“实用生物信息技术”课程小组讨论总结报告

组：G3　**次：**R3　**组长：**陈伟辉　**执笔：**王紫贤

1. 时间

2026年4月22日星期三

2. 方式

面对面交流

3. 主题

课后复习与组内互助

4. 内容

（一）血红蛋白基础知识查阅与物种演化关系分析

（二）蛋白质与DNA序列比对分析及参数影响探讨

（三）UniProt高级检索实践与数据库交叉链接应用

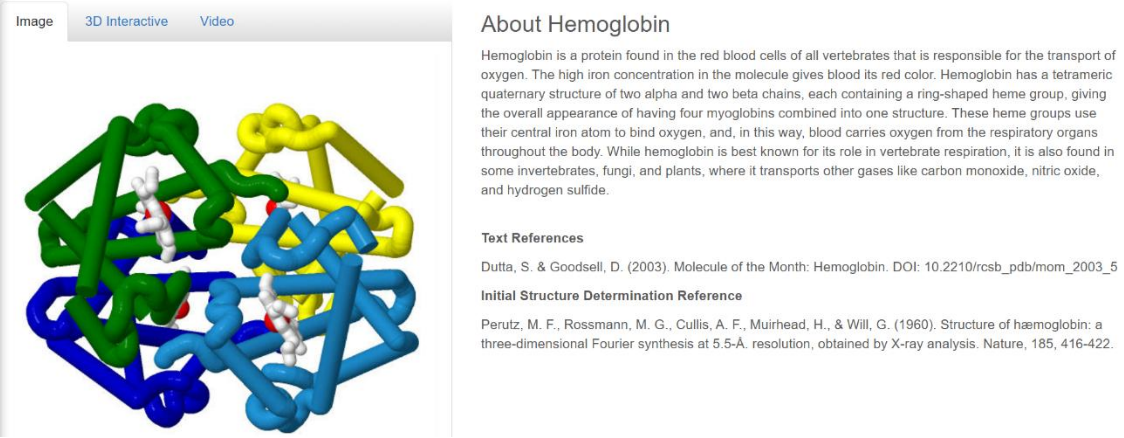
5. 问题

见报告后面

（一）血红蛋白基础知识查阅与物种演化关系分析

（1）阅读分子月报、蛋白质精选以及维基百科等网站中有关血红蛋白的介绍，了解血红蛋白的生理功能、空间结构、亚基组成等基本知识。

分子月报：



About Hemoglobin

Hemoglobin is a protein found in the red blood cells of all vertebrates that is responsible for the transport of oxygen. The high iron concentration in the molecule gives blood its red color. Hemoglobin has a tetrameric quaternary structure of two alpha and two beta chains, each containing a ring-shaped heme group, giving the overall appearance of having four myoglobins combined into one structure. These heme groups use their central iron atom to bind oxygen, and, in this way, blood carries oxygen from the respiratory organs throughout the body. While hemoglobin is best known for its role in vertebrate respiration, it is also found in some invertebrates, fungi, and plants, where it transports other gases like carbon monoxide, nitric oxide, and hydrogen sulfide.

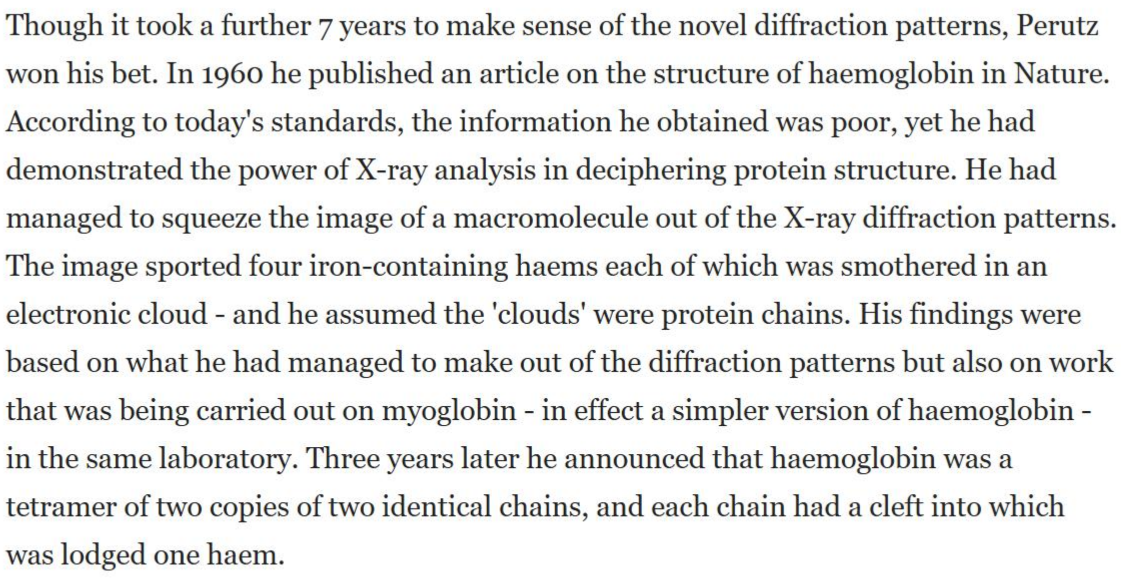
Text References

Dutta, S. & Goodsell, D. (2003). Molecule of the Month: Hemoglobin. DOI: 10.2210/rcsb_pdb/mom_2003_5

Initial Structure Determination Reference

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., & Will, G. (1960). Structure of hæmoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis. Nature, 185, 416-422.

蛋白质精选：



Though it took a further 7 years to make sense of the novel diffraction patterns, Perutz won his bet. In 1960 he published an article on the structure of haemoglobin in Nature. According to today's standards, the information he obtained was poor, yet he had demonstrated the power of X-ray analysis in deciphering protein structure. He had managed to squeeze the image of a macromolecule out of the X-ray diffraction patterns. The image sported four iron-containing haems each of which was smothered in an electronic cloud - and he assumed the 'clouds' were protein chains. His findings were based on what he had managed to make out of the diffraction patterns but also on work that was being carried out on myoglobin - in effect a simpler version of haemoglobin - in the same laboratory. Three years later he announced that haemoglobin was a tetramer of two copies of two identical chains, and each chain had a cleft into which was lodged one haem.

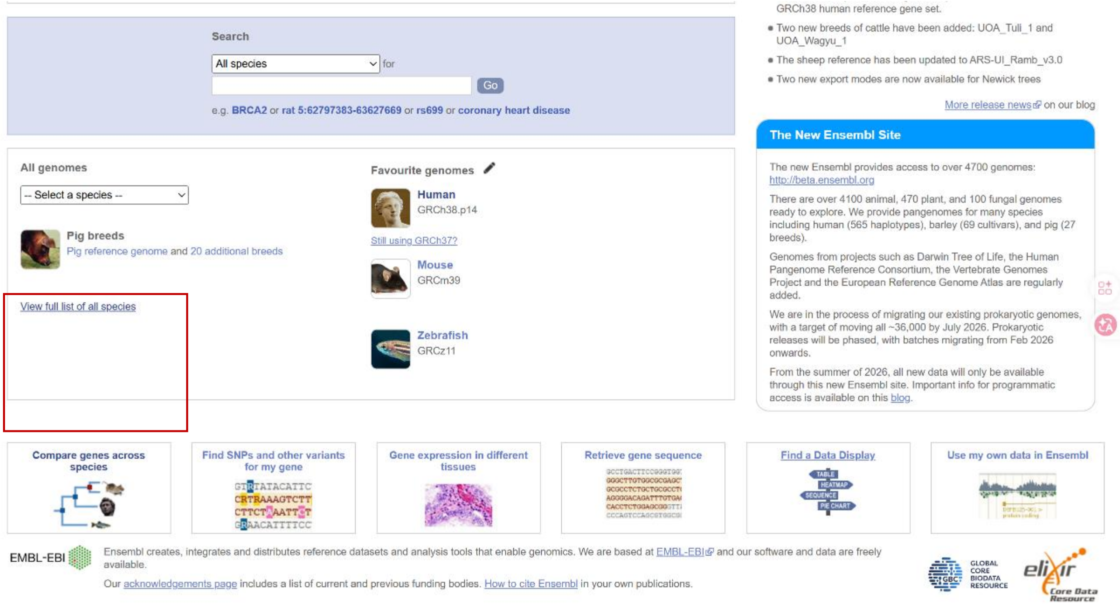
（2）查阅 ENSEMBL 基因组数据库中已经或正在进行基因组测序的物种树，了解人、小鼠、大鼠三个物种之间演化关系；检索物种分歧时间数据库 TimeTree，了解人和小鼠、小鼠和大鼠之间的分歧时间。

系统发育树的核心逻辑：两个物种的共同祖先节点越靠下（分化时间越晚），亲缘关系越近；分支长度代表演化距离 / 进化速率。

第一步分化：灵长类（人）和啮齿类（鼠）的共同祖先，先发生了一次物种分化，形成了两个独立的演化支（对应图中 Taxon: Primates and Rodents -84 MYA 这个

节点，标注了约 8400 万年前的分化时间）。

第二步分化：在啮齿类分支内部，小鼠和大鼠的共同祖先，再发生一次物种分化，形成了小鼠和大鼠两个独立物种（这个节点远晚于人和啮齿类的分化节点）。



Search

All species for

Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

All genomes

-- Select a species --

Pig breeds

Pig reference genome and 20 additional breeds

View full list of all species

Favourite genomes

Human

GRCh38.p14

Still using GRCh37?

Mouse

GRCm39

Zebrafish

GRCz11

GRCh38 human reference gene set.

- Two new breeds of cattle have been added: UOA_Tuli_1 and UOA_Wagyu_1
- The sheep reference has been updated to ARS-UI_Ramb_v3.0
- Two new export modes are now available for Newick trees

More release news on our blog

The New Ensembl Site

The new Ensembl provides access to over 4700 genomes: http://beta.ensembl.org

There are over 4100 animal, 470 plant, and 100 fungal genomes ready to explore. We provide pangenomes for many species including human (565 haplotypes), barley (69 cultivars), and pig (27 breeds).

Genomes from projects such as Darwin Tree of Life, the Human Pangenome Reference Consortium, the Vertebrate Genomes Project and the European Reference Genome Atlas are regularly added.

We are in the process of migrating our existing prokaryotic genomes, with a target of moving all ~36,000 by July 2026. Prokaryotic releases will be phased, with batches migrating from Feb 2026 onwards.

From the summer of 2026, all new data will only be available through this new Ensembl site. Important info for programmatic access is available on this blog.

