系统生物学重点

第一章 系统生物学概论

1.系统生物学定义：

Hood，2004：系统生物学是研究一个生物系统中所有组成成分（基因、mRNA、蛋白质等）的构成，以及在特定条件下这些组分间的相互关系，并通过计算生物学建立一个数学模型来定量描述和预测生物功能、表型和行为的学科。

杨胜利，2004：系统生物学是在细胞、组织、器官和生物体水平上研究结构和功能各异的生物分子及其相互作用，并通过计算生物学定量阐明和预测生物功能、表型和行为。

系统生物学将在基因组测序基础上完成 DNA 序列到生命的过程，这是逐步整合、优化的过程，系统生物学的发展预计需要一个世纪或更长时期，因此常把系统生物学称为 21 世纪的生物学。

2.系统生物学内容：

湿实验：采用高通量实验技术，通过众多组学，在整体和动态研究水平上积累数据并在挖掘数据时发现新规律、新知识，提出新概念。

干实验：通过计算生物学建立生物模型，根据被研究的真实系统的模型，利用计算机进行实验研究。

3.系统生物学流程：

①研究组分，构建模型
②改变条件，观测变化
③比较结果，修订模型
④重新实验，继续修订

4.系统生物学方法：

整合（incorporation）：把系统内不同性质的构成要素（DNA、RNA、蛋白质和生物小分子等）或不同层次的构成要素整合在一起进行研究。

干涉（perturbation）：人为地设定某种或某些条件去作用于被实验的对象，从而研究特定的生命系统在不同时间和空间条件下具有的动力学特征。

5.系统生物学整合策略：

自下而上（hypothesis based）：使用独立的实验数据，适用于大多数基因和它们的调控关系相对比较清楚的情况。

自上而下（data driven）：利用高通量的 DNA 芯片和其他新的测试技术获得数据来研究。

混合使用：自下而上 + 自上而下

第二章 基因组学

1.基因（gene）：是编码某种特定多肽链、tRNA、rRNA 和 ncRNA 的 DNA 区段，是 DNA 上的功能单位。

2.基因组（genome）:是一种生物体或个体细胞所具有的一套完整的基因及其调控序列。

3.基因组学（genomics）:是研究基因组的结构组成、时序表达模式和功能，并提供有关生物物种及其细胞功能的进化信息，是研究生物基因组和如何利用基因的一门学问。

4.基因组学的主要工具和方法：生物信息学，遗传分析，基因表达测量，基因功能鉴定

5.基因组学的特点：强调进行细胞中全部基因及非编码区的整体性考查和系统性研究，从而全面揭示基因与基因间的相互关系、基因与非编码序列的关系、基因与基因组的相互关系。

6.结构基因组学（structural genomics）：是基因组学的一个重要组成部分和研究领域，它是一门通过基因作图、核苷酸序列分析确定基因组成、基因定位的学科。

7.功能基因组学（functional genomics）：又称为后基因组学（postgenomics），它是利用结构基因组学提供的信息和产物，在基因组或系统水平上全面分析基因的功能和相互作用。

功能基因组学是利用结构基因组学所获得的各种信息，建立与发展各种 技术和实验模型来测定基因及基因非编码序列的生物学功能。

8.比较基因组学（comparative genomics）：在基因组图谱和测序技术的基础上，对已知的基因特征和基因组结构进行比较以了解基因的功能、表达机制和不同物种亲缘关系的生物学研究。

比较基因组学的基础是相关生物基因组的相似性。全基因组比对是比较基因组学的经典方法。比较基因组学的研究成果催生了水平基因转移理论，支持细胞器起源的内共生学说。

9.元基因组或总体基因组学（metagenomics）：是一门直接取得环境中所有遗传物质的研究，意指直接研究环境中微生物群落基因组学的应用，而非于实验室中进行单一个体纯化与培养的实验方式。

10.测序方法：

①第一代

桑格测序法：测序过程需要先做一个聚合酶连锁反应。PCR 过程中，双脱氧核糖核苷酸可能随机的被加入到正在合成中的 DNA 片段里。由于双脱氧核糖核苷酸少了一个氧原子，一旦它被加入到 DNA 链上，这个 DNA 链就不能继续增加长度。最终的结果是获得所有可能获得的、不同长度的 DNA 片段。

目前最普遍最先进的方法，是将双脱氧核糖核苷酸进行不同荧光标记。将 PCR 反应获得的总 DNA 通过毛细管电泳分离，跑到最末端的 DNA 就可以在激光的作用下发出荧光。由于 ddATP, ddGTP, ddCTP, ddTTP（4 种双脱氧核糖核苷酸）荧光标记不同，计算机可以自动根据颜色判断该位置上碱基究竟是 A，T，G，C 中的哪一个。

原理：双脱氧链终止法采用 DNA 复制原理。Sanger 测序反应体系中包括目标 DNA 片断、脱氧三磷酸核苷酸（dNTP）、双脱氧三磷酸核苷酸（ddNTP）、测序引物及 DNA 聚合酶等。测序反应的核心就是其使用的 ddNTP：由于缺少 3’-OH 基团，不具有与另一个 dNTP 连接形成磷酸二酯键的能力，这些 ddNTP 可用来中止 DNA 链的延伸。此外，这些 ddNTP 上连接有放射性同位素或荧光标记基团，因此可以被自动化的仪器或凝胶成像系统所检测到。

优点：最长可测定 600-1000bp 的 DNA 片断；对重复序列和多聚序列的处理较好；序列准确性高，高达 99.999%

缺点：通量较低（在 24h 内可测定的 DNA 分子数一般不超过 10000 个）；每碱基测序成本较高；不适合大规模平行测序

②第二代

Roche/454——焦磷酸测序:是一种基于聚合原理的 DNA 测序方法，它依赖于核苷酸掺入中焦磷酸盐的释放，而非双脱氧三磷酸核苷酸参与的链终止反应。是由 4 种酶催化的同一反应体系中的酶级联化学发光反应。在每一轮测序中，只加入一种 dNTP，若该 dNTP 与模板配对，聚合酶就可以将其掺入到引物链中并释放出等摩尔数的焦磷酸基团(PPi)。PPi 可最终转化为可见光信号，并由 PyrogramTM 转化为一个峰值。每个峰值的高度与反应中掺入的核苷酸数目成正比。然后加入下一种 dNTP，继续 DNA 链的合成。

优点:读长长，使得后继的序列拼接工作更加高效、准确;速度快，一个测序反应耗时 10 个

小时，获得 4-6 亿个碱基对;特别适合从头拼接和宏基因组学应用，多用于新的细菌基因组
缺点:无法准确测量同聚物的长度，所以检测插入缺失突变的误差率高;通量小且费用高;对重测序来说太贵，不适合

Illumina/Solexa——边合成边测序（sequencing by synthesis，SBS）：以 DNA 单链为模板，在合成互补链的时候，利用带荧光标记的 dNTP 发出不同的荧光来确定碱基类型。这种测序技术通过将基因组 DNA 的随机片断附着到光学透明的表面，这些 DNA 片断通过延长和桥梁扩增，形成了具有数以亿计 cluster 的 Flowcell，每个 cluster 具有约 1000 拷贝的相同 DNA 模板，然后用 4 种末端被封闭的不同荧光标记的碱基进行边合成边测序。这种新方法确保了高精确度和真实的一个碱基接一个碱基的测序，排除了序列方面的特殊错误，能够测序同聚物和重复序列。

桥式扩增（bridge amplification）：随机打断的单链 DNA 片段通过两端接头与寡核苷酸的互补固定在芯片表面，形成桥形结构，之后以寡核苷酸为引物进行 PCR 扩增，得到单克隆的 DNA 簇群。

