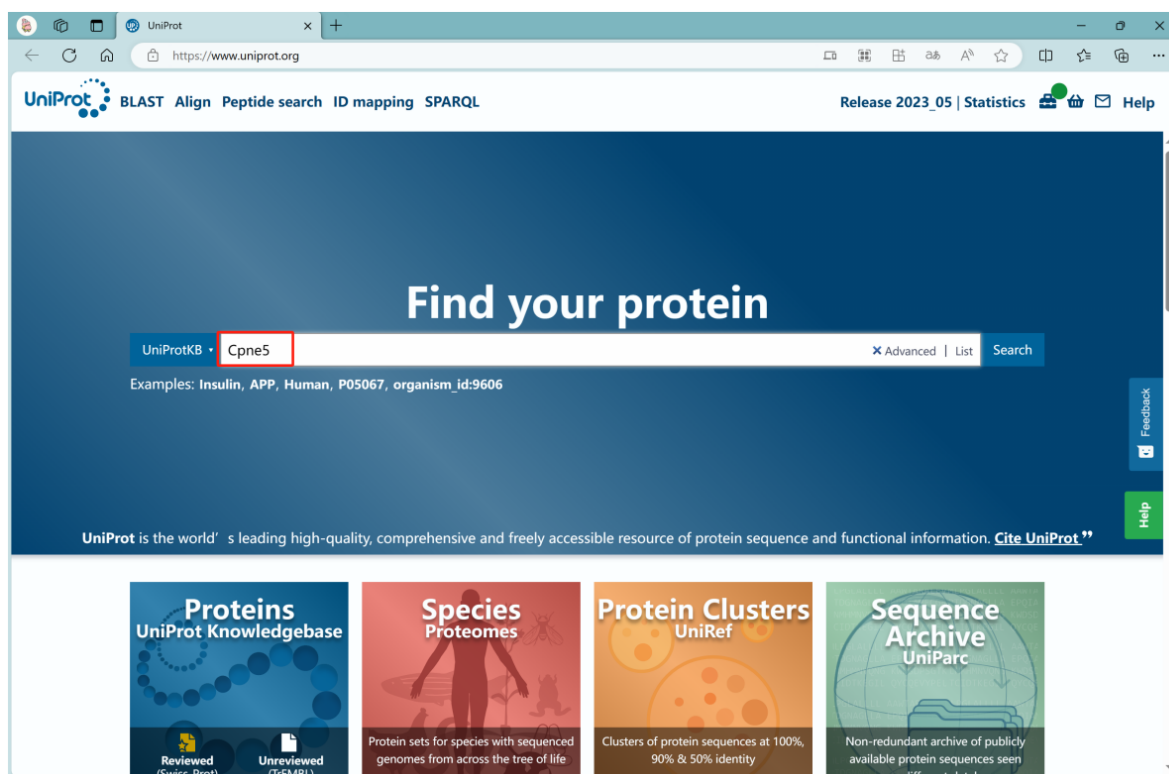


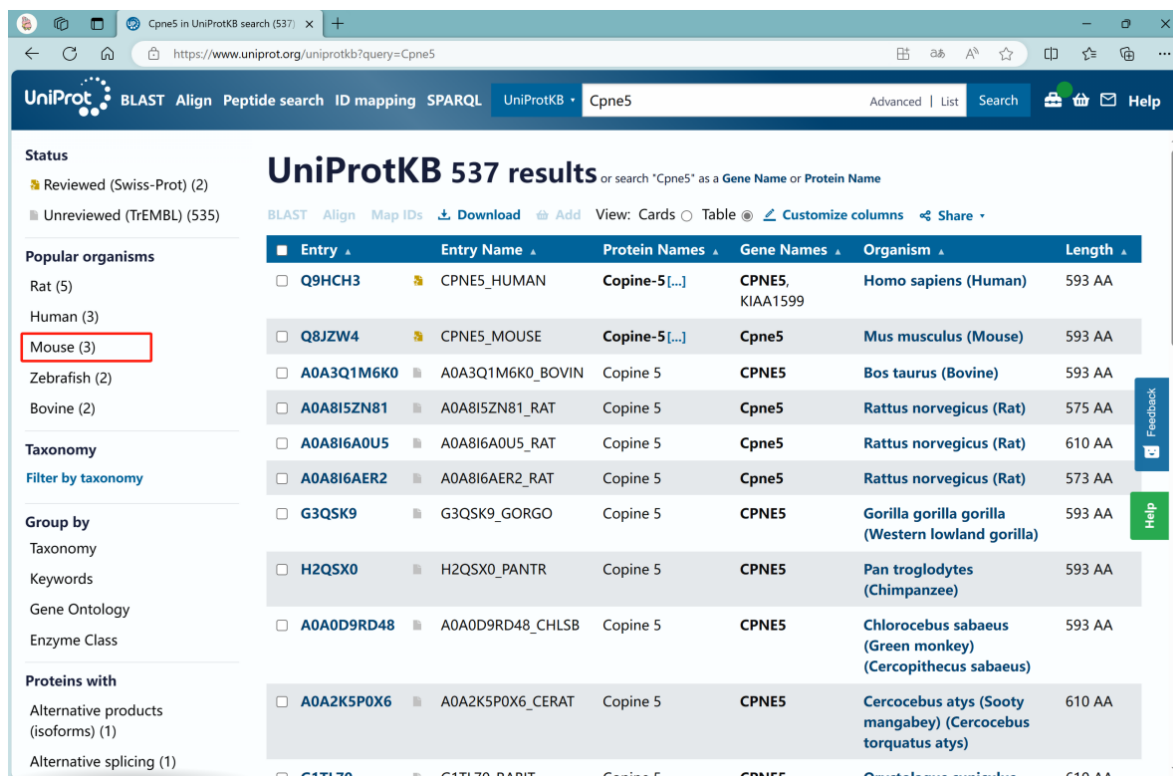
这部分教程将指引你如何获得基因的 ESM-2 编码

查询

进入 [UniProt](https://www.uniprot.org) 网站后，在搜索框搜索基因名（这里以 Cpne5 为例），然后回车：



然后在左边筛选框点 Mouse：



在确定红色框为 Cpne5 的情况下，点击黄色框的选中（Gene Names 相同的情况下选第一个），然后点绿色框的下载；

UniProtKB BLAST Align Peptide search ID mapping SPARQL UniProtKB Cpne5 Advanced | List Search

UniProtKB 3 results

Status
Reviewed (Swiss-Prot) (1)
Unreviewed (TrEMBL) (2)

Popular organisms
Mouse (3)

Taxonomy
Filter by taxonomy

Group by
Taxonomy
Keywords
Gene Ontology
Enzyme Class

Proteins with
Binding site (1)
Chain (1)
Cofactors (1)
Compositional bias (1)
Developmental stage (1)
More items

BLAST Align Map ID **Download** Add View: Cards Table Customize columns Share 1 row selected out of 3

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input checked="" type="checkbox"/> Q8JZW4	CPNE5_MOUSE	Copine-5[...]	Cpne5	Mus musculus (Mouse)	593 AA
<input type="checkbox"/> E9Q1V2	E9Q1V2_MOUSE	Copine V	Cpne5	Mus musculus (Mouse)	307 AA
<input type="checkbox"/> Q8BVA1	Q8BVA1_MOUSE	C2 domain-containing protein	Cpne5	Mus musculus (Mouse)	153 AA

Feedback Help

在确定红色框为下载选中的 1 个基因的 FASTA 格式（一般默认就是这样）后，点击黄色框下载

Download

☒ Download selected (1)
☐ Download all (3)

Format
FASTA (canonical)

Compressedⁱ
☒ Yes
☐ No

Generate URL for API Preview 1 Cancel **Download**

Advanced | List Search

Customize columns Share 1 row selected out of 3

Gene Names	Organism	Length
Cpne5	Mus musculus (Mouse)	593 AA
Cpne5	Mus musculus (Mouse)	307 AA
Cpne5	Mus musculus (Mouse)	153 AA

