第一次上机实验

一、实验目的

用SARS-CoV-2参考基因组进行在线BLAST。使用在线工具基于氨基酸序列分析蛋白质的二级结构和三级结构。

1. 学习和掌握常见生物数据库的检索
2. 学习和掌握在线blast的基本使用方法
3. 学习和掌握基于序列的蛋白质理化性质分析
4. 学习和掌握基于序列的蛋白质二级和三级结构预测

二、实验内容

1. NCBI Blast在线工具的使用及结果说明

- 获取SARS-CoV-2参考基因组的accession number 打开NCBI主页（http://www.ncbi.nlm.nih.gov/），在核酸（Nucleotide）数据库查询SARS-CoV-2，找到RefSeq后面的SARS-CoV-2参考基因组的accession number。
- 在线blast操作 打开blast页面https://blast.ncbi.nlm.nih.gov/Blast.cgi，进入Nucleotide BLAST。在Enter Query Sequence框中输入SARS-CoV-+ 2参考基因组的accession number。在Choose Search Set部分选择默认的数据库。在Program Selection部分默认选择Highly similar sequences (megablast)。Algorithm parameters部分可调整Max target sequences 改变最终展现的序列的数量。Expect threshold 调整E值（默认0.05，改小一点？）。点击blast运行。
- 阅读blast结果
- 下载blast结果并保存为FASTA (complete sequence)格式

2. 基于序列对蛋白质进行理化性质分析

(1) 计算序列等电点和分子量

Compute pI/MW(theoretical pI (isoelectric point) and Mw (molecular weight))是计算序列的等电点和分子量的小工具。

基本操作

- 打开网页；
- 输入的格式可以是UniProtKB/Swiss-Prot的ID号，如ALBU_HUMAN，或UniProt的ID号，如P04406。也可以直接输入序列。注意，仅仅是序列就行了，不是Fasta格式。
- 运行
- submit查看整个蛋白质序列的预测结果，也可以选择某一片段。

(2) 氨基酸的理化性质分析

ProtParam这个工具可以计算氨基酸个数(Number of amino acids),蛋白质分子量(Molecular weight)、理论等电点(Theoretical pI)、氨基酸的组成(Amino acid composition)、原子组成(Atomic composition)、总原子数(Total number of atoms)、分子式(Formula)、消光系数(Extinction coefficients)、估计半衰期(Estimated half-life)、不稳定指数(Instability index < 40 -> Stable；>40 -> Unstalbe)、脂肪系数(Aliphatic index)、亲水性平均系数(Grand average of hydropathicity)等等。需要注意的是，如果序列有修饰过的氨基酸残基，则不计算在内。也可以分析蛋白质的亲疏水性分析。

- 打开ProtParam网站。
- 输入序列。如果分析SWISS-PROT和TrEMBL数据库中序列，直接填写Swiss-Prot/TrEMBL AC号（accession number）；如果分析新序列，直接在搜索框中粘贴氨基酸序列。
- 运行
- submit查看整个蛋白质序列的预测结果，也可以选择某一片段。

(3) 特定蛋白酶或化学物质可能的裂解位点预测

PeptideCutter是一个根据蛋白质序列预测特定蛋白酶或化学物质可能的裂解位点。

- 打开PeptideCutter网站。
- 输入序列。如果分析SWISS-PROT和TrEMBL数据库中序列，直接填写Swiss-Prot/TrEMBL AC号（accession number）；如果分析新序列，直接在搜索框中粘贴氨基酸序列。
- 运行。
- 查看运行结果。切割发生在标记氨基酸的右侧（C端方向）。您可以通过在地图上单击相应的酶名称来显示单个酶的结果。

(4) 蛋白的亲疏水性分析

ProtScale可计算和展示所选蛋白的50多种氨基酸标度（Amino acid scale）的图谱，氨基酸标度是给每种氨基酸指定的数值，比如Molecular weight（分子质量）、Number of codon（密码子数）、Bulkiness（膨胀度）等等。最常用的标度是疏水性标度，主要来源于多肽在非极性和极性溶剂中的分配实验数据。如果你对所选标度感兴趣可点击标度名称可进入详情页面，查看具体的氨基酸的标度数值和参考文献，如下图，这里正值的氨基酸具有更大的疏水性，负值越小的氨基酸则更加亲水。

