**生物信息学知识点汇总**

1. 生物信息学的概念（教材）：生物信息学是研究生物信息的采集、处理、储存、传播、分析和解释等的一门学科。它通过综合利用分子生物学、遗传学、计算机科学与技术，来揭示大量且复杂的生物数据所赋有的生物学奥秘。其根本目标是增加对生物学过程的认识。
2. 生物信息学与传统生物学的关系：可以概括为互补和协同。传统生物学侧重于实验方法和直接观察来研究生物体、细胞、生物分子等，而生物信息学则利用计算机科学、数学和统计学的方法来分析和解释生物数据，尤其是大规模的分子数据。生物信息学的出现极大地扩展了传统生物学的研究范围，使得科学家能够处理和解释前所未有的大量复杂数据，从而在基因组学、蛋白质组学和系统生物学等领域取得了重大进展。同时，生物信息学的发展也依赖于传统生物学的实验数据和理论基础，两者相互依存，共同推动生命科学的发展。**简而言之，生物信息学为传统生物学提供了强大的数据分析工具和理论模型，而传统生物学则为生物信息学提供了实验验证和应用场景。**
3. 生物信息的数据类型：核酸序列数据、蛋白质序列和结构数据、分子标记数据、生物芯片数据和生物表型数据等。
4. 第一代测序技术又叫Sanger测序，是英国科学家Sanger发明的，目前一般有效平均读长约为500~800 bp，总体准确度约为99.999％。由于其准确性极高，目前依然是**测序技术的金标准**，广泛应用于基因克隆、载体构建、变异检测、小基因组测序和医学检测等。
5. 第一代测序数据的基本原理：利用一种DNA聚合酶来延伸结合在特定序列模板上的引物，直到掺入一种链终止核苷酸为止。每一次序列测定由一套四个单独的反应构成，每个反应含有所有四种脱氧核苷酸三磷酸（dNTP），并混入限量的一种不同的**双脱氧核苷三磷酸**（**ddNTP**）。由于ddNTP缺乏延伸所需要的3-OH基团，使延长的寡聚核苷酸选择性地在A、T、G或C处终止并在每个碱基后进行标记，产生以A、T、G、C等结束的四组不同长度的一系列核苷酸，然后在尿素变性的PAGE胶上进行电泳检测，从而获得可见的DNA碱基序列。
6. 二代测序的代表性平台Illumina：2006年，Solexa公司推出第二代测序系统GA，其中**DNA簇**（DNA cluster）、**桥式PCR**（bridge PCR）和**可逆阻断**（reversible terminator）等核心技术使GA系统具有高通量、低成本的优势。后来，该公司被Illumina收购，陆续推出了多款低通量的桌面和高通量的台式测序仪，目前已经成为二代市场最主流的测序平台。
7. Illumina测序的原理主要是**边合成边测序**。
8. Q值是用来评估测序碱基准确性的标准之一，一般Q值越高，碱基准确性越高。
9. 三代测序又称为单分子测序技术，它能够直接对单个DNA分子进行测序，而无需像传统的二代测序技术那样进行PCR扩增。这种技术的主要优势在于其读长更长，能够提供更为连续和完整的基因组信息，尤其适合用于大型、复杂基因组的组装和结构变异的研究。
10. 目前主流的三代测序分别是PacBio公司的单分子荧光测序和Nanopore公司的纳米孔测序。单分子荧光测序是通过在单个DNA分子上直接添加带有荧光标记的核苷酸，随后去除荧光标记并重复此过程来实现测序。而纳米孔测序则是利用电导率的变化来检测通过纳米孔的单个DNA分子上的碱基序列。
11. 三代测序PacBio的主要三种模式：①**CLR测序**（Continuous Long Read）：PacBio最初的测序模式，产生非常长的读段，平均读长可达10-20kb，最长可超过60kb。它可以生成大量的长读段数据，非常适合用于新基因组的组装和复杂基因组区域的解析。②**CCS测序**（Circular Consensus Sequencing）：通过对同一分子进行多次测序来提高序列准确性的技术。单个DNA分子形成一个闭环，多次读取，通过一致性分析生成高质量的序列。CCS产生的读段比CLR短，但准确性更高。③**HiFi测序**：基于CCS的进一步发展，它结合了长读段的优势和高准确度的需求。HiFi读段具有极高的准确度（>99.9%），同时保持较长的读长，是理想的全基因组测序解决方案。
12. 不同世代测序技术的异同点：