优点：通量大；测序方式灵活；分析软件多样化

缺点：样本制备过程复杂；样本要求相对较高

ABI/SOLiD——边连接边测序（sequencing by ligation）：基于连接酶法，即利用 DNA 连接酶在连接过程之中测序。SOLiD 连接反应的底物是 8 碱基单链荧光探针混合物（3'-XXnnnzzz-5'），其中第 1 和第 2 位碱基（XX）上的碱基是确定的，并根据种类的不同在 6-8 位（zzz）上加上 CY5、Texas Red、CY3、6-FAM 四种不同的荧光标记。这是 SOLiD 的独特测序法，两个碱基确定一个荧光信号，相当于一次能决定两个碱基，因此也称为两碱基测序法。当荧光探针能够与 DNA 模板链配对而连接上时，就会发出代表第 1、2 位碱基的荧光信号。在记录下荧光信号后，通过化学方法在第 5 和第 6 位碱基之间进行切割，这样就能移除荧光信号，以便进行下一个位 置的测序。这种测序方法每次测序的位置都相差 5 位：即第一次是第 1、2 位，第二次是第 6、7 位……在测到末尾后，要将新合成的链变性，洗脱。接着用引物 n-1 进行第二轮测序。引物 n-1 与引物 n 的区别是，二者在与接头配对的位置上相差一个碱基。也即是，通过引物 n-1 在引物 n 的基础上将测序位置往 3'端移动一个碱基位置，因而就能测定第 0、1 位和第 5、6 位……第二轮测序完成，依此类推，直至第五轮测序，最终可以完成所有位置的碱基测序，并且每个位置的碱基均被检测了两次。

优点：高准确性，每个 DNA 碱基检测 2 次，增加了序列读取的准确性

缺点：运行时间长，检测碱基替换突变的误差率高

11.NGS 数据库：SRA、GEO、TCGA

12.NGS 数据格式：

Reference sequences：FASTA、2bit

Reads：FASTQ (FASTA with quality scores)

Alignments：SAM (Sequence Alignment/Map format)、BAM (Binary version of SAM)

Features, annotation, coverage, scores：GFF3/GTF (General Feature Format, Gene Transfer Format) 、BED/bigBed (Browser Extensible Data) 、WIG/bigWig (Wiggle format) bedGraph

Variations：VCF (Variant Call Format)、BCF (Binary version of VCF)

13.深度（depth）：也叫乘数，衡量测序量的首要参数；测序得到的总碱基数与待测基因组大小的比值；每个碱基被测序的平均次数。

14.覆盖度（coverage）：测序获得的序列占整个基因组的比例。由于基因组中的高 GC、重复序列等复杂结构的存在，测序最终拼接组装获得的序列往往无法覆盖所有的区域，这部分没有获得的区域就称为 Gap。

15.质控工具：FastQC、NGS QC Toolkit 、SolexaQA

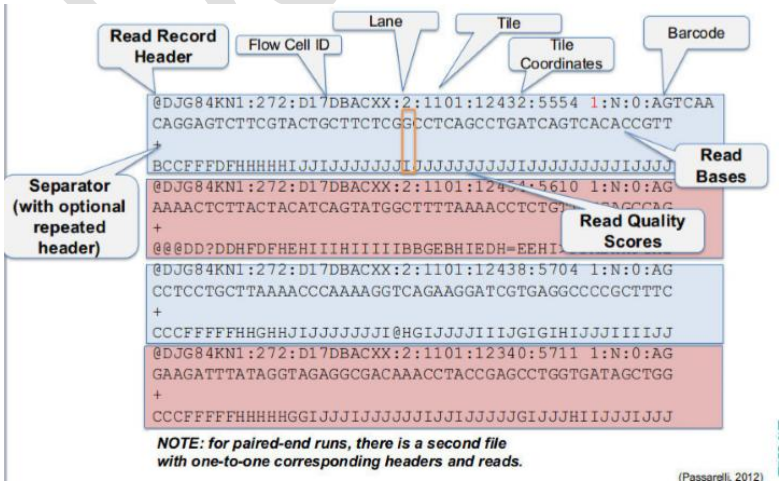
16.预处理工具：FASTX-Toolkit、PRINSEQ、cutadapt
17.比对工具（BWT 算法）：BWA、Bowtie、SOAP
18.提取变异工具：SAMtools、GATK、VarScan
19.注释变异工具：SnpEff、ANNOVAR、SeattleSeq Annotation、SIFT、PolyPhen-2
20.可视化工具：Genome Browser、IGV、Tablet、Circos
21.工作流：Bpipe、Galaxy、Taverna、BioX::Workflow
22.第三代测序技术：tSMS、SMRT、FRET、纳米孔测序、TEM
23.三代测序技术比较：

第一代：	第二代：	第三代：
优点： ✓准确 ✓读长长	优点： ✓通量高 ✓成本低 ✓时间短	优点： ✓无扩增 ✓直接观察 ✓速度快 ✓长读长
缺点： ✓通量低 ✓速度慢 ✓成本高	缺点： ✓读长短 ✓效率不一致	缺点： ✓错误率高 ✓可靠性差

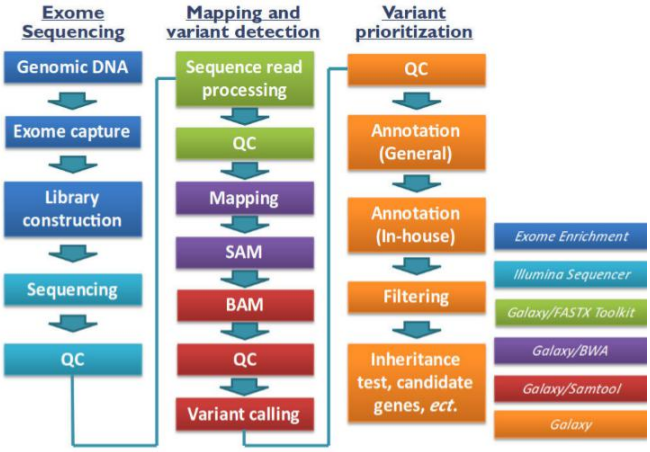
23.fastq 格式包含信息：

```
@SEQUENCE_ID1
ATGCGCGCGCGCGCGCGCGGGTAGCAGATGACGACACAGAGCGAGGATGCGCTGAGAGTA
GTGTGACGACGATGACGGAAAATCAGA
+
BBBBBPPPPPPXXXXX^^^^^^^^^^^^^^^^^^^^^^^^^__^^^^^^^^^^^^_eeeeeee
[[[[[^^^^]]]]XXXXXPPPPPBBBB
```

1. **Single line ID** with at symbol ("@") in the first column.
2. There should be not space between "@" symbol and the first letter of the identifier.
3. Sequences are in multiple lines after the ID line
4. Single line with plus symbol ("+") in the first column to represent the quality line.
5. Quality ID line can have or have not ID
6. Quality values are in multiple lines after the + line



24.外显子组测序数据分析流程：



第三章 转录组学

1.基因表达（gene expression）：用基因中的信息来合成基因产物的过程。

2.基因表达谱（gene expression profile）：一种在分子生物学领域，借助 cDNA、表达序列标签（EST）或寡核苷酸芯片来测定细胞基因表达情况（包括特定基因是否表达、表达丰度、不同组织、不同发育阶段以及不同生理状态下的表达差异）的方法。

3.转录组（transcriptome）：也称为“转录物组”，广义上指在相同环境（或生理条件）下的在一个细胞、或一群细胞中所能转录出的所有 RNA 的总和，包括信使 RNA（mRNA）、核糖体 RNA（rRNA）、转运 RNA（tRNA）及非编码 RNA；狭义上则指细胞所能转录出的所有信使 RNA（mRNA）。

4.转录组学（transcriptomics）：对转录水平上发生的事件及其相互关系和意义进行整体研究的一门学科，负责研究在单个细胞或一个细胞群的特定细胞类型内所生产的 mRNA 分子。

