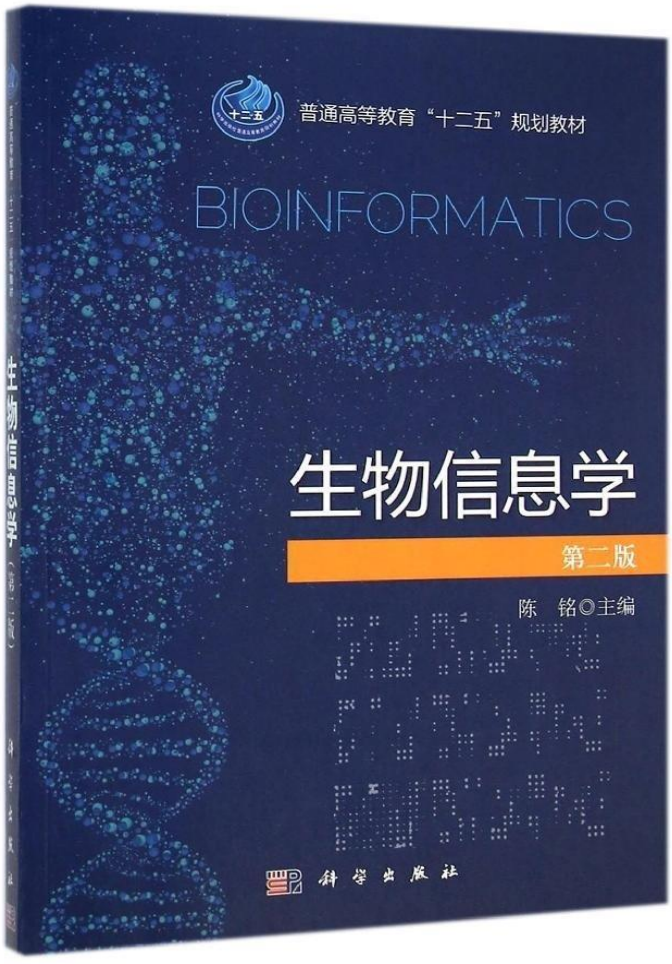
课程学习笔记

之

《生物信息学》第二版

（陈铭 科学出版社）



第一版

（如有改进或者建议，请发到邮箱 bixiaan@hnu.edu.cn）

2019 年 12 月 8 日

目　录

第一章 生物信息学的概念及其发展历史.....6

第一节 生物信息学的发展历史.....6

第二节 生物信息学的研究领域.....8

第三节 生物信息学的主要应用.....14

第四节 生物信息学面临的挑战.....24

第二章 生物学数据库及其检索 24

第一节 生物学数据库简介.....24

第二节 生物学数据库的内容与结构.....25

第三节 生物学数据库的检索.....27

第三章 序列比对原理.....33

第一节 序列比对相关概念.....33

第二节 序列比对打分方法.....35

第三节 序列比对算法.....38

第四节 序列比对工具.....39

第五节 多序列比对.....39

第四章 蛋白质结构预测与分析 40

第一节 蛋白质结构组织层次.....40

第二节 蛋白质的测定与理论预测.....47

第三节 蛋白质对接.....49

第五章 真核生物基因组的注释 53

第一节 蛋白质编码基因的注释.....53

第二节 RNA 基因的注释 57
第三节 重复序列的注释 58
第四节 假基因的注释 59
第五节 案例分析：黄瓜基因组的注释 59
第六章 蛋白质组学 61
第一节 蛋白质组学概述 62
第二节 蛋白质的大规模分离鉴定技术 64
第三节 蛋白质的翻译后修饰 69
第四节 蛋白质分选 69
第五节 蛋白质相互作用 71
第七章 系统生物学 72
第一节 系统生物学基本概念 72
第二节 系统生物学基本技术与方法 72
第三节 先进的成像技术 73
第四节 基因表达调控网络 73
第五节 代谢网络 74
第六节 信号传导途径 75
第七节 蛋白质-蛋白质相互作用网络 76
第八节 虚拟细胞 76
第八章 合成生物学 76
第一节 合成生物学概述 76
第二节 合成生物学基础研究经典实例 78

第三节 合成生物学应用研究经典实例.....80
第九章 分子进化与系统发育.....80
第一节 分子进化与系统发育.....80
第二节 分子系统发育树的构建方法.....81
第三节 系统发育树构建及应用.....81
第十章 统计学习与推理.....81
第一节 统计学习与推理基础.....81
第二节 统计模型与参数推断.....81
第三节 聚类分析、主成分分析与 Fisher 判别.....82
第四节 贝叶斯推理.....82
第五节 隐马尔科夫模型.....82
第六节 动态神经网络.....82
第七节 支持向量机.....82
第八节 MATLAB 的应用实例.....82
第十一章 生物信息学编程基础.....82
第一节 Linux 操作系统.....82
第二节 生物信息学中的编程语言.....82
第三节 SQL 及数据库编程.....82
第四节 并行计算.....82
第十二章 第二代测序技术及其应用.....83
第一节 测序技术概述.....83
第二节 第二代测序原理.....85

第三节 第二代测序技术的应用.....86

第四节 生物信息学在第二代测序中的应用.....87

第一章 生物信息学的概念及其发展历史

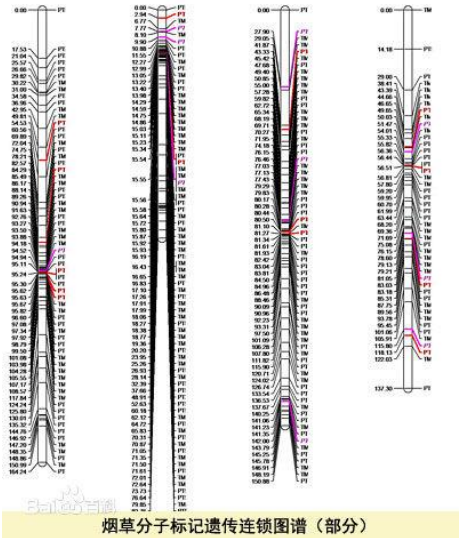
第一节 生物信息学的发展历史

1、人类基因组计划（P1）

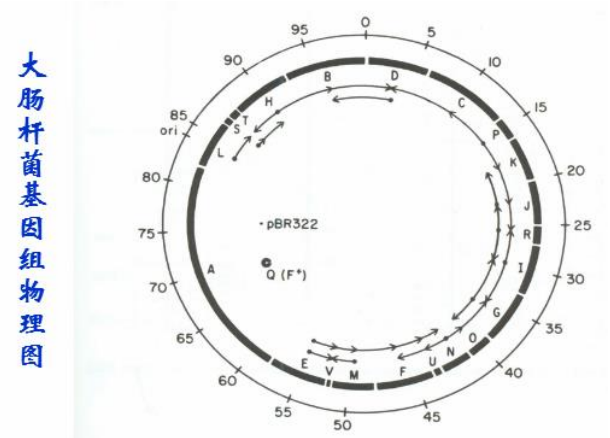
人类基因组计划（human genome project， HGP）是由美国科学家于 1985 年率先提出，于 1990 年正式启动的。其宗旨在于测定组成人类染色体（指单倍体）中所包含的 30 亿个碱基对组成的核苷酸序列，从而绘制人类基因组图谱，并且辨识其载有的基因及其序列，达到破译人类遗传信息的最终目的。人类基因组计划与曼哈顿原子弹计划和阿波罗计划并称为三大科学计划。人类基因组计划被誉为生命科学的“登月计划”。

基因组图谱：

定义：展示一种生物全基因组结构的图谱。按建立图谱的研究目的方法和精细程度，可以有不同的形式，包括以遗传学方法建立的遗传连锁图谱，按距离绘出基因位置分布的物理图谱，经测定核酸序列建立的核苷酸序列图谱，以及标记出可表达序列的转录图谱等。



烟草分子标记遗传连锁图谱（部分）



2、DNA 序列（P1）

DNA 序列或基因序列是使用一串字母表示的真实的或者假设的携带基因信息的 DNA 分子的一级结构。

例如，序列 AAAGTCTGAC。任意长度大于 4 的一串核苷酸被称做一个序列。（A：腺嘌呤，C：胞嘧啶，G：鸟嘌呤和 T：胸腺嘧啶，分别代表组成 DNA 的四种核苷酸）

3、GenBank、EMBC、DDBJ（P1）

GenBank, 美国国家生物技术信息中心建立的 DNA 序列数据库，从公共资源中获取序列数据，主要是科研人员直接提供或来源于大规模基因组测序计划。

EMBC，欧洲分子生物学会议，英文全称为 European Molecular Biology Conference。该组织于 1964 年 2 月在日内瓦成立，目的是促进欧洲分子生物学的发展和开展年轻科学家的培训工作。

DBJ，日本 DNA 数据库，英文全称为 DNA Data Bank of Japan，于 1984 年建立，是世界三大 DNA 数据库之一，与 NCBI 的 GenBank，EMBL 的 EBI 数据库共同组成国际 DNA 数据库。

4、碱基、碱基对、核酸、核苷酸（P1）

碱基：碱基，又称核碱基、含氮碱基，是形成核苷的含氮化合物，核苷又是核苷酸的组分。碱基、核苷和核苷酸等单体构成了核酸的基

本构件。

碱基对：碱基对是一对相互匹配的碱基（即 A—T， G—C，A—U 相互作用）被氢键连接起来

核酸：核酸是一类生物聚合物，是所有已知生命形式必不可少的组成物质。核酸是脱氧核糖核酸（DNA）和核糖核酸（RNA）的总称。

核苷酸：核苷酸（hé gān suān） Nucleotide，一类由嘌呤碱或嘧啶碱、核糖或脱氧核糖以及磷酸三种物质组成的化合物。又称核甙酸。戊糖与有机碱合成核苷，核苷与磷酸合成核苷酸，4 种核苷酸组成核酸。核苷酸主要参与构成核酸，许多单核苷酸也具有多种重要的生物学功能，如与能量代谢有关的三磷酸腺苷（ATP）、脱氢辅酶等。

5、核苷酸序列（P1）

核苷酸序列：核苷酸序列，就是指 DNA 或 RNA 中碱基的排列顺序。这意味着 DNA 中 A，T，G，C 的排列循序，或者 mRNA 中 A，U，G，C 的排列循序，也包括 rRNA，tRNA 中碱基的排列顺序。

第二节 生物信息学的研究领域

1、后基因组时代（P4）

后基因组时代指的就是生物界，特别是从事分子生物学的科学工作者，将研究方向由结构基因组方向向功能基因组方向转移的时代趋势。

2、表型、表型相关基因（P5）

丹麦遗传学家 W.L.约翰森于 1911 年提出的两个遗传学名词，有基因型和表型。

基因型，是指某一生物个体全部基因组合的总称。

表型，是具有特定基因型的个体，在一定环境条件下，所表现出来的性状特征的总和。简单来说，指个体形态、功能等各方面的表现，如身高、肤色、血型、酶活力、药物耐受力乃至性格等等。表型是基因型和环境共同作用的结果。

性状，是指生物体所有特征的总和，由基因决定，必须是可以遗传的。而表型则是这些基因决定的性状在环境作用下的具体表现，与性状的概念有着本质区别，表型是不可遗传的。

表型和性状的区别：性状由基因决定，而表型是基因和环境共同作用的结果。比如 A 和 B 生的孩子 C 有身高为 175cm 的性状，但可能 C 由于外界环境的影响，比如 C 有喜欢打篮球之类能够增高的行为，那么他可能长到 180cm，这就叫表型，然而他遗传到下一代的性状还是 175cm。

3、细胞核、染色体（P5）

细胞核：真核细胞内最大、最重要的细胞结构，是细胞遗传与代谢的调控中心，是真核细胞区别于原核细胞最显著的标志之一（极少数真核细胞无细胞核，如哺乳动物的成熟的红细胞，高等植物成熟的筛管细胞等）。它主要由核膜（nuclear membrane）、染色质

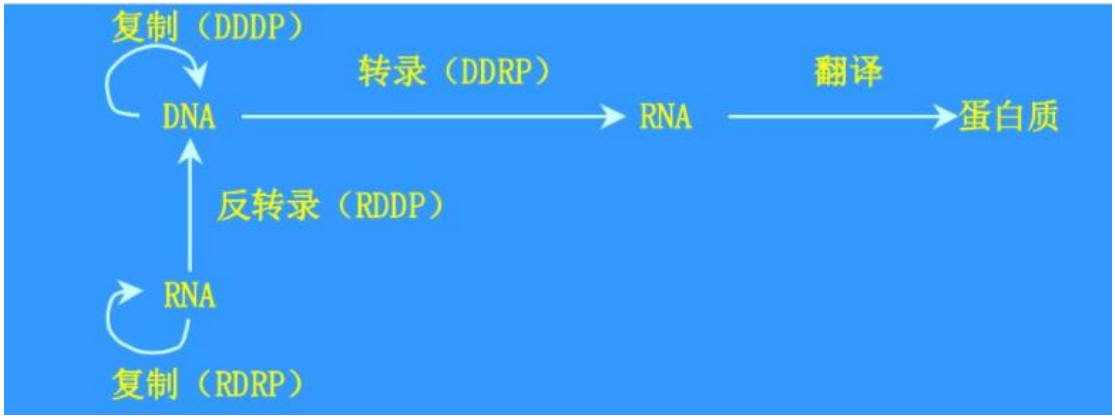
（chromatin）、核仁(nucleolus)、核基质（nuclear matrix） 等组成。

染色体：细胞在有丝分裂或减数分裂时 DNA 存在的特定形式。

4、DNA-RNA-蛋白质（P5）

DNA-RNA-蛋白质：一句话讲，DNA 决定 RNA，RNA 决定蛋白质

三句话讲：DNA 是一切生物的遗传物质的基础，DNA 经过转录称为 RNA，RNA 经过翻译称为蛋白质。



转录：转录 (Transcription) 是遗传信息从 DNA 流向 RNA 的过程。即以双链 DNA 中的确定的一条链（模板链用于转录，编码链不用于转录）为模板，以 A,U,C,G 四种核苷酸为原料，在 RNA 聚合酶催化下合成 RNA 的过程。作为蛋白质生物合成的第一步，进行转录时，一个基因会被读取并被复制为 mRNA，即特定的 DNA 片断作为遗传信息模板，以依赖 DNA 的 RNA 聚合酶作为催化剂，通过碱基互补的原则合成前体 mRNA。RNA 聚合酶通过与一系列组分构成动态复合体，完成转录起始、延伸、终止等过程。生成的 mRNA 携有的密码子，进入核糖体后可以实现蛋白质的合成。转录仅以 DNA 的一条链作为模

板，被选为模板的单链称为模板链，亦称无义链；另一条单链称为非模板链，即编码链，因编码链与转录生成的 RNA 序列 T 变为 U 外其他序列一致，所以又称有义链。DNA 上的转录区域称为转录单位。

翻译：翻译是蛋白质生物合成（基因表达中的一部分，基因表达还包括转录）过程中的第二步（转录为第一步），翻译是根据遗传密码的中心法则，将成熟的信使 RNA 分子（由 DNA 通过转录而生成）中“碱基的排列顺序”（核苷酸序列）解码，并生成对应的特定氨基酸序列的过程。但也有许多转录生成的 RNA，如转运 RNA（tRNA）、核糖体 RNA（rRNA）和小核 RNA（snRNA）等并不被翻译为氨基酸序列。

5、编码区、非编码区（P5）

基因在结构上，分为编码区和非编码区两部分（其中非编码区对基因的表达主要起调控作用，如启动子等位于该区）。真核生物基因的编码区是不连续的，分为外显子和内含子（其中外显子是可以最终实现表达（表现在蛋白质的一级结构上），内含子则最终不能表达（所以真核生物基因表达过程中，转录产物——信使 RNA 不能直接进行翻译，而是要修剪掉内含子部分后才能去指导翻译）。原核生物的基因是连续的，所以谈不上外显子、内含子的区分。

6、神经递质、神经受体（P5）

神经递质：神经递质（英文 neurotransmitter）在突触传递中是担

当“信使”的特定化学物质，简称递质。随着神经生物学的发展，陆续在神经系统中发现了大量神经活性物质。

神经受体：无神经受体，有神经递质受体。

神经递质受体：在化学突触传递中担当信使的特定化学物质，简称递质。神经递质受体都是一些跨越细胞膜的蛋白质复合体。

7、药物作用靶（P5）

无药物作用靶，有靶点。

靶点， 医学上进行某些放射治疗时，放射线从不同方位照射，汇集病变部位，这个病变部位叫做靶点。

8、分子生物学（P4）、细胞生物学（P5）、神经生物学（P5）、系统生物学（P8）、网络生物学、进化生物学（P5）、计算生物学（P7）

分子生物学 :分子生物学是生命科学发展过程中诞生的一门实验性极强的新兴学科。美国著名分子生物学家 Robert F.Weaver 遵循这一学科发展的特点，于 1999 年出版了 Molecular Biology 一书。《分子生物学（第五版）》以原始研究论文为基础，通过对实验的设计、对结果的分析而逐步展开对分子生物学理论的讲述，文字通俗流畅，叙述由浅入深。随着学科的迅速发展，几经修订再版的《分子生物学》第五版共有导论，分子生物学方法，原核生物的转录，真核生物的转录，转录后加工，翻译，DNA 复制、重组和转座，以及基因组等 8 个部分共 25 章的内容，书后还附有术语表。每一章节都以提出科学问

题、展开研究过程开始，以提供思考习题、推荐阅读文献结束。

细胞生物学：细胞生物学(Cell Biology)是在显微、亚显微和分子水平三个层次上，研究细胞的结构、功能和各种生命规律的一门科学。细胞生物学由 Cytology 发展而来，Cytology 是关于细胞结构与功能(特别是染色体)的研究。现代细胞生物学从显微水平、超微水平和分子水平等不同层次研究细胞的结构、功能及生命活动。在我国基础学科发展规划中，细胞生物学与分子生物学、神经生物学和生态学并列为生命科学的四大基础学科

神经生物学：神经生物学，21 世纪的明星学科。神经生物学是生物学中研究神经系统的解剖，生理， 神经生物学病理方面内容的一个分支。从上个世纪 90 年代以来，世界科研强国加快了对神经生物学研究的投入。神经科学是专门研究神经系统的结构、功能、发育、遗传学、生物化学、生理学、药理学及病理学的一门科学。对行为及学习的研究都是神经科学的分支。

系统生物学：研究生物系统组成成分的构成与相互关系的结构、动态与发生，以系统论和实验、计算方法整合研究为特征的生物学。系统生物学不同于以往仅仅关心个别的基因和蛋白质的分子生物学，在于研究细胞信号传导和基因调控网路、生物系统组成之间相互关系的结构和系统功能的涌现。

进化生物学：进化生物学（evolutionary biology） 进化生物学是生物学最基本的理论之一，由查尔斯·达尔文提出的进化论。他第一次提出自然选择是进化的一个机制。 研究生物进化过程和生物群落特征

的综合学科，包括生态学、行为学和分类学等。进化生物学，包括进化遗传学、进化基因组学（evolutional genomics），进化论是实验生物学的理论依据，基于系统论的生物系统泛进化论建立系统生物学的理论基础，遗传学与基因组研究是进化生物学的方法体系，尤其是系统遗传学开创了对基因组自组织进化与生物体发育自组织化及其相互关系的细胞发生系统动力学研究，为进化生物学的发展开拓了新的途径。

计算生物学：指开发和应用数据分析及理论的方法、数学建模、计算机仿真技术等。计算生物学的最终目的不仅仅局限于测序，而是运用计算机的思维解决生物问题，用计算机的语言和数学的逻辑构建和描述并模拟出生物世界。相对于生物信息学，计算生物学的层次更高。虽然两者之间界限模糊，但生物信息学略微偏向于生物而计算生物学略微偏向计算机。生物信息学侧重于数据的提取、挖掘，而计算生物学侧重对数据的处理、运用。计算生物学的最终目的不只局限于测序，而是运用计算机的思维解决生物问题，用计算机的语言和数学的逻辑构建和描述并模拟出生物世界。