1. 蛋白质结构测定的主要方法：X射线晶体学、核磁共振（NMR）光谱学和冷冻电子显微镜（Cryo-EM）。
2. 分子数据库：分子数据库是存储和管理生物分子信息的电子数据库系统，包括储存有核苷酸序列、蛋白质序列和结构等的**初级数据库**和基于初级数据库建立起来的**二级数据库**，为生物信息学研究提供数据支持。
3. 分子生物信息数据库的格式种类繁多，较常见的有FASTA（序列文件）、FASTQ（测序文件）、GFF3（注释文件）和GenBank（注释文件）等格式。
4. **FASTA格式**：一种文本格式，用于表示核酸序列或蛋白质序列。它由一个以">"字符开始的标题行，后跟序列数据的行组成。标题行通常包含序列的标识符和描述信息，而序列数据是连续的，可以跨多行。
5. **FASTQ格式**：用于存储序列数据及其对应的质量分数。它包括四行表示一个序列记录：第一行以"@"开始，后跟序列标识符；第二行是原始序列；第三行以"+"开始，可能后跟与第一行相同的序列标识符；第四行表示每个序列字符的质量分数，通常使用ASCII码表示。
6. **GFF3格式**：GFF（General Feature Format）格式版本3，用于描述基因和其他功能区的位置及属性。它是一个九列的表格格式，包括：序列ID、源、特征类型、起始位置、结束位置、得分、链方向、相位以及属性。
7. **GenBank格式**：一种用于存储广泛类型的遗传标记的详细注释信息的格式。它提供了关于特定DNA或蛋白质序列的详细信息，包括序列的来源、分类、文献引用、特征及其位置等。
8. 国际三大主要生物信息数据库：**NCBI**、**EMBL-EBI**和**DDBJ**。
9. **NCBI数据库**：国家生物技术信息中心（National Center for Biotechnology Information, NCBI）成立于1988年，是美国国立卫生研究院（NIH）下属的一个机构。NCBI的主要任务是开发生物信息学工具，提供生物大数据的存储和检索服务，以支持生命科学和医学研究。NCBI承担着收集、整理、分析和发布生物学分子数据的重要职责，为科研人员提供了一个重要的数据资源和分析工具库。
10. NCBI重要子库：
	* 1. **GenBank**：一个综合性的公共DNA序列数据库，收录了从全球范围内提交的所有公开可用的核酸序列。
		2. **PubMed**：一个免费的搜索引擎，主要用于检索生命科学和医学领域的文献信息。它提供对MEDLINE数据库及其他相关文献的访问。
		3. **BLAST** ：Basic Local Alignment Search Tool是一个序列比对工具，用于比较给定的核酸序列或蛋白质序列与数据库中的序列，以找出序列之间的相似性。
11. 我国建立的主要生物信息数据库：CNCB数据库（China National Center for Bioinformation），中国国家生物信息中心，提供一个综合性的生物大数据服务平台，旨在收集、整理和分享生物学数据，支持全球生命科学研究。
12. UniProt数据库：一个综合性的蛋白质序列和功能信息数据库，旨在提供一个关于蛋白质的结构、功能和序列信息资源。它由三个主要组成部分构成：UniProtKB、UniRef 和 UniParc。是由欧洲生物信息研究所(EBI)、瑞士生物信息学研究所(SIB)和蛋白质信息资源(PIR)共同维护的。
13. PDB数据库：PDB（Protein Data Bank）是一个蛋白质结构数据库，专门收集存储蛋白质和核酸的三维结构数据。这些结构数据主要通过X射线晶体学、核磁共振（NMR）光谱学和冷冻电子显微镜（Cryo-EM）等实验方法获得。
14. Ensembl数据库：Ensembl是脊椎动物基因组的基因组浏览器，支持比较基因组学、进化、序列变异和转录调控的研究。
15. EnsemblPlants数据库：EnsemblPlants 是 Ensembl 数据库的一个专门针对植物的子数据库。它包含了超过50种植物的基因组数据，包括水稻、小麦、玉米、大豆、拟南芥等。
16. Phytozome数据库：一个植物比较基因组学数据库，由美国能源部旗下的联合基因组研究所（JGI）维护。它专注于提供植物基因组、基因模型、功能信息和分析工具。
17. **TAIR数据库**：The Arabidopsis Information Resource 维护有拟南芥（*Arabidopsis thaliana*）的遗传和分子生物学数据集，包括基因模型、分子证据、生物学功能和表型数据。
18. **Gene Ontology基因本体**（GO）是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表来全面描述生物体中基因和基因产物的属性。
19. **GO**主要包含三个方面的内容：
	* 1. **分子功能**：描述基因产物（如蛋白质）的生物化学功能，例如催化反应、结合其他分子等。
		2. **细胞组分**：描述基因产物所在的细胞结构，例如细胞膜、细胞核等。
		3. **生物过程**：描述基因参与的生物学过程，例如细胞分裂、信号转导等。
