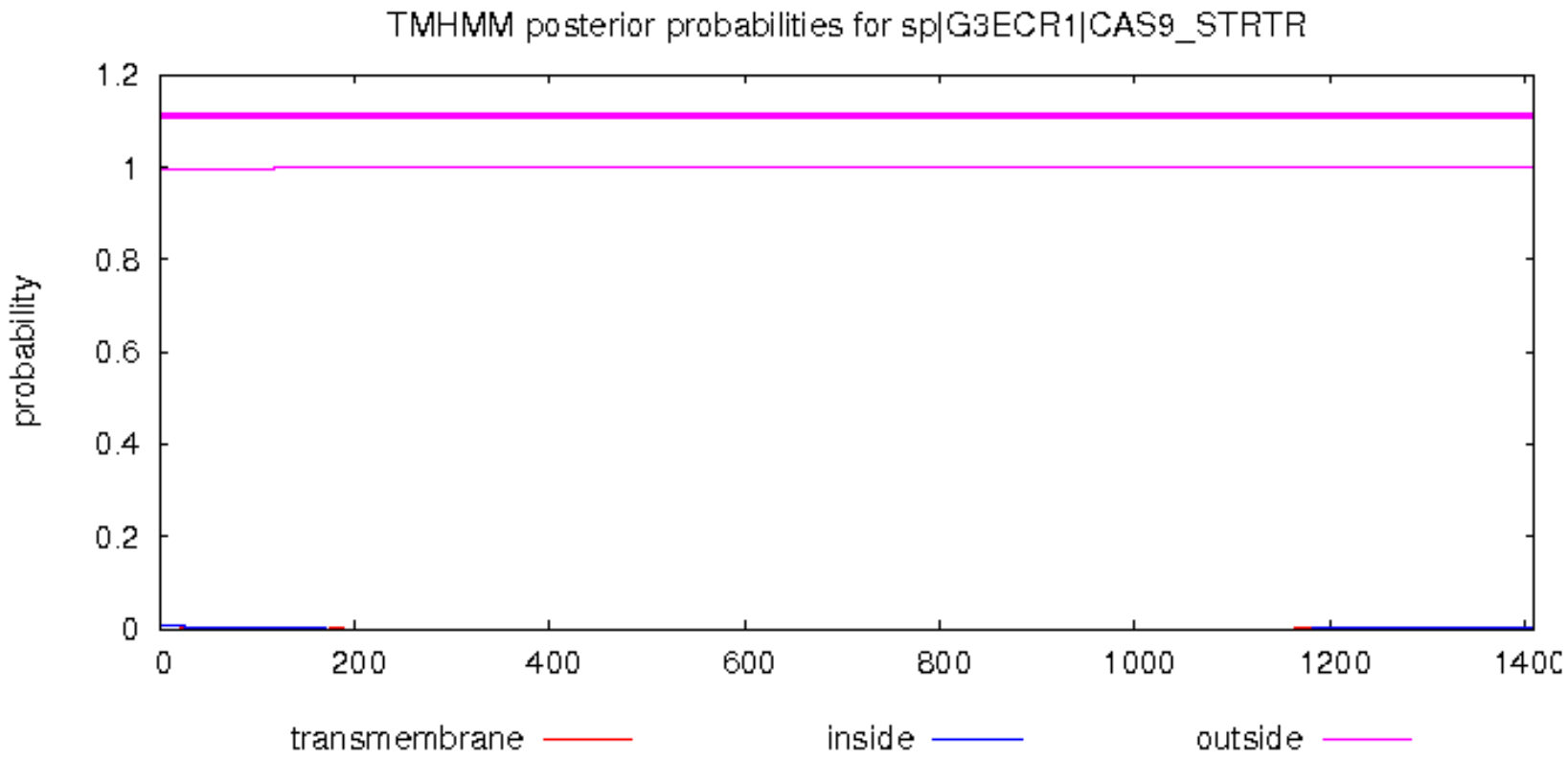
蛋白质分析


Pepstats分析氨基酸组成
180
160
140
120
100
80
60
40
20
0
64
4
99
105
74
72
25
110
159
152
16
90
34
40
75
100
53
61
7
69
ala
cys
asp
glu
phe
gly
his
ile
lys
leu
met
asn
pro
gln
arg
ser
thr
val
trp
tyr