第三节 生物信息学的主要应用

1、EMBL（P6）

EMBL,欧洲分子生物学实验室 EMBL（European Molecular Biology Laboratory）1974 年由欧洲 14 个国家加上亚洲的以色列共同发起建立，现在由欧洲 30 个成员国政府支持组成，目的在于促进欧洲国家

之间的合作来发展分子生物学的基础研究和改进仪器设备、教育工作等。EMBL 分为 7 个部分：结构、分化、物理仪器、生化仪器、生物仪器、计算机和应用数学。它的宗旨是：从事结构分子生物学及分子医学方面的基础研究；为科学家、学生及访问学者提供高层次的培训；为成员国的科学家提供必需的科研服务；在生命科学领域开发新型的科研仪器及研究方法；积极参与生物技术的转化及应用。

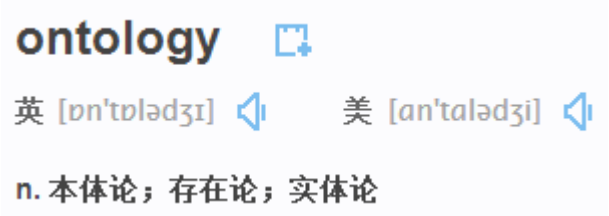
2、KEGG（P6）

KEGG，京都基因与基因组百科全书，英文全称为 Kyoto Encyclopedia of Genes and Genomes。KEGG 由日本京都大学生物信息学中心的 Kanehisa 实验室于 1995 年建立，是国际最常用的生物信息数据库之一，以“理解生物系统的高级功能和实用程序资源库”著称。

3、联合数据库系统（P6）

联合数据库系统：许多组织的计算资源，包括在部门或分部作为独立系统存在的局域网(LAN)、小型计算机和大型计算机，这些系统有各自的数据库，当建造一个企业计算平台来互连一个组织时，组织内分散的异构数据库系统就应结合成联合系统以供多个用户访问数据。

4、Ontology（P6）



定义：

Ontology 这个词来自于哲学，称为：本体论、实体论或存在论，在哲学中指的是对客观存在的系统的解释和说明。在 AI 领域，Neches 等人将 Ontology 解释为“Ontology 定义了包含相关领域词汇的基本术语和关系，以及组合这些术语和关系定义词汇外延的规则”。B. Chandrasekaran 等人的观点和 Neches 类似，认为“Ontology 属于人工智能领域中的内容理论，它研究特定领域指示的对象、对象属性和对象间的关系，它为领域知识的描述提供术语”。

基本思想：

先建立相关领域的 Ontology，根据 Ontology 对收集的信息进行标注，用户的检索请求按照 Ontology 转换成规定的格式，在 Ontology 的帮助下匹配符合条件的数据集合返回给用户。

功能：

主要是实现某种程度的知识共享和重用，它能使得计算机对信息和对语言的理解上升到语义层次。因此，Ontology 在一些涉及到信息的互操作、知识理解等方面的领域具有很大的应用前景。

5、鸟枪法（P6）

鸟枪法是将目的 DNA 随机地处理成大小不同的片段，再将这些片段的序列连接起来的测序方法，属于第一代测序技术。

“鸟枪法”是将基因组按染色体分开后，将它全部打乱，切成碎片，进行随机测序，测序后再将它们拼接起来。它的好处是速度快，在很

短的时间里就能拿到生物细胞基因组。该法的思路独特，好像树林里停了一大群鸟，很多人乱枪射击，在很短的时间内，就可以将林子中的大部分鸟打中。“鸟枪法”又有点类似人们玩的拼图游戏。拼图游戏是将一个完整的画面分成杂乱无章的碎块,然后重新拼装复原。而“鸟枪法”则是先将整个基因组打乱，切成随机碎片，然后测定每个小片段序列，最终利用计算机对这些切片进行排序和组装，并确定它们在基因组中的正确位置。

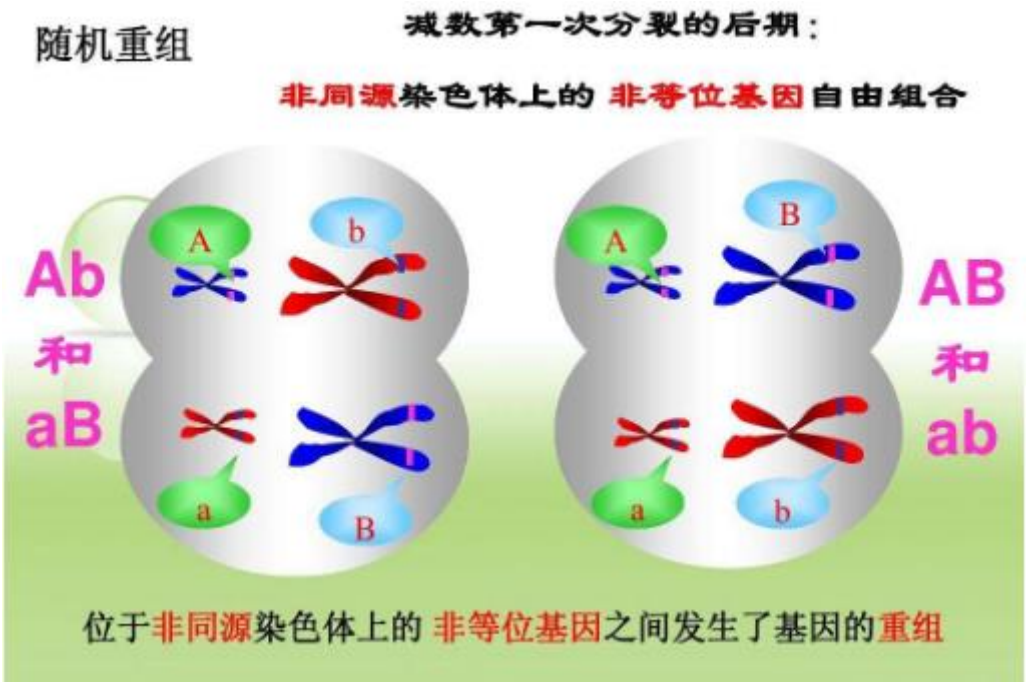
6、PSI-BLAST（P6）

BLAST :Bell Labs Layered Space-Time，即贝尔实验室分层空时，是贝尔实验室提出的一种能够达到极高数据传输速率的无线通信技术。这种技术在一个带限无线信道内，通过多天线技术充分利用空间复用，使得信道容量随着天线数的增加呈线性增长，有着极高的频谱效率。

PSI-BLAST: Position-Specific Iterated BALST，即特殊位置重复BLAST。

7、基因重组算法（P6）

基因重组，是指在生物体进行有性生殖的过程中，控制不同性状的基因重新组合。主要发生在减数第一次分裂前期的交叉互换和后期的非同源染色体自由组合。



基因重组算法的定义没有找到。

非同源染色体：同源染色体指减数分裂时发生配对的染色体。基因数目一样，而同样的基因或等位基因又是以同样顺序排列着的称为完全同源的染色体。部分相同的染色体则称为部分同源的染色体（partially homologous chro-mosomes）。部分同源的染色体只是在同源部分能够配对，非同源的部分是分离的，所以常形成一端接在一起的二价染色体。二倍体是具有二个染色体组，分别来自双亲的配子。多倍体具有二个以上的染色体组。自然的多倍体多具有部分同源染色体。

减数第一次分裂：减数分裂是生物细胞中染色体数目减半的分裂方式。生殖细胞分裂时，染色体只复制一次，细胞连续分裂两次，这是染色体数目减半的一种特殊分裂方式。减数分裂不仅是保证物种染色体数目稳定的机制，同时也是物种适应环境变化不断进化的机制。减数分裂的结果是：成熟生殖细胞中的染色体数目比原始生殖细胞的

减少一半。减数分裂（Meiosis） 范围是进行有性生殖的生物；时期是从原始生殖细胞发展到成熟生殖细胞。

8、基因注释（P7）

基因组测序只能测出整个 DNA 的碱基对排列顺序，不能直接测出 DNA 上的基因及其功能，必须通过生物信息学方法，结合蛋白组学、转录组学，对测出来的序列进行分析，将基因及其功能加以挖掘，这称作基因注释。

9、基因组（P7）

基因组是指细胞内所有遗传信息，这种遗传信息以核苷酸序列形式存储。细胞或生物体中，一套完整单体的遗传物质的总和即为基因组。

10、基因组学（P7）

基因组学阐明整个基因组的结构、结构与功能的关系以及基因之间相互作用的科学。

换言之，基因组学是以分子生物学技术、电子计算机技术和信息网络技术为手段，以生物体内基因组的全部基因为研究对象，从整体水平上探索全基因组在生命活动的作用及其内在规律和内外环境影响机制的科学。

11、比较基因组学（P7）

比较基因组学是基于基因组图谱和测序基础上，对已知的基因和基因组结构进行比较，来了解基因的功能、表达机理和物种进化的学科。

12、全基因组（P7）

全基因组和基因组是一个意思，只不过针对的问题有些不同，全基因组一般是说测序测全部的基因，因而一般都会说要测全基因组；而基因组一般只是说大小多少，几倍体，跟要测的内容没多大关系。

13、全基因组学（P7）

全基因组学，是以生物体内所有基因组的全部基因为研究对象，从整体水平而不是单个基因水平上探索基因组在生命活动的作用及其内在规律和内外环境影响机制的科学。

14、微阵列（P7）

微阵列，也叫寡核苷酸阵列，是生物芯片中的一种。微阵列技术是人类基因组计划的逐步实施和分子生物学的迅猛发展及运用的产物，是第三次革命（基因组革命）中的主要技术之一。

15、生物芯片（P7）

生物芯片，又称蛋白芯片或基因芯片，所谓生物芯片一般指高密

度固定在互相支持介质上的生物信息分子（如基因片段、DNA 片段或多肽、蛋白质、糖分子、组织等）的微阵列杂交型芯片，阵列中每个分子的序列及位置都是已知的，并且是预先设定好的序列点阵。

简单说，生物芯片就是在一块玻璃片、硅片、尼龙膜等材料上放上生物样品，然后由一种仪器收集信号，用计算机分析数据结果。人们可能很容易把生物芯片与电子芯片联系起来。事实上，两者确有一个最基本的共同点：在微小尺寸上具有海量的数据信息。但它们是完全不同的两种东西，电子芯片上布列的是一个个半导体电子单元，而生物芯片上布列的是一个个生物探针分子。

16、表达谱（P7）

表达谱有两种，分别是基因表达谱和蛋白质表达谱。

以基因表达谱为例，基因表达谱是指细胞中所有基因表达的格局。通过比较分析基因表达谱，可从整体水平研究代谢机制，认识基因相互作用的网络关系，发现重要基因。

17、高通量技术（P7）

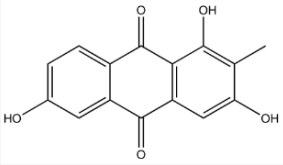
只有高通量测序技术，无高通量技术。

高通量测序技术（High-throughput sequencing）又称“下一代”测序技术（"Next-generation" sequencing technology），以能一次并行对几十万到几百万条 DNA 分子进行序列测定和一般读长较短等为标志。

18、质谱数据（P7）

质谱，又叫质谱法，是一种与光谱并列的谱学方法，通常意义上是指广泛应用于各个学科领域中通过制备、分离、检测气相离子来鉴定化合物的一种专门技术。通过质谱法能够得到丰富的结构信息。

质谱数据就是通过质谱法得到的数据。如下所示：

质谱详细信息	
化合物中文名：	1，3，6-三羟基-2-甲基蒽醌
化合物英文名：	6-Hydroxyrubiadin
化合物分子式：	$C_{15}H_{10}O_5$
化合物分子量：	270.0528
化合物紫外(UV)：	-
正离子碎片：	215,187,159,131,115
负离子碎片：	
化学结构式：	

19、表达分析（P7）

生物学中的表达是指细胞在生命过程中，把储存在 DNA 序列中的遗传信息，经过转录和翻译，转变成具有生物活性的蛋白质分子的过程。

表达分析就是对这种表达进行分析。

20、蛋白组（P7）

蛋白组是指一种生物或一个细胞里单套所有蛋白质品种的总和。

21、转录组（P7）

转录组广义上指某一生理条件下，细胞内所有转录产物的集合，

包括信使 RNA、核糖体 RNA、转运 RNA 及非编码 RNA；狭义上指所有信使 RNA 的集合。

22、代谢网络（P8）

有关细胞的代谢网络可以分不同层次来讨论：基因组（DNA 层次）、代谢途径及生化反应网络（蛋白质层次）、代谢流（物流层次）、代谢生理（微生物细胞层次）等等。以蛋白质层次的代谢网络来说，一个代谢物分子就是一个节点，而节点之间的连结则是生化反应。

23、SBML（P9）

SBML，系统生物学标记语言，英文全称为 Systems Biology Markup Language。主要功能是：SBML 可以编码由生化分子及其相互反应而成的生化反应网络组成的模型。

24、虚拟细胞模型（P9）

虚拟细胞，亦称电子细胞。它是通过数学的计算和分析，对细胞的结构和功能进行分析、整合和应用，以模拟和再现细胞和生命的现象的一门新兴学科。因此，虚拟细胞亦称人工细胞或人工生命。

虚拟细胞的构建一般需要以下几个步骤：1、选择建模对象，即细胞的种类、虚拟的功能等；2、建立相关数据库；3、设计与建模；4、组装、测试和维护。

现在，国际上已有两个虚拟细胞问世，一个是日本的原核虚拟细

胞模型，一个是美国的真核虚拟细胞模型。

意义：它不仅可以模拟人体细胞结构和功能，阐明生命活动的反应和规律，了解疾病发病的过程和机理，进行疾病的辅助诊断和治疗；而且可以用于药物设计和虚拟试验，还可以用于能源的开发和利用，推动仿生科学的进展，应对内外界环境的变化，改变和完善人们的生存条件。

第四节 生物信息学面临的挑战

无。

第二章 生物学数据库及其检索

第一节 生物学数据库简介

1、平面文件（P11）

平面文件是去除了所有特定应用（程序）格式的电子记录，从而使数据元素可以迁移到其他的应用上进行处理。这种去除电子数据格式的模式可以避免因为硬件和专有软件的过时而导致数据丢失。

2、Swiss-Prot 数据库（P12）

Swiss-Prot 是经过注释的蛋白质序列数据库。数据库由蛋白质序列条目构成，每个条目包含蛋白质序列、引用文献信息、分类学信息、注释等。

3、PubMed 数据库（P12）

PubMed 是美国家医学图书馆(NLM)下属的国家生物技术信息中心(NCBI)开发的、基于 WWW 的查询系统，其内容包括：DNA 与蛋白质序列，基因图数据等。

4、PROSITE 数据库（P12）

PROSITE 数据库收集了生物学有显著意义的蛋白质位点和序列模式,并能根据这些位点和模式快速和可靠地鉴别一个未知功能的蛋白质序列应该属于哪一个蛋白质家族。

5、Ensembl 数据库（P12）

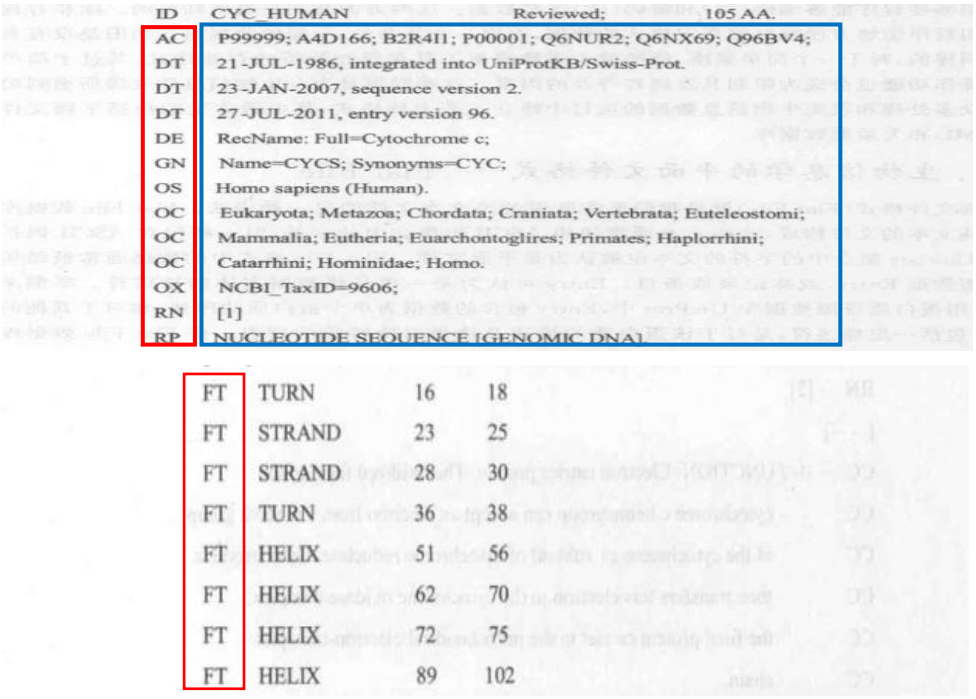
Ensembl 概念的核心是能够自动生成基因以及其他基因组数据与参考基因组比对的图形视图。

第二节 生物学数据库的内容与结构

1、Flat File 文件格式（P19）；

以纯文本的方式描述数据的一种数据格式。

在这种数据格式中，首先以 Entry 作为数据库的基本数据单位，例如蛋白质数据库中，包含若干种蛋白质，那么每一个蛋白质的所有信息都是作为一个 Entry。以下面的例子为例：



ID	CYC_HUMAN Reviewed; 105 AA.
AC	P99999; A4D166; B2R411; P00001; Q6NUR2; Q6NX69; Q96BV4;
DT	21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT	23-JAN-2007, sequence version 2.
DT	27-JUL-2011, entry version 96.
DE	RecName: Full=Cytochrome c;
GN	Name=CYCS; Synonyms=CYC;
OS	Homo sapiens (Human).
OC	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC	Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC	Catarrhini; Hominidae; Homo.
OX	NCBI_TaxID=9606;
RN	[1]
RP	NUCLEOTIDE SEQUENCE [GENOMIC DNA].

