生物医学信息学概论

一、总论

1、什么是生物信息学？

广义：应用信息科学的方法和技术，研究生物体系和生物过程中信息的存贮、信息的内涵和信息的传递，研究和分析生物体细胞、组织、器官的生理、病理、药理过程中的各种生物信息，或者也可以说成是生命科学中的信息科学。

狭义：应用信息科学的理论、方法和技术，管理、分析和利用生物分子数据。

生物信息学是生物科学与信息科学的交叉学科，是利用计算机科学（信息学）的技术手段来研究生物学的数据，如对生物数据进行获取(retrival)，存储(storage)，传输(transfer)，计算(manuipulation)，分析(analysis)，模拟(simulation)，预测(prediction)等等的一门新兴学科，是 21 世纪科学发展的热点之一。

2、生物信息学的研究领域和研究手段有哪些？

研究领域：

Database 数据库建设

Data Mining & Integration 数据库整合和数据挖掘

Sequence Analysis 序列分析

Structural Analysis & Functional Prediction 结构分析与功能预测

Large Scale Expressional Profile Analysis 大规模功能表达谱的分析

Modeling and Simulation of BioPathways 代谢网络建模分析

Reconstruction 预测调控网络

Network Analysis 网络普遍性分析

Modeling 模型分析

Program Development 程序开发

Commercialization 商业化

Nucleic Acid 核酸

Protein 蛋白质

研究手段：

①二代测序：基因组测序、甲基化测序、RNA 测序、ATAC-seq、ChIP-seq、Hi-C

ATAC-seq(Assay for Transposase Accessible Chromatin with high-throughput sequencing)：利用转座酶研究染色质可进入性的高通量测序技术，该技术通过转座酶对某种特定时空下开放的核染色质区域进行切割，进而获得在该特定时空下基因组中所有活跃转录的调控序列

ChIP-seq(Chromatin Immunoprecipitation)：染色质免疫共沉淀技术

Hi-C(High-through chromosome conformation capture)：以整个细胞核为研究对象，利用高通量测序技术，结合生物信息分析方法， 研究全基因组范围内整个染色质 DNA 在空间位置上的关系， 获得高分辨率的染色质调控元件相互作用图谱

②芯片：表达谱芯片、甲基化芯片、SNP 芯片、CNV 芯片

SNP(Single Nucleotide Polymorphisms)：单核苷酸多态性

CNP(copy number variation)：拷贝数变异

③质谱：代谢组、蛋白组……

3、什么是医学信息学？研究内容有哪些？
医学信息学（Medical Informatics）是研究如何通过现代信息技术来地有效收集、存储、检索、分析和更好地利用患者的医疗信息、临床研究信息和医学教育信息，从而提高医疗卫生机构的管理与决策水平、医疗质量和医学教育效果的一门学科。
研究内容：

4、什么是生物医学信息学？
生物医学信息学是在充分利用信息学、技术和人力（如：研究人员，从业者和用户等）资源的基础上，提高个体健康、医疗保健、公共卫生和生物医学研究的领域。

二、生物学数据库及其检索

（一）数据库简介
数据库：被组织起来的大量的、整合的生物数据，这些数据通过计算机可以被方便的访问、管理以及更新
数据库管理系统（Database Management System）：一种用于存储、管理和方便访问数据库的软件系统
注意：www 不是 DBMS，www 数据大多是非结构化和无类型的，不能修改数据，无法获得摘要，数据的复杂组合很少能保证数据新鲜度、跨数据项的一致性、容错性，网站在后台有一个 DBMS 来提供这些功能。
知识内容：表示信息（数据模型）、用于查询数据的语言和系统、数据操作的并发控制、可靠的数据存储

数据模型（data model）是用于描述数据的概念集合，模式（schema）是使用给定数据模型对特定数据集合的描述，数据的关系模型（relational model of data）是当今使用最广泛的模型

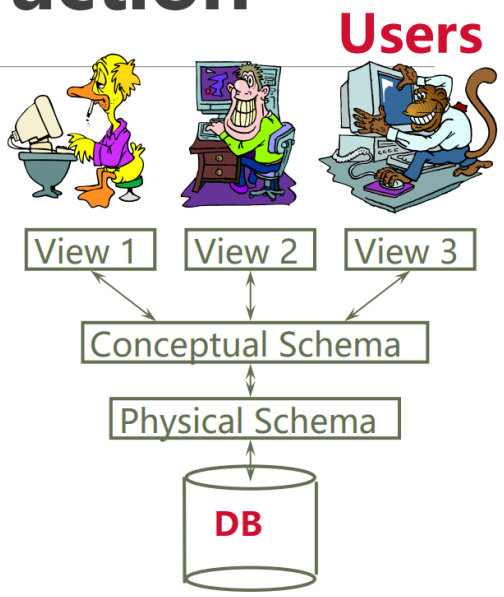
抽象层次（levels of abstraction）

Levels of Abstraction

Views describe how users see the data.

Conceptual schema defines logical structure

Physical schema describes the files and indexes used.



一、数据库基础

1.为什么我们需要数据库？
①从计算转向信息：
◦始终真实的企业计算
◦Web 为个人电脑制造了这一点
◦越来越真实的科学计算
②在过去的几年中，对 DBMS 的需求激增：
◦企业:零售滑动/点击流，“客户关系”、“供应链管理”、“数据仓库”等
◦科学:数字图书馆，人类基因组计划，美国宇航局，地球任务，物理传感器，网格物理，网络
③DBMS 在实践学科中包含了许多 CS：
◦操作系统，语言，理论，AI，多媒体，逻辑
◦传统的关注现实世界的应用程序

2.DBMS 的主要任务
①数据**定义**（Data Definition）：DBMS 负责定义数据库中的数据结构和模式，包括创建表格、定义字段、设置数据类型、建立索引等。它提供了一种方便的方式来描述和管理数据库的结构。
②数据**操作**（Data Manipulation）：DBMS 提供了各种**操作和查询**语言，如 SQL（Structured Query Language），用于插入、更新、删除和查询数据库中的数据。它允许用户和应用程序对数据进行灵活和高效的操作。
③数据库安全（Database Security）：DBMS 负责确保数据库的安全性和保密性。它提供用户身份验证和授权机制，以控制谁可以访问数据库和执行特定操作。它还提供数据加密、审计和访问控制等功能来保护数据库免受未经授权的访问和恶意操作。
④数据完整性（Data Integrity）：DBMS 确保数据库中的数据保持一致、准确和有

效。它通过实施各种约束条件（如主键、外键、唯一性约束）和触发器来强制执行数据的完整性规则。这有助于防止无效或不一致的数据进入数据库。
⑤数据并发控制（Concurrency Control）：DBMS 处理多个用户或应用程序同时访问数据库的情况。它实施并发控制机制，如锁定和事务管理，以确保并发操作的正确性和一致性，避免数据冲突和丢失更新。
⑥数据备份和恢复（Data Backup and Recovery）：DBMS 负责数据库的备份和恢复，以防止数据丢失和灾难恢复。它提供了备份和还原工具，允许将数据库的副本保存在可靠的存储介质上，并在需要时恢复到先前的状态。
⑦性能优化（Performance Optimization）：DBMS 通过优化查询执行计划、索引设计、数据存储和缓存管理等技术，提高数据库的性能和响应时间。它通过使用各种优化技术和策略来减少查询时间、提高并发处理和最大化数据库的吞吐量。

3.关系数据库的基本形式
格（Table）：关系数据库由一个或多个表格组成，每个表格代表一个实体类型或关系。每个表格由行（记录）和列（字段）组成，用于存储实体的属性和关系之间的联系。
行（Record）：表格中的每一行代表一个实体或一个实体的实例。每一行包含一组字段值，表示实体的属性。
列（Field）：表格中的每一列代表一个属性或字段。每一列具有特定的数据类型和约束规则，用于定义属性的特征和取值范围。
键（Key）：表格中的键用于唯一标识表格中的每个记录。主键（Primary Key）是一列或一组列，其值在表格中是唯一的，用于确保记录的唯一性和标识。外键（Foreign Key）是一个表格的列，它引用另一个表格的主键，用于建立表格之间的关联关系。
关系（Relationship）：关系数据库通过表格之间的关系来表示实体之间的联系和依赖关系。关系可以通过主键和外键之间的匹配来建立。常见的关系包括一对一关系、一对多关系和多对多关系。
约束（Constraint）：约束用于定义和强制数据库中数据的规则和完整性。常见的约束包括主键约束、唯一性约束、非空约束、默认值约束和外键约束。它们帮助确保数据的有效性和一致性。
4.DBMS 的优缺点
优点：①抽象层次提供数据独立性；
②高效的数据访问；
③数据完整性和安全性；
④数据管理；
⑤并发访问，可从崩溃系统中恢复；
⑥缩短应用程序开发时间（快速应用开发）
缺点：①昂贵且复杂的设置和维护；
②这种成本和复杂性必须由需求来抵；
③通用，不适合特殊用途的任务

二、NCBI 数据库简介

NCBI：National Center for Biotechnology Information 美国国家生物技术信息中心,

隶属于美国国立卫生研究院（National Institutes of Health, NIH）

网址：www.ncbi.nlm.nih.gov

服务：①GenBank 最大的序列数据库
②免费公开获取的生物医学文献（PubMed）
③Entrez 整合分子和文献数据库
④BLAST 最高容量序列搜索服务
⑤VAST 结构相似度搜索
⑥软件和数据库

1.任务：①建立公共数据库；②计算生物学研究；③开发序列分析软件工具④传播生物医学信息

2.数据库类型：

（1）主数据库：①实验者提交的原始材料
②提交者控制的内容：GenBank, SNP, GEO

（2）衍生数据库：①从原始数据建构；②第三方控制的内容：Refseq, TPA, RefSNP, UniGene, NCBI Protein, Structure, Conserved Domain

3.数据库内容

（1）GenBank 简介

a)国际核酸序列数据库合作联盟：GenBank, DDBJ（DNA database of Japan）, EMBL 开发并维护的 ENA（European Nucleotide Archive）

b)GenBank 是 NCBI 的原始序列数据库，属于核酸序列数据库，GenBank 将数据按高通量基因组序列（HighThroughput Genomic Sequences，HTG）、表达序列标记（Expressed Sequence Tags，EST）、序列标记位点（SequenceTaggedSites，STS）和基因组概览序列（Genome Survey Sequences，GSS）单独分类

档案性质：历史、主观（反映提交者的观点）、冗余

c)数据：①直接提交；②批量提交（EST/GSS/STS）；③FTP 账号（基因组数据）

d)主要内容和格式：完整的 GenBank 数据库包括序列文件，索引文件以及其它有关文件。索引文件是根据数据库中作者、参考文献等字段建立的，用于数据库查询。GenPept 是由 GenBank 中的核酸序列翻译而得到的蛋白质序列数据库，其数据格式为 FastA

