2021年生物信息学期末作业

营养与健康研究所 刘可钦　202028019230006

1.请阐述人类基因组计划和精准医疗计划的各自意义.每个计划请写出你认为最重要的五点（每点字数不少于100字）

人类基因组计划（Human Genome Project, HGP）是一项于1985年提出、于2005年完成的规模宏大，跨国跨学科的科学探索工程。人类基因组计划旨在测定组成人类染色体的30亿个碱基对的核苷酸序列，从而绘制人类基因组图谱，并辨识其载有的基因及其序列，达到破译人类遗传信息的最终目的。人类基因组计划与“曼哈顿原子弹计划”和“阿波罗计划”并称为三大科学计划，是人类科学史上的又一个伟大工程，被誉为生命科学的“登月计划”。而我认为，人类基因组计划的重大意义主要在于以下五点：

1. 人类基因组计划创造的技术和资源对整个生命科学的研究产生了重大影响。

   人类基因组序列开启了对大多数人类基因组"元件"的全面发现，同时为寻找并发现遗传编码系统内的其他重要元件奠定了基础（如发现非编码调节RNA等）。由于理解一个复杂的生物系统需要了解各个部分，它们之间的连接方式，它们的动态变化以及所有这些与功能的关系；所以这些元件的发现对于 "系统生物学 "的出现至关重要，它改变了我们对生物学和医学的研究方法。

   人类基因组计划的完成还导致了蛋白质组学的出现，这是一门专注于识别和量化存在于离散生物区间的蛋白质的学科。蛋白质构成了生物体基因组部件列表中的细胞特定功能，而人类基因组计划通过提供人类蛋白质组中所有胰蛋白酶的参考序列和预测质量，促进了质谱分析这一关键分析工具的使用。这种基于质谱技术的蛋白质组的可及性推动了引人注目的新应用，如定向蛋白质组学。

2. 人类基因组计划开启了颠覆传统的“学科交叉”科研模式。

   由于蛋白质组学需要极其复杂的计算技术，人类基因组计划同时也推动了复杂的数据计算和数学方法的发展。因此使计算机科学家、数学家、工程师和理论物理学家与生物学家通力合作，促进了更多的跨学科交叉研究。事实上，人类基因组计划从根本上引发了研究方式的变革，甚至可以说是一场科研方式的革命。过去主流的研究方式是PI制下的假设驱动研究（Hypothesis-driven research），而人类基因组计划的成果则推动这一研究制度逐渐转变为大数据驱动的新型科研模式（big-data-driven research）。

   在人类基因组计划带来的新型科研模式的推动下，数千种疾病致病机理的发现、癌症的重新定义和精准个体化治疗、基因编辑技术和干细胞治疗的应用等重大成果接踵而至，一系列精彩纷呈的生物医学成就把人类带入了一个全新的世界。

3. 人类基因组计划促进了数据开放分享。

   人类基因组计划改变了生物医学研究中数据分享的旧例。人类基因组计划主要领导人同意在基因序列生成后的24小时内，将达到一定规模的基因组序列整合提交到一个公共数据库。2014年，一项全面的基因组数据分享政策（Genomic Data Sharing Policy）开始实施，要求几乎所有在NIH（美国国立卫生院）基金资助下获得或分析的基因组数据都必须拿出来共享。

事实上，人类基因组计划普及了在GenBank和UCSC基因组浏览器等用户友好型数据库中立即向公众提供数据的理念。此外，人类基因组计划还推广了开源软件的理念，即程序的源代码可供有兴趣扩展和改进的人编辑。Linux的开源操作系统和它所催生的社区已经显示了这种方法的力量。数据的可及性是未来生物学文化和成功的一个关键概念，因为"数据的民主化"对于吸引研究人员专注于具有内在复杂性的生物系统的挑战性问题至关重要。这在医学上将更为关键，因为科学家需要获得每个人的数据云，以挖掘未来的预测性医学——这对未来医学来说意义重大。

不过，人类基因数据的分享也带来了新挑战，例如海量数据在计算、传输上的难题，以及如何保护研究中被试者的隐私。研究者们仍然需要不断探索新方案以解决不断出现的新问题。

4. 人类基因组计划带来了“优先技术发展”的共识。

   上世纪九十年代初期，人类基因组计划刚一启动，核心科学家就清醒地认识到，必须把基因组测序与绘制图谱的工具与方法作为一个更大的项目去开发。在过去，人们常常认为，是先有科学发现再有技术发明，科学发现和技术发明存在单向关系，其实并非如此，大多数惊人的科学发现都是由新型技术发明所推动的，比如，由于玻璃磨制技术的改进，发明出了望远镜，天文学才得以空前进步。事实上，有了DNA测序技术的进步和工具的发展，基因组科学才有如此惊人的突破性发展。参与人类基因组计划的研究者们把多种渐进性的技术创新整合在一起，取得了革命性的进步，从毛细管电泳法到桑格DNA测序法就是一个极好的例证，DNA测序法最终成为绘制人类首个基因组图谱的关键技术。工具和技术不断创新的同时，该计划也催生了无数的遗传学技术，并引发了分子生物学、化学、物理学和计算科学的重大革新。此后，随着新一代高通量测序技术的发展，基因组测序价格逐步下降，时至今日已开始成为大多数实验室能够负担的科研项目；而随着单细胞测序技术的出现与发展，可供研究的基因组数据立刻增长至海量。

5. 人类基因组计划的成功使各国研究者意识到合作型大型科学项目的重要性。

   此外，人类基因组计划的成功引领了未来一系列大型科学计划的发展。人类基因组计划的设想和实施是生物学中大型科学计划的第一个成功例子，它清楚地表明了这种方法在处理其综合生物和技术目标时的力量和必要性。人类基因组计划的特点是有一套明确的目标和实现这些目标的计划；受资助的研究人员数量有限，通常围绕中心或联合体组织；承诺公开数据/资源；需要大量资金支持项目基础设施和新技术开发。大型科学计划和范围较小的面向个人研究者的科学具有强大的互补性，因为前者产生的资源是所有研究者的基础，而后者对具体问题进行了详细的实验澄清，并对大科学产生的数据进行了分析深度和细节。所以，大科学项目对于全面、综合地解决生物学和医学研究的复杂性至关重要。

   而人类基因组计划打破了研究学者独自探索，为少数几个科学问题寻求答案的格局，这是势所必然的。同时，还一反“假说驱动研究”（Hypothesis-driven research）的科研传统，而是基于大人群和海量样本、数据开展研究。人类基因组计划汇集了来自六个国家、多种学科背景、不同资历的2000多名研究人员，他们分成若干小组，经费来源多种多样，但最后的成功都源自几个共同点：各小组负责人的有力领导、对这项工作重要性的共识、以及研究者们甘愿为了共同利益而放弃个人成就的意愿。此后完成的重大基因组学项目，包括“千人基因组项目”（the 1000 Genomes Project），鉴定致癌突变的“癌症基因组图谱项目”（Cancer Genome Atlas），以及“人类微生物组计划”（Human Microbiome Project），都继承并实践了这一优秀理念。当前，由于新冠疫情，各国间的交流壁垒日益增厚，因此更加需要强调大科学研究中开放合作的重要性。