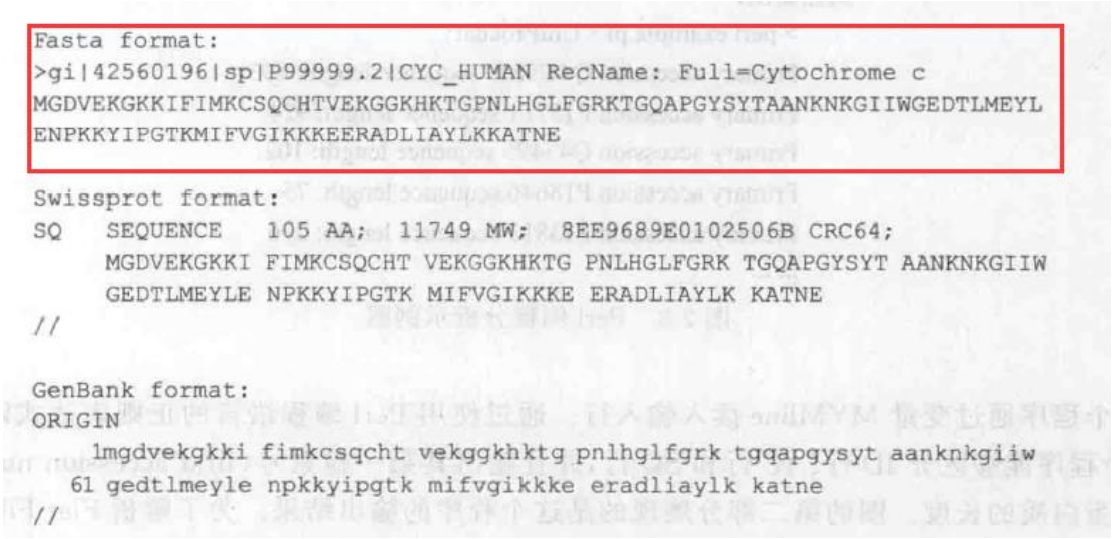
FT	TURN	16	18
FT	STRAND	23	25
FT	STRAND	28	30
FT	TURN	36	38
FT	HELIX	51	56
FT	HELIX	62	70
FT	HELIX	72	75
FT	HELIX	89	102

// （表示一个 Entry 信息的结束）

上图就是以人类细胞色素 c 作为一个 Entry(截图只展示了部分)，左边红色框的开头两个大写字母字段标识符表示人类细胞色素 c 的各个不同属性，如 RN 表示编号，RP 表示位置，右边蓝色框图的部分表示各个不同属性的具体描述。同时，一个属性可能具有多个子属性，就将该属性用的子属性用多行来表示，见图上第二个红框中部分。

2、序列格式（P24）

主要就是用一个字母来代表一个碱基，用碱基字母的顺序表示蛋白质的信息或其他生物信号的信息。以下图为例：



```
Fasta format:
>gi|42560196|sp|P99999.2|CYC_HUMAN RecName: Full=Cytochrome c
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYL
ENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

Swissprot format:
SQ   SEQUENCE   105 AA;  11749 MW;  8EE9689E0102506B CRC64;
     MGDVEKGKKI FIMKCSQCHT VEKGGKHKTG PNLHGLFGRK TGQAPGYSYT AANKNKGIIW
     GEDTLMEYLE NPKKYIPGTK MIFVGIKKKE ERADLIAYLK KATNE
//

GenBank format:
ORIGIN
     1mgdvekgkki fimkcsqcht vekggkhktg pnlhglfgrk tgqapgysyt aanknkgiiw
    61 gedtlmeyle npkkyipgtk mifvgikkke eradliaylk katne
//
```

红框中表示 P32234 蛋白的序列表示方法，第一行为序列格式申明，第二行为数据序列构建指令，余下部分碱基序列。

第三节 生物学数据库的检索

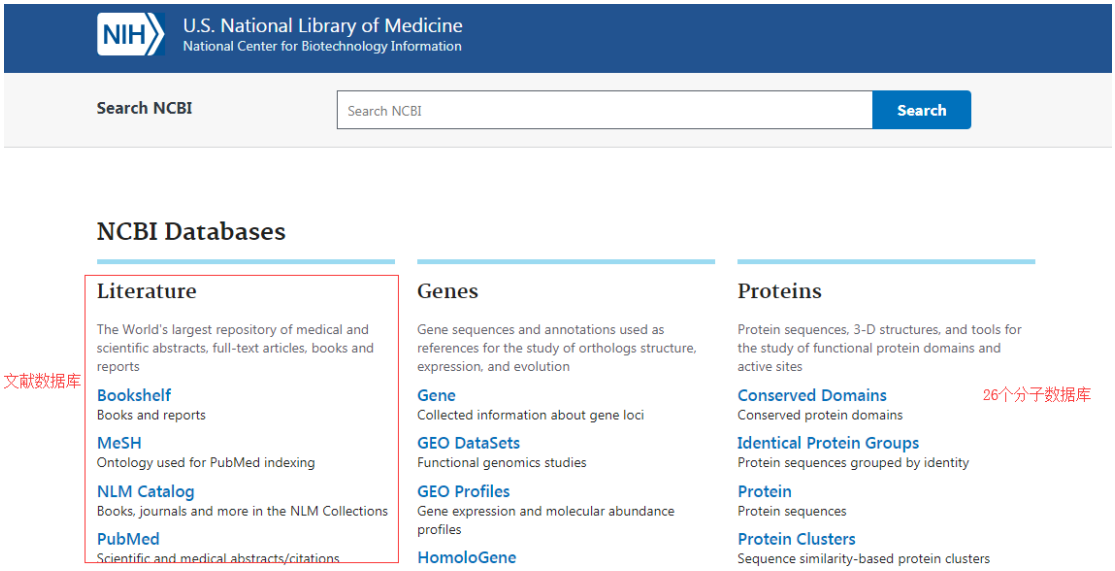
1、硬连接（P30）

硬连接是对于同一文件系统的一个文件的连接,节点是文件存在的唯一标识，在文件系统中是唯一的，但在不同的文件系统则有可能重复，硬连接就是记录了文件的地址，你在文件管理器中看到的每个文件都是磁盘里文件的“硬连接”。当一个文件的所有硬连接都被删除的时候，该文件占用的空间才可以被新加入的数据覆盖。

2、Entrez 检索（P31）

P31

（该主页已经改版，与书中描述有差异。）



全局搜索框如下：

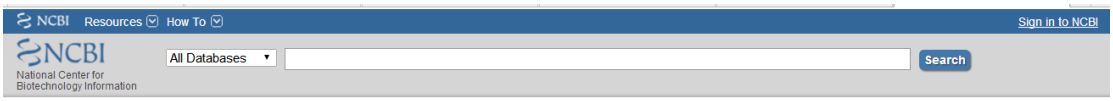


1、仍然可以输入一个简单的搜索词，进行搜索。

2、仍然可以使用布尔操作符将一个或多个搜索域连接起来，进行精确搜索。（AND：交集；OR：并集；NOT：差集。）

P32

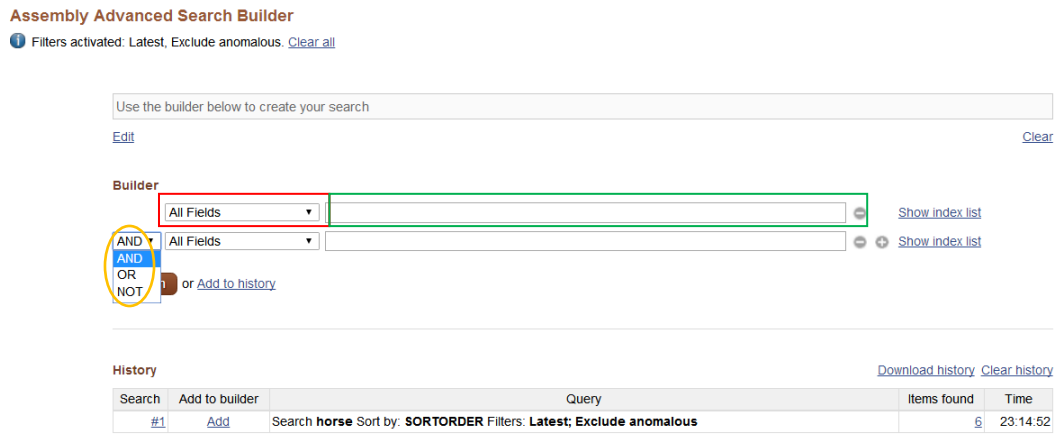
书上图 2-9 也已改版，网址为 https://www.ncbi.nlm.nih.gov/



方法与书上一样，即关键字+[特定的域]，如 horse[Organism]。

P35

复杂高精度查询语句建立



在红色框图中输入第一个主题的搜索域，然后在绿色框图中输入第一个主题，以此类推输入更多地搜索主题，然后在图中橙色框图中选择不同主题之间的关系，建立复杂高精度查询语句。

2、MeSH

《医学主题词表》(Medical Subject Headings，简称 MeSh)，它是一部规范化的可扩充的动态性叙词表，汇集约 18,000 多个医学主题词。

3、GI 号

GI 编号是 NCBI 网站的所有序列相关数据库的流水编号，其最有用的特征就是唯一性。对于每一条递交给 NCBI 的序列，都会赋给一个编号，而且这个编号对应的序列不可更改。这个编号对应这个唯一的一条序列，类似与我们用的身份证号。因此，利用 GI 在 NCBI 中查询时，你只要把数据库（蛋白质/核苷酸）选对，只要输入这个号码就可以把相应的序列调出来。

4、EBI 的 SRS 系统

4.1、快速文本检索

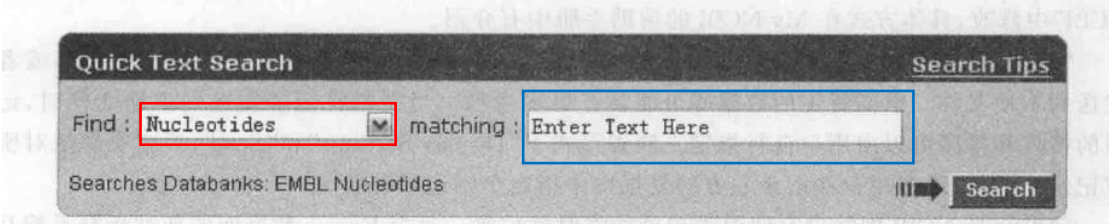


图 2-14　SRS 快速文本检索界面

首先在红色框中选择搜索的数据库，然后在蓝色的输入框中输入搜索的关键字进行搜索，因为只是采用简单匹配，所以搜索结果不是特别好。

4.2、list match 批量搜索

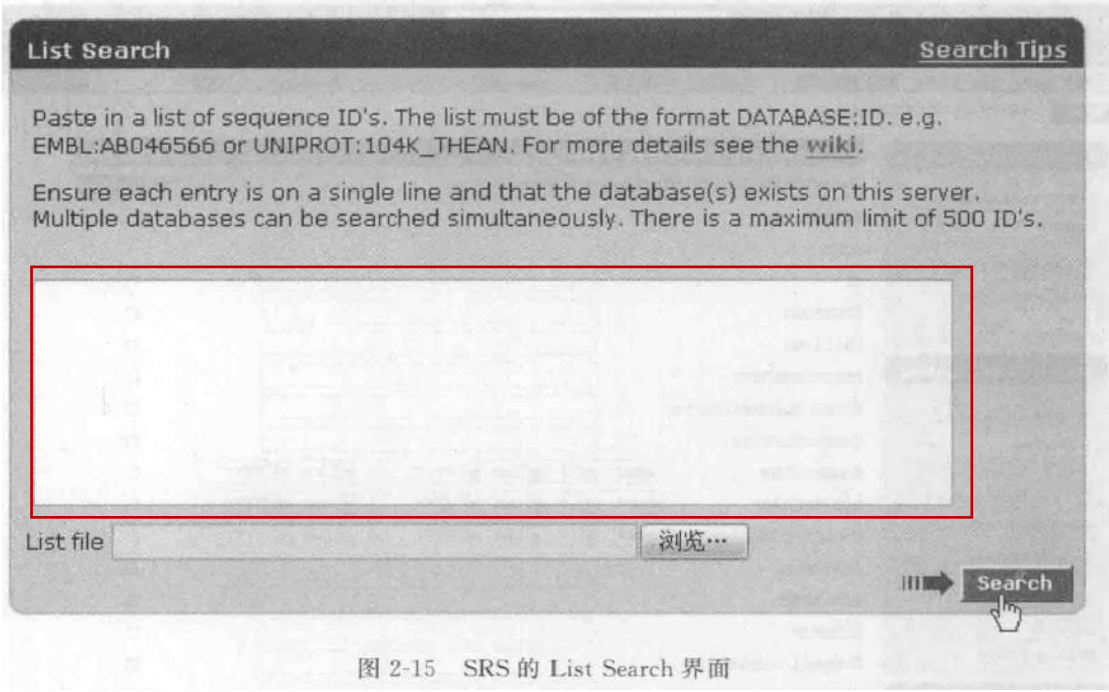


图 2-15　SRS 的 List Search 界面

在图中红色框图中输入想要的搜索的内容，搜索内容输入按照如下格式进行，即数据库名：数据库对应序列的 ID，（如：EMBL：AB046566）即可进行搜索，每次默认显示搜索结果的前 500 行。

4.3、标准表单查询

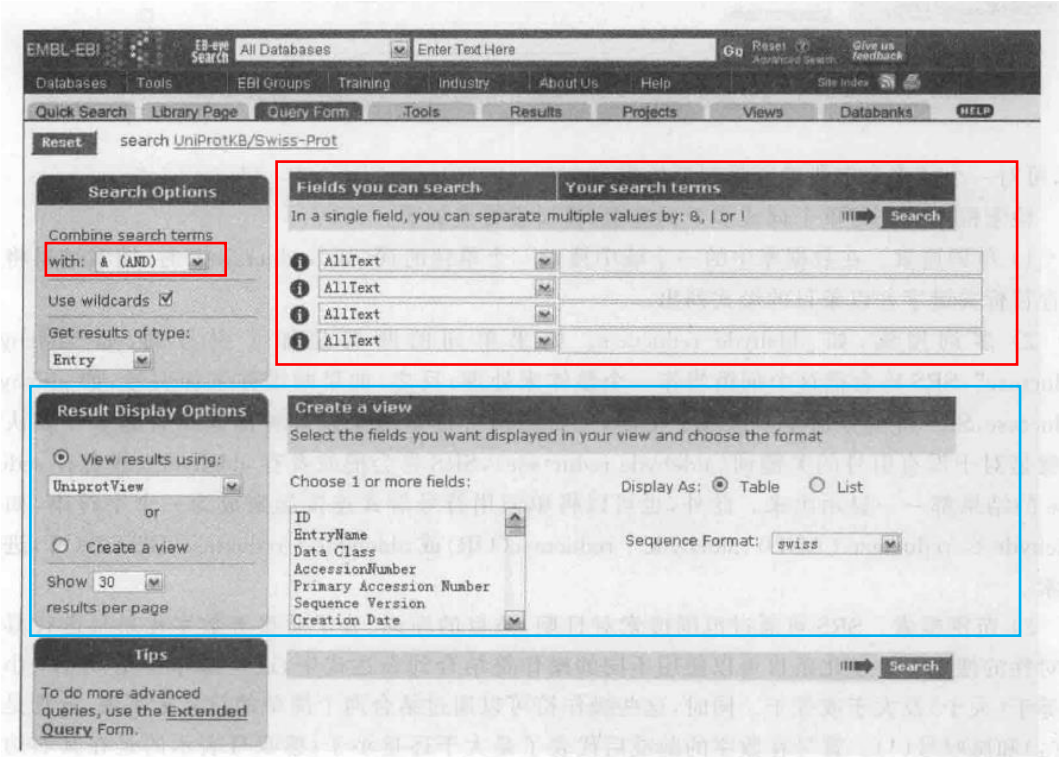


图 2-16　UniProt/Swiss-Prot 数据库标准查询表单模式

首先右边红色框图中选择要搜索的域，然后填写要搜寻的关键字，如果需要进行多个域与关键字的复合搜索，再在左边红色框图中选择相关关键字的关系（AND、OR、NOT），然后在蓝色框图中设定搜索结果的结果显示样式，点击搜索即可。

4.4、扩展表单查询

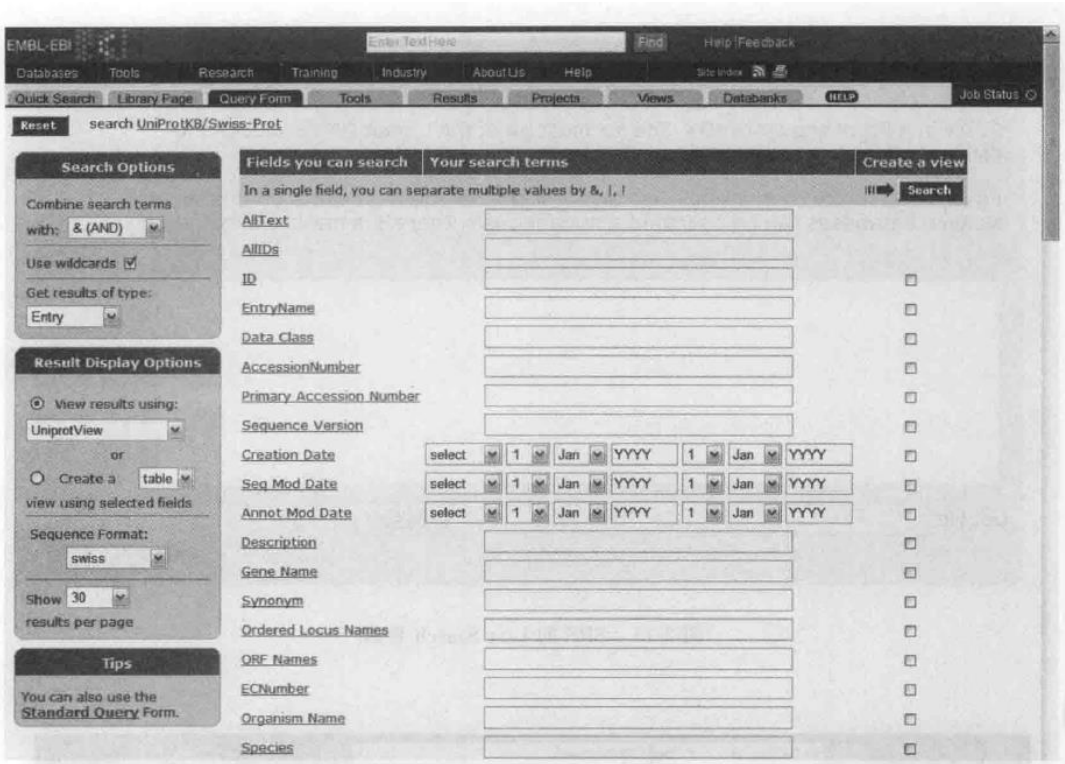


图 2-17　UniProt/Swiss-Prot 数据库扩展查询表单模式(部分字段)

扩展表单查询与标准表单查询的主要区别，在于扩展表单可以限制搜索的信息更加全面，可以对关键字和时间等其他进行更加准确的限定，并且可以使用*和？等通配符进行搜索，搜索结果更加精确。

4.5、搜索式查询



图 2-19　搜索式查询界面

在红色框图中输入搜索内容 Q1，得到 Q1 的搜索结果，然后输入搜索内容 Q2，得到 Q2 的搜索结果,之后输入 Q1+布尔操作符+Q2 即可得到 Q1 和 Q2 的共同搜索结果。

4.6、字段索引检索

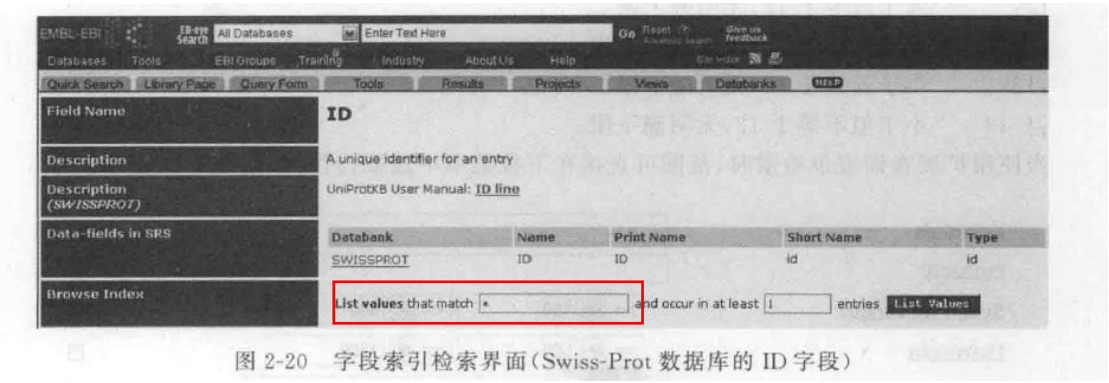


图 2-20　字段索引检索界面(Swiss-Prot 数据库的 ID 字段)