（2）GenBank 的衍生序列数据库

①GenPept：GenPept 是由 GenBank 中的核酸序列翻译而得到的蛋白质序列数据库，其数据格式为 FastA

②RefSeq：NCBI 的参考序列（RefSeq）计划，为多种生物提供序列的数据信息及相关资料，用于医学、基因功能和基因功能比较研究。RefSeq 数据库中所有的数据是一个非冗余的、提供参考标准的数据，包括染色体、基因组（细胞器、病毒、质粒）、蛋白、RNA 等。RefSeq 数据库被设计成每个人类位点挑出一个代表序列来减少重复，是 NCBI 提供的校正的序列数据和相关的信息。数据库包括构建的基因组 contig、mRNA、蛋白和整个染色体。refseq 序列是 NCBI 筛选过的非冗余数据库，一般可信度比较高。

数据：精选转录本和蛋白质、模型转录本和蛋白质、组装基因组区域（contigs）、染色体的记录

（3）其他 NCBI 数据库

a) 分子建模数据库（Molecular Modeling Database, MMDB）
矢量对齐搜索工具（Vector Alignment Search Tools, VAST）
b) NCBI 的保守域数据库
①基于 PSI-BLAST 的得分矩阵
②可使用 PRS-BLAST 进行搜索
③来源：SMART、PFAM、COGs、NCBI 策划域
（4）Entrez 数据库
序列来源于 Refseq 数据库；
详尽的注释信息，包括基因在基因组的定位，基因名称、蛋白质名称，基因结构等；基因的命名主要来自权威命名委员会的官方符号以及 Refseq 记录中的基因名，由 NCBI 工作人员进行数据收集并注释。NLM 的索引部门对基因功能进行阐述。沿用人类孟德尔遗传网（OMIM）中的疾病名称并与 NCBI 其他数据库形成交互链接。

三、序列同源性比对（参考 https://zhuanlan.zhihu.com/p/106727231）

1.序列比对的意义和合理性
有助于发现功能和进化关系：①相似序列→进化关系②进化关系→相关功能③同源物→不同生物体中相同（几乎相同）的功能。
2.点图法作用：
①可视化基因组之间的相似性和差异性：通过点图，可以直观地比较不同基因组之间的相似性和差异性。相似的片段在点图中会显示为对角线或近似对角线的模式，而不相似的片段则显示为散布在其他位置的点。
②发现基因组重排和重复序列：点图可以帮助检测基因组重排（基因组内部序列顺序的改变）和重复序列（在同一基因组中出现多次的相似片段）。这对于研究基因组结构和进化过程非常重要。
③寻找基因组中的基因和功能元素：通过比较不同基因组的点图，可以定位基因和其他功能元素在基因组中的位置。相似的功能元素通常在点图中显示为具有相似模式的片段。
④基因组注释和比较基因组学研究：基因组点图是进行基因组注释和比较基因组学研究的重要工具之一。它可以帮助研究人员理解基因组的结构、功能和演化，并揭示基因组之间的关系。
局部比对与全局比对
①局部比对从被比较的两个序列中寻找长度未指定的相似片段
②严格的方法是局部动态规划，时间正比于它所比较的序列长度的乘积

3.不同罚分系统进一步提升了简单对比，不同的罚分系统适用于不同的场景

搜索序列数据库
搜索引擎主要由 3 个部分组成：①评分功能②算法③统计模型恢复显著结果
重要问题：速度
NCBI 的 BLAST 算法是线性时间启发式算法

4.BLAST: Basic Local Alignment Search Tool

（https://zhuanlan.zhihu.com/p/150579075）
①简介：BLAST（Basic Local Alignment Search Tool）是一个最流行的序列搜索程序，包括 Basic BLAST, gap BLAST, PSI BLAST 等
②**主要思想**（基本 BLAST）：同源序列可能包含一个短的**高分相似区域**，每次命中都会产生一个种子，BLAST 会尝试**向两侧延伸**（BLAST 程序首先查询 query 序列的所有子序列，储存在哈希表中。收索数据库中所有与子序列精确匹配的序列，作为种子，向两个方向继续延伸每个精确匹配。期间不允许有空位和错配的情况。然后在限制性区域内；连接延伸的匹配序列，期间允许空位和错配，比对分值要大于设定的阈值。阈值越大，需要匹配的计算越小，软件计算速度越快。仅仅对对延伸匹配进行连接的区域(限制性区域)，而不是整个矩阵，是 BLAST 相对于其他算法速度提高的关键，是以牺牲对角线带以外的任何匹配信息为代价，因此并不能确保 query 序列与数据库比对结果是最优的比对结果）
③**主要步骤：**
步骤 1:给定查询序列 Q，编制与 Q 高分词对中的词构成的可能词的列表
步骤 2:扫描数据库，与步骤 1 中编写的单词列表精确匹配
步骤 3:从步骤 2 延伸命中
步骤 4:评估步骤 3 中延伸命中的重要性
④网址：BLAST: Basic Local Alignment Search Tool (nih.gov)

4、多序列比对
多序列比对是双序列比对推广，即把两个以上字符序列对齐，逐列比较其字符的异同，使得每一列字符尽可能一致，以发现其共同的结构特征的方法。其主要思想是通过比对多个序列的相似性和差异性，揭示它们之间的共同特征和进化关系。

代表性工作：
①Clustal：Clustal 是一种经典的多序列比对算法，最早由 Thompson 等人提出。它基于序列的相似性构建一个序列树，然后通过逐步的多重比对来对齐序列。Clustal 系列算法包括 ClustalW 和 Clustal Omega，广泛用于常见的多序列比对任务。
②Muscle：Muscle 是一种高效的多序列比对算法，由 Edgar 提出。它采用迭代算法，通过逐步改进序列的比对，寻找全局最优比对。Muscle 具有较高的比对准确性和速度，适用于大规模序列比对和高度相似序列的比对。
③MAFFT：MAFFT 是一种多序列比对工具，由 Katoh 等人开发。它通过迭代算法和快速傅里叶变换（FFT）技术，在保持计算效率的同时提供高质量的比对结果。MAFFT 支持多种比对模式，包括常规比对、迭代比对和基于轮廓的比对。
④T-Coffee：T-Coffee 是一种基于一致性的多序列比对方法，由 Notredame 等人提出。它通过比对序列对的子集，并将结果组合成最终的多序列比对。T-Coffee 能够在保持准确性的同时处理结构域和重复序列等复杂情况。
⑤ProbCons：ProbCons 是一种基于概率的多序列比对算法，由 Do 等人开发。它利用统计模型和动态规划方法，通过最大化比对的概率来生成最优的比对结果。ProbCons 在处理高度变异序列和结构域比对等方面具有较好的性能。

5.FASTA 序列格式：
FASTA 格式是一种用于记录核酸序列或肽序列的文本格式，其中的核酸或氨基

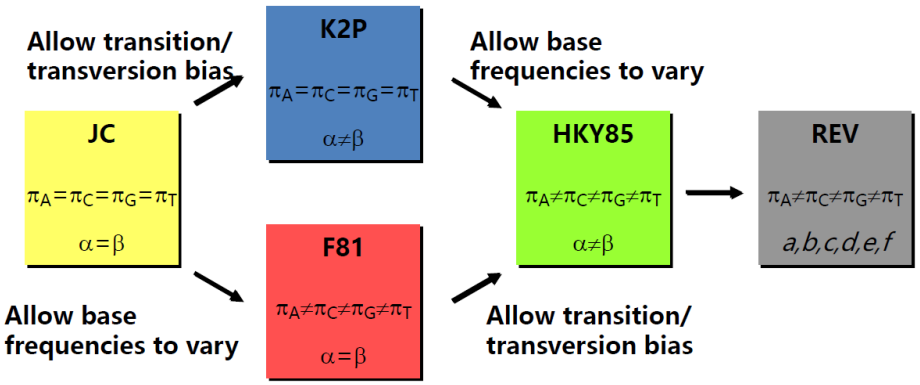
酸均以单个字母编码呈现。该格式同时还允许在序列之前定义名称和编写注释。
格式：FASTA 格式是一种基于 ASCII 码的文本的格式，可以存储一个或多个核苷酸序列或肽序列数据。在 FASTA 格式中，每一个序列数据以单行描述开始（必须单行），后跟紧跟一行或多行序列数据。下一个序列数据也是如此，循环往复。
FASTA 格式文件中的每个序列信息由两个部分组成：
1. 描述行 (The description line, Defline, Header or Identifier line): 以一个大于号(">")开头，内容可以随意，但不能有重复，相当于身份识别信息。
2. 序列行 (Sequence Line)：一行或多行的核苷酸序列或肽序列，其中碱基对或氨基酸使用单字母代码表示。

四、序列分析基础与分子进化

1、碱基替换类型
转换(transition)是由嘌呤置换嘌呤或嘧啶置换嘧啶
颠换(transversion) 是指嘌呤置换嘧啶或嘧啶置换嘌呤

2.遗传差异度量中的“饱和现象”：
最简单的测量方法是计算不同位点的数量
测量差：一些位点可能经历重复替换，而随着序列的分化，测量变得不那么准确
最终发生饱和：大多数位点变化之前已经改变过

3.几种碱基替换模型的特点及其发展沿革
（1）Jukes-Cantor(JC)模型
假设所有四种碱基都有相同的频率并且所有替换的可能性都是一样的
（2）Kimura’s 2 parameter (K2P)模型
考虑到不同频率的转换和转换
（3）Felsenstein (1981)：F81 模型
考虑到碱基组成的差异：①百分比(G + C)可以从 25% - 75%范围内②F81 模型允许四个核苷酸的频率不同③不允许基因/物种之间的差异
（4）Hasegawa, Kishino and Yano (1985)：HKY85 模型
基本上合并了 K2P 和 F81 模型，以允许以不同的速率发生转换和翻转，以及允许基频变化
（5）General reversible model ：REV
最一般的模型-每个替换都有自己的概率



4.有根树和无根树的特点和区别（https://zhuanlan.zhihu.com/p/141835886）
系统发育进化树 (Phylogenetic tree)： 一般也叫系统进化树，进化树。它可以利用树状分支图形来表示各物种或基因间的亲缘关系
有根树：可以从树中找到共同的祖先
无根树：顾名思义，没有根，也就找不到共同的祖先