Compare genes across species

Find SNPs and other variants for my gene

Gene expression in different tissues

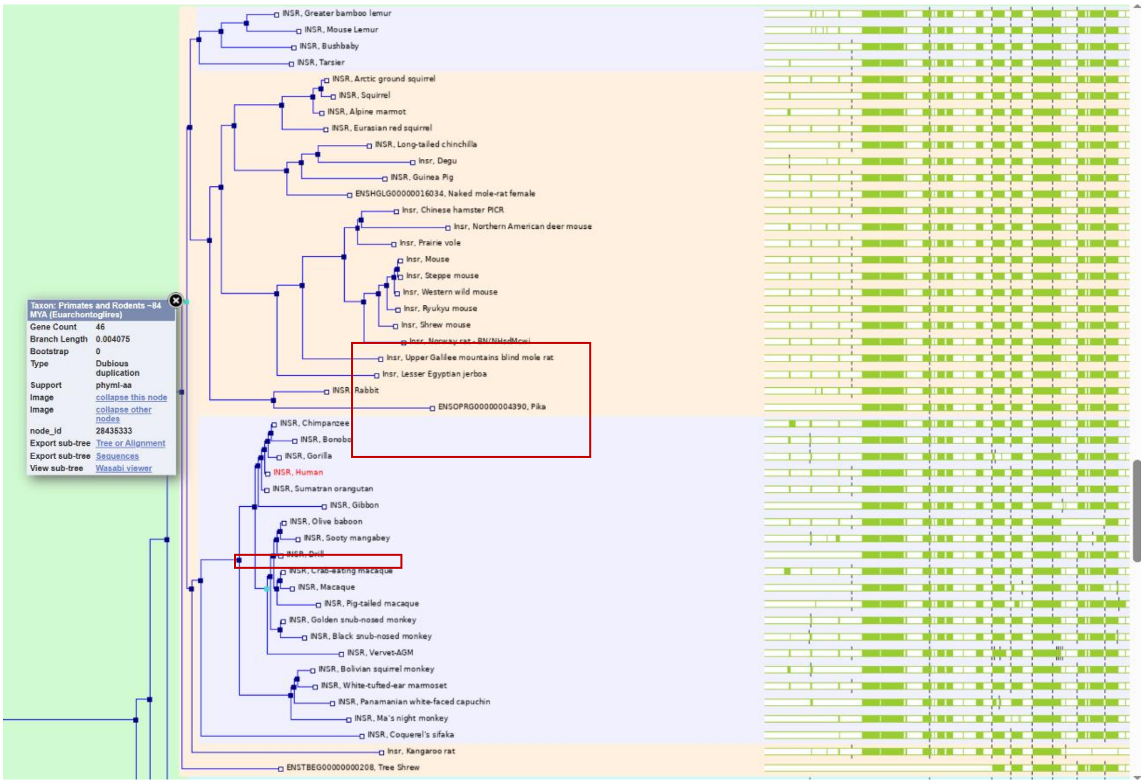
Retrieve gene sequence

Find a Data Display

Use my own data in Ensembl

Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI and our software and data are freely available.

Our acknowledgements page includes a list of current and previous funding bodies. How to cite Ensembl in your own publications.



Taxon: Primates and Rodents ~84 MYA (Euarchontoglires)

Gene Count 46

Branch Length 0.004075

Bootstrap 0

Type Dubious duplication

Support phyml-aa

Image collapse this node

Image collapse other nodes

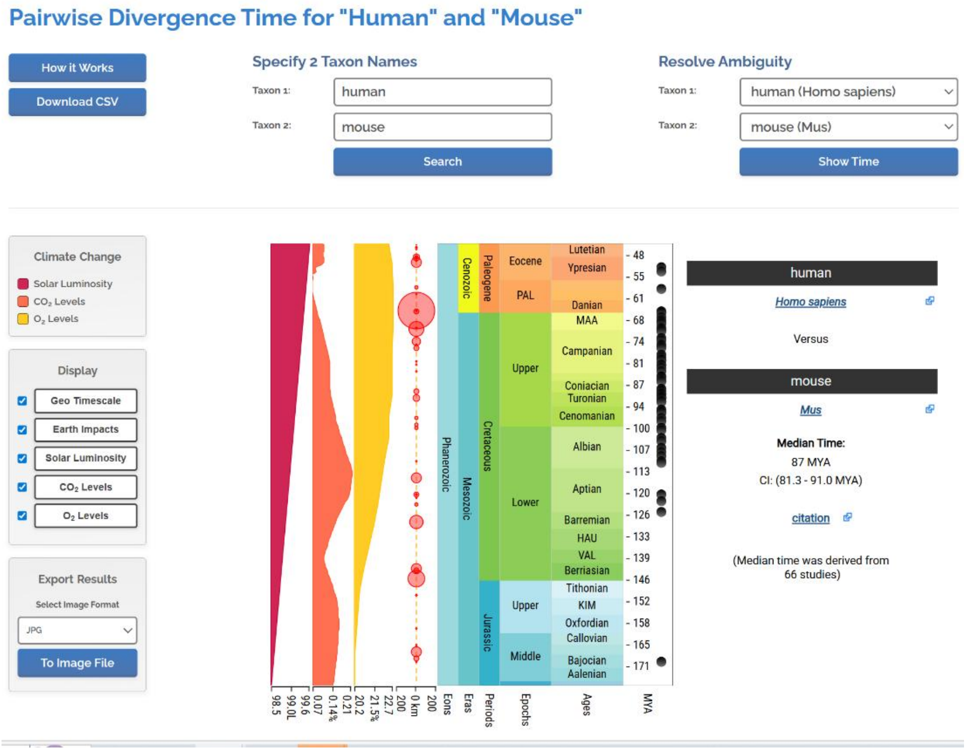
node_id 28435333

Export sub-tree Tree or Alignment

Export sub-tree Sequences

View sub-tree Wasabi viewer

① 人与小鼠：



Pairwise Divergence Time for "Human" and "Mouse"

Specify 2 Taxon Names

Taxon 1: human

Taxon 2: mouse

Resolve Ambiguity

Taxon 1: human (Homo sapiens)

Taxon 2: mouse (Mus)

human

Homo sapiens

Versus

mouse

Mus

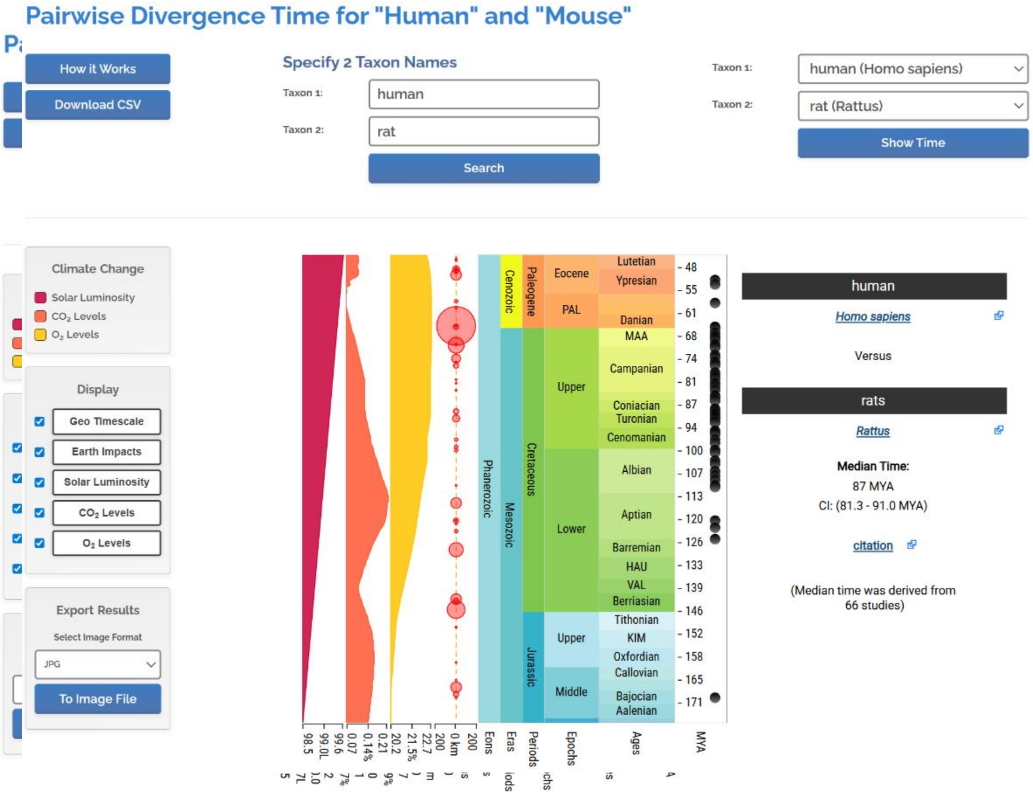
Median Time:

87 MYA

CI: (81.3 - 91.0 MYA)

citation

(Median time was derived from 66 studies)



Pairwise Divergence Time for "Human" and "Mouse"

Specify 2 Taxon Names

Taxon 1: human

Taxon 2: rat

Taxon 1: human (Homo sapiens)

Taxon 2: rat (Rattus)

human

Homo sapiens

Versus

rats

Rattus

Median Time:

87 MYA

CI: (81.3 - 91.0 MYA)

citation

(Median time was derived from 66 studies)

② 人与大鼠

③ 大鼠与小鼠：

TimeTree 物种两两比对可知：人与小鼠分歧时间点距今约 87MYA（8700 万年）；人与大鼠分歧时间点距今同样约 87MYA；小鼠与大鼠分歧时间点距今约 13.1MYA（1310 万年）。

（3）分析上述结果，进化过程中基因动态特征

重复与丢失：啮齿类（小鼠、大鼠）的 α-珠蛋白基因数量多于人类，这是谱系特异性串联重复的结果；而人类在 ζ 下游保留了 θ 假基因，啮齿类中 θ 已退化。

基因转换：所有哺乳动物 α-珠蛋白基因簇中均检测到基因转换（gene conversion），使同一簇内的不同 α 基因保持序列相似性。小鼠和大鼠的转换事件尤其频繁，导致 Hba-a1 和 Hba-a2 几乎相同。

调控差异：尽管结构保守，小鼠和人 MRE 的远端增强子组成存在差异（如小鼠有额外的内含子增强子），这可能解释两者对 MRE 缺失表型的不同（人缺失导致严重 α-地中海贫血，小鼠表型较轻）。

α-珠蛋白基因簇在脊椎动物中结构保守，但各谱系通过独立的基因重复、转换和假基因化，产生了适应各自生理需求的基因拷贝数差异。啮齿类（小鼠、大鼠）的多次重复可能与其高代谢率和红细胞生成需求有关，而人类仅保留两个近乎相同的 α 基因，足以满足氧运输需求。

（二）蛋白质与 DNA 序列比对分析及参数影响讨论

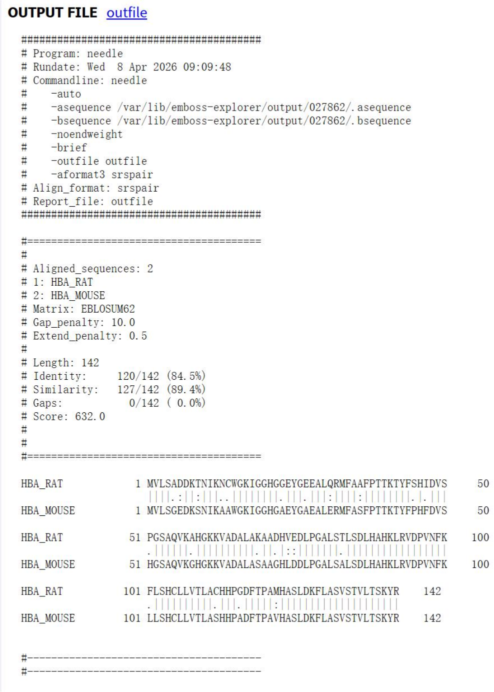
（1）提取 RefSeq 数据库中人 HBA1、小鼠 Hba1-a1、大鼠 Hba1 编码区序列进行序列比对，注意选择恰当的计分矩阵和空位罚分，分析不同空位罚分值时的比对结果。

	PAM 矩阵	BLOSUM 矩阵
数据来源	基于相似度极高（>85%）的序列家族外推	基于无空位局部比对的保守区块
数字含义	数字越大，进化距离越远	数字越小，进化距离越远
常用推荐	PAM30/70：寻找保守结构域	BLOSUM62：数据库搜索金标准（BLAST 默认）
适用场景	系统发育分析、验证同源性	BLOSUM80：近缘物种比对（小鼠 vs 大鼠）