5.转录组学研究内容：对特定细胞的转录与加工机制进行研究；对转录物编制目录便于进一步归类研究；绘制动态的转录物图形；转录物调控网络

6.转录组学研究方法：EST、SAGE、MPSS、Microarray、RNA-Seq

7.RNA 测序（RNAsequencing，简称 RNA-Seq，也被称为全转录物组鸟枪法测序，Whole Transcriptome Shotgun Sequencing，简称 WTSS）：基于第二代测序技术的转录组学研究方法。RNA 测序是使用第二代测序的能力，在给定时刻从一个基因组中，揭示 RNA 的存在和数量的一个快照的技术。

首先提取生物样品的全部转录的 RNA，然后反转录为 cDNA 后进行二代高通量测序，在此基础上进行片段的重叠组装，从而可得到一个个的转录本。

8.RNA-Seq 转录本组装的两种思路：genome-guided 、de novo

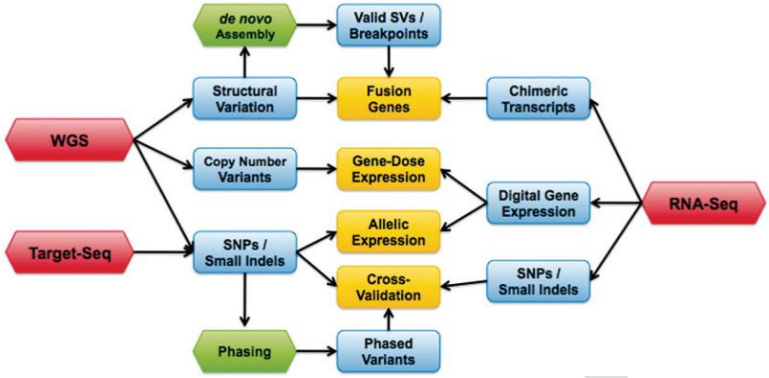
9. RNA-Seq Application：

- **Annotation**: Identify novel genes, transcripts, exons, splicing events, ncRNAs
- Detecting RNA editing and SNPs
- **Measurements: RNA quantification and differential gene expression**

10.RNA-Seq 分析工具：

Quality control：FastQC、NGSQC、RNA-SeQC、RSeQC

Trimming and adapters removal：FASTX、PRINSEQ、cutadapt
Alignment：TopHat
Transcriptome assemblers：Cufflinks、Scripture
Expression：Cufflinks/Cuffdiff、DESeq、EdgeR、DEGseq、baySeq
Workbench：easyRNASeq、Galaxy、GenePattern、Taverna
Visualization：ngs.plot、GBrowse、IGB、IGV、SeqMonk
Databases：ENCODE、RNA-Seq Atlas、SRA
11.顺反组（cistrome）：用于定义一个反式（trans）调控因子在基因组 （genome）范围内的作用对象——一组顺式 (cis) 作用元素。
研究方法：染色质免疫沉淀-测序（ChIP-seq）
12.RNA-Seq 分析流程：



13.Tuxedo 工具套装：

Tool		Tool description
Bowtie		Ultrafast short read aligner
Tophat		Aligns RNA-seq reads to the genome using Bowtie. Discovers splice sites
Cufflinks package	Cufflinks	Assembles transcripts
	Cuffcompare	Compares transcript assemblies to annotation
	Cuffmerge	Merges two or more transcript assemblies
	Cuffdiff	Finds differentially expressed genes and transcripts. Detects differential splicing and promoter use

14.R 中差异表达分析的工具：

Differential gene expression analysis
edgeR

- edgeR [12]
  - Differential gene expression analysis
  - Free R Package (GPL2)
  - Galaxy wrapper does normalizations for you
    - Use raw reads, do NOT use FPKM/RPKM!
- "Limma" for count data
  - Not Gaussian (normal) distributed like e.g. micro-array data – but negative binomial

15.RPKM：

RPKM:
Reads Per Kilobase and Million mapped reads

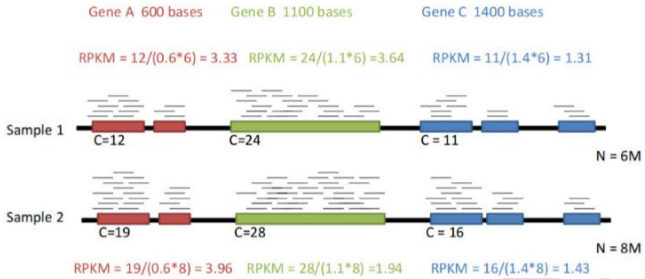
Unit of measurement

$$RPKM = \#MappedReads * \frac{1000bases * 10^6}{length\ of\ transcript * Total\ number\ of\ mapped\ reads}$$

- RPKM reflects the molar concentration of a transcript in the starting sample by normalizing for
  - RNA length
  - Total read number in the measurement
- This facilitates transparent comparison of transcript levels within and between samples

$$RPKM = \frac{number\ of\ reads\ of\ the\ region}{\frac{total\ reads}{1{,}000{,}000}\ x\ \frac{region\ length}{1{,}000}}$$