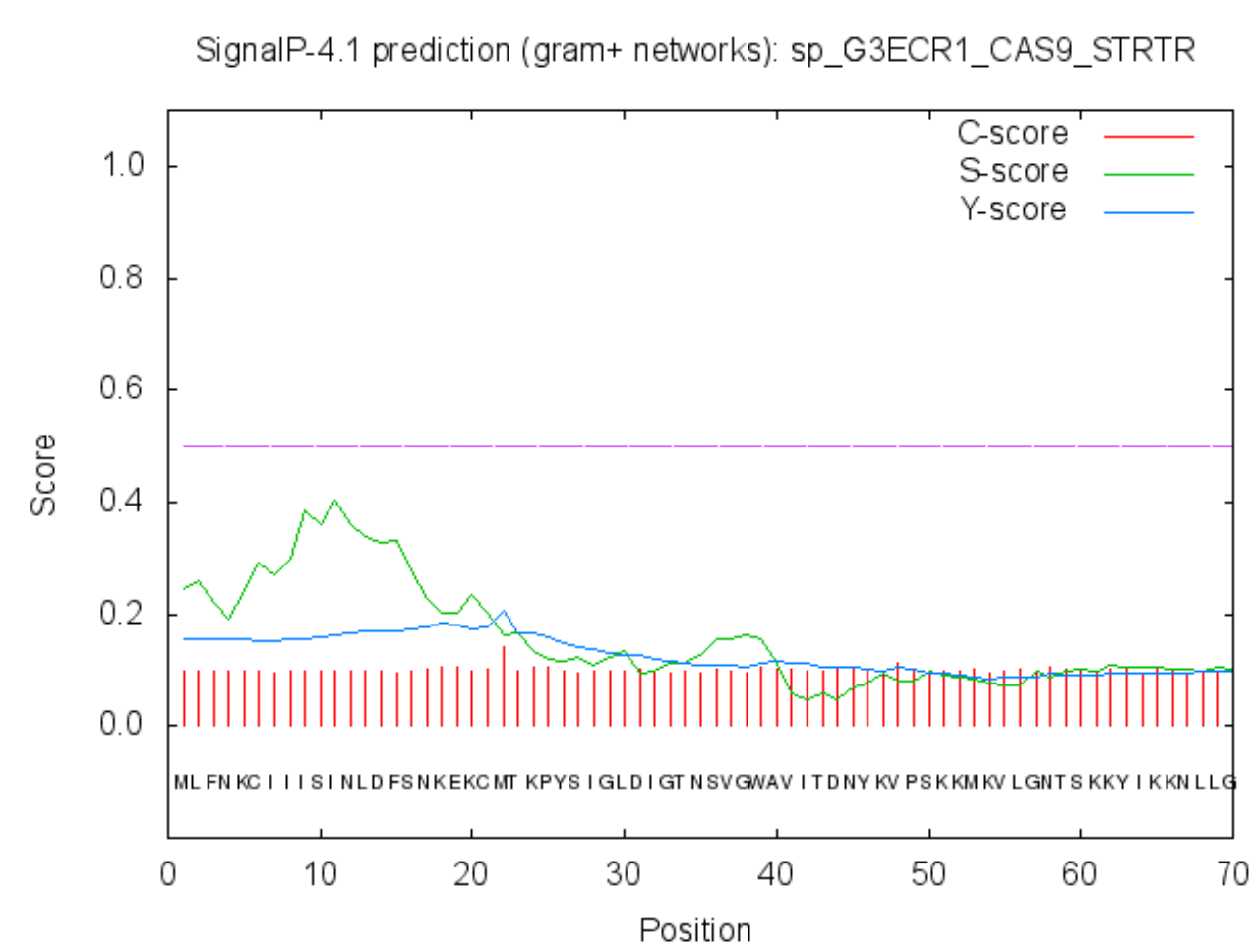
蛋白亲疏水性分析


ProtScale output for user sequence
Hphob. / Kyte & Doolittle
Score
Position
3
2
1
0
-1
-2
-3
-4
0
200
400
600
800
1000
1200
1400

跨膜区分析


TMHMM posterior probabilities for sp|G3ECR1|CAS9_STRTR
probability
1.2
1
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000
1200
1400
transmembrane
inside
outside