高罚分：让比对结果紧凑，打断连续的片段，适用于寻找高度保守的局部区域。

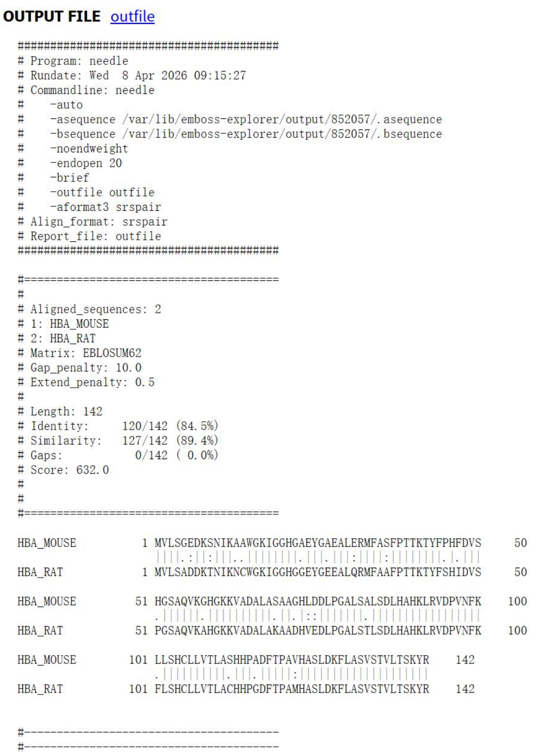
低罚分：允许引入长片段插入/缺失，适用于比对含有大段内含子的基因组 DNA。

比对需求	推荐设置
搜索近缘物种基因差异	蛋白质：BLOSUM80；DNA：+1/-3
数据库同源搜索（BLAST）	蛋白质：BLOSUM62；Gap Costs: 11/1
寻找极远缘的保守结构域	蛋白质：PAM250 或 BLOSUM45；Gap Costs: 12/2
比对含内含子的长 DNA 序列	Gap Open: 5， Gap Extend: 2 (低罚分)



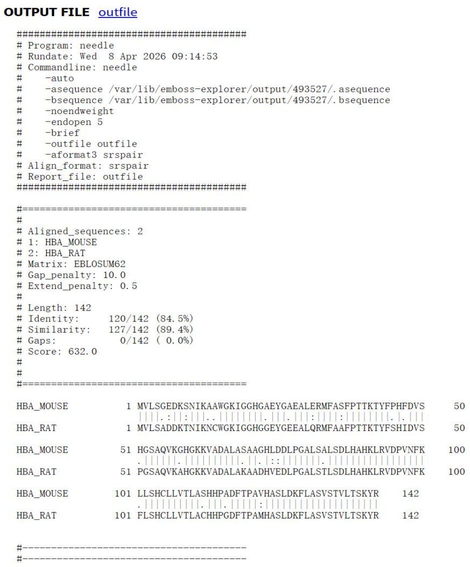
OUTPUT FILE outfile

```
########################################
# Program: needle
# Rundate: Wed  8 Apr 2026 09:09:48
# Commandline: needle
#    -auto
#    -asequence /var/lib/emboss-explorer/output/027862/.asequence
#    -bsequence /var/lib/emboss-explorer/output/027862/.bsequence
#    -noendweight
#    -brief
#    -outfile outfile
#    -aformat3 srspair
# Align_format: srspair
# Report_file: outfile
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: HBA_RAT
# 2: HBA_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     120/142 (84.5%)
# Similarity:   127/142 (89.4%)
# Gaps:           0/142 ( 0.0%)
# Score: 632.0
#
#
#=======================================

HBA_RAT            1 MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVS     50
                     ||||.:||:|||..||||||||.|||.|||:||||:||||||||.|.|||
HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50

HBA_RAT           51 PGSAQVKAHGKKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFK    100
                     .||||||.||||||||||.||.|::|||||||.|||||||||||||||||
HBA_MOUSE         51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100

HBA_RAT          101 FLSHCLLVTLACHHPGDFTPAMHASLDKFLASVSTVLTSKYR    142
                     .||||||||||.|||.|||||:||||||||||||||||||||
HBA_MOUSE        101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142


#---------------------------------------
#---------------------------------------
```



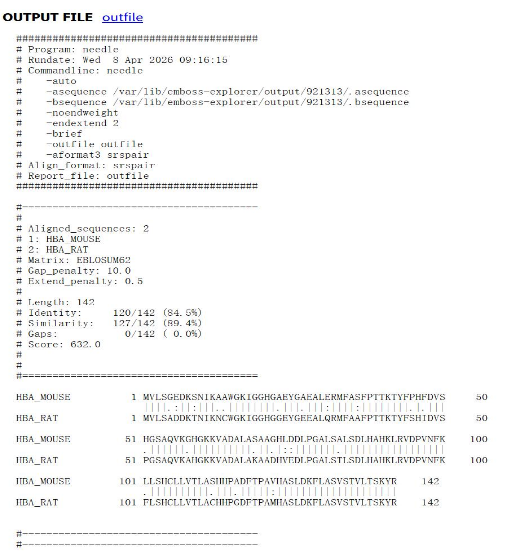
OUTPUT FILE outfile

```
########################################
# Program: needle
# Rundate: Wed  8 Apr 2026 09:15:27
# Commandline: needle
#    -auto
#    -asequence /var/lib/emboss-explorer/output/852057/.asequence
#    -bsequence /var/lib/emboss-explorer/output/852057/.bsequence
#    -noendweight
#    -endopen 20
#    -brief
#    -outfile outfile
#    -aformat3 srspair
# Align_format: srspair
# Report_file: outfile
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: HBA_MOUSE
# 2: HBA_RAT
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     120/142 (84.5%)
# Similarity:   127/142 (89.4%)
# Gaps:           0/142 ( 0.0%)
# Score: 632.0
#
#
#=======================================

HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50
                     ||||.:||:|||..||||||||.|||.|||:||||:||||||||.|.|||
HBA_RAT            1 MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVS     50

HBA_MOUSE         51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100
                     .||||||.||||||||||.||.|::|||||||.|||||||||||||||||
HBA_RAT           51 PGSAQVKAHGKKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFK    100

HBA_MOUSE        101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142
                     .||||||||||.|||.|||||:||||||||||||||||||||
HBA_RAT          101 FLSHCLLVTLACHHPGDFTPAMHASLDKFLASVSTVLTSKYR    142


#---------------------------------------
#---------------------------------------
```



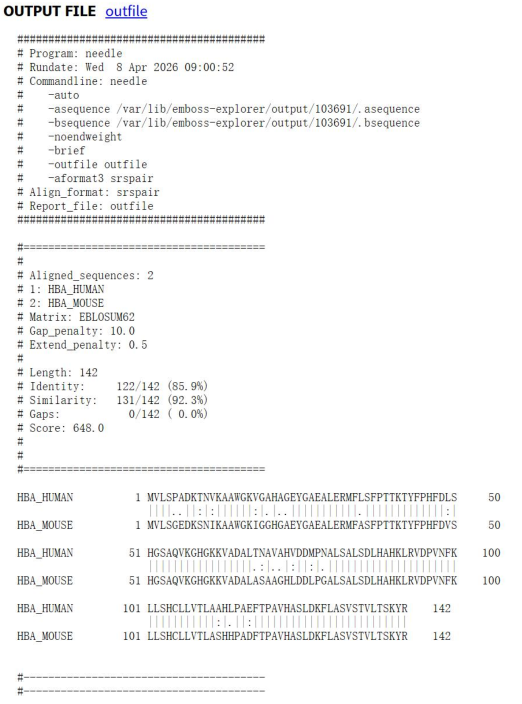
OUTPUT FILE outfile

```
########################################
# Program: needle
# Rundate: Wed  8 Apr 2026 09:14:53
# Commandline: needle
#    -auto
#    -asequence /var/lib/emboss-explorer/output/493527/.asequence
#    -bsequence /var/lib/emboss-explorer/output/493527/.bsequence
#    -noendweight
#    -endopen 5
#    -brief
#    -outfile outfile
#    -aformat3 srspair
# Align_format: srspair
# Report_file: outfile
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: HBA_MOUSE
# 2: HBA_RAT
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     120/142 (84.5%)
# Similarity:   127/142 (89.4%)
# Gaps:           0/142 ( 0.0%)
# Score: 632.0
#
#
#=======================================

HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50
                     ||||.:||:|||..||||||||.|||.|||:||||:||||||||.|.|||
HBA_RAT            1 MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVS     50

HBA_MOUSE         51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100
                     .||||||.||||||||||.||.|::|||||||.|||||||||||||||||
HBA_RAT           51 PGSAQVKAHGKKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFK    100

HBA_MOUSE        101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142
                     .||||||||||.|||.|||||:||||||||||||||||||||
HBA_RAT          101 FLSHCLLVTLACHHPGDFTPAMHASLDKFLASVSTVLTSKYR    142


#---------------------------------------
#---------------------------------------
```



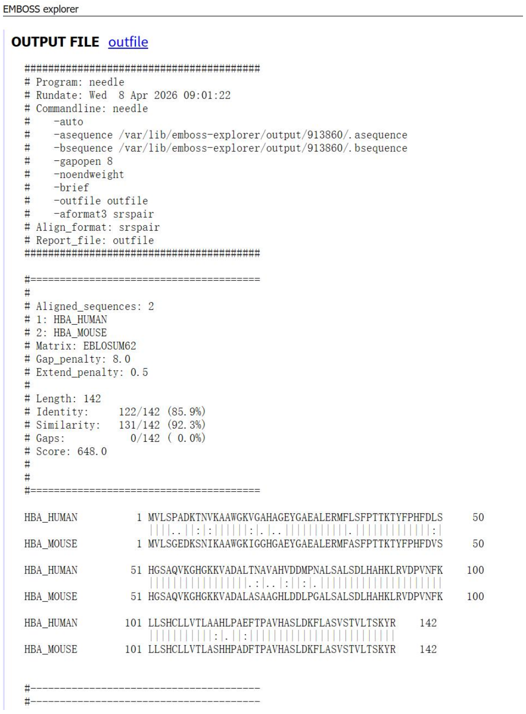
OUTPUT FILE outfile

```
########################################
# Program: needle
# Rundate: Wed  8 Apr 2026 09:16:15
# Commandline: needle
#    -auto
#    -asequence /var/lib/emboss-explorer/output/921313/.asequence
#    -bsequence /var/lib/emboss-explorer/output/921313/.bsequence
#    -noendweight
#    -endextend 2
#    -brief
#    -outfile outfile
#    -aformat3 srspair
# Align_format: srspair
# Report_file: outfile
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: HBA_MOUSE
# 2: HBA_RAT
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     120/142 (84.5%)
# Similarity:   127/142 (89.4%)
# Gaps:           0/142 ( 0.0%)
# Score: 632.0
#
#
#=======================================

HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50
                     ||||.:||:|||..||||||||.|||.|||:||||:||||||||.|.|||
HBA_RAT            1 MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVS     50

HBA_MOUSE         51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100
                     .||||||.||||||||||.||.|::|||||||.|||||||||||||||||
HBA_RAT           51 PGSAQVKAHGKKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFK    100

HBA_MOUSE        101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142
                     .||||||||||.|||.|||||:||||||||||||||||||||
HBA_RAT          101 FLSHCLLVTLACHHPGDFTPAMHASLDKFLASVSTVLTSKYR    142


#---------------------------------------
#---------------------------------------
```



