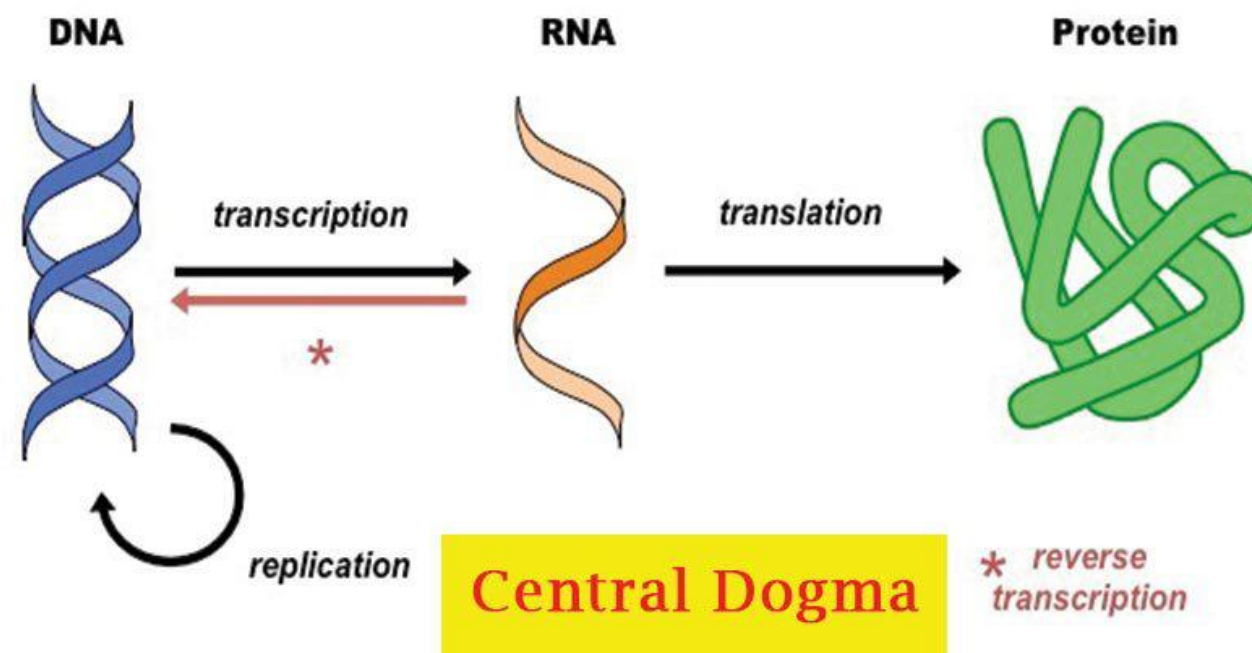
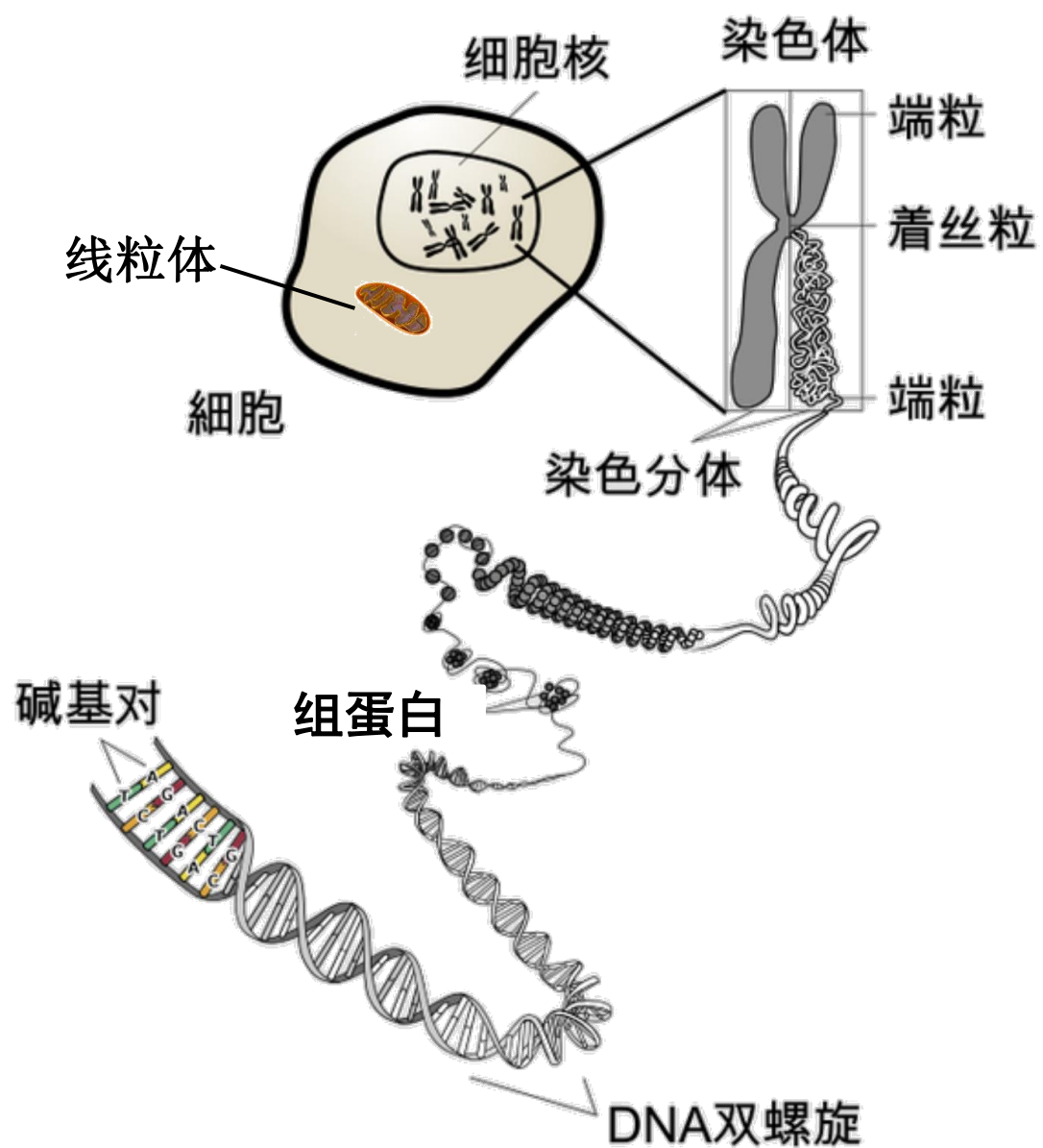