在红色的输入框中输入关键字对应的 ID，进行搜索即可得到该关键字的搜索结果。

第三章 序列比对原理

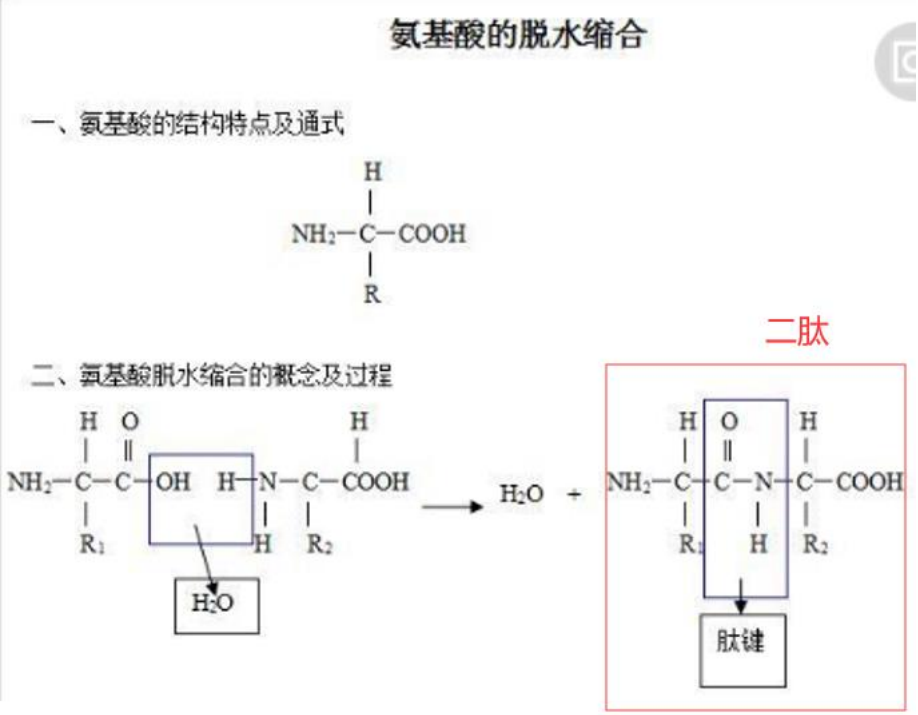
第一节 序列比对相关概念

1、残基（P42）

在化学中，是指化学大分子上的一个部位，例如甲基。而在生物化学与分子生物学里，则是指一个聚合物中，如多糖、蛋白质或核酸上的某个特定单元。

氨基酸残基：

在蛋白质的序列中，氨基酸之间的氨基和羧基脱水成键，氨基酸由于其部分基团参与了肽键的形成，剩余的结构部分则称氨基酸残基。



2、编辑距离（P43）

编辑距离，是指两个字符串之间，由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。一般来说，编辑距离越小，两个字符串的相似度越大。

举个简单的例子：从字符串“kitten”修改为字符串“sitting”只需 3 次单字符编辑操作，如下：①将 k 改为 s；②将 e 改为 i；③在结尾处添加字母 g

因此“kitten”和“sitting”的编辑距离为 3。

3、位点（P45）

位点是指染色体上一个基因或者标记的位置。有时特指 DNA 上

有表达功能的部分。

4、直系同源、旁系同源（P45）

直系同源就是指不同物种中具有相同功能的同源基因；旁系同源就是一个物种内由源基因变异而来的不同基因。

第二节 序列比对打分方法

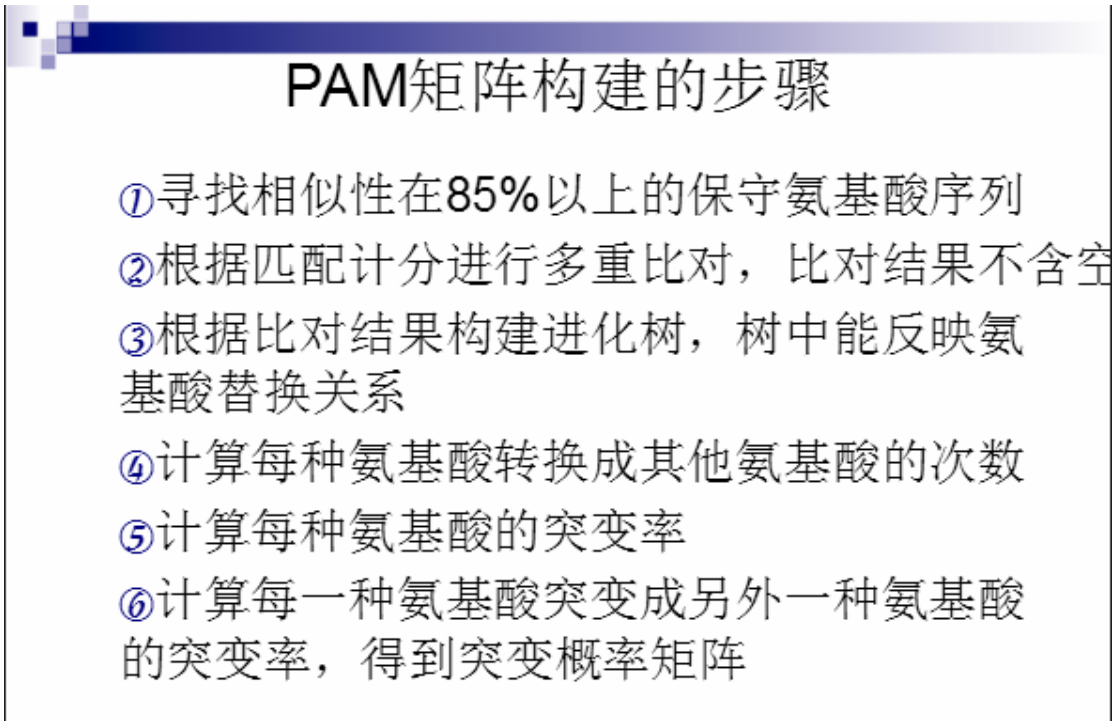
1、20 种氨基酸对应的符号（P47）

	缩写	中文译名	支链	分子量	等电点	羧基解离常数	氨基解离常数	Pkr(R)	R基
Gly	G	甘氨酸	亲水性	75.07	6.06	2.35	9.78		-H
Ala	A	丙氨酸	疏水性	89.09	6.11	2.35	9.87		$-CH_3$
Val	V	缬氨酸	疏水性	117.15	6.00	2.39	9.74		$-CH-(CH_3)_2$
Leu	L	亮氨酸	疏水性	131.17	6.01	2.33	9.74		$-CH_2-CH(CH_3)_2$
Ile	I	异亮氨酸	疏水性	131.17	6.05	2.32	9.76		$-CH(CH_3)-CH_2-CH_3$
Phe	F	苯丙氨酸	疏水性	165.19	5.49	2.20	9.31		$-CH_2-C_6H_5$
Trp	W	色氨酸	疏水性	204.23	5.89	2.46	9.41		$-C_8NH_6$
Tyr	Y	酪氨酸	疏水性	181.19	5.64	2.20	9.21	10.46	$-CH_2-C_6H_4-OH$
Asp	D	天冬氨酸	酸性	133.10	2.85	1.99	9.90	3.90	$-CH_2-COOH$
Asn	N	天冬酰胺	亲水性	132.12	5.41	2.14	8.72		$-CH_2-CONH_2$
Glu	E	谷氨酸	酸性	147.13	3.15	2.10	9.47	4.07	$-(CH_2)_2-COOH$
Lys	K	赖氨酸	碱性	146.19	9.60	2.16	9.06	10.54	$-(CH_2)_4-NH_2$
Gln	Q	谷氨酰胺	亲水性	146.15	5.65	2.17	9.13		$-(CH_2)_2-CONH_2$
Met	M	甲硫氨酸	疏水性	149.21	5.74	2.13	9.28		$-(CH_2)-S-CH_3$
Ser	S	丝氨酸	亲水性	105.09	5.68	2.19	9.21		$-CH_2-OH$
Thr	T	苏氨酸	亲水性	119.12	5.60	2.09	9.10		$-CH(CH_3)-OH$
Cys	C	半胱氨酸	亲水性	121.16	5.05	1.92	10.70	8.37	$-CH_2-SH$
Pro	P	脯氨酸	疏水性	115.13	6.30	1.95	10.64		$-C_3H_6$
His	H	组氨酸	碱性	155.16	7.60	1.80	9.33	6.04	
Arg	R	精氨酸	碱性	174.20	10.76	1.82	8.99	12.48	

2、PAM 打分矩阵（P47）

PAM 打分矩阵是生物信息学中用来判断两个序列相似性的矩阵。这个矩阵的各个格点上的数值为这一氨基酸被其他氨基酸替换的概

率。



由 PAM1 到 PAM250：

PAM250 是 PAM1 自身相乘 250 次得到的，表明现存所有蛋白质序列是经历了 250 次的变异后而呈现的概率最适宜对现有的蛋白质比对做区分，Dayhoff 并不以进化的初期的蛋白质群作为衡量的标准，也不让矩阵自乘至无穷，而是选择了她假定的一个过程，PAM1 自身相乘了 250 次这个过程表明了地球现有物种基因的任何一个氨基酸都有突变成另一个氨基酸的可能性，但大小各不相同。

这种矩阵的缺点是一旦 PAM1 的矩阵有误差，那么自乘 250 后得到的 PAM250 矩阵的误差就会变得很大。

这个矩阵现在已经很少使用，而被 BLOSUM 代替。

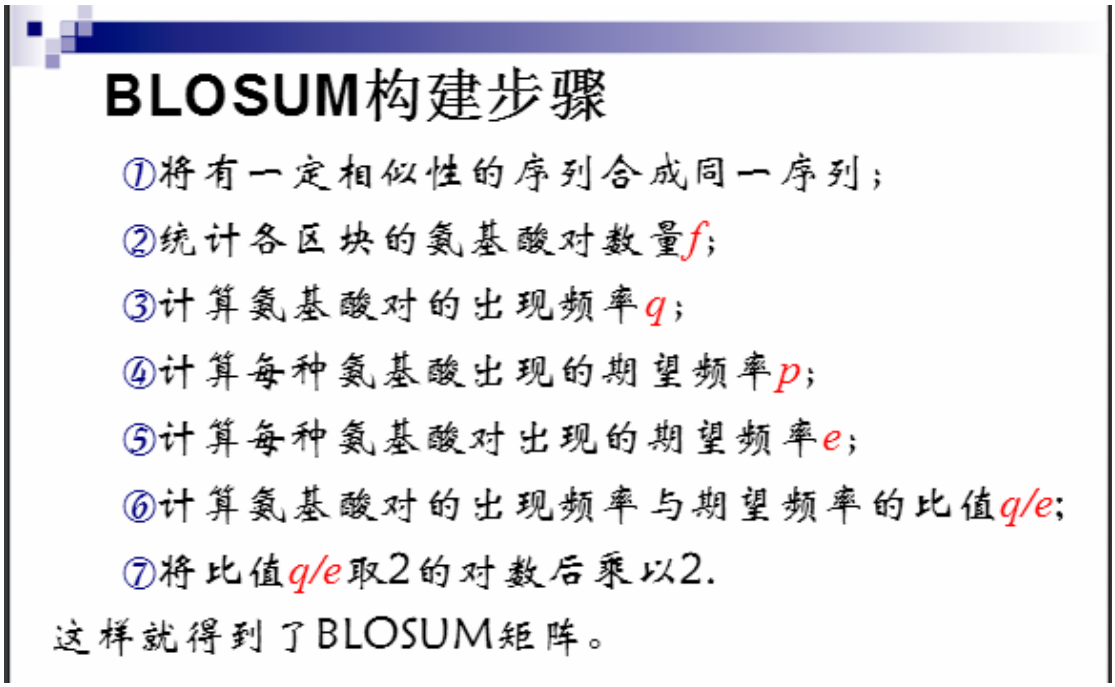
3、进化距离（P49）

从 20 世纪中期开始，分子生物学和分子遗传学兴起。通过对 DNA 的研究发现，生物的进化，归根结蒂是由于 DNA 的变异。

科学家经过研究，发现 DNA 的自发变异具有固定的频率，就是说，过一段时间，DNA 上的某一个位点的碱基就会被另一个碱基所取代。通过对不同物种某一特定基因进行测序，就可以知道这些物种的这一基因中不同的碱基发生取代的大致年代。由此上溯，就可以知道不同物种在进化过程中互相分离的时间了，而这个时间就是生物之间的进化距离。

4、BLOSUM 打分矩阵（P49）

BLOSUM 矩阵，称为模块替换矩阵，英文全称为 Blocks Substitution Matrix。BLOSUM 打分矩阵是生物信息学中用来判断两个序列相似性的矩阵。矩阵中的每个点的分值来自蛋白质比对的局部块中的替代频率的观察。



其中，BLOSUM-62 用来比较 62％相似度的序列，BLOSUM-80 用来比较 80％相似性的序列。

第三节 序列比对算法

1、回溯法（P53）

回溯法（探索与回溯法）是一种选优搜索法，又称为试探法，按选优条件向前搜索，以达到目标。

基本思想：在包含问题的所有解的解空间树中，按照深度优先搜索的策略，从根结点出发深度探索解空间树。当探索到某一结点时，要先判断该结点是否包含问题的解，如果包含，就从该结点出发继续探索下去，如果该结点不包含问题的解，则逐层向其祖先结点回溯。若用回溯法求问题的所有解时，要回溯到根，且根结点的所有可行的子树都要已被搜索遍才结束。而若使用回溯法求任一个解时，只要搜索到问题的一个解就可以结束。

2、BLAST 算法（P54）

Blast 算法是 1990 年由 Altschul 等人提出的两序列局部比对算法，采用了一种短片段匹配算法和一种有效的统计模型来找出目的序列和数据库之间的最佳局部比对效果。广泛被使用在蛋白质 DNA 序列的分析问题中，在其他序列相似性比对中也有应用。

基本思想是：通过产生数量更少的但质量更好的增强点来提高匹配的精确度。

步骤：首先采用哈希法对查询序列以碱基的位置为索引建立哈希表，然后将查询序列和数据库中所有序列联配，找出精确匹配的“种子”，以“种子”为中心，使用动态规划法向两边扩展成更长的联配，最后在一定精度范围内选取符合条件的联配按序输出。得分最高的联配序列就是最优比对序列。

第四节 序列比对工具

1． FASTA 算法（P55）

FASTA 程序是第一个广泛使用的数据库相似性搜索程序。程序引用取代矩阵实行局部比对以获得最佳搜索。但众所周知，使用这种策略会非常耗费工作时，为了提高速度，在实施耗时的最佳搜索之前，程序使用已知的字串检索出可能的匹配。在速度和敏感度之间权衡选择依赖于 ktup 参数。它决定了字串的大小。增大 ktup 参数就会减少字串命中的数目，也就会减少所需要的最佳搜索的数目和搜索的速度。

第五节 多序列比对

1． 傅里叶变换（P62）

表示能将满足一定条件的某个函数表示成三角函数（正弦和/或余弦函数）或者它们的积分的线性组合。

第四章 蛋白质结构预测与分析

第一节 蛋白质结构组织层次

1、靶分子（Biotarget molecule）（P64）

靶分子即外源性物质进入生物体后进攻并与之结合的生物大分子（如蛋白质，脱氧核糖核酸），靶分子与何者结合除了与物质的结构有关还与热力学和动力学等因素有关。

2、肽键、多肽链、多肽（P64）

肽键是指一个氨基酸的羧基与另一个氨基酸的氨基脱水缩合形成的酰胺键，其化学式为-CO-NH-。

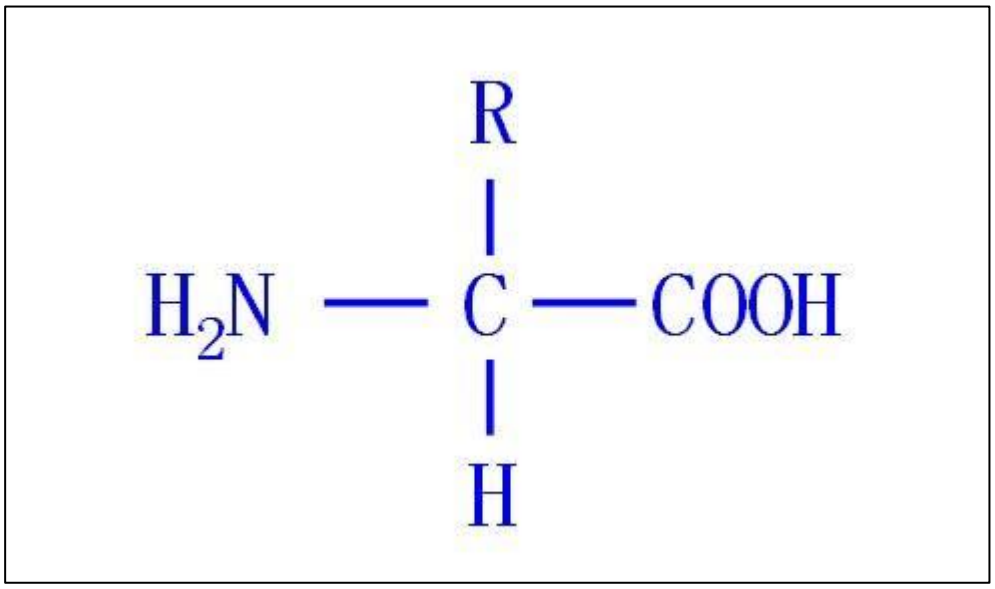
每两个氨基酸相互连接形成一个肽键，多个氨基酸相互连接就形成了多个肽键，由多个氨基酸相互连接形成的含有多个肽键的一条链状结构称为多肽链。

两个或以上的氨基酸脱水缩合形成若干个肽键从而组成一条多肽链，多条多肽链进行多级折叠就组成一个蛋白质分子。蛋白质有时也被称为“多肽”。

3、氨基酸残基（amino acid residue）（P64）

简单的说，氨基酸残基就是指不完整的氨基酸。一个完整的氨基酸包括一个羧基（—COOH），一个氨基(—NH2)，一个 H，一个 R 基。组成多肽的氨基酸在相互结合时，由于其部分基团参与了肽键的形成

而失去一分子水，因此把多肽中的氨基酸单位称为氨基酸残基，即由肽键连接的氨基酸失水部分。



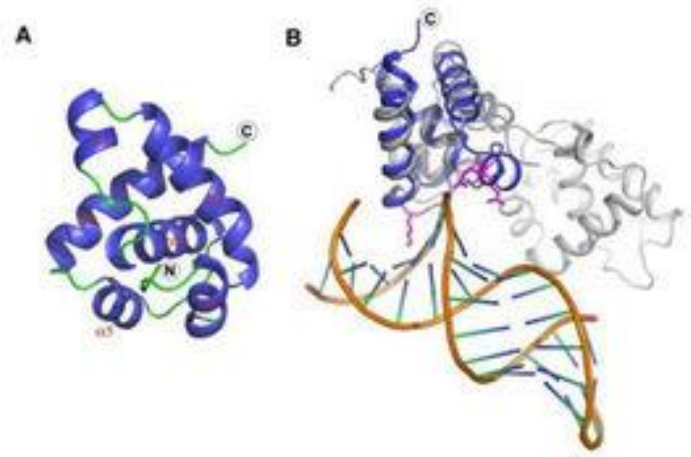
氨基酸构成图

4、氢键（P65）

氢原子与电负性大的原子 X 以共价键结合，若与电负性大、半径小的原子 Y（O、F、N 等）接近，在 X 与 Y 之间以氢为媒介，生成 X-H···Y 形式的一种特殊的分子间或分子内相互作用，称为氢键。

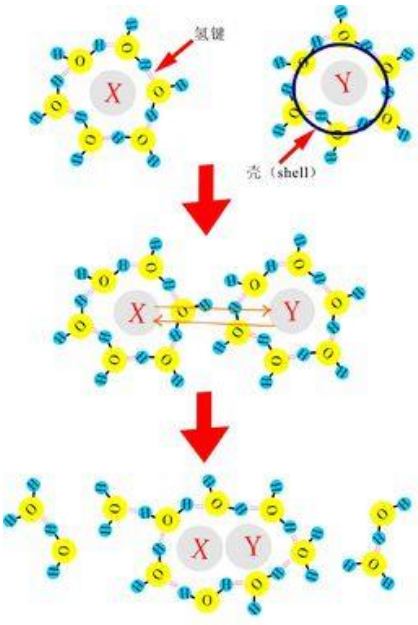
5、亚基（P67）

每一条多肽链都有其完整的三级结构，称为亚基（subunit）。亚基之间呈特定的三维空间排布，并以非共价键链接。蛋白质分子中各个亚基的空间排布及亚基接触部位的布局和相互作用，成为蛋白质的四级结构。