精准医疗是“一种新兴的疾病治疗和预防方法，它考虑到每个人的基因、环境和生活方式的个体差异”。精确医学指的是根据每个病人的个人特征来定制医疗。它并不意味着创造出对病人来说独一无二的药物或医疗设备，而是指将个人划分为不同的亚群的能力，这些亚群在对特定疾病的易感性、可能发生的疾病的生物学或预后、或对特定治疗的反应方面有所不同。然后，预防或治疗干预措施可以集中在那些将

受益的人身上，为那些不会受益的人节省费用和副作用。尽管""个性化医疗""一词也被用来表达这一含义，但该词有时被误解为暗示可以为每个人设计独特的治疗方法。这种方法将使医生和研究人员能够更准确地预测哪种特定疾病的治疗和预防策略会对哪些人群有效。精准医疗与传统医疗"一刀切"的方法相反，后者是为普通人制定疾病治疗和预防策略，较少考虑个人之间的差异。在我看来，精准医疗计划的重大意义在于以下五点：

1. 精准医疗的实施有助于改进预防、诊断和治疗各种疾病的方法，也有助于更好地理解各种疾病发生的基本机制。

   在精准医疗蓬勃发展的十年中，精准医学领域已经取得了一系列重大发展。有关生物标志物的研究发现了可以解释长期学习现象的标志物，可以帮助理解多动症和注意力缺失症等症状。而药物基因组学的治疗领域则发现了负责确定影响病人药物反应的基因，这可以帮助将病人的基因特征与他们的可变药物反应联系起来。而从血液转录组学和代谢组学获得的数据则可被用于开发模型，用以帮助了解免疫系统和途径，从而治疗与之相关的疾病。这些成就可能对许多过去难以诊断和治疗的神经系统疾病和其他慢性疾病的患者产生深远影响。

2. 对于病人个体而言，精准医学在诊疗中可以更好地整合电子健康记录，这将使医生和研究人员更容易获得医疗数据。而由于医生有更广泛的能力使用病人的基因和其他分子信息作为常规医疗的一部分，这可以提高预测针对病人个体有效治疗方法的能力。

   精准医学的基本原理是，根据基因组结构、生活方式和环境条件，一种特定的疾病在不同的人身上会产生不同的身体症状。例如，两个不同种族的人对一种药物治疗的反应可能不一样。精准医疗的"医疗模式 "旨在根据上述三个管理因素来定制治疗，这也是个人对疾病反应的特点。系统生物学和泛基因组学的使用通过在分子水平上诊断疾病来帮助这一过程。一旦确定了病因，就可以给他们开出定制的药物。这样的个性化医疗摆脱了医生以前所坚持的 "一刀切 "模式。

3. 精准医疗的实施可以改进食品药品监督管理局对测试、药物和其他技术的监督，以支持创新，同时确保这些产品的安全和有效。

   虽然为每个人定制药物可能很昂贵，但收集个人的基因构成可能有助于将他们归类。然后，这些群体将有自己的定制治疗。收集人口的基因组信息也将有助于未来的人口健康研究，这可以帮助在早期阶段预防疾病。精准医学还可以通过减少重复用药和减轻相关的副作用来降低治疗成本。在分子水平上治疗疾病将帮助生物学家了解其发生背后的根本原因，并相应地开发药物来彻底根除它们。基因靶向可以推动药物的开发，而不是漫长而艰巨的试验和错误程序。

4. 精准医疗计划需要设计新的工具来建立、分析和分享大量的医疗数据集，这将会引发一系列的技术进步；同时，也会为各种专业的科学家，以及来自患者权益团体、大学、制药公司和其他方面的人员带来新的合作研究关系。

   虽然精准医疗在医疗保健领域具有潜在优势，但在实施方面也有一些伦理问题的限制。由于数据需要在不同的平台上共享，因此也需要技术手段确保成千上万人的遗传信息安全。现在，新兴的区块链技术和人工智能为这种网络安全防护提供了可能。在处理人口基因组结构的巨大数据时，在分布式网络和共享账簿上工作的区块链技术可以用来确保数据的安全性和使用的道德性，同时防止误操作。而人工智能则可以实施，以分析数据中的模式，并向医疗专业人士提供关于个人状况的见解。基因组测序也可以利用人工智能加速进行。人工智能还可用于聚类，并将人口隔离成群体，以帮助生产定制药物。

5. 精准医疗计划的实施将会引发新的法规和政策出台，从而保护研究参与者的健康数据隐私，特别是病人的个体隐私和他们数据的保密性，而这可以确保法律层面上对患者遗传信息的隐私保护。

针对患者个体的精准医疗是信息密集型的。利用基因组学和其他“全息”技术创建的高维数据是个性化医疗的许多预测、诊断和治疗应用的核心。然而，这种方法所需的个人健康信息的大量增加也是个性化医疗的伦理、法律和社会问题的主要来源之一。在临床上利用基因组信息的能力在很大程度上取决于卫生信息技术。电子健康档案和电子档案网络正在许多开始试行精准医疗的发达国家被广泛采用。传统上只由医疗服务提供者掌握的健康信息，也越来越多地被个人（以个人健康记录的形式）和第三方（根据病人签署的授权获得）所掌握。这引发了一系列的法律与伦理层面的讨论。而在此前，针对这类遗传信息数据的保护法案尚不全面，而精准医疗的提出将会推动新的法律法规出现，从而更好地保护患者隐私权。

2.请列举十种你在本课程中所学到的生物信息学算法或模型的基本概念、适用生物学问题和适用数据及工具（每种不少于200字）

1. 尼德曼-翁施算法（Needleman-Wunsch Algorithm）

   Needleman-Wunsch (NW) 算法常用于基因序列匹配以及单词匹配，是基于生物信息学的知识来匹配蛋白序列或者DNA序列的算法，被认为是一种实现全局最优匹配的动态规划算法。 NW算法也被称为优化匹配算法和整体序列比较法其最差情况下的时间复杂度是$O(mn)$, 空间复杂度是$O(mn)$，其中m和n是两个序列的长度。

   NW算法步骤一般为：

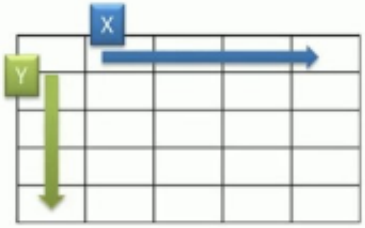
   构造一个打分矩阵；

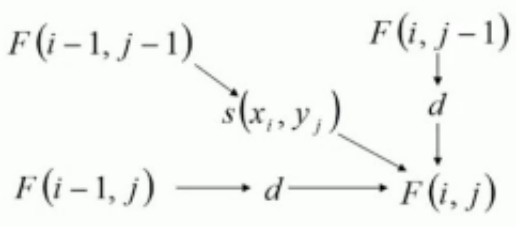
   设计得分矩阵。例如：第i行第j列字母匹配+1，不匹配-1，插入和删除（字母与空白对比）操作-1；