基因组分析

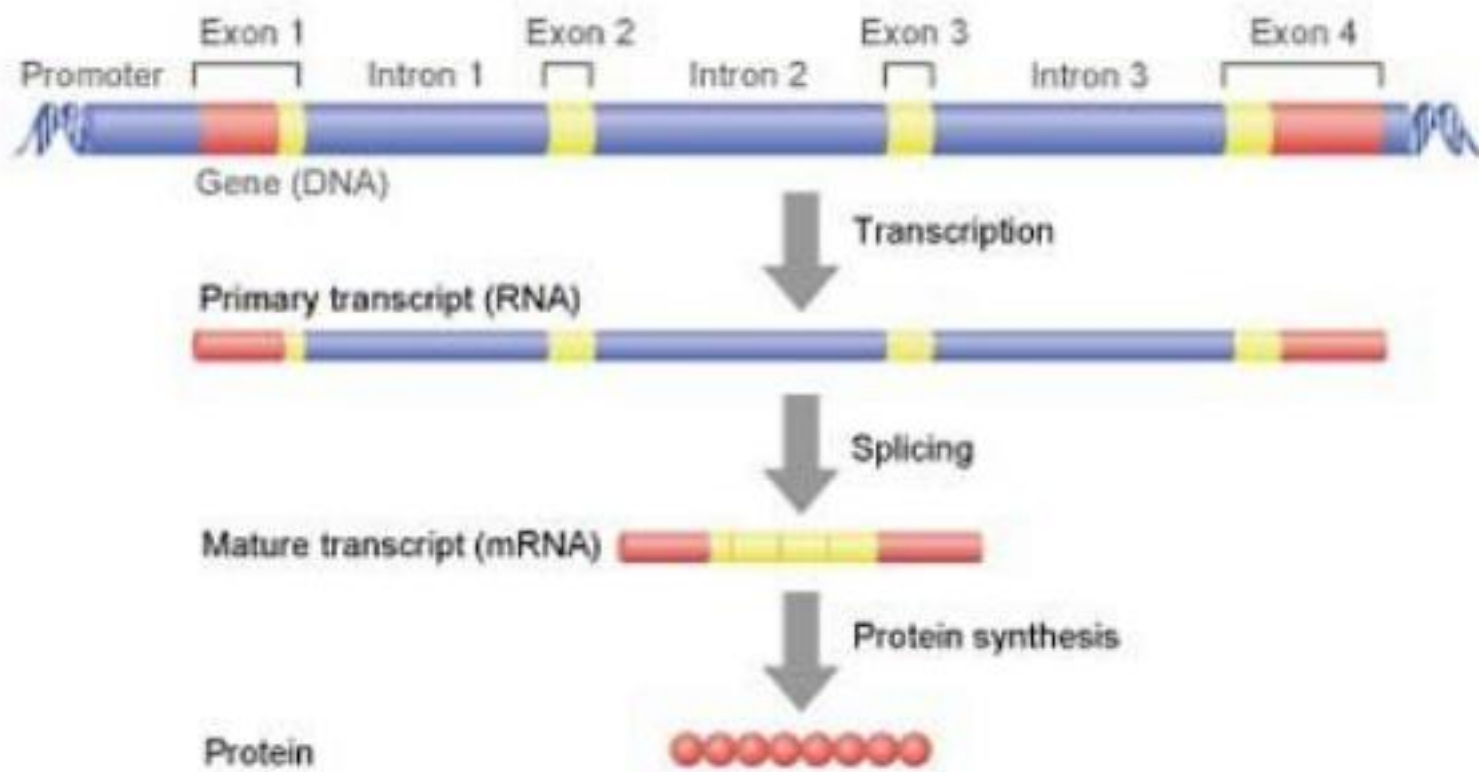
基因组研究内容

- **基因组测序与拼装**
 - 基因组重复序列分析
 - 两个基因组比较
 - 多个基因组比较（重测序）
 - 群体遗传分析
- **编码基因预测及其功能注释**
- **非编码RNA鉴定及功能预测**
 - 小RNA: miRNA, siRNA
 - 长RNA: lncRNA, circRNA
- **基因转录与调控网络**



基因的结构

Structure of a Gene

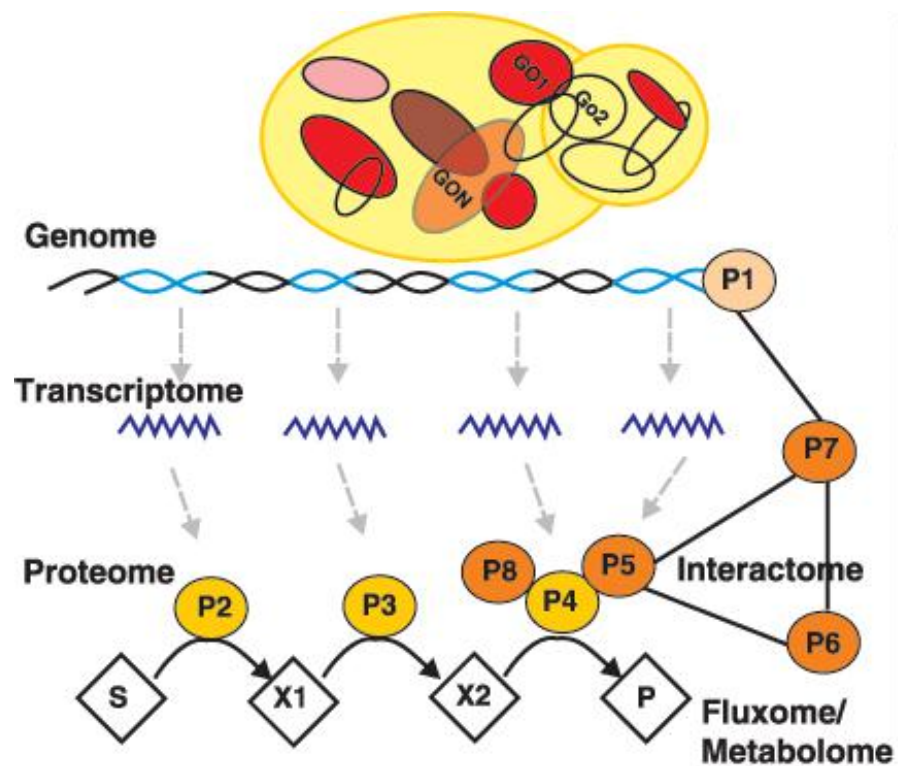


基因组、转录组和蛋白质组

基因组 转录组

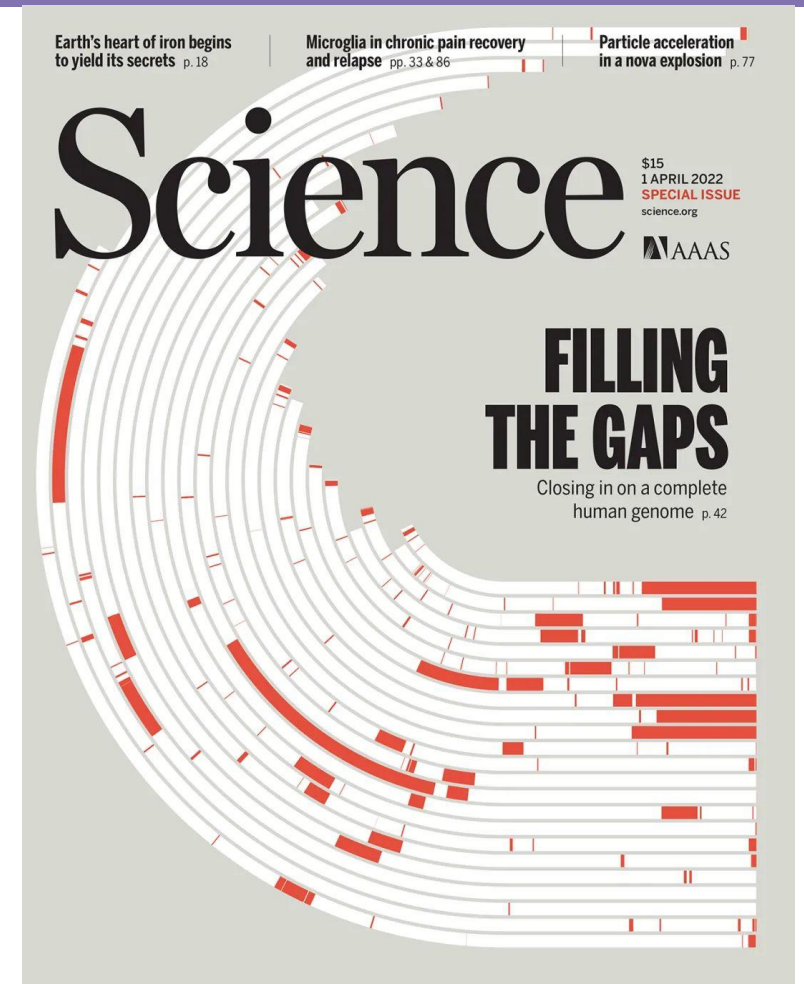
蛋白质组

化学生物学



a. GO	Global functional annotation
Interaction	Level of information
b. TF	Transcriptional regulation
c. Protein	Flow of information through PPI
d. Complexes	Agglomerative functional modules
e. Metabolites	Metabolic network regulation

