生 物 信 息 学

实 验 讲 义

生物医学工程学院

基础医学院

2012 年

目　　录

实验一 DNA 序列的基本信息及特征分析

一、实验目的

人类 CD9 抗原基因属于 4 跨膜蛋白超家族，与癌细胞黏附、迁移及增生有关，具有肿瘤转移抑制作用。应用集成工具 EMBOSS 的网页版 EMBOSS Explorer，对人类 CD9 抗原基因（序列号为 AY422198）进行核酸序列组分分析，同时对大肠杆菌基因组序列（U00096）进行 ORF 的分析，从中识别真实的 ORF。

1. 学习和掌握应用 NCBI 查询核酸序列信息。
2. 学习和掌握 EMBOSS 的基本使用方法。
3. 学习和掌握核酸序列 ORF 的性质。

二、实验内容

1. 人类 CD9 基因的序列组分分析

（1）获取序列。打开 NCBI 主页（http://www.ncbi.nlm.nih.gov/），在核酸（Nucleotide）数据库查询人类 CD9 抗原基因（AY422198），将其序列以 FASTA 格式保存到本地计算机上。

（2） 打开 EMBOSS。有 http://emboss.bioinformatics.nl/和 http://genome.csdb.cn/emboss/多个 EMBOSS Explorer 可供选用，任选其一打开链接即可使用。

（3）碱基组成分析。找到程序 compseq，在 Input section 项目中，使用 upload 上传 CD9 的序列；在 Required section 项目中，把 Word size to consider 修改为“1”；其他参数默认即可。最后，点击 Run compseq 获得 ATGC 四种碱基的数目及百分比含量。

（4）GC 含量分析。找到程序 geecee，以上传文件的方式提交 DNA 序列，之后点击 Run geecee 得到 CD9 基因的 GC 含量。

（5）序列转换。找到程序 revseq，如前所述上传 CD9 基因的 DNA 序列，点击 Run revseq 后，得到其反向互补序列。通过调整 Advanced section 项目中的参数，可以仅获得反向序列或互补序列。

（6）比较 AY422198 及其互补序列的组成成分。

2. 大肠杆菌基因组序列的 ORF 分析

（1）获取序列。登录 NCBI 主页（http://www.ncbi.nlm.nih.gov/），在核酸（Nucleotide）数据库查询大肠杆菌基因组序列（U00096），将其以 FASTA 格式保存到本地计算机上。

（2）截取序列。找到 EMBOSS 中的程序 extractseq，在 Required section 项目中，通过设定 Regions to extract 为 1-3000 截取大肠杆菌基因组序列的前 3000bp。

（4）ORF 预测。找到程序 getorf，在序列输入框中粘贴上一步截取的前 3000bp 序列片段，Code to use 中选择“Bacterial”， Type of output 中选择“Nucleic sequences between START and STOP codons”，其他默认即可。之后点击 Run getorf 得到 ORF 的预测结果。

（5）和 NCBI 的 ORF Finder 预测结果进行比较分析。

实验二 原核生物和真核生物的基因识别

一、实验目的

原核生物的基因结构简单，属连续基因，识别过程相对简单；而真核基因的编码区是不连续的，结构比较复杂，需要根据基因的整体结构进行基因预测。应用 GeneMarkS 识别大肠杆菌基因组序列（U00096）中的基因，并用 GENSCAN 分析人类 CD9 抗原基因（序列号为 AY422198）的基因结构。

1. 学习和掌握原核基因和真核基因的结构特征。
2. 学习和了解隐马尔可夫模型在基因识别中的应用。
3. 学习和掌握应用 GeneMarkS、GENSCAN 识别基因。

二、实验内容

1. 大肠杆菌基因组序列的基因识别

（1）获取序列。登录 NCBI 主页（http://www.ncbi.nlm.nih.gov/），在核酸（Nucleotide）数据库查询大肠杆菌基因组序列（U00096），将其以 FASTA 格式保存到本地计算机上。同时保存一份 GenBank 格式的基因信息，用于后续比较。

（2）截取序列。利用 EMBOSS 中的程序 extractseq 截取大肠杆菌基因组序列的前 10000bp。

（3）打开 GeneMarkS 网站。通过搜索引擎搜索 GeneMarkS，或者直接在浏览器的地址栏中输入网址，打开 http://exon.gatech.edu/GeneMark/genemarks.cgi 即可。

（4）基因识别。在 Input sequence 输入框中粘贴截取的前 10000bp 序列片段，其他参数默认即可。点击 Start GeneMarkS 后，在新的页面中点击 gms.out 即可看到基因的识别结果。

（5）结果分析。打开保存的 GneBank 格式文件，找到并查看前 10000bp 中基因的位置信息。将其与 GeneMarkS 的预测结果进行比较分析。

2. 人类 CD9 抗原基因的结构分析

（1） 获取序列。登录 NCBI 主页（http://www.ncbi.nlm.nih.gov/），在核酸（Nucleotide）数据

库查询人类 CD9 抗原基因（AY422198），将其序列以 FASTA 格式保存到本地计算机上。同时保存一份 GenBank 格式的基因信息，用于后续比较。

（2） 打开 GENSCAN 网站。通过搜索引擎搜索 GENSCAN，或者直接在浏览器的地址栏中输入网址，打开 http://genes.mit.edu/GENSCAN.html 即可。

