

第十一章 系统生物学

廖明帜

第一节 系统生物学的概述

第二节 生物网络类型及数据资源

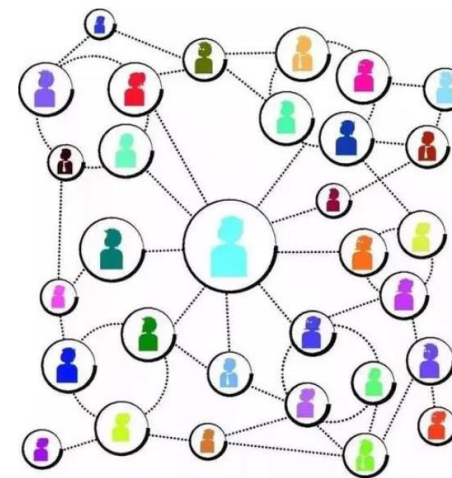
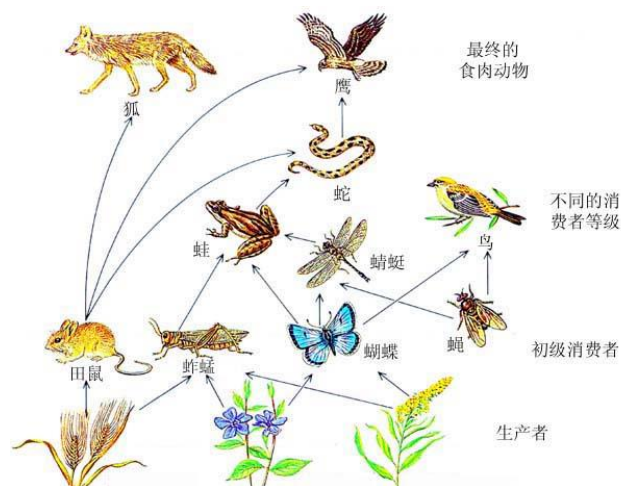
第三节 生物分子网络及特征

第四节 生物网络的构建

第五节 生物系统建模

第六节 总结

系统生物学是一门主要从分子的微观层次研究生物系统的学科，它将生物体视为一个复杂的系统，并通过整合不同层次和尺度的信息来理解生物体的功能和行为。



□整体论

- ✓ 系统生物学认为生物体是一个整体，其功能不能简单地由其组成部分的性质相加得到

□整合性

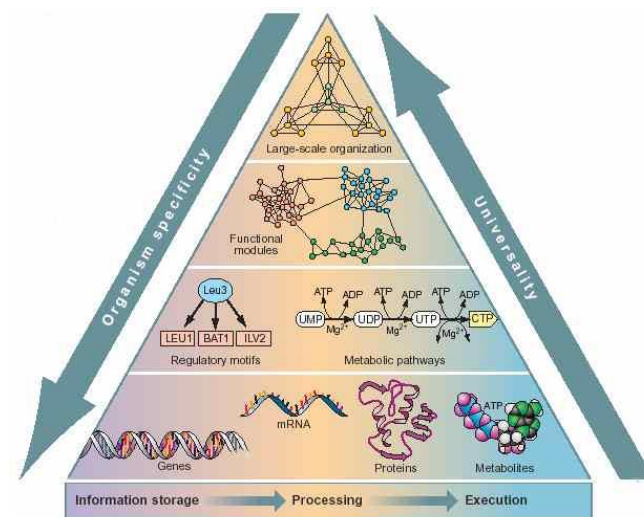
- ✓ 系统生物学整合来自不同学科和层次的信息，包括基因组学、蛋白质组学、代谢组学、细胞生物学等

□定量化

- ✓ 系统生物学使用数学模型来模拟和分析生物系统

□动态性

- ✓ 系统生物学关注生物系统的动态变化



□20世纪初，随着生物学研究的深入，科学家们逐渐认识到，生物系统是一个由多个相互作用的部分组成的复杂整体

✓传统的还原论方法已经无法有效地解释生物系统的复杂性

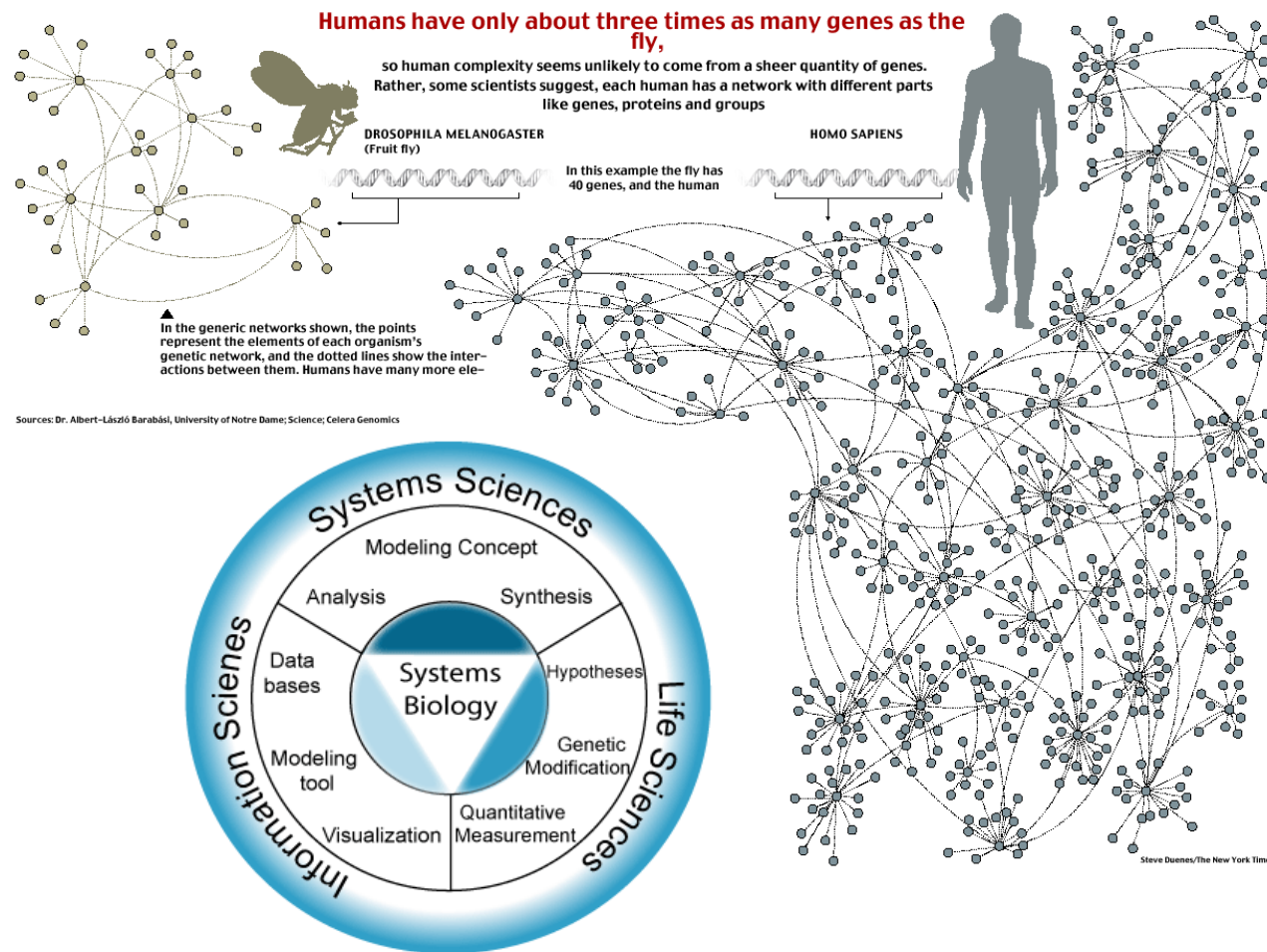
□卡尔·路德维希·冯·贝塔郎菲（Karl Ludwig von Bertalanffy）提出一般系统论（General Systems Theory），认为不同领域的系统都存在着一些共同的规律，这为从系统角度研究生物提供了新的理论基础