OUTPUT FILE outfile

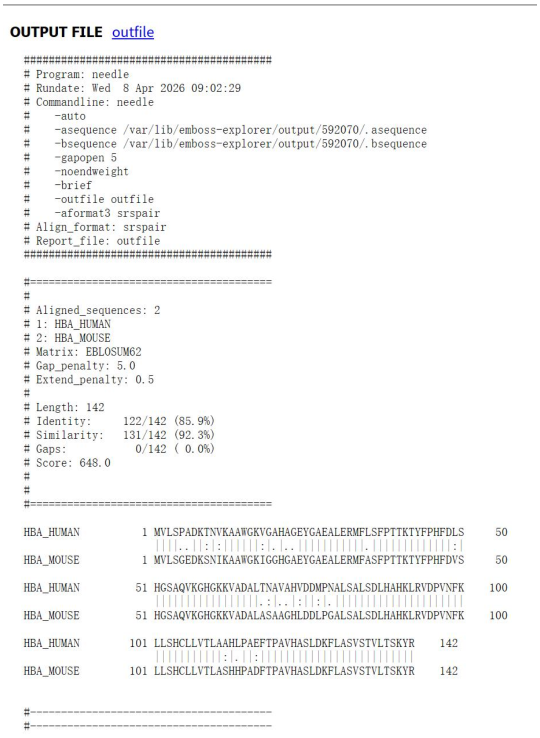
```
########################################
# Program: needle
# Rundate: Wed  8 Apr 2026 09:00:52
# Commandline: needle
#    -auto
#    -asequence /var/lib/emboss-explorer/output/103691/.asequence
#    -bsequence /var/lib/emboss-explorer/output/103691/.bsequence
#    -noendweight
#    -brief
#    -outfile outfile
#    -aformat3 srspair
# Align_format: srspair
# Report_file: outfile
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBA_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     122/142 (85.9%)
# Similarity:   131/142 (92.3%)
# Gaps:           0/142 ( 0.0%)
# Score: 648.0
#
#
#=======================================

HBA_HUMAN          1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS     50
                     ||||..||:|:||||||:|.|..|||||||||||.|||||||||||||:|
HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50

HBA_HUMAN         51 HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK    100
                     |||||||||||||||||.:|..|:||:|.|||||||||||||||||||||
HBA_MOUSE         51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100

HBA_HUMAN        101 LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     |||||||||||:|.||:|||||||||||||||||||||||||
HBA_MOUSE        101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142


#---------------------------------------
#---------------------------------------
```



EMBOSS explorer

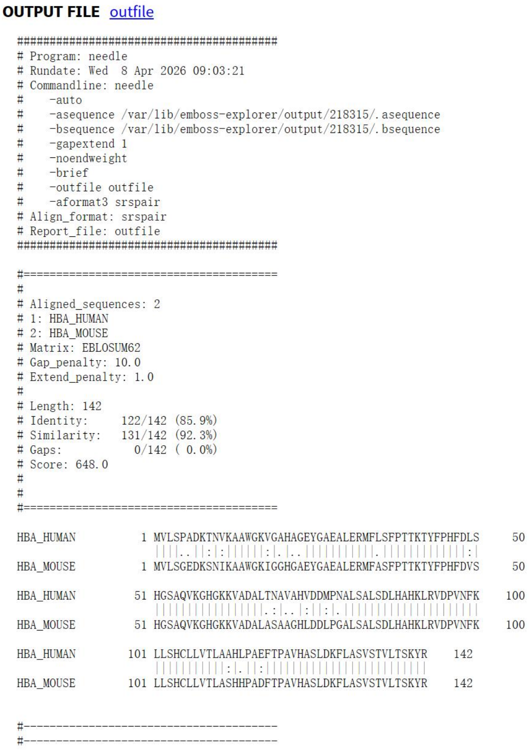
OUTPUT FILE outfile

```
########################################
# Program: needle
# Rundate: Wed  8 Apr 2026 09:01:22
# Commandline: needle
#    -auto
#    -asequence /var/lib/emboss-explorer/output/913860/.asequence
#    -bsequence /var/lib/emboss-explorer/output/913860/.bsequence
#    -gapopen 8
#    -noendweight
#    -brief
#    -outfile outfile
#    -aformat3 srspair
# Align_format: srspair
# Report_file: outfile
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBA_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     122/142 (85.9%)
# Similarity:   131/142 (92.3%)
# Gaps:           0/142 ( 0.0%)
# Score: 648.0
#
#
#=======================================

HBA_HUMAN          1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS     50
                     ||||..||:|:||||||:|.|..|||||||||||.|||||||||||||:|
HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50

HBA_HUMAN         51 HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK    100
                     |||||||||||||||||.:|..|:||:|.|||||||||||||||||||||
HBA_MOUSE         51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100

HBA_HUMAN        101 LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     |||||||||||:|.||:|||||||||||||||||||||||||
HBA_MOUSE        101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142


#---------------------------------------
#---------------------------------------
```



OUTPUT FILE outfile

```
########################################
# Program: needle
# Rundate: Wed  8 Apr 2026 09:02:29
# Commandline: needle
#    -auto
#    -asequence /var/lib/emboss-explorer/output/592070/.asequence
#    -bsequence /var/lib/emboss-explorer/output/592070/.bsequence
#    -gapopen 5
#    -noendweight
#    -brief
#    -outfile outfile
#    -aformat3 srspair
# Align_format: srspair
# Report_file: outfile
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBA_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 5.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     122/142 (85.9%)
# Similarity:   131/142 (92.3%)
# Gaps:           0/142 ( 0.0%)
# Score: 648.0
#
#
#=======================================

HBA_HUMAN          1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS     50
                     ||||..||:|:||||||:|.|..|||||||||||.|||||||||||||:|
HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50

HBA_HUMAN         51 HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK    100
                     |||||||||||||||||.:|..|:||:|.|||||||||||||||||||||
HBA_MOUSE         51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100

HBA_HUMAN        101 LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     |||||||||||:|.||:|||||||||||||||||||||||||
HBA_MOUSE        101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142


#---------------------------------------
#---------------------------------------
```



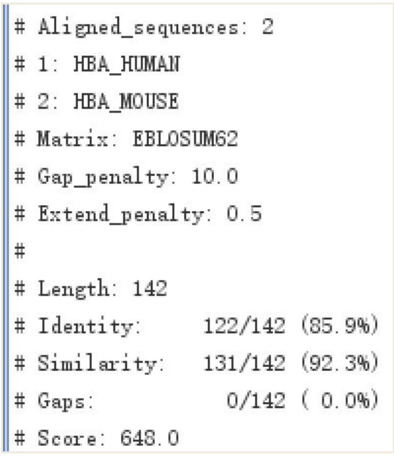
OUTPUT FILE outfile

```
########################################
# Program: needle
# Rundate: Wed  8 Apr 2026 09:03:21
# Commandline: needle
#    -auto
#    -asequence /var/lib/emboss-explorer/output/218315/.asequence
#    -bsequence /var/lib/emboss-explorer/output/218315/.bsequence
#    -gapextend 1
#    -noendweight
#    -brief
#    -outfile outfile
#    -aformat3 srspair
# Align_format: srspair
# Report_file: outfile
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBA_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 1.0
#
# Length: 142
# Identity:     122/142 (85.9%)
# Similarity:   131/142 (92.3%)
# Gaps:           0/142 ( 0.0%)
# Score: 648.0
#
#
#=======================================

HBA_HUMAN          1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS     50
                     ||||..||:|:||||||:|.|..|||||||||||.|||||||||||||:|
HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50

HBA_HUMAN         51 HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK    100
                     |||||||||||||||||.:|..|:||:|.|||||||||||||||||||||
HBA_MOUSE         51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100

HBA_HUMAN        101 LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     |||||||||||:|.||:|||||||||||||||||||||||||
HBA_MOUSE        101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142


#---------------------------------------
#---------------------------------------
```

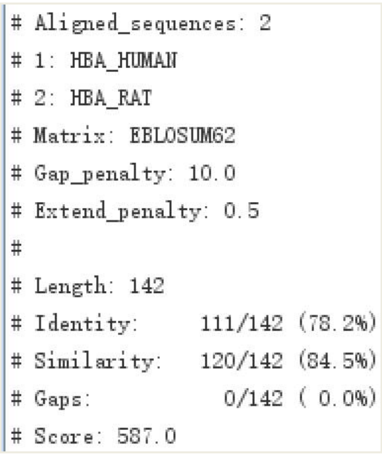
（三）UniProt 高级检索实践与数据库交叉链接应用

（1）从 UniProt 数据库中提取人、小鼠、大鼠血红蛋白 alpha 亚基蛋白质序列，进行双序列全局比对。分析比对结果，说明所得结果和预期不符的原因和进一步分析思路。

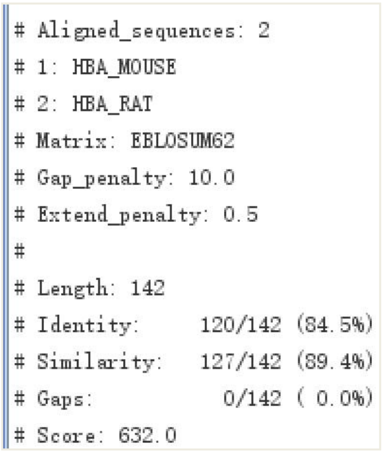
<table>
<tr><td>>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens OX=9606 GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR</td></tr>
<tr><td>>sp|P01942|HBA_MOUSE Hemoglobin subunit alpha OS=Mus musculus OX=10090 GN=Hba PE=1 SV=2
MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
AVHASLDKFLASVSTVLTSKYR</td></tr>
<tr><td>>sp|P01946|HBA_RAT Hemoglobin subunit alpha-1/2 OS=Rattus norvegicus OX=10116 GN=Hba1 PE=1 SV=3
MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG
KKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFKFLSHCLLVTLACHHPGDFTP
AMHASLDKFLASVSTVLTSKYR</td></tr>
</table>



```
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBA_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     122/142 (85.9%)
# Similarity:   131/142 (92.3%)
# Gaps:           0/142 ( 0.0%)
# Score: 648.0
```



```
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBA_RAT
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     111/142 (78.2%)
# Similarity:   120/142 (84.5%)
# Gaps:           0/142 ( 0.0%)
# Score: 587.0
```



```
# Aligned_sequences: 2
# 1: HBA_MOUSE
# 2: HBA_RAT
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     120/142 (84.5%)
# Similarity:   127/142 (89.4%)
# Gaps:           0/142 ( 0.0%)
# Score: 632.0
```

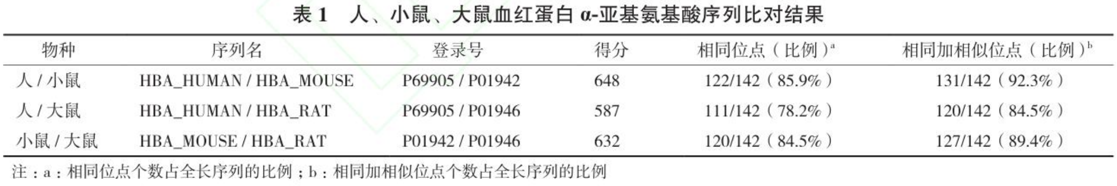


表 1　人、小鼠、大鼠血红蛋白 α-亚基氨基酸序列比对结果

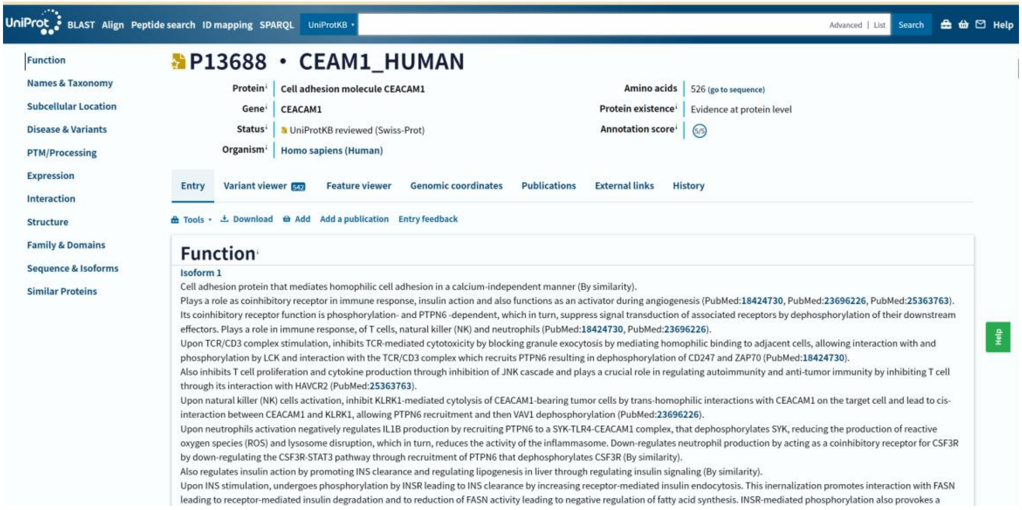
物种	序列名	登录号	得分	相同位点（比例）[a]	相同加相似位点（比例）[b]
人 / 小鼠	HBA_HUMAN / HBA_MOUSE	P69905 / P01942	648	122/142（85.9%）	131/142（92.3%）
人 / 大鼠	HBA_HUMAN / HBA_RAT	P69905 / P01946	587	111/142（78.2%）	120/142（84.5%）
小鼠 / 大鼠	HBA_MOUSE / HBA_RAT	P01942 / P01946	632	120/142（84.5%）	127/142（89.4%）

注：a：相同位点个数占全长序列的比例；b：相同加相似位点个数占全长序列的比例

Emboss Needle 比对结果如上

经查阅，比对结果与课程文献内给出结果相同

（2）以人癌胚抗原 CEAM1_HUMAN 的恒定结构域 Ig-like C2-type 1（145-232）搜索 Swiss-Prot 中人的 CEA 家族成员， 分析搜索结果，说明最大分值（Max Score）和总体分值（Total Score）的含义。



P13688 • CEAM1_HUMAN

Protein: Cell adhesion molecule CEACAM1
Gene: CEACAM1
Status: UniProtKB reviewed (Swiss-Prot)
Organism: Homo sapiens (Human)
Amino acids: 526 (go to sequence)
Protein existence: Evidence at protein level
Annotation score: 5/5

Function

Isoform 1

Cell adhesion protein that mediates homophilic cell adhesion in a calcium-independent manner (By similarity).