6、疏水相互作用（P67）

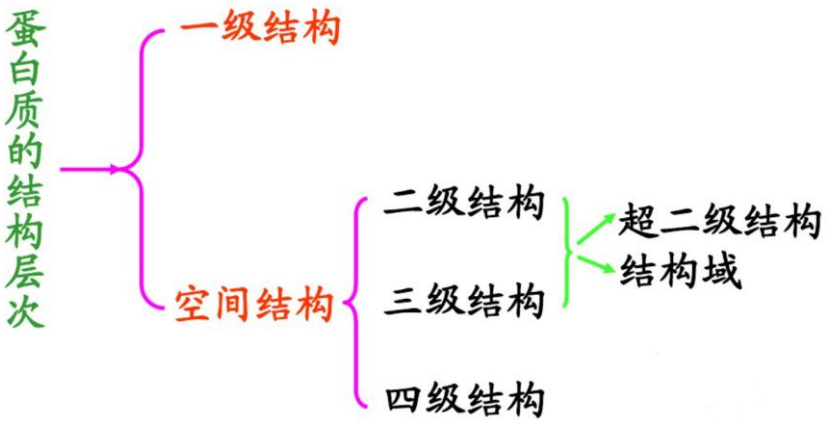
疏水相互作用是蛋白质折叠的主要驱动力。疏水基团彼此靠近聚集以避开水的现象称为疏水相互作用(hydrophobic interaction)。



疏水相互作用过程示意图(其中X、Y为微溶于水的分子)

小结：

一、蛋白质的结构层次



一级结构：多肽链氨基酸残基排列顺序（顺序结构）

二级结构：多肽链局部的空间结构（空间结构）

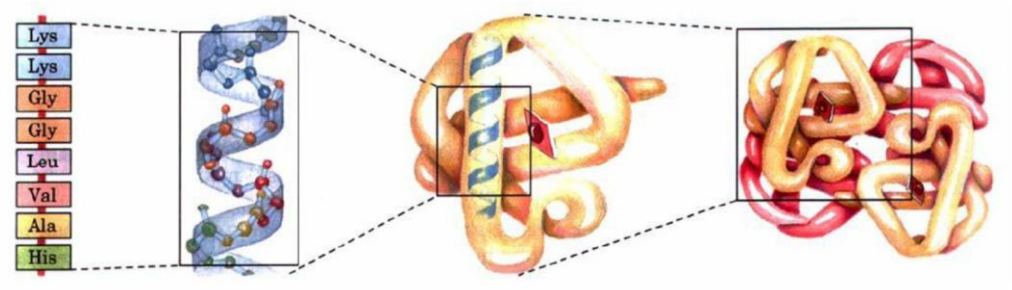
超二级结构：相邻的二级结形成的空间结构

结构域：干个超二级结构排列在一起，形成的相对独立的局部折叠区域

三级结构：整条多肽链的三维结构

四级结构：亚基（多肽链）与亚基之间通过疏水作用等次级键结合成为有序排列的特定空间结构。

蛋白质结构的主要层次



一级结构　二级结构　　三级结构　　四级结构

一级结构：指蛋白质分子中的氨基酸的排列顺序。

二级结构：指多肽链本身通过氢键沿一定方向盘绕、折叠而形成的构象。

三级结构：指多肽链在二级结构的基础上，各原子的空间排布，是二级结构的基础上范围更广的盘旋和折叠。

四级结构：亚基与亚基之间通过疏水作用等次级键结合成为有序排列的特定空间结构。

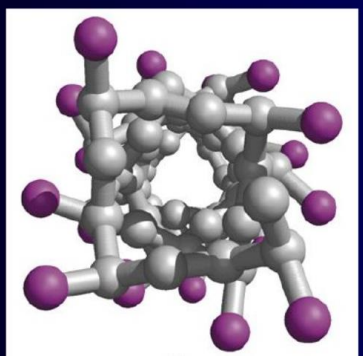
视频网址：https://www.bilibili.com/video/av12700308?p=2

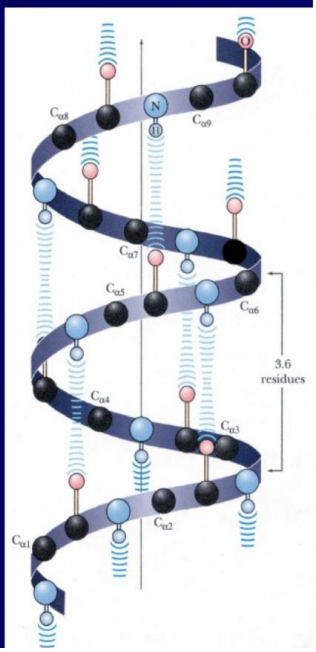
次级键：除了典型的强化学键(共价键、离子键和金属键)等，依靠氢键以及弱的共价键和范德华作用力（即分子间作用力）相结合的各种分子内和分子间作用力的总称。

二级结构的α螺旋、β折叠、β转角和无规卷曲四种形式的介绍

1. α-螺旋(α-helix)

- 肽链中的酰胺平面绕C(相继旋转一定角度形成α-螺旋，呈右手螺旋。酰胺平面平行于中心轴；
- 螺旋体中所有氨基酸残基侧链都伸向外侧；



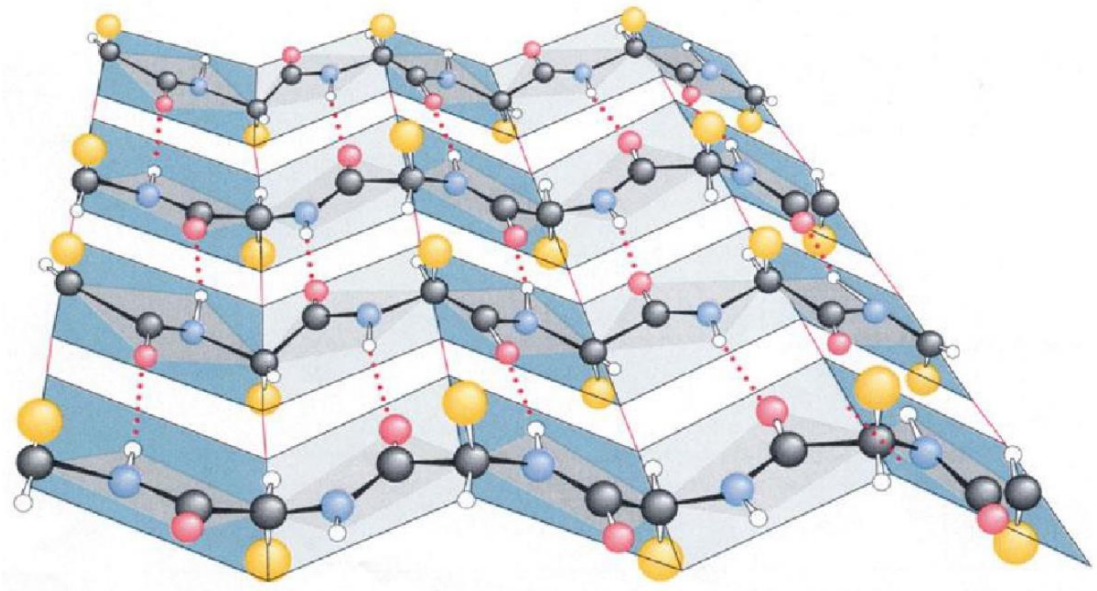


2、β-折叠(β-pleated sheet)

β-折叠是由两条或多条伸展的多肽链靠氢键联结而成的锯齿状片状结构。

侧链基团与C_α间的键几乎垂直于折叠平面，R基团交替地分布于片层平面两侧。

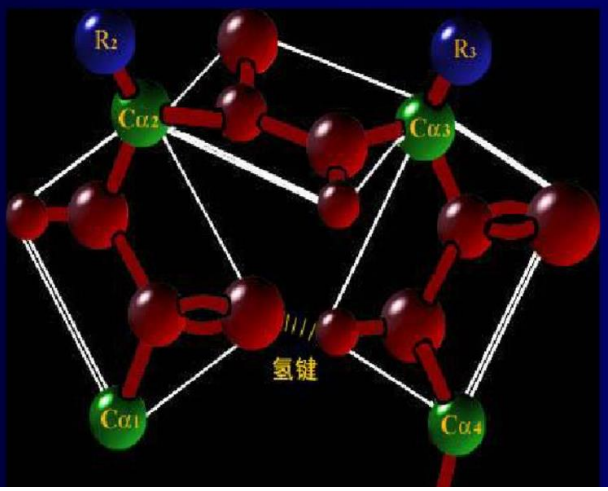
β-折叠分平行式和反平行式，后者更为稳定。



3、β -转角（β -turn）

也称β -回折，存在于球状蛋白中。其特点是肽链回折180°，使得氨基酸残基的C=O和与第四个残基的N-H形成氢键。

β -转角都在蛋白质分子的表面。



4、无规则卷曲（nonregular coil）

又称自由卷曲，是指没有一定规律的松散肽链结构。酶的功能部位常常处于这种构象区域。

无规卷曲常出现在a-螺旋与a-螺旋、a-螺旋与β -折叠、 β -折叠与β -折叠之间。它是形成蛋白质三级结构所必需的。

7、蛋白质结构分类数据库 SCOP2 和 CATH（P68）

（1）SCOP2 数据库

蛋白质被分为四个层次：家族 ➡ 超家族 ➡ 折叠 ➡ 结构类型

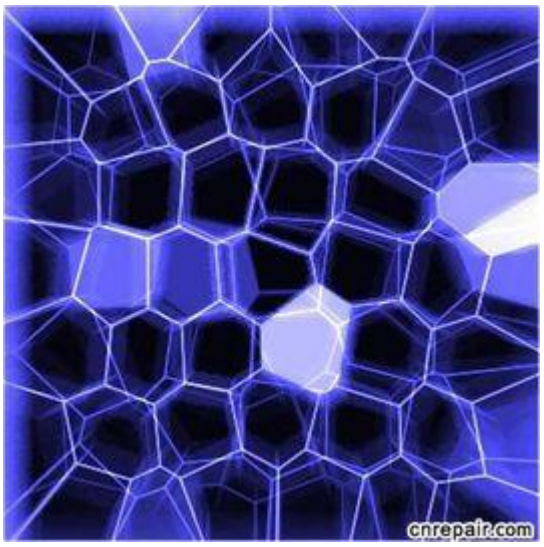
（2）CATH 数据库

蛋白质被分为四个层次：类型 ➡ 构架 ➡ 拓扑结构 ➡ 同源性

第二节 蛋白质的测定与理论预测

1、晶格（P70）

晶体内部原子是按一定的几何规律排列的。为了便于理解，把原子看成是一个球体，则金属晶体就是由这些小球有规律堆积而成的物体。为了形象地表示晶体中原子排列的规律，可以将原子简化成一个点，用假想的线将这些点连接起来，构成有明显规律性的空间格架。这种表示原子在晶体中排列规律的空间格架叫做晶格，又称晶架。



2、结构基因组学（P70）

结构基因组学（structural genomics ）是以全基因组测序为目标，确定基因组的组织结构、基因组成及基因定位的基因组学的一个分支；代表基因组分析的早期阶段，以建立具有高分辨率的生物体基因组的遗传图谱、物理图谱及转录图谱为主要内容；以及研究蛋白质组成和结构的学科。

3、射频（P70）

射频（radio frequency，RF）表示可以辐射到空间的电磁频率，频率范围从 300kHz ~ 300GHz 之间。射频就是射频电流，它是一种高频交流变化电磁波的简称。每秒变化小于 1000 次的交流电称为低频电流，大于 10000 次的称为高频电流，而射频就是这样一种高频电流。

4、同源蛋白质（P71）

同源蛋白质指进化上相关的蛋白质。即不同物种中具有相同或相似功能的蛋白质或具有明显序列同源性的蛋白质。

序列同源性：同源蛋白质的氨基酸序列具有明显的相似性，这种相似性称为序列同源性。

5、金标准（P71）

所谓“金标准”是指当前临床医学界公认的诊断疾病的最可靠、最准确、最好的诊断方法。临床上常用的金标准有组织病理学检查（活检、尸检）、手术发现、影像诊断（CT、核磁共振、彩色 B 超）、病原体的分离培养以及长期随访所得的结论。金标准一般是特异性诊断方法，可以正确区分为“有病”和“无病”。但目前金标准在诊断医学中的应用不是很多。

6、蛋白质折叠（P76）

蛋白质的基本单位为氨基酸，而蛋白质的一级结构指的就是其氨基酸序列，蛋白质会由所含氨基酸残基的亲水性、疏水性、带正电、带负电等特性通过残基间的相互作用而折叠成一立体的三级结构。蛋白质在短时间内从一级结构折叠至立体结构的过程称为折叠。

第三节 蛋白质对接

1、载体蛋白、支架蛋白（P84）

载体蛋白又称作载体（carrier）、通透酶（permease）或转运器（transporter）。能够与特异性溶质结合，通过自身构象的变化，将与它结合的溶质转移到膜的另一侧。

支架蛋白在生物学中是许多关键信号通路中的重要调控因子。尽管支架蛋白的功能没有被严格定义，但是它们与信号通路中的多个成分互相作用或结合，形成复合体。在这些信号通路中，支架蛋白调控信号转导，并将通路组分（以复合体的形式）定位在细胞的特定区域，诸如细胞膜、细胞质基质、细胞核、高尔基体、内体及线粒体。

2、转录调控

定义：指通过改变转录速率从而改变基因表达的水平。

作用：控制转录何时发生以及产生多少 RNA。

3、蛋白质对接的分析软件

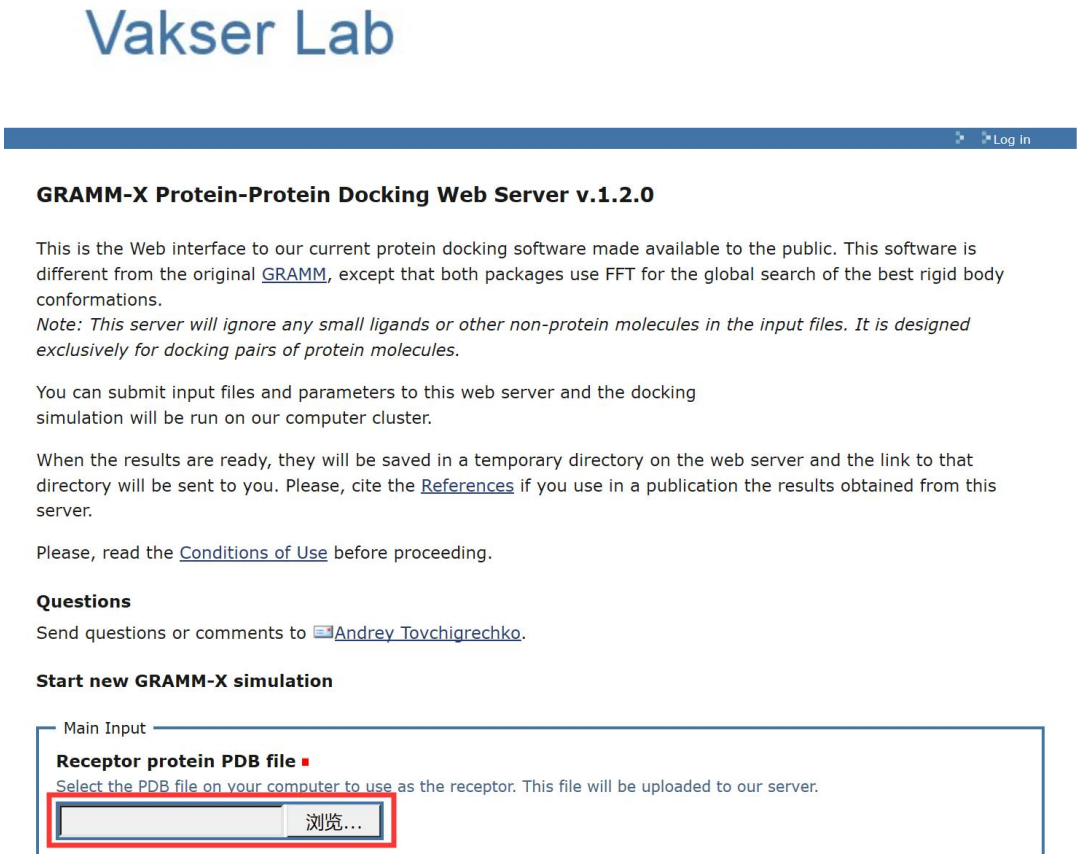
（1）刚性对接软件—ZDOCK（http://zdock.umassmed.edu/）



在红框中输入蛋白质结构文件的 PDB ID。

（2）GRAMMX--刚性对接软件

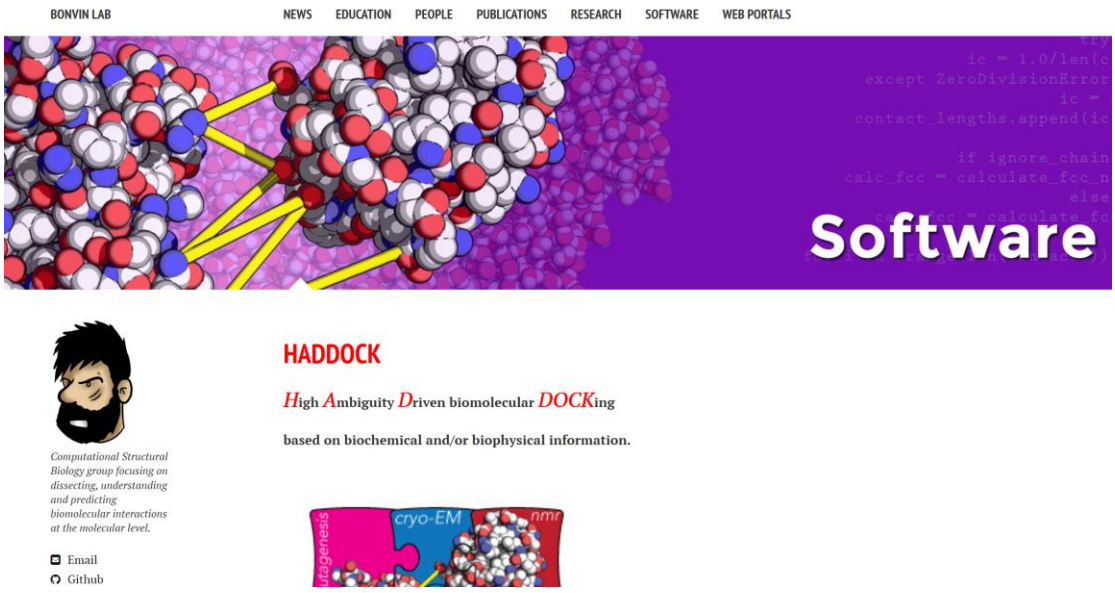
（http://vakser.compbio.ku.edu/resources/gramm/grammx）



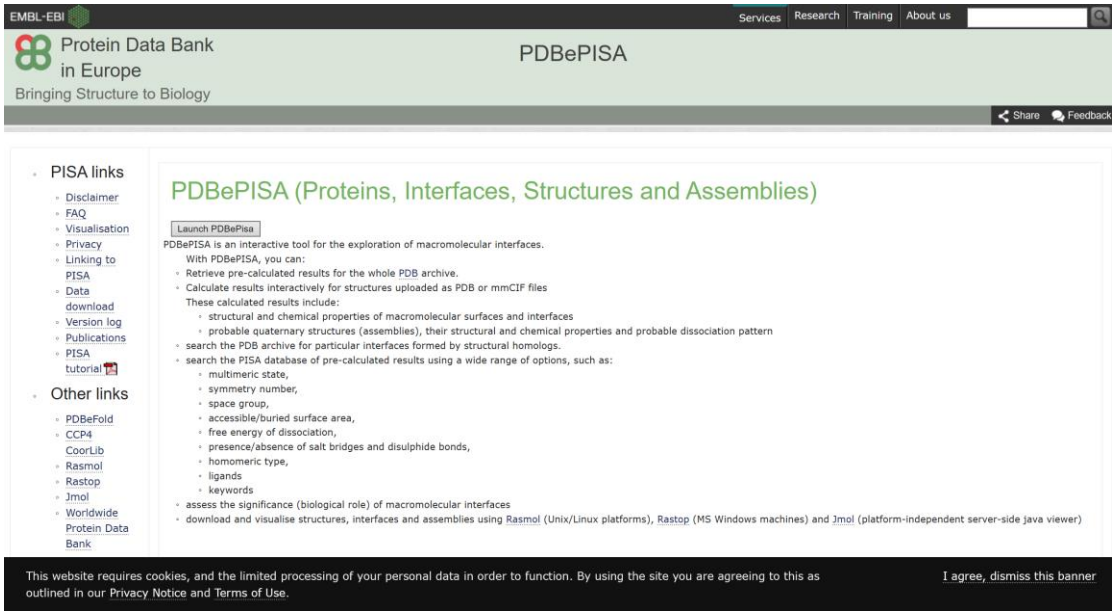
在红框中输入蛋白质结构文件的 PDB ID。

（3）HADDOCK--半柔性对接软件

（http://www.bonvinlab.org/software/haddock2.2/）



（4） PDBePISA—分子相互作用探测分析工具（http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html）