5.树的拓扑结构复杂性
6.基于**距离**的建树方法的代表：UPGMA 和 NJ
UPGMA：Unweighted Pair-Group Method with Arithmetic means，非加权组平均法
采用顺序聚类算法
①通过计算所有序列两两间的距离，再根据距离远近构建系统发生树
②形成有根树，末端分支相等
③计算过程：
NJ：Neighbour Joining，邻接法
①从一个距离矩阵开始，采用一定的标准，递归的合并矩阵中距离最短的两个节点，并重构新的距离矩阵，如此反复，直到剩下最后一个分类单元
②常用分析软件为 MEGA，形成无根树
③优点：假设少，算法简单易懂，运算速度快，树的构建相对准确，可以分析较多的序列；缺点：序列上的所有位点等同对待，只能应用于进化距离小，序列相似度较高的序列
④特点：不做分子钟假设/产生无根树/假设可加性：对叶子之间的距离=连接它们的边的长度之和/像 UPGMA 一样，通过顺序连接子树来构建树

由于 UPGMA 方法假定演化速率相等，因此分支末端相等，NJ 方法允许不相等的演化速率，因此分支长度与变化量成正比

7.基于模型的建树方法的代表：ML,
Maximum Likelihood method, 最大似然法

五、群体异质性及现代人群历史

1.“波利尼西亚特快”与踏脚石模型

踏脚石模型（Stepping Stone Model）是一种用于解释物种扩散和遗传流动的概念模型。这个模型描述了在一个连续的空间中，物种或基因型通过在相邻地点之间进行逐步扩散的方式来传播。

踏脚石模型假设了一个连续的生境，例如陆地或水域，被称为连续的"踏脚石"。物种或基因型可以从一个踏脚石跳跃到相邻踏脚石，但在跳跃过程中无法跨越较大的空间障碍。这种逐步的扩散过程可以导致物种或基因型在整个空间中逐渐传播和分布。

踏脚石模型通常用于解释生物群落的形成和遗传结构的形成。它可以描述不同地区之间物种或基因型的相似性和差异性，并对物种的迁移、遗传流动和种群扩散进行建模和预测。踏脚石模型可以为生态学和遗传学研究提供重要的理论基础，帮助我们理解生物在空间上的分布和遗传变异的产生。

2.群体遗传构成发生分化的根本原因

群体遗传构成发生分化的根本原因是遗传漂变和自然选择的作用。

①遗传漂变（Genetic Drift）：遗传漂变是指在小样本群体中随机发生的遗传变异。由于有限的个体数量，每一代中的遗传变异可能会在群体中产生随机的频率变动。这种随机性可以导致不同群体之间的遗传差异逐渐积累。遗传漂变在小规模群体、孤立种群或新建群体中起着重要的作用，可以导致基因频率的随机改变。

②自然选择（Natural Selection）：自然选择是指环境对个体适应度的选择过程。当环境条件发生变化时，某些个体可能具有更高的适应度，并能够更好地生存和繁殖。这些具有有利基因型的个体会相对于其他个体在繁殖中更成功地传递其基因给下一代。随着时间的推移，自然选择会导致有利基因型在群体中的频率增加，从而引发分化。

遗传漂变和自然选择相互作用，共同推动群体遗传构成的分化。遗传漂变可以导致随机性的遗传变异积累，而自然选择则通过筛选适应度更高的个体和基因型来驱动某些变异的固定或消失。在不同环境条件下，不同的变异可能会在群体中获得不同的适应度，从而导致分化。

3.现代人类的遗传历史和遗传混合

①人类起源和扩散：根据现有的考古学、人类学和遗传学证据，现代人类起源于非洲，并在约 20 万年前开始扩散到其他大陆。这一过程涉及到不同人群的迁移和扩散，并在全球范围内形成了不同的人群和种群。

②古人类亚种的遗传贡献：早期的人类亚种，如尼安德特人和丹尼索瓦人，与现代人类之间有遗传交流。通过 DNA 研究，发现现代人类的基因组中携带有来自这些古人类亚种的遗传物质，表明了早期人类之间的混合。

③地理隔离和遗传差异：由于地理隔离和人口迁移的影响，不同地区的人群在遗传上逐渐形成了差异。这种差异可以通过遗传标记（如单核苷酸多态性 SNP）的分析来检测和描述，有助于研究人类群体之间的遗传关系和历史。

④近代人类的遗传混合：近代人类历史中的迁徙、贸易、征服和殖民等事件导致了不同人群之间的遗传混合。例如，欧洲、非洲和亚洲之间的交流导致了现代人

类的遗传混合，形成了新的遗传组合和多样性。
⑤遗传学研究方法的进展：随着 DNA 测序技术的发展，研究人员能够更深入地研究人类的遗传历史和遗传混合。通过对现代人类基因组的广泛测序和比较，可以揭示不同人群之间的遗传关系、迁徙路径和混合事件。

综上所述，现代人类的遗传历史和遗传混合是一个复杂且多样的过程，涉及到起源、扩散、地理隔离、遗传贡献和近代交流等多个因素。通过遗传学研究方法的应用，我们可以更好地理解人类的遗传多样性和人类群体之间的关系。

六、遗传形状在群体中的传递
1.遗传关联分析的用途和基本思想
用途：①精细定位候选基因研究；②全基因组关联分析；③基因-环境相互作用；④药物基因组学
基本思想/步骤：
①样本收集：收集包括患者和对照组的个体样本，这些个体可能具有不同的表型特征，例如疾病患者和非患者。
②基因型测定：对收集的个体样本进行基因型测定，通常使用分子生物学技术（如 PCR 和测序）来确定个体在特定基因位点上的遗传变异。
③统计分析：将个体的基因型与其表型特征进行比较和分析。常见的统计方法包括卡方检验、逻辑回归分析、线性回归分析等。这些方法可以评估基因型与表型之间的关联性，并确定是否存在显著的关联。
④修正和验证：在进行统计分析时，需要考虑多重比较的问题，以避免误诊的结果。此外，为了验证结果的可靠性，通常需要在独立的样本集中进行复制实验。

2.遗传关联研究结果的解释
①存在显著关联的可能原因:
- 目标等位型是致病遗传因素
- 目标等位型与致病遗传因素间存在连锁不平衡
- 遗传分层造成的假象
- 抽样方法不合理造成偏差
- 统计假阳性 Type I error

②未发现显著关联的可能原因:
- 目标等位型非致病因素
- 目标等位型与致病遗传因素间不存在连锁不平衡
- 统计假阴性 Type II error

3.常见的遗传关联研究的设计方向
①横断面研究；②病例对照研究；③队列研究；④临床试验 (drug response hypothesis is really a case-control)；⑤家系研究 (trios, sibs, extended families)；⑥仅病例研究

4.遗传关联研究中需要考虑的主要因素
①抽样策略：
- 以无关联个体为基础
- 以家系为基础

②分析方式：
- 单一多态位点分析
- 单体型分析

③统计方法考虑：
- 卡方检验；似然比检验
- 基因-基因和基因-环境相互作用
- 群体分层效应的矫正
- 对多重检验的矫正

七、遗传性状在家系中的传递

1.连锁遗传的背景、原理和根本要点

背景：①不同个体的基因组中存在数以百万计的遗传差异，因此逐一进行研究是非常困难的；②对于家系来说，这数以百万级的遗传差异组合形成了有限的几十种单体型，因此通过检测相对有限的遗传标记，就足以确定不同的染色体单体型在家系中的传递情况

要点：①在不同家系中追踪已知位置的遗传标记和位置未知的遗传性状的传递；
②在不同的家系中都能观察到性状和遗传标记的共分离现象；
③如果性状和某个遗传标记呈现显著地共分离，则性状有关的基因一定位于该标记所在区域

性状和遗传标记的共分离现象是指在有性繁殖的后代,假如基因附近有一紧密连锁的分子标记，在细胞减数分裂时分子标记与基因之间由于相距太近很少有机会发生交换的现象

2.IBS 和 IBD

IBS（Identity by State，状态同源）指的是两个个体在某个基因位点上的基因型相同，但这种相似性并**不一定是由共同祖先遗传而来的**。换句话说，这两个个体在某个位点上的基因型是偶然一致的。在基因组研究中，IBS 可以用于衡量个体之间的遗传相似性，从而推断它们之间的关系或进行种群结构分析。

IBD（Identity by Descent，血缘同源）则指的是两个个体在某个基因位点上的基因型相同，并且这种相似性是**由共同祖先遗传而来的**。换句话说，这两个个体在某个位点上的基因型是由共同祖先传递给他们的。IBD 在遗传学研究中常用于确定个体之间的亲缘关系，例如确定家系或进行遗传连锁分析等。

通常IBD无法直接观测，但IBS可以通过两个体基因型算出。

个体 1	个体2	IBS
AA	AA	2
AA	Aa	1
AA	aa	0

在某一基因座，两个体可能有 0个，1个，或2个相同的等位基因

IBD可以让我们了解两个体间的亲缘关系，虽然**无法直接测得**，但可以根据IBS以及等位基因频率的分布来推定。

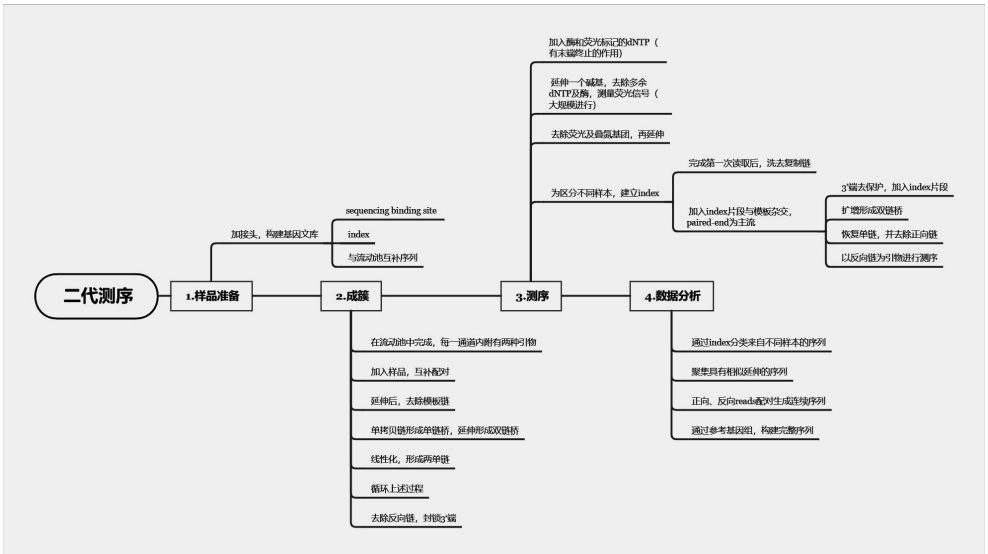
3.“受累亲属对分析”的主要思想

受累亲属对分析是一种遗传研究方法，用于识别遗传病和复杂疾病中的致病基因或易感基因。它的主要思想是通过**研究患有特定疾病的个体及其家族成员之**

间的遗传关系，来寻找与该疾病相关的遗传变异。
①建立家系：首先，需要建立一个包括患者和其家族成员的家系。这些家族成员可能包括患者的父母、兄弟姐妹、子女等。建立家系的目的是确定患者和非患者之间的亲属关系，从而确定他们之间的遗传关联。
②遗传标记分析：接下来，对家系成员进行遗传标记分析。这涉及到对 DNA 样本进行基因型测定，通常使用分子生物学技术（如 PCR 和测序）来确定个体在特定位点上的遗传变异。常用的遗传标记包括单核苷酸多态性（SNP）、微卫星标记等。
③遗传连锁分析：利用遗传标记数据，进行遗传连锁分析来确定家系中与疾病相关的遗传变异。这可以通过计算遗传标记之间的连锁不平衡（Linkage Disequilibrium）来实现。连锁不平衡是指两个或多个遗传标记之间的非随机关联，当一个遗传变异与疾病相关时，与之连锁的遗传标记也可能与疾病相关。
④候选基因筛选：根据遗传连锁分析的结果，确定与疾病相关的遗传位点。这些位点可能位于具体的基因中，因此可以进行候选基因筛选，即确定与疾病相关的候选基因。
⑤功能研究和验证：最后，对候选基因进行功能研究和验证。这可以包括体外实验、动物模型和人群的功能研究，以评估候选基因对疾病发生和发展的影响。