Plays a role as coinhibitory receptor in immune response, insulin action and also functions as an activator during angiogenesis (PubMed:18424730, PubMed:23696226, PubMed:25363763). Its coinhibitory receptor function is phosphorylation- and PTPN6 -dependent, which in turn, suppress signal transduction of associated receptors by dephosphorylation of their downstream effectors. Plays a role in immune response, of T cells, natural killer (NK) and neutrophils (PubMed:18424730, PubMed:23696226).

Upon TCR/CD3 complex stimulation, inhibits TCR-mediated cytotoxicity by blocking granule exocytosis by mediating homophilic binding to adjacent cells, allowing interaction with and phosphorylation by LCK and interaction with the TCR/CD3 complex which recruits PTPN6 resulting in dephosphorylation of CD247 and ZAP70 (PubMed:18424730).

Also inhibits T cell proliferation and cytokine production through inhibition of JNK cascade and plays a crucial role in regulating autoimmunity and anti-tumor immunity by inhibiting T cell through its interaction with HAVCR2 (PubMed:25363763).

Upon natural killer (NK) cells activation, inhibit KLRK1-mediated cytolysis of CEACAM1-bearing tumor cells by trans-homophilic interactions with CEACAM1 on the target cell and lead to cis-interaction between CEACAM1 and KLRK1, allowing PTPN6 recruitment and then VAV1 dephosphorylation (PubMed:23696226).

Upon neutrophils activation negatively regulates IL1B production by recruiting PTPN6 to a SYK-TLR4-CEACAM1 complex, that dephosphorylates SYK, reducing the production of reactive oxygen species (ROS) and lysosome disruption, which in turn, reduces the activity of the inflammasome. Down-regulates neutrophil production by acting as a coinhibitory receptor for CSF3R by down-regulating the CSF3R-STAT3 pathway through recruitment of PTPN6 that dephosphorylates CSF3R (By similarity).

Also regulates insulin action by promoting INS clearance and regulating lipogenesis in liver through regulating insulin signaling (By similarity).

Upon INS stimulation, undergoes phosphorylation by INSR leading to INS clearance by increasing receptor-mediated insulin endocytosis. This inernalization promotes interaction with FASN leading to receptor-mediated insulin degradation and to reduction of FASN activity leading to negative regulation of fatty acid synthesis. INSR-mediated phosphorylation also provokes a

Status: UniProtKB reviewed (Swiss-Prot)

Organism: Homo sapiens (Human)

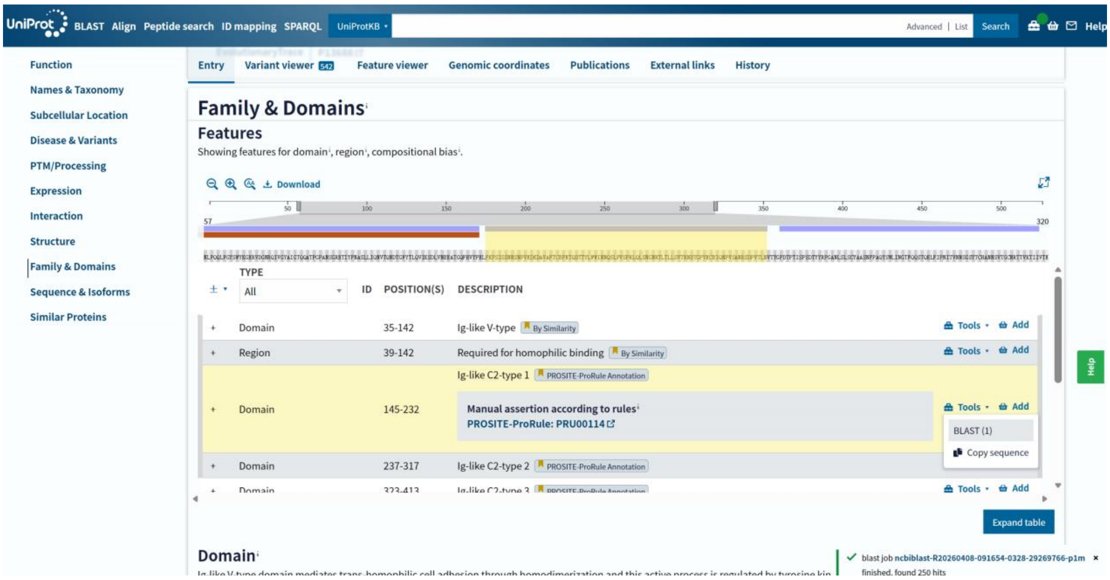
Annotation score: 5/5

Feature viewer

DOMAIN 145-232

Description
Ig-like C2-type 1

Evidence
Sequence Analysis: PRU00114 (PROSITE-ProRule)



Family & Domains

Features

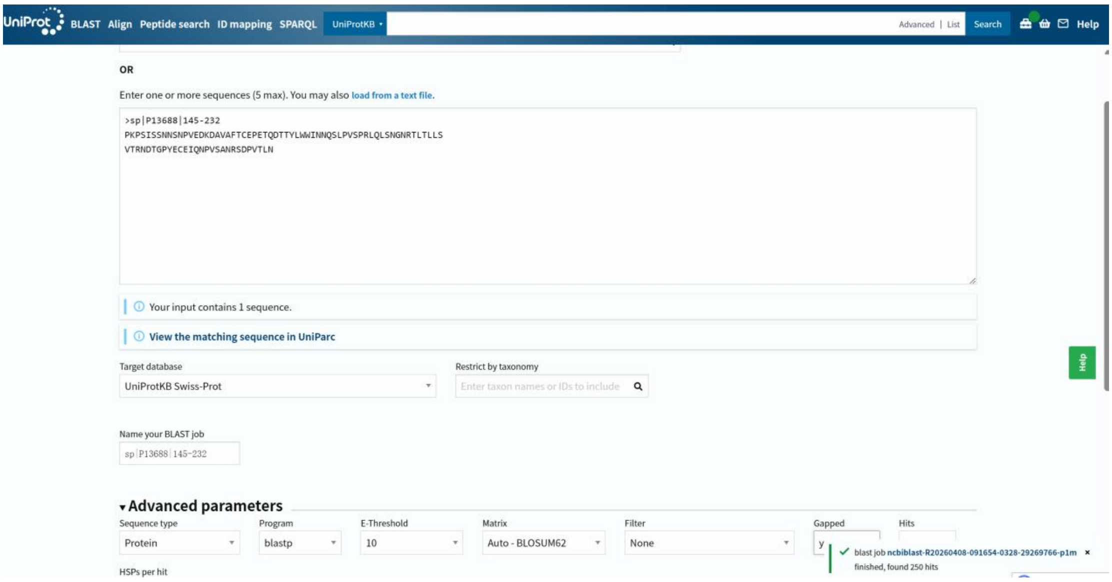
Showing features for domain, region, compositional bias.

TYPE	ID	POSITION(S)	DESCRIPTION
Domain		35-142	Ig-like V-type By Similarity
Region		39-142	Required for homophilic binding By Similarity
Domain		145-232	Ig-like C2-type 1 PROSITE-ProRule Annotation Manual assertion according to rules PROSITE-ProRule: PRU00114
Domain		237-317	Ig-like C2-type 2 PROSITE-ProRule Annotation
Domain		323-413	Ig-like C2-type 3 PROSITE-ProRule Annotation

✓ blast job ncbiblast-R20260408-091654-0328-29269766-p1m finished. found 250 hits

Domain

Ig-like V-type domain mediates trans-homophilic cell adhesion through homodimerization and this active process is regulated by tyrosine kin



OR

Enter one or more sequences (5 max). You may also load from a text file.

```
>sp|P13688|145-232
PKPSISSNNSNPVEDKDAVAFTCEPETQDTTYLWWINNQSLPVSPRLQLSNGNRTLTLLS
VTRNDTGPYECEIQNPVSANRSDPVTLN
```

ⓘ Your input contains 1 sequence.

ⓘ View the matching sequence in UniParc

Target database: UniProtKB Swiss-Prot

Restrict by taxonomy: Enter taxon names or IDs to include

Name your BLAST job: sp|P13688|145-232

▾ Advanced parameters

Sequence type: Protein

Program: blastp

E-Threshold: 10

Matrix: Auto - BLOSUM62

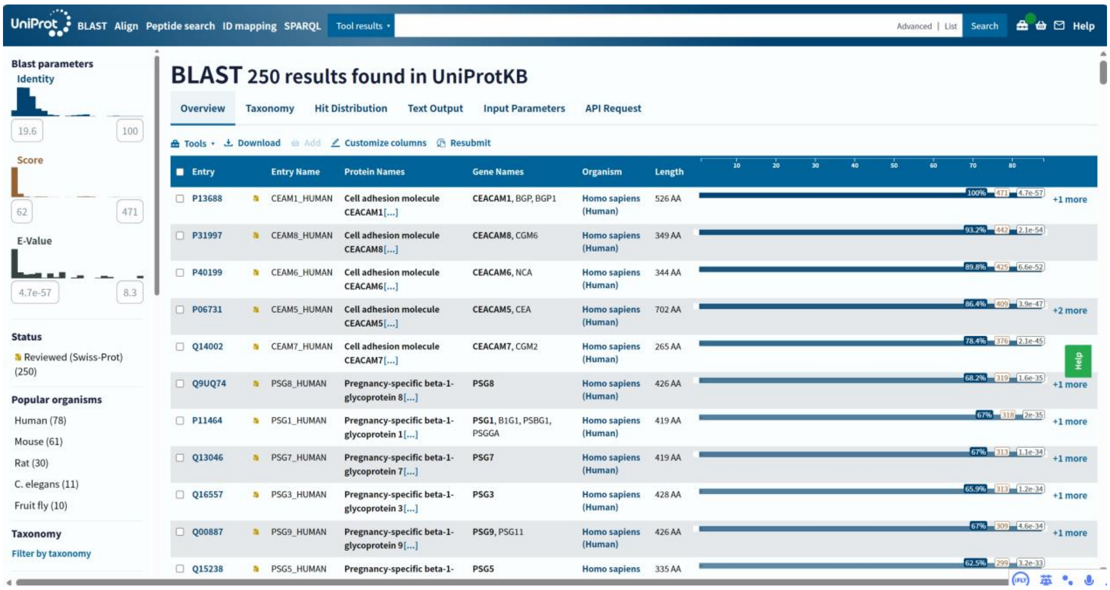
Filter: None

Gapped: y

Hits

HSPs per hit

✓ blast job ncbiblast-R20260408-091654-0328-29269766-p1m finished, found 250 hits