   将以上表格的第二行第二列的初始得分设为0，通过公式计算填充整个表格，并记录得分的方向。特别的，当限制每个网格的最低分为0时，此时NW算法就变成了寻找局部最优匹配的Smith-Waterman算法；

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) & x_i \text{ aligned to } y_j \\ F(i-1,j) + d & x_i \text{ aligned to a gap} \\ F(i,j-1) + d & y_j \text{ aligned to a gap} \end{cases}$$





   从表格的右下角开始，根据之前记录的路径逐级返回，并通过对应的操作得到最优匹配的序列。

   python程序实现为：

```
# -*- coding: utf-8 -*-
```

```
import itertools
import copy
seq1 = ""#输入序列1
seq2 = ""#输入序列2

#定义得分矩阵
score_dic = {'match': 1, 'mismatch': -1, 'gap': -1}
# initial score_matrix
score_matrix = [[0 for column in range(len(seq1))] for row in
range(len(seq2))]
trace_back = [[[]for column in range(len(seq1))] for row in
range(len(seq2))]
for i in range(len(score_matrix[0])):
    score_matrix[0][i] = i * -1
    if i > 0:
        trace_back[0][i].append('left')
for i in range(len(score_matrix)):
    score_matrix[i][0] = i * -1
    if i > 0:
        trace_back[i][0].append('up')
trace_back[0][0].append('done')
# fill the table
for i in range(1, len(score_matrix)):
    for j in range(1, len(score_matrix[0])):
        if seq1[j] == seq2[i]:
            char_score = score_dic['match']
        else:
            char_score = score_dic['mismatch']
        top_score = score_matrix[i - 1][j] + score_dic['gap']
        left_score = score_matrix[i][j - 1] + score_dic['gap']
        diag_score = score_matrix[i - 1][j - 1] + char_score
        score = max(top_score, left_score, diag_score)
        score_matrix[i][j] = score
        if top_score == score:
            trace_back[i][j].append('up')
        if left_score == score:
            trace_back[i][j].append('left')
        if diag_score == score:
            trace_back[i][j].append('diag')

# 根据结果计算最优匹配的序列
# pointer = [seq2_index, seq1_index]
pointer = [len(score_matrix) - 1, len(score_matrix[0]) - 1]
align_seq1 = []
align_seq2 = []
arrow = trace_back[pointer[0]][pointer[1]]


def seq_letter_finder(current_arrow, current_pointer):
```

```
    if current_arrow == 'diag':
        letter = [seq1[current_pointer[1]], seq2[current_pointer[0]]]
        next_pointer = [current_pointer[0] - 1, current_pointer[1] - 1]
        next_arrow = trace_back[next_pointer[0]][next_pointer[1]]
        return letter, next_arrow, next_pointer
    elif current_arrow == 'left':
        letter = [seq1[current_pointer[1]], '-']
        next_pointer = [current_pointer[0], current_pointer[1] - 1]
        next_arrow = trace_back[next_pointer[0]][next_pointer[1]]
        return letter, next_arrow, next_pointer
    else:
        letter = ['-', seq2[current_pointer[0]]]
        next_pointer = [current_pointer[0] - 1, current_pointer[1]]
        next_arrow = trace_back[next_pointer[0]][next_pointer[1]]
        return letter, next_arrow, next_pointer


def align_seq_finder(rec_arrow, rec_pointer, rec_ls):
    if rec_arrow[0] == 'done':
        rec_ls = [rec_ls[0][::-1], rec_ls[1][::-1]]
        return rec_ls
    else:
        if len(rec_arrow) == 1:
            letter, rec_arrow, rec_pointer =
seq_letter_finder(rec_arrow[0], rec_pointer)
            rec_ls[0] += letter[0]
            rec_ls[1] += letter[1]
            return align_seq_finder(rec_arrow, rec_pointer, rec_ls)

        elif len(rec_arrow) == 2:
            arrow1 = copy.deepcopy(rec_arrow[0])
            pointer1 = copy.deepcopy(rec_pointer)
            ls1 = copy.deepcopy(rec_ls)
            arrow2 = copy.deepcopy(rec_arrow[1])
            pointer2 = copy.deepcopy(rec_pointer)
            ls2 = copy.deepcopy(rec_ls)
            letter1, arrow1, pointer1 = seq_letter_finder(arrow1, pointer1)
            letter2, arrow2, pointer2 = seq_letter_finder(arrow2, pointer2)
            ls1[0] += letter1[0]
            ls1[1] += letter1[1]
            ls2[0] += letter2[0]
            ls2[1] += letter2[1]
            return list(itertools.chain(align_seq_finder(arrow1, pointer1,
ls1),
                                        align_seq_finder(arrow2, pointer2,
ls2)))
        else:
            arrow1 = copy.deepcopy(rec_arrow[0])
            pointer1 = copy.deepcopy(rec_pointer)
```

```
            pointer2 = copy.deepcopy(rec_pointer)
            pointer3 = copy.deepcopy(rec_pointer)
            ls1 = copy.deepcopy(rec_ls)
            ls2 = copy.deepcopy(rec_ls)
            ls3 = copy.deepcopy(rec_ls)
            letter, arrow1, pointer1 = seq_letter_finder(arrow1, pointer1)
            ls1[0] += letter[0]
            ls1[1] += letter[1]
            arrow2 = rec_arrow[1]
            letter, arrow2, pointer2 = seq_letter_finder(arrow2, pointer2)
            ls2[0] += letter[0]
            ls2[1] += letter[1]
            arrow3 = rec_arrow[2]
            letter, arrow3, pointer3 = seq_letter_finder(arrow3, pointer3)
            ls3[0] += letter[0]
            ls3[1] += letter[1]
            return list(itertools.chain(align_seq_finder(arrow1, pointer1,
ls1),
                                        align_seq_finder(arrow2, pointer2,
ls2),
                                        align_seq_finder(arrow3, pointer3,
ls3)))
```

2. 史密斯-沃特曼算法（Smith-Waterman algorithm）

   史密斯-沃特曼算法是一种进行局部序列比对（相对于全局比对）的算法，用于找出两个核苷酸序列或蛋白质序列之间的相似区域。该算法的目的不是进行全序列的比对，而是找出两个序列中具有高相似度的片段。该算法是NW算法的一个变体，二者都是动态规划算法。这一算法的优势在于可以在给定的打分方法下找出两个序列的最优的局部比对（打分方法使用了置换矩阵和空位罚分）。该算法和尼德曼-翁施算法的主要区别在于该算法不存在负分（负分被替换为零），因此局部比对成为可能。