钟凡老师部分

一、表达谱组学分析（上）
参考链接：https://www.jianshu.com/p/c57e518ff507
1.二代测序技术的基本原理和关键技术
第二代测序（Next-generation sequencing，NGS）又称为高通量测序（High-throughput sequencing），是基于 PCR 和基因芯片发展而来的 DNA 测序技术。我们都知道一代测序为合成终止测序，而二代测序开创性的引入了可逆终止末端，从而实现边合成边测序（Sequencing by Synthesis）。
基本原理：https://zhuanlan.zhihu.com/p/58708887
步骤：样本准备-簇生成-测序-数据分析
关键技术：边合成边测序（SBS）



2.三代测序技术的基本原理和关键技术

PacBio 公司的 SMAT

（1）原理概述

应用了边合成边测序的思想，以 SMRT 芯片为测序载体，DNA 聚合酶和模板结合，用 4 色荧光标记 4 种碱基。在碱基配对阶段，不同碱基的加入，会发出不同光，根据光的波长与峰值可判断进入的碱基类型。保持酶活性，区**别反应信号与周围游离碱基荧光背景是关键技术**。

（2）特点

读长长，测序速度快，测序错误率较高，达到 15%，但是出错是随机的，可以通过多次测序来进行有效的纠错。

Oxford 公司的 Nanopore

（1）原理概述

根据碱基所影响的电流变化幅度的不同，设计了一种特殊的纳米孔，孔内共价结合有分子接头。当 DNA 碱基通过纳米孔时，它们使电荷发生变化，从而短暂地影响流过纳米孔的电流强度，灵敏的电子设备检测到这些变化从而鉴定所通过的碱基，是一种基于电信号而不是光信号的测序技术。

（2）特点

读长很长，达到几十 kb，甚至 100kb，错误率在 1%到 4%之间，且是随机错误，通量较高，能够直接读取出甲基化的胞嘧啶。

3.三种测序方法的比较

①**从头组装首选长读本**，它们能提高组装步骤的效率。大多数短读本没有跨越共享区域或共享外显子交界处，使得组装步骤模糊不清。全长转录本测序则无需进行组装。

②由于短读本的通量较高，因此在对转录本进行定量时，短读本是首选。不过，将短读本分配给转录本需要更复杂的概率和统计方法。长读本的通量较低。

4.存储 RNA-seq 下机结果读本质量信息的文件类型及其格式

在 illumina 公司测得的序列文件经过处理以 fastq 文件协议存储为*.fastq 格式文件，在 fastq 文件中每 4 行存储一个 read。

第一行：以@开头接 ReadID 和其他信息

第二行：read 测序信息

第三行：规定必须以“+”开头，后面跟着可选的 ID 标识符和可选的描述内容，如果“+”后面有内容，该内容必须与第一行“@”后的内容相同

第四行：每个碱基的质量得分。记分方法是利用 ERROR P 经过对数和运算分为 40 个级别分别与 ASCII 码的第 33 号!和第 73 号 I 对应。用 ASCII 码表示碱基质量是为了减少文件空间占据和防止移码导致的数据损失

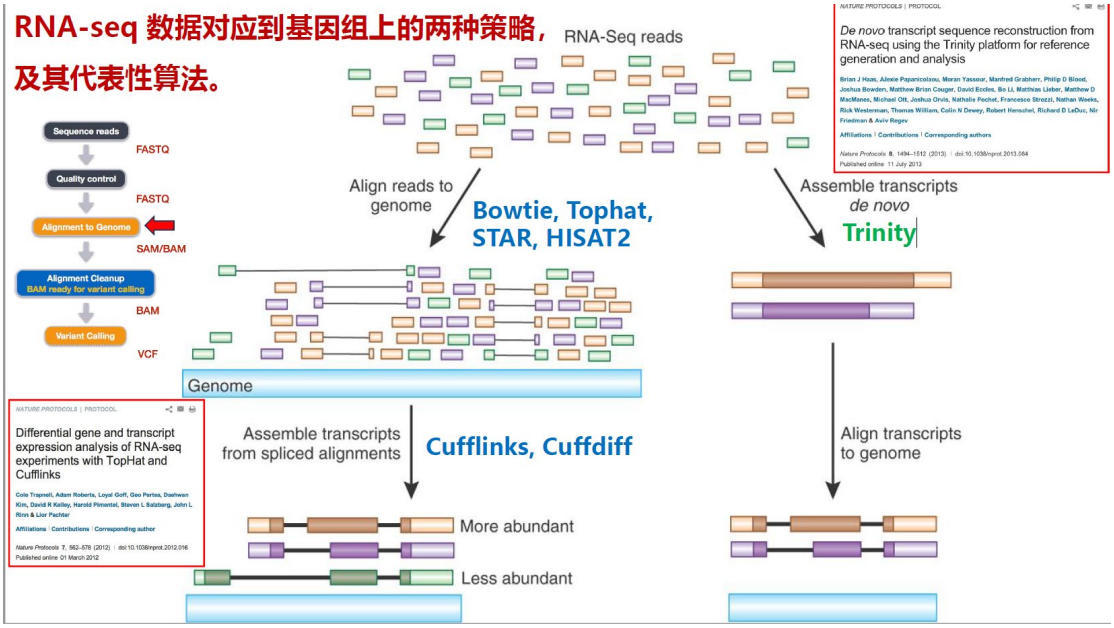
5.Phred 质量分数的意义以及表示方式

表示方式：Q = -10lgP（P 为错误率）

意义：Q_{30}>80% 表示读本 80%的碱基 Phred 质量分数不低于 30，通常用这个标准过滤有效读本进入后续分析。

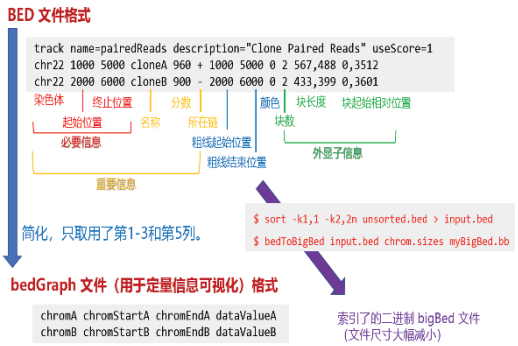
6. FASTQ 文件质控，修剪低质量碱基和过滤低质量读本的原则
读本质控前后，读本内碱基 Phred；质量分数分布，以及长度分布状况。

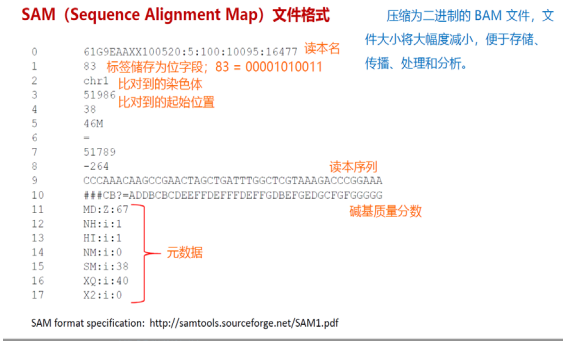
7. 二代测序读本比对到参考基因组上的代表性算法、结果文件类型、结果得到的转录本定量值的表征方法类型
代表性算法：
①Tophat, SATR, Bowtie, HISAT2；Cufflinks, Cuffdiff 等
②Trinity



结果文件类型：
①SAM/BAM 文件：SAM（Sequence Alignment/Map）和它的二进制形式 BAM（Binary Alignment/Map）是存储比对后的测序数据的标准格式。SAM 文件是文本格式，而 BAM 文件是其压缩且索引化的二进制版本，对于大型数据集更高效。

②BED 文件：BED 格式用于描述基因组上的区域、特征和注释信息，比如基因的位置、外显子、染色体范围等。





转录本定量值表征方法：
① FPKM（每百万读数的预期碎片数）/ TPM（每百万转录本数）：常用的一种方法是使用 FPKM（Fragments Per Kilobase Million）或 TPM（Transcripts Per Kilobase

Million），这些是一种标准化方法，用于估计转录本的相对丰度或表达水平，考虑到了转录本长度和测序深度的因素。
②RPKM（每百万读数的预期碱基数）：在基因表达研究中也常用 RPKM（Reads Per Kilobase Million），它与 FPKM 类似，但以读数而不是片段数为单位。
③Counts（计数）：另一种常见的方法是直接使用转录本的读数或计数。这种方法简单直接，计数越高代表转录本表达量越高。

