肽段可检测性预测软件使用说明书

DeepDetect V1.0.0

Last revised February 26, 2023

杨婧涵 [1,2]，程志远 [1,2]，巩馥洲 [1,2]，付岩 [1,2*]

[1]CEMS, NCMIS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

[2]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.

[*]通讯作者:

付岩，邮箱：yfu@amss.ac.cn

1 软件介绍

现在生物质谱技术已经成为了高通量蛋白质组学研究的中心，但是基于此的肽段检测机制却尚不明确，从而阻碍了人们对质谱数据进行大规模、高精度的解析。近年来，研究者们提出了不同的肽段可检测性预测方法，但都不尽如人意。一方面，大部分基于深度学习的方法没有完全发挥算法本身的优势。众所周知，神经网络是一种端到端的模型，能够从原始数据中学习到深层特征，而现有的深度学习方法大多囿于人工设计的特征，导致模型性能存在不足。[1,2] 另一方面，基于传统机器学习的方法没有充分利用蛋白酶切的信息。我们知道，典型的鸟枪法蛋白质组学实验包括蛋白酶切和肽段检测两个过程，而普遍存在的漏切现象导致了蛋白酶切的不确定性，从而在极大程度上影响到了进入检测过程的肽段。此外，目前已有的肽段可检测性预测工具均只针对胰蛋白酶，也难以满足广泛的实际应用需求。为了解决上述问题，我们开发了支持八种常用水解酶的肽段可检测性预测软件 DeepDetect。首先，我们利用长短时记忆网络提取序列特征，再通过预测的肽段酶切概率特征对网络预测值进行校正。与已发表的肽段可检测性预测软件 AP3[3] 和 PepFormer[4] 相比，该软件具有更高的精度和更强的泛化能力。该软件仅需输入蛋白序列就可以完成理论酶切肽段可检测性的准确预测，在蛋白质组学的很多领域都能发挥重要作用，例如：辅助评估鉴定肽段的可信度、靶向蛋白质组学的实验设计以及蛋白质定量等。DeepDetect 需要以命令行的方式调用，可从如下网址免费下载使用：http://fugroup.amss.ac.cn/software/DeepDetect/DeepDetect.html。

2 软件使用环境

2.1 硬件环境

- CPU：2.2GHZ 以上
- 内存：2G 以上
- 硬盘：100G 以上

2.2 软件环境

- 已经验证的操作系统
  - Windows 7（64 位）
  - Windows 10（64 位）
- Python 3.5.2 版本
  - TensorFlow 1.10.0 版本库
  - Keras 2.2.4 版本库
  - Numpy 1.14.5 版本库