RPKM Example

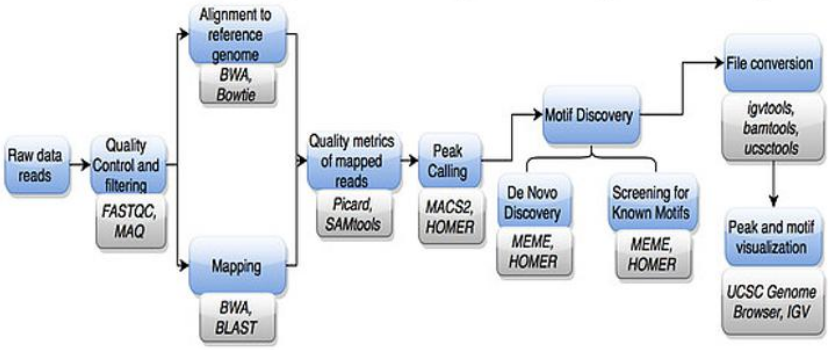


16.ChIP-Seq：

提取 Peak：PeakSeq

处理 Motif：HOMER

Motif Discovery and Analysis Using ChIP-seq



第四章 蛋白质组学

1.蛋白质组(Proteome)：由一个基因组或一个细胞、组织表达的所有蛋白质。

特点：

①对应于基因组的所有蛋白质构成的整体，不是局限于一个或几个蛋白质。

②同一基因组在不同细胞、不同组织中的表达情况各不相同。

③在空间和时间上动态变化着的整体。

2.蛋白质组学(proteomics)：在蛋白质水平上定量、动态、整体性地研究生物体。（分为结构蛋白质组学和功能蛋白质组学）

主要研究内容：

①了解某种特定的细胞、组织或器官制造的蛋白质种类。

②明确各种蛋白质分子是如何形成类似于电路的网络的。

③描绘蛋白质的精确三维结构，揭示其结构上的关键部位，如与药物结合并且决定其活性的部位。

3.功能蛋白质组：细胞在一定阶段或与某一生理现象相关的所有蛋白质。

4.蛋白组学技术的优点：大规模蛋白质分离；高通量蛋白质鉴定。

5.蛋白质组学研究的主要方面：

组成性蛋白质组学:对某个体系的蛋白质进行鉴定并详细阐述其翻译后修饰的特性。

比较蛋白质组学:以重要生命过程或人类重大疾病为对象，进行重要生理和病理体系或过程的蛋白质表达的比较。

相互作用蛋白质组学:通过各种先进技术研究蛋白质之间的相互作用，绘制某个体系的蛋白质作用的网络图谱。

6.蛋白质组表达模式的研究方法：

①双向凝胶电泳(two-dimensional electrophoresis，2-DE)：利用蛋白质的等电点和分子量，结合凝胶化学特性，分离各种蛋白质的方法。

原理：第一向在高压电场下对蛋白质进行等电聚焦(IEF), 再在第一向垂直方向上进行第二向SDS-聚丙烯酰胺凝胶电泳(SDS-PAGE)。

特点：可分离 10～100 kD 分子量的蛋白质；高灵敏度和高分辨率；便于计算机进行图像分析处理；与质谱分析匹配

缺点：极酸、极碱性蛋白质，疏水性蛋白质，极大蛋白质、极小蛋白质以及低丰度蛋白质用此种技术难于有效分离；胶内酶解过程费时、费力，难于与质谱联用实现自动化。

新型非凝胶技术：液相色谱法（LC）、毛细管电泳（CE）、液质联用技术（LC-MS/MS）、多维色谱技术（LC/LC-MS/MS）

②质谱（MS）法：

基本原理：样品分子离子化后，根据离子间质荷比（m/z）的差异来分离并确定样品的分子量。

分类：基质辅助激光解吸/电离飞行时间质谱（MALDI-TOF MS）、电喷雾质谱（ESI-MS）

鉴定和注释蛋白质的路线：通过肽质谱指纹图（peptide mass fingerprinting，PMF）和数据库搜寻匹配；通过测出样品中部分肽段二级质谱信息或氨基酸序列标签和数据库搜寻匹配。

③生物信息学：构建和分析双向凝胶电泳图谱；数据库的搜索与构建

数据库：SWISS-PROT、dbEST

7.定量蛋白质组学研究：把一个基因组表达的全部蛋白质或一个复杂体系中所有的蛋白质进行精确的定量和鉴定。

①基于双向凝胶电泳的定量蛋白质组研究策略：双向凝胶电泳(2-DE)、荧光差异显示双向电泳（F-2D-DIGE）

②基于生物质谱的定量蛋白质组研究策略：对于具有相同离子化能力的蛋白质或多肽，可以通过比较质谱峰的强度（或峰面积）得到待比较蛋白质的相对量。

8.蛋白质组功能模式的研究方法：

研究目标：揭示蛋白质组成员间的相互作用、相互协调的关系，并深入了解蛋白质的结构与功能的相互关系，以及基因结构与蛋白质结构功能的关系。

蛋白质相互作用研究技术：酵母双杂交系统、噬菌体展示技术、亲和层析耦联质谱技术、细胞内蛋白质共定位、免疫共沉淀耦联质谱技术、等离子表面共振技术、蛋白质芯片

第五章 系统生物学分析技术

1.泡利不相容原理(Pauli exclusion principle)：原子中不能有两个或两个以上的电子处于完全相同的状态。

2.在原子中完全确定一个电子的状态需要四个量子数，即**主量子数，角量子数，磁量子数和自旋量子数。**

3.核磁共振与紫外光谱、红外光谱的比较：

共同点：吸收光谱

不同点：

	紫外-可见	红外	核磁共振
吸收能量	紫外可见光 200 ~750nm	红外光 0.75 ~1000μm	无线电波1~100m
跃迁类型	电子能级跃迁	振转能级跃迁	自旋原子核能级跃迁

4.核磁共振：处于低能级状态的原子核吸收电磁辐射能量而跃迁至高能级状态。

5.化学位移（chemical shift）：因核所处化学环境改变而引起的共振条件（核的共振频率或外磁场强度）变化的现象。

6.质谱原理：用外力使较大的分子破裂，每一部分都带有电荷，因而其在电场或磁场中运动规律不同，可以被分离检测。

7.核磁共振波谱法（NMR）：利用核磁共振光谱进行结构测定，定性与定量分析的方法。

优点：

①迅速、准确、分辨率高。

②分析测定时，可深入物质内部而不破坏样品，属于无破损分析方法。

③确定氢原子在有机物分子中的位置、各种官能团和母核骨架上氢原子的相对数目以及空间结构等。

7.分子离子（奇电子离子)：化合物分子失去一个外层电子而形成的带正电荷的离子。（一般质谱图上质荷比最大的峰为分子离子峰。）

8.碎片离子：有机化合物在电场作用下，其结构裂解，产生各种“碎片”离子。

9.

红外光谱特点

1）红外吸收只有振-转跃迁，能量低；

2）应用范围广：除单原子分子及单核分子外，几乎所有有机物均有红外吸收；

3）分子结构更为精细的表征：通过*IR*谱的波数位置、波峰数目及强度确定分子基团、分子结构；

4）定量分析；

5）固、液、气态样均可用，且用量少、不破坏样品；

6）分析速度快；

7）与色谱等联用（*GC-FTIR*）具有强大的定性功能。

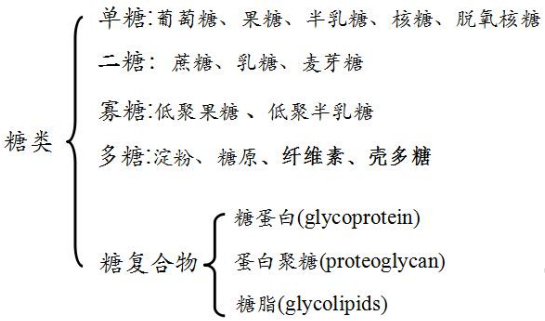
10.色谱原理：溶于流动相中的各组分经过固定相时，由于与固定相发生作用（吸附、分配、离子吸引、排阻、亲和）的大小、强弱不同，在固定相中滞留时间不同，从而先后从固定相中流出。

第六章 糖组学

1.糖组(glycome):生物体内所有游离糖分子及其复合物的集合。

2.糖组学(glycomics):研究生物体内所有游离糖分子及其复合物的组成和功能的学科。

3.糖的种类：



超过85%的蛋白质都存在糖基化。

4.糖的生物学作用：

①生物体的结构成分
②生物体内的主要能源物质
③生物体内合成其他物质的碳源
④生物体内的信息分子和调节分子
⑤糖链影响糖蛋白的分泌和稳定性
⑥与血浆中衰老蛋白的清除有关
⑦与细胞粘着有关

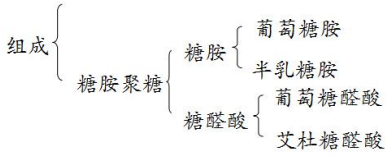
5.糖蛋白中聚糖与蛋白质的主要连接(糖基化)类型：N-连接（高甘露糖型、复杂型、杂合型）、O-连接

6.糖蛋白分子中聚糖的功能：

①聚糖可影响糖蛋白生物活性
②糖蛋白聚糖加工可参与新生肽链折叠
③糖蛋白聚糖可参与维系亚基聚合
④聚糖对蛋白质在细胞内的分拣、投送和分泌中的作用
⑤糖蛋白聚糖具有分子间的识别作用

7.蛋白聚糖（Proteoglycan）：一条或多条糖胺聚糖以共价键与核心蛋白形成的大分子糖复合物化合物。

8.蛋白聚糖的组成：



9.蛋白聚糖的分类：大分子聚集型胞外基质蛋白聚糖；小分子富含亮氨酸胞外基质蛋白聚糖；跨膜胞内蛋白聚糖

10.糖组学研究策略：

①分析单物种生物所产生的所有聚糖

②以糖肽为研究对象确认编码糖蛋白的基因

③结合有效的理化和生化性质,研究糖蛋白糖链的性质

研究内容：结构糖组学；功能糖组学；生物信息学

步骤：

①糖蛋白的提取和分离纯化

②糖链的释放（化学法、酶释放法）

③糖链的分离纯化（层析法、电泳法）

④糖链结构的解析（荧光糖电泳、质谱分析法、核磁共振技术）

第七章 脂质组学

1.脂质:生物体内低溶于水,高溶于非极性溶剂并能为机体利用的有机化合物。脂质可以简单地分为脂肪和类脂。

2.脂类的组成、分布、生理功能：

分类	含量	分布	生理功能
脂肪	95%	脂肪组织、血浆	1. 储脂供能 2. 提供必需脂酸 3. 促脂溶性维生素吸收 4. 热垫作用 5. 保护垫作用 6. 构成血浆脂蛋白
类脂	5%	生物膜、神经、血浆	1. 维持生物膜的结构和功能 2. 胆固醇可转变成类固醇激 素、维生素、胆汁酸等 3. 构成血浆脂蛋白