人类基因组计划



2003年人类基因组草图公布。但所得到的的是一个非最终版本，只覆盖了基因组中常染色质的部分，而非常重要的异染色质区域尚未完成。为了完成基因组的最后剩余区域，2022年4月1日科学家们利用PacBio HiFi和Oxford Nanopore超长测序的互补方面来组装均匀纯合CHM13hTERT细胞系中的人类基因组。由此产生的T2T-CHM13参考装配弥补了这20多年来人类基因组中8%的空白，最终汇总为***The complete sequence of a human genome***。

人类基因组基本数据

1Gb (Gigabases) = 1 000 000 000bp = 10^9 bp

- 人类基因组大约有30多亿 (3Gb) 个碱基对 (A,C,G,T) .
- 大约20000个基因；蛋白质的编码序列，只占总长度的约1.5%。
- 清楚功能的3%；不清楚功能约97%（“暗物质”）。



已测序最大的**动物**基因组(2021-1-25)：澳大利亚肺鱼： Australian lungfish (Neoceratodus forsteri) 430亿碱基对（43Gb）

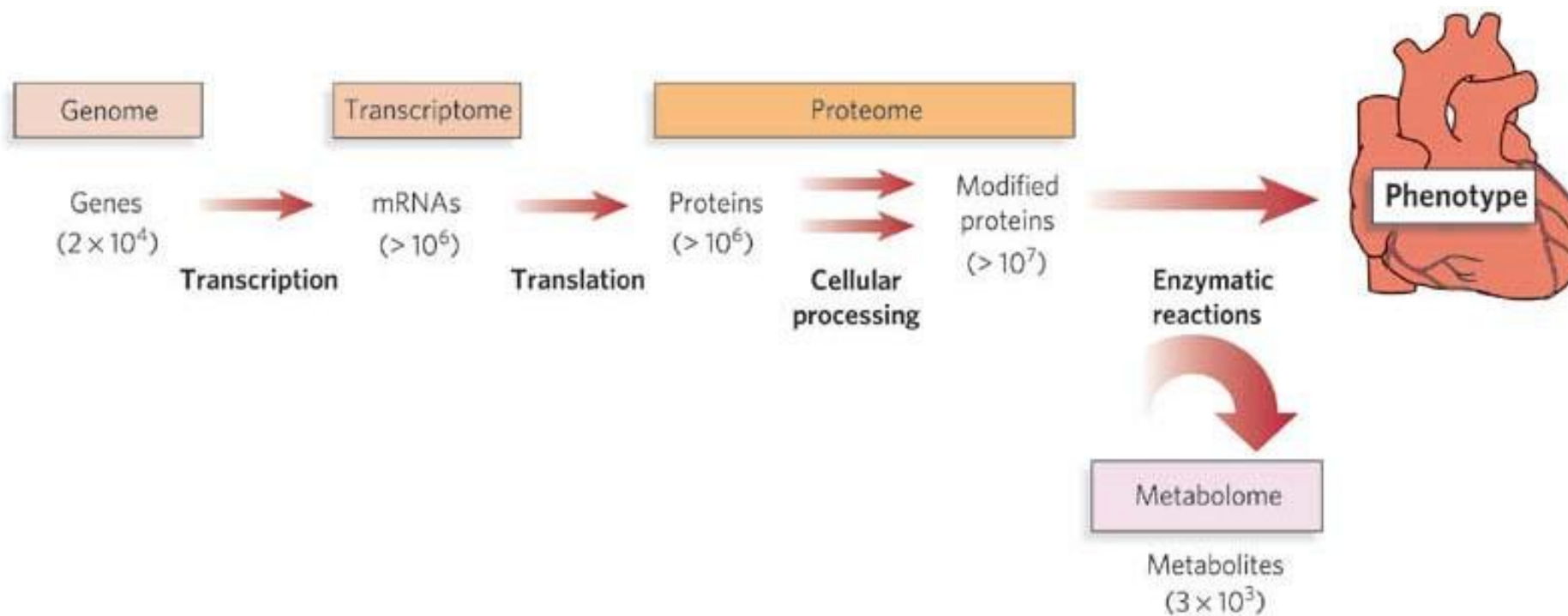
其他物种基因组大小

物种	染色体个数（单倍体）	基因组大小
人	22,X,Y	3.2Gb
黑猩猩	23,X,Y	2.73Gb
家犬	39	2.4Gb
猫	19	2.5Gb
家猪	18， X,Y	2.8Gb
鸡	38， Z,W	1.2Gb
水稻（籼稻）	12	400Mb
玉米	20	2Gb
大豆	20	1.1Gb
酿酒酵母	16	12Mb
火炬松(loblolly pine)	12	22.1Gb

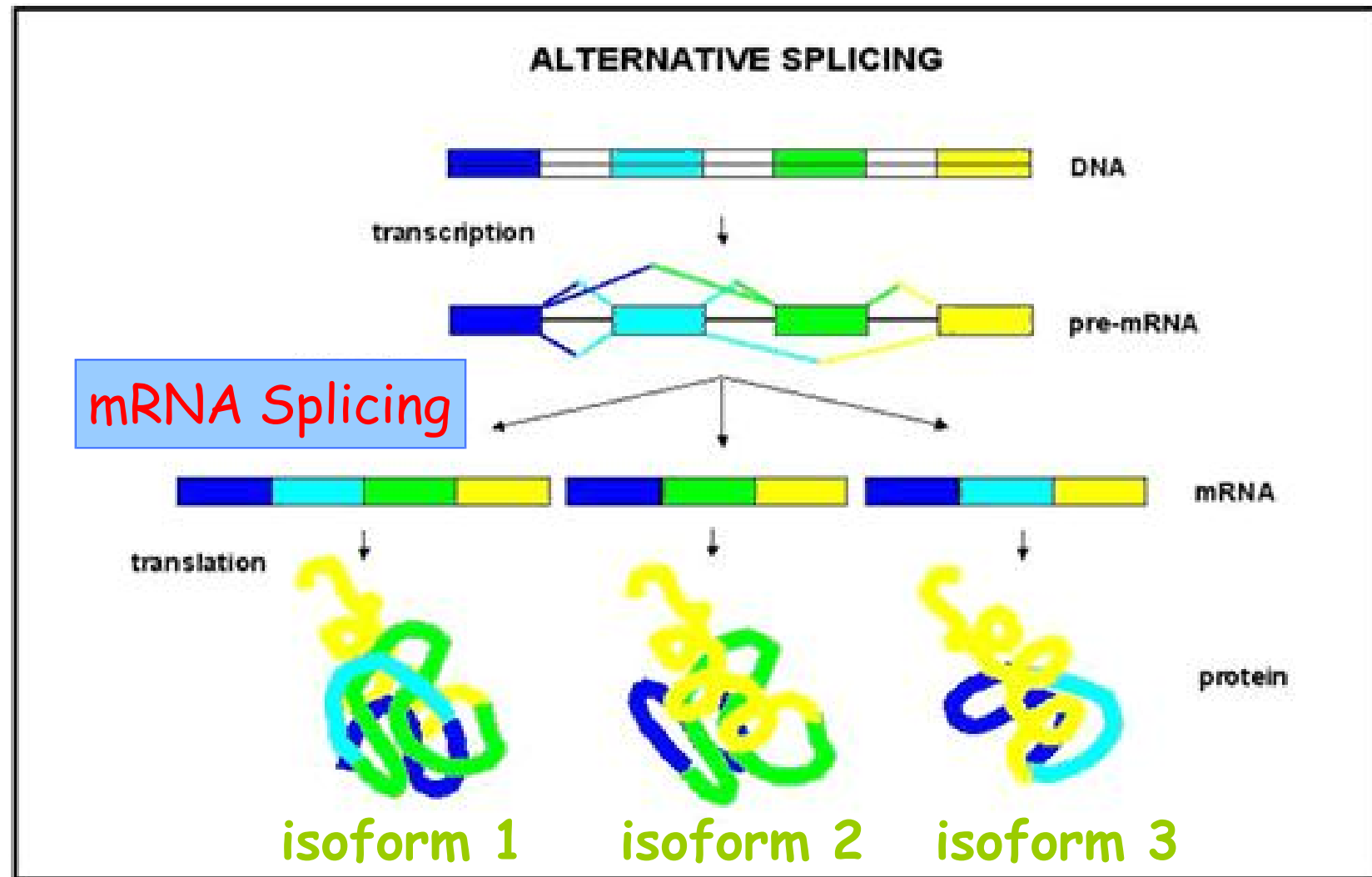
基因数量 -> 生物复杂性?