□由于支撑系统研究的大量元素（基因、蛋白等）通过组学研究才得以快速测定获得，所以真正的系统生物学一般认为是从21世纪开始

✓尤其是在莱诺伊·胡德（Leroy Hood）和北野宏明（Hiroaki Kitano）的引领下得到了发展

- 大量实验数据收集与整合
- 建立刻画该数据集的数学模型
- 数学方程的精确计算及数值预测
- 基于数值模拟与实验数据的模型质量评估

兼具生物学和信息科学的特点！

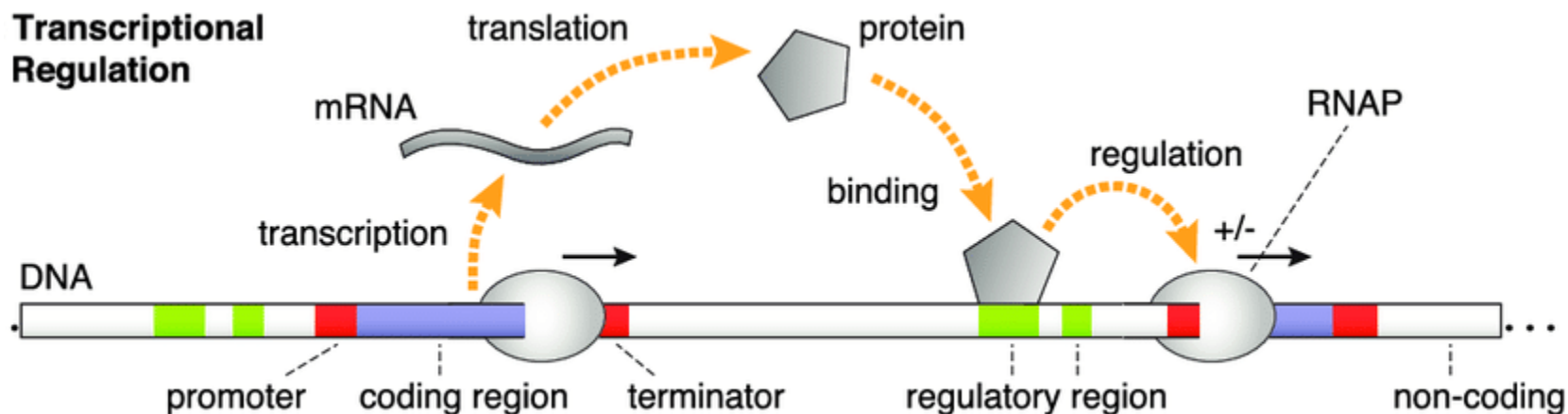


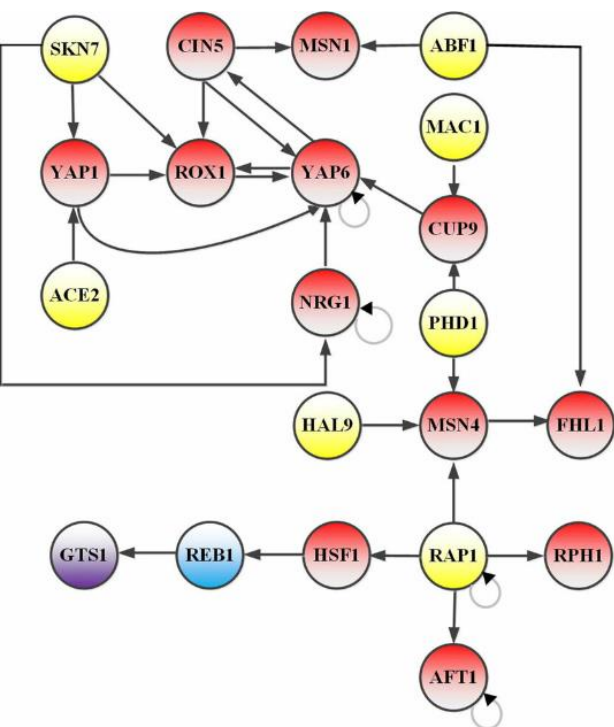
□生物分子网络通常由参与生物过程的众多分子元件组成

✓ 其中基因和蛋白质等分子是关键的构成要素

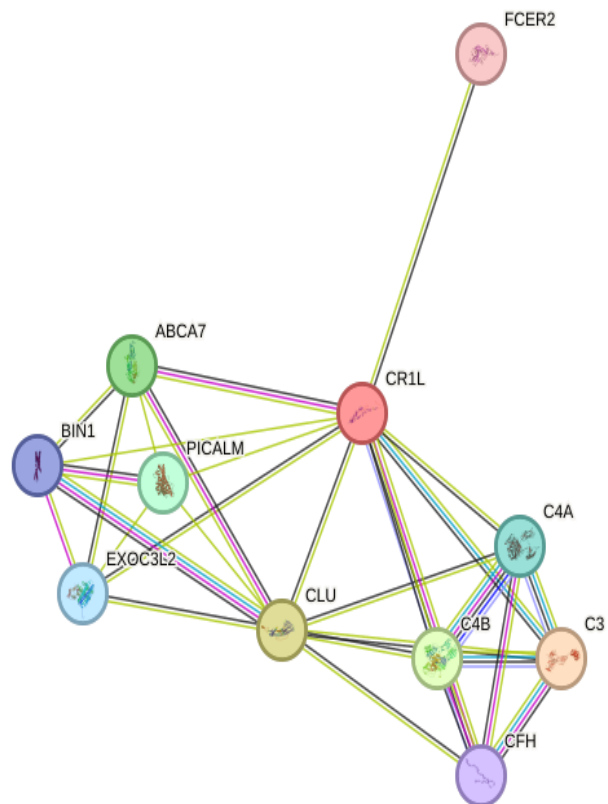
□从系统的角度来看，关键在于元件之间的相互关系，而非元件本身