- 打开ProtScale网站。
- 输入序列。如果分析SWISS-PROT和TrEMBL数据库中序列，直接填写Swiss-Prot/TrEMBL AC号（accession number）；如果分析新序列，直接在搜索框中粘贴氨基酸序列。
- 选择scale，设定窗口大小等参数。
- 查看运行结果。

3. 蛋白质序列特征分析

(1) 跨膜区分析

以下是几个蛋白质的跨膜螺旋预测的工具。以G蛋白偶联受体

- TMHMM基于隐马尔可夫模型(HMM)的蛋白质跨膜区预测工具.TMHMM综合了跨膜区疏水性、电荷偏倚、螺旋长度和膜蛋白拓扑学限制等性质，采用隐马氏模型(Hidden Markov Models)，对跨膜区及膜内外区进行整体的预测。TMHMM是目前最好的进行跨膜区预测的软件,它尤其长于区分可溶性蛋白和膜蛋白，因此首选它来判定一个蛋白是否为膜蛋白。
- 基本操作
    - 打开网址TMHMM。
    - 上传本地的fasta文件或以fasta格式粘贴到输入框即可。如果只有序列，在最上面新建一行，以 > 开头，后面写点啥都行。
    - Output format保持默认
    - submit
    - 结果：序列长度，跨膜螺旋的个数，蛋白质的胞内段、跨膜螺旋、胞外端的氨基酸位置，图示。
- TMPred基于对 Tmbase 数据库的统计分析来预测蛋白质跨膜区和跨膜方向。
- 基本操作
    - 打开网址TMPred。
    - 选择输出格式，跨膜螺旋的最大和最小长度，输入蛋白质ID或者氨基酸序列
    - submit
    - 结果：直接查看 **-----> SSTRONGLY prefered model** 部分。看有多少跨膜螺旋。同时有图示。
- CCTOP也是基于隐马尔可夫模型(HMM)的蛋白质跨膜区预测工具。
- 基本操作
    - 打开网址CCTOP。
    - 新建prediction，输入氨基酸序列
    - submit(可分析信号肽)
    - 结果：不同维度展示预测跨膜结构的结果

比较不同的跨膜结构预测方法的差异

(2) 信号肽分析

信号肽是蛋白质N-末端一段编码长度为5-30的疏水性氨基酸序列，用于引导新合成蛋白质向通路转移的短肽链。信号肽存在于分泌蛋白、跨膜蛋白和真核生物细胞器内的蛋白中。

信号肽指引蛋白质转移的方式有两种：（1）常规的分泌（Sec/secretory）通路；（2）双精氨酸转移（Tat/twin-arginine）通路。前者存在于原核生物蛋白质转移到质膜过程中，以及真核生物蛋白质转移到内质网膜的过程中。后者存在于细菌、古菌、叶绿体和线粒体中，信号肽序列较长、疏水性较弱且尾部区含有两个连续精氨酸。相比于前者转运非折叠蛋白质，后者能转运折叠蛋白质跨越双层脂质膜 。

信号肽指引蛋白质转运后，将由信号肽酶进行切除。信号肽酶有三种：（1）一型信号肽酶（SPaseI）；（2）二型信号肽酶（SPaseII）；（3）三型信号肽酶（SPaseIII）。大部分信号肽由SPaseI进行移除，SPaseI存在古菌、细菌和真核生物中，且在真核生物的内质网膜上仅存在一型信号肽酶。细菌和古菌脂蛋白的信号肽C端含有一段称为 lipobox 的保守区域，由SPaseII切除其信号肽，且lipobox紧邻切除位点（CS/Cleavage Site）的氨基酸是半胱氨酸，这和锚定到膜的功能是相关的。细菌的四型菌毛蛋白信号肽由SPaseIII进行切除。此外：分泌通路（Sec）相关信号肽能由SPaseI、SPaseII和SPaseIII切除，但是双精氨酸转移（Tat）通路相关信号肽仅由 SPaseI和SPaseII切除。