- ❑ 基因数量的变化，无法解释生物学功能、调控机理以及物种多样性和复杂性的巨大变化
- ❑ 当前解释：蛋白质组的多样性和复杂性 -> 物种的多样性和复杂性；
~10,000,000种蛋白质分子
- ❑ 两种观点：
 - ⚙ 转录后层面，mRNA剪切，产生拼接异构体
 - ⚙ 蛋白质层面，蛋白质序列上一个或多个位点上发生的翻译后修饰

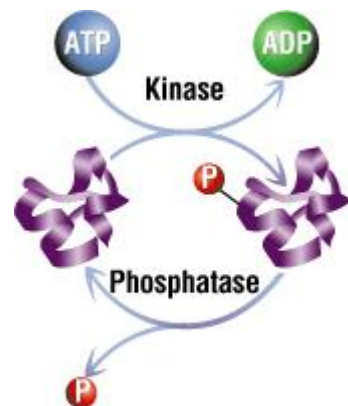
基因型到表型



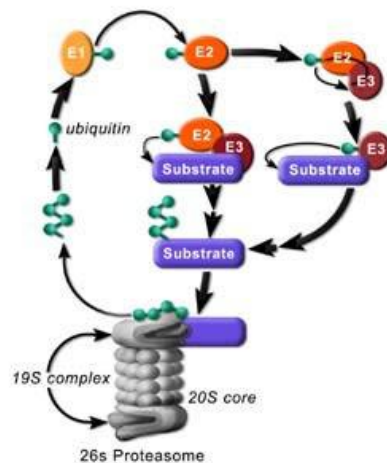
转录后层面: mRNA Splicing



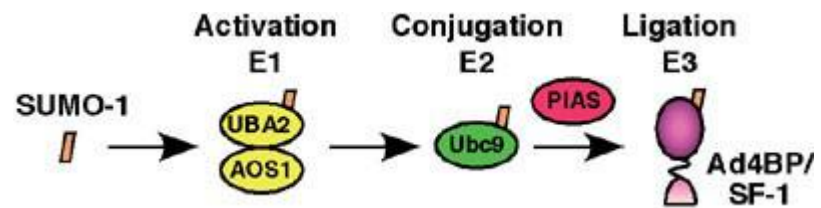
蛋白质层面：翻译后修饰



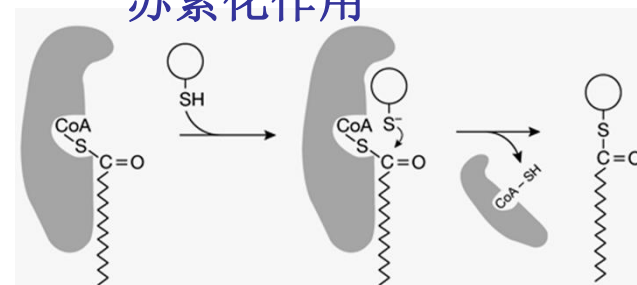
Phosphorylation
磷酸化作用



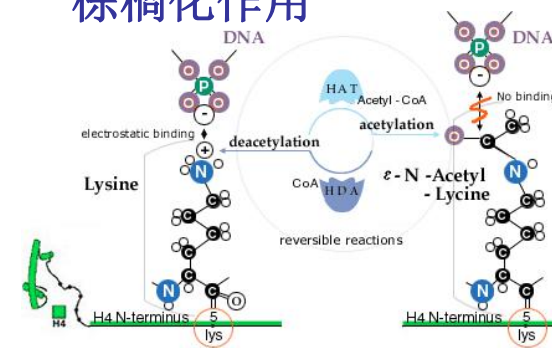
Ubiquitination 泛素化作用



Sumoylation
苏素化作用

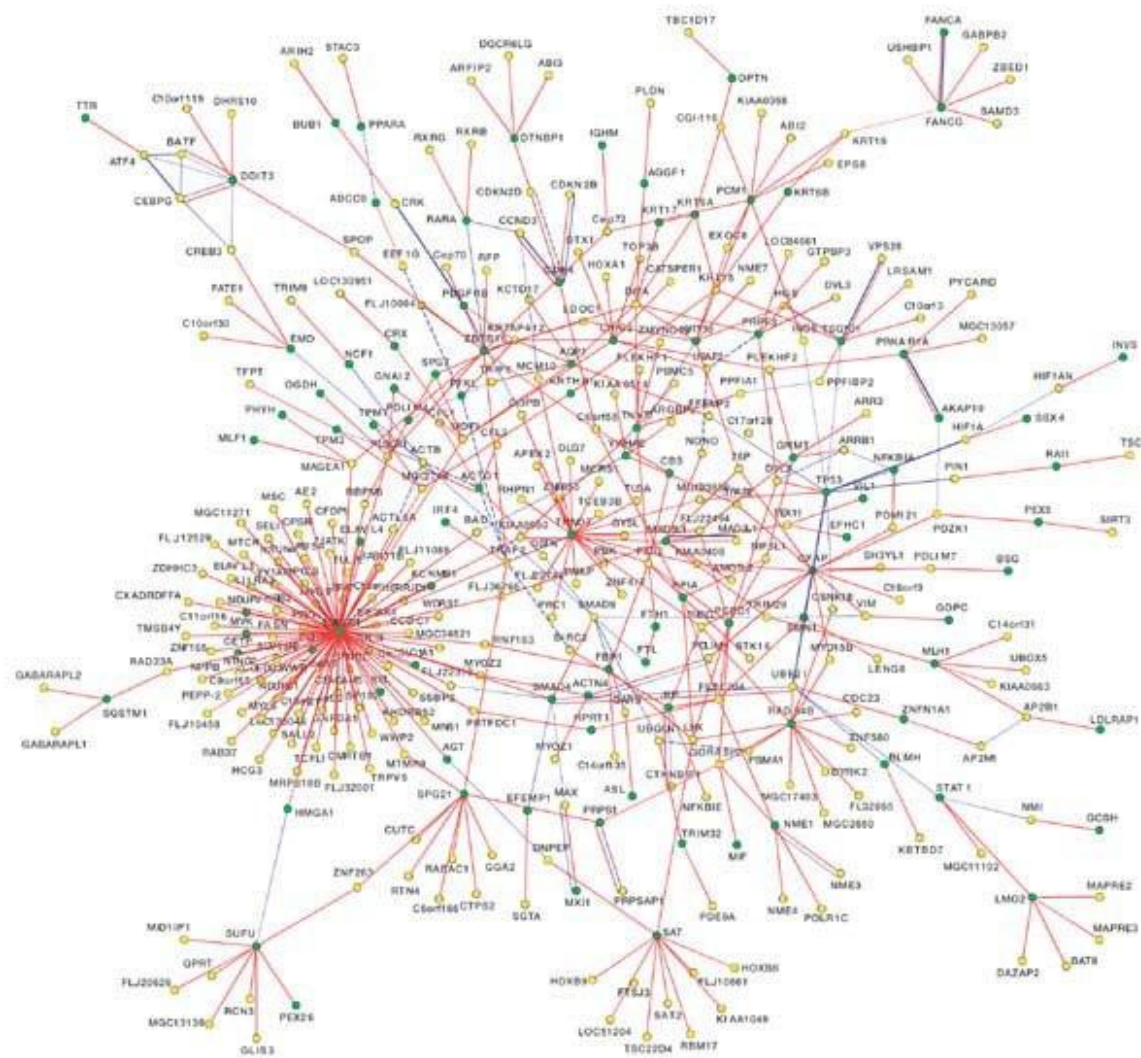


Palmitoylation
棕榈化作用



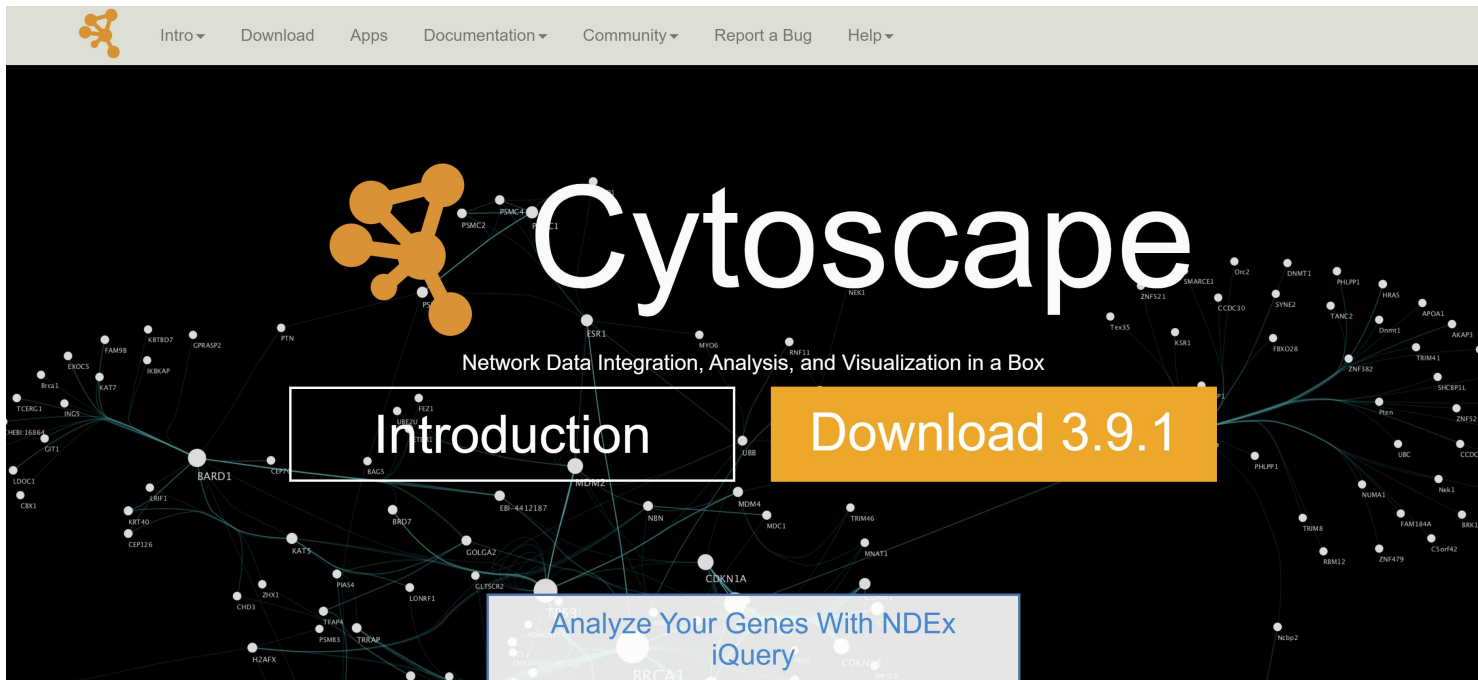
Acetylation
乙酰化作用

相互作用网络

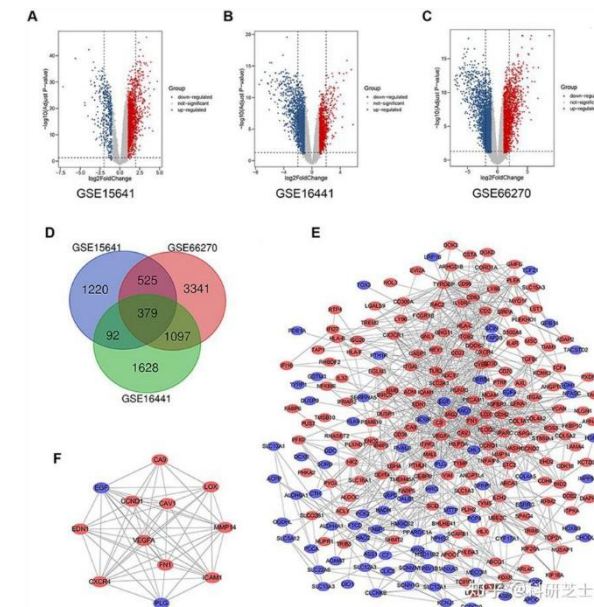