✓ 在生物分子网络中，关系可以体现为分子与分子之间的相互作用，或某种化学反应

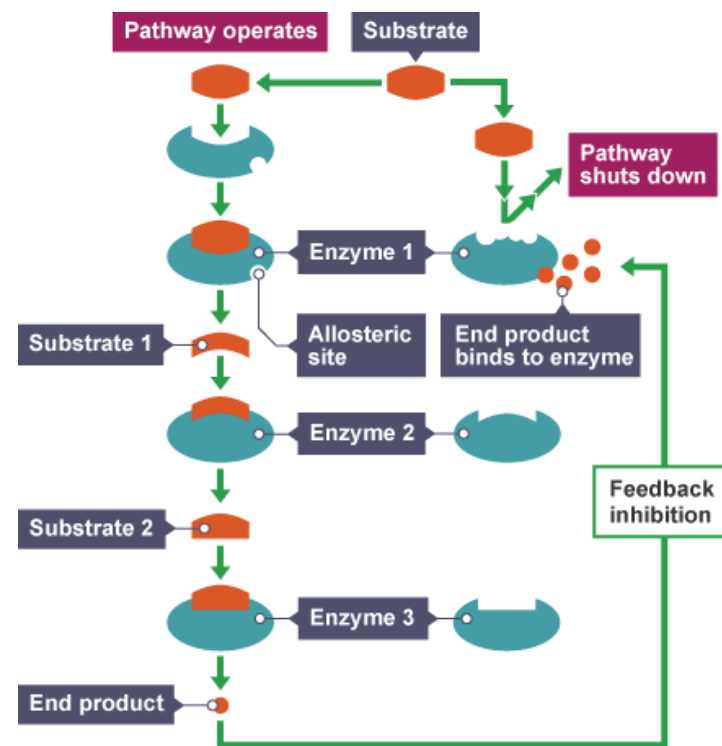




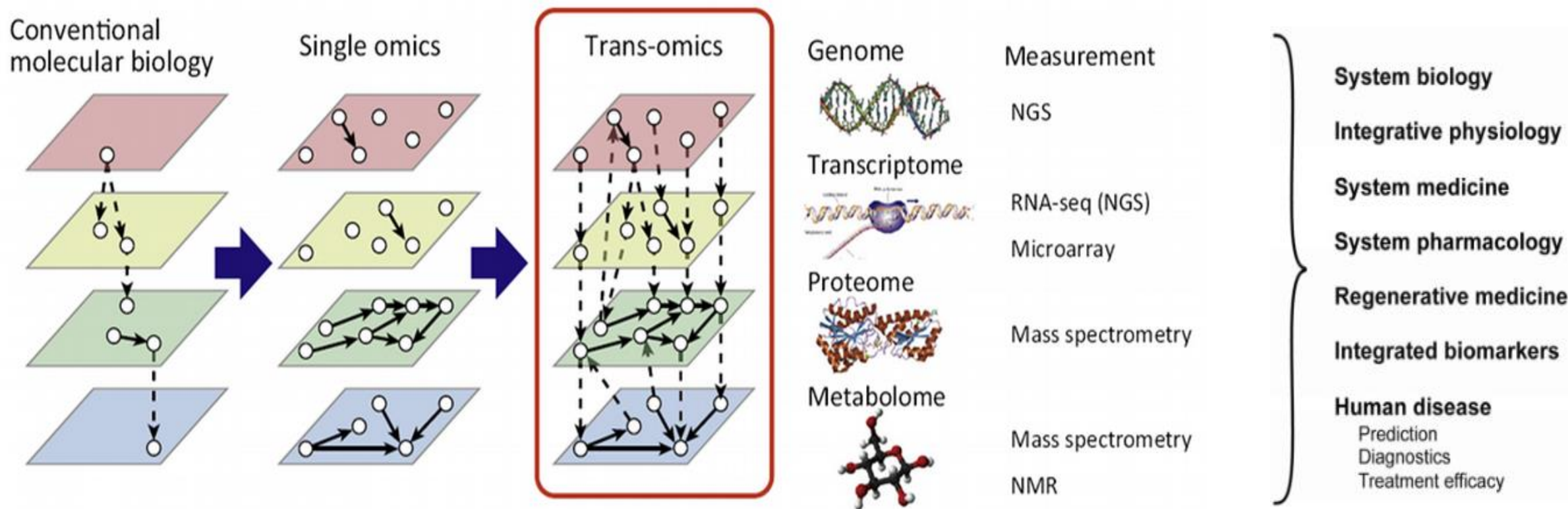
转录调控网络



蛋白互作网络



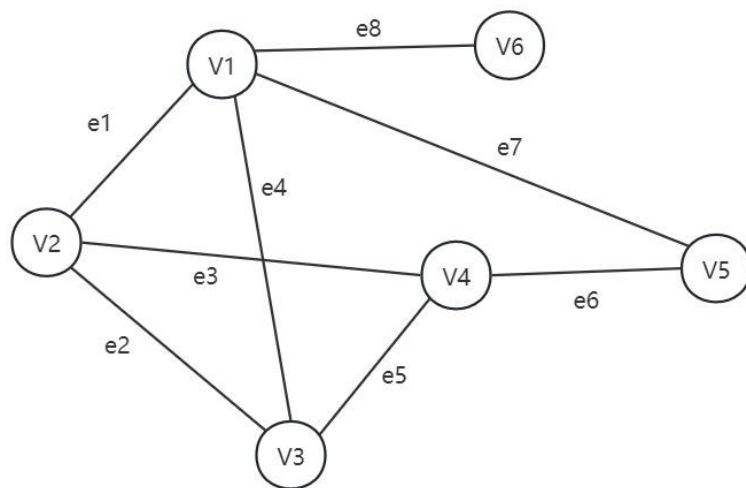
代谢网络



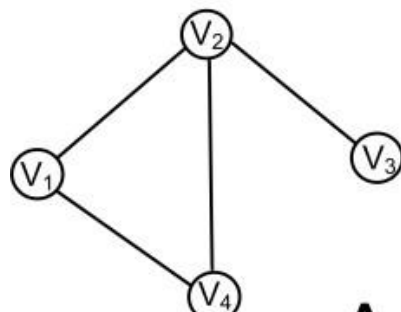
□ 通常用图 $G=(V, E)$ 表示网络(network)

✓其中 V 是网络中节点的集合，每个节点代表一个生物分子； E 是边的集合

□当 V 中的两个节点 $v1$ 与 $v2$ 之间存在一条属于 E 的边 $e1$ 时，称边 $e1$ 连接 $v1$ 与 $v2$ ，或者称 $v1$ 连接于 $v2$ ，也称作 $v2$ 是 $v1$ 的邻居