（3） 基因识别。上传人类 CD9 抗原基因的序列，其他参数默认即可。点击 Run GENSCAN 得到基因结构的预测结果。

（4） 结果分析。打开保存的 GneBank 格式文件，找到并查看该基因的 CDS 区域信息。将其与 GENSCAN 的预测结果进行比较分析。

实验三 基于 Galaxy 的基因组数据处理

一、实验目的

Galaxy 是一个开放的、基于网络的生物信息学分析平台，整合了各种生物信息学分析工具，包括基因组注释的常用工具。完全基于 Galaxy 平台，利用各种工具及相关操作，寻找人类基因组中 22 号染色体（chr22）上含有 SNP 数量最多的外显子。

1. 学习和掌握基因组注释中的常用格式。
2. 学习和掌握基因组坐标的逻辑运算模式。
3. 学习和掌握 Galaxy 的基本使用方法。

二、实验内容

1. 打开 Galaxy 网站。通过搜索引擎搜索“Galaxy UCSC”，或者直接在浏览器的地址栏中输入网址，打开 https://main.g2.bx.psu.edu/即可。

2. 获取所需数据。

（1）chr22 上的外显子数据。打开 Get Data 工具集中的 UCSC Main 工具，调整参数提取人类 hg19 基因组中 22 号染色体（chr22）上的所有外显子信息，以 BED 格式进行存储。

（2）chr22 上的 SNP 数据。同样使用 UCSC Main 工具提取 dbSNP37 中 chr22 上的所有 SNP 信息，也以 BED 格式进行存储。

3. 提取含有 SNP 的外显子。打开 Operate on Genomic Intervals 工具集中的 Join 工具，以外显子数据为第一套数据集、SNP 数据为第二套数据集，通过坐标比较提取出含有 SNP 的所有外显子。坐标联合操作的结果中前半部分是外显子的信息，后半部分是 SNP 的信息。结果显示，不少外显子上含有多个 SNP。

4. 对外显子上的 SNP 进行计数。使用 Join, Subtract and Group 工具集中的 Group 工具，针对第 4 列外显子的 ID 进行计数，得到的结果中第一列是外显子的 ID，第二列则是对应外显子上的 SNP 数目。

5. 对 SNP 数目进行排序。使用 Filter and Sort 工具集中的 Sort 工具，对第 2 列中的 SNP 计数

进行排序，采用降序方式，最终含有最多 SNP 的外显子将排在最前面。

6. 筛选至少含有 10 个 SNP 的外显子。使用 Filter and Sort 工具集中的 Filter 工具，使用“c2>=10”条件提取第 2 列 SNP 计数大于等于 10 的所有行，即可筛选出至少含有 10 个 SNP 的外显子。

7. 附加外显子的原始信息。使用 Join, Subtract and Group 工具集中的 Join two Datasets 工具，根据最初外显子数据中的第 4 列和上一步结果中的第 1 列（两者均为外显子的 ID），把两套数据整合起来，恢复外显子的染色体号、起始终止位置等信息。最终结果中，前面部分是外显子的相关信息，最后一列是相应外显子上的 SNP 数目。

8. 尝试对其他染色体或全基因组上的外显子进行类似的分析，进一步熟悉在 Galaxy 中进行基因组注释分析的相关操作。

实验四 蛋白质序列的基本信息分析

一．实验目的

对于蛋白质的序列特征进行分析可以了解蛋白质的基本信息，如蛋白质的分子量、等电点、氨基酸的组成、亲水性和疏水性等。ExPASy 数据库（Expert Protein Analysis System）提供了一系列蛋白质理化分析工具，如 AACompIdent 可以进行蛋白质氨基酸组成信息分析；ProtParam 可以进行蛋白质基本的物理化学参数的计算；ProtScale 可以进行氨基酸亲/疏水性的分析等。

1 学习和掌握蛋白质氨基酸组成分析方法。

2 学习和掌握蛋白质理化性质分析方法。

3 学习和掌握蛋白质亲疏水性分析方法。

二．实验内容

1. AACompIdent 分析蛋白质的氨基酸组成

（1）登陆 AACompIdent 工具网址为 http://www.expasy.org/tools/aacomp/，或从 ExPASy 工具页面选取 AACompIdent 工具。

（2）选择用于匹配的相应氨基酸组群（Constellation）。对于标准方法测定氨基酸组成，选用组群 2。这个组群显示 16 个氨基酸，不考虑半胱氨酸和色氨酸，并且将天冬酰胺（Asn）和天冬氨酸（Asp）一起按照 Asx 计算，谷氨酸（Glu）和谷氨酰胺（Gln）一起按照 Glx 计算。

（3）在 e-mail 地址栏中填写接收发送结果的电子邮件地址，然后向下滚动至“Unknown Protein”区域。在此处注明一个搜索用名称，这也将作为电子邮件的主题；通过实验测得或估算（如蛋白质双向电泳凝胶中的位置）的等电点（PI）和分子质量（Mw），以及它们的准确性误差范围。

（4）输入一个或多个在 Swiss-Prot OS（种）或 OC（分类）中的匹配词使搜索限定在一个或一系列生物种类内（Swiss-Prot 种属缩写列表见 http://www.expasy.org/cgi-bin/speclist），以及在 Keyword from KW lines (for search in Swiss-Prot only) 方框中注明关键字（Swiss-Prot 关键字列表见 http://www.expasy.org/cgi-bin/keywlist.pl）。

（5）在氨基酸组成（Amino acid composition）方框中应注明实验测定的蛋白质各个氨基酸组成，以摩尔百分比（molar percent）形式表示的成分数据。如果在氨基酸分析

步骤中有校准蛋白质（Calibration protein）与未知蛋白质进行平行分析，则校准蛋白质组成可作为未知蛋白质分析步骤中系统误差的补偿矫正。可在“Calibration Protein”区域中，注明校准蛋白质的 Swiss-Prot ID 名称（如，牛血清白蛋白 ALBU_BOVIN）并输入实验测定的蛋白质氨基酸组成以及摩尔百分比形式表示的数据。