利用Signal P预测是否有信号肽

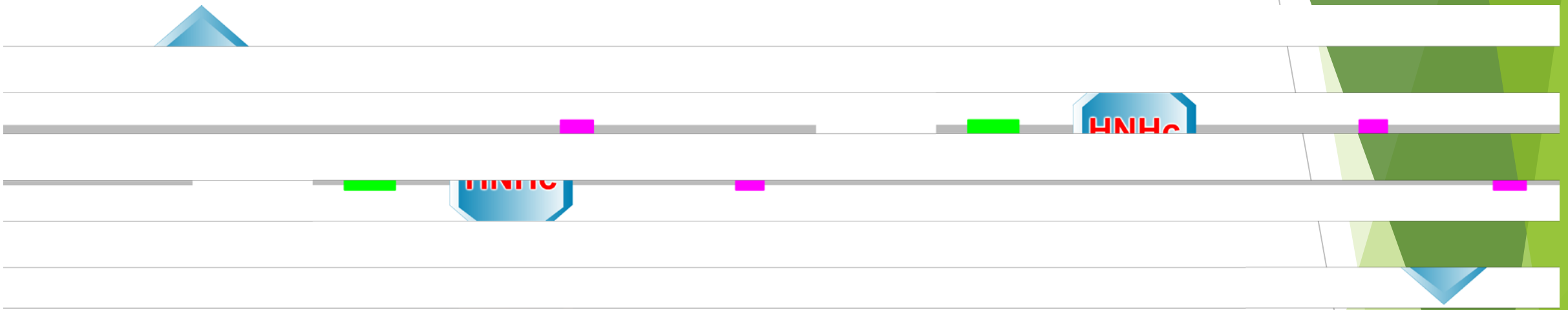


# Measure	Position	Value	Cutoff	signal peptide?
max. C	22	0.142		
max. Y	22	0.205		
max. S	11	0.403		
mean S	1-21	0.280		
D	1-21	0.234	0.450	NO

亚细胞定位

```
SeqID: sp|G3ECR1|CAS9_STRTR CRISPR-associated endonuclease Cas9 OS=Streptococcus thermophilus GN=cas9 PE=1 SV=2
  Analysis Report:
    CMSVM+              Unknown                     [No details]
    CWSVM+              Unknown                     [No details]
    CytoSVM+            Unknown                     [No details]
    ECSVM+              Unknown                     [No details]
    ModHMM+             Unknown                     [No internal helices found]
    Motif+              Unknown                     [No motifs found]
    Profile+            Unknown                     [No matches to profiles found]
    SCL-BLAST+          Unknown                     [No matches against database]
    SCL-BLASTe+         Unknown                     [No matches against database]
    Signal+             Unknown                     [No signal peptide detected]
  Localization Scores:
    Cytoplasmic             2.50
    CytoplasmicMembrane     2.50
    Cellwall                2.50
    Extracellular           2.50
  Final Prediction:
    Unknown
```

SMART分析保守结构域



Confidently predicted domains, repeats, motifs and features:

Name	Start ▲	End	E-value
low complexity	83	99	N/A
coiled coil	792	813	N/A
HNHc	838	892	3.64
low complexity	964	975	N/A
low complexity	1297	1310	N/A

Protein of unknown function (DUF1524) Provide feedback

This family of uncharacterised proteins contain a conserved HXXP motif. A similar motif is seen in protein families in the His-Me finger endonuclease superfamily which suggests this family of proteins may also act as endonucleases.

HNHc domain

This is a SMART **HNHc** domain (full annotation).

Position:	**838** to **892**
E-value:	**3.64453454691318** (HMMER2)

SMART ACC:	SM000507
Definition:	HNH nucleases
Description:	
Interpro abstract (IPR003615):	This domain is found in HNH family of nucleases that includes yeast intron 1 protein, MutS, and bacterial colicins and pyocins. They are found in bacteria, viruses and eukaryotes.