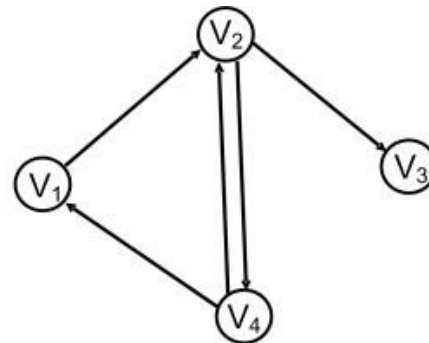


无向网络



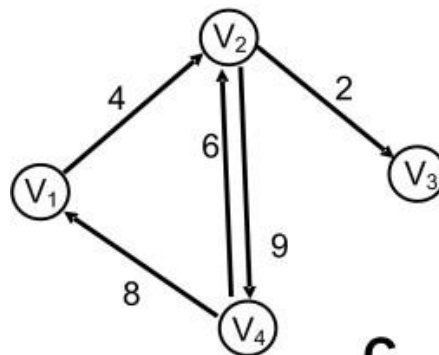
A

有向网络



B

加权网络

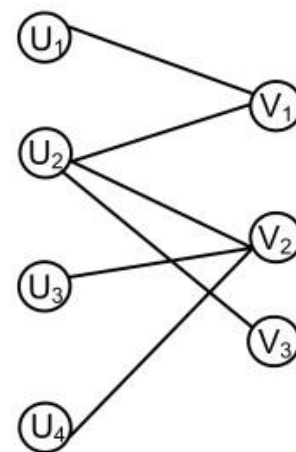


C

二分网络

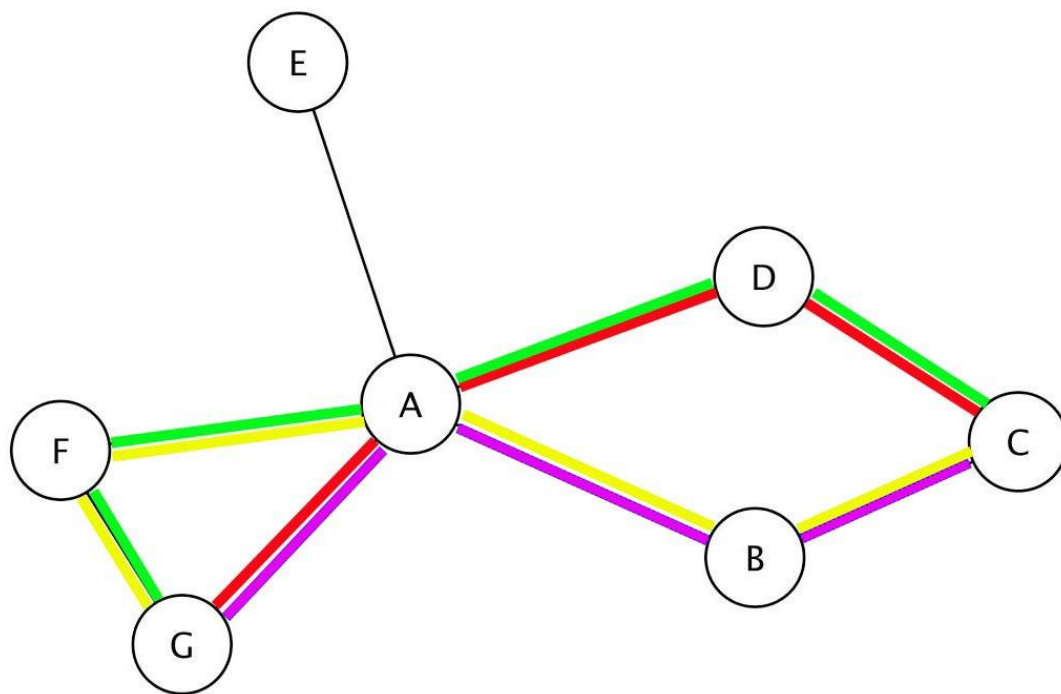
Set 1

Set 2



D

- 网络中的**路径**是指一系列节点，其中每个节点都有一条边连接到紧随其后的节点。
- 对包含节点数目有限的路径来说，第一个节点称为**起点**，最后一个节点称为**终点**，二者均可称为路径的**端点**
- 路径中经过边的权重之和称为路径的权重，也称为路径的**长度**
- 在连接两个节点的所有路径中，长度最短的路径称为**最短路径**，最短路径的长度称为从起点到终点的**距离**



网络中节点G到节点C的路径有：

$I_1 = \{G, A, B, C\}$

$I_2 = \{G, A, D, C\}$

$I_3 = \{G, F, A, B, C\}$

$I_4 = \{G, F, A, D, C\}$

从节点G到节点C的路径中， I_1 和 I_2 的长度为3， I_3 和 I_4 的长度为4
从节点G到节点C的距离为3

□基因调控网络

- ✓TRANSFAC、TRRD、RegulonDB和AnimalTFDB
- ✓JASPAR、DBD

□蛋白质互作网络

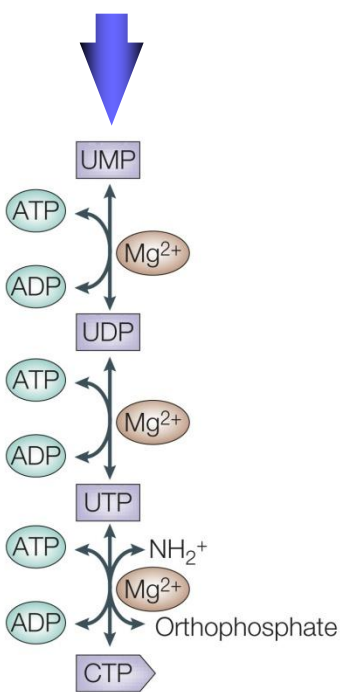
- ✓蛋白质互作网络是研究最充分的生物分子网络，往往也是规模最大的生物分子网络，常常包含数千甚至上万个节点以及为数更多的边
- ✓BIND、DIP、MIPS、BioGRID和String等

□代谢网络和信号传导网络

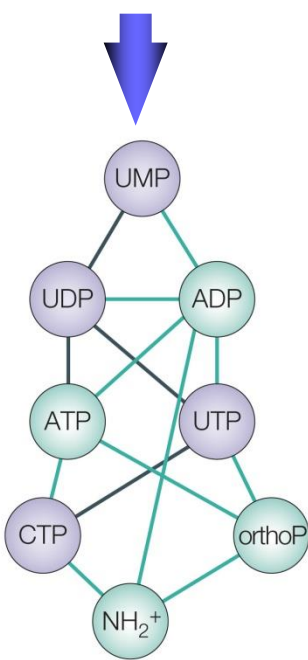
□细胞间通讯网络

- 代谢通路** 指细胞中代谢物在酶的作用下转化为新的代谢物过程中所发生的一系列生物化学反应
- 代谢网络** 指由代谢反应以及调节这些反应的调控机制所组成的描述细胞内代谢和生理过程的网络
- 信号传导** 指细胞将一种类型的生物信号或刺激转换为其他生物信号最终激活细胞反应的过程
- 信号传导网络** 指参与信号传导通路的分子和酶以及其间所发生的生化反应所构成的网络