第四节 蛋白质折叠与疾病

1、自由能（P86）

自由能是指在某一个热力学过程中，系统减少的内能中可以转化为对外做功的部分，它衡量的是：在一个特定的热力学过程中，系统可对外输出的“有用能量”。可分为亥姆霍兹自由能和吉布斯自由能。

2、Prion（P88）

朊（ruǎn）病毒蛋白（PRION）是生物体正常基因编码的产物，本不具有感染性和致病性，但是遗传突变可以产生传染型朊病毒，可以将正常的朊病毒异构为传染型朊病毒，其因为结构特殊，无法被细胞内溶酶体中的蛋白酶分解，而在溶酶体中大量积累，最终涨破溶酶体，使其中的蛋白酶流出而对细胞造成破坏，使神经细胞大量死亡而产生

海绵状空洞。疯牛病、羊瘙痒症、库鲁病都是由朊病毒引起。

第五章 真核生物基因组的注释

基因组注释：是利用生物信息学方法和工具，对基因组所有基因的生物学功能进行高通量注释。基因组注释的研究内容包括基因识别和基因功能注释两个方面。基因识别的核心是确定全基因组序列中所有基因的确切位置。

第一节 蛋白质编码基因的注释

1、真核生物、原核生物（P90）

真核生物是指细胞中具有细胞核的生物的总称。

原核生物是指一类细胞核无核膜包裹，只有称作核区的裸露DNA的原始单细胞生物。如细菌、放线菌等。

真核生物与原核生物的根本性区别是前者的细胞内含有成形的细胞核，因此以真核来命名这一类细胞。

2、功能基因组学（P90）

功能基因组学从基因组信息与外界环境相互作用的高度，阐明基因组的功能。功能基因组学的研究内容：人类基因组 DNA 序列变异性研究、基因组表达调控的研究、模式生物体的研究和生物信息学的研究等。

3、假基因（P90）

假基因：或称为伪基因，是基因组中与编码基因序列非常相似的非功能性基因组 DNA 拷贝，一般情况都不被转录，且没有明确生理意义。假基因与正常基因相似，但丧失正常功能的 DNA 序列，他是基因家族在进化过程中形成的无功能的残留物。

4、外显子-内含子结构（P90）

外显子是基因中在 mRNA 剪切后保留的片段，绝大部分的外显子为编码序列。剪切后拼接在一起的外显子序列形成为肽链编码的成熟 mRNA。基因组所含的所有外显子的总和又称外显子组。

内含子是基因中在 mRNA 剪切时切除的部分。现知大部分内含子是无功能的，但也有的基因的内含子中含调节序列。

5、cDNA（P90）

cDNA，互补脱氧核糖核酸，英文全称为 complementary DNA。与 mRNA 链互补的单链 DNA，以其 mRNA 为模板，在适当引物的存在下，由 mRNA 与 DNA 进行一定条件下合成的，就是 cDNA。

与 DNA 的区别：

DNA 指的是生物体的主要遗传物质，内部有内含子等结构。cDNA 的内部无内含子等结构，在基因克隆中利于在原核生物中表达。

6、表达序列标签（P90）

表达序列标签（EST, Expressed Sequence Tag）从互补 DNA(cDNA)

分子所测得部分序列的短段 DNA（通常 300~500bp），代表在一定的发育时期或特定的环境条件下，特定的组织细胞基因表达的序列。可用于验证基因在特定组织中的表达，推导全长 cDNA 序列，或作为标签标志基因组中的特殊位点以确定基因的位置等。

7、RT—PCR（P91）

RT—PCR（Reverse Transcription-Polymerase Chain Reaction）是将 RNA 的反转录（RT）和 cDNA 的聚合酶链式扩增（PCR）相结合的技术。首先经反转录酶的作用，从 RNA 合成 cDNA，再以 cDNA 为模板，在 DNA 聚合酶作用下扩增合成目的片段。RT—PCR 技术灵敏而且用途广泛，可用于检测细胞中基因表达水平、细胞中 RNA 病毒的含量和直接克隆特定基因的 cDNA 序列。

8、开放阅读框（P92）

在分子生物学中，开放阅读框（Open Reading Frame，ORF）从起始密码子开始，是 DNA 序列中具有编码蛋白质潜能，一段无终止密码子打断的碱基序列。

ORF 的识别是证明一个新的 DNA 序列为特定的蛋白质编码基因的部分或全部的先决条件。

9、GC 含量（P92）

GC 含量是在 DNA4 种碱基中，鸟嘌呤（G）和胞嘧啶（C）所占

的比率称为 GC 含量。在双链 DNA 中，腺嘌呤与胸腺嘧啶（A/T）之比，以及鸟嘌呤与胞嘧啶（G/C）之比都是 1。但是，（A+T）/（G+C）之比则随 DNA 的种类不同而异。一种生物的基因组或特定 DNA、RNA 片段有特定的 GC 含量。

10、CpG 岛（P93）

CpG 双核苷酸在人类基因组中的分布很不均一，而在基因组的某些区段，CpG 保持或高于正常概率。CpG 岛主要位于基因的启动子（promotor）和外显子区域，是富含 CpG 二核苷酸的一些区域，长度为 300—3000bp。这里 CpG 是胞嘧啶（C）—磷酸（p）—鸟嘌呤（G）的缩写。

11、负选择（P93）

正选择，是指当一个群体中出现能够提高个体生存力或育性的突变时，具有该基因的个体将比其它个体留下更多的子代，而突变基因最终在整个群体中扩散。

负选择，也叫纯化选择，是指群体中出现有害基因时，携带该基因的个体会因为生存力或育性降低而从群体中淘汰。

12、沉默突变（P93）

沉默突变指基因突变导致了多肽产物氨基酸序列改变，但对蛋白质性质的生理功能没有影响的一种突变。这种突变可以在子代保留，

并以变异体基因或等位基因的形式出现在群体中，是形成蛋白质遗传多态性的主要原因。

13、亚细胞定位（P94）

亚细胞定位是指某种蛋白或表达产物在细胞内的具体存在部位。例如在核内、胞质内或者细胞膜上存在。GFP 是绿色荧光蛋白，在扫描共聚焦显微镜的激光照射下会发出绿色荧光，从而可以精确地定位蛋白质的位置。

14、趋同和趋异（P94）

趋同是指两种或两种以上亲缘关系甚远的生物，由于栖居于同一类型的环境之中，从而演化成具有相似的形态特征或构造的现象。

趋异是指同源生物进化成多种不同类的后代，以适应不同环境的现象。例如，古代一种具有五趾的短腿食虫性哺乳动物，由于适应不同环境而进化成当今各种哺乳动物，如能奔跑的斑马、能滑翔的鼯鼠和蝙蝠、能打洞的犰狳、能游泳的鲸等。

14、水平基因转移（P95）

水平基因的转移，主要发生在微生物中，常常是由质粒介导，是仅次于基因垂直传递重要的基因交流。

第二节 RNA 基因的注释

1、 剪接位点（P95）

剪接体可识别的 RNA 前体中内含子和外显子连接边界的序列和接头位点。根据位置不同可以分为供体和接纳体剪接位点。

第三节 重复序列的注释

1、着丝粒、端粒（P97）

着丝粒是指中期染色体的两条姐妹染色单体的连接处，位于染色体的主缢痕处，着丝粒将两条染色单体分为短臂(p)和长臂(q)，由高度重复的异染色质组成，其主要成分为 DNA 和蛋白质。

端粒是存在于真核细胞线状染色体末端的一小段 DNA-蛋白质复合体，它与端粒结合蛋白一起构成了特殊的“帽子”结构，作用是保持染色体的完整性和控制细胞分裂周期。端粒、着丝粒和复制原点是染色体保持完整和稳定的三大要素。

1、转座、转座元件（P97）

转座：

转座现象是 20 世纪 40 年代由 Barbara McClintock 在进行玉米的遗传学研究时首次发现（B. McClintock 因这一发现荣获了 1983 年诺贝尔奖），并在细菌中得到了最广泛深入的研究。

一段 DNA 片段可以从原位上单独复制或断裂下来，环化后插入另一位点，并对其后的基因起调控作用，此过程称转座。

转座元件：

生物体内非游离的、能自复制或自剪切拷贝的、并能以相同或不相同拷贝在该生物体基因组内不断移动位置的功能性 DNA 片段。

第四节 假基因的注释

无

第五节 案例分析：黄瓜基因组的注释

1、植物性别（P98）

植物和动物一样，植物也存在着性的差别。例如，以雌蕊、雄蕊为特征，可分为雌雄同花植物、雌雄同株异花植物、雌雄异株植物等。

2、图 5-2（P99）

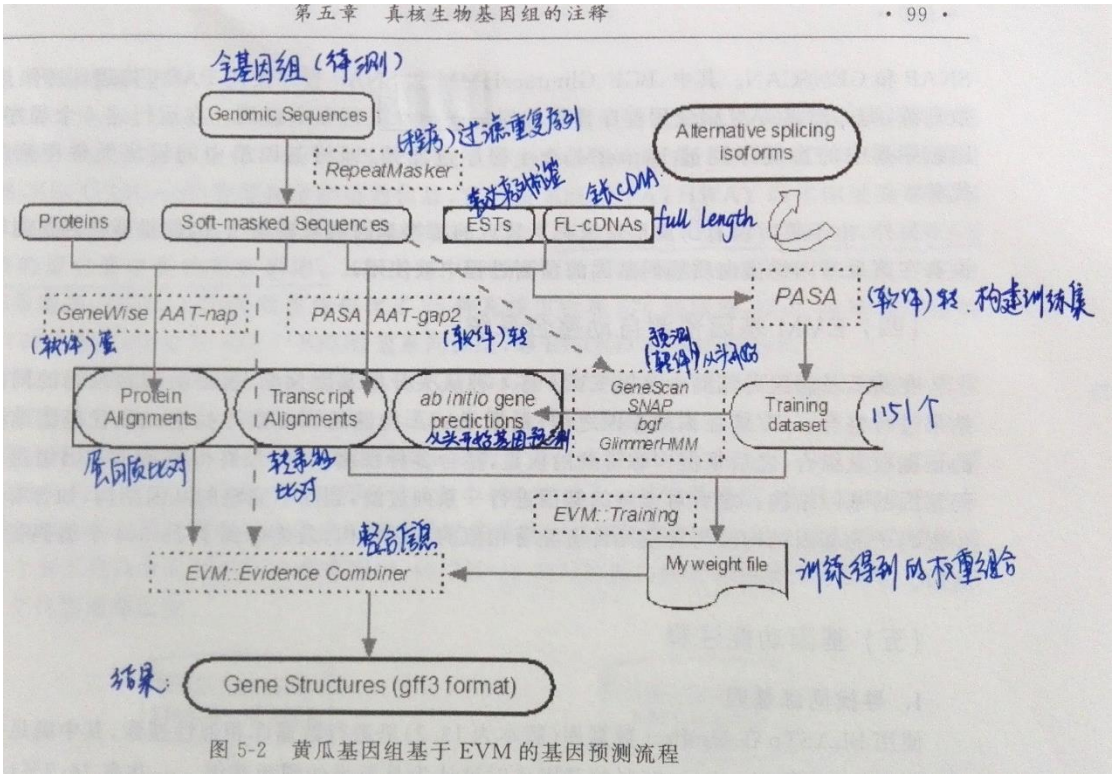
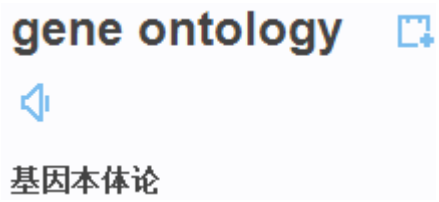


图 5-2 黄瓜基因组基于 EVM 的基因预测流程

3、结构域（P100）

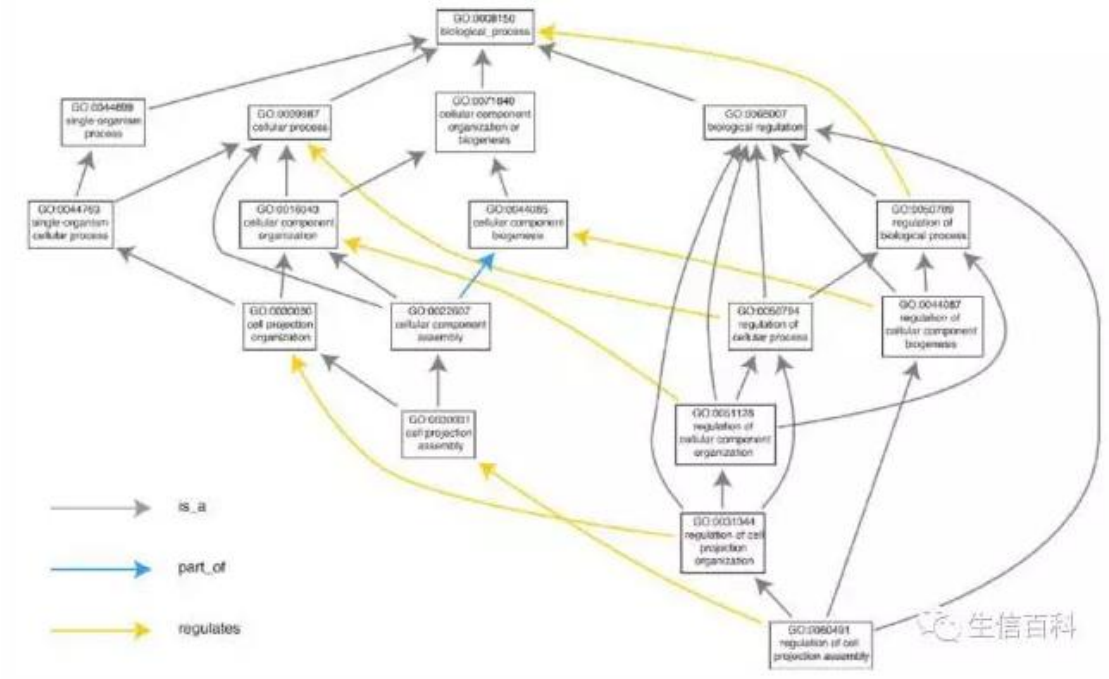
结构域是生物大分子中具有特异结构和独立功能的区域，特别指蛋白质中这样的区域。蛋白质分子中不同的结构域常由基因的不同外显子所编码。结构域也是蛋白质功能单元。在多结构域蛋白质中，不同的结构域常常与不同的功能相联系。

4、GO 注释（P100）



相似的基因在不同物种中，其功能往往保守的。显然，需要一个统一的术语用于描述这些跨物种的同源基因及其基因产物的功能，否则，不同的实验室对相同的基因的功能的描述不同，将极大限制学术的交流。而 GO 项目正是为了能够使对各种数据库中基因获基因产物功能描述相一致的努力结果。

GO，Gene Ontology，是生物学功能注释的一个标准词汇表术语（GO term），不同的 GO term 通过有向无环图关联起来，如下图所示：



目前，GO 注释主要有两种方法：(1) 序列相似性比对；(2) 结构域相似性比对。

第六章 蛋白质组学

本章提要

1、分子生物学（P108）

分子生物学 (molecular biology)是从分子水平研究生物大分子的结构与功能从而阐明生命现象本质的科学。自 20 世纪 50 年代以来，分子生物学是生物学的前沿与生长点，其主要研究领域包括蛋白质体系、蛋白质-核酸体系（中心是分子遗传学）和蛋白质-脂质体系（即生物膜）。1953 年沃森、克里克提出 DNA 分子的双螺旋结构模型是分子生物学诞生的标志。

2、DNA 序列（P108）

DNA 序列或基因序列是使用一串字母表示的真实的或者假设的携带基因信息的 DNA 分子的一级结构。可能的字母只有 A，C，G 和 T，分别代表组成 DNA 的四种核苷酸——腺嘌呤，胞嘧啶，鸟嘌呤，胸腺嘧啶。每个字母代表一种碱基，两个碱基形成一个碱基对，碱基对的配对规律是固定的，即是 :A-T,C-G。他们无间隔的排列在一起，例如序列 AAAGTCTGAC。任意长度大于 4 的一串核苷酸被称作一个序列。

proteomics

英 [p'rəʊti:əʊmɪks] 美 [p'roʊti:oʊmɪks]

蛋白质组学

第一节 蛋白质组学概述

1、孟德尔遗传定律（P108）

孟德尔遗传定律由奥地利帝国遗传学家格里哥·孟德尔在 1865 年发表并催生了遗传学诞生的著名定律。他揭示出遗传学的两个基本定律——分离定律和自由组合定律，统称为孟德尔遗传规律。

2、遗传密码（P108）

遗传密码又称密码子、遗传密码子、三联体密码。遗传密码是一组规则，将 DNA 或 RNA 序列以三个核苷酸为一组的密码子转译为蛋白质的氨基酸序列，以用于蛋白质合成。它决定肽链上每一个氨基酸和各氨基酸的合成顺序，以及蛋白质合成的起始、延伸和终止。

3、聚合酶链反应（P108）

聚合酶链式反应（Polymerase Chain Reaction，PCR）是一种用于放大扩增特定的 DNA 片段的分子生物学技术，它可看作是生物体外的特殊 DNA 复制，PCR 的最大特点是能将微量的 DNA 大幅增加。

4、基因芯片（P110）

基因芯片(genechip)（又称 DNA 芯片、生物芯片）的原型是 80 年代中期提出的。基因芯片的测序原理是杂交测序方法，即通过与一组已知序列的核酸探针杂交进行核酸序列测定的方法，在一块基片表面固定了序列已知的靶核苷酸的探针。当溶液中带有荧光标记的核酸序列 TATGCAATCTAG，与基因芯片上对应位置的核酸探针产生互补匹配时，通过确定荧光强度最强的探针位置，获得一组序列完全互补的探针序列。据此可重组出靶核酸的序列。

5、基因表达（P110）

基因表达是指基因指导下的蛋白质合成过程。生物体生命活动中并不是所有的基因都同时表达，代谢过程中所需要的各种酶和蛋白质的基因以及构成细胞化学成分的各种编码基因，正常情况下是经常表达的，而与生物发育过程有关的基因则要在特定的时空才表达。

6、mRNA 丰度和蛋白质丰度（P110）

mRNA 丰度, 指一种特定的 mRNA 在某个细胞中的平均分子数。

蛋白质丰度即某种蛋白质在蛋白质组中含量的高低。

7、蛋白质翻译（P111）

翻译是蛋白质生物合成（基因表达中的一部分，基因表达还包括转录）过程中的第二步（转录为第一步），翻译是根据遗传密码的中心法则，将成熟的信使 RNA 分子（由 DNA 通过转录而生成）中“碱基的排列顺序”（核苷酸序列）解码，并生成对应的特定氨基酸序列的过程。

8、蛋白质翻译后修饰（P111）

翻译后修饰是指蛋白质在翻译后的化学修饰。对于大部分的蛋白质来说，这是蛋白质生物合成的较后步骤。

9、双向凝胶电泳技术（P111）

双向凝胶电泳由 O'Farrel 以及 Klose 和 Scheele 等人于 1975 年发明的，原理是第一相基于蛋白质的等电点不同用等电聚焦分离，具有相同等电点的蛋白质无论其分子大小，在电场的作用下都会聚焦在某一特定位置即等电点处 ;第二相则按分子量的不同用 SDS-PAGE 分离，把复杂蛋白混合物中的蛋白质在二维平面上分开。