Blast parameters

Identity 19.6 100

Score 62 471

E-Value 4.7e-57 8.3

Status

Reviewed (Swiss-Prot) (250)

Popular organisms

Human (78)

Mouse (61)

Rat (30)

C. elegans (11)

Fruit fly (10)

Taxonomy

Filter by taxonomy

BLAST 250 results found in UniProtKB

☐	Entry	Entry Name	Protein Names	Gene Names	Organism	Length				
☐	P13688	CEAM1_HUMAN	Cell adhesion molecule CEACAM1[...]	CEACAM1, BGP, BGP1	Homo sapiens (Human)	526 AA	100%	471	4.7e-57	+1 more
☐	P31997	CEAM8_HUMAN	Cell adhesion molecule CEACAM8[...]	CEACAM8, CGM6	Homo sapiens (Human)	349 AA	93.2%	442	2.1e-54	
☐	P40199	CEAM6_HUMAN	Cell adhesion molecule CEACAM6[...]	CEACAM6, NCA	Homo sapiens (Human)	344 AA	89.8%	425	6.6e-52	
☐	P06731	CEAM5_HUMAN	Cell adhesion molecule CEACAM5[...]	CEACAM5, CEA	Homo sapiens (Human)	702 AA	86.4%	409	3.9e-47	+2 more
☐	Q14002	CEAM7_HUMAN	Cell adhesion molecule CEACAM7[...]	CEACAM7, CGM2	Homo sapiens (Human)	265 AA	78.4%	376	2.1e-45	
☐	Q9UQ74	PSG8_HUMAN	Pregnancy-specific beta-1-glycoprotein 8[...]	PSG8	Homo sapiens (Human)	426 AA	68.2%	319	1.6e-35	+1 more
☐	P11464	PSG1_HUMAN	Pregnancy-specific beta-1-glycoprotein 1[...]	PSG1, B1G1, PSBG1, PSGGA	Homo sapiens (Human)	419 AA	67%	318	2e-35	+1 more
☐	Q13046	PSG7_HUMAN	Pregnancy-specific beta-1-glycoprotein 7[...]	PSG7	Homo sapiens (Human)	419 AA	67%	313	1.1e-34	+1 more
☐	Q16557	PSG3_HUMAN	Pregnancy-specific beta-1-glycoprotein 3[...]	PSG3	Homo sapiens (Human)	428 AA	65.9%	313	1.2e-34	+1 more
☐	Q00887	PSG9_HUMAN	Pregnancy-specific beta-1-glycoprotein 9[...]	PSG9, PSG11	Homo sapiens (Human)	426 AA	67%	309	4.6e-34	+1 more
☐	Q15238	PSG5_HUMAN	Pregnancy-specific beta-1-	PSG5	Homo sapiens	335 AA	62.5%	299	3.2e-33	

1. Max Score（最大分值）

含义：在本次查询序列与该条目的所有可能局部比对中，得分最高的那个片段所获得的分值。

简单理解：这个分值代表“这段序列中与查询序列最相似的那一块有多像”。

对于我们的查询（单个结构域，长度约 88 个氨基酸），绝大多数命中条目都只有 一个 显著的比对区域（即 Ig-like C2-type 结构域本身）。因此 Max Score 通常等于 Total Score。

如果某个蛋白质有两个或更多相似的结构域（例如 CEA 家族某些成员有多个 Ig 结构域），BLAST 可能会分别比对到不同区域，此时 Max Score 仍然只取 最好那个区域 的分值。

2. Total Score（总体分值）

含义：所有比对上区域的得分总和。

当查询序列与目标序列 只有一个连续匹配区域 时，Total Score = Max Score。

如果查询序列与目标序列 有多个不连续的相似片段（例如查询序列较长且包含两个独立结构域），BLAST 会分别计算每个片段的得分并相加，得到 Total Score。

在我们这个例子中，因为查询序列只是一个连续的结构域，而 CEA 家族的每个成员通常只含 一个 与该结构域同源的区域，所以 Total Score 几乎总是等于 Max Score。

3. 分值高低的意义

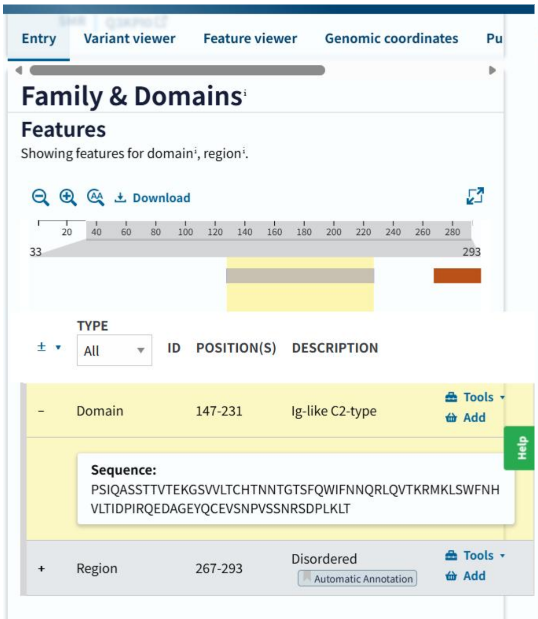
分值越高，表示比对越可靠、序列相似性越高。

250 分左右：完全一致（自身比对）或极高相似性。

140–200 分：同源结构域，但有一定氨基酸差异（例如 CEACAM 家族成员之间约 40–60% 的序列同一性）。

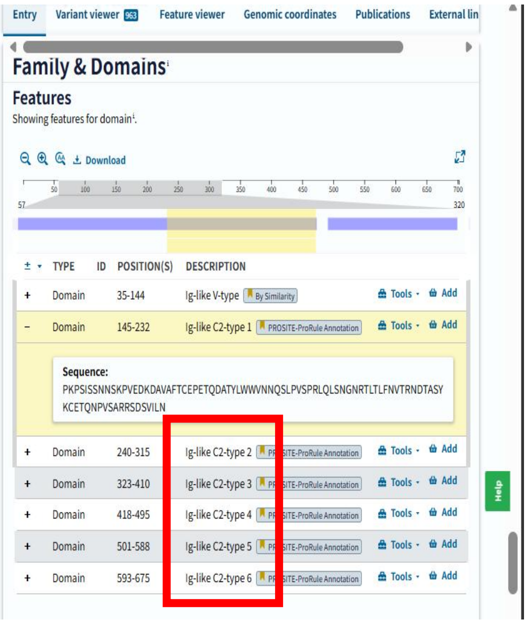
<50 分：可能为随机噪音或远缘同源。

（3）以人癌胚抗原 CEA21_HUMAN 的恒定结构域 Ig-like C2-type （147-231）与 CEAM5_HUMA 进行双序列比对，分析比对结果， 说明 Blast 与 EMBOSS 软件包中程序 water 运行结果的差别。



TYPE	ID	POSITION(S)	DESCRIPTION
− Domain		147-231	Ig-like C2-type
+ Region		267-293	Disordered Automatic Annotation

Sequence:
PSIQASSTTVTEKGSVVLTCHTNNTGTSFQWIFNNQRLQVTKRMKLSWFNH
VLTIDPIRQEDAGEYQCEVSNPVSSNRSDPLKLT

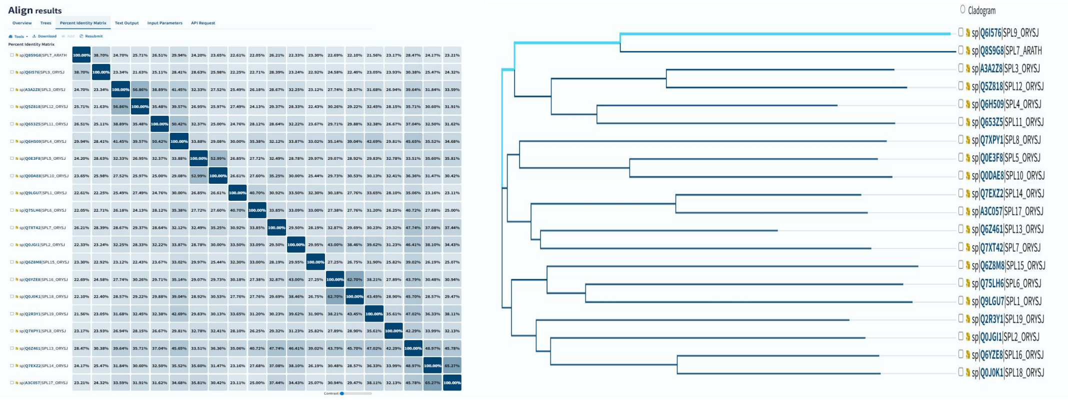


TYPE	ID	POSITION(S)	DESCRIPTION
+ Domain		35-144	Ig-like V-type By Similarity
− Domain		145-232	Ig-like C2-type 1 PROSITE-ProRule Annotation
+ Domain		240-315	Ig-like C2-type 2 PROSITE-ProRule Annotation
+ Domain		323-410	Ig-like C2-type 3 PROSITE-ProRule Annotation
+ Domain		418-495	Ig-like C2-type 4 PROSITE-ProRule Annotation
+ Domain		501-588	Ig-like C2-type 5 PROSITE-ProRule Annotation
+ Domain		593-675	Ig-like C2-type 6 PROSITE-ProRule Annotation

Sequence:
PKPSISSNNSKPVEDKDAVAFTCEPETQDATYLWWVNNQSLPVSPRLQLSNGNRTLTLFNVTRNDTASY
KCETQNPVSARRSDSVILN

（4）以蛋白名“Squamosa Promoter-binding-like protein”检索 UniProt 数据库，

下载拟南芥 SPL7 转录因子和 19 个粳稻（分类学登录号 39947）转录因子，利用 Blast 找出水稻中与拟南芥中 SPL7_ARATH 最可能的直系同源蛋白。SPL9_ORYSJ