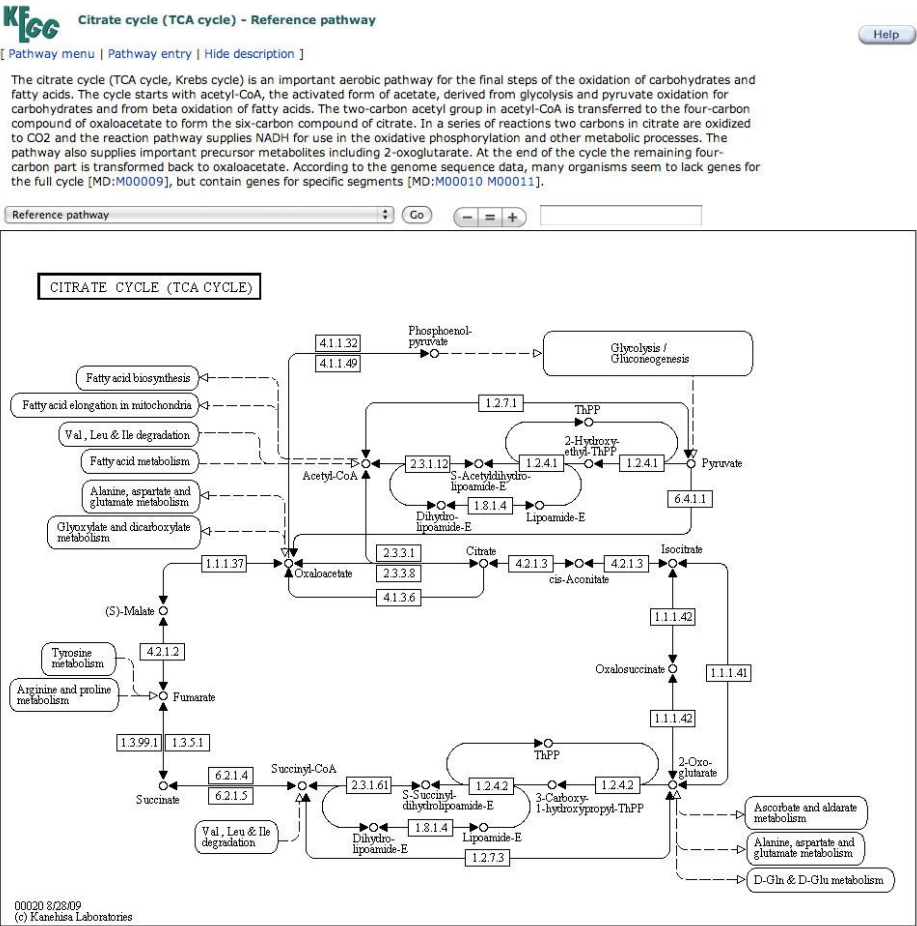
完全网络



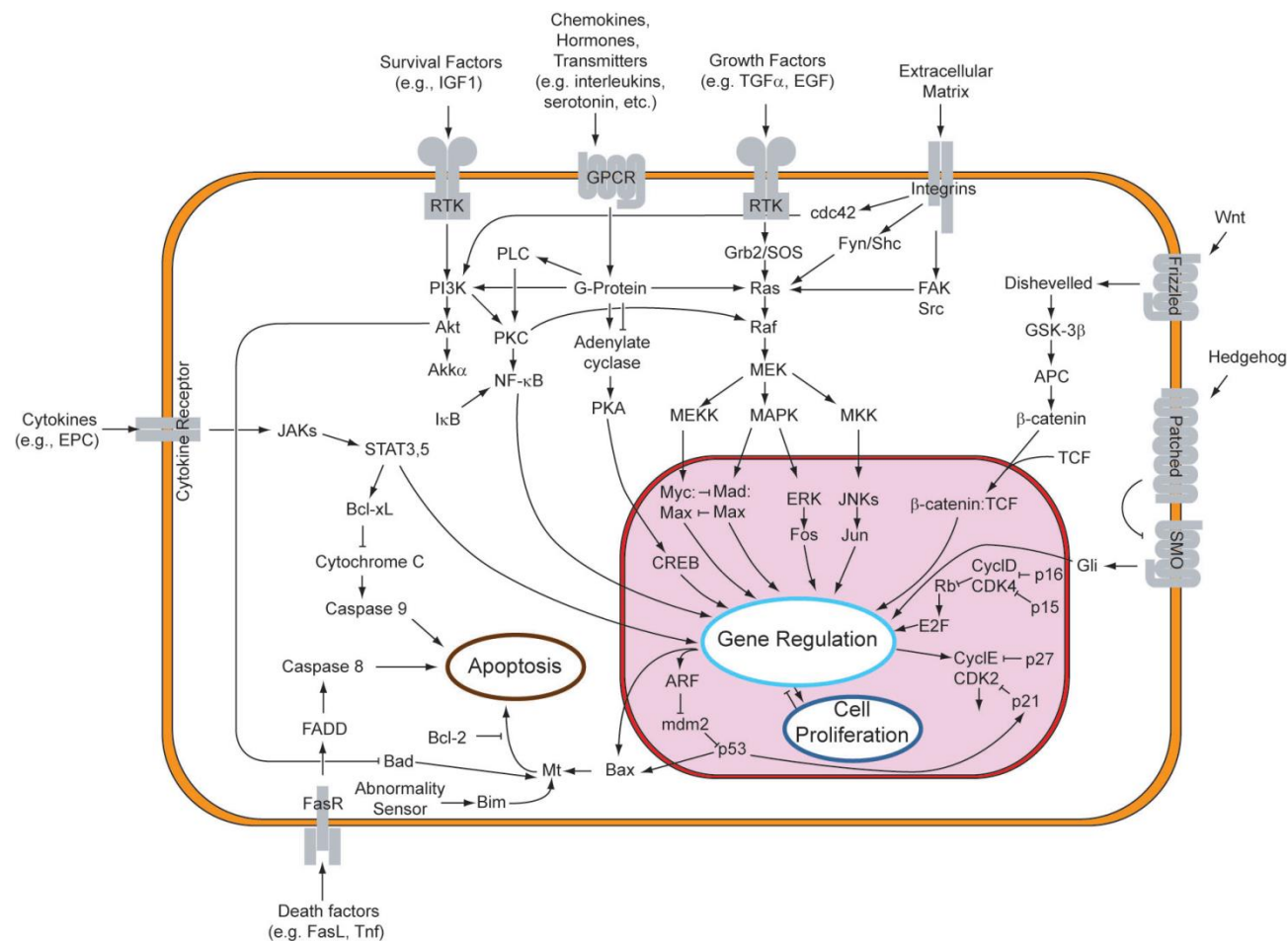
多反应物网络



主要反应物网络



信号传导通路



- ❑ 生物信号的传递不仅发生在细胞内部，同样也出现在不同细胞之间
- ❑ 细胞间通讯是生物体调节细胞基本活性、协调细胞活动的信息传递机制

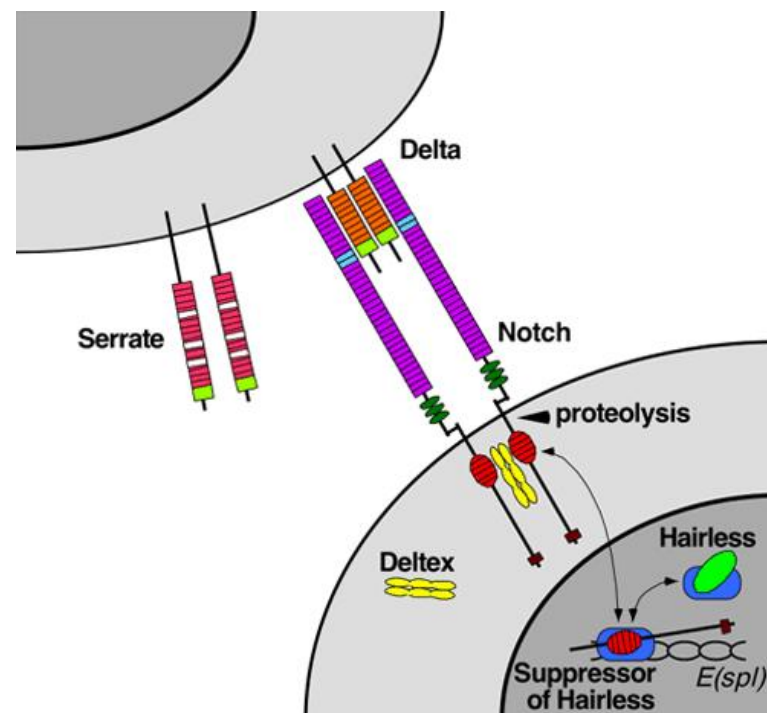
根据细胞间的距离可以分为：

直接接触（邻分泌通讯）

近距离通讯（旁分泌通讯）

远距离通讯（内分泌通讯）

自分泌通讯



□节点v的**连通度** (degree) 是指网络中直接与v相连的边的数目

✓连通度是描述单一节点的最基本的拓扑性质

✓对于有向网络往往还要区分边的方向，由节点v发出的边的数目称为节点v的**出度**，指向节点V的边数则称为节点V的**入度**

□在生物分子网络中，连通度较大的节点往往发挥着重要的作用，可以作为一种重要的指标来识别生物分子网络中的关键节点

✓**中心节点** (hub, 连通度较高的节点) 往往是必需基因或其翻译产物的编码基因，它们对于支持生命基本活动至关重要

□网络中**聚类系数** (clustering coefficient, CC) 是用来衡量网络中三角关系数量的指标

✓如果节点v1连接于节点v2，节点v2连接于节点v3，那么节点v3很可能与v1相连接。这种现象被称为三角关系，它体现了部分节点间存在的密集连接性质。

□对于无向网络，节点v的聚类系数定义为：

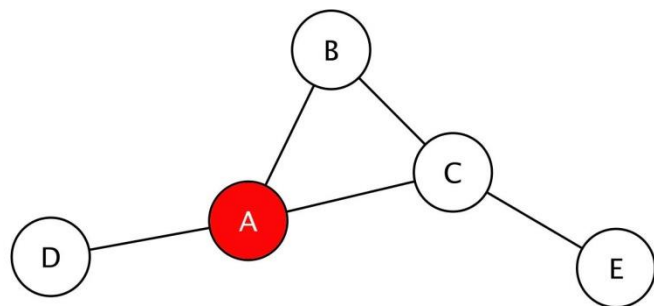
$$CC_v = \frac{n}{C_k^2} = \frac{2n}{k(k-1)}$$

□对于有向网络，聚类系数定义为：

$$CC_v = \frac{n}{P_k^2} = \frac{n}{k_{\text{out}}(k_{\text{out}} - 1)}$$

□网络中**聚类系数** (clustering coefficient, CC) 是用来衡量网络中三角关系数量的指标

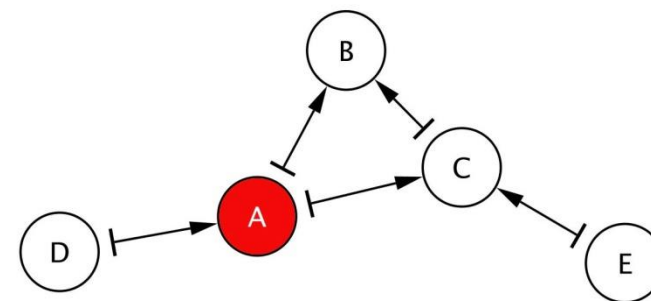
无向网络



节点A的连通度为3

$$CC_v = \frac{n}{C_k^2} = \frac{2n}{k(k-1)} \quad CC_A = \frac{2 \times 1}{3 \times (3-1)} = \frac{1}{3}$$