HNHc domain sequence (55 aa):

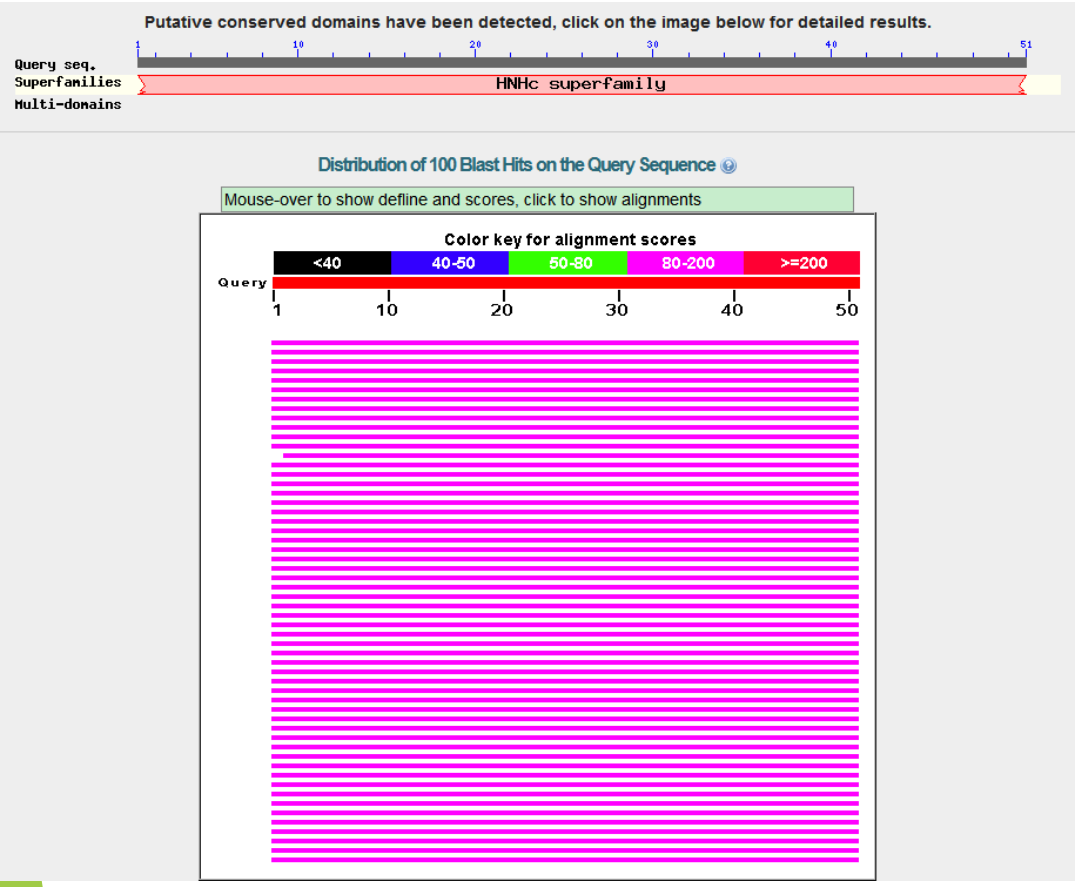
Submit to BLAST | Align with the SMART alignment | Copy to clipboard