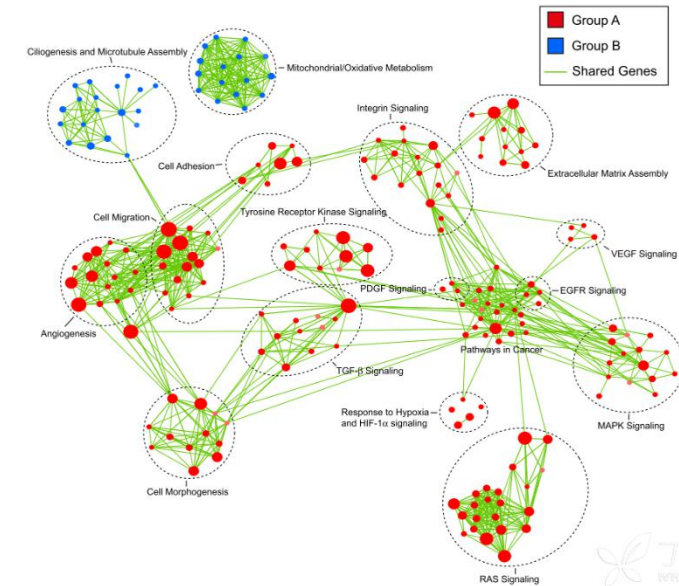


蛋白质-蛋白质相互作用网络

Cytoscape: 网络构建和分析工具



The screenshot shows the Cytoscape website with a navigation bar at the top containing links: Intro, Download, Apps, Documentation, Community, Report a Bug, and Help. The main content area features the Cytoscape logo and the text "Cytoscape Network Data Integration, Analysis, and Visualization in a Box". Below this, there are two prominent buttons: "Introduction" and "Download 3.9.1". At the bottom, a banner reads "Analyze Your Genes With NDEx iQuery". The background of the website is a complex network graph with various nodes and edges.