Percent Identity Matrix

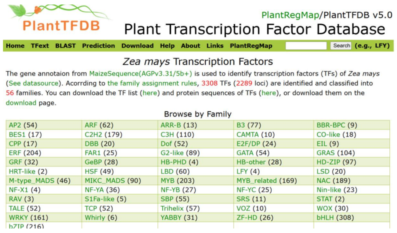
sp|Q8S9G8|SPL7_ARATH	100.00%	38.70%	24.70%	25.71%	26.51%	29.94%	24.20%	23.65%	22.61%	22.05%	26.21%	22.33%	23.30%	22.69%	22.10%	21.56%	23.17%	28.47%	24.17%	23.21%
sp|Q6I576|SPL9_ORYSJ	38.70%	100.00%	23.34%	21.63%	25.11%	28.41%	28.63%	25.98%	22.25%	22.71%	28.39%	23.24%	22.92%	24.58%	22.40%	23.05%	23.93%	30.38%	25.47%	24.32%
sp|A3A2Z8|SPL3_ORYSJ	24.70%	23.34%	100.00%	56.86%	38.89%	41.45%	32.33%	27.52%	25.49%	26.18%	28.67%	32.25%	23.12%	27.74%	28.57%	31.68%	26.94%	39.64%	31.84%	33.59%
sp|Q5Z818|SPL12_ORYSJ	25.71%	21.63%	56.86%	100.00%	35.48%	39.57%	26.95%	25.97%	27.49%	24.13%	29.37%	28.33%	22.43%	30.26%	29.22%	32.45%	28.15%	35.71%	30.60%	31.91%
sp|Q653Z5|SPL11_ORYSJ	26.51%	25.11%	38.89%	35.48%	100.00%	50.42%	32.37%	25.00%	24.76%	28.12%	28.64%	32.22%	23.67%	29.71%	29.88%	32.38%	26.67%	37.04%	32.50%	31.62%
sp|Q6H509|SPL4_ORYSJ	29.94%	28.41%	41.45%	39.57%	50.42%	100.00%	33.88%	29.08%	30.00%	35.38%	32.12%	33.87%	33.02%	35.14%	39.04%	42.69%	29.81%	45.65%	35.52%	34.68%
sp|Q0E3F8|SPL5_ORYSJ	24.20%	28.63%	32.33%	26.95%	32.37%	33.88%	100.00%	52.99%	26.85%	27.72%	32.49%	28.78%	29.97%	29.07%	28.92%	29.83%	32.78%	33.51%	35.60%	35.81%
sp|Q0DAE8|SPL10_ORYSJ	23.65%	25.98%	27.52%	25.97%	25.00%	29.08%	52.99%	100.00%	26.61%	27.60%	35.25%	30.00%	25.44%	29.73%	30.53%	30.13%	32.41%	36.36%	31.47%	30.42%
sp|Q9LGU7|SPL1_ORYSJ	22.61%	22.25%	25.49%	27.49%	24.76%	30.00%	26.85%	26.61%	100.00%	40.70%	30.92%	33.50%	32.30%	30.18%	27.76%	33.65%	28.10%	35.06%	23.16%	23.11%
sp|Q75LH6|SPL6_ORYSJ	22.05%	22.71%	26.18%	24.13%	28.12%	35.38%	27.72%	27.60%	40.70%	100.00%	33.85%	33.09%	33.00%	27.38%	27.76%	31.20%	26.25%	40.72%	27.68%	25.00%
sp|Q7XT42|SPL7_ORYSJ	26.21%	28.39%	28.67%	29.37%	28.64%	32.12%	32.49%	35.25%	30.92%	33.85%	100.00%	29.50%	28.19%	32.87%	29.69%	30.23%	29.32%	47.74%	37.08%	37.44%
sp|Q0JGI1|SPL2_ORYSJ	22.33%	23.24%	32.25%	28.33%	32.22%	33.87%	28.78%	30.00%	33.50%	33.09%	29.50%	100.00%	29.95%	43.00%	38.46%	39.62%	31.23%	46.41%	38.10%	34.43%
sp|Q6Z8M8|SPL15_ORYSJ	23.30%	22.92%	23.12%	22.43%	23.67%	33.02%	29.97%	25.44%	32.30%	33.00%	28.19%	29.95%	100.00%	27.25%	26.75%	31.90%	25.82%	39.02%	26.19%	25.07%
sp|Q6YZE8|SPL16_ORYSJ	22.69%	24.58%	27.74%	30.26%	29.71%	35.14%	29.07%	29.73%	30.18%	27.38%	32.87%	43.00%	27.25%	100.00%	62.70%	38.21%	27.89%	43.79%	30.48%	30.94%
sp|Q0J0K1|SPL18_ORYSJ	22.10%	22.40%	28.57%	29.22%	29.88%	39.04%	28.92%	30.53%	27.76%	27.76%	29.69%	38.46%	26.75%	62.70%	100.00%	43.45%	28.90%	45.70%	28.57%	29.47%
sp|Q2R3Y1|SPL19_ORYSJ	21.56%	23.05%	31.68%	32.45%	32.38%	42.69%	29.83%	30.13%	33.65%	31.20%	30.23%	39.62%	31.90%	38.21%	43.45%	100.00%	35.61%	47.02%	36.33%	38.11%
sp|Q7XPY1|SPL8_ORYSJ	23.17%	23.93%	26.94%	28.15%	26.67%	29.81%	32.78%	32.41%	28.10%	26.25%	29.32%	31.23%	25.82%	27.89%	28.90%	35.61%	100.00%	42.29%	33.99%	32.13%
sp|Q6Z461|SPL13_ORYSJ	28.47%	30.38%	39.64%	35.71%	37.04%	45.65%	33.51%	36.36%	35.06%	40.72%	47.74%	46.41%	39.02%	43.79%	45.70%	47.02%	42.29%	100.00%	48.97%	45.78%
sp|Q7EXZ2|SPL14_ORYSJ	24.17%	25.47%	31.84%	30.60%	32.50%	35.52%	35.60%	31.47%	23.16%	27.68%	37.08%	38.10%	26.19%	30.48%	28.57%	36.33%	33.99%	48.97%	100.00%	65.27%
sp|A3C057|SPL17_ORYSJ	23.21%	24.32%	33.59%	31.91%	31.62%	34.68%	35.81%	30.42%	23.11%	25.00%	37.44%	34.43%	25.07%	30.94%	29.47%	38.11%	32.13%	45.78%	65.27%	100.00%

首先，左图是一张同源性百分比矩阵中，拟南芥 SPL7_ARATH（第一行序列）与各水稻 SPL 蛋白的序列一致性里，除自身比对的 100% 外，最高值为 38.70%，对应 SPL9_ORYSJ（第一行第二个单元格），且该单元格颜色显著更深；其余水稻 SPL 与它的一致性均仅 23%-24% 左右，远低于该数值，因此从图中可直接判定 SPL9_ORYSJ 是 SPL7_ARATH 最可能的水稻直系同源蛋白。

其次，右图的系统发育树直观展示了 SPL 家族蛋白的进化关系，拟南芥 SPL7_ARATH 与水稻 SPL9_ORYSJ 在树中直接聚类、亲缘关系最近，进一步验证了二者为直系同源蛋白。

（5）下载玉米转录因子蛋白质序列和编码区核苷酸序列数据，构建本地 BLAST 数据库。以拟南芥转录因子 SPL3 蛋白质序列为检索序列， 用 BlastP 搜索玉米转录因子蛋白质序列，用 tBlastN 搜索玉米转录因子编码区序列并分析结果；以拟南芥转录因子 SPL3 编码区序列为检索序列， 用 BlastN 搜索玉米转录因子编码区序列，用 BlastX 搜索玉米转录因子蛋白质序列中相似序列并分析结果。

①网站：https://planttfdb.gao-lab.org/index.php?sp=Zma，下载玉米转录因子蛋白序列 & CDS 序列。



Zea mays Transcription Factors

The gene annotaion from MaizeSequence(AGPv3.31/5b+) is used to identify transcription factors (TFs) of *Zea mays* (See datasource). Acorrding to the family assignment rules, 3308 TFs (2289 loci) are identified and classified into 56 families. You can download the TF list (here) and protein sequences of TFs (here), or download them on the download page.

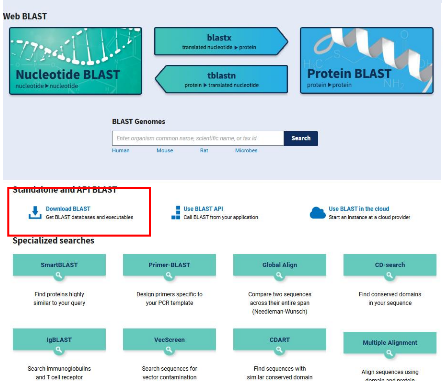
Browse by Family

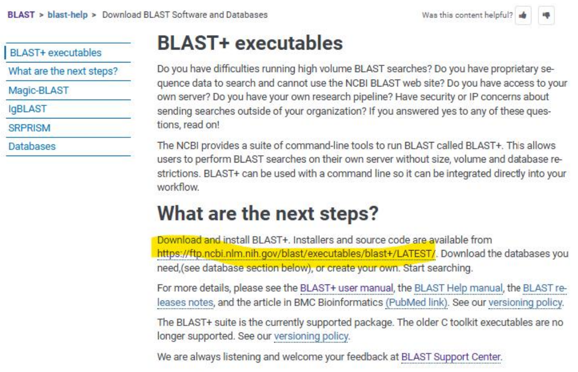
AP2 (54)	ARF (62)	ARR-B (13)	B3 (77)	BBR-BPC (9)
BES1 (17)	C2H2 (179)	C3H (110)	CAMTA (10)	CO-like (18)
CPP (17)	DBB (20)	Dof (52)	E2F/DP (24)	EIL (9)
ERF (204)	FAR1 (25)	G2-like (89)	GATA (54)	GRAS (104)
GRF (32)	GeBP (28)	HB-PHD (4)	HB-other (28)	HD-ZIP (97)
HRT-like (2)	HSF (49)	LBD (60)	LFY (4)	LSD (20)
M-type_MADS (46)	MIKC_MADS (90)	MYB (203)	MYB_related (169)	NAC (189)
NF-X1 (4)	NF-YA (36)	NF-YB (27)	NF-YC (25)	Nin-like (23)
RAV (3)	S1Fa-like (5)	SBP (55)	SRS (11)	STAT (2)
TALE (52)	TCP (52)	Trihelix (57)	VOZ (10)	WOX (30)
WRKY (161)	Whirly (6)	YABBY (31)	ZF-HD (26)	bHLH (308)
bZIP (216)				

名称	修改日期	类型	大小
Zma_cds.fas	2026/4/8 16:43	压缩存档文件夹	1,104 KB
Zma_pep.fas	2026/4/8 16:43	压缩存档文件夹	723 KB

②安装 blast+

下载网址：Index of /blast/executables/blast+/LATEST。





BLAST+ executables

Do you have difficulties running high volume BLAST searches? Do you have proprietary sequence data to search and cannot use the NCBI BLAST web site? Do you have access to your own server? Do you have your own research pipeline? Have security or IP concerns about sending searches outside of your organization? If you answered yes to any of these questions, read on!

The NCBI provides a suite of command-line tools to run BLAST called BLAST+. This allows users to perform BLAST searches on their own server without size, volume and database restrictions. BLAST+ can be used with a command line so it can be integrated directly into your workflow.

What are the next steps?

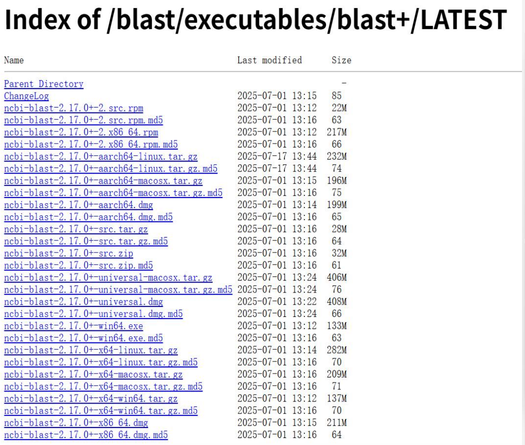
Download and install BLAST+. Installers and source code are available from https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/. Download the databases you need,(see database section below), or create your own. Start searching.

For more details, please see the BLAST+ user manual, the BLAST Help manual, the BLAST releases notes, and the article in BMC Bioinformatics (PubMed link). See our versioning policy.

The BLAST+ suite is the currently supported package. The older C toolkit executables are no longer supported. See our versioning policy.

We are always listening and welcome your feedback at BLAST Support Center.