   史密斯-沃特曼算法先用迭代方法计算出两个序列的所有可能相似性比较的分值，然后通过动态规划的方法回溯寻找最优相似性比较。假设比对的两序列为: $A = a_1a_2a_3...a_n$ 和 $B = b_1b_2b_3...b_n$ ，则两 序列的长度分别为 $len(A) = n$ ， $Len(B) = m$;

   $s(a, b)$:字符a和字符b的相似分数;

   H :匹配分数矩阵

   $W1 = 2$:一个空位罚分，这里设置为2，那么先初始化算法分数矩阵 H ，使行 i 表示字符 a_i ，列 j 表示字符 b_j ;

   计算矩阵中每一项的 $H_{i\ j}$:

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j), \\ H_{i-1,j} - W_1, \\ H_{i,j-1} - W_1, \\ 0 \end{cases} \qquad S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$$

从而得到整个得分矩阵，回溯，从矩阵 H 中分数最大的一项开始:若 $a_i = b_j$，则回溯到左上角单元格，若 $a_i \neq b_j$ ，回溯到左上角、上边、左边中值最大的单元格，若有相同最大值的单元格，优先级按照左上角、上边、左边的顺序，根据回溯路径，写出匹配字符串: 若回溯到左上角单元格，将 a_i 添加到匹配字串 A，将 b_j 添加到匹配字串 B; 回溯到上边单元格，将 a_i 添加到匹配字串 A ，将_添加到匹配字串 B ;若回溯到左边单元格，将_添加到匹配字串 A，将 b_j 添加到匹配字串 B，最终得到局部最优匹配序列。

3. 最大简约法(MP)

   最大简约法(maximum parsimony，MP)最早源于形态性状研究，现在已经推广到分子序列的进化分析中。最大简约法的理论基础是奥卡姆（Ockham）哲学原则，这个原则认为：解释一个过程的最好理论是所需假设 数目最少的那一个。对所有可能的拓扑结构进行计算，并计算出所需替代数最小的那个拓扑结构，作为最优树。Felsenstein指出，在试图使进化事件的次数最小时，简约法隐含地假定这类事件 是不可能的。如果在进化时间范围内碱基变更的量较小，则简约法是很合理的，但对于存在大量变更的情形，随着所用资料的增加，简约法可能给出实际上更为错误的系统树。

   最大简约法的优点： 最大简约法不需要在处理核苷酸或者氨基酸替代的时候引入假设（替代模型）。此外，最大简约法对于分析某些特殊的分子数据如插入、缺失等序列有用。
   缺点：在分析的序列位点上没有回复突变或平行突变，且被检验的序列位点数很大的时候，最大简约法能够推导获得一个很好的进化树。然而在分析序列上存在较多的回复突变或平行突变，而被检验的序列位点数又比较少的时候，最大简约法可能会给出一个不合理的或者错误的进化树推导结果。

4. 最大似然法(ML)

   最大似然法(maximum likelihood,ML)最早应用于系统发育分析是在对基因频率数据的分析上，后来基于分子序列的分析中也已经引入了最大似然法的分析方法。
   最大似然法分析中，选取一个特定的替代模型来分析给定的一组序列数据，使得获得的每一个拓扑结构的似然率都为最大值，然后再挑出其中似然率最大的拓扑结构作为 最优树。在最大似然法的分析中，所考虑的参数并不是拓扑结构而是每个拓扑结构的枝长，并对似然率球最大值来估计枝长。最大似然法的建树过程是个很费时的过 程，因为在分析过程中有很大的计算量，每个步骤都要考虑内部节点的所有可能性。
   最大似然法也是一个比较成熟的参数估计的统计学方法，具有很好的统计学理论基础，在当样本量很大的时候，似然法可以获得参数统计的最小方差。只要使用了一个合理的、正确的替代模型，最大似然法可以推导出一个很好的进化树结果。但是对于相似度很低的序列，NJ往往出现Long-branch attraction（LBA，长枝吸引现象），有时严重干扰进化树的构建。

5. 非加权分组平均法(UPGMA)

   前提条件：在进化过程中，每一代发生趋异的次数相同，即碱基或氨基酸的替换速率是均等且恒等的。

   UPGMA法计算原理和过程:

   1. 以已求得的距离系数,所有比较的分类单元的成对距离构成一个t×t方阵,即建立一个距离矩阵M。
   2. 对于一个给定的距离矩阵,寻求最小距离值 Dpq 。
   3. 定义类群p和q之间的分支深度 $Lpq = Dpq/2$ 。
   4. 若p和q是最后一个类群,侧聚类过程完成,否侧合并 p 和 q 成一个新类群 r 。

5. 定义并计算新类群 r 到其他各类群 i (i≠p和q)的距离 $Dir = (Dpi + Dqi)/2$ 。
6. 回到第一步,在矩阵中消除 p 和 q ,加入新类群 r ,矩阵减少一阶,重复进行直至达到最后归群。

UPGMA法比较直观和简单,运算速度快,应用很广，但缺点在于当分子进化速率较大时,在建树过程会引入系统误差。

6. 邻接法(neighbor-joining)

邻接法(Neighbor-joiningMethod)由Saitou和Nei(1987)提出。该方法通过确定距离最近(或相邻)的成对分类单位来使系统树的总距离达到最小。相邻是指两个分类 单位在某一无根分叉树中仅通过一个节点(node)相连。通过循序地将相邻点合并成新的点，就可以建立一个相应的拓扑树。

NJ法的步骤如下：

1. 对于给定距离矩阵中的每一端结 i , 用下式计算与其它分类单元之间的净趋异量 (Ri) (t:矩阵中的分类单元数)
2. 建立一个速率校正距离矩阵 M :
3. 定义一个新节点 u , u 的三个分支分别与节点 i , j 和树的其余部分相连,并且 $D_{i\ j}$ 为矩阵中距离最小者
4. 定义 u 到树的其它节点 k(k≠i和j外的所有节点)的距离:
5. 从距离矩阵中删除 i 和 j 的距离,矩阵减少一阶。
6. 如果矩阵仍然多于两个的节点,重复第①-⑤步,否测除最外两个节点的分支长度来确定外,树上其余节点都确定,最后是剩余的2个的分支长度 $Sy = D_{i\ j}$

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

邻接法在概念上与UPGMA法相同，但是仍有以下四点区别：

- NJ法不要求距离符合超度量特性,但要求数据应非常接近或符合叠加性条件,即该方法要求对距离进行校正。
- 邻接法在成聚过程中连接的是分类单元之间的节点(node),而不是分类单元本身。
- NJ法中原始距离数据用于估算系统树上所有端结分类单元之间的距离矩阵,校正后的距离用于确定节点之间的连接顺序。