非编码RNA

非编码RNA

❑ 不翻译成蛋白质，具有重要的调控功能

❑ 分类：

- ✿ transfer RNA (tRNA)

- ✿ ribosomal RNA (rRNA)

- ✿ snoRNAs: Small nucleolar RNAs; 介导其他RNA分子的化学修饰，例如甲基化

- ✿ microRNAs

- ✿ siRNAs

- ✿ piRNAs: 与piwi相互作用的RNA

- ✿ long ncRNAs

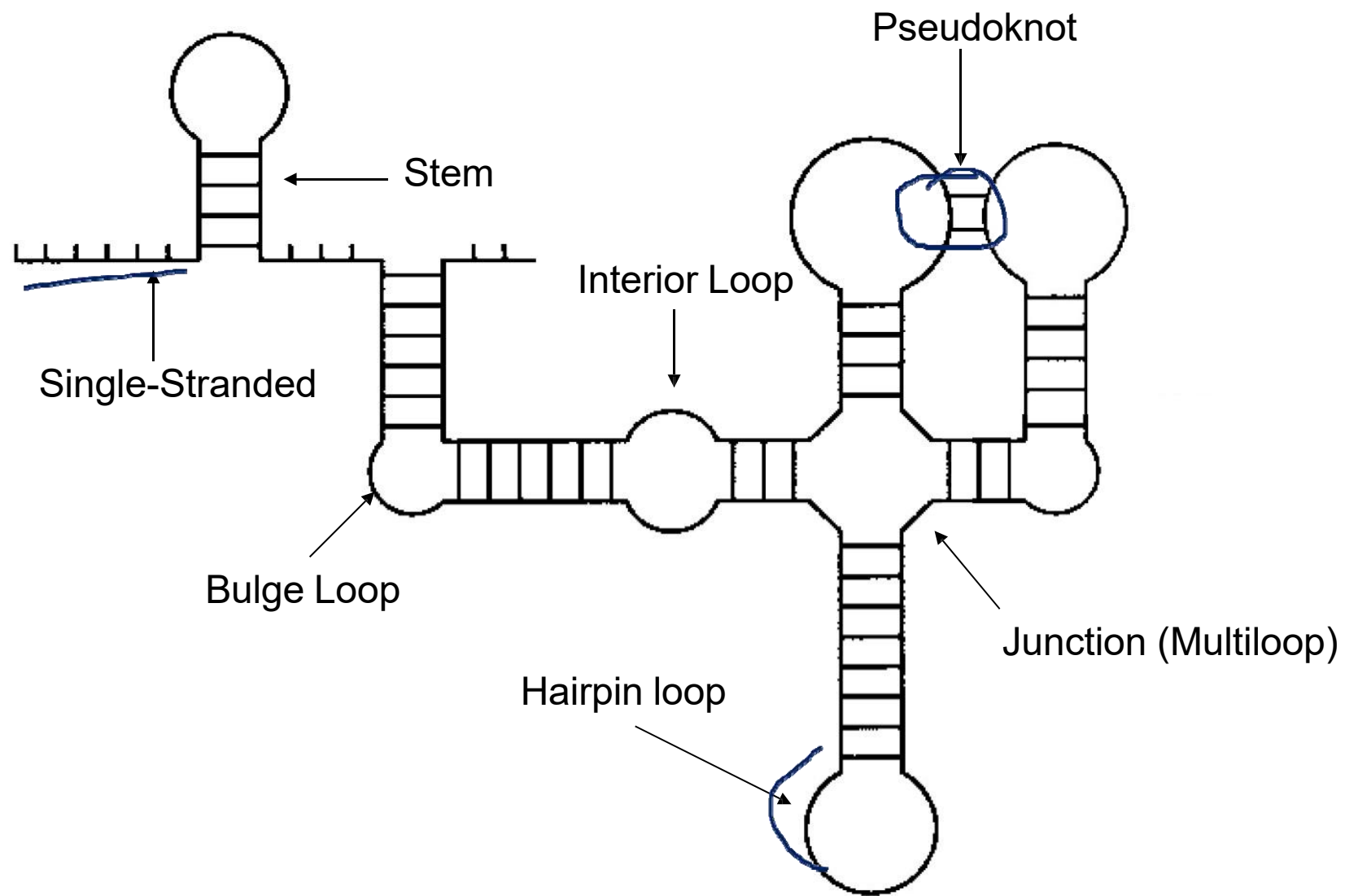
- ✿ circRNAs

❑ 本章主要介绍：

小RNA: miRNA, siRNA

长RNA: lncRNA, circRNA

RNA二级结构

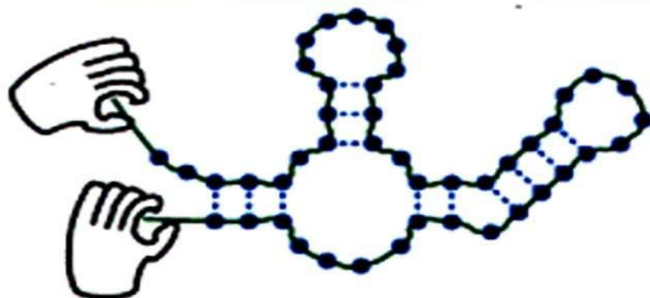


RNA二级结构的显示方式

- Grammatically correct **string of parentheses**

..(((.(((.....))).((((((.....))))).)).....)))
AGCTACGGAGCGATCTCCGAGCTTTTCGAGAAAGCCTCTATTAGC

- Planar graph**



- Arch diagram**



- Mountain diagram**

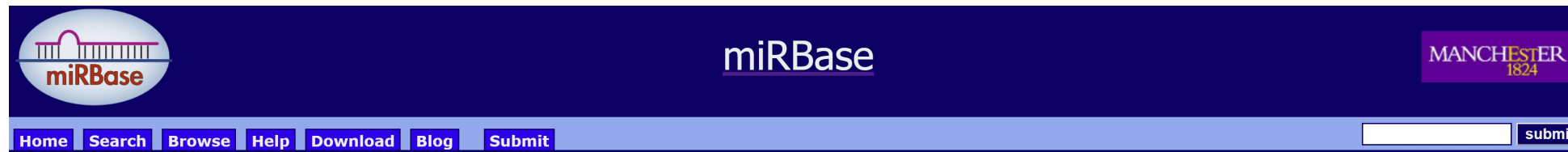


microRNA/miRNA

- ❑ **MicroRNA (miRNA):** 一种非编码小的RNA (~21–23 bp), 通过**Dicer**剪切其前体RNA (~70-90) 所得
- ❑ **miRNA**以RNA-蛋白质复合物的形式, 在动物和植物的细胞中广泛的表达, 也称为**miRISCs**
- ❑ 在发育的过程中起着关键性的作用, 能够促使与**miRNA**序列同源的靶基因的**mRNA**的降解或者抑制翻译。

miRNAs的多样性

- ❑ miRNA 数据库 miRbase: <https://mirbase.org/>
- ❑ Release 12.0, 收录8619个具有发夹结构的前体 miRNAs, 能够表达出8273个miRNAs



miRBase: the microRNA database

miRBase provides the following services:

- The [miRBase database](#) is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for [searching](#) and [browsing](#), and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also [available for download](#).
- The [miRBase Registry](#) provides miRNA gene hunters with unique names for novel miRNA genes prior to publication of results. Visit the [help pages](#) for more information about the naming service.