（6）选择“Run AACompIdent”将数据上传至 ExPASy 服务器，结果将以电子邮件的形式发送给用户。通过 e-mail 返回的结果中包括三级列表：第一张列表中的蛋白都基于特定的物种分类而不考虑 pI 和分子量；第二张列表包含了不考虑物种分类、pI 和分子量的全体蛋白；第三张列表中的蛋白不但基于特定物种分类，并且将 pI 和分子量也考虑在内。数据库中的所有蛋白按照分数由低（最优匹配）向高（最差匹配）排列。

2. ProtParam 分析蛋白质的理化性质。

(1) Expasy 系统中点击 ProtParam 工具，或直接登陆 ProtParam 网址，其网址为 http://www.expasy.org/tools/protparam.html。

（2）输入某一种蛋白质的 Swiss-Prot/TrEMBL 录入号或 ID，也可以是该蛋白质的氨基酸序列。如 P10599（Thioredoxin ，人硫氧还蛋白）或其蛋白质序列。

（3）点击 Compute Parameters，根据得到的分析结果，熟悉 ProtParam 各个参数及其意义，如分子量、理论等电点、氨基酸组成、原子组成、消光系数、体内半衰期、不稳定指数、脂肪指数和总平均亲水性等。

3. ProtScale 分析蛋白质的亲疏水性。

（1）在 ExPASy 系统中点击 ProtScale 程序或直接登陆其网站，其网址为 http://www.expasy.org/tools/protscale.html。

（2）输入某一种蛋白质的 Swiss-Prot/TrTMBL 的记录号或其氨基酸序列。如 P10599（Thioredoxin ，人硫氧还蛋白）或其蛋白质序列，观察 ProtScale 的预测结果。

（3）ProtScale 默认的标度值为 Hphob.Kyte & Doolittle 标度。选择不同的标度，观察分析结果有何不同。

实验五 蛋白质序列的特征信息分析

一. 实验目的

对蛋白质序列的特征信息进行分析，以更全面的了解蛋白质的性质和功能。蛋白质序列的特征信息包括跨膜区、信号肽、卷曲螺旋等。

1 学习和掌握蛋白质跨膜区的分析方法。

2 学习和掌握蛋白质信号肽的分析方法。

3 学习和掌握蛋白质卷曲螺旋的分析方法。

二. 实验内容

1. TMpred 预测蛋白质跨膜区。

（1）登陆 http://www.ch.embnet.org/software/TMPRED_form.html。

（2）将某一种蛋白质的氨基酸序列粘贴在 TMpred 主操作界面的查询序列框中。如：P35462 蛋白（多巴胺 D3 受体，dopamine D3 receptor），将其氨基酸序列粘贴在 TMpred 主操作界面的查询序列框中，选择跨膜螺旋疏水区的最小长度和最大长度值分别为 17 和 33（软件默认值），输出格式（Output format）选择“html”，输入序列格式（Input sequence format）选择“Plain Text”（纯文本格式），然后点击“Run TMpred”，即可得到 TMpred 软件对 P35462 序列的分析结果。

（3）根据输入序列所得到的可能的跨膜螺旋区，分析由膜内到膜外（inside->outside）的跨膜螺旋有几个；由膜外到膜内（outside->inside）跨膜螺旋有几个，分别是什么。

（4）分析每个跨膜螺旋的得分及中心位点。分析跨膜螺旋在某个方向（膜内向膜外还是膜外向膜内）的偏好性，符号“+”表示这个跨膜螺旋在此方向上有偏好性，符号“++”表示这个跨膜螺旋在此方向有很强的偏好性。

2. SignalP 分析蛋白质信号肽。

（1）登陆 SignalP 网站，网址为 http://genome.cbs.dtu.dk/services/SignalP/。

（2）输入某一种蛋白质的氨基酸序列。如人载脂蛋白 A5（Apolipoprotein A5），其在 UniProtKB/Swiss-Prot 中的编号为 Q6Q788。从 UniProtKB/Swiss-Prot 数据库中下载 Q6Q788 蛋白质序列并粘贴到指定的方框中。

（3）Organism group 参数选择“Eukaryotes”， Method 参数选择“Input sequences may include TM regions”， Graphics 参数选择“PNG（inline）”，Output format 参数选择“Standard”，Optional - User defined D-cutoff values 和 Truncate sequence 两个参数不选择

（即软件默认值），最后点击“Submit”，等待软件运行。

（4）根据返回的结果进行分析。结果中有 3 个分值，分别为 C 分值（C-score）、S 分值（S-score）和 Y 分值（Y-score），理解 3 个分值的意义。

（5）分析 2 个指标及其意义：mean S 和 D 分值。

（6）分析蛋白质序列的剪切位点所在氨基酸的位置。

3. COILS 分析蛋白质卷曲螺旋。

（1）登陆 COILS，其网址为 http://www.ch.embnet.org/software/COILS_form.html。

（2）输入某一种蛋白质的氨基酸序列。如：碱性亮氨酸拉链核因子Ⅰ（Basic leucine zipper nuclear factor 1，GO45_HUMAN）蛋白序列为例，其在 UniProtKB/Swiss-Prot 中的编号为 Q9H2G9。将 UniProtKB/Swiss-Prot 的编号 Q9H2G9 粘贴到查询序列框中。

（3）Input sequence format 参数选择“SwissProt ID or AC”， Window width 参数选择“all”， matrix 参数选择“MTIDK”，2.5fold weighting of positions a,d 参数选择“no”，将“GO45_HUMAN”填写到 Query title 框中，然后点击“Run Coils”按钮，返回结果。