3 软件使用手册

3.1 软件分析流程

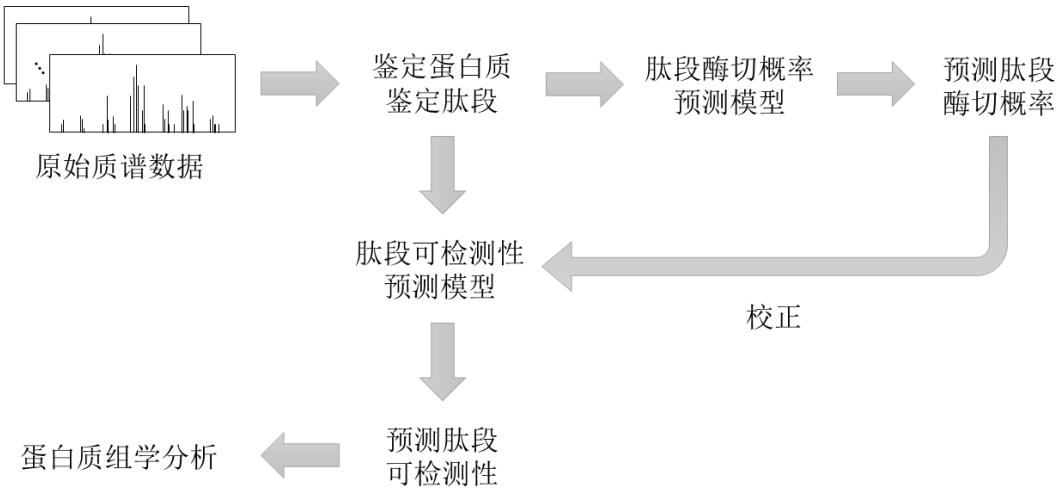


图 1 肽段可检测性预测软件流程图

软件分析流程如图 1 所示。主要包括两个部分：1）肽段酶切概率预测；2）肽段可检测性预测。首先，通过蛋白质鉴定软件（如：MaxQuant[5,6] 等）对原始质谱数据进行分析，得到鉴定蛋白质和鉴定肽段。然后，根据鉴定蛋白质的谱图计数（Spectral Counts, SCs）和序列覆盖度筛选高可信鉴定蛋白，再对其所有候选酶切位点提取 31-连子（包含位点），即分别取位点 N 端和 C 端的 15 位氨基酸加上位点本身所构成的序列，输入到肽段酶切概率预测模型中预测所有候选位点的

酶切概率值。进一步地，按照用户指定的参数设置对所有高可信鉴定蛋白进行理论酶切，并计算所有理论酶切肽段的酶切概率值。同时，按照字母表顺序对常见的 20 种氨基酸进行编码，通过词嵌入层把各氨基酸表示为学习到的向量，再使用一个双向长短时记忆网络层提取抽象特征，经过一个批标准化层后，由一个 sigmoid 函数激活的全连接层进行概率预测。最后，经由肽段酶切概率预测值的校正后，得到所有理论酶切肽段的可检测性预测值，进而完成下游的蛋白质组学分析。

3.2 软件环境配置

DeepDetect 是基于 Python 语言打包的可执行文件(.exe)，在调用该软件之前最好配置相应版本下的 Python 环境，避免出现不必要的错误。为了方便用户使用，建议直接在如下官方网址下载并安装 Anaconda：https://www.anaconda.com/。下载页面如图 2 所示：