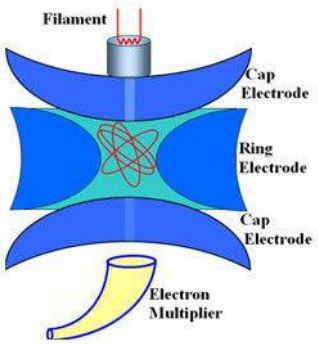
第二节 蛋白质的大规模分离鉴定技术

1、等电点（P112）

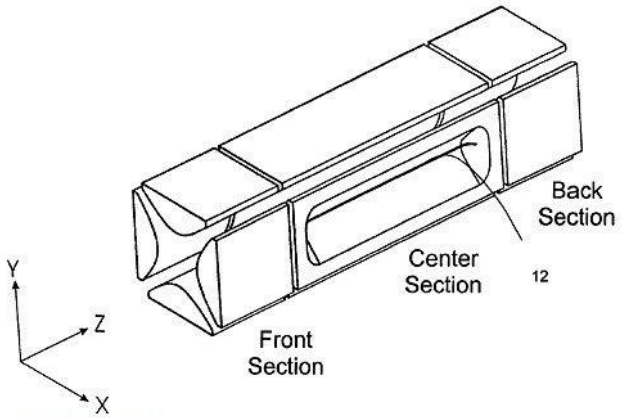
等电点是两性电解质所带电荷因溶液的 pH 值不同而改变，当两性电解质正负电荷数值相等时，溶液的 pH 值即称为该物质的等电点。蛋白质是两性电解质，其等电点和它所含的酸性氨基酸和碱性氨基酸的数量比例有关。各种蛋白质因氨基酸残基组成不同，等电点也不一样。

2、离子阱（P113）

离子阱（Ion trap），大致分为三维离子阱（3D Ion Trap）、线性离子阱（Linear Ion Trap）、轨道离子阱（Orbitrap）三种。除轨道离子阱外，离子阱使用电磁场将离子限定在特定的空间内，通过改变电场的参数，使特定的离子进入不稳定状态，最终导致离子从预留的孔或窄缝中射出到达检测器。除轨道离子阱，离子阱质量选择器因为其特有的空间限定特点，使其在多级质谱分析(MS-MS)中占据了重要的地位。



三维离子阱



线性离子阱

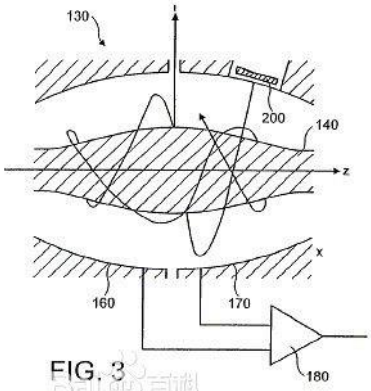


FIG. 3

轨道离子阱

3、免疫沉淀（P113）

免疫沉淀是利用抗体特异性反应纯化富集目的蛋白的一种方法。

4、免疫亲和（P113）

以抗原抗体中的一方作为配基亲和吸附另一方的分离系统称为免疫亲和层析（IAFC）。

5、神经递质受体（P113）

在化学突触传递中担当信使的特定化学物质，简称递质。神经递质受体都是一些跨越细胞膜的蛋白质复合体。

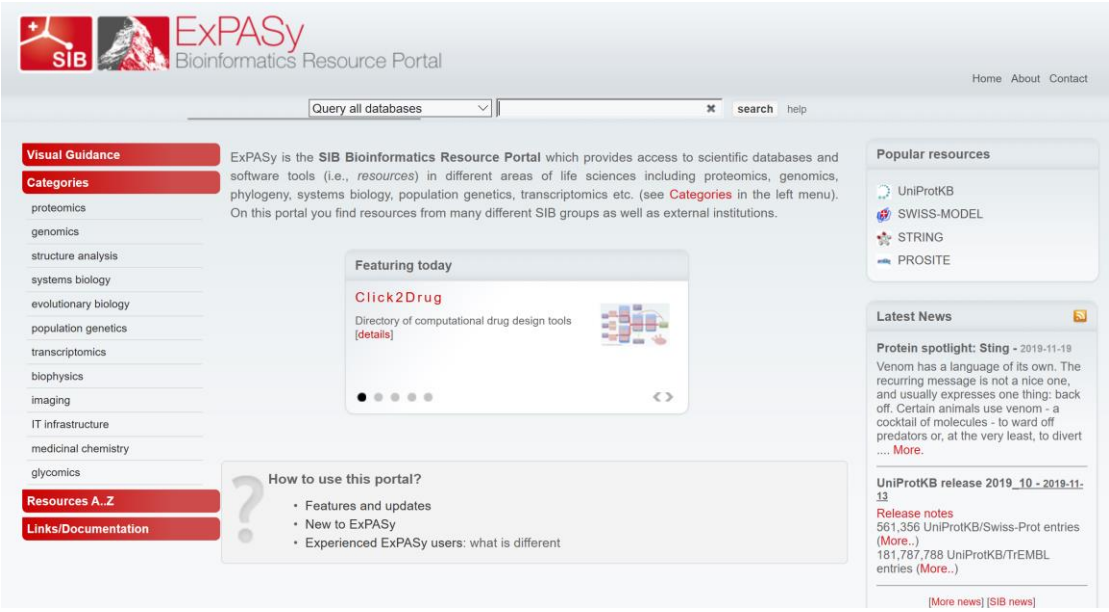
6、配体（P116）

同锚定蛋白结合的任何分子都称为配体（ligand）。在受体介导的内吞中，与细胞质膜受体蛋白结合，最后被吞入细胞的即是配体。根据配体的性质以及被细胞内吞后的作用，将配体分为四大类: Ⅰ.营养物，如转铁蛋白、低密度脂蛋白(LDL)等；Ⅱ.有害物质，如某些细菌；Ⅲ.免疫物质，如免疫球蛋白、抗原等；Ⅳ.信号物质，如胰岛素等多种肽类激素等。

7、肽质量指纹（P117）

肽质量指纹图谱理论上每个蛋白消化后有不同的肽段，这些肽段的质量（分子量）就是这个蛋白的肽指纹图谱。用质谱可以检测出其中所有肽段的质量，然后将这些质量与数据库中所有蛋白指纹进行匹配，就可确定一种未知蛋白。

8. 专家蛋白质分析系统 ExPASy



9、蛋白质谱

蛋白质谱技术简单来说就是一种将质谱仪用于研究蛋白质的技术。它的基本原理是蛋白质经过蛋白酶的酶切消化后成肽段混合物，在质谱仪中肽段混合物电离形成带电离子，质谱分析器的电场、磁场将具有特定质量与电荷比值（即质荷比，M/Z）的肽段离子分离开来，经过检测器收集分离的离子，确定每个离子的 M/Z 值。经过质量分析器可分析出每个肽段的 M/Z，得到蛋白质所有肽段的 M/Z 图谱，即蛋白质的一级质谱峰图。离子选择装置自动选取强度较大肽段离子进行二级质谱分析，输出选取肽段的二级质谱峰图，通过和理论上蛋白质经过胰蛋白酶消化后产生的一级质谱峰图和二级质谱峰图进行比对而鉴定蛋白质。

10、色谱（物理化学分析方法）

色谱又称色层法或层析法，是一种物理化学分析方法，它利用不同溶质（样品）与固定相和流动相之间的作用力（分配、吸附、离子交换等）的差别，当两相做相对移动时，各溶质在两相间进行多次平

衡，使各溶质达到相互分离。

第三节 蛋白质的翻译后修饰

1、亚细胞（P117）

亚细胞（subcell）是比细胞结构更细化的结构。亚细胞结构则指一般是电子显微镜((EMs)下才可见的细胞结构如线粒体，中心体，高尔基体，细胞壁上的纹孔等。

2、基序（P117）

基序，亦称模序、模体、结构域。是指 DNA、蛋白质等生物大分子中的保守序列，介于二级和三级结构之间的另一种结构层次。所谓结构域指蛋白质亚基结构中明显分开的紧密球状结构区域，亦称辖区。

3、假阳性（P117）

一般来说，阳性（+）是表示疾病或体内生理的变化有一定的结果。相反，化验单或报告单上的阴性（－），则多数基本上否定或排除某种病变的可能性。然而，由于环境因素、操作因素、实验方法或者患者自身因素等等可能把不具备阳性症状的人检测出阳性的结果。

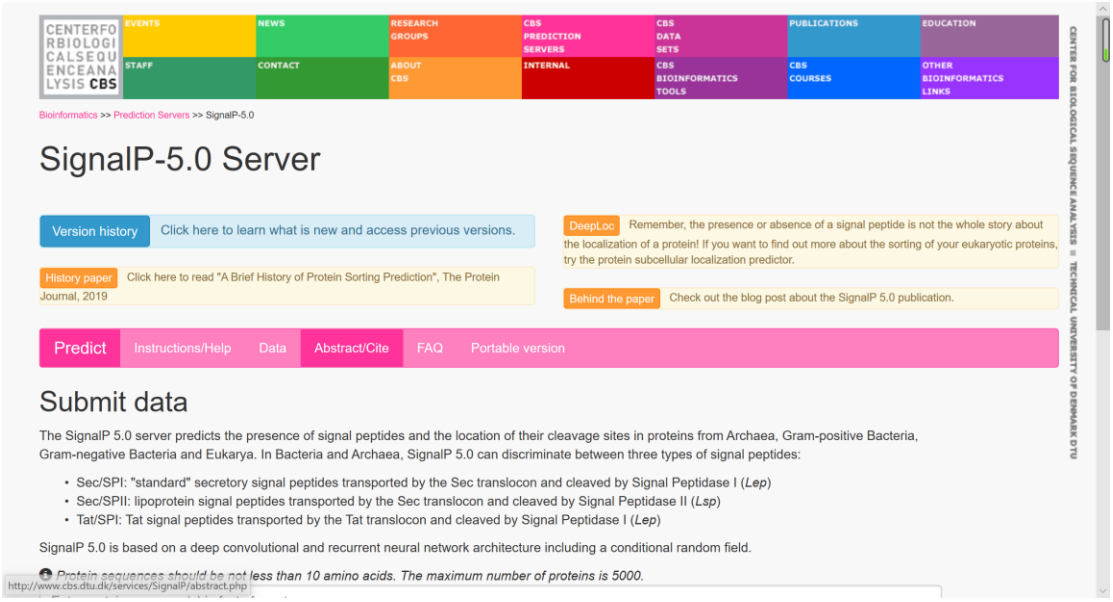
第四节 蛋白质分选

1、革兰氏阳性菌和革兰氏阴性菌（P118）

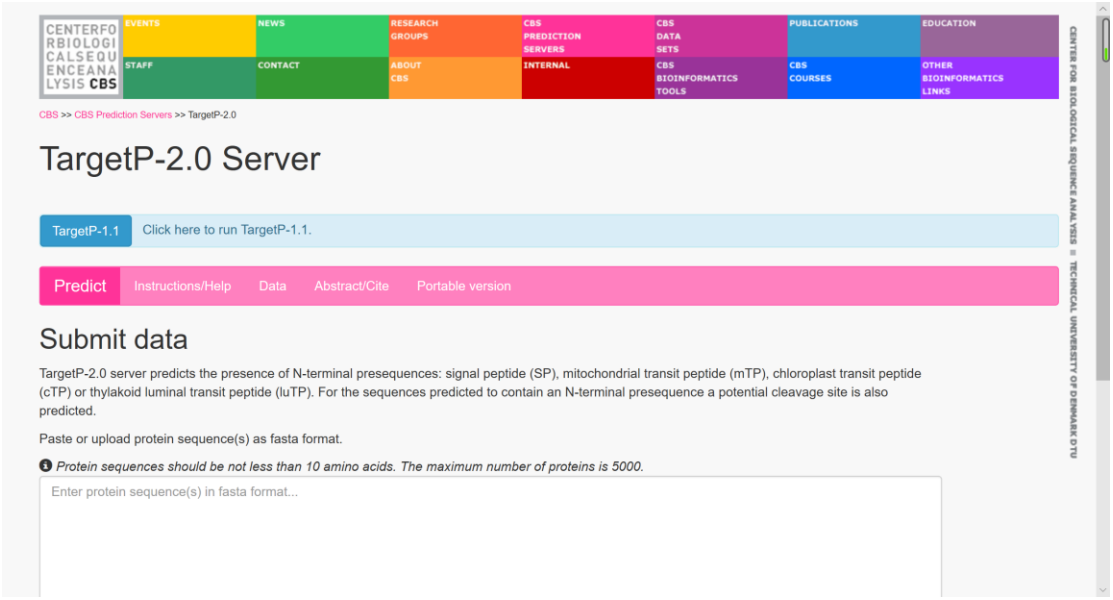
革兰氏阳性菌和革兰氏阴性菌是利用革兰氏染色法来鉴别的两大类细菌。大多数化脓性球菌都属于革兰氏阳性菌，它们能产生外毒素使人致病，而大多数肠道菌多属于革兰氏阴性菌，它们产生内毒素，靠内毒素使人致病。

2、三个亚细胞定位程序（P119）

SignalP(http://www.cbs.dtu.dk/services/SignalP/)



TargetP(http://www.cbs.dtu.dk/services/TargetP/)



PSORT(已改版，现网址为 www.ims.u-tokyo.ac.jp/imsut/jp/)



第五节 蛋白质相互作用

1、操纵子（P122）

操纵子，是指启动基因、操纵基因和一系列紧密连锁的结构基因的总称，是转录的功能单位。操纵子通常由两个以上的编码序列与启

动序列、操纵序列以及其他调节序列在基因组中成簇串联组成。启动序列是与 RNA 聚合酶结合并启动转录的特异 DNA 序列。操纵序列是原核阻遏蛋白的结合位点。当操纵序列结合阻遏蛋白时会阻碍 RNA 聚合酶与启动序列的结合，或使 RNA 聚合酶不能沿 DNA 向前移动，阻遏转录，介导负性调节；另有一种特异 DNA 序列可与激活蛋白结合，使转录激活，介导正性调节。原核生物大多数基因表达调控是通过操纵子机制实现的，如乳糖操纵子、阿拉伯糖操纵子、组氨酸操纵子、色氨酸操纵子等。

第七章 系统生物学

第一节 系统生物学基本概念

1、 鲁棒性（P129）

鲁棒是 Robust 的音译，也就是健壮和强壮的意思。它也是在异常和危险情况下系统生存的能力。比如说，计算机软件在输入错误、磁盘故障、网络过载或有意攻击情况下，能否不死机、不崩溃，就是该软件的鲁棒性。所谓“鲁棒性”，也是指控制系统在一定（结构，大小）的参数摄动下，维持其它某些性能的特性。

第二节 系统生物学基本技术与方法

1、野生型基因（P132）

在自然群体中往往有一种占多数的等位基因，称为野生型基因。与之相对的是突变型基因。

在目前的研究中是把从大自然中获得的个体，也就是非人工诱变的，作为野生型，那么它所携带的就是野生型的基因组。但事实上，自然界本身就发生了很多的突变，然后进行选择，得到的同一个物种会有许多不同的生态型，通常是取自不同的产地。对于同一个基因，不同生态型的序列会不一样，但都是野生型的。

第三节 先进的成像技术

无

第四节 基因表达调控网络

1、基因的差异表达（P140）

在细胞分化的过程中，有些基因是表达的，有些基因是关闭的。细胞的分化就是基因选择性表达的结果，也就是基因的差异表达。

现代分子生物学研究表明，人类基因组约由 10 万左右的不同基因组成，这些基因选择性的表达决定了机体整个生命过程，基因表达的变化处于控制生物学调节机制的中心位置。差异表达不仅有助于阐明生命的奥秘，而且还能为基因诊断与治疗提供重要的理论依据。近几年来，差异表达基因克隆技术不断完善与发展，已成为研究肿瘤和疾病等相关基因的重要手段。

2、野生型基因与突变型基因（P140）

野生型基因指自然界中占多数的等位基因，在生物学实验中常作

为标准对照基因。

突变型基因为 DNA 分子中发生碱基对的增添、缺失或改变，而引起的基因结构的改变。

第五节 代谢网络

1、代谢网络（P143）

有关细胞的代谢网络可以分不同层次来讨论：基因组（DNA 层次）、代谢途径及生化反应网络（蛋白质层次）、代谢流（物流层次）、代谢生理（微生物细胞层次）等等。

对蛋白质层次的代谢网络来说，一个代谢物分子就是一个节点，而节点之间的连结则是生化反应。大部分的分子只参加一种或两种反应，但少数分子参与许多反应，它们实际上就是代谢流的集散中心或通用代谢物。

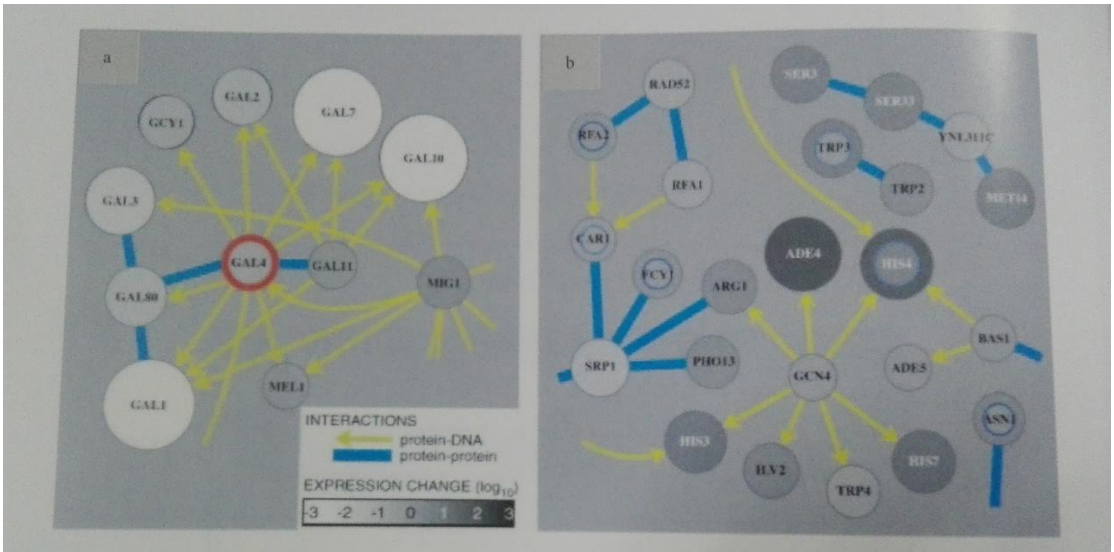


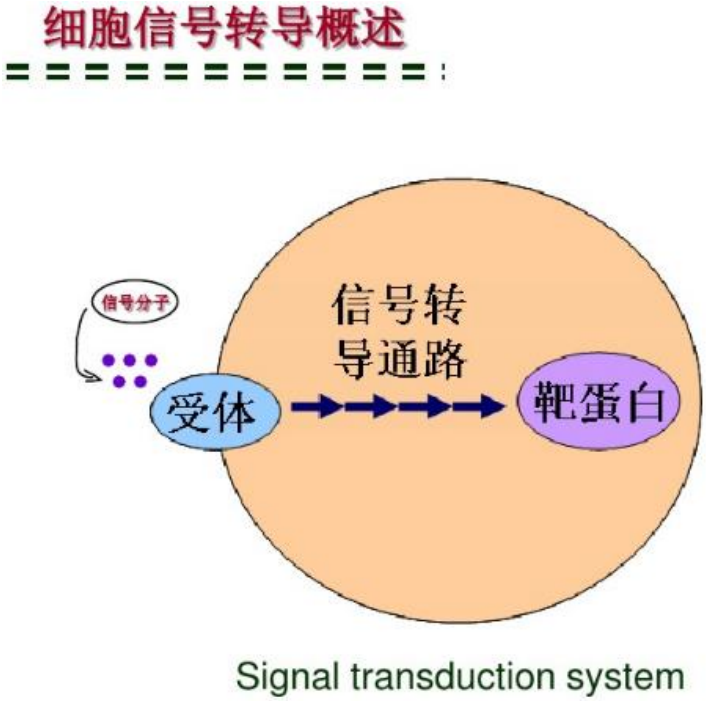
图 7-14　整合的物理相互作用网络实例。a.半乳糖利用；b.氨基酸生物代谢。节点表示基因，从一个节点到另一个节点的黄色有向箭头表示蛋白质-DNA相互作用，两节点之间的蓝线表示对应的蛋白质有物理相互作用。每一节点的灰度表示对应基因mRNA表达的变化，中间灰度表示无变化，暗色和亮色表示表达的增强和减弱。图b的节点有两个不同区域，外环表示mRNA表达的变化，内环表示蛋白质表达的变化。图a中用红色外沿圆环表示对GAL4的外部扰动 (Ideker et al.，2001)