（4）分析 GO45_HUMAN 卷曲螺旋的图形显示结果，理解三条曲线分别代表 Window width 为 14、21、28 时预测出的卷曲螺旋的位置，预测出了几个卷曲螺旋。由于预测时 Window width 参数选值不同，预测出来的卷曲螺旋的位置和大小也有差别。

实验六　四种蛋白质二级结构预测方法比较

一. 实验目的

蛋白质二级结构的预测通常被认为是蛋白结构预测的第一步。蛋白质二级结构是指α螺旋和β折叠等规则的蛋白质局部结构元件。不同氨基酸残基对于形成不同的二级结构元件具有不同的倾向性。按蛋白质中二级结构的成分可以把球形蛋白分为全α蛋白、全β蛋白、α＋β蛋白等几类折叠类型。目前已有多种不同算法用于蛋白质二级结构预测，本次实验选取理论课已经讲解的四种常用预测方法：Chou-Fasman 方法、PHD 方法、NNSSP 方法和 CONCORD 方法，对属于三类不同折叠类型的蛋白质（全α蛋白、全β蛋白和α+β蛋白）二级结构进行预测，并将预测结果与蛋白质真实的二级结构进行对比。

1. 学习和掌握四种常用蛋白质二级结构预测方法（Chou-Fasman 方法，PHD 方法、NNSSP 方法和 CONCORD 方法）的使用。
2. 学习和掌握蛋白质二级结构预测方法的评价方法。

二. 实验内容

3. 从 PDB 数据库中选择分别属于三类蛋白结构（全α蛋白、全β蛋白和α+β蛋白）各 3 个蛋白质，下载氨基酸序列信息（Fasta 格式）和实验测定的二级结构信息。

(1) 登录 PDB 蛋白质结构数据库(www.rcsb.org)，点击数据库界面右侧的“Browse”。

(2) 在新打开的数据库界面中选择“SCOP”标签，打开不同折叠类型蛋白类别显示界面，分别点击全α蛋白、全β蛋白和α＋β蛋白类别名称，将打开属于不同折叠类型的蛋白结构列表显示界面。

(3) 在蛋白结构列表显示界面，选择单聚物蛋白，并点击要预测的蛋白质的名称，打开对于该蛋白结构的详细介绍界面。

(4) 在新打开的界面中选择“Sequence”标签，在界面中点击“FASTA”和“Sequence & DSSP”标签，下载蛋白质的氨基酸序列信息（Fasta 格式）和实验测定的二级结构信息。

(5) 重复以上步骤，分别选择三类不同折叠类型蛋白结构（全α蛋白、全β蛋白和α＋β蛋白）各 3 个蛋白质，下载氨基酸序列信息（Fasta 格式）和实验测定的二级结构信息。

2. 分别运用四种蛋白质二级结构预测方法（Chou-Fasman 方法，PHD 方法、NNSSP 方法和 CONCORD 方法）预测已经选择的九个蛋白质的二级结构。

(1) 登录 Chou-Fasman 方法在线预测服务器的用户界面（http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=misc1）。单击“Program”中的下拉菜单，选择 Chos-Fasman 方法。然后在下方的“Protein sequence”中的空白处粘贴上需要预测的蛋白质氨基酸信息，格式应为 FASTA。最后点击“Submit Sequence”进行预测。

(2) 登录 PHD 方法在线预测服务器的用户界面(http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_phd.html)。输入 Fasta 格式的蛋白质氨基酸序列信息，并对蛋白质二级结构进行预测。

(3) 登录 NNSSP 方法在线预测服务器的用户界面（http://linux1.softberry.com/berry.phtml?topic=nnssp&group=programs&subgroup=propt）。在输入框内输入 Fasta 格式的蛋白质氨基酸序列信息，对蛋白质二级结构进行预测。

(4) 登录 CONCORD 方法在线预测服务器的用户界面（http://helios.princeton.edu/CONCORD）。在输入框内输入 Fasta 格式的蛋白质氨基酸序列信息，对蛋白质二级结构进行预测。

3. 将预测结果与实际二级结构相比较，利用以下公式计算每种方法的预测准确率 P。

$$P = \frac{N\alpha + N\beta}{N} \times 100\%$$

式中 Nα 和 Nβ 分别表示为正确预测 α -螺旋和 β -折叠的氨基酸个数，N 为蛋白质中总氨基酸个数。

4. 分别计算并列表比较四种方法对于每类蛋白质二级结构的预测准确率。

PDB 号	蛋白折叠类型	预测准确率比较			
		Chou-Fasman	PHD 方法	NNSSP 方法	CONCORD

实验七　使用同源建模软件 Modeller 预测蛋白质空间结构

一. 实验目的

同源建模方法是蛋白质空间结构预测的基本方法，该方法的基本理论依据是具有相似氨基酸序列的蛋白质之间具有相似的空间结构，如果要预测一个蛋白质的空间结构可以通过与之序列相似性较高的蛋白质结构进行预测。本次实验通过使用同源建模软件 Modeller 中两种不同预测策略来预测蛋白质空间结构。

4．学习和掌握运用 Modeller 软件基于单模板方法预测蛋白质空间结构。

5．学习和掌握运用 Modeller 软件基于多模板方法预测蛋白质空间结构。

二. 实验内容

6．登录 NCBI 数据库（http://www.ncbi.nlm.nih.gov/protein/），在搜索框中输入待预测蛋白质的名称，下载得到该蛋白质氨基酸序列信息（Fasta 格式）。

7．登录 NCBI 数据库的 BLAST 网站（http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome），运用 BLAST 方法依据蛋白质序列信息从 PDB 数据库中获得与之序列相近的蛋白质信息。