有向网络



节点A的入度为1，出度为2

$$CC_v = \frac{n}{P_k^2} = \frac{n}{k_{out}(k_{out}-1)} \quad CC_A = \frac{1}{2 \times (2-1)} = \frac{1}{2}$$

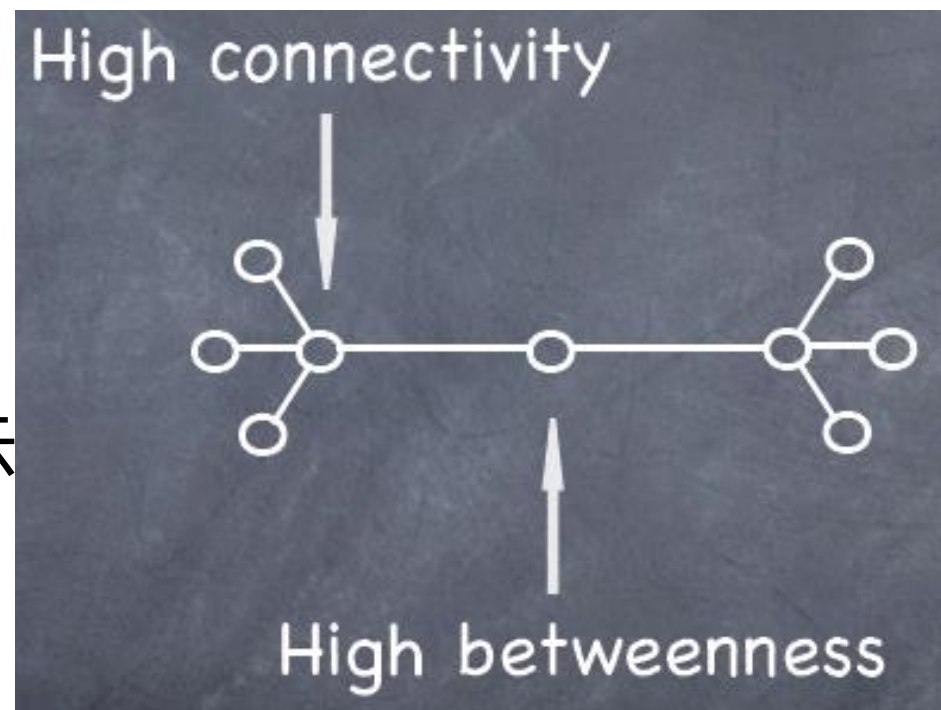
□ 一个节点的**介数** (betweenness) 是衡量这个节点出现在其他节点间最短路径中的比例

□ 节点v的介数定义为:

$$B_v = \sum_{i \neq j \neq v \in V} \frac{\sigma_{ivj}}{\sigma_{ij}}$$

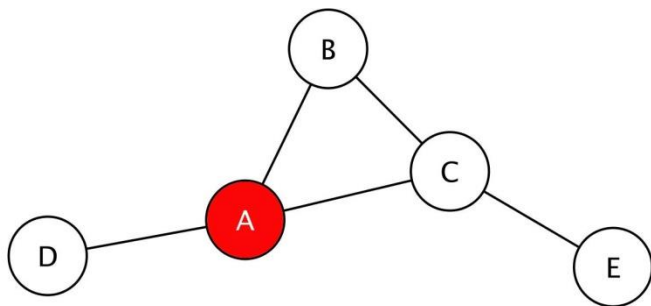
□ 标准化至[0, 1]区间的形式表示

$$B_v = \frac{1}{(N-1)(N-2)} \sum_{i \neq j \neq v \in V} \frac{\sigma_{ivj}}{\sigma_{ij}}$$



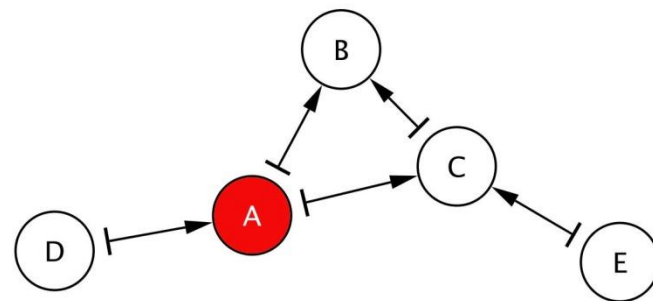
□ 一个节点的**介数** (betweenness) 是衡量这个节点出现在其他节点间最短路径中的比例

无向网络



有{B, A, D}, {C, A, D}, {D, A, C, E}以及它们的逆序路径共6条最短路径通过节点A, 节点A的介数为6。

有向网络



真正连通的路径只有{C, B}, {D, A, B}, {D, A, C}, {D, A, C, B}, {E, C}, {E, C, B}。其中经过节点A的路径有2条, 则节点A的介数为2。

□ **紧密度**(closeness)是描述一个节点到网络中其他所有节点平均距离的指标

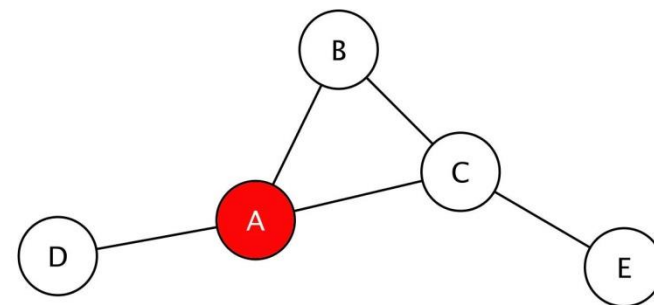
□ 节点v的紧密度定义为:

$$C_v = \frac{1}{N-1} \sum_{j \neq v \in V} d_{vj}$$

□ **拓扑系数**(topology coefficient)是反映互作节点间共享连接比例的测度

□ 节点v的拓扑系数定义为:

$$T_v = \frac{1}{|M_v|} \sum_{t \in M_v} C_{v,t} / \min \{k_v, k_t\}$$



节点A到B、C、D、E的距离分别为1、1、1、2。
节点A的紧密度为1.25。

与A节点共享邻居的节点共有3个，则 $M_A = \{B, C, E\}$ 则节点A的拓扑系数

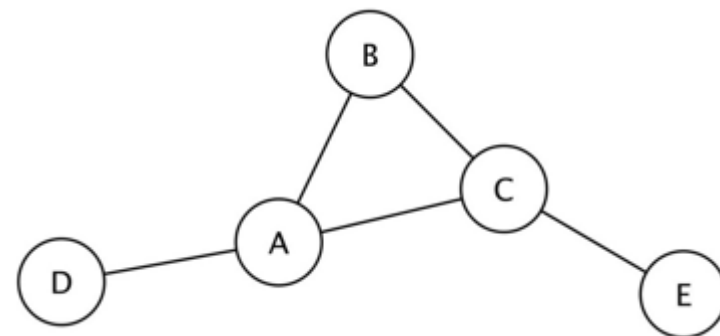
$$T_A = \frac{1}{3} \left(\frac{1}{2} + \frac{1}{3} + 1 \right) = \frac{11}{18}$$