To receive email notification of data updates and feature changes please subscribe to the [miRBase announcements mailing list](#). Any queries about the website or naming service should be directed at mirbase@manchester.ac.uk.

miRBase is managed by the [Griffiths-Jones lab](#) at the [Faculty of Biology, Medicine and Health, University of Manchester](#) with funding from the [BBSRC](#). miRBase was previously hosted and supported by the [Wellcome Trust Sanger Institute](#).

miRNA count: 38589 entries

[Release 22.1](#)

Search by miRNA name or keyword

Download published miRNA data

[Download page](#)

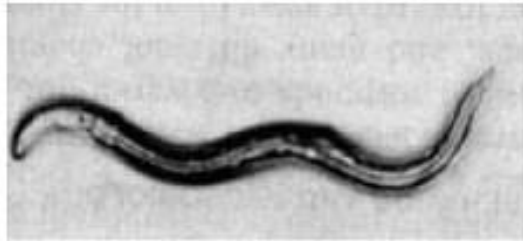
miRNAs的多样性

MicroRNAs (miRNAs)



*A. thaliana/
O. sativa*

拟南芥: **187**



C. elegans

154



D. melanogaster

152



*H. sapiens/
M. musculus*

人: **695**

The miRBase Sequence Database -- Release 12.0

miRNAs的计算鉴定

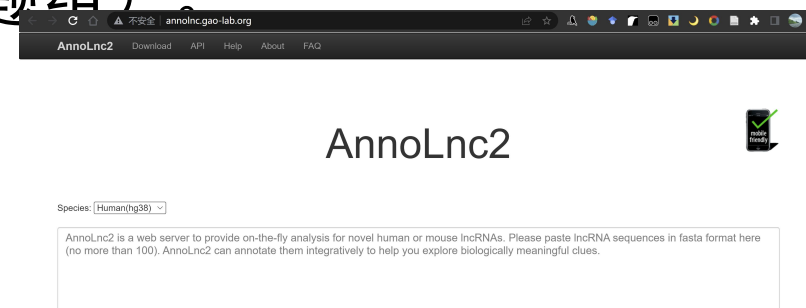
- ❑ 同源比对。
 - ❑ BLAST, miRAlign(2005), miRNAminer (2008)
- ❑ 邻近茎环结构搜索。
 - ❑ 动物miRNA经常成簇存在于基因组，通过已知miRNA附近区域进行茎环结构预测，可用来发现miRNA。
 - ❑ 植物中miRNA成簇比较少。例如拟南芥25%的miRNA成簇。
- ❑ 基于比较基因组学算法。
 - ❑ 亲缘关系很近的生物，它们的基因序列很保守。MIRcheck软件（2004）
- ❑ 高通量miRNA测序数据的方法。
 - ❑ 从高通量测序数据获得的小RNA数据中预测miRNA。例如MIREAP（2012），sRNAanno (www.plantsRNA.org)

miRNA靶基因预测

- ❑ 动物miRNA结合靶基因机制较复杂，植物miRNA主要通过互补配对结合到靶位点，对目标mRNA切割。
 - ✿ 植物：psRNATarget (2011),
TAPIR (2013, <http://bioinformatics.psb.ugent.be/webtools/tapir/>)
 - ✿ 动物：TargetScan, miRecords, PicTar等。
- ❑ 靶序列特征
 - ✿ 长度较短 (~21 nt)
 - ✿ G-U配对
 - ✿ 错配和空位 (bulges, 凸起膨胀)
 - ✿ 假阳性高
- ❑ 提高准确性：考虑mRNA UTR的二级结构

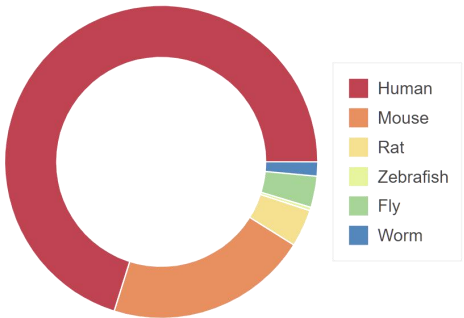
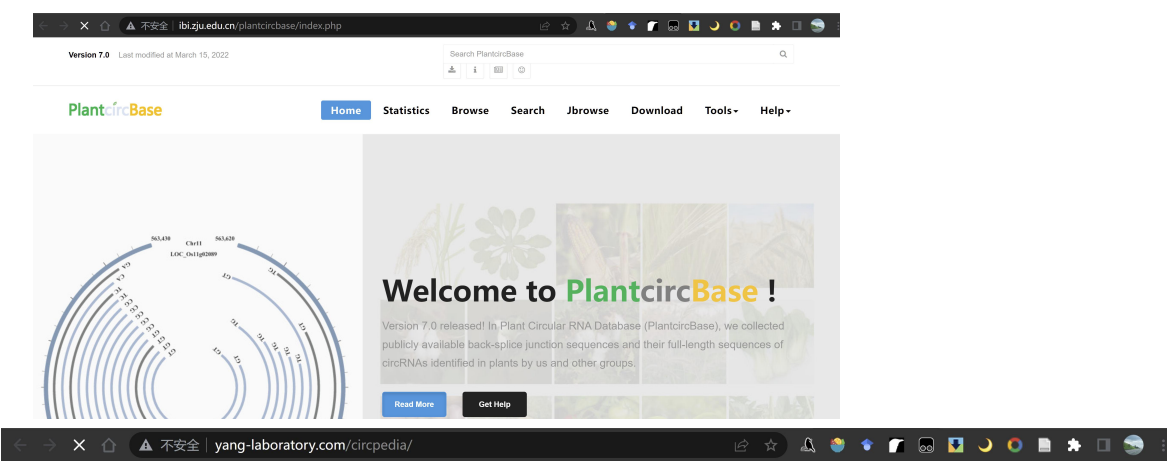
长非编码RNA鉴定和功能预测

- **LncRNA，是一类长度大于200个核苷酸，并且不能翻译成蛋白质的转录本。**
 - 鉴定LncRNA最大难点是确定转录组的非编码性。主要通过排除编码蛋白质的转录本来实现。
- **LncRNA功能预测：**
 - LncRNA与miRNA分子相互作用。LncRNA诱捕miRNA序列，miRNA先与AGO蛋白形成复合物，同碱基互补配对绑定特定的信使RNA序列，导致信使RNA序列翻译受阻或者在特定位点被剪切。
 - LncRNA与其他RNA分子互相作用。INTARNA (2008) 基于分子间互相作用能。
 - LncRNA与蛋白质分子互相作用。（1）基于序列方法 （2）基于结构和序列的方法。（3）基于实验数据的方法。 代表方法有： RPI-Seq, IncPRO
 - LncRNA功能注释平台：AnnoLnc （北京大学高歌课题组）



circRNA 鉴定和预测

- 环状RNA分子是一类由反向剪切形成的非编码RNA.分为：
 - 外显子类型环状RNA
 - 内含子类型环状RNA
- 数据库：
 - plantcircBase <http://ibi.zju.edu.cn/plantcircbase/index.php>
 - CIRCpedia <http://yang-laboratory.com/circpedia/>



Species	No. of samples			No. of circRNAs
	Ribo ⁻	poly(A) ⁻	RNase R	
Human (hg38)	54	13	3	183,943
Mouse (mm10)	51	1	1	55,312
Rat (rn6)	6	0	0	10,197
Zebrafish (danRer10)	1	0	1	911
Fly (dm6)	33	0	9	8560
Worm (ce10)	6	0	6	3859