8．从比对结果中分别选择序列相似度最高的蛋白质和较高的三个蛋白质作为同源建模的模板，记录相应的 PDB 号码。

9．登录 PDB 蛋白质结构数据库(www.rcsb.org)，根据蛋白质的 PDB 号码下载得到蛋白质的 PDB 结构文件（PDB File）。

10．　使用单模板蛋白结构，运用 Modeller 软件基于氨基酸序列信息构建蛋白结构。

（1）将待预测蛋白质的序列信息编辑成 Modeller 软件识别的 PIR 格式序列文件，并命名为 PP.ali ，该文件格式如下图：

```
>P1;TvLDH
sequence:TvLDH:::::::0.00: 0.00
MSEAAHVLITGAAGQIGYILSHWIASGELYGDRQVYLHLLDIPPAMNRLTALTMELEDCAFPHLAGFVATTDPKA
AFKDIDCAFLVASMPLKPGQVRADLISSNSVIFKNTGEYLSKWAKPSVKVLVIGNPDNTNCEIAMLHAKNLKPEN
FSSLSMLDQNRAYYEVASKLGVDVKDVHDIIVWGNHGESMVADLTQATFTKEGKTQKVVDVLDHDYVFDTFFKKI
GHRAWDILEHRGFTSAASPTKAAIQHMKAWLFGTAPGEVLSMGIPVPEGNPYGIKPGVVFSFPCNVDKEGKIHVV
EGFKVNDWLREKLDFTEKDLFHEKEIALNHLAQGG*
```

PIR 格式的序列文件有三部分组成：

① 文件第一行显示序列标记，其中“;”之后的名称应与待预测蛋白名称保持一致。

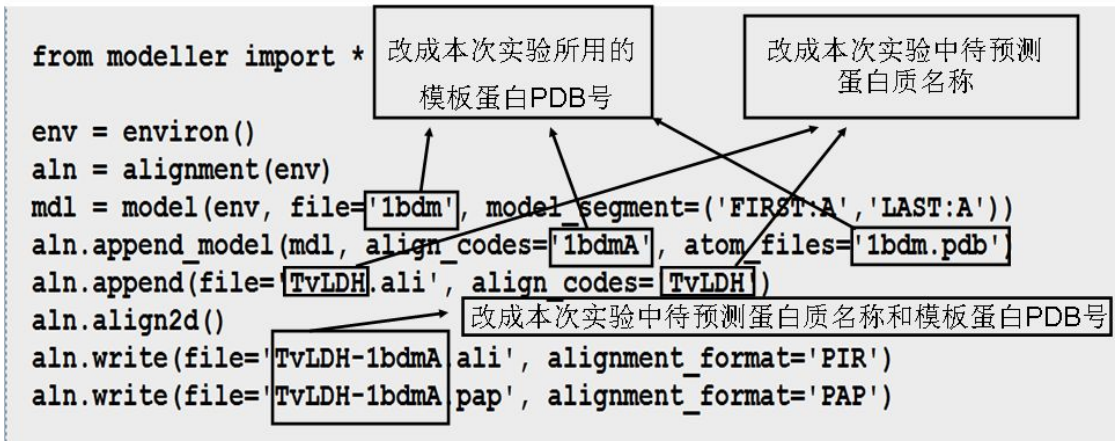
② 文件第二行由九个“:”组成。其中第一个“:”之前的 sequence 表示该文件中的蛋白质序列是待预测蛋白质的序列。前两个“:”之间的名称应与待预测蛋白名称保持一致。第八个和九个“:”之后的数字代表蛋白质结构的参数。由于该文件是待预测蛋白质的序列信息，无蛋白结构存在，因此数字取值为 0。

③ 从文件的第三行开始，是待预测蛋白质的序列信息。其中每行包含的最大字符数是 75 个，氨基酸序列结束的位置用“*”标记。

（2）待预测蛋白质的氨基酸序列与模板蛋白的序列及结构进行比对

① 将文件 align2d.py,模板蛋白的 PDB 文件和前一步得到的待预测蛋白的序列文件拷贝到同一文件夹下。

② 用写字板软件将文件 align2d.py 打开进行修改，修改内容如下图所示。



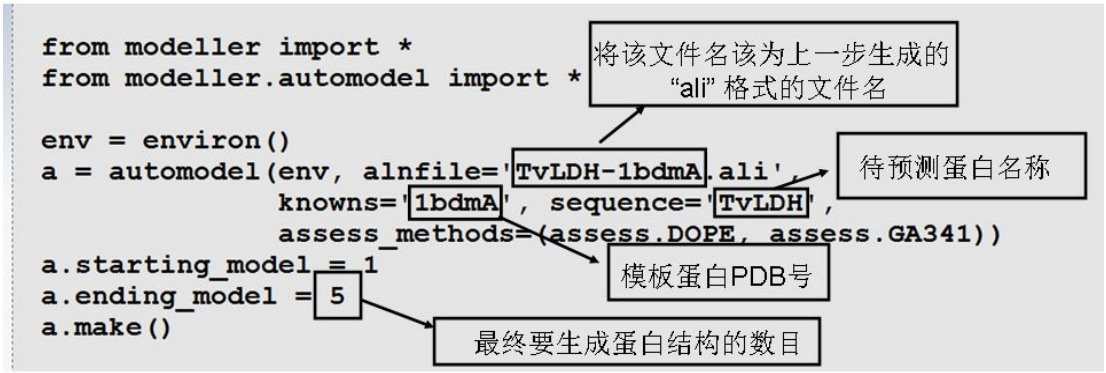
③ 打开 modeller 软件,用”cd”命令将当前目录调整到保存以上文件的文件夹下，在 DOS 输入界面中输入以下命令：