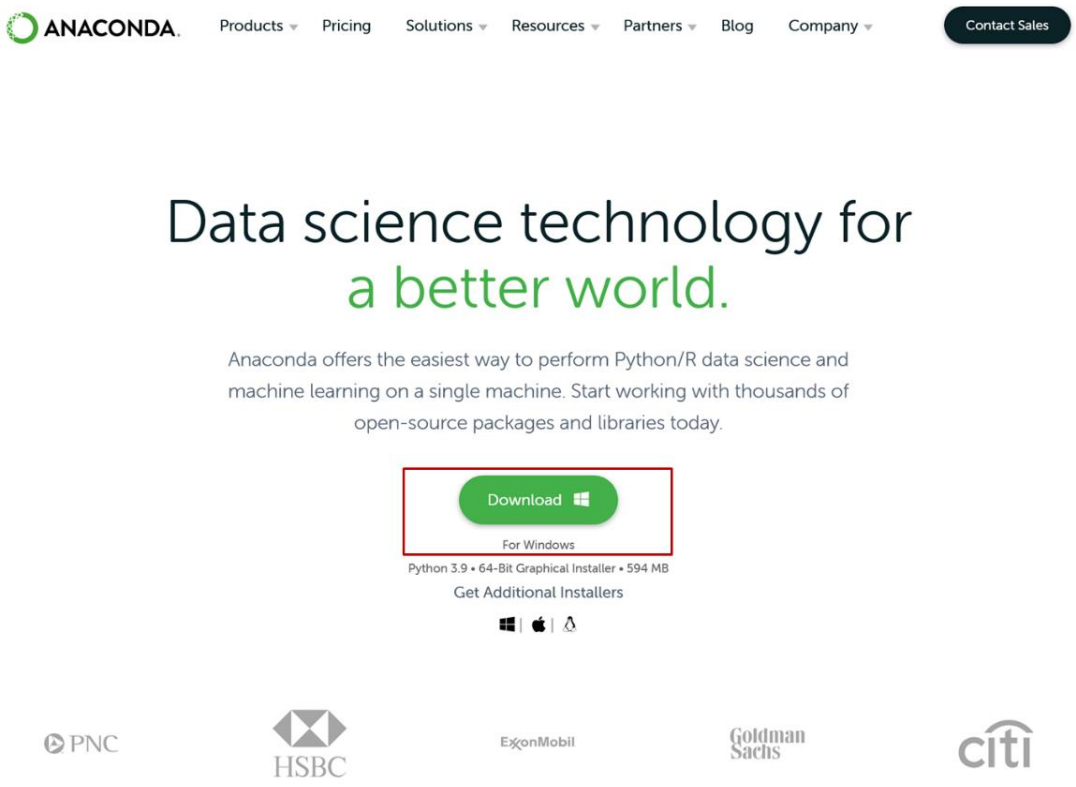


图 2 官网下载 Anaconda 示意图

双击用户使用的电脑操作系统对应的下载文件即可自动安装，并自动配置系统环境变量。值得注意的是，官网上下载的 Anaconda 通常只会自动安装最新版

本的 Python 环境，用户还需要自行创建运行 DeepDetect 所对应的 Python 版本。安装 Anaconda 的一大便利之处就在于，只需简单几个步骤就可以完成多版本 Python 环境的配置，可供用户自由切换。

首先，点击桌面左下角“开始”按钮，键入“cmd.exe”打开命令行解释器。然后，输入“conda --version”命令查看当前的 conda 版本，并确认当前安装的 Anaconda 为最新版（图 3a）。若不确定是否为最新版本，可通过“conda update conda”和“conda update anaconda”两个命令进行升级（图 3b）。

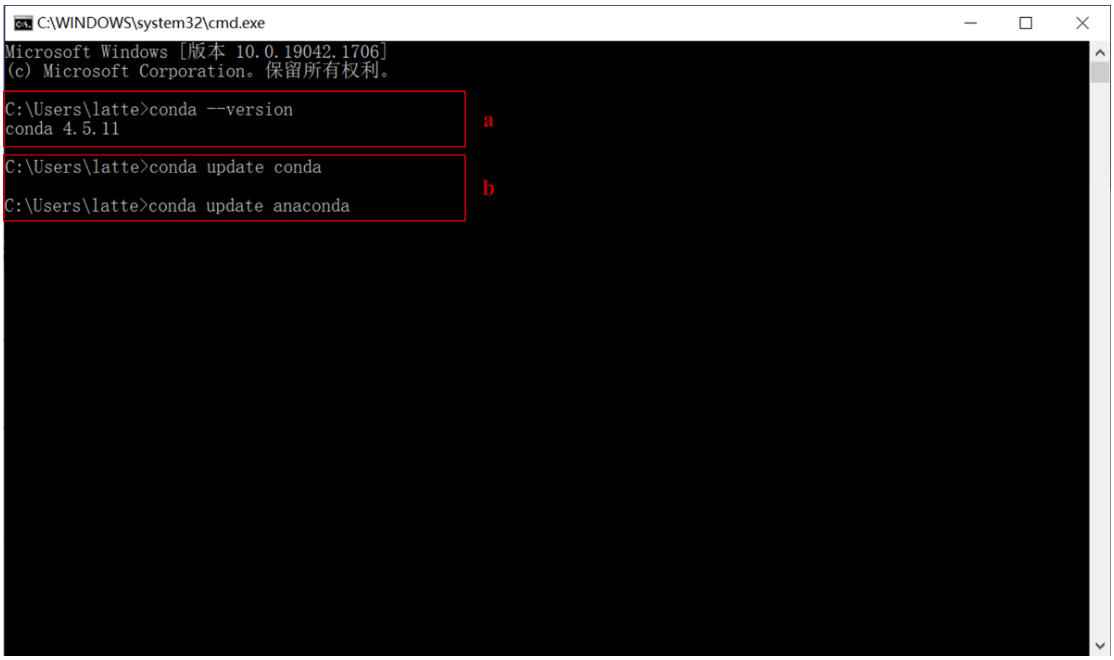


图 3 查看并升级 Anaconda 版本示意图

接着，使用“conda create --name python3.5.2 python=3.5.2”命令创建一个新的 Python 3.5.2 版本环境（图 4）。