CIRCpedia v2

SearchBrowseDownloadToolAbout

CIRCpedia v1

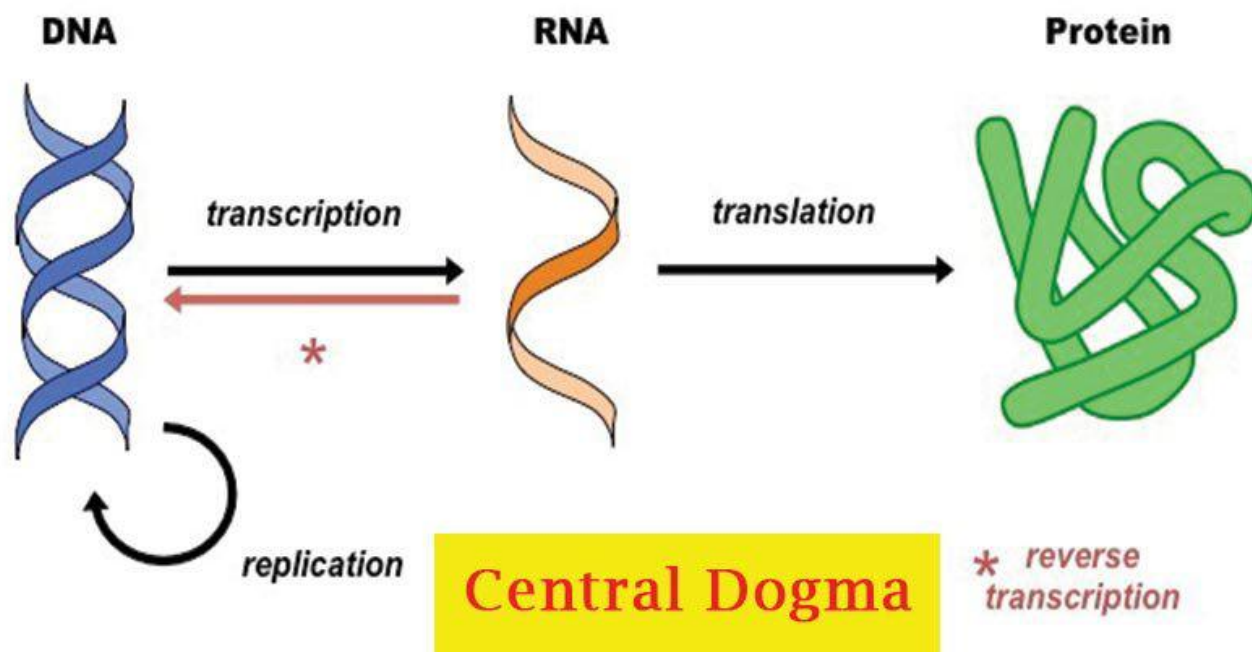
Welcome to CIRCpedia v2

CIRCpedia v2 is an updated comprehensive database containing circRNA annotations from over 180 RNA-seq datasets across six different species. This atlas allows users to search, browse and download circRNAs with expression characteristics/features in various cell types/tissues, including disease samples. In addition, the updated database incorporates conservation analysis of circRNAs between humans and mice. Finally, the web interface also contains computational tools for the comparison of circRNA expression between samples. The updated database is accessible at <http://www.picb.ac.cn/nomics/circpedia>.

circRNA功能

- **circRNA在生物的生长发育过程起到重要作用。**
- **circRNA和miRNA相互作用**
 - 预测动物circRNA: miRanda 软件
 - 预测植物circRNA: eTM_finder 软件
 - 在线服务器: circBase, circaltas, circInteractome
- **circRNA 编码潜能预测。若circRNA能翻译蛋白质，其氨基酸序列一定覆盖circRNA的反向剪接位点。这是区分线性RNA与circRNA翻译的蛋白质的关键。常见的方法有：**
 - 通过判断是否有内在的核糖体进入结合位点来预测: IRESfinder
 - 是否具有开放读码框来预测: cORF_pipeline, ORFfinder 等方法。

基因转录与调控网络



- RNA是桥梁。
- 转录组：
狭义上讲，特定环境下，一个细胞或者一群细胞的基因组转录出来的所有mRNA的总和。
广义来讲，是特定组织或细胞在某一发育阶段或功能状态下转录出来的所有RNA的总和。包括mRNA和非编码RNA。

基因表达分析

- 当我们拿到基因的表达数据之后，不能直接比较两个基因的表达差异。基因长度越长，测序深度越深，比对到该基因的读序数量越多。所以需要`对基因上的读序数据进行标准化`。
- 常见的标准化方法有
 - RPKM: reads per kilobase per million mapped reads
 - FPKM: fragments per kilobase per million mapped reads
 - TPM: transcripts per kilobase per million mapped reads

标准化方法：RPKM, FPKM

- RPKM: 每百万比对上的**读序**中, 比对到每千个碱基长度的外显子 (或转录本) 上的读序数量。 (适用单端测序)

- $$RPKM = \frac{\text{落在基因上的总读序数}}{\frac{\text{全部读序数}}{1\,000\,000} \times \frac{\text{基因长度}}{1\,000}}$$

- FPKM: 每百万比对上的**片段**中, 比对到每千个碱基长度的外显子 (或转录本) 上的片段数量。 (适用双端测序)

- $$FPKM = \frac{\text{落在基因上的总片段数}}{\frac{\text{全部片段数}}{1\,000\,000} \times \frac{\text{基因长度}}{1\,000}}$$

标准化方法: TPM

- TPM 先对基因长度进行标准化, 然后在对测序深度进行标准化。
- $TPM = \frac{A \times 10^6}{\sum A}$, $A = \frac{\text{落在基因上的总读序数}}{\frac{\text{基因长度}}{1\,000}}$
- 常见的标准化软件 Range(计算RPKM) , StringTie(计算 FPKM, TPM) 。
- 对基因表达数据进行分析时, 直接使用R软件包DESeq2, edgeR, 对原始的读序数(reads count)做处理。

表达差异基因的鉴定

- 使用R软件包 DESeq, edgeR, baySeq, 基于负二项分布模型, 来鉴定差异表达基因。
- 鉴定差异表达基因之后, 需要做多重检验问题, 需要用 Benjamini检验方法+ BH校正(Benjamini and Hochberg 1995)。

差异表达基因富集分析

- **基因功能富集分析是筛选出两组或多组表达水平有差异的基因集合（富集基因集）。**
- **富集分析中常用的注释数据库**
 - GO (Gene Ontology)
 - KEGG (Kyoto Encyclopedia of Genes and Genomes)
- **常用基因富集分析方法**
 - Fisher 精确概率方法。

基因调控网络分析

网络分类

- **随机网络** random network: 平均节点度遵循一个简单参数。如渔网，每个节点具有同样的链路数量。
 - 随机网络的平均节点度分布，泊松分布。
- **无标度网络** scale-free network: 网络中某些节点与其他节点有很多相连的链路，但大多数节点与其他节点的链路很少。这些很多链路的节点称为集散节点（Hub）。如生物网络
 - 无标度网络平均节点度分布（x-轴：链路数；y-轴节点数量），遵循幂次定律（连续递减曲线）。

聚合系数 Clustering coefficient

- 聚合系数为，给定节点与其邻近节点的实际连接数 n 和所有可能连接数的比值。
- $$C = \frac{n}{\frac{k(k-1)}{2}}$$
- 如果一个网络的平均聚合系数远远高于具有同样节点数的随机网络平均聚合系数同时网络的平均最短路径都很短，则称该网络为小世界网络(small world network)

网络模块 Modularity

- 网络模块是指负责网络中网结之间显著增多的链接图形和规则。
- 常用的网络模型分析与可视化软件：
- Cytoscape 插件： NeMo; MCODE
- Pajek: 大型网络分析软件

基因调控网络

- **基因调控网络建模方法包括：布尔网络、贝叶斯网络、神经网络、线性模型、微分方程以及其他随机模型。**
- **以布尔网络为例，基因表达状态离散为0,1.基因之间的相互作用通过布尔函数描述。具有以下优点：**
 - 模型重点研究系统基本原理，而不是生化反应细节。
 - 确定基因之间相互作用的定性关系。
 - 确定网络动态行为及其生物现象。
 - 研究网络的干预效果。

基因预测

❑ 直接的，序列高度匹配

- 同一或近缘物种中，与**EST**，**cDNA**，蛋白质 等序列完美或近似完美的匹配

❑ 间接的，基于统计学的

- 序列比对 (**Homology**)
- 从头预测 (*ab initio*)
- 以上两种方法的结合

真核生物基因结构