mod9.10　align2d.py

该命令执行后将在所在文件夹下产生两个新文件，文件后缀名分别是“ali”和“pap”，两个文件即是预测蛋白和模板蛋白的比对文件。

（3）构建待预测蛋白的结构

① 将文件 model-single.py 拷贝到当前文件夹下，并用写字板软件将该文件打开进行修改，修改内容如下图：



② 在 DOS 输入界面中输入以下命令：

mod9.10　model-single.py

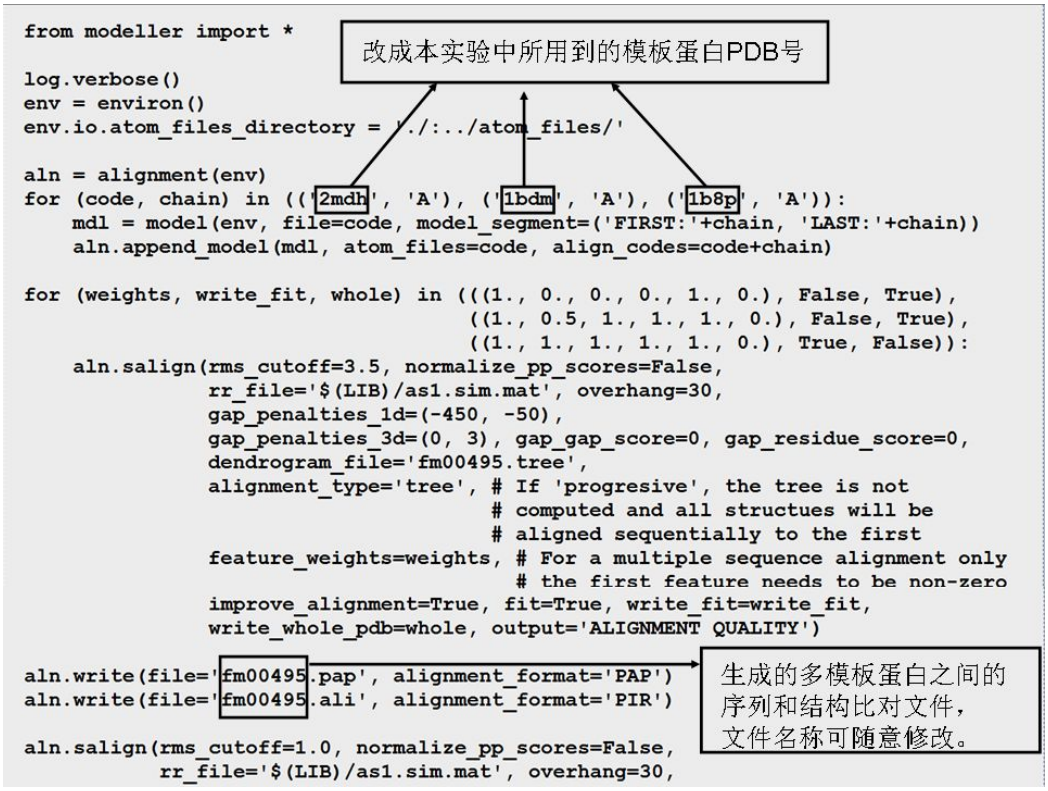
该命令执行后将在所在文件夹下产生后缀名为“pdb”的预测蛋白结构文件。

11.　　使用三个模板蛋白结构，运用 Modeller 软件基于氨基酸序列信息构建蛋白结构。

（1）多条模板蛋白之间的序列及结构进行比对

① 将文件 salign.py,三个模板蛋白的 PDB 文件拷贝到同一文件夹下。

② 用写字板软件将文件 salign.py 打开进行修改，修改内容如下图所示。



③ 在 DOS 输入界面中输入以下命令：

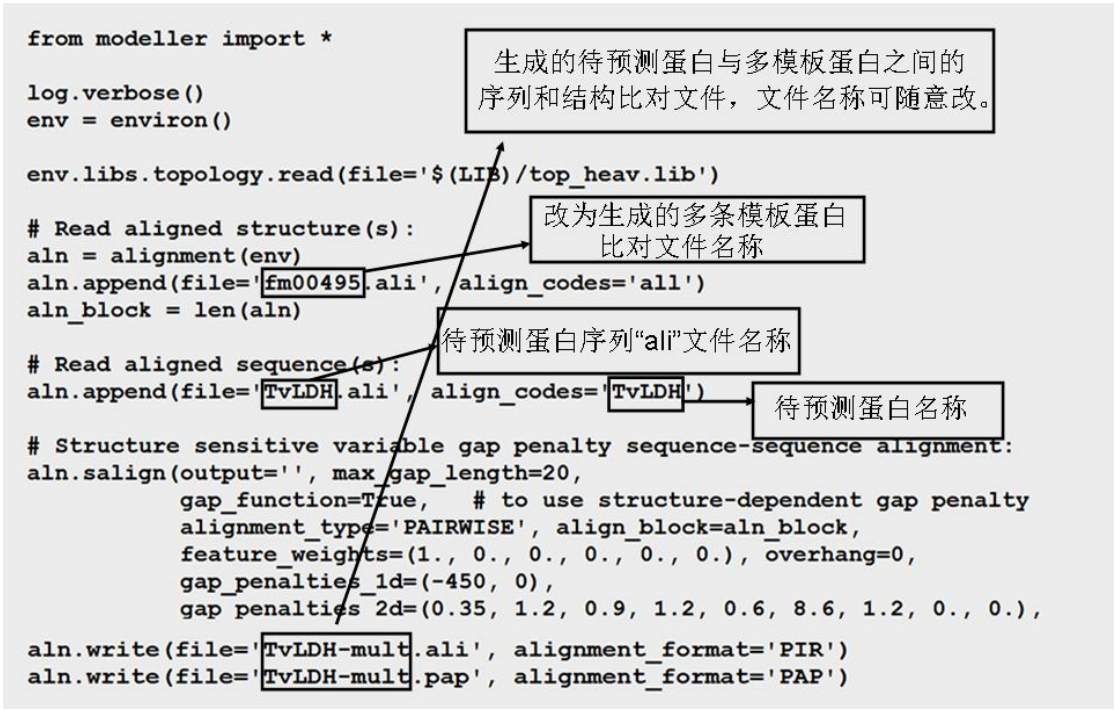
mod9.10　salign.py

该命令执行后将在所在文件夹下产生两个新文件，文件后缀名分别是“ali”和“pap”，两个文件即是多个模板蛋白之间的比对文件。

（2）待预测蛋白与多模板蛋白进行比对

① 将文件 align2d-mult.py,待预测蛋白的序列文件拷贝到前一步文件夹下。

② 用写字板软件将文件 align2d-mult.py 打开进行修改，修改内容如下图所示。




```
from modeller import *

log.verbose()
env = environ()

env.libs.topology.read(file='$(LIB)/top_heav.lib')

# Read aligned structure(s):
aln = alignment(env)
aln.append(file='fm00495.ali', align_codes='all')
aln_block = len(aln)

# Read aligned sequence(s):
aln.append(file='TvLDH.ali', align_codes='TvLDH')

# Structure sensitive variable gap penalty sequence-sequence alignment:
aln.salign(output='', max_gap_length=20,
           gap_function=True,   # to use structure-dependent gap penalty
           alignment_type='PAIRWISE', align_block=aln_block,
           feature_weights=(1., 0., 0., 0., 0., 0.), overhang=0,
           gap_penalties_1d=(-450, 0),