这些方法都是用于评估转录本的表达水平或丰度，并且在研究中经常被用来比较不同条件下的基因或转录本表达差异。

8.存储基因组、转录组、甲基化组等核酸组原始数据的数据库：GEO

二、表达谱组学分析（下）：以表达矩阵为基础的下游分析

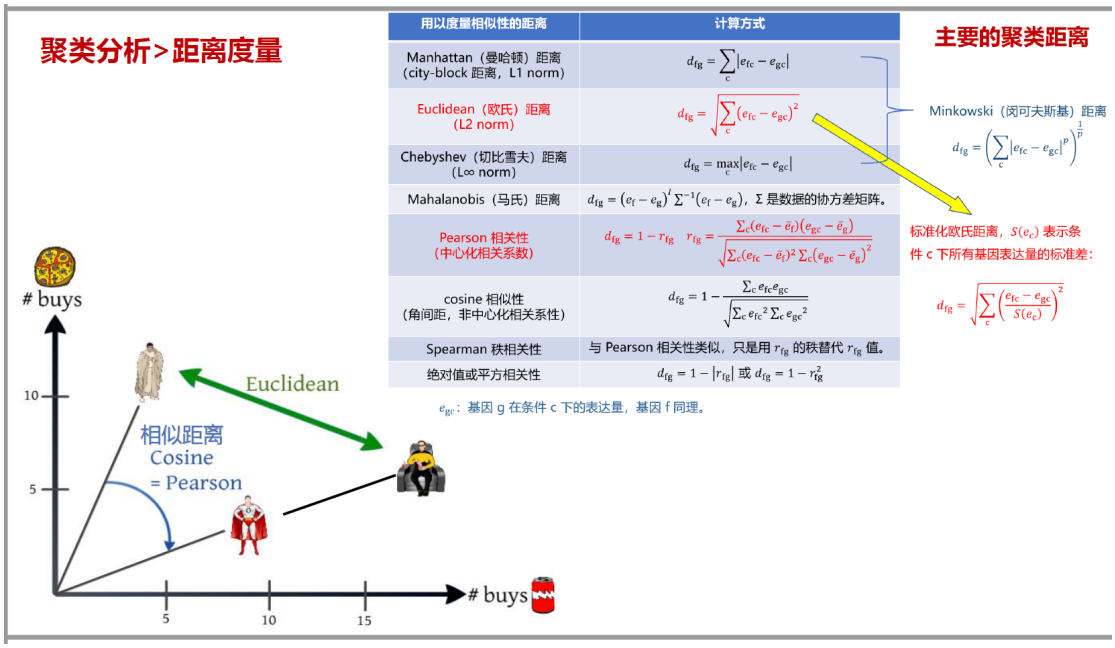
1. 二代测序读本数背景噪音的分布模型及其特点
①泊松分布：泊松分布是一种常用的离散型概率分布，适用于描述单位时间或单位空间内事件发生的次数的分布情况。在二代测序中，泊松分布模型可以描述读本数背景噪音的分布情况。
特点：事件发生的概率是固定的，事件之间是独立的。描述高深度测序数据中的背景噪音。
②负二项分布：负二项分布也是一种离散型概率分布，适用于描述在一系列独立重复试验中，成功次数达到指定次数之前的失败次数的分布情况。在二代测序中，负二项分布模型可以更准确地描述读本数背景噪音的分布情况，因为它考虑了失败次数和成功次数之间的关系。
特点：事件的发生概率可以不固定，事件之间是独立的。描述低深度测序数据中的背景噪音。
应用：Poisson 分布适用于 RNA-seq 的技术重复分析；生物学重复则具有更高的可变性而遵循负二项式（Negative Binomial）分布

2. 组学差异结果筛选的两方面标准及其典型统计量
①显著性（可信度）标准：显著性 p 值，以及通过 Benjamini Hochberg 校正得到的 FDR
②效应强弱标准：差异倍数（Fold Change，FC）、LogRatio、Delta 等

3. Gene Ontology 数据库的特点
①层次结构：GO 数据库采用**树状的层次结构**来组织和表示基因功能、细胞组分和生物过程。这种层次结构使得用户可以从总体到具体地浏览和理解基因功能的关系
②标准化分类体系：GO 数据库提供了一套标准化的分类体系，其中的术语和关系已经经过严格定义和组织。这使得不同研究中的基因功能和生物过程可以进行比较和统一注释
③三个主要领域：GO 数据库分为三个主要领域，包括分子功能（Molecular Function）、细胞组分（Cellular Component）和生物过程（Biological Process）。每个领域都包含一系列术语和关系，用于描述基因在不同方面的功能和参与的生物过程。

④注释数据：GO 数据库与许多其他数据库和资源相连接，可以获取大量的注释数据。这些数据包括基因与特定 GO 术语的关联，以及基因在特定功能、组分和过程中的参与程度
⑤统计分析：GO 数据库还提供了一些统计分析工具，用于研究基因功能和生物过程的富集情况。通过比较实验数据中的基因集与预期随机分布之间的差异，可以揭示基因集在特定功能和过程中的富集情况
⑥动态更新：GO 数据库是一个持续更新和维护的资源。随着科学研究的进展和新的发现，GO 数据库不断更新术语和关系，以确保其与最新的生物学知识保持一致

4.主要聚类距离在生物学问题上的适用性



①欧氏距离：欧氏距离是最为常见和广泛使用的距离度量。它适用于**数值型数据**，例如基因表达数据或蛋白质表达数据。欧氏距离在比较样本之间的数值差异时很有用，可以用来发现相似的表达模式或样本聚类。
②曼哈顿距离：曼哈顿距离是通过计算**两个向量之间的绝对差异**来度量它们之间的距离。它适用于数值型数据，特别是当特征之间的度量单位不同或存在离群值时。曼哈顿距离在基因表达数据分析、代谢组学和蛋白质互作网络分析中广泛应用。
③余弦（cosine）相似度：余弦相似度是通过计算两个向量之间的夹角来度量它们之间的相似性。它适用于描述基因表达数据中的**样本相似性**，尤其是当数据具有高维度且稀疏性较高时。余弦相似度常用于聚类分析、文本挖掘和基因共表达网络分析。

5. Metascape 分析的主要应用场景
①**基因组学和转录组学**分析：Metascape 可以用于对基因组学和转录组学数据进行功能富集分析，帮助研究人员理解在给定条件下基因表达的生物学意义。通过对差异表达基因进行富集分析，可以鉴定与特定生物过程、细胞组分和分子功能

相关的基因集。
②**蛋白质组学**分析：适用于蛋白质组学数据的功能富集分析。它可以帮助研究人员对蛋白质集合进行注释和解释，揭示它们在生物学过程和功能中的作用。
③组织学和病理学研究：可用于对组织学和病理学数据进行分析。通过对基因表达、蛋白质表达和临床数据进行集成分析，可以识别与特定组织类型或疾病相关的生物学过程、细胞组分和分子功能。
④生物网络分析：提供了生物网络分析的功能，可以帮助研究人员在生物网络中鉴定功能模块、富集亚网络和预测功能相关的基因。
⑤**数据可视化和结果解释**：还提供了数据可视化工具，用于呈现富集分析的结果。这些可视化工具有助于研究人员更好地理解和解释分析结果，并将其与其他数据集整合。

6. GSEA 的主要应用场景和配套知识库，以及其优于“基因交盖富集分析”之处
（1）主要应用场景：
①基因表达分析：GSEA 可用于对基因表达数据进行功能富集分析，帮助研究人员理解在不同生物条件下基因表达的生物学意义。通过对基因集的富集分析，可以发现与特定生物过程、细胞组分和分子功能相关的基因集。
②转录因子分析：GSEA 也可以应用于转录因子分析，帮助研究人员识别在特定生物条件下调控基因表达的转录因子。通过对转录因子关联基因集的富集分析，可以揭示转录因子在特定生物过程和信号通路中的作用。
③组织或疾病特异性分析：GSEA 可用于分析组织或疾病特异性的基因表达模式。通过将不同组织或疾病样本的基因表达数据进行比较，可以发现与特定组织或疾病相关的功能模块或信号通路。

（2）配套知识库：
①Gene Ontology（GO）：GO 知识库提供了基因功能、细胞组分和生物过程的标准化分类体系，可以用于对基因集进行功能富集分析。
②KEGG（Kyoto Encyclopedia of Genes and Genomes）：KEGG 知识库包含了基因的信号通路和生物化学反应网络，可用于富集分析和功能注释。
③Reactome：Reactome 知识库是一个基于人类生物学过程的数据库，用于描述和注释基因集在生物过程中的功能富集。
④MSigDB（Molecular Signatures Database）：MSigDB 是一个包含多种基因集的数据库，包括 GO、KEGG、Reactome 等，可用于 GSEA 和其他富集分析方法。

（3）GSEA 相对于"基因交叉富集分析"的优势：
①考虑基因集整体性：GSEA 将基因集作为整体进行分析，而不是仅考虑基因集中的个别基因。这使得 GSEA 能够更全面地捕捉基因集在生物学过程中的相关性和协同作用。
②考虑基因表达的整体分布：GSEA 基于基因表达的整体分布来计算富集分数，而不依赖于特定的显著性阈值。这种方法可以更好地处理基因表达数据中的噪声和变异性。
③不受基因集大小的限制：GSEA 对基因集的大小没有限制，可以处理包含数个基因到数千个基因的基因集。这使得 GSEA 可以发现小型基因集中的相关信号。
④可发现微弱但一致的信号：GSEA 可以发现在单个基因水平上可能不显著的微

弱但一致的信号，通过考虑整个基因集的表达模式来增加富集的敏感性。
7.Connectivity Map（CMap）的主要应用场景
①药物发现和再利用：CMap 可以用于识别已知药物和化合物对于特定疾病或生物过程的影响。通过将基因表达数据与 CMap 数据库中的药物基因表达模式进行比较，可以找到与已知药物相似的化合物或药物，从而发现新的治疗候选物或重新利用现有药物。
②疾病机制研究：CMap 可以帮助研究人员理解疾病的分子机制。通过比较疾病样本与正常样本的基因表达模式，可以识别与疾病相关的基因集。然后，将这些基因集与 CMap 数据库中的基因表达模式进行比较，可以找到与疾病相关的生物过程、信号通路和药物响应模式。
③药物机制研究：CMap 可用于揭示药物的作用机制。通过将药物处理后的基因表达数据与 CMap 数据库中的基因表达模式进行比较，可以找到与药物处理相关的生物过程和信号通路。这有助于理解药物的分子机制和作用靶点。
④基因功能注释：CMap 可以用于对基因功能进行注释。通过将感兴趣的基因与 CMap 数据库中的基因表达模式进行比较，可以发现与这些基因共同调控的生物过程和信号通路，从而帮助解释这些基因的功能和相互作用。

三、生物通路与网络分析
1. 网络及其节点、边、度等的基本定义
网络：相互连接的节点
节点：基因、基因产物（RNA 、蛋白、肽）、其它分子、功能等
边：生物学关系如相互作用、调控、反应、转换、激活、抑制等

2. 生物分子网络的特性
度（degree）的指数分布：“贫富”差距悬殊。
小世界模型：具有小的平均路径长度（节点到节点间的平均最短路径）。
稳健性：有**弹性**并对**随机攻击具有很强的抵抗力，但对靶向攻击则很脆弱**。
层级模块性：具有大的聚集系数（一个节点所有相邻节点之间边的数目站可能的最大边数目的比例）