RLYLYYLQNGKDMYTGDDLDIDRLSNYDIDHIIPQAFLKDNSIDNKVLVSSASNR

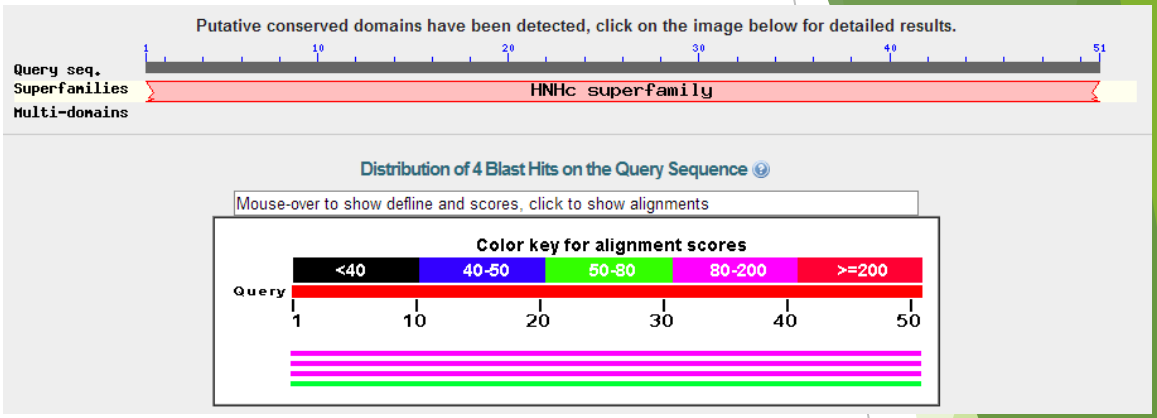
预测的HNHc结构域序列及位置与uniprot里的结果有误差。

HNH Cas9-type(51aa)
DMYTGDDLDIDRLSNYDIDHIIPQAFLKDNSIDNKVL
VSSASNRGKSDDFP

对HNH结构域序列blast分析



数据库：refseq_protein
E-value：0.001
Organism:bacteria



数据库：swissprot
E-value：0.001
Organism:bacteria

经HNH结构域只出现在cas9蛋白？

序列比对

```
GRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEH------PVENTQ   805   Q99ZW2   CAS9_STRP1
GGRKPESIVVEMARENQYTNQGKSNSQQRLKRLEKSLKELGSKILKENIPAKLSKIDNNA   812   Q03JI6   CAS9B_STRTD
GY-PPQTIVVEMARENQTTGKGKNNSRPRYKSLEKAIKEFGSQILKEH------PTDNQE   808   Q927P4   CAS9_LISIN
GGRKPESIVVEMARENQYTNQGKSNSQQRLKRLEKSLKELGSKILKENIPAKLSKIDNNA   833   G3ECR1   CAS9_STRTR
:    *:.**:******* * :*:.**: * * :*:.:**:**:****.        :*

.QNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRG   865   Q99ZW2   CAS9_STRP1
.QNDRLYLYYLQNGKDMYTGDDLDIDRLSNYDIDHIIPQAFLKDNSIDNKVLVSSASNRG   872   Q03JI6   CAS9B_STRTD
.RNNRLYLYYLQNGKDMYTGQDLDIHNLSNYDIDHIVPQSFITDNSIDNLVLTSSAGNRE   868   Q927P4   CAS9_LISIN
.QNDRLYLYYLQNGKDMYTGDDLDIDRLSNYDIDHIIPQAFLKDNSIDNKVLVSSASNRG   893   G3ECR1   CAS9_STRTR
::*::**********:***.  ::***..**:**:***:**:*:.*:**** **.  *   **

LSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETR   925   Q99ZW2   CAS9_STRP1
LSDDVPSLEVVKKRKTFWYQLLKSKLISQRKFDNLTKAERGGLSPEDKAGFIQRQLVETR   932   Q03JI6   CAS9B_STRTD
LGDDVPPLEIVRKRKVFWEKLYQGNLMSKRKFDYLTKAERGGLTEADKARFIHRQLVETR   928   Q927P4   CAS9_LISIN
LSDDFPSLEVVKKRKTFWYQLLKSKLISQRKFDNLTKAERGGLLPEDKAGFIQRQLVETR   953   G3ECR1   CAS9_STRTR
:.*:.*  *:*:* * :* :* :.:*:::**** **********   *** **:********

QITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAH   985   Q99ZW2   CAS9_STRP1
```

Phyre 2

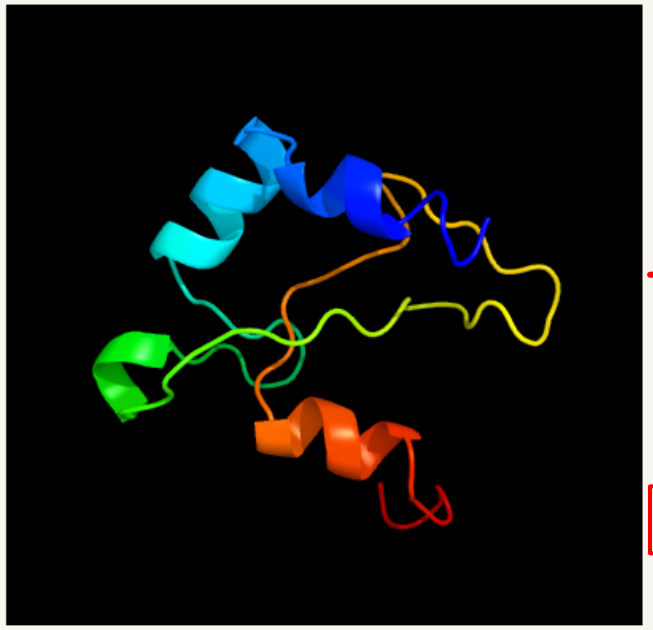


Image coloured by rainbow N → C terminus

PDB header:hydrolase
Chain: A: **PDB Molecule:**hnh endonuclease;
PDBTitle: x-ray structure of the nhn endonuclease from geobacter2 metallireducens. northeast structural genomics consortium target3 gmr87.