（图片来自书“彩板”图 7-14）

第六节 信号传导途径

1、细胞的信号传导途径（P146）

在生物体中，细胞之间是相互联系的，相互作用的。机体产生的各种各样的信号分子，例如激素和细胞因子，在细胞膜上结合之后，就会与细胞膜上的受体结合，激活细胞内的一系列生化反应，使细胞能够产生一定的反应。从细胞膜到细胞内的这样的反应途径，就是信号传导途径。



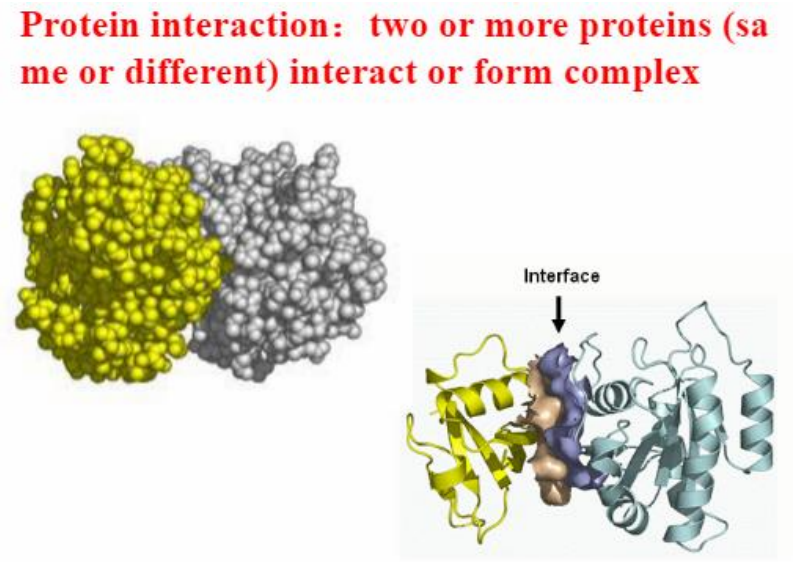
2.趋向性（P147）

趋向性(taxis)是一生物（或细胞）天生的行为反应，指其对一指向性刺激（由特定方向给的刺激），而会有趋进（正趋向性）或远离（负趋向性）刺激源的动作。

第七节 蛋白质-蛋白质相互作用网络

1、蛋白质的相互作用（P154）

蛋白质的相互作用是指蛋白质分子之间的相关性，并从生物化学、信号传导和遗传网络的角度研究这种相关性。



蛋白质相互作用在许多生物过程和研究防治疾病中发挥着非常重要的作用，其研究比基因网络更为复杂和困难。

第八节 虚拟细胞

无

第八章 合成生物学

第一节 合成生物学概述

1、细胞生物学（P169）

细胞生物学(Cell Biology)是在显微、亚显微和分子水平三个层次

上，研究细胞的结构、功能和各种生命规律的一门科学。细胞生物学由细胞学（Cytology）发展而来，Cytology 是关于细胞结构与功能(特别是染色体)的研究。现代细胞生物学从显微水平、超微水平和分子水平等不同层次研究细胞的结构、功能及生命活动。在我国基础学科发展规划中，细胞生物学与分子生物学、神经生物学和生态学并列为生命科学的四大基础学科。

2、分子生物学（P169）

分子生物学（molecular biology)是从分子水平研究生物大分子的结构与功能从而阐明生命现象本质的科学。自 20 世纪 50 年代以来，分子生物学是生物学的前沿与生长点，其主要研究领域包括蛋白质体系、蛋白质-核酸体系（中心是分子遗传学）和蛋白质-脂质体系（即生物膜）。1953 年沃森、克里克提出 DNA 分子的双螺旋结构模型是分子生物学诞生的标志。

3、终止子（P170）

终止子(terminator T)是给予 RNA 聚合酶转录终止信号的 DNA 序列。在一个操纵元中至少在构基因群最后一个基因的后面有一个终止子。

4、启动子（P170）

启动子是 RNA 聚合酶识别、结合和开始转录的一段 DNA 序列，

它含有 RNA 聚合酶特异性结合和转录起始所需的保守序列，启动子本身不被转录。启动子的特性最初是通过能增加或降低基因转录速率的突变而鉴定的。启动子一般位于转录起始位点的上游。

5、酶切位点（P170）

DNA 上一段碱基的特定序列，限制性内切酶能够识别出这个序列并在此将 DNA 序列切成两段。蛋白质酶切位点可为蛋白酶提供特定的作用位点，使蛋白质本身在该处断裂，转变成具有活性的形式，从而完成其生物学功能。

第二节 合成生物学基础研究经典实例

1、双稳态开关的基因线路设计（P172）

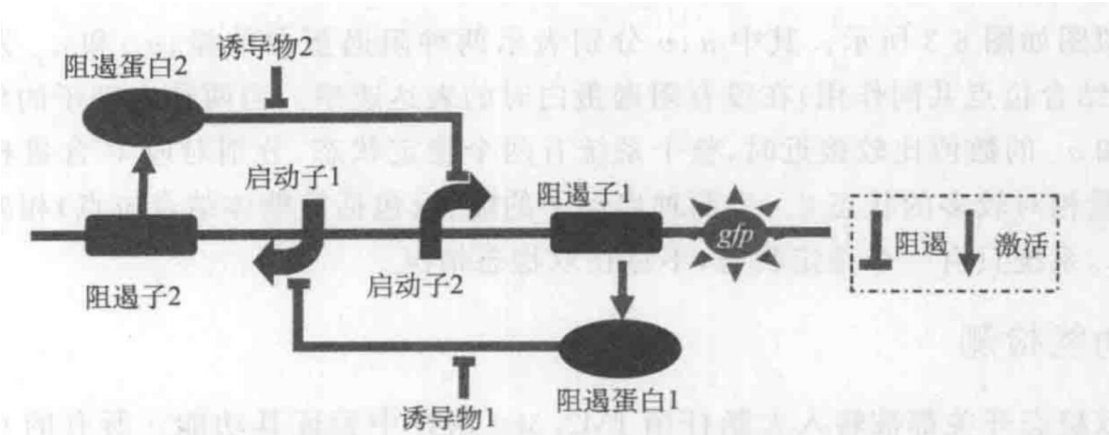


图 8-2 双稳态开关的基因线路设计(Gardner et al.,2000)

由上图可看出：

① 启动子 1 和启动子 2 分别控制着阻遏子 2 和阻遏子 1 基因序列的表达状态；

② 阻遏子 2 和阻遏子 1 基因表达的产物会分别阻遏启动子 2 和启动

子 1 的启动；

③ 整个系统有两种稳定的输出状态：

状态 A：启动子 1 启动而启动子 2 关闭的状态，此时，启动子 2 下游的报告基因无法开启，系统无特定的输出产物；

状态 B：启动子 2 启动而启动子 1 被关闭的状态，此时，启动子 2 下游的报告基因同时被开启，可以检测到报告基因的产物。

④ 此时，加入诱导子 1 和诱导子 2 可以实现系统在两种状态之间的转换，达到人工控制状态切换的目的。

当系统稳定于 A 状态，加入诱导子 2，可以解除阻遏子 2 对启动子 2 的阻遏作用，使阻遏子 1 表达，阻遏启动子 1，终止阻遏子 2 的表达，一段时间后系统从 A 状态翻转为 B 状态；

同理加入诱导子 1 作用一段时间，也可以使系统从状态 B 切换到状态 A。

2、振荡器（Repressilator）的基因线路设计（P174）

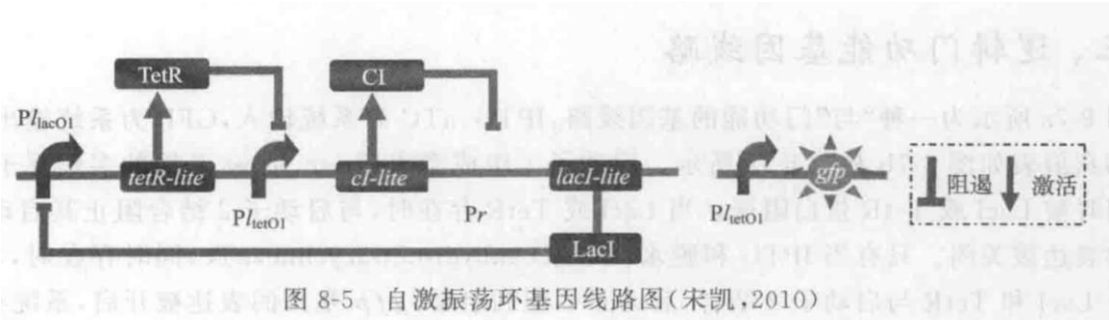


图 8-5 自激振荡环基因线路图(宋凯,2010)

上述振荡器的状态转换可具体表述如下：

① 当 Pl_{lacO1} 启动子启动时，产生的 TetR-lite 蛋白会阻遏下游 Pl_{tetO1} 启动子，CI-lite 蛋白无法产生，Pr 启动子的阻遏被解除，Pr 下游的基因

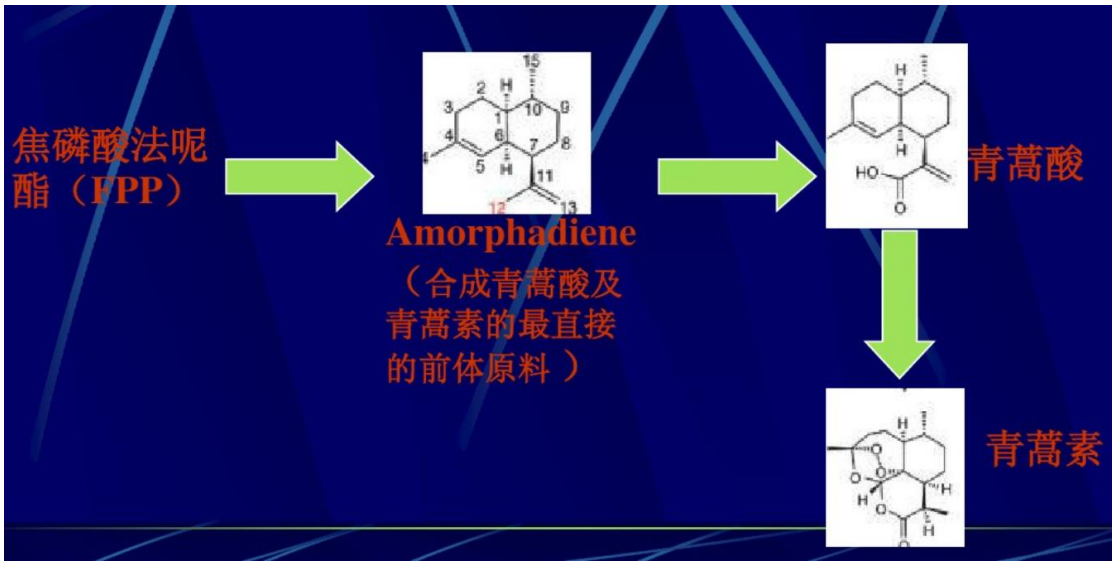
顺利表达，产生 LacI-lite 蛋白并阻遏下游 PI_{lacO1} 启动子，PI_{lacO1} 启动子的状态从“开启”转换为“关闭”；

② 下游基因停止产生 TetR-lite 蛋白，PI_{tetO1} 启动子的阻遏被解除，状态由“关闭”转换为“开启”，下游基因产生 CI-lite 蛋白去阻遏 PR 启动子，导致 PR 启动子的状态从“开启”转换为“关闭”；Pr 启动子的被阻遏使得 LacI-lite 蛋白无法继续产生，PI_{lacO1} 启动子的阻遏被消除，其状态由“关闭”翻转为“开启”；

③ 下游基因表达 TetR-lite 蛋白阻遏 PI_{tetO1} 启动子。如此往复，实现整个振荡器周而复始的状态转换。

第三节 合成生物学应用研究经典实例

1、Amorphadiene——紫穗槐二烯



第九章 分子进化与系统发育

第一节 分子进化与系统发育

1、生物大分子（P184）

生物大分子是指生物体细胞内存在的蛋白质、核酸、多糖等大分子。生物大分子的结构很复杂，但其基本的结构单元并不复杂。蛋白质分子是由氨基酸分子以一定的顺序排列成的长链。氨基酸分子是大部分生命物质的组成材料，不同的氨基酸分子有好几十种。生物体内的绝大多数酶就属于蛋白质，是生物体维持正常代谢功能所不可缺少的。

第二节 分子系统发育树的构建方法



第三节 系统发育树构建及应用

无。

第十章 统计学习与推理

第一节 统计学习与推理基础

第二节 统计模型与参数推断

第三节 聚类分析、主成分分析与 Fisher 判别

第四节 贝叶斯推理

第五节 隐马尔科夫模型

第六节 动态神经网络

第七节 支持向量机

第八节 MATLAB 的应用实例

第十一章 生物信息学编程基础

第一节 Linux 操作系统

第二节 生物信息学中的编程语言

第三节 SQL 及数据库编程

第四节 并行计算

第十二章 第二代测序技术及其应用

第一节 测序技术概述

1、第一代测序——Sanger 测序法（P288）

Sanger 法是根据核苷酸在某一固定的点开始，随机在某一个特定的碱基处终止，并且在每个碱基后面进行荧光标记，产生以 A、T、C、G 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 胶上电泳进行检测，从而获得可见 DNA 碱基序列的一种方法。

2、第二代测序技术的原理

在第二代测序中，片段化的基因组 DNA 两侧连上接头，随后运用不同的步骤来产生几百万个空间固定的 PCR 克隆序列。每个克隆由单个文库片段的多个拷贝组成，之后进行引物杂交和酶延伸反应。由于所有的克隆都在同一平面上，这些反应就能够大规模平行进行。同样地，每个延伸所掺入的荧光标记的成像检测也能同时进行，以获取测序数据。

3、第三代测序技术的原理（P289）

第三代测序技术原理主要分为两大技术阵营：

第一大阵营是单分子荧光测序，代表性的技术为美国螺旋生物(Helicos)的 SMS 技术和美国太平洋生物(Pacific Bioscience)的 SMRT 技术。脱氧核苷酸用荧光标记，显微镜可以实时记录荧光的强度变化。当荧光标记的脱氧核苷酸被掺入 DNA 链的时候，它的荧光就同时能

在 DNA 链上探测到。当它与 DNA 链形成化学键的时候，它的荧光基团就被 DNA 聚合酶切除，荧光消失。这种荧光标记的脱氧核苷酸不会影响 DNA 聚合酶的活性，并且在荧光被切除之后，合成的 DNA 链和天然的 DNA 链完全一样。

第二大阵营为纳米孔测序，代表性的公司为英国牛津纳米孔公司。新型纳米孔测序法（nanopore sequencing）是采用电泳技术，借助电泳驱动单个分子逐一通过纳米孔来实现测序的。由于纳米孔的直径非常细小，仅允许单个核酸聚合物通过，而 ATCG 单个碱基的带电性质不一样，通过电信号的差异就能检测出通过的碱基类别，从而实现测序。

3、第一代、第二代、第三代测序的优缺点（P289）

第一代测序技术以其可达 1000 bp 的测序读长、99.999%的高准确性帮助人们完成了大量的测序工作，但其测试速度慢、成本高、通量低。

第二代测序技术又称高通量测序技术、大量并行测序技术，以低成本、99% 以上的准确度，1 次可对几百、几千个样本的几十万至几百万条 DNA 分子同时进行快速测序分析。但是相对于第三代，都仍然需要扩增的方法放大信号，扩增后再检测。

第三代测序技术是指单分子测序技术。DNA 测序时，不需要经过 PCR 扩增，实现了对每一条 DNA 分子的单独测序。第三代测序技术也叫从头测序技术，即单分子实时 DNA 测序。核心特点是无需扩增

所以成本更低。

4、噬菌体（P288）

噬菌体（phage）是侵袭细菌的病毒，也是赋予宿主菌生物学性状的遗传物质。噬菌体必须在活菌内寄生，有严格的宿主特异性，其取决于噬菌体吸附器官和受体菌表面受体的分子结构和互补性。噬菌体是病毒中最为普遍和分布最广的群体。通常在一些充满细菌群落的地方，如：泥土、动物的肠道里，都可以找到噬菌体。

5、自噬

自噬是一个吞噬自身细胞质蛋白或细胞器并使其包被进入囊泡，并与溶酶体融合形成自噬溶酶体，降解其所包裹的内容物的过程，藉此实现细胞本身的代谢需要和某些细胞器的更新。

自噬在机体的生理和病理过程中都能见到，其所起的作用是正面还是负面的尚未完全阐明，对肿瘤的研究尤其如此，值得关注。

第二节 第二代测序原理

1、Roche454 测序仪的原理

①文库制备 :首先将待测的目的 DNA 分子打断成 300-800 bp 的片段，然后在 DNA 片段的 5′端加上一个磷酸基团，3′变成平端，在两端分别加上 44bp 的 A、B 两个衔接子，组成目的 DNA 的样品文库。加上这样两个衔接子 A、B 的目的是为了其在生物素和链霉亲和

素的作用下，同含有过量链霉亲和素的磁珠特异性结合。

②乳液 PCR 扩增：目的 DNA 片段固定到一个磁珠上之后，将磁珠包被在单个油水混合小滴(乳滴)，在这个乳滴里进行独立的扩增，而没有其他的竞争性或者污染性序列的影响，从而实现了所有目的 DNA 片段进行平行扩增乳滴 PCR。

③测序：将得到的携带有 DNA 扩增片段的珠子放入 PTP 板上，由于 PTP 孔直径较小，每个孔只允许结合一个磁珠。将 PTP 板置于 GSFLX 中开始测序。每当有核苷酸加入到 DNA 片段上，荧光素酶就会催化相关反应释放出相应强度的荧光信号，通过信号转换后可以准确快速的活的模板的碱基序列。

第三节 第二代测序技术的应用

1、DNA 甲基化

DNA 甲基化（DNA methylation）为 DNA 化学修饰的一种形式，能够在不改变 DNA 序列的前提下，改变遗传表现。

2、长非编码 RNA

长链非编码 RNA(Long non-coding RNA, lncRNA)是长度大于 200 个核苷酸的非编码 RNA。研究表明，lncRNA 在剂量补偿效应(Dosage compensationeffect)、表观遗传调控、细胞周期调控和细胞分化调控等众多生命活动中发挥重要作用，成为遗传学研究热点。

目前的非编码 RNA 测序，主要是针对小的非编码 RNA。因为

LncRNA的转录本多数没有polyA尾，而大部分RNA-seq是利用polyA尾来捕获转录本后构建文库。

3、小非编码 RNA

长度小于50 nt的非编码RNA称为小非编码RNA，包括microRNA，siRNA，piRNA。

其中，kb=千碱基 kilobase，nt = 核苷酸 nucleotide，bp = 碱基对 base pair。

4、靶基因

目的基因（gene of interest），也称靶标基因，是指在实验中研究或操纵的特定基因。在基因克隆过程中目的基因就是要分离、纯化、克隆并转化到生物体的那个可以带来预期表型性状如抗虫或耐除草剂的基因。

第四节 生物信息学在第二代测序中的应用

无。