3. 四种类型生物分子网络的基本概念
I 型：信号转导网络的节点分子是蛋白质或信号小分子，呈现长程级联调控
II 型：转录调控网络会从转录调控因子跨越 DNA 到靶基因蛋白，呈现两层多对多的密集调控
上述两种又被称为基因表达调控网络。
III 型：代谢网络具有与前两者完全不同的拓扑关系：实际中的物质流是代谢分子，它们在以基因产物为节点的网络中表现为连接的作用边。也即代谢网络节点间实际上**不存在直接的相互作用**，而是在通路中起到协同调控代谢分子流的作用。针对特定的数据挖掘目的，代谢网络可能会涉及节点与作用边的交换变换。
IV 型：基于不完备实验信息的 IV 型生物分子网络，即蛋白相互作用（PPI）网络，其本质上是一种实验数据网络，数据量巨大，但缺乏作用方向和效应方向等细节信息，而且作用边也并非 100% 可信。当 IV 型网络中的作用及其细节被确定之后，就会升级成基因表达调控网络。

4. 生物分子通路和生物分子通路网络相比较，各自的特点

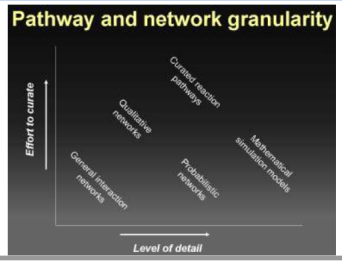
通路（pathway）和网络（network）各自的特点

通路可以看成是网络的一种特殊亚型，其对应于了解较为透彻的生物学进程，具有深度的注释和知识加工，较为详尽的上下文环境，以及可视化美化。

网络
- 具有相互作用或共表达证据的基因及其产物簇；
- 连接通常表示相互作用或共表达的程度；
- 对知识的深度加工不是必需的；
- 预测力较低。

通路
- 一系列链状的化学反应；
- 连接通常表示分子间可量化描述的关系；
- 酶学过程可以被阐明；
- 可预测扰动后的下游变化。

	网络	通路
注释	相对简单：自动的和手工的	困难：绝大数为手工
节点	基因或基因产物（RNA、蛋白）	任何可能的分子
边	作用或定量关系 / 共表达水平	展示元件间可能的可量化机制
保真度	低 – 通常具有很少细节	高 – 针对特定过程
预测力	相对低	相对高



5. KEGG、WikiPathways、Reactome、STRING 四个数据库各自的特点

（1）KEGG (Kyoto Encyclopedia of Genes and Genomes)：

KEGG 是一个综合性的生物信息学数据库，提供了关于基因组、生物化学反应、代谢途径、信号通路和疾病等方面的信息。

特点：①其广泛的生物信息资源，包括基因和蛋白质序列、代谢途径、基因调控网络、药物信息等，为生物学研究和药物开发提供了重要的参考资料；②提供了一套用于数据分析和可视化的工具，如 KegArray、KegEnrich 和 KegGlycan 等，方便用户对数据进行深入的分析和解释。

（2）WikiPathways：

WikiPathways 是一个开放的、协作的在线路径数据库，由科学家社区创建和维护。

特点：①开放性和协作性，任何人都可以贡献和编辑路径信息，使其成为一个不断更新和增长的资源；②WikiPathways 主要关注基因调控网络、代谢途径和信号传导通路等方面的信息，提供基因、蛋白质和代谢物等多个层次的信息。

（3）Reactome：

Reactome 是一个专注于人类生物学过程的开放性数据库，涵盖了代谢、信号传导、基因表达调控、免疫系统等多个生物学领域。

特点：①详细和精确的生物学过程注释，提供了丰富的生物学反应、分子参与者和调控信息；②提供了数据分析工具和可视化工具，如 Pathway Browser、Pathway Analysis 等，支持用户进行数据解析和可视化分析。

（4）STRING：

STRING 是一个蛋白质相互作用网络数据库，提供了基于已知和预测相互作用的蛋白质网络。

特点：①重点关注蛋白质相互作用和功能注释，可以帮助研究人员理解蛋白质间的相互作用关系以及功能模块的组织；②提供了丰富的功能注释、蛋白质复合物信息和生物学通路的链接，方便用户进行更深入的研究和分析。

6. CytoScape 软件的主要应用场景
①生物分子网络分析：Cytoscape 广泛用于分析和可视化生物分子网络，如蛋白质相互作用网络、基因调控网络、信号通路等。它可以帮助研究人员理解网络的拓扑结构、节点间的相互作用关系以及功能模块的组织。
②数据整合和可视化：Cytoscape 可以整合多种生物学数据，如基因表达数据、蛋白质互作数据、基因组数据等。通过将这些数据与网络结构关联起来，Cytoscape 可以提供多维度的可视化展示，帮助用户观察和解释数据。
③生物标志物发现：Cytoscape 可以用于生物标志物（biomarker）的发现和分析。通过结合基因表达数据和网络拓扑信息，研究人员可以在网络中识别出与特定生物过程或疾病相关的关键节点或功能模块，从而发现潜在的生物标志物。
④药物靶点识别：Cytoscape 在药物研究中也有应用。通过整合化合物-靶点互作信息和蛋白质相互作用网络，研究人员可以预测和识别潜在的药物靶点，帮助药物研发的目标鉴定和药物作用机制的解析。
⑤网络模拟和分析：Cytoscape 提供了一系列的插件和工具，支持网络模拟、拓扑分析、网络聚类等功能。用户可以进行网络模型的构建和模拟，以及一系列的网络分析和统计，帮助理解网络的动态特性和调控机制。

7. 造成生物分子网络、通路数据库异质性问题，数据库间交盖度不高的原因
①通路的定义
②中间步骤的数目
③可变底物的数目
④反应物 ID 号的丢失
⑤基因产物不正确的定位信息

8. 基于组学数据从头网络构建和聚类分析的两种代表性算法，及其各自特点
WGCNA（Weighted Gene Co-expression Network Analysis，加权基因共表达网络分析）：
特点：
WGCNA 算法通过计算基因间的共表达关系构建基因共表达网络。它基于基因之间的相关性来发现共表达模块，并将相似的基因聚类在一起。
WGCNA 使用基于 Pearson 相关系数的加权方法来衡量基因之间的关联性，可以捕捉到基因之间的非线性相关关系。
WGCNA 将基因共表达网络转化为模块，每个模块代表一组高度相关的基因，有助于发现基因功能模块和关键调控基因。
WGCNA 还可以对模块与临床特征之间的关系进行分析，帮助研究人员理解基因网络与表型之间的关联性。
SCENIC（Single-Cell Regulatory Network Inference and Clustering，单细胞调控网络推断和聚类）：

特点：
SCENIC 算法用于从单细胞转录组数据中推断细胞调控网络。它可以鉴定细胞类型特异的转录因子及其调控的基因。
SCENIC 通过整合基因共表达网络和转录因子的调控信息，发现细胞类型特异的转录因子-基因调控模块。

SCENIC 使用基于基因表达数据的因子化方法，将细胞类型转录因子的调控网络拆分为离散的调控基因集合，从而识别细胞类型的转录因子调控模式。
SCENIC 还可以进行细胞聚类分析，将具有相似调控模式的细胞聚类在一起，以揭示细胞类型的异质性和调控网络的变化。

四、机器学习方法概要
1.有监督和非监督机器学习方法的区别，及其包括的常见算法
（1）有监督学习（Supervised Learning）：
①特点：有监督学习使用已标记的训练数据集，其中包含输入特征和相应的目标变量（标签）。算法通过学习输入特征与目标变量之间的关系来训练模型，以便对新的未标记数据进行预测或分类。
②常见算法：
线性回归（Linear Regression）
逻辑回归（Logistic Regression）
支持向量机（Support Vector Machines，SVM）
决策树（Decision Trees）
随机森林（Random Forests）
梯度提升树（Gradient Boosting Trees）
人工神经网络（Artificial Neural Networks）

（2）非监督学习（Unsupervised Learning）：
①特点：非监督学习使用未标记的训练数据集，其中只包含输入特征，没有对应的目标变量。算法通过学习数据内部的结构、模式或关系来发现隐藏的模式或进行数据降维等任务。
②常见算法：
聚类分析（Clustering）
K 均值聚类（K-means Clustering）
层次聚类（Hierarchical Clustering）
DBSCAN（Density-Based Spatial Clustering of Applications with Noise）
关联规则学习（Association Rule Learning）
Apriori 算法
FP-Growth 算法
主成分分析（Principal Component Analysis，PCA）
自组织映射（Self-Organizing Maps，SOM）
高斯混合模型（Gaussian Mixture Models，GMM）

2.回归方法的应用场景
①基因表达预测：回归方法可用于预测基因表达水平。通过建立基因表达与相关特征（如 DNA 序列、甲基化状态等）之间的回归模型，可以预测基因在不同条件下的表达水平，帮助理解基因调控机制和疾病发生的潜在机理。
②蛋白质结构和功能预测：回归方法可用于预测蛋白质结构和功能。通过将蛋白质序列和结构特征作为输入，建立回归模型来预测蛋白质的二级结构、折叠状态、配体结合位点等信息，帮助理解蛋白质的功能和相互作用。
③药物活性预测：回归方法可用于预测药物分子的活性和亲和性。通过将分子结

构和化学描述符作为输入，建立回归模型来预测药物分子与靶点之间的结合亲和性，帮助筛选候选药物和设计新的药物分子。
④疾病预后和生存分析：回归方法可用于预测疾病患者的预后和生存情况。通过将临床特征、基因表达数据和其他相关数据作为输入，建立回归模型来预测患者的生存时间、治疗反应和疾病进展情况，帮助个体化医疗决策和治疗策略制定。
⑤疾病风险预测：回归方法可用于预测个体患某种疾病的风险。通过将个体的遗传变异、环境因素和生活方式等作为输入，建立回归模型来预测患病的概率或风险评分，帮助早期筛查、个体化预防和干预措施。

3.决策树和随机森林的关系

决策树：决策树是一种基于树状结构的分类和回归算法。它通过将数据集分割成不同的子集，每个子集对应树中的一个节点，最终形成一个树形结构。决策树通过在每个节点上选择最佳的特征进行分割，以最小化样本的不纯度（如熵或基尼指数）来进行分类或预测。决策树具有易于理解和解释的优点，并且可以处理数值型和类别型数据。

随机森林：随机森林是一种基于集成学习的算法，它通过构建多个决策树来进行分类和回归。随机森林中的每个决策树都是独立生成的，通过对训练数据进行有放回的随机抽样（袋装法）来创建不同的训练集，然后针对每个训练集生成一个决策树。最后，对于分类问题，随机森林通过投票或平均预测结果来确定最终的分类结果；对于回归问题，随机森林通过平均预测结果来确定最终的预测值。