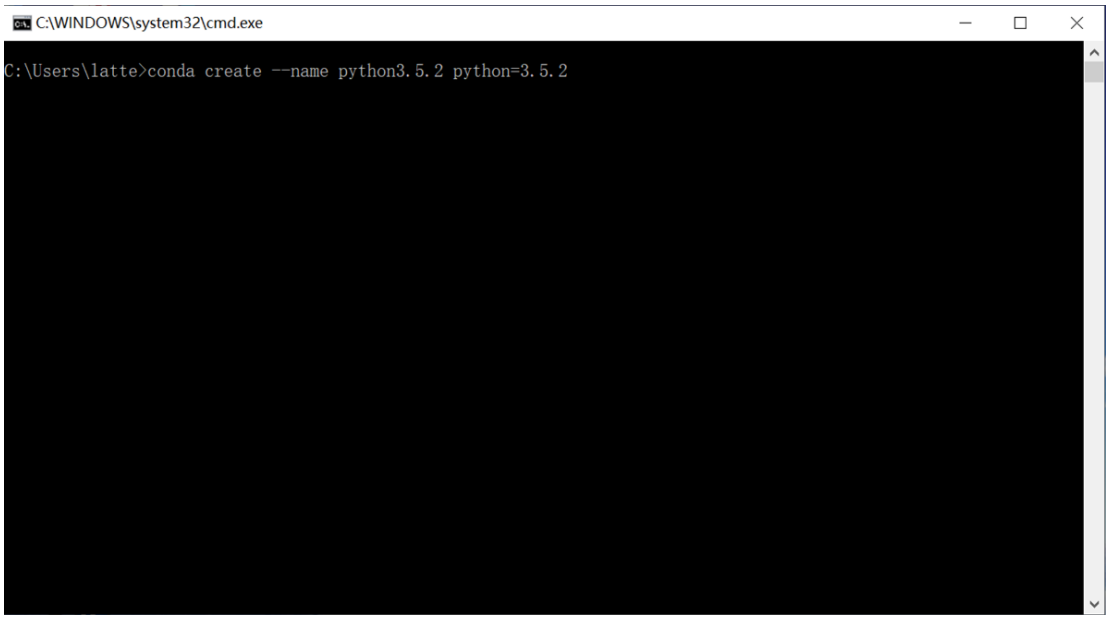


图 4 创建新版本 Python 环境示意图

创建完成后，可以通过“conda info -e”命令查看当前已经安装的所有 Python 环境（图 5a）。在使用 DeepDetect 前，只需要在命令行解释器中使用“activate python3.5.2”命令激活 Python 3.5.2 环境即可（图 5b）。