Confidence and coverage

Confidence: **99.3%** Coverage: **5%**

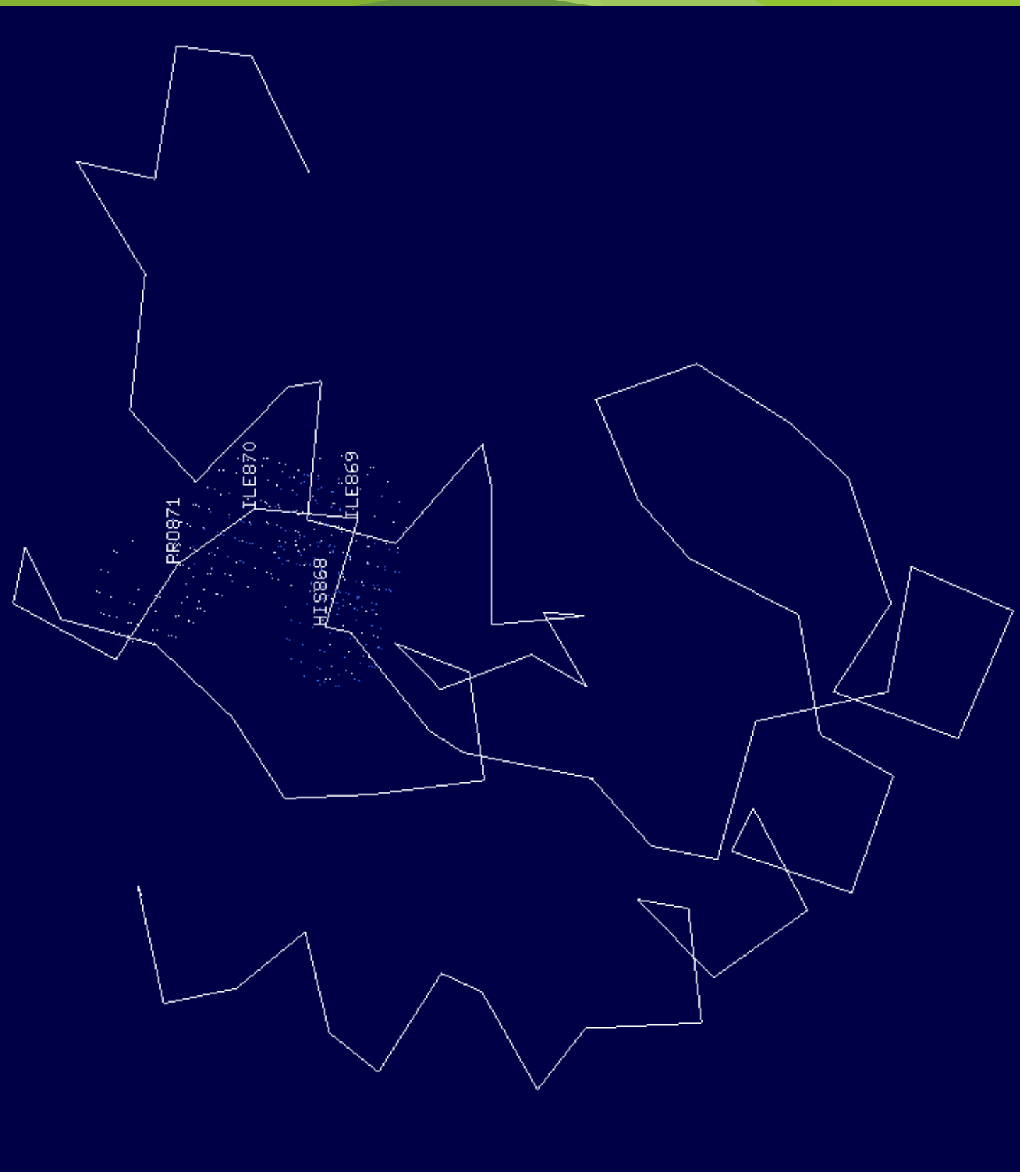
73 residues (5% of your sequence) have been modelled with 99.3% confidence by the single highest scoring template.