在重建系统发育树时，NJ法取消了UPGMA法所做的假定，认为在此进化分支上,发生趋异的次数可以不同。

7. 贝叶斯法（Bayesian）

贝叶斯法是一种新的利用贝叶斯演绎法预测种系发生史的系统进化分析方法。它基于统计特征的系统发生树构建方法。该算法同时考虑模型参数和树拓扑结构的概率分布，通过马尔科夫—蒙特卡洛模拟”(MCMC)算法从后验概率分布中抽样来确定统计性质，贝叶斯法根据多种分子进化模型，利用马尔科夫链的蒙特卡洛方法产生所有参数的后验概率(posterior probability)估计值，这些参数包括拓扑结构、分支长度和替代模型各参数的估计。该方法不仅可以对模型的参数进行直接量化，而且可以分析很大的数据集，因其以后验概率来表示各分支的可信性，而不需用自举法(bootstrap)进行检验。

与ML和NJ相比，贝叶斯法效率更高，对于同一组数据的分析, 贝叶斯方法分析结果中的节点支持率高于其它算法中的相应结果。贝叶斯法的其他优点在于：推导系统树、评估系统树的不确定性、检测选择作用、比较系统树、参考化石记录计算分歧时间和检测分子钟。贝叶斯法得到的系统进化树不需要利用自引导法进行检验，其后验概率直观地反映了系统进化树的可信程度，是一种系统进化分析的好方法。

8. K 最近邻法(k-Nearest Neighbor)

   K最近邻(k-NearestNeighbour，KNN)分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的核心原理是：如果一个样本在特征空间中的k个最相似 (即特征空间中最邻近) 的样本中的大多数属于某一个类别，则该样本也属于这个类别。KNN算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。KNN方法虽然从原理上也依赖于极限定理，但在类别决策时，只与极少量的相邻样本有关。由于KNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN方法较其他方法更为适合。

   KNN算法主要涉及 3 个主要因素: 训练集、距离与相似的衡量、k 的大小。KNN算法的主要考虑因素: 距离与相似度。

9. K均值算法 (k-means cluster)

   k-均值算法（英文：k-means clustering）源于信号处理中的一种向量量化方法，现在则更多地作为一种聚类分析方法流行于数据挖掘领域。k-均值聚类的目的是：把 n 个点（可以是样本的一次观察或一个实例）划分到 k 个聚类中，使得每个点都属于离他最近的均值（此即聚类中心）对应的聚类，以之作为聚类的标准。这个问题将归结为一个把数据空间划分为Voronoi cells的问题。

   其具体步骤为：

   任意选取 K 个基因表达向量作为初始聚类中心 $Z_1, Z_2, \ldots, Z_k$ ，反复迭代计算，如果 $||X - Z_j(l)|| < ||X - Z_i(l)|| (i = 1, 2, \ldots, K, i \neq j)$，则将 X 所代表的基因归于第 j 类。按照上述办法处理所有的基因;经过上述处理，聚类可能发生变化，因此需要重新计算 K个新聚类中心: 对于所有的聚类中心，如果 $Z_j(l+1) = Z_j(l) (j = 1, 2, \ldots, K)$ ，则迭代结束，得到最后的聚类结果; 否则继续进行迭代计算。

   公式为： $$Z_j(l+1) = \frac{1}{N_j} \sum_{X \in f_j(l)} X$$

10. 主成分分析 (Principal Component Analysis，PCA)

    是一种分析、简化数据集的技术，可以用于群体分层分析和推断进化关系。PCA 通过正交变换将一组数量庞大且可能存在相关性的变量转换为一组低维的线性不相关的变量, 实现保留低阶主成分，忽略高阶主成分，这样就减少数据集的维数，同时保持数据集中对方差贡献最大的特征，保留的低阶成分往往也能够保留住数据的最重要方面。

    主成分分析基本思想：

    1. 将坐标轴中心移到数据的中心，然后旋转坐标轴，使得数据在C1轴上的方差最大，即全部n个数据个体在该方向上的投影最为分散。意味着更多的信息被保留下来。C1成为第一主成分。
    2. C2第二主成分：找一个C2，使得C2与C1的协方差（相关系数）为0，以免与C1信息重叠，并

且使数据在该方向的方差尽量最大。

3. 以此类推，找到第三主成分，第四主成分……第p个主成分。p个随机变量可以有p个主成分。

3.请阐述5种高通量组学测序技术的基本原理，及其优缺点（不少于500字）。

1. 454焦磷酸测序

   基本原理：

   454焦磷酸测序技术原理主要是将打碎的DNA片段结合到微珠上，使测序PCR反应发生在固相的微珠（resin beads）上，并且整个PCR反应和相关的酶被油包水的液滴包裹，并且每个油滴系统只包含1个DNA模板。扩增后，每个DNA分子可以得到富集，每个微珠只能形成一个克隆集落。454测序仪的测序通道体积非常狭小，只能容纳一个微珠。测序过程中，GS FLX系统会将引物上dNTP的聚合与荧光信号释放偶联起来。通过检测荧光信号，就可以达到DNA测序的目的。454可以提供中等的读长和适中的价格，适合de novo 测序、转录组测序、宏基因组研究等。

   - 454焦磷酸测序从双链DNA样本开始，首先需要使用限制性酶将DNA分解成大约400到600个碱基对的片段；然后，adaptors的DNA短序列将连接到DNA片段上。随后，resin beads被添加到混合物中。由于resin beads上的DNA序列与adaptors上的序列是互补的，因此DNA片段可以直接与resin beads结合，每个resin beads上有一个DNA片段。
   - 当adaptors上的DNA片段与resin beads上的DNA序列结合时，原样本DNA连接双链的纽带断裂，双链分离，成为单链的DNA。然后，通过聚合酶链式反应，DNA片段在每个resin beads上被复制了无数次，这一步骤将产生数百万个相同的DNA序列副本。
     之后需要对resin beads进行过滤，以去除那些未能附着在任何DNA上或含有一种以上的DNA片段的resin beads。
   - 接下来，剩余的resin beads与包含测序反应所需的DNA聚合酶和引物的酶珠一起被放入测序板的孔中（每孔包含一个resin beads）。聚合酶和引物附着在resin beads上的DNA片段上。核苷酸碱基被按照ACGT的次序轮流地添加到孔中。当每个碱基被加入到DNA中时，激发出的荧光可由相机记录。通过记录荧光强度模式，可以解码原始DNA片段的序列。

   优点：

   在454测序技术中，每个测序反应都在PTP板上独立的小孔中进行，因而能大大降低相互间的干扰和测序偏差；另外，454测序技术的读长较长，平均读长可达400bp。

   缺点：

   454技术会在测序过程中引入插入和缺失的测序错误。由于454测序技术无法准确测量同聚物的长度，因此，例如，当序列中存在类似于PolyA的情况时，测序反应会一次加入多个T，而所加入的T的个数只能通过荧光强度推测获得，有可能导致测序结果不准确。