找到电脑适配的版本下载安装，并且配置环境变量与测试安装成功与否。



Index of /blast/executables/blast+/LATEST

Name	Last modified	Size
Parent Directory		-
ChangeLog	2025-07-01 13:15	85
ncbi-blast-2.17.0+-2.src.rpm	2025-07-01 13:12	22M
ncbi-blast-2.17.0+-2.src.rpm.md5	2025-07-01 13:16	63
ncbi-blast-2.17.0+-2.x86_64.rpm	2025-07-01 13:12	217M
ncbi-blast-2.17.0+-2.x86_64.rpm.md5	2025-07-01 13:16	66
ncbi-blast-2.17.0+-aarch64-linux.tar.gz	2025-07-17 13:44	232M
ncbi-blast-2.17.0+-aarch64-linux.tar.gz.md5	2025-07-17 13:44	74
ncbi-blast-2.17.0+-aarch64-macosx.tar.gz	2025-07-01 13:15	196M
ncbi-blast-2.17.0+-aarch64-macosx.tar.gz.md5	2025-07-01 13:16	75
ncbi-blast-2.17.0+-aarch64.dmg	2025-07-01 13:14	199M
ncbi-blast-2.17.0+-aarch64.dmg.md5	2025-07-01 13:16	65
ncbi-blast-2.17.0+-src.tar.gz	2025-07-01 13:16	28M
ncbi-blast-2.17.0+-src.tar.gz.md5	2025-07-01 13:16	64
ncbi-blast-2.17.0+-src.zip	2025-07-01 13:16	32M
ncbi-blast-2.17.0+-src.zip.md5	2025-07-01 13:16	61
ncbi-blast-2.17.0+-universal-macosx.tar.gz	2025-07-01 13:24	406M
ncbi-blast-2.17.0+-universal-macosx.tar.gz.md5	2025-07-01 13:24	76
ncbi-blast-2.17.0+-universal.dmg	2025-07-01 13:22	408M
ncbi-blast-2.17.0+-universal.dmg.md5	2025-07-01 13:24	66
ncbi-blast-2.17.0+-win64.exe	2025-07-01 13:12	133M
ncbi-blast-2.17.0+-win64.exe.md5	2025-07-01 13:16	63
ncbi-blast-2.17.0+-x64-linux.tar.gz	2025-07-01 13:14	282M
ncbi-blast-2.17.0+-x64-linux.tar.gz.md5	2025-07-01 13:16	70
ncbi-blast-2.17.0+-x64-macosx.tar.gz	2025-07-01 13:16	209M
ncbi-blast-2.17.0+-x64-macosx.tar.gz.md5	2025-07-01 13:16	71
ncbi-blast-2.17.0+-x64-win64.tar.gz	2025-07-01 13:12	137M
ncbi-blast-2.17.0+-x64-win64.tar.gz.md5	2025-07-01 13:16	70
ncbi-blast-2.17.0+-x86_64.dmg	2025-07-01 13:15	211M
ncbi-blast-2.17.0+-x86_64.dmg.md5	2025-07-01 13:16	64

要进行大多数更改，你必须作为管理员登录。

视觉效果，处理器计划，内存使用，以及虚拟内存

与登录帐户相关的桌面设置

系统启动、系统故障和调试信息

Sherr 的用户变量(U)

变量	值
OneDrive	C:\Users\Sherr\OneDrive
Path	D:\软件\blast-2.17.0+\bin;C:\Users\Sherr\AppData\Local\
TEMP	C:\Users\Sherr\AppData\Local\Temp
TMP	C:\Users\Sherr\AppData\Local\Temp

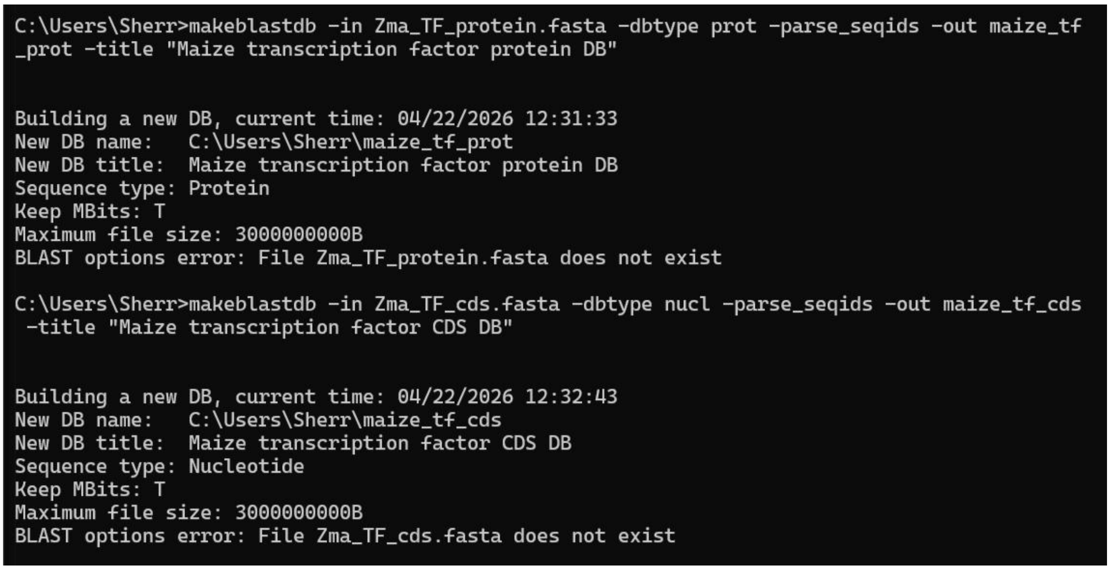
系统变量(S)

变量	值
blast2.17.0	D:\软件\blast-2.17.0+\bin
ComSpec	C:\Windows\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
EMBOSS_ACDROOT	C:\mEMBOSS\acd
EMBOSS_DATA	C:\mEMBOSS\data
EMBOSS_ROOT	C:\mEMBOSS
NUMBER_OF_PROCESSORS	12

```
C:\Users\Sherr>blastp -version
blastp: 2.17.0+
 Package: blast 2.17.0, build Jul  1 2025 08:57:20

C:\Users\Sherr>makeblastdb -version
makeblastdb: 2.17.0+
 Package: blast 2.17.0, build Jul  1 2025 08:57:20
```

③用 makeblastdb 构建本地库：

[蛋白质数据库（用于 blastp/blastx）;核酸（CDS）数据库（用于 blastn/tblastn）]

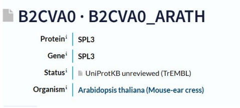


```
C:\Users\Sherr>makeblastdb -in Zma_TF_protein.fasta -dbtype prot -parse_seqids -out maize_tf
_prot -title "Maize transcription factor protein DB"


Building a new DB, current time: 04/22/2026 12:31:33
New DB name:   C:\Users\Sherr\maize_tf_prot
New DB title:  Maize transcription factor protein DB
Sequence type: Protein
Keep MBits: T
Maximum file size: 3000000000B
BLAST options error: File Zma_TF_protein.fasta does not exist

C:\Users\Sherr>makeblastdb -in Zma_TF_cds.fasta -dbtype nucl -parse_seqids -out maize_tf_cds
 -title "Maize transcription factor CDS DB"


Building a new DB, current time: 04/22/2026 12:32:43
New DB name:   C:\Users\Sherr\maize_tf_cds
New DB title:  Maize transcription factor CDS DB
Sequence type: Nucleotide
Keep MBits: T
Maximum file size: 3000000000B
BLAST options error: File Zma_TF_cds.fasta does not exist
```

参数说明

- -in：输入 FASTA
- -dbtype：prot（蛋白）/ nucl（核酸）
- -parse_seqids：保留原序列 ID（必加，方便后续检索）
- -out：数据库前缀（生成 .pin/.psq/.phr 或 .nin/.nsq/.nhr）

④uniprot 检索出拟南芥转录因子 SPL3 蛋白质序列， 用 BlastP 搜索玉米转录因子蛋白质序列，用 tBlastN 搜索玉米转录因子编码区序列并分析结果。



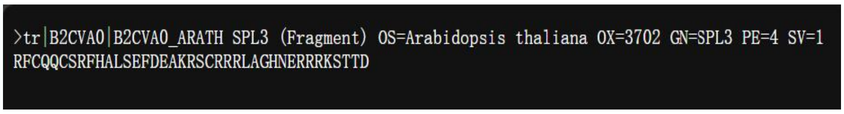
B2CVA0 · B2CVA0_ARATH

Protein: SPL3

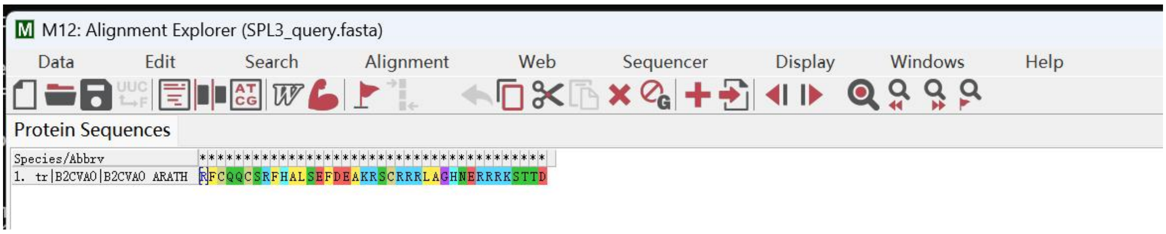
Gene: SPL3

Status: UniProtKB unreviewed (TrEMBL)

Organism: Arabidopsis thaliana (Mouse-ear cress)



```
>tr|B2CVA0|B2CVA0_ARATH SPL3 (Fragment) OS=Arabidopsis thaliana OX=3702 GN=SPL3 PE=4 SV=1
RFCQQCSRFHALSEFDEAKRSCRRRLAGHNERRRKSTTD
```



```
Species/Abbrv               ***************************************
1. tr|B2CVA0|B2CVA0 ARATH   RFCQQCSRFHALSEFDEAKRSCRRRLAGHNERRRKSTTD
```

后面比对不出来了，一直显示路径错误。

心得体会：

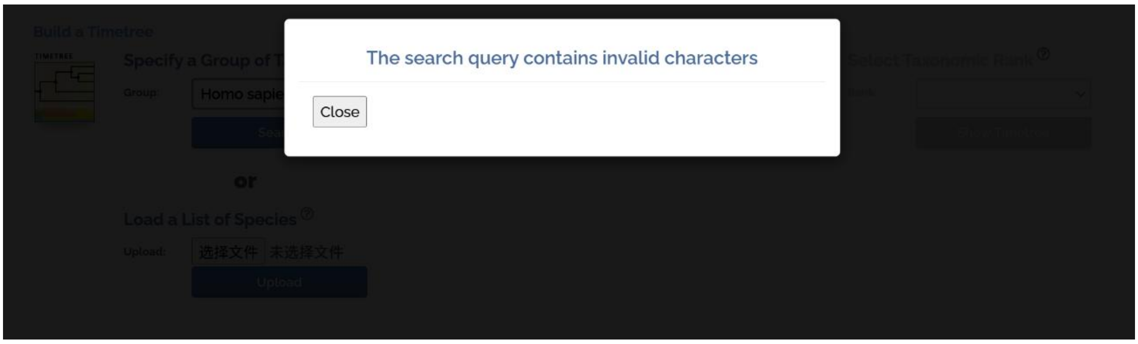
当在亲缘关系较近的物种间寻找高度同源的基因时，优先使用 blastn（核酸对核酸）或 blastp（蛋白对蛋白），速度快且结果精确。

当在亲缘关系较远的物种间挖掘保守基因时，强烈推荐使用 tblastn 或 blastx。它们通过蛋白质层面的比对，能突破核酸序列的差异，发现功能保守但序列分歧的同源基因。

四种 BLAST 程序的联合使用，能从不同角度提供证据，使结果更可信、更全面。

问题：

1.尝试利用 TimeTree5 平台的“Build a Timetree”模块构建＞2 个物种 TimeTree，但 TimeTree，Group 输入“Homo sapiens,Mus musculus,Rattus norvegicus”页面如下图，尝试提交 Text 导入物种名，同样无法识别。


The search query contains invalid characters

2.在实际操作在使用人，小鼠，大鼠 alpha 蛋白进行序列比对之后发现在改变空位罚分之后没有显著的区别，想要了解对于空位罚分的改变在序列比对中的意义。

3.课外练习 3 的第八题，题目不完整，缺少 CEAM5_HUMAN 的恒定结构域的具体信息。Ig-like C2-type 一共检索出来六种。