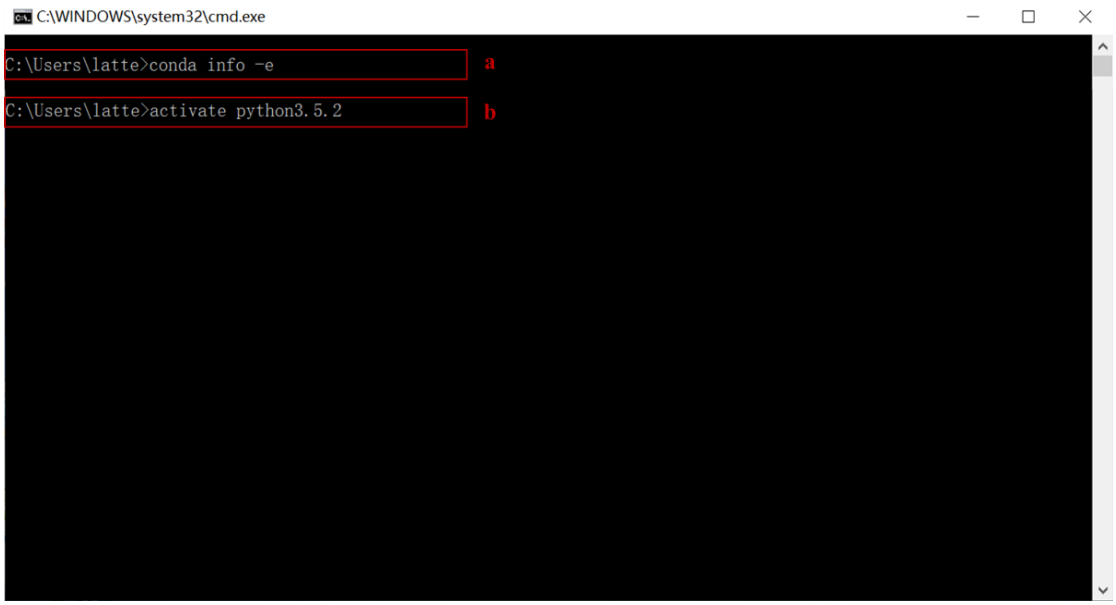


图 5 查看所有 Python 环境并激活新版本 Python 环境示意图

3.3 使用命令行运行软件

3.3.1 参数设置

在命令行中调用 DeepDetect 时，用户可以自定义一系列参数设置，具体参数及其含义如下表所示：

表 1 肽段可检测性预测软件 DeepDetect 的参数说明

参数名称	含义
input	蛋白质序列文件路径（.fasta）
output	输出文件路径（.txt）
regular	在蛋白质序列文件中提取蛋白质标识名的正则表达式（默认值：">(.*?)\s"）
protease	水解酶类型（默认值：Trypsin）
missed_cleavages	蛋白质理论酶切肽段的允许最大漏切位点数（默认值：2）
min_len	蛋白质理论酶切肽段的允许最小长度（默认值：7）
max_len	蛋白质理论酶切肽段的允许最大长度（默认值：47）

3.3.2 运行命令

首先，点击桌面左下角“开始”按钮，键入“cmd.exe”打开命令行解释器。然后，通过 cd 命令设置 DeepDetect.exe 所在目录为当前工作目录。最后，按照如下格式调用 DeepDetect：

>DeepDetect.exe --input=*the path of protein sequence file* --output=*the path of output file* --regular=">(.*?)\s" --protease=Trypsin --missed_cleavages=2 --min_len =7 --max_len=47

图 6 为一个实例：

```
E:\DeepDetect\tool\exe>DeepDetect.exe --input=.\SampleData\Sample_Data.fasta --output=.\SampleData\Sample_Data.txt --regular=">(.*?)\s" --protease=Trypsin --missed_cleavages=2 --min_len=7 --max_len=47
```

图 6 命令行调用 DeepDetect 实例

3.4 结果文件说明

DeepDetect 成功运行后会在用户设置的目录下生成一个结果文件（.txt），表 2 为结果文件内容说明。

表 2 DeepDetect 结果文件说明

列名	含义
Protein id	蛋白质标识名
Peptide sequence	蛋白质的理论酶切肽段
Peptide detectability	理论酶切肽段的可检测性预测值

4 参考文献

1. Zimmer, D., Schneider, K., Sommer, F., Schroda, M. & Mühlhaus, T. Artificial Intelligence Understands Peptide Observability and Assists With Absolute Protein Quantification. *Front. Plant Sci.* **9**, 1559 (2018).

2. Yu, M., Duan, Y., Li, Z. & Zhang, Y. Prediction of peptide detectability based on capsnet and convolutional block attention module. *Int. J. Mol. Sci.* **22**, 12080 (2021).

3. Gao, Z., Chang, C., Yang, J., Zhu, Y. & Fu, Y. AP3: An Advanced Proteotypic Peptide Predictor for Targeted Proteomics by Incorporating Peptide Digestibility. *Anal. Chem.* **91**, 8705–8711 (2019).

4. Cheng, H. *et al.* PepFormer: End-to-End Transformer-Based Siamese Network to Predict and Enhance Peptide Detectability Based on Sequence only. *Anal. Chem.* **93**, 6481–6490 (2021).

5. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

6. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).