Additional confident templates have been detected (see Domain analysis) which cover other regions of your sequence.

178 residues (13%) could be modelled at >90% confidence using multiple-templates.

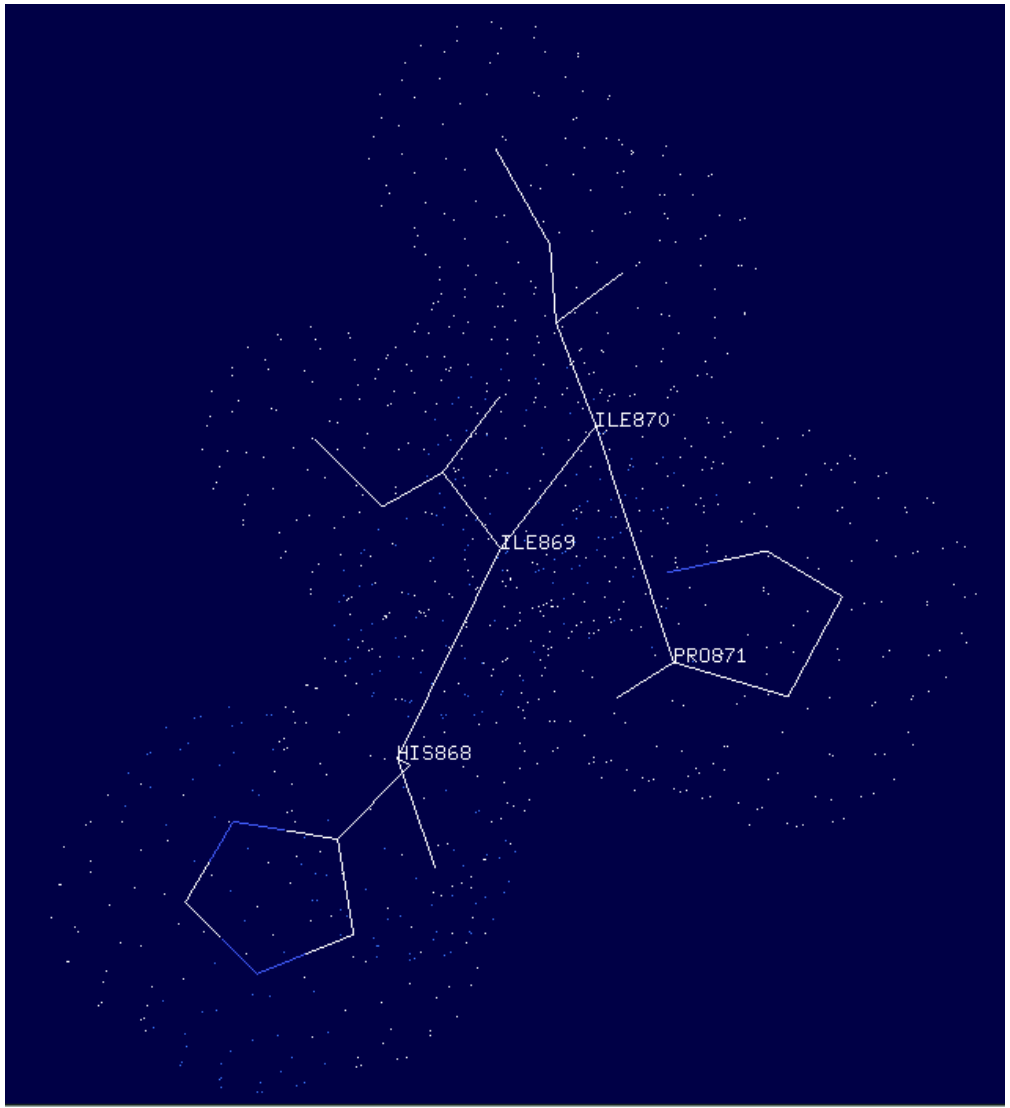
You may wish to try resubmitting your sequence in "intensive" mode to model more of your sequence.