20. GO是一个有向无环图（DAG）结构，每个条目都与其它的同域或者不同域的条目定义了关系。GO的主要应用包括：基因功能注释、基因组分析和生物学研究等。
21. KEGG（Kyoto Encyclopedia of Genes and Genomes）分析是一种常用于系统生物学的分析方法，它主要用于研究基因和蛋白质参与的代谢途径和信号传导途径。KEGG是一个综合数据库资源，提供了生物化学途径、疾病信息、药物信息、生物系统的组件等数据。通过KEGG分析，研究人员可以将实验得到的基因或蛋白质数据与KEGG数据库中的生物途径进行比较，从而识别出在特定生物学条件或疾病状态下变化的关键途径，进一步理解基因或蛋白质在细胞生物学过程中的作用。
22. 序列联配（Sequence Alignment）：也称为序列比对，是生物信息学中一种基本且关键的分析方法。它涉及到将两个或多个生物序列（DNA、RNA或蛋白质）按照某种方式排列起来，以便最大限度地匹配相同或相似的序列区域。
23. **序列相似性（Similarity）**：指两个或多个生物序列（DNA、RNA、蛋白质）在结构、功能或进化上的相似度。通常通过比对算法计算得到，用百分比表示。
24. **序列一致性（Identity）**：指两个或多个序列在相同位置上具有完全相同的核苷酸或氨基酸残基的程度。一致性是相似性的一个子集，通常也用百分比表示。
25. **计算两条序列的一致性和相似性。**
26. 计分矩阵（Scoring Matrix）：又称打分矩阵、替换矩阵（Substitution Matrix），是一种用于衡量两个序列相似程度的工具。它是一个二维矩阵，其中每个位置的值代表一对氨基酸或核苷酸的匹配或错配得分。
27. 计分矩阵构建的一般原则：
	* 1. 匹配得分：两个序列中相同碱基或氨基酸的得分。一般来说，在序列比对中，匹配得分最高。
		2. 错配得（罚）分：两条序列中不同碱基或氨基酸匹配的得（罚）分。一般来说，错配得分越高（罚分越低），表明两个序列越相似。对氨基酸而言，要特别注意氨基酸性质的差异对得（罚）分的影响，一般性质相似的氨基酸之间的配对得分较高，性质差异较大的氨基酸之间的配对得分较低。
		3. 空位罚分：序列中插入或删除空位的得分。一般来说，空位得分应该为负数，以惩罚空位的插入或删除，且空位设置（开gap）的罚分一般大于空位扩展（gap延伸）罚值。如何设定罚分目前没有明确的理论，但大的空位设置罚分和小的空位延伸罚分被普遍接受。
28. PAM计分矩阵：Dayhoff等人通过对蛋白质演化模式的研究，基于所选择的蛋白序列中共观测到1572次替换事件，建立了一组被广泛应用的氨基酸计分矩阵，常被称为PAM（Point Accepted Mutation）矩阵或Dayhoff矩阵。其是一种用于量化蛋白质序列比对时氨基酸替换的概率矩阵。它们是在考虑进化距离的基础上构建的，即一个点接受突变的距离。PAM单位用于描述序列发生一次可接受突变的平均比例。
29. PAM1计分矩阵：PAM1矩阵是PAM（Point Accepted Mutation）矩阵家族中最基础的矩阵，用来表示蛋白质序列在1%的进化距离下，每种氨基酸变为另一种氨基酸的概率。这种1%的进化距离意味着在100个氨基酸中，预期会有1个氨基酸发生改变。其适用于进化距离较近的蛋白质序列比对。
30. PAM250计分矩阵：PAM250矩阵是PAM矩阵家族的一员，是将PAM1矩阵自身相乘250次得到的，代表250%的氨基酸替换。其推导基于以下假设：一是氨基酸序列的进化是通过一系列随机的点突变发生的；二是每一次点突变发生的概率是相同的。它适用于进化距离较远的蛋白质序列比对。
31. BLOSUM计分矩阵：BLOSUM矩阵（Blocks Substitution Matrix）是由Steven Henikoff等人于1992年提出的一种用于衡量蛋白质序列之间相似性的打分矩阵。他们直接利用多序列联配分析亲缘关系较远的而不是PAM矩阵中所用的亲缘关系较近的蛋白质序列。在一定程度上避免了类似PAM矩阵这种基于随机序列进化的模型构建的，并未充分考虑蛋白质序列的结构和功能信息的问题。BLOSUM矩阵的优点是较符合实际观测效果，不足之处是它不能与演化挂钩。
32. BLOSUM62计分矩阵：BLOSUM62矩阵是基于Henikoff等人于1992年构建的BLOSUM矩阵系列中的第一个矩阵。它基于至少62%序列相似度的区块中氨基酸替换的概率。它是使用最广泛的BLOSUM矩阵，适用于进化距离较近的蛋白质序列比对。矩阵中大于0的匹配为相同或相似氨基酸匹配。
33. PAM矩阵与BLOSUM矩阵的关系：
	* 1. PAM矩阵：后面的数字代表了进化时间。例如，PAM250矩阵代表了进化时间为250个单位的氨基酸替换可能性。PAM矩阵假设氨基酸替换的概率与进化时间成正比，因此数字越大，代表氨基酸替换的可能性越高。
		2. BLOSUM矩阵：后面的数字代表了氨基酸替换的百分比。例如，BLOSUM62矩阵代表了氨基酸替换可能性为62%的同源蛋白序列。BLOSUM矩阵是基于蛋白质家族的序列比对构建的，因此数字越大，代表氨基酸替换的可能性越高，说明两条序列在该位置的氨基酸越相似。
34. 总结PAM矩阵与BLOSUM矩阵的比较：