2. Illumina (Solexa) 测序

   基本原理：

   Illumina (Solexa) 测序的基本原理是可逆终止化学反应。DNA片段加上接头之后，可以随机的附着于flow cell（流动池）表面，并且在固相的表面经过桥式扩增。这样就形成了数千份相同的单分子簇，被用做测序模板。测序采取边合成边测序的方法，和模板配对的ddNTP原料被添加上去，不配对的ddNTP原料被洗去，成像系统能够捕捉荧光标记的核苷酸。随着DNA 3'端的阻断剂的去除，下

一轮的延伸就可以进行。和焦磷酸测序不同，每次DNA的只能延伸一个核苷酸。Solexa的读长在100-150bp之间，适合小RNA鉴定、甲基化和表观遗传学研究。

- Illumina (Solexa) 测序测序技术的第一步是将DNA分解成约200至600个碱基对的更易处理的片段。然后，adaptors的DNA短序列将连接到DNA片段上。接着，连接到适配体的DNA片段被制成单链。
- 一旦制备好DNA单链，DNA片段就会在流动池中被清洗。互补的DNA与流动池中的引物结合，没有结合的DNA则会被洗掉。附着在流动池上的DNA随后被复制，形成具有相同序列的DNA小簇。当进行测序时，每个DNA分子簇将发出足够强的信号，以便被相机检测到。
- 随后，向池中加入未标记的核苷酸碱基和DNA聚合酶来延长和连接附着在的DNA链。这在流动池表面的引物之间形成了双链DNA的“桥”。然后利用热量将双链DNA分解成单链DNA，留下几百万个相同的DNA序列的密集簇。
- 之后，测序采取边合成边测序的方法，引物和荧光标记的终止剂被添加到流动池中。引物附着在被测序的DNA上，DNA聚合酶与引物结合，将第一个荧光标记的终止子加入到新的DNA链中。一旦添加了一个碱基，就不能再向DNA链上添加更多的碱基，直到终止子碱基从DNA上被切断。
- 激光穿过流动池以激活核苷酸碱基上的荧光标签。这种荧光被相机检测到并记录在电脑上。每个终止碱基（A、C、G和T）都会发出不同的颜色。
- 然后将荧光标记的终止基团从第一个碱基上移除，并在旁边添加下一个荧光标记的终止基团。就这样，这个过程一直持续到数百万个集群被测序。
- Illumina测序过程中会对DNA序列进行逐个碱基分析，使其成为一种高度精确的方法。然后，生成的序列可以与参考序列进行比对，这将寻找被测序的DNA中的匹配或变化。

优点：

Illumina (Solexa) 测序技术所需的样品量少，测序通量高，精确性高，其高自动化系统操作较为简单，读取片段多、适合大量小片段的测序（microRNA、lncRNA等）

缺点：

读取序列较短、不适于de novo sequence。

3. ABI SOLiD 测序

基本原理：

ABI Solid 测序技术的核心是4种荧光标记寡核苷酸的连接反应测序。测序之前，DNA模板通过乳化PCR扩增，和Roche 454的基本相同，只是Solid的微珠更小，只有1μm。3'端修饰的微珠可以沉淀在玻片上。连接测序所用的底物是8个碱基荧光探针混合物，根据序列的位置，样品DNA就可以被探针标记。DNA连接酶优先连接和模板配对的探针，并引发该位点的荧光信号的产生。Solid的读长只有50-75bp，精确度可达Q40，适于基因组重测序和SNP检测。

- 首先根据实际需要制备片段文库(fragment library)或末端配对文库(mate-paired library)。制备片段文库就是在短DNA片段（60～110 bp）两端加上SOLiD接头（P1、P2 adapter）。而制备末端配对文库，先通过DNA环化、Ecop15I酶切等步骤截取长DNA片段（600bp到10kb）两末端各25 bp进行连接，然后在该连接产物两端加上SOLiD接头。两种文库的最终产物都是两端分别带有P1、P2 adapter的DNA双链，插入片段及测序接头总长为120～180 bp。
- 文库制备得到大量末端带P1、P2 adapter但内部插入序列不同的DNA双链模板。P1引物固定在P1磁珠球形表面 (SOLiD将这种表面固定着大量P1引物的磁珠称为P1磁珠)。PCR反应过程中磁珠表面的P1引物可以和变性模板的P1 adapter负链结合，引导模板合成，这样一来，P1引

物引导合成的DNA链也就被固定到P1磁珠表面了。

- 然后将包含PCR所有反应成分的水溶液注入到高速旋转的矿物油表面，水溶液瞬间形成无数个被矿物油包裹的小水滴。这些小水滴就构成了独立的PCR反应空间。理想状态下，每个小水滴只含一个DNA模板和一个P1磁珠。PCR反应结束后，P1磁珠表面就固定有拷贝数目巨大的同来源DNA模板扩增产物。
- SOLiD测序反应在SOLiD玻片表面进行。含有DNA模板的P1磁珠共价结合在SOLiD玻片表面。磁珠是SOLiD测序的最小单元。每个磁珠SOLiD测序后形成一条序列。
- 接着，以8碱基单链荧光探针混合物作为连接反应底物，按照荧光探针的颜色类型与探针编码区碱基对的对应关系，使这些探针按照碱基互补规则与单链DNA模板链配对。SOLiD测序完成后，获得了由颜色编码组成的SOLiD原始序列。

优点：

ABI SOLiD 测序技术是目前二代测序技术中准确度最高的，其原始碱基数据的准确度大于99.94%，而在15X覆盖率时的准确度可以达到99.999%。而除了测序和重测序之外，ABI SOLiD 测序技术还能进行全基因组表达图谱分析、SNP、microRNA、ChIP、甲基化等分析。

缺点：

ABI Solid 测序技术读长短，拼接复杂；双碱基对应一个荧光信号，如果发生读码错误，将发生连锁读码反应；此外，ABI Solid 测序易受回文序列影响。

4. Ion torrent测序

基本原理：

Ion Torrent测序技术的主要原理是使用半导体技术在化学和数字信息之间建立直接的联系。测序反应仅是在一张半导体芯片上实现的。芯片上布满成千上亿个小孔，每一个小孔中的PH电极，当A、C、G、T四种dNTP的溶液，分别地、依次地流过芯片的表面，每个dNTP分子有3个磷酸基团，当dNTP被聚合酶结合到DNA链上时，会掉下来的一分子的焦磷酸，1个焦磷酸分子会被酶再进一步分解成2个磷酸分子，这样，在测序小孔的微环境中，就会多出两个酸性分子，一个珠子上有几千、几百条DNA链，每次发生聚合反应，就会多出几千、几百个酸分子。这样，小孔微环境的pH值就会短暂地下降，小孔的pH变化不断的被记录下来，并将信号值传给计算机，从而实现碱基的测序分析。