3.脂质组：细胞、组织、器官或生物个体内的全部脂类物质的 集合。

4.脂质组学：对生物样品中脂质总体进行定性和定量分析的学科，研究各种脂质分子的结构、功能，以及脂质分子与脂质分子、蛋白质分子或其他代谢物之间的相互作用。它通过比较不同生理状态下脂代谢状况，识别代谢调控中关键的脂生物标志物，阐明脂质在各种生命活动中的作用机制。在疾病研究中，脂质组学的主要目的是发现某种疾病的脂质生物标记物特性，或作为内部或外部干扰的标记物。

第八章 代谢组学

1.代谢(metabolism)：生物体内所发生的用于维持生命的一系列有序的化学反应的总称，包

括物质代谢和能量代谢两个方面。

作用：为生物体的生长和繁殖提供物质及能量、保持生物体的结构并对外界环境做出反应、消除或排泄体内不需要的废弃物。

分类：

分解代谢--对较大的分子进行分解以获得能量

合成代谢--利用能量来合成生命活动所需的组分

场所：

原核生物--细胞质

真核生物—线粒体、叶绿体、内质网、细胞质等

2.代谢组(metabolome)：一个生物样品内所有小分子物质的集合。生物样品可以是细胞、细胞器、组织、器官、组织提取物、生物体液或整个生物体。

3.一次代谢物(Primary metabolite)：与生物体的生长、发育和繁殖等直接相关的代谢物，又称为中心代谢物；通常存在于多种生物或同一生物体的不同器官、组织中；比如乳酸、部分氨基酸、某些维生素等。

二次代谢物(Secondary metabolite)：与生物体的生长、发育和繁殖等不直接相关的代谢物，缺乏时不会很快导致个体死亡或其他明显后果，但长期缺乏有可能有不良后果；通常存在于亲缘关系较近的物种中；比如某些植物产生的生物碱、酚类。

4.代谢组学研究内容：

①代谢物靶标分析：直接研究基因、蛋白质表达变化对代谢物的影响，比如分析特定底物与相应编码基因和蛋白质之间的关系。

优点：可以详细而精确地完成某个或某几个特定组分的分析。

困难：容易被其他相似的混合物所干扰。

②代谢轮廓分析（Metabolic profiling）： 代谢轮廓分析是少数预设的一些代谢产物的识别和定量分析，如某一类结构、性质相关的化合物（氨基酸、有机酸、顺二醇类）或某一代谢途径的所有中间产物或多条代谢途径的标志性组分。

③代谢组学（Metabolomics）：代谢组学分析揭示了研究的生物学系统中对限定条件下的特定生物样品中所有代谢组分的全面的定性和定量变化。

④代谢物指纹分析（Metabolic fingerprinting）：物图谱有其特质性，类似样品的“指纹”一样；对这种特质性进行区分、鉴定，被称为“代谢指纹分析”，帮助找出机体代谢的共性与个性。

目的：根据其来源和生物学相关性对样品进行快速分类。

5.代谢组学研究的平台与技术：

红外线光谱技术（IR），
核磁共振技术（NMR），
薄层色谱技术（TLC），
高效液相色谱技术（HPLC），
高效毛细管电泳技术（HPCE），
毛细管电泳与紫外线吸光率检测连用技术（CE/UV），
毛细管电泳与激光诱导荧光检测连用技术（CE/LIF），
毛细管电泳与质谱连用技术（CE/MS），
气相色谱与质谱共用技术（GC/MS），
液相色谱与质谱共用技术（LC/MS），
液相色谱与质谱先后使用技术（LC/MS/MS），
高效液相色谱与质谱和核磁共振技术功用（LC/NMR/MS）

6.代谢组学的主要应用：

研究基因功能

发掘疾病新的诊断方法

研究疾病的发病机制
研究疾病发生、发展、恢复的进程
开发新药

第九章 相互作用组学

1.定义：

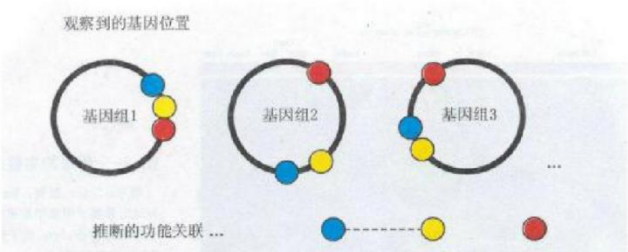
- ***相互作用组***
  - 体内各种分子相互作用的整体被称为相互作用组.
- ***相互作用组学***
  - 相互作用组学系统地研究各种分子相互作用，包括蛋白质-蛋白质、蛋白质-核酸、蛋白质-代谢物等的相互作用和这些作用形成的分子机制、途径和网络。
  - 相互作用组学研究可用于构建生物系统中的各种途径和网络，鉴别参与网络和途径的生物元件，形成系统生物学研究中的模块，进一步通过模块的相互作用研究构建完整的生命活动线路图。

2.研究蛋白质-蛋白质直接相互作用：

免疫共沉淀和抗原抗体结合
亲和作用
酵母双杂交和噬菌体展示
蛋白质组学技术

3.研究蛋白质-蛋白质相互作用的生物信息学技术：

- **系统发育谱法简介**
  - 通过系统发育谱，把功能相关的蛋白很好的聚类在一起，从而进行功能注释。
  - 它是一种基于非同源性的方法，基于的原理是功能相关的基因在进化过程中通常是一起被遗传下来或是同时都被丢失的，所以可以通过基因在物种间的分布模式来进行基因功能的预测。
- **基因邻居法**
  - 在原核生物中，如果两个基因在一个共同的操纵子内，且在不同的其他基因组也出现相邻现象，就可以推断它们编码的蛋白质之间具有功能联系。



域融合分析法（domain fusion analysis, DFA）

- 这是一种应用纯计算机分析技术从基因组序列中识别蛋白质与蛋白质之间的相互作用的方法
- 例如：在甲基因组中有分别编码A、B两种蛋白质的基因，而在乙基因组中发现A和B的基因融合成一条单链，这条单链就叫罗塞达碑序列（rosetta stone sequence），那么可以认为A、B是甲基因组中相互作用的两种蛋白质。

4.遗传相互作用分析：

遗传学阵列分析技术

- 获得感兴趣的目标基因的突变
- 将其它多种基因的缺失突变制成微阵列
- 将二者进行杂交，从而形成双突变体阵列
- 随后根据特异性表型进行打分
- 双突变联合作用引起细胞死亡或生存能力下降，为致死性或致病相互作用。
- 可以通过识别某一基因表达的蛋白质，受到另一基因表达产物的影响，说明它们参与了相同的重要生物过程。

第十章 表型组学

1.表型(phenotype)：又称表现型，指生物个体具有的可观测的特征或性状，包括所有肉眼可看到的形态特征和可用物理化学方法测定的与形态、发育、行为等相关的生理生化特性。