Feedback Help

下载并解压后，得到这样的一个文件，其中红色框部分为基因名，黄色框部分为蛋白表达。

```
uniprotkb_accession_Q8JZW4_2024_01_06.fasta - 记事本
文件(E) 编辑(E) 格式(O) 查看(V) 帮助(H)
>sp|Q8JZW4|CPNE5_MOUSE Copine-5 OS=Mus musculus OX=10090 GN=Cpne5 PE=1 SV=1
MEQPEDMASLSEFDSLGSIPATKVEITVSCRNLLDKDMFSKSDPLCVMYTQGMENKQWR
EFGRTVIDNTLNPDFVRKFIVDYFFEEKQNLRFDLVDVSKSPDLSKHDFLGQAFCTLG
EIVGSSGSRLEKPLTIGTFLNSRTGKPMFPAVSNGGVPGKKCGTIILSAEELSNCRDVAT
MQFCANKLDKKDFFGKSDPFLVFYRSNEDGFTTICKTEVMKNTLNPVWQTFSIPVRALC
NGDYDRTIKVEVYDWDGSHDFIGFTTSYRELARGQSQFNIYEVINPKKKMKKKKYVN
SGTVTLLSFAVESESTFLDYIKGGTQINFTVAIDFTASNGNPSQSTSLHYMSPYQLNAYA
LALTAVGEIIGHYDSDKMFALGFAGLPPDGRVSHEFPLNGNQENPSCCGIDGILEAYH
SSLRTVQLYGPTNFAPVVTHVARNAAAVQDGSQYSVLLIITDGVISDMAQTKEAIVNAAK
LPMSIIIVGVGQAEFDAMVELDGDVRISSRGKLAERDIVQFVPRDYVDRTGNHVL SMA
RLARDVLAIPDQLVSYMKAAQGIRPRPPPAAPAQSPQSPAHSPPGSPVHHTHI
```

将基因名和蛋白表达提取出来后，得到下述文件：

```
uniprotkb_accession_Q8JZW4_2024_04_16.fasta - 记事本
文件(E) 编辑(E) 格式(O) 查看(V) 帮助(H)
>Cpne5
MEQPEDMASLSEFDSLGSIPATKVEITVSCRNLLDKDMFSKSDPLCVMYTQGMENKQWREFGRTEVIDNTLNPDFVRKFIVE
```

ESM-2 编码

安装好 ESM-2（安装过程可参考 [ESM-2](#)）后，输入如下指令：

```
cd esm-main/

python scripts/extract.py esm2_t36_3B_UR50D
uniprotkb_accession_Q8JZW4_2024_04_16.fasta examples/data/some_proteins_emb_esm2
--repr_layers 36 --include mean_per_tok

CUDA_VISIBLE_DEVICES=1 python scripts/extract.py esm2_t36_3B_UR50D
/home/songyj/uniprotkb_accession_Q8JZW4_2024_04_16.fasta
examples/data/some_proteins_emb_esm2 --repr_layers 36 --include mean_per_tok
```

其中 uniprotkb_accession_Q8JZW4_2024_04_16.fasta 为上文中处理好的基因名和蛋白表达文件；运行成功后会在 esm-main/examples/data/some_proteins_emb_esm2/ 目录下生成基因的编码文件（在这里为 Cpne5.pt）

把你的数据集的基因全部转化为 *.pt 文件后，将其至于一个文件夹下（如 ./pt/），之后运行以下代码：

```
import os
import torch
import pickle
import pandas as pd

df = pd.DataFrame()

for path, dir_lst, file_lst in os.walk(r'./pt'):
    for file_name in file_lst:
        data = torch.load(open(os.path.join(path, file_name), 'rb'))
        print(data['label'])
        print(data['representations'][36][-1])
        df.insert(df.shape[1], data['label'], data['representations'][36]
[-1].numpy())
```

```
pickle.dump(df, open('emb.pkl', 'wb'))
```

即可生成基因的 ESM-2 编码文件