优点：

由于硬件设备无需光学检测和扫描系统，并且使用天然核苷酸和聚合酶、无需焦磷酸酶化学级联，无需标记荧光染料和化学发光的配套试剂，因此测序成本低；此外，Ion Torrent测序技术的应用范围也很广，几乎涵盖Sanger方法和已有高通量测序技术的应用，例如基因组DNA序列测定（微生物基因组测序、线粒体测序、靶向测序）、DNA扩增子测序、体细胞突变测序、 De novo测序、小RNA和基因表达研究、 ChIP-seq、农业SNP应用等；另外，Ion Torrent测序技术的一大重要优势是测序反应时间短，其标准的测序时间仅为2-3小时，弥补了高通量测序周期长的缺陷。

缺点：

Homopolymer问题难以解决。当Ion torrent测序仪在测到一连串相同的碱基时，通常难以读准碱基的具体数目。尽管Ion torrent测序仪能分辨出一串相同碱基的强信号，但在对碱基数目进行具体判断时容易发生错误。

5. 单分子纳米孔测序

基本原理：

由牛津纳米孔科技公司开发的单分子纳米孔测序技术的核心原理是芯片测序。单链DNA分子穿过纳米孔时，由于不同的碱基的形状大小有差异，与孔内环糊精分子发生特异性反应从而引起电阻变化。纳米孔的两侧有一恒定电压，因此可以检测到纳米孔中电流的变化，从而反映出通过纳米孔的DNA分子的碱基排列情况。

目前，该纳米孔测序平台主要使用1D和1D2两种测序策略，其中1D测序原理是：基因组DNA或cDNA分子经接头帮助到达纳米孔附近，在解旋酶的作用下双链DNA分子解开为单链，通过孔道蛋白；传感器检测到不同核苷酸通过所引起的电流变化的差异并将其转换为电信号；最后，根据电信号变化的频谱，应用模式识别算法得到碱基类型。与1D测序策略不同的是，1D2测序策略在建库时会在两条DNA分子上加上一种特殊的接头，使得在读取模板链的同时互补链可以附着到膜上，在第一条链离开纳米孔后不久，互补链就有一定概率接着被测序，两条链的数据相互校正，可以帮助提高测序的准确率。

优点：

单分子纳米孔测序技术在测序过程中不需要通过PCR进行信号放大，因此避免了PCR反应过程中引入的碱基错配；在整个反应中也不涉及酶的催化反应，理论上只要核酸提取步骤可以得到足够长度的序列，测序步骤就可以对其进行检测；此外，由于甲基化等修饰前后的核苷酸所引起的电阻变化是不同的，所以该测序平台可通过对电信号的识别来判断碱基的甲基化修饰情况。总体来说，单分子纳米孔测序技术具有高通量、长读长、可以直接检测碱基甲基化修饰和体积较小便于携带等优势，应用前景非常广阔。

缺点：

单个核苷酸通过纳米孔的速度及纳米孔的厚度可能引起电流差异特征性的不明显，可能会降低测序的精确度。事实上，单分子纳米孔测序的测序准确率仍然不及第二代测序技术准确率。

4.（1）请叙述5种生物网络模型的基本原理及其优缺点(不少于500字)；

1. 马尔可夫模型

   马尔可夫链是一种随机过程，适用于分析时间序列的基因 表达数据。在马尔可夫模型中，马尔可夫链假设某一时刻的基因表达水平决定了下一时刻的基因表达水平，公式为: $C(t) = JC(t-1)$，构建GRN过程中，基于 马尔可夫模型对gene expression profile的特征提取和聚类都表现出良好的适应性。如果要提高模型的准确性，可提高马尔可夫模型的阶数。

2. 加权矩阵模型

   加权矩阵模型可以看作是线性组合模型的推广。在该模型中，一个基因的表达水平是其他基因表达水平的函数，Weaver 等用一个加权矩 阵表示基因彼此之间的相互调控影响，含有n个基因的转录调控网络的基因表达状态，用n维空间中的向量$u(t)$表示，$u(t)$的每一个元素代表一个基因在状态t或时刻t的表达水平，以一个加权矩阵 W 表示基因之间的相互调控作用$r_i(t) = \Sigma_j W_{i\ j} u_j(t)$，W 的每一行代表一个基因的所有调控输入，$w_{i\ j}$代表基因$j$的表达水平对基因$i$的影响，在时刻$t$，基因$j$对基因$i$的调控效应为$j$的表达水平，即$u_j(t)$乘以$j$对$i$的调控影响程度$W_{i\ j}$，若$W_{i\ j}$为正，则基因$j$激发基因$i$的表达， 负值表示基因$j$抑制基因$i$的表达，0 表示无影响。该模型具有稳点过的周期和稳定的基因表达水平，在这种模型中还可以加入新的变量，模拟环境条件变化对基因表达水平的影响。

3. 布尔网络模型

布尔网络模型是刻画基因调控网络一种最简单的模型，在 布尔网络中，每个基因所处的状态或者是“开”，或者是“关”，状态“开”表示 一个基因转录表达，形成基因产物，状态“关”则代表一个基因未转录，基因之 间的相互作用关系由布尔表达式来表示，即基因之间的关系由逻辑算子 and、or 和 not 刻画。查询简单，便于计算，容易理解，但是不能很好地反映细胞中基因表达的实际情况。

4. 线性组合模型

   线性组合模型是一种连续网络模型，在这种模型中，一个基因的表达值是若干个其它基因表达值的加权和，$X_i(t+\Delta t)=\Sigma_j W_{i\ j}X(t)$，其中，$X_i(t+\Delta t)$是基因$i$在$t+\Delta t$时刻的表达水平，$X_j(t)$是基因$j$在$t$时刻的表达水平,而$W_{i\ j}$代表基因$j$的表达水平对基因$i$的影响，正值表示激活，负值表示抑制，0 表示无影响。还可以增加其他数据项，以逼近基因调控的实际情况，该模 型简单，但只能处理具有线性关系的基因表达数据，应用范围小。

5. 贝叶斯网络模型

   该模型以贝叶斯定理和假设为理论基础，用有向无环图 (DAG) 的形式表示随机变量间的概率关系，网络中每个基因是一个节点，每个调控关系是一条边。该模型可以处理随机事件，控制噪声，可以获得变量间的因果关系，在 GRN 模型中，贝叶斯网络比其他模型更有优势。

4.（2）请用表格总结现有生物网络数据库、工具、软件包及各自特点与参考文献（不少于10种）。

名称	类别	特点	参考文献
HPRD	蛋白质相互作用数据库	该数据库是包含蛋白质相互作用、蛋白质注释、结构域、转 录后修饰、亚细胞定位等多种信息的综合数据库。该数据库只收录人类蛋白质相互作用，是来源于文献挖掘的最大的人 类 PPI 数据库。HPRD 对蛋白质相互作用数据有两种分类方式：1，根据相互作用的拓扑结构和参与数目，将蛋白质相互作用分成二元相互作用和复杂相互作用(复合物); 2，根据实验类型，将蛋白质相互作用分为体内(invivo),体外(in vitro)和酵母双杂交(Y2H)三类相互作用。	Keshava Prasad, T.T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A. and Balakrishnan, L., 2009. Human protein reference database—2009 update. Nucleic acids research, 37(suppl_1), pp.D767-D772.
I2D	蛋白质相互作用数据	该数据库是已知和预测的哺乳动物和真核蛋白质-蛋白质相互作用的在线数据库，是已知和预测的真核 PPI 的最全面来源之一，还具有相互作用网络图。	Interologous Interaction Database; DOI: https://doi.org/10.25504/FAIRsharing.k56rjs; Last edited: June 23, 2021, 2 p.m.; Last accessed: Jun 24 2021 2:21 p.m.