2.基因型与表现型之间的关系：

①基因型决定表现型

②表现型相同，基因型未必相同

③通常表现型与基因型间关系复杂

3.表现型的复杂性：

①同一表型的多样性

②同一个体表型的广泛性

③表现型受环境影响

4.表型组(phenome):是指细胞、组织、器官、或整个生物体的的全部可观测的性状特征。

5.表型组学（phenomics）：是系统研究某一生物、组织、器官或细胞在各种不同环境条件下所有表型的学科。

6.研究表型及表型组学的意义：

①表现型虽然主要由基因型决定，但是表现型是决定生物生存、繁殖等能力和机会的直接因素，也是自然选择或人工选择的主要依据。

②研究表现型，可以更好地认识基因型的作用。

③更好地认识基因型、表现型及环境之间的关系。

7.表型组学的挑战：

①表型本身的复杂性和动态变化，研究者通常只专注于少数几个表型，进行静态粗略的研究。

②传统的表型研究效率很低，不同研究者具有主观性，导致研究结果偏差大。

③表型研究技术发展相对滞后，导致表型研究严重滞后于各种组学研究。

8.基因型、表型和环境三者间的关系是表型组学的主要研究内容。

9.表型组学的研究方法：

①成像技术

②细胞芯片和组织芯片

③生物信息学技术

第十一章 系统生物建模

1.研究系统的策略：

①还原主义(Reductionism)：通过把一个整体分解为相关但相对独立且较简单的多个部分，对其分别加以研究。

②整体主义(Holism)：把事物作为整体来研究其特点和功能。

③分解与综合策略的统一，事物的部分特性，需要从整体层次认识；其他特性，则需要对各个部分深入研究分析，再从系统层次进行综合。

2.生物系统的特点：系统尺度及时间跨度的多样性；系统模型的多样性

3.系统生物建模基础：确定研究对象；系统中的组成部分；微分方程及方程组

4.反馈：又称回馈，指将系统的输出返回到输入端并以某种方式改变输入，进而影响系统功能的过程 。

5.单分子体系：

假设一个体系中，A物质的分子以恒定的速率常数生成和降解：

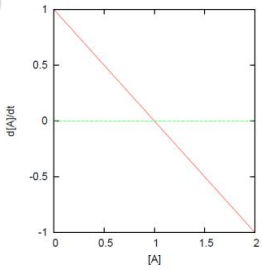
$$\emptyset \underset{k_2}{\overset{k_1}{\rightleftharpoons}} A$$

其中k_1、k_2反应速率常数，表示化学反应进行的快慢，用单位时间内反应物浓度的减少或生成物浓度的增加量来表示，在此处分别是A物质生成和降解的速率，均为恒定值。显然，二者均为正数。则系统中A物质的浓度可以表示为：

$$\frac{d[A]}{dt} = k_1 - k_2[A]$$

当 $d[A]/dt = 0$ 时，A物质的浓度不随时间变化，有：

$$[A] = k_1/k_2$$



横轴为浓度，各点对应浓度为数值的k_1/k_2倍；纵轴为浓度对时间的导数，为对应数值的k_1倍。

图中，[A]<k_1/k_2时，导数为正数，即系统中A物质的浓度在增加，但随着[A]的增大导数值趋近零；[A]>$k1/k2$时，导数为负数，即系统中A物质的浓度在减小，但随着[A]的减少导数值趋近零。

也就是说，不论初始浓度如何，结果足够长的时间，[A]都将不再随时间变化，达到系统稳定状态。

6.自催化体系：

假设在一个系统中，X物质可以与A物质反应，生成X，即

$$A + X \underset{k_2}{\overset{k_1}{\rightleftharpoons}} 2X$$

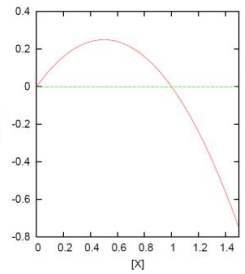
则：$$\frac{d[X]}{dt} = k_1[A][X] - k_2[X]^2$$

假定系统中A物质足够多，其浓度可以认为在反应中不变化，令$K = k_1[A]$，则：

$$\frac{d[X]}{dt} = k_1[A][X] - k_2[X]^2 = K[X] - k_2[X]^2 = [X](K - k_2[X])$$

当$d[X]/dt = 0$ 时，X物质的浓度不随时间变化，有：

$$[X] = 0 \quad 或 \quad [X] = K/k_2$$



横轴为浓度，各点对应浓度为数值的K/k_2倍；纵轴为浓度对时间的导数，为对应数值的K倍。

图中，[X] = 0时，导数为零，但是系统不稳定；[X] = K/$k2$时，系统稳定。

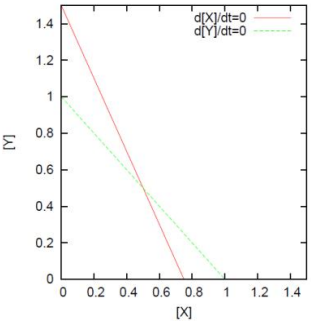
7.双分子自催化体系：

假设在一个系统中，X物质可以自A物质生成，Y物质可以自B物质生成，且X与Y可以形成复合物C，即：

$$A + X \underset{k_2}{\overset{k_1}{\rightleftharpoons}} 2X,$$
$$B + Y \underset{k_5}{\overset{k_4}{\rightleftharpoons}} 2Y,$$
$$X + Y \xrightarrow{k_3} C,$$

假定系统中A物质和B物质足够多，其浓度可以认为在反应中不变化，令$K_1 = k_1[A]$， $K_2 = k_4[B]$则：

$$\begin{aligned}\frac{d[X]}{dt} &= k_1[A][X] - k_2[X]^2 - k_3[X][Y] = K_1[X] - k_2[X]^2 - k_3[X][Y] \\ &= [X](K_1 - k_2[X] - k_3[Y]), \\ \frac{d[Y]}{dt} &= k_4[B][Y] - k_5[Y]^2 - k_3[X][Y] = K_2[Y] - k_5[Y]^2 \quad -k_3[X][Y] \\ &= [Y](K_2 - k_5[Y] - k_3[X]),\end{aligned}$$



系统中唯一稳定的点为两条直线的交点

当$d[X]/dt = 0$时，X物质的浓度不随时间变化，有：

$$[X] = 0 \quad 或 \quad [X] = (K_1 - k_3[Y])/k_2$$

当[X] = 0时，X物质的浓度不随时间变化，但[Y]可以为≥0的任意值；因此，对该系统不存在使浓度导数为零的确定的点，而是存在两条零渐进线。

同样，当$d[Y]/dt = 0$时，Y物质的浓度不随时间变化，有：

$$[Y] = 0 \quad 或 \quad [Y] = (K_2 - k_3[X])/k_5$$

当[Y] = 0时，Y物质的浓度不随时间变化，但[X]可以为≥0的任意值；因此，对该系统不存在使浓度导数为零的确定的点，而是存在两条零渐进线。

8.米氏方程：

生化反应的一般形式：

$$s_1S_1 + s_2S_2 + \ldots \xrightarrow{k_f} p_1P_1 + p_2P_2 + \ldots$$

S_1、S_2…为反应物，P_1、P_2…为产物， s_1、s_2…以及 p_1、p2…为计量化学系数

对下列反应：

$$A + B \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C.$$

$$\frac{d[A]}{dt} = \frac{d[B]}{dt} = -\frac{d[C]}{dt} = -k_f[A][B] + k_b[C]$$

若 $\frac{d[A]}{dt} = \frac{d[B]}{dt} = -\frac{d[C]}{dt} = 0$，则有 $[C]/[A][B] = k_f/k_b$

那么可以得到：

$$K = k_1/(k_2 + k_3) = [SE]/[S][E]$$

进一步有：

$$\begin{aligned}[SE] = K[S][E] &= K[S](E_0 - [SE]) \\ [SE]\,(1 + K[S]) &= KE_0[S] \\ [SE] = \frac{KE_0[S]}{1 + K[S]} &= \frac{E_0[S]}{(1/K + [S])}\end{aligned}$$

对反应产物P，则有：

$$\frac{d[P]}{dt} = k_3[SE] = \frac{k_3E_0[S]}{1/K + [S]}$$

上述方程通常写为：

$$\frac{d[P]}{dt} = \frac{V_{max}[S]}{K_m + [S]} \qquad \begin{cases} V_{max} = k_3E_0 \\ K_m = 1/K \end{cases}$$