□直径(diameter)指网络中任意两个连通节点间距离的最大值，是描述网络总体性质的一个属性

✓代表了网络中节点连接可能出现的最远距离，标志着网络紧密的程度

□网络的平均距离是指网络中任意两个连通节点距离的平均值，是描述网络总体性质的一个属性，也是衡量网络紧密程度的重要指标

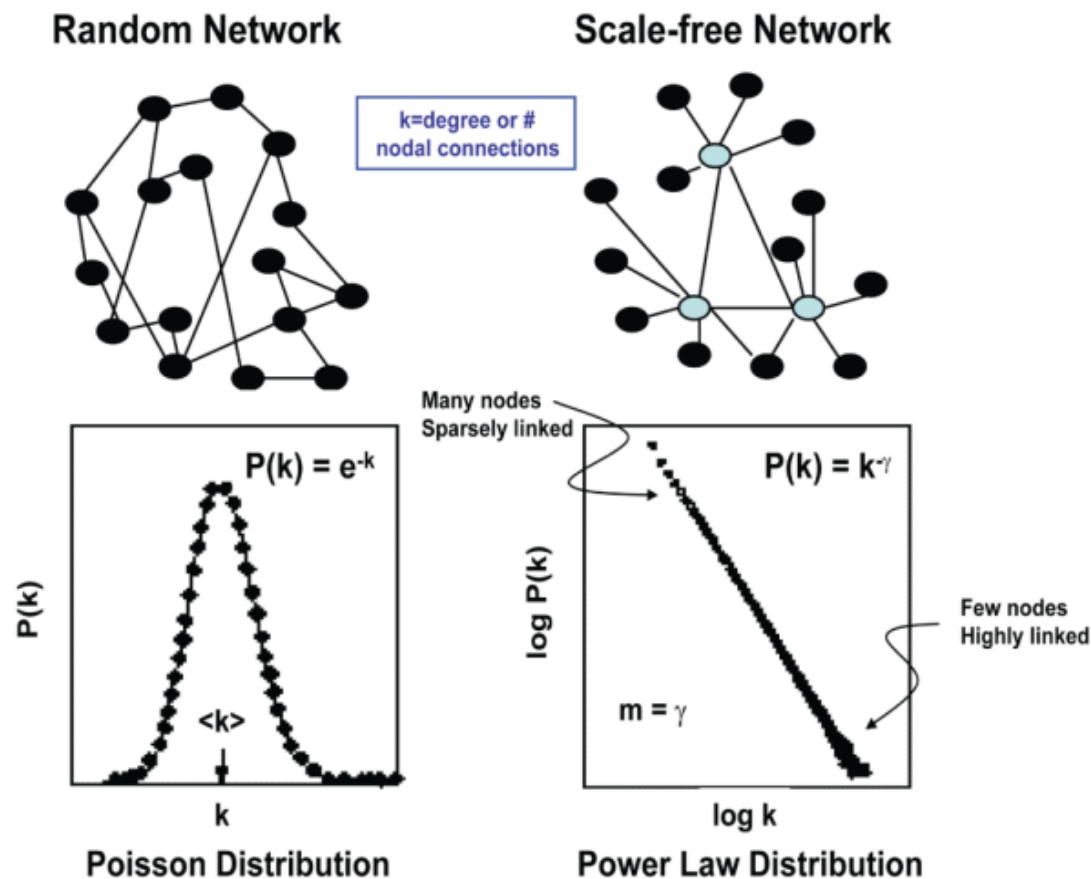
□连通度的分布函数和聚类系数分布函数



D与E间的距离最远，则网络直径为3

无标度 (scale free) 网络，是指网络中连通度的分布符合幂律分布 $P(k) \sim k^{-\gamma}$ 的网络，通常 $2 < \gamma < 3$

- ✓ 无标度网络中大部分节点的连通度较低，但存在少数连通度非常高的枢纽节点 (hub) 使网络连接在一起
- ✓ 枢纽节点的存在使得无标度网络对**意外故障**具有很强的鲁棒性。因为即使少数枢纽节点失效，网络整体的连通性仍然能够保持
- ✓ 无标度网络也因此对**协同攻击**更加脆弱。因为攻击者只需要集中攻击少数枢纽节点，就能够使整个网络瘫痪

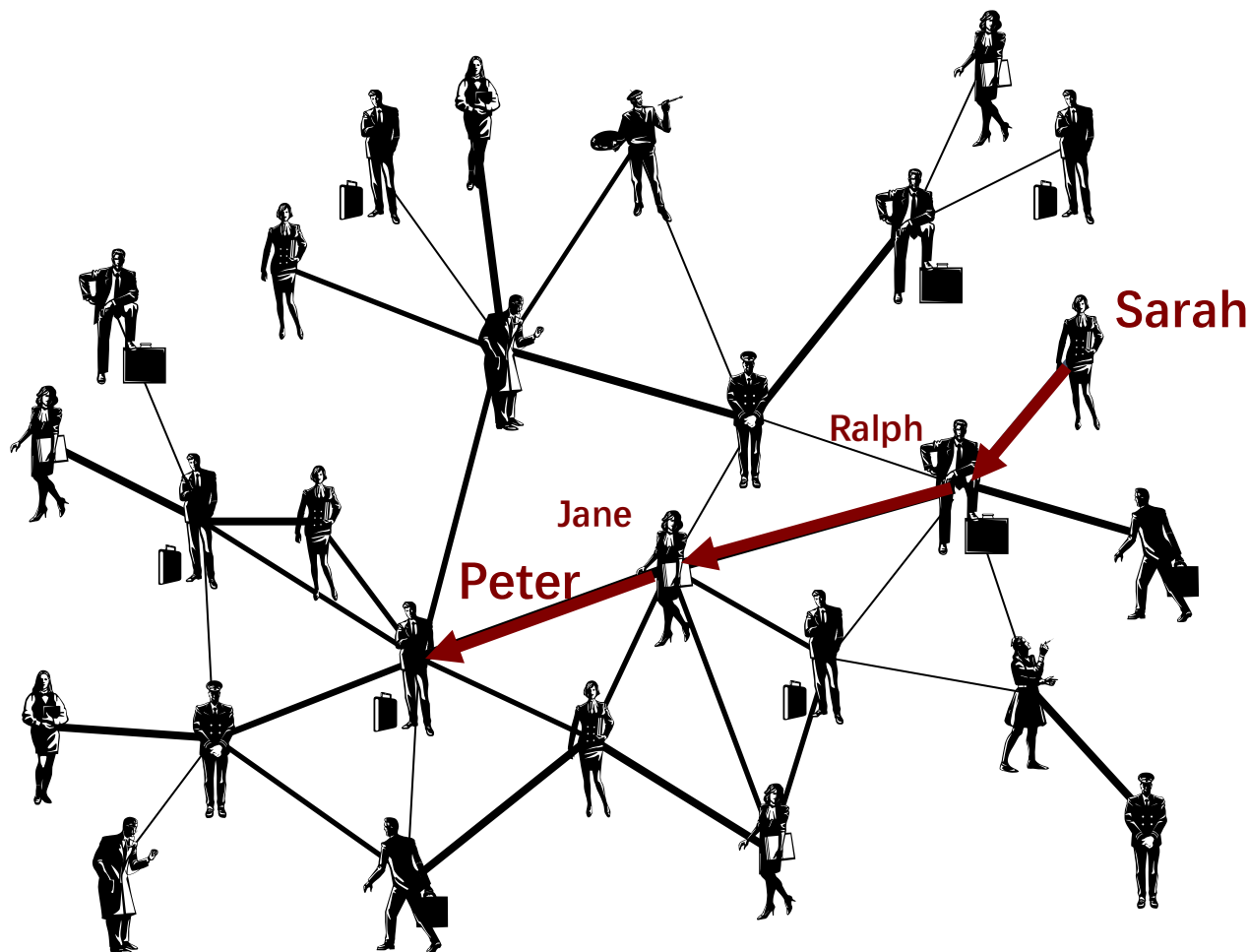


□ **小世界网络** 中直径或网络平均距离与节点数的对数成正比 ($\sim \log(N)$)