使用SignalP 5.0能对原核生物的信号肽Sec/SPI、Sec/SPII和Tat/SPI，和对真核生物仅含有 Sec/SPI信号肽进行预测 。 SignalP 5.0目前不能对Tat/SPII进行预测。此外，由于没有足够大的数据进行训练，SignalP 5.0 也不能对Sec/SPIII进行分析。

基本操作：

- 打开网址。
- 序列输入(以胰岛素为例Insulin_uniport)。一种是点击选择文件直接上传FASTA文件；另一种是将氨基酸序列复制粘贴到“文本框”中(PTM / Processingi; Sequences)。
- 物种选择。真核生物（ Eukarya）、革兰氏阳性细菌（Gram-positive）、革兰氏阳性细菌（ Gram-negative）和古细菌（Archaea）
- submit查看信号肽的预测结果。

3. 蛋白质结构预测

(1) 蛋白质二级结构预测

蛋白质二级结构包括α螺旋，β折叠， β转角，无规卷曲以及模体等蛋白质局部结构组件。

PredictProtein可基于氨基酸序列预测蛋白质二级结构。可以获得功能预测、二级结构、基序、二硫键结构、结构域等许多蛋白质序列的结构信息。该方法的平均准确率超过72%，最佳残基预测准确率达90%以上。因此，被视为蛋白质二级结构预测的标准。用户需要注册ID、验证E-mail后，才能使用PredictProtein工具。

基本操作

- 打开网址PredictProtein并注册
- 输入序列
- submit，关闭窗口，在My Predictions 里查看。
- 结果显示
    - Secondary Structure and Solvent Accessibility： 蛋白质二级结构预测
    - Transmembrane Helices： 跨膜螺旋预测
    - Disulphide Bridges： 二硫键预测

(2) 蛋白质三级结构预测

SWISS-MODEL 原理：相似的氨基酸序列应该对应着相似的蛋白质结构。 要求：找到与目标序列一致度≥30%已知结构作为模板 基本操作

- 打开网站SWISS-MODEL
- 粘贴序列，填写Project Title和邮箱，点击build model

> 一般耗时几分钟到半小时不等。运行成功后，所留下的邮箱会收到通知。

- 结果查看
    - 先看搜到的模板与你的序列一致度(Seq Identity)是否大于30%，如果不大于，此结果就应放弃，
    - 如果序列一致度大于30%，再看swissmodel自带的评分高低，主要是GMQE和QMEAN。

> GMQE ：可信度范围为 0-1，值越大表明质量越好; QMEAN：区间-4-0，越接近0，评估待测蛋白与模板蛋白的匹配度越好。
>
> GMQE（全球模型质量估计）是一种结合目标-模板对齐方式和模板搜索方法的属性的质量估计。所得的GMQE分数表示为0到1之间的数字，反映了使用该对齐方式和模板构建的模型的预期准确性以及目标的覆盖范围。数字越高表示可靠性越高。
>
> QMEAN该模型的得分可与相似大小的实验结构所期望的得分相媲美。0值附近的QMEAN 得分表明模型结构与相似大小的实验结构之间具有良好的一致性。分数为-4.0或以下表示模型的质量较低。

参考资料

- Compute pI/Mw:计算序列等电点和分子量的工具
- ProtParam预测蛋白质基本理化性质
- 蛋白质分析工具
- 蛋白亲疏水性分析工具
- TMHMM 2.0--蛋白质的跨膜螺旋预测
- 如何使用TMpred进行蛋白质跨膜结构预测
- 使用SignalP对蛋白序列进行信号肽预测
- 蛋白质二级结构预测方法
- 如何预测蛋白质三维结构（SWISS-MODEL）
- 蛋白质三维结构预测、结果解读与评分