关系：随机森林是基于决策树的集成学习方法，它通过集成多个决策树的预测结果来提高模型的准确性和鲁棒性。每个决策树都是独立生成的，它们之间没有直接的联系。随机森林通过决策树的多样性来减少过拟合风险，并通过集成多个决策树的预测结果来提供更稳定和可靠的预测。由于每个决策树都是基于随机样本和随机特征选择生成的，随机森林还具有一定的抗噪能力和泛化能力。

总结而言，决策树是随机森林的基本组成单元，而随机森林则是通过集成多个决策树来提高模型性能和稳定性的集成学习方法。

4.使用 K 近邻（KNN）时应注意的问题

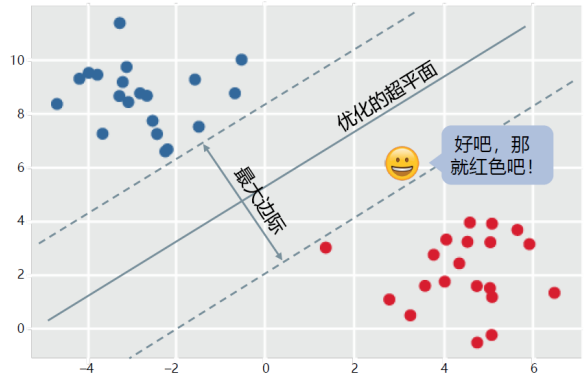
①距离度量：KNN 算法使用距离度量来计算样本之间的相似性。选择合适的距离度量方法对算法的性能影响很大。常用的距离度量方法包括欧氏距离、曼哈顿距离、闵可夫斯基距离等，应根据具体问题的特点选择适当的距离度量方法。
②K 值选择：KNN 算法中的 K 值代表着**选取最近邻居的数量**。选择合适的 K 值对算法的性能也有重要影响。较小的 K 值可能对噪声敏感，容易过拟合；较大的 K 值可能会平滑决策边界，容易欠拟合。选择适当的 K 值需要进行交叉验证或者其他模型选择方法。
③特征归一化：在应用 KNN 算法之前，通常需要对特征进行归一化处理。因为 KNN 算法是基于样本之间的距离进行分类或回归的，如果特征具有不同的尺度或者范围，那么距离计算可能会受到影响。常见的归一化方法包括将特征缩放到[0, 1]范围内或者使用标准化方法。
④处理数据不平衡：当样本类别不平衡时，KNN 算法可能会受到影响。具有较大数量的类别可能对预测结果产生更大的影响，并且可能导致预测偏向于占多数的类别。在处理数据不平衡时，可以考虑使用加权 KNN 或者采用过采样和欠采样等方法来平衡类别分布。

⑤高维问题：KNN 算法在高维数据集上可能会遇到维度灾难问题。随着维度的增加，样本之间的距离变得更加稀疏，导致 KNN 算法的性能下降。可以通过特征选择、降维等方法来解决高维问题，以提高算法的效率和准确性。

5.支持向量机的基本概念和特点

支持向量机（SVM）

SVM 算法找到最佳超平面（hyperplane），以尽可能大的余量将数据点分开。支持向量一般就是那些与分界点较近的训练样本点，它们在训练 SVM 或者确定超平面时起关键作用。



SVM 是一种以监督分类学习方式对数据进行二元的广义线性分类器，具有稀疏性和稳健性，且具有良好的泛化能力。

我们把分离超平面到最近的表达向量的距离定义为超平面的边际，那么 SVM 就会选择最大边际的分离的超平面。选择这个特定的超平面可以使 SVM 的预测正确分类以前未见过的例子。

支持向量机（SVM）

支持向量机具有高准确率，为避免过拟合提供了很好的理论保证，而且就算数据在原特征空间线性不可分，只要给个合适的核函数，它就能运行得很好。在动辄超高维的文本分类问题中特别受欢迎。

优点：

- 可以解决高维问题，即大型特征空间；
- 解决小样本下机器学习问题；
- 能够处理非线性特征的相互作用；
- 无局部极小值问题（相对于神经网络等算法）；
- 无需依赖整个数据；
- 泛化能力比较强。

缺点：

- 当观测样本很多时，效率并不是很高；
- 对非线性问题没有通用解决方案，有时候很难找到一个合适的核函数；
- 对于核函数的高维映射解释力不强，尤其是径向基函数；
- 常规 SVM 只支持二分类；
- 对缺失数据敏感。

6.过拟合和欠拟合的概念，及其发生的可能原因

（1）过拟合（Overfitting）：

概念：过拟合指的是模型在训练集上表现良好，但在未见过的测试数据上表现较差的情况。它表示模型过于复杂，过度拟合了训练集中的噪声和细节，从而导致对新数据的泛化能力下降。

可能原因：

①模型复杂度过高：模型具有过多的参数或复杂的拟合函数，使其可以在训练集上完美地拟合每个样本的细节和噪声，但对新数据的泛化能力较差。

②数据不平衡：训练集中某些类别的样本数量过少，导致模型在这些类别上过度

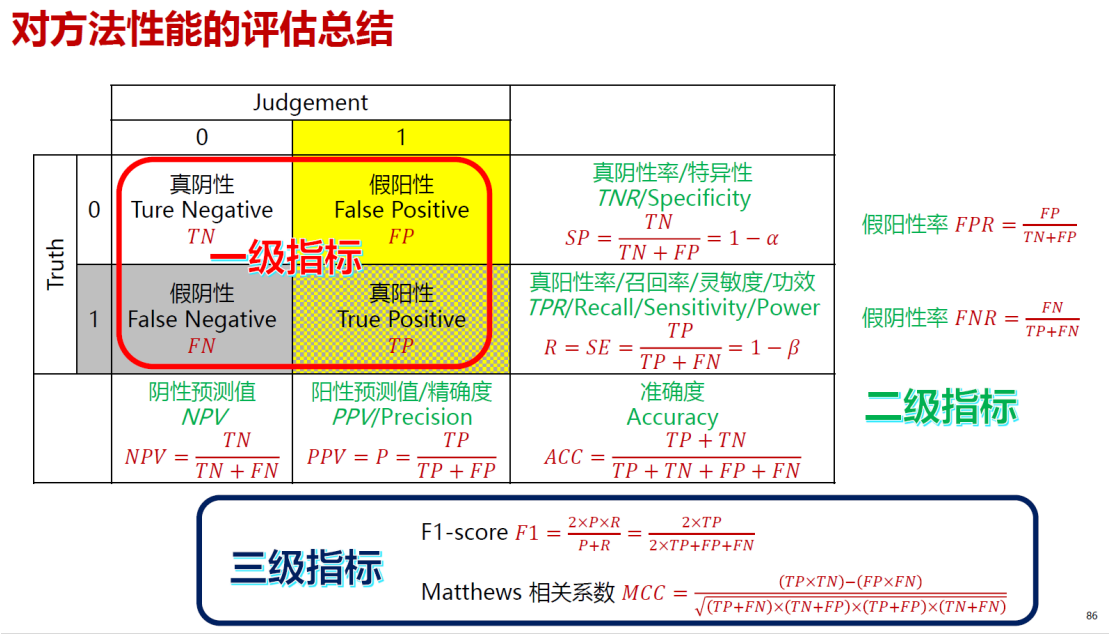
拟合。
③噪声数据：训练集中存在噪声或异常值，模型可能会过度拟合这些异常值。
过度训练：迭代次数过多或训练集上训练时间过长，模型可能过度学习训练集中的细节和噪声。

（2）欠拟合（Underfitting）：

概念：欠拟合指的是模型无法很好地拟合训练数据，表现为在训练集和测试集上都表现较差的情况。它表示模型过于简单，不能捕捉到数据的复杂关系和模式。
可能原因：
①模型复杂度过低：模型的容量不足，无法拟合数据中的复杂关系。
特征选择不当：特征提取或选择的方式不合适，导致模型无法捕捉到数据中的重要特征。
②数据量不足：训练集过小，不足以充分训练模型，导致欠拟合。
③数据噪声过多：训练集中存在大量噪声或异常值，干扰了模型的学习过程。

（3）解决过拟合和欠拟合问题的方法包括：
过拟合：增加训练数据量、减小模型复杂度（如减少参数、使用正则化）、特征选择、降维、数据增强、提前停止训练等。
欠拟合：增加模型复杂度（如增加参数、引入更多特征）、改进特征选择、增加训练数据量、使用更复杂的模型等。

7.对机器学习方法性能评估的二级指标



8.深度学习神经网络的基本概念
深度学习神经网络是一种基于人工神经网络的机器学习方法，通过多层非线性变换实现高度抽象的特征表示和学习。它可以自动从大量数据中学习到复杂的模式和规律，被广泛应用于图像识别、自然语言处理、语音识别等领域。
9.深度学习神经网络包含的主要类型，及其各自适用的场景
①生成对抗网络（Generative Adversarial Network，GAN）：适用于**生成具有逼真性的新样本**，如图像生成、图像增强、生成对抗攻击等任务。

②循环神经网络（Recurrent Neural Network，RNN）：适用于**处理序列数据**（如语音、文本等），如自然语言处理、语音识别、机器翻译等任务
③卷积神经网络（Convolutional Neural Network，CNN）：适用于处理具有网**格结构数据（如图像）**的任务，如图像识别、目标检测、图像生成等
④自编码器（Autoencoder）：适用于数据降维、特征提取、异常检测等任务

刘雷老师部分：

1.临床信息学的基本概念：

临床信息学是将生物医学信息学应用于患者护理领域的学科，结合了计算机科学、信息科学和临床科学，旨在管理和处理临床数据、信息和知识，以支持临床实践。

2.医院信息系统包括哪些：

医院信息系统包括医院管理信息系统（HIS）、电子病历系统（EHR）、检验信息系统（LIS）和临床专科信息系统等。

HIS 用于医院及其各部门的综合管理，EHR 是电子病历数据的主要来源，LIS 用于检验闭环全流程管理，临床专科信息系统则针对专科诊疗服务和科研信息需求。

3.真实世界数据应用的特点与挑战：

特点：数据来源多样、数据量大、数据质量不一、数据结构复杂，并且数据的获取和整合涉及多个系统和部门。

挑战：数据隐私和安全保护、数据质量控制、数据整合和标准化、数据分析和解释等。

4.数据库设计的基本概念：

数据库设计是指按照特定需求和目标，确定数据库的结构、关系和约束的过程。基本概念包括实体、属性、关系、主键、外键、范式等。

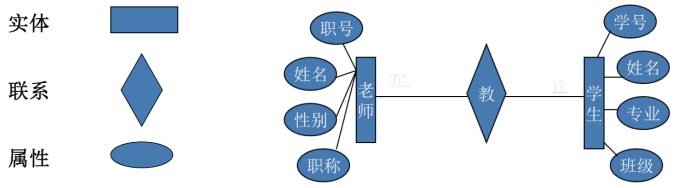
5.实体关系图（ER 图）：

ER 图（Entity-Relationship Diagram）是一种用于表示实体、属性和实体之间关系的图形工具，常用于数据库设计和模型表示。

实体关系图

ER图有三个基本成分：

- 矩形框，用于表示实体类型（考虑问题的对象）
- 菱形框，用于表示联系类型（实体间联系）
- 椭圆形框，用于表示实体类型和联系类型的属性。



基数性质（Cardinality）：基数性质描述了实体之间关系的数量关系。常见的基数性质有一对一（1:1）、一对多（1:N）和多对多（N:M）。基数性质可以在关系线上用箭头或标注来表示。