           gap_penalties_2d=(0.35, 1.2, 0.9, 1.2, 0.6, 8.6, 1.2, 0., 0.),

aln.write(file='TvLDH-mult.ali', alignment_format='PIR')
aln.write(file='TvLDH-mult.pap', alignment_format='PAP')
```

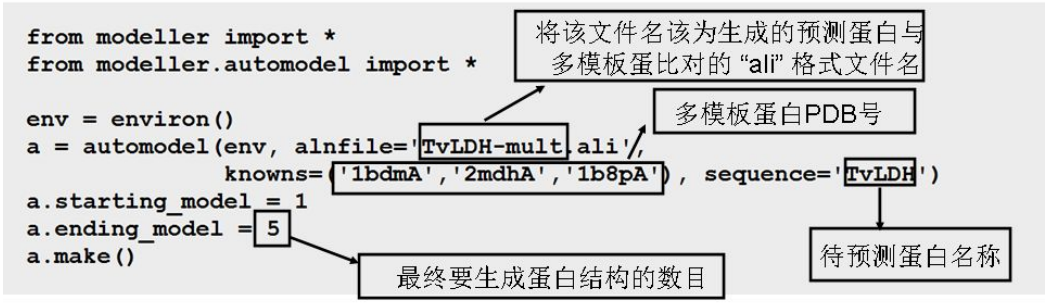
③ 在 DOS 输入界面中输入以下命令：

Mod9.10　align2d-mult.py

该命令执行后将在所在文件夹下产生两个新文件，文件后缀名分别是“ali”和“pap”，两个文件即是待预测蛋白与多个模板蛋白之间的比对文件。

（3）基于多模板蛋白构建待预测蛋白的结构

① 将文件 model-mult.py 拷贝到当前文件夹下，并用写字板软件将该文件打开进行修改，修改内容如下图：



② 在 DOS 输入界面中输入以下命令：

Mod9.10　model-mult.py

该命令执行后将在所在文件夹下产生后缀名为“pdb”的预测蛋白结构文件。

7.利用 RasWin 软件，显示预测得到的蛋白质结构。

① 打开 RasWin 软件，点击工具栏上“File”—“Open”，在弹出对话框中选择实验产生后缀名为“pdb”的预测蛋白结构文件，将蛋白结构文件读入到 Raswin 软件中。

② 点击工具栏上“Display” — “Ribbons”，将蛋白结构的显示方式为条带式。

③ 点击工具栏上“Colours” — “Structure”，将蛋白结构按照二级结构进行染色。

④ 点击工具栏上“File” — “Print”，打印蛋白结构到PDF文件中。

实验八 基因芯片数据的可视化分析

一. 实验目的

基因芯片可以大规模测定基因表达状态。由于这类实验所产生的数据量大，而且数据结构复杂，对数据进行严格的质量控制和分析是保证结果有效性和可靠性的关键步骤。通过图像显示的方法对基因表达数据进行可视化分析，可以快速直观地检验实验数据质量和数据间的关系，对于从总体上把握实验效果和制定进一步的分析策略是很有用的。

本实验以 MATLAB 的 Bioinformatics 工具箱为平台，通过对一组基因芯片测定的基因表达数据进行一系列的可视化分析，学习和掌握有关的分析工具。

1. 学习和掌握在 MATLAB 平台上读取基因芯片数据，获取数据结构信息并提取特定数据列的方法。

2. 学习和掌握在 MATLAB 平台上应用 Bioinformatics 工具箱显示和分析基因芯片数据空间分布图像的方法。

3. 学习和掌握在 MATLAB 平台上应用 Bioinformatics 工具箱对基因芯片数据进行统计分析和显示的方法。

二. 实验内容

1. 本实验分析一组用基因芯片测定的帕金森症小鼠模型的脑部基因表达数据。实验和分析的完整报告可参考 Vanessa Brown et al., Genome Research 12: 868-884 (2002)。完整的基因芯片实验数据可以从以下网址获取：