Michaelis–Menten方程

考虑一个酶催化的反应：

$$S + E \underset{k_2}{\overset{k_1}{\rightleftharpoons}} SE \xrightarrow{k_3} P + E$$

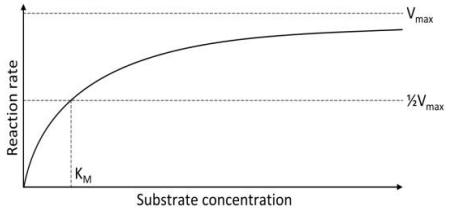
则有：

$$\begin{aligned}\frac{d[S]}{dt} &= -k_1[S][E] + k_2[SE] \\ \frac{d[E]}{dt} &= -k_1[S][E] + k_2[SE] + k_3[SE] \\ \frac{d[SE]}{dt} &= k_1[S][E] - k_2[SE] - k_3[SE] \\ \frac{d[P]}{dt} &= k_3[SE]\end{aligned}$$

第一步反应 $S + E \underset{k_2}{\overset{k_1}{\rightleftharpoons}} SE$ 通常速度很快且基本处于平衡状态，

可以假定： $d[SE]/dt \approx 0$

V_{max}及K_m的意义如下所示：



上述方程假定酶的浓度为恒定的，如果酶浓度动态变化，则：

$$\frac{d[P]}{dt} = \frac{V'_{max}[S][E]}{K_m + [S]}$$

9.蛋白质的激活与失活：

假定蛋白质X*在酶催化下被激活为X，X又可以被酶催化失活变为X*；其中催化激活酶的浓度是动态变化的，而催化失活酶的浓度则近似恒定。则：

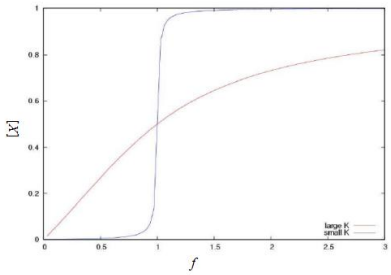
$$X^* \underset{(V_1,K_1)}{\overset{(V_2E,K_2)}{\rightleftharpoons}} X$$

假定蛋白质的总浓度固定，记为[X*]+[X]=1，有

$$\frac{d[X]}{dt} = -\frac{V_1[X]}{K_1+[X]} + \frac{V_2[E](1-[X])}{K_2+(1-[X])}$$

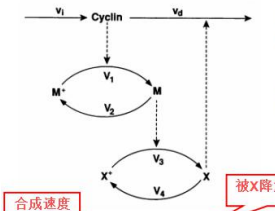
在系统处于平衡状态时，有 $\frac{d[X]}{dt} = 0$

$$\frac{V_1[X]}{K_1+[X]} = \frac{V_2[E](1-[X])}{K_2+(1-[X])}$$



可以看到，K值较大时，激活过程相当于一个信号放大器；而K较小时，则相当于一个开关。

10.细胞周期模型：



在细胞周期模型中，细胞周期蛋白cyclin (C)以固定的速率被合成。它可以诱导细胞周期蛋白激酶M的激活，而激活的M又诱导细胞周期蛋白酶X的激活。激活的X则会降解细胞周期蛋白C，从而形成一个反馈循环。

合成速度　被X降解速度　被其他因素降解速度

$$\frac{dC}{dt} = v_i - v_d X\frac{C}{K_d+C} - k_d C$$

$$\frac{dM}{dt} = V_1 C\frac{(1-M)}{K_1+(1-M)} - V_2\frac{M}{K_2+M}$$

$$\frac{dX}{dt} = V_3 M\frac{(1-X)}{K_3+(1-X)} - V_4\frac{X}{K_4+X}$$

11.基因调控模型：

转录/翻译过程可以简化为：转录因子(TF)与DNA结合形成复合物，从而促进或抑制蛋白质的合成。如果只考虑促进蛋白质合成，模型可以描述为：

$$TF + DNA \underset{k_2}{\overset{k_1}{\rightleftharpoons}} TFDNA \xrightarrow{k_3} P + TFDNA$$

如果假定同蛋白质合成相比，TF与DNA的结合及分离都十分迅速，那么上述过程与前面讲述的酶催化反应很类似，这里把DNA“想像”为酶。则可以用类似的米氏方程描述该过程。

$$\frac{d[TF]}{dt} = -k_1[TF][DNA] + k_2[TFDNA]$$

$$\frac{d[DNA]}{dt} = -k_1[TF][DNA] + k_2[TFDNA]$$

$$\frac{d[TFDNA]}{dt} = k_1[TF][DNA] - k_2[TFDNA]$$

$$\frac{d[P]}{dt} = k_3[TFDNA]$$

$$\frac{[X]}{K_1+[X]} = \frac{V_2[E]}{V_1}\frac{(1-[X])}{K_2+(1-[X])}$$

令 $K_1 = K_2 = K$ 及 $f = V_2[E]/V_1$

$$\frac{[X]}{K+[X]} = f\frac{(1-[X])}{K+(1-[X])}$$

该式描述了激活蛋白质浓度[X]与f之间的关系。同时，K值对其关系有直接影响。

由于DNA在一个细胞中只有一个拷贝，即一个基因要么被TF结合，要么没有结合，即：

$$[DNA]+[TFDNA]=1$$

从而有：

$$K = \frac{k_2}{k_1} = \frac{[TF][DNA]}{[TFDNA]}$$

$$[TFDNA] = \frac{[TF]}{K+[TF]}$$

K值越大，则基因更有可能不被TF结合而处于非转录状态，则[TFDNA]也越小，因此，[TFDNA]可以理解为某段时间内基因被TF结合的比率。

$$\begin{cases} \frac{d[P]}{dt} = k_3[TFDNA] = k_3\frac{[TF]}{K+[TF]} = V_{\max}\frac{[TF]}{K+[TF]} \\ V_{\max} = k_3 \end{cases}$$

那么，如果TF抑制基因转录时，该怎么建模？

$$TF + DNA \underset{k_2}{\overset{k_1}{\rightleftharpoons}} TFDNA$$

$$DNA \xrightarrow{k_3} P + DNA$$

$$\frac{d[TF]}{dt} = -k_1[TF][DNA] + k_2[TFDNA]$$

$$\frac{d[DNA]}{dt} = -k_1[TF][DNA] + k_2[TFDNA]$$

$$\frac{d[TFDNA]}{dt} = k_1[TF][DNA] - k_2[TFDNA]$$

$$\frac{d[P]}{dt} = k_3[DNA]$$

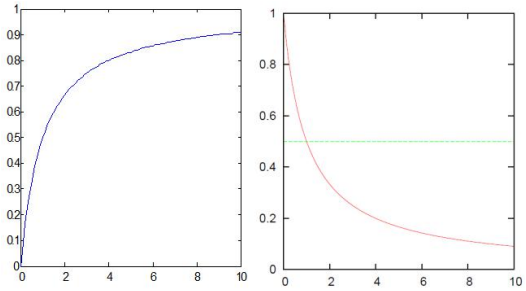
同样，假定同蛋白质合成相比，TF与DNA的结合及分离都十分迅速，且：

$$[DNA] + [TFDNA] = 1$$

$$K = \frac{k_2}{k_1} = \frac{[TF][DNA]}{[TFDNA]}$$

$$[DNA] = \frac{K}{K + [TF]}$$

$$\begin{cases} \frac{d[P]}{dt} = k_3[DNA] = k_3\frac{K}{K+[TF]} = V_{\max}\frac{K}{K+[TF]} \\ V_{\max} = k_3 \end{cases}$$



TF促使基因转录

TF抑制基因转录

如果基因的转录需要结合同一转录因子的多个分子才能进行，该如何建模？

$$X + X + DNA \underset{k_2}{\overset{k_1}{\rightleftharpoons}} TFDNA \xrightarrow{k_3} P + TFDNA$$