1. 为什么序列比对选择动态规划算法：具体来说，一对比对只有两种方式，或残基间相互对应，或一个残基对应空位，根据所选择的打分矩阵和gap罚分，可以得到一对比对的分数，而最终的比对分数即为各个比对得分之和。根据动态规划算法，我们可以推出，当前最好的比对，就是之前最好的比对加上当前位置最好的比对。
2. 联配矩阵：在双序列比对中，每个序列的每个部分都可能与另一个序列的部分对应，因此产生大量可能的对齐方式。矩阵提供了一种系统的方法来探索这些可能性，并找到最佳解。人们将两条用于联配的序列分别放置在矩阵的第一行和第一列，通过对矩阵的填充，代表序列局部或全局联所有的可能性。



1. **全局比对之Needleman-Wunsch算法**：Needleman-Wunsch算法是一种在生物信息学中广泛应用的算法，特别是在序列比对中。它是由Saul B. Needleman和Christian D. Wunsch在1970年提出的。该算法使用动态规划技术来寻找两个序列（通常是蛋白质或核酸）之间的最佳全局对齐。全局对齐指的是将两个序列从头到尾进行对齐，以便最大化它们之间的相似性。该算法的目的是确定序列间的相似性程度，这对于理解生物分子的功能和进化关系非常重要。
2. Needleman-Wunsch算法的主要步骤：
	* 1. 初始化：创建一个矩阵来表示两个序列之间所有可能的对齐方式。矩阵的行和列分别代表两个序列。矩阵的第一行和第一列用于初始化，代表与空序列的对齐，其值基于罚分来设定。这个罚分通常是负值，代表插入或删除操作的成本。
		2. 填充矩阵：算法填充矩阵的其余部分。每个单元格的值根据一定的得分规则计算得出，这些规则考虑了相邻单元格（上方、左侧和对角线上方）的值以及对应操作的罚分（匹配、不匹配、插入或删除）。通常，对于匹配，会给予正得分；对于不匹配或indels，会施加罚分。
		3. 追踪回溯：从矩阵的右下角开始，根据矩阵中的得分逆向追踪到左上角，以找到得分最高的路径。这条路径代表了两个序列之间的最佳全局对齐。追踪过程中，根据移动的方向（对角线、上方或左方），可以确定是匹配/不匹配还是插入/删除操作。
3. Needleman-Wunsch算法比对公式：



1. Needleman-Wunsch算法比对公式：



1. **Needleman-Wunsch比对例题1**：





1. **Needleman-Wunsch比对例题2**：





1. 从全局比对到局部比对：
	* 1. 全局比对（Global Alignment）：将两个序列的全部字符进行匹配，并尽可能地使匹配的字符数量最大化。全局比对通常用于比较具有相似长度和结构的序列。
		2. 局部比对（Local Alignment）：找到两个序列之间所有局部相似区域的匹配。局部比对通常用于比较具有不同长度或结构的序列。
2. 从全局比对到局部比对：