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30

本实验分析上述研究工作的部分数据，即小鼠 A1 脑区对应的数据。数据分两组，其中一组来自帕金森症(以下简称 PD)小鼠大脑，另一组来自普通对照组(以下简称 WT)小鼠。芯片采用双通道双染色芯片模式，A1 脑区对应的 RNA 样品用 Cy3(绿色荧光染料；对应波长为 532nm)标记，而对照样品(正常小鼠整个大脑中提取的 RNA)用 Cy5(红色荧光染料；对应波长为 635nm)标记。

2. 在 MATLAB 平台上读取基因芯片数据并检查数据结构。

① 用“gprread”函数读取 A1 区 PD 组数据：

```
pd = gprread('exp_mouse_a1pd.gpr');
```

② 检查数据结构：

芯片数据的基本信息：Pd

芯片数据每一列对应的信息：pd.ColumnNames

任意行对应的基因名称(以前20行为例)：pd.Names(1:20)

3. 可以通过基因芯片表达数据的伪色彩图来直观显示芯片不同位置对应的基因表达信号强度。“maimage”函数可以实现此功能。

① 红色染料对应的表达强度数据列中值(所对应列名称为 F635 Median), 在 MATLAB 命令窗口输入：

```
Figure；
maimage(pd,'F635 Median')；
```

② 绿色染料对应的表达强度数据列中值(所对应列名称为 F532 Median), 在 MATLAB 命令窗口输入：

```
Figure；
maimage(pd,'F532 Median')；
```

③ 红色染料对应的背景数据列中值(所对应列名称为 B635 Median), 在 MATLAB 命令窗口输入：

```
Figure；
maimage(pd,'B635 Median')；
```

④ 绿色染料对应的背景数据列中值(所对应列名称为 B532 Median), 在 MATLAB 命令窗口输入：

```
Figure；
maimage(pd,'B532 Median')；
```

⑤ 以上为帕金森症小鼠(PD)对应的数据，按照同样步骤读取对照组的芯片数据：

```
wt = gprread('exp_mouse_a1wt.gpr')；
```

检查数据结构，确保 wt 和 pd 有同样的数据结构。

⑥ 同样用伪色彩图检查数据分布，可以把多幅图形同时显示。

```
Figure；
subplot(2,2,1);
maimage(wt,'F635 Median')；
subplot(2,2,2);
maimage(wt,'F532 Median')；
subplot(2,2,3);
maimage(wt,'B635 Median')；
subplot(2,2,4);
maimage(wt,'B532 Median')；
```

4. 基因芯片数据的统计分析

① 这类芯片表面被分成 8 个较小的区域，成为 Block。可以用”maboxplot”函数检查每个 block 对应的数据的统计分布。

```
Figure；
subplot(2,1,1)；
maboxplot(pd,'F532 Median','title','Parkinson''s Disease Model Mouse')；
subplot(2,1,2)；
maboxplot(pd,'B532 Median','title','Parkinson''s Disease Model Mouse')；
figure；
subplot(2,1,1)；
maboxplot(wt,'F532 Median','title','Untreated Mouse')；
subplot(2,1,2)；
maboxplot(wt,'B532 Median','title','Untreated Mouse')；
```

可以看出，不论是 PD 组还是 WT 组，红色染料通道对应的数据分布比较均一，而绿色染料通道对应的数据，在不同 block 之间有明显差别。

② 两个通道数据的比较

上面我们分析了每个通道对应的信号强度(即 F635 Median 和 F532 Median 列对应的数值)，以及每个通道对应的背景噪音强度(即 B635 Median 和 B532 Median 列对应的数值)。而真正的基因表达数据应该是信号强度减掉背景噪音后的数值。在所读入的基因芯片数据中，数据列“F635 Median - B635”和“F532 Median - B532”分别为红色和绿色染料通道对应的去掉背景噪音的表达强度数值。

先找到这两列在数据表中的位置：

```
cy5DataCol = find(strcmp(pd.ColumnNames,'F635 Median - B635'))；
cy3DataCol = find(strcmp(pd.ColumnNames,'F532 Median - B532'))；
```

把数据转移到相应的变量中：

```
cy5Data = pd.Data(:,cy5DataCol);
cy3Data = pd.Data(:,cy3DataCol);
```

由于数据分布范围比较广，需要用“maloglog”函数来显示其分布：

```
Figure；
maloglog(cy5Data,cy3Data)；
xlabel('F635 Median - B635 (Control)');
ylabel('F532 Median - B532 (Voxel A1)');
```

如果表达数据为零、负数或很小的正数，则没有生物学意义，而且在对数图中也不能正常显示，需要过滤掉。在本实验中，小于或等于 10 的表达数据都被过滤掉。

```
threshold = 10;
badPoints = (cy5Data <= threshold) | (cy3Data <= threshold);
cy5Data(badPoints) = [];
cy3Data(badPoints) = [];
```

然后再次检查数据分布：

```
maloglog(cy5Data,cy3Data);
xlabel('F635 Median - B635 (Control)');
ylabel('F532 Median - B532 (Voxel A1)');
```

③ 数据的标准化

可以用“manorm”函数进行数据标准化：

```
normcy5 = manorm(cy5Data);
normcy3 = manorm(cy3Data);
```

标准化的数据可以用函数“mairplot”来显示基因表达水平的强度-比值之间的关系。

```
mairplot(normcy5,normcy3,'labels',wt.IDs(~badPoints), 'factorlines',2);
```

点击每个点可以看到相应基因的名称，也可以通过图中的标尺直线筛选表达水平变化的基因。