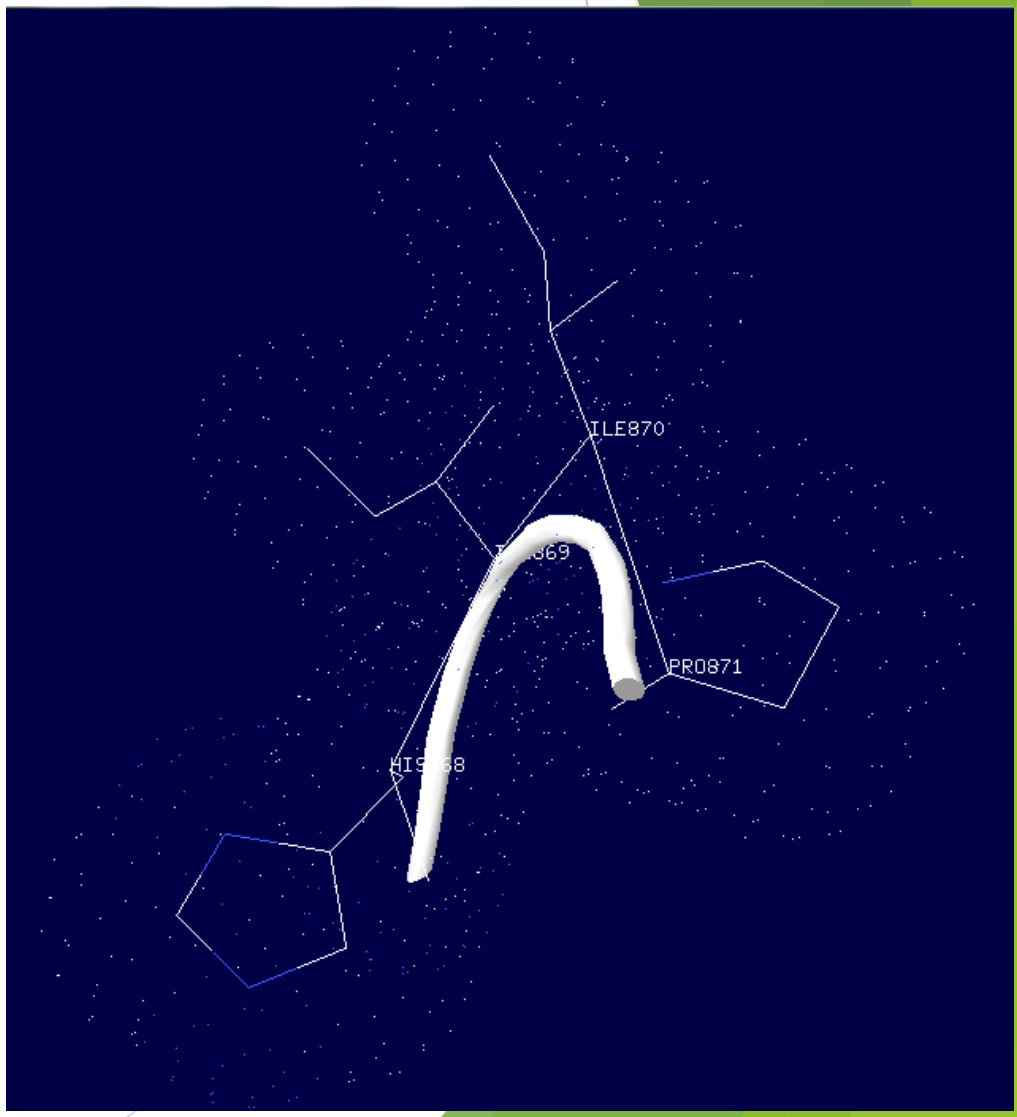
实际上只对838-902位置进行预测



使用SPDB-Viewer寻找铁指结构

使用SPDB-Viewer寻找锌指结构





参考文献

- Philippe Horvath, Rodolphe Barrangou. CRISPR/Cas, the Immune System of Bacteria and Archaea. Science 327, 167(2010).
- Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell. 2009 Nov 25;139(5):945-56.
- Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A. 2012 Sep 25;109(39):E2579-86.
- Krishna SS, Majumdar I, Grishin NV. Structural classification of zinc fingers: survey and summary. Nucleic Acids Res. 2003 Jan 15;31(2):532-50.

Thank you!