6、数据库的基本术语：

数据库的基本术语包括表（Table）、字段（Field）、记录（Record）、主键（Primary Key）、外键（Foreign Key）、索引（Index）等。

①关系（Relation）：关系是关系模型中的基本组成单元，类似于数据表。它由一组具有相同结构的元组（Tuple）组成，每个元组表示一个记录。

②元组（Tuple）：元组是关系模型中的一行，表示一个实体或数据项。每个元组由一组属性值组成，每个属性值对应关系模型中的一个属性。

③属性（Attribute）：属性是关系模型中的一列，表示元组的特定特征或数据项。每个属性具有特定的数据类型，如文本、数字、日期等。

④域（Domain）：域是属性的取值范围或数据类型定义。它规定了属性可以包含的值的类型和约束条件。

⑤主键（Primary Key）：主键是关系模型中用于唯一标识每个元组的属性或属性组合。主键的值在关系中必须是**唯一且不为空**，用于确保数据的唯一性和引用完整性。

⑥外键（Foreign Key）：外键是一个属性或属性组合，用于建立关系之间的关联。它引用另一个关系的主键，用于维护关系之间的引用完整性。

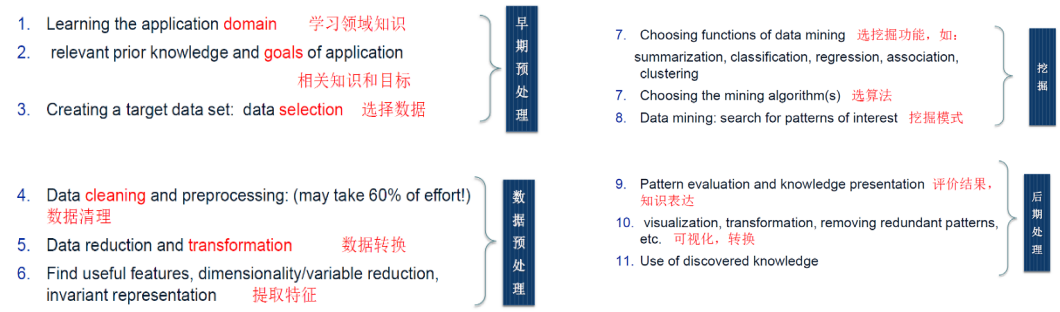
⑦候选键（Candidate Key）：候选键是满足唯一性要求的属性或属性组合。一个关系可以有多个候选键，其中选择一个作为主键。

关系模式（Relation Schema）：关系模式定义了关系的结构和属性的集合。它包括关系的名称、属性的名称和属性的数据类型。

7.SQL 语法：

SQL（Structured Query Language）是用于管理和操作关系型数据库的语言。SQL语法包括数据查询（SELECT）、数据插入（INSERT）、数据更新（UPDATE）、数据删除（DELETE）等命令。

8.数据挖掘的基本流程包括每一大步骤里的小步骤：



9.OHDSI 的基本理念：

OHDSI（Observational Health Data Sciences and Informatics）是一个全球性的协作社区，致力于利用大规模医疗健康数据进行观察性研究和医疗信息学的发展。它的基本理念包括以下几个方面：

①开放和透明：OHDSI 倡导开放和透明的研究环境，鼓励共享数据、工具和研究结果。成员可以共享匿名化的健康数据集、分析代码和方法，促进知识的共享和协作。

②观察性研究：OHDSI 关注观察性研究，即基于现有的临床实践和健康数据进行研究。它利用电子医疗记录、医保记录和其他现实世界数据，为研究人员提供了机会对医疗决策和治疗效果进行评估。

③**开放标准**：OHDSI 积极推动和采用开放标准化的方法和工具，以实现数据互操作性和比较性研究。它使用 OMOP（Observational Medical Outcomes Partnership）标准化数据模型和一致的术语表，使不同机构和数据源的数据能够进行集成和比较分析。

④协作和跨学科：OHDSI 鼓励跨学科的合作，将医学、生物统计学、计算机科学和其他相关领域的专业知识结合起来。成员可以共同开发和改进分析工具、方法和最佳实践，以推动观察性研究的发展。

⑤知识生成和决策支持：OHDSI 旨在生成高质量的证据，为医疗决策和政策制定提供支持。它通过大规模数据分析和比较效果研究等方法，帮助揭示治疗效果、药物安全性和其他与健康相关的问题。

⑥教育和培训：OHDSI 提供教育和培训资源，帮助研究人员和临床专业人员学习和应用 OHDSI 的方法和工具。它组织培训活动、举办会议和提供在线资源，支持成员的能力建设和学习。

10.医学术语集有哪些？

医学术语集包括标准化的医学词汇、术语和编码系统，如国际疾病分类（ICD）、医学主题词汇（MeSH）、医学术语系统化标准 SNOMED、统一医学诊断编码（SNOMED CT）等。

11.ETL 的步骤：

ETL（Extract, Transform, Load）是指从源系统中提取数据、对数据进行转换和清洗，并将数据加载到目标系统的过程。步骤包括数据抽取、数据转换和数据加载。

①提取（Extract）：

- 识别和连接数据源：确定要提取数据的源系统或数据源，并与之建立连接。

• 定义提取范围：确定要提取的数据范围，例如特定的表、文件或查询条件。
• 提取数据：从源系统中抽取数据，可以通过数据库查询、API 调用、文件导出等方式进行。
②转换（Transform）：
• 数据清洗：处理缺失值、重复值、异常值和格式错误等数据质量问题。
• 数据整合：将多个数据源的数据进行合并和集成，创建一个一致的数据集。
• 数据转换：对数据进行转换和重塑，包括数据格式转换、数据合并、数据拆分数据计算和数据衍生等。
• 数据标准化：将数据转换为一致的格式和单位，使其符合目标系统的要求。
• 数据验证和校验：验证转换后的数据的准确性和完整性，进行数据质量控制。
③加载（Load）：
• 目标定义：确定目标系统或数据仓库的结构和模型。
• 数据映射：将转换后的数据映射到目标系统的表、字段或模型中。
• 数据加载：将转换后的数据加载到目标系统中，可以使用数据库插入、更新或追加等方式进行。
• 数据索引和优化：根据目标系统的要求创建索引和优化数据加载性能。

13、队列的基本概念和建立队列的规则：

队列是一种常见的数据结构，用于按照先进先出（FIFO）的原则管理和操作数据。它基本上是一个有序的元素集合，新元素被添加到队列的一端，称为队尾（rear），而现有元素则从队列的另一端被移除，称为队首（front）。

以下是建立队列的基本规则：

①入队（Enqueue）：将新元素添加到队列的队尾。新元素成为队列中的最后一个元素。如果队列为空，则新元素同时成为队首和队尾。

②出队（Dequeue）：从队列的队首移除一个元素。移除后，队列中的下一个元素成为新的队首。如果队列只有一个元素，那么在出队后，队列将为空。

③队首元素（Front）：获取队列中的第一个元素，即队首元素，但不对其进行移除。

④队列空（Empty）：判断队列是否为空。如果队列中没有任何元素，即为空。

⑤队列长度（Size）：获取队列中元素的数量。

建立队列时，需要遵守以上规则来维护队列的特性。入队操作仅在队列的队尾添加元素，而出队操作仅从队列的队首移除元素。这种方式确保了先进先出的顺序，即最先入队的元素最先被移除。

队列可以通过不同的方式实现，例如使用数组或链表等数据结构。在实现队列时，还需要考虑队列的空间限制和性能需求，确保队列操作的效率和可靠性。

需要注意的是，队列并不支持在队列中间插入或删除元素，它只允许在一端进行插入和删除操作。如果需要在任意位置进行插入和删除操作，可以考虑其他数据结构，如链表或数组列表。

16、个体水平预测是什么？临床问题的例子：

个体水平预测是指基于患者的个体特征和临床数据，通过建立模型来预测患者的个体结果，如患病风险、治疗效果等。例如，根据患者的年龄、性别、家族病史等信息，预测其患某种疾病的概率。

17、群体水平评估是什么？临床问题的例子：

群体水平评估是指对一群患者或人群的特征和临床数据进行分析和评估，从而了解群体的整体情况和统计特征。例如，对某种疾病的患者群体进行统计分析，得出平均患病年龄、不同年龄段患病比例等信息。

18、个体水平预测与群体水平评估的差别：

个体水平预测和群体水平评估是在医学和流行病学领域中两种不同的分析方法，它们的差别在于关注的**对象和目的。**

个体水平预测（Individual-level prediction）关注的是对单个个体进行预测和评估。它基于个体的特征、病史、生物标志物等信息，通过建立模型来预测个体的风险、治疗效果、转归等。个体水平预测的目的是为个体提供个性化的医疗决策和临床管理，以实现精准医学的目标。个体水平预测可以为医生和患者提供定制的治疗方案、药物选择、预后评估等，以提高医疗效果和个体健康状况。

群体水平评估（Population-level assessment）关注的是对整个人群或群体进行评估和比较。它通常基于大规模的流行病学数据，用于描述和推断人群的疾病发病率、死亡率、流行趋势等。群体水平评估的目的是了解整个人群的健康状况和疾病负担，以制定公共卫生政策、预防策略和资源分配。群体水平评估可以帮助决策者了解人群的健康需求、制定公共健康干预措施、监测疾病流行趋势等。

总结起来，个体水平预测关注的是单个个体的预测和个性化医学，目的是为个体提供个性化的医疗决策和管理；而群体水平评估关注的是整体人群的评估和比较，目的是了解人群的健康状况和制定公共卫生政策。两者在数据来源、分析方法和应用领域上有所不同，但都对医学和流行病学的研究和实践具有重要意义。