1. Smith–Waterman算法：是一种用于生物信息学中的序列比对，尤其是用于局部比对的算法。这种算法由 Temple F. Smith 和 Michael S. Waterman 在 1981 年提出。它基于动态规划算法，旨在找到两条序列中相似度最高的子序列，这使它特别适合于鉴定序列中的局部相似性或功能性区域。基本思想是将两个序列进行逐个比较，并根据匹配、错配和插入/缺失三种操作来计算每个位置的得分。最终，通过找到所有可能的路径中的最大得分路径，得到两个序列的最佳局部比对。其与Needleman-Wunsch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。
2. Smith–Waterman算法比对公式：



1. BLAST（Basic Local Alignment Search Tool）：生物信息学中最广泛使用的工具之一，主要用于序列比对。BLAST允许研究者快速比对核酸或蛋白质序列，以找到数据库中与给定序列相似的序列。这个工具由Stephen Altschul等人在1990年开发，目的是优化传统的Smith-Waterman算法，使得在处理大规模数据库时更高效。
2. 多序列联配（Multiple Sequence Alignment，MSA）：指将多个生物序列进行排列，以便显示出它们之间的相似性和差异性。多序列比对可以用于研究序列的进化关系、预测蛋白质结构和功能等。
3. 分子钟模型：是一种假设进化速率恒定的模型。该模型认为，每个谱系中的分子（例如DNA或蛋白质）的演化速率是恒定的。这意味着在给定时间内，所有谱系中的分子都会发生相同数量的突变。分子钟模型可以用于估计谱系的分化时间和进化速率。它通常用于构建系统发育树，并为化石记录提供时间校准。
4. 分子钟模型与系统发育树的关系：分子钟模型可以用于构建系统发育树，并为化石记录提供时间校准。具体来说，它可以用于：估计谱系的分化时间：通过比较不同谱系中同源分子的序列差异，可以估计这些谱系在进化过程中分化的年代。
5. 分子钟模型的局限性：分子钟模型是一种简化的模型，它有一些局限性。例如，它假设进化速率是恒定的，而实际上进化速率可能因谱系、时间和基因而异。此外，分子钟模型还受到突变饱和、测序误差和其他因素的影响。
6. 系统发育树：综合来看，系统发育树是一种模型，用于描述和推断生物实体之间基于共同祖先的演化关系。这些实体可以是物种，也可以是分子序列。系统发育树通过节点、分支和叶的图形结构来描绘演化过程中的分化和关系，不仅反映了物种或分子间的历史联系，也提供了进化动态的视角。此外，系统发育树还是理解生物多样性、指导分类学、推动保护生物学和进化研究的重要工具。通过这种方式，系统发育树桥接了分子生物学和系统生物学的视角，使得从基因到生态系统的各个层面的研究都能找到共同的理论基础和实证方法。
7. **系统发育树的构成**：系统发育树是一种图形表示，用来描绘物种、群体或基因之间基于进化关系的历史。这种图形表达的核心目的是展示这些实体如何从一个或多个共同祖先分化而来。其主要构成元素包括：
	* 1. **节点（Node）**：节点代表分支点，其中每个节点都表示一个演化分化事件。在树中，最底部的节点（根节点）通常代表最近的共同祖先。根据树是否有一个明确的起点，系统发育树可以是有根的（rooted）或无根的（unrooted）。有根树明确指出了演化的方向和时间顺序，而无根树则主要展示群体间的关系而不强调方向。
		2. **分支（Branch）**：分支连接节点，表示从一个演化事件到另一个演化事件的路径。分支的长度在很多系统发育树中有特定含义，比如可以代表时间的长短、遗传变异的数量或演化的速率。不同的研究可能使用不同的标准来定义分支长度。
		3. **叶（Leaf）**：树的末端通常称为叶或外节点，代表当前存在的物种、已知的群体或观察到的基因序列。在有根树中，这些叶节点通常排列在树的外围，显示出从共同祖先到现代物种的演化路径。
		4. **树的拓扑结构（Topology）**：这是指树的形状，也就是节点和分支如何连接的具体方式。拓扑结构决定了树的解读，不同的拓扑结构可能暗示着不同的演化假设。
		5. **树根（Root）**：在有根系统发育树中，树根是指位于树底部的节点，代表所有分析中的物种或群体的最近共同祖先。确定树的根通常需要外部信息，如化石记录或已知的演化事件。
8. 系统发育树的构建方法：距离法、最大简约法、最大似然法、贝叶斯法等。