	库		
3DID	蛋白质相互作用数据库	该数据库用于搜集 3D 结构已知的蛋白质的互作信息，可通过结构域名称、基序名称、蛋白质序列、GO 编码、PDB ID、Pfam 编码进行检索。	Roberto Mosca, Arnaud Ceol, Amelie Stein, Roger Olivella & Patrick Aloy, 3did: a catalogue of domain-based interactions of known three-dimensional structure, Nucleic Acids Research 2014, 42(D1):D374-D379, doi:10.1093/nar/gkt887
STRING	蛋白质相互作用数据库	该数据库可应用于 2031 个物种，包含 960 万种蛋白和1380 万中蛋白质之间的相互作用。它除了包含有实验数据、从 PubMed 摘要中文本挖掘的结果和综合其他数据库数据外，还有利用生物信息学的方法预测的结果，覆盖的物种最多，相互作用信息最多。	Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M. and Bork, P., 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic acids research, 37(suppl_1), pp.D412-D416.
DIP	蛋白质相互作用数据库	该数据库专门存储经实验证实的来自文献报道的二元蛋白质相互作用，以及来自PDB(protein data bank)数据库的蛋白质复合物,其目的在于建立一个简单易用高度可信的PPI 公共数据库。DIP 数据库包含 DIP CORE 和 DIP FULL 两部分。DIP 数据库的蛋白质相互作用包含果蝇、酵母、老鼠、人类等多个物种,提供多种查询方式,用户可直接基于蛋白质名称、物种查询相互作用蛋白质;也可基于序列匹配的BLAST 搜索和模体(motif)搜索查询相互作用蛋白质。JDIP是该数据库提供的一个基于 Java 语言的可视化应用工具,把蛋白质相互作用数据以网络形式更加直观的展现出来。	Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., & Eisenberg, D. (2000). DIP: the database of interacting proteins. Nucleic acids research, 28(1), 289-291.
BioGRID	蛋白和遗传相互	该数据库是-个包含了模式生物(如酿酒酵母、裂殖酵母、黑腹果蝇、线虫等)蛋白质相互作用和基因相互作用的数据库。其包含了高通量的实验数据和传统的实验数据。数据库中包数含了原始文献中的相互作用,而综述和未发表的论文中的数据没有被收录。由于BioGrid 未将来源于不同论文和通过不	Biological General Repository for Interaction Datasets; DOI: https://doi.org/10.25504/FAIRsharing.9d5f5r; Last edited: March

	作用数据库	同实验手段所产生的实验数据进行整合,因而数据库中包含了很多冗余的蛋白质相互作用数据。BioGrid 数据库中只包含了反应物、实验手段PubMed ID和Evidence Code 等信息。	9, 2021, 4:07 p.m.; Last accessed: Jun 18 2021 10:33 a.m.
CSNDB	细胞信号网络数据库	该数据库由日本国立健康科学研究所建立，是人类细胞中信号途径的数据和知识库，汇编了有关信号传输的生物分子、序列、结构、功能和生物化学反应，并能够自动绘图表示信号途径。	/
KEGG	代谢途径数据库	KEGG，（Kyoto Encyclopedia of Genes and Genomes,京都基因与基因组百科全书）。是一个整合了基因组、化学和系统功能信息的数据库，旨在揭示生命现象的遗传与化学蓝图。它是由人工创建的一个知识库，是基于使用一种可计算的形式捕捉和组织实验得到的知识而形成的系统功能知识库。另外，KEGG具有强大的图形功能，它利用图形来介绍众多的代谢途径以及各途径之间的关系。基因组信息存储在GENES数据库里，包括完整和部分测序的基因组序列；功能信息存储在PATHWAY数据库里，包括图解的细胞生化过程，如代谢、膜转运、信号传递、细胞周期，以及同系保守的子通路信息；LIGAND数据库包括关于化学物质、酶分子、酶反应等信息。KEGG提供了Java的图形工具来访问基因组图谱，比较基因组图谱和操作表达图谱，以及其他序列比较、图形比较和通路计算的工具。	Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.
GO	代谢途径数据库	GO是Gene ontology的缩写，GO数据库分别从功能、参与的生物途径及细胞中的定位对基因产物进行了标准化描述，即对基因产物进行简单注释，通过GO富集分析可以粗略了解差异基因富集在哪些生物学功能、途径或者细胞定位。	Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. Jan 2019;47(D1):D419-D426.
	代谢	一种细胞网络分析工具，前身是FluxAnalyzer，是	Klamt, S., Saez-Rodriguez, J., & Gilles,

CellNetAnalyzer	与信号分析软件	基于MatLab的代谢网络和信号传导网络分析模块，。这是一个典型的代谢流分析工具，可以进行代谢流的计算、预测、目标函数的优化，端途径分析、元素模式分析，以及代谢流之间的对比等。可以基本满足研究一个中型代谢网络的结构尤其是计算流分配的要求。	E. D. (2007). Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC systems biology, 1(1), 1-13.

5.请描述什么是分子标志物、网络标志物、动态网络标志物。

生物标志物（Biomarker）是近年来随着免疫学、分子生物学和基因组学技术的发展而提出的一类与细胞生长、增殖、疾病发生等有关的标志物，主要是指可以标记系统、器官、组织、细胞及亚细胞结构或功能的改变或可能发生的改变的生化指标，具有非常广泛的用途，目前生物标志物能反映正常生理过程或病理过程或对治疗干预的药物反应，可用于疾病诊断、判断疾病分期或者用来评价新药或新疗法在目标人群中的安全性及有效性，在早期诊断、疾病预防、药物靶点确定、药物反应以及其他方面发挥作用。

网络标志物是一类新的生物标志物，与蛋白质间关系有关，通过整合蛋白质注解、联系和信号通路的相关知识进行研究。由于人类疾病涉及多种基因、蛋白质、多肽等生理/化学变量，因此，有关疾病生物标志物的研究需要从单个标志物推广到多个标志物，从表达到功能指标，从网络到动态网络。

但是，网络生物标志物在疾病发展的不同阶段和时间点会有所改变，标志物的改变可以被检测和评估，即所谓的动态网络生物标志物，是一种针对人类疾病标志物的新型研究方法。**动态网络生物标志物**是一组分子(即基因或蛋白质)或分子模块，可以在复杂疾病的急剧恶化之前发出临界点或临界状态的信号。