仍然假定：$[DNA] + [TFDNA] = 1$

由于：$\frac{d[TFDNA]}{dt} = k_1[X]^2[DNA] - k_2[TFDNA]$

$\frac{d[P]}{dt} = k_3[TFDNA]$

故：

$$\left.\begin{aligned} [TFDNA] &= \frac{[X]^2}{K+[X]^2} \\ \frac{dP}{dt} &= V_{max}\frac{X^2}{K+X^2} \end{aligned}\right\} \quad K = k_2/k_1 \qquad k_3 = V_{max}$$

如果基因的转录需要结合同一转录因子的多个分子才能被抑制，该如何建模？

$$X + X + DNA \underset{k_2}{\overset{k_1}{\rightleftharpoons}} TFDNA$$

$$DNA \xrightarrow{k_3} P + DNA$$

从而有：

$$\frac{d[p]}{dt} = V_{\max}\frac{K}{K+[X]^2}$$

如果转录1(TF1)与基因结合时，转录启动，另一个转录因子(TF2)结合时，转录被抑制，该如何建模？

$$P_{TF1bound} = \frac{[TF1]}{K_1+[TF1]} = \frac{[TF1]/K_1}{1+[TF1]/K_1}$$

$$P_{TF2notbound} = \frac{K_2}{K_2+[TF2]} = \frac{1}{1+[TF2]/K_2}$$

$$P_{TF1boundANDTF2notbound} = P_{TF1bound}P_{TF2notbound} = \frac{[TF1]/K_1}{1+[TF1]/K_1+[TF2]/K_2+[TF1][TF2]/K_1K_2}$$

如果我们只看Raf，该如何建模？

$$Raf + RasGTP \underset{k_{-28}}{\overset{k_{28}}{\rightleftharpoons}} RafRasGTP$$

$$Raf^*P1 \xrightarrow{k_{43}} Raf + P1$$

从而有：

$$\frac{d[Raf]}{dt} = -k_{28}[Raf][RasGTP] + k_{-28}[RafRasGTP] + k_{43}[Raf^*P1]$$

12.系统建模工具：化学动力学数据库（BRENDA）；生物系统建模工具（CellML）

第十二章 分子进化与系统发育分析

1.研究生物进化的途径：

①化石：可以获取与已经消亡生物最直接的证据和信息，是研究进化的理想材料。

②比较形态学和比较生理学：确定大致的进化框架。

③分子进化：研究的 DNA、RNA 和蛋白质分子的序列、结构的变化以及这些变化与生物进化的关系。

2.分子进化：细胞中的 DNA、RNA 和蛋白质等大分子的序列发生改变，并随着物种繁衍在代间传递的过程。

3.分子进化速率：核酸或蛋白质等分子在进化过程中核苷酸或氨基酸残基发生替换的频度，

即单位时间、单位长度序列上发生的变化。

4.分子进化速率的计算：

①氨基酸差异比例的计算：Pd=daa/Naa

daa 为用于比较的同源蛋白质之间有差异的氨基酸数目

Naa 为所比较的蛋白质的氨基酸序列长度

②氨基酸差异比例的校正：Kaa=-2.3Log10(1-Pd)

③进化速率的计算：kaa=Kaa/2T

进化速率用氨基酸替代频度表示。氨基酸替代频度是指每年每个氨基酸座位发生替代的可能性, 即用替换率表示。

5.分子进化的理论基础：

①中性突变理论：核酸或蛋白质序列中发生的分子突变大多数是中性的, 这些突变通过随机的遗传漂变在群体里固定下来, 分子进化是遗传漂变的结果, 在分子进化上自然选择不起作用。

②分子钟理论：同一核酸或蛋白质分子序列在不同物种间的差别的大小与这些物种在进化中的分化时间之间具有近似正线性关系，即分子变化和生物形态、生理等特征具有关联。

6.分子进化的机制：

①DNA 突变：替代，插入，缺失，倒位

②核苷酸替代：转换 (Transition) & 颠换 (Transversion)

③基因复制：多基因家族的产生以及伪基因的产生；单个基因复制–重组或者逆转录；染色体片断复制；基因组复制

7.分子进化的研究内容：

①核苷酸替代等的速率及其影响

②进化理论(中性进化与自然选择)

③基因的起源

④复杂性状的遗传规律

⑤物种的分化

⑥进化对基因组及表型等的影响

⑦基因组的结构及演化

⑧分子的系统发育

8.分子进化的特点：

①普适性：对所有物种，均可以分析对应的由 4 种核苷酸及 20 种氨基酸所衍生的核酸及蛋白质序列演化和蛋白质结构演化。

②可比较性：通过分子进化分析，可以通过建立核酸序列、氨基酸序列、蛋白质结构及其与形态、性状演化关系的模型，对不同物种进行比较。

③信息丰富性：与形态、性状包含的信息相比，基因组序列包含的信息更全面和多样，可以通过分子进化对物种进行更系统的研究。

④进化速率的恒定性：不同物种同源大分子的分子进化速率大体相同，例如人与马的血红蛋白氨基酸序列差异 0.8×10-9/AA.a，人与鲤鱼 0.6×10-9/AA.a，分子进化速率远远比表型进化速率稳定。

⑤进化保守性：功能重要的分子的进化速率上明显低于功能不太重要的分子引起表型发生显著改变的突变频率要低于无明显表型改变的频率。

9.分子进化的应用：

①物种进化关系重建：以一些生物大分子为基础，构建多个物种的生物系统发生的关系-- tree of life，或对物种进行分类。

②大分子功能与结构的分析：同一家族的大分子，具有相似的三级结构及生化功能，通过序列同源性分析，分析大分子结构、功能之间的关联。

③进化速率分析：例如，HIV 的高突变性；哪些位点易发生突变。

10.作为进化标尺的生物大分子的选择原则：

①在所需研究的种群范围内，它必须是普遍存的。

②在所有物种中该分子的功能是相同的。

③分子上序列的改变（突变）频率应与进化的测量尺度相适应。

11.分子进化分析的基本步骤：

①序列比较：源于同一祖先 DNA/氨基酸序列的两条 DNA/氨基酸序列，考察二者的差异。

②序列差异：进化过程中分子突变的痕迹。

③分子进化：以累计在 DNA/氨基酸分子上的历史信息为基础研究分子水平的生物进化过程和机制。

12.分子钟：两个物种的同源基因之间的差异程度与它们的共同祖先的存在时间(即两者的分歧时间)有一定的数量关系。

分子钟成立的先决条件：分子进化速率恒定。

13.建立分子钟的步骤：

①选择所要比较的生物大分子种类

②确定各物种的比较组合及其所代表的进化事件

③获得生物大分子一级结构资料

④获得有关的代表性进化事件发生的地质时间数据

⑤通过比较大分子一级结构，选择合适的数学模型计算得到进化产生的分子差异 d，通过回归分析等统计方法得到大分子的进化速率 r(t)

⑥由此可以推断未知进化事件的发生时间

14.系统发育树：用一种类似树状分支的图形来概括各种（类）生物之间的亲缘关系。

15.分子系统树：通过比较生物大分子序列差异的数值构建的系统树。

16.系统发育树重建分析步骤：

①多序列比对（自动比对，手工比对）

②建立取代模型（建树方法）

③建立进化树

④进化树评估

17.系统发育树重建的基本方法：

①最大简约法(maximum parsimony,MP)：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。

②距离法(distance)：又称距离矩阵法，首先通过各个物种之间的比较，根据一定的假设（进化距离模型）推导得出分类群之间的进化距离，构建一个进化距离矩阵。进化树的构建则是基于这个矩阵中的进化距离关系。（包括：FM 法、NJ 法/邻接法、邻居关系法、UPGMA 法）

③最大似然法(maximum likelihood,ML)：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。