□ **大世界网络** 的平均距离正比于 N

□ 无标度网络的直径相对较小，一般来说无标度网络直径的大小正比于网络中节点数目的对数值的对数值即 $\sim \log(\log(N))$

✓ 无标度网络比一般小世界网络直径更小，联系更紧密

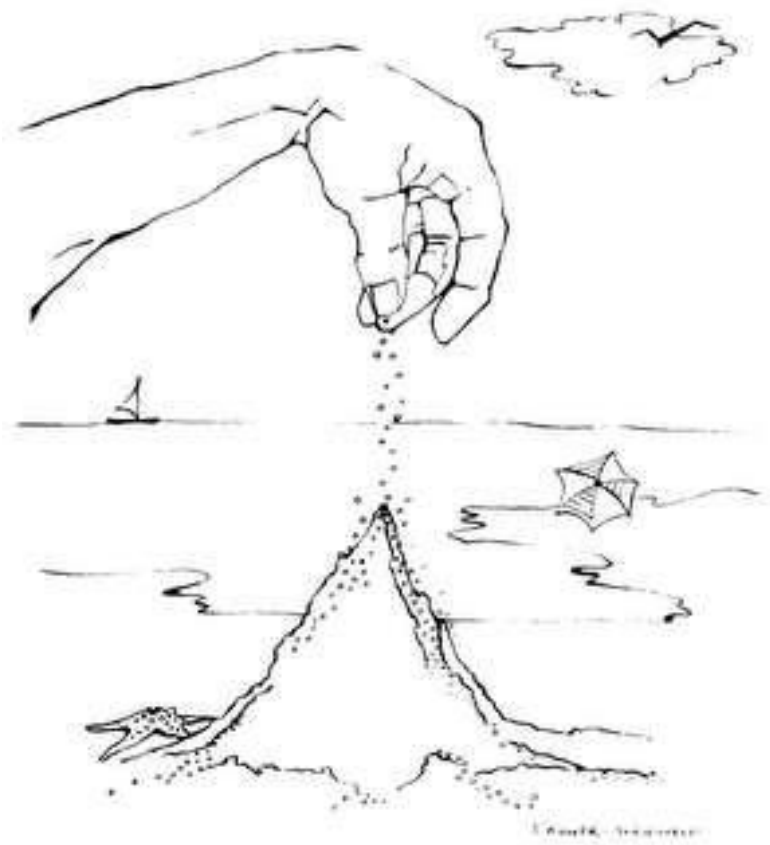


社会网络:
6度分离

WWW:
16次点击分离

□Barabási-Albert模型

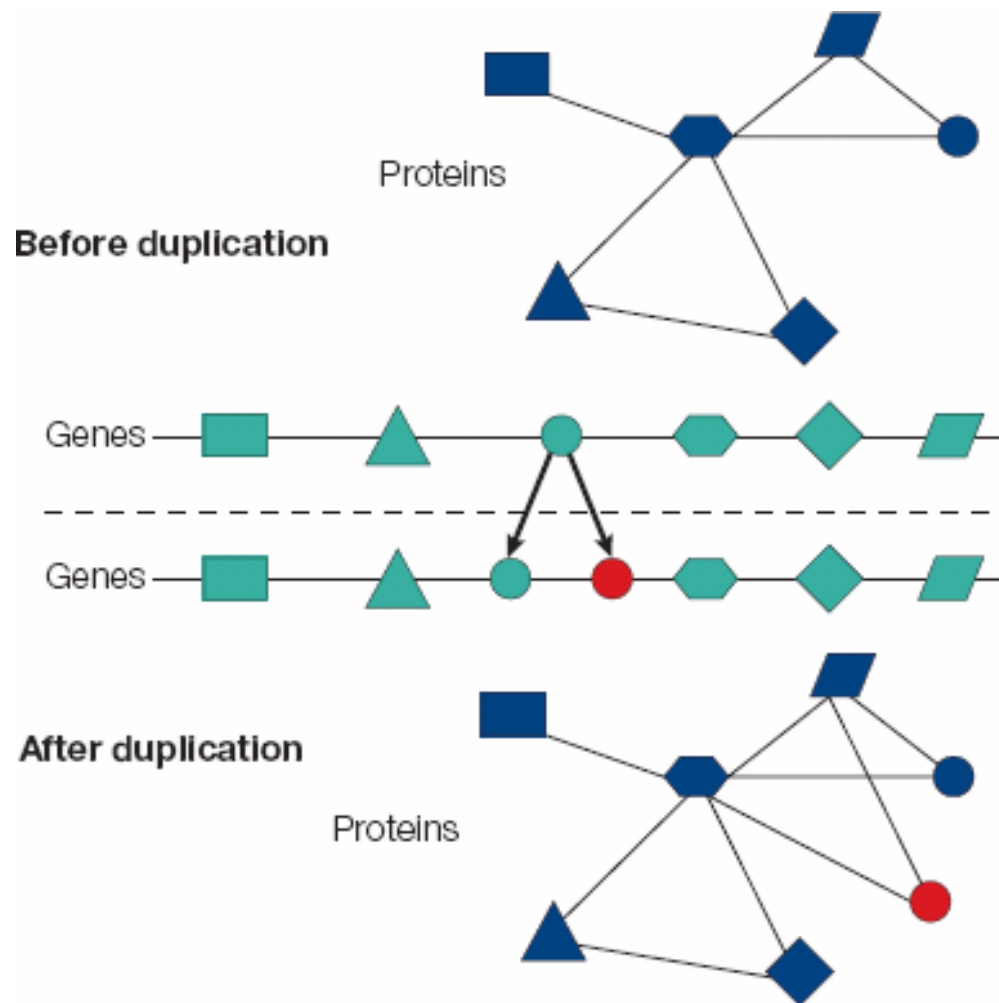
- 首先从一个包含 m_0 个节点的网络开始，其中 $m_0 \geq 2$
初始网络中每个节点的连通度都应该大于零，否则在后续过程中将无法与网络连接
- 而后，通过一个循环过程扩大网络，在每次循环中只增加一个节点，并依次按照概率 $\pi_i = \frac{k_i}{\sum_v k_v}$ 决定是否与原网络中节点 i 建立连接，其中 k_i 是节点 i 的连通度
- 因此，原有连通度较高的节点将更有机会与新加入节点连接，从而获得更高的连通度



□ Barabási-Albert模型

□ 蛋白质互作网络中出现无标度特性的原因在于基因复制

- ✓ 在细胞分裂过程中复制产生的基因产物会与相同的蛋白发生相互作用。
- ✓ 与发生复制的蛋白连接的蛋白节点将会获得新的连接
- ✓ 因此，高度连接的节点更有可能与发生复制的基因产物发生互作，从而获得额外的连接，最终形成蛋白互作网络的无标度特性

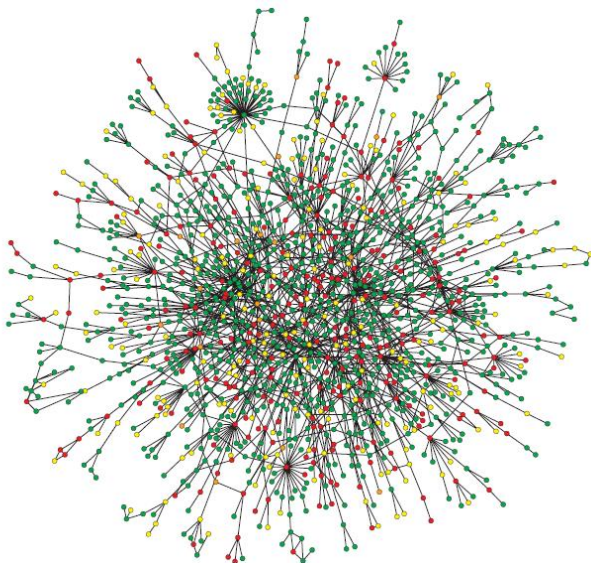


- 诱饵-猎物模式的蛋白质相互作用检测方式?
- 目前还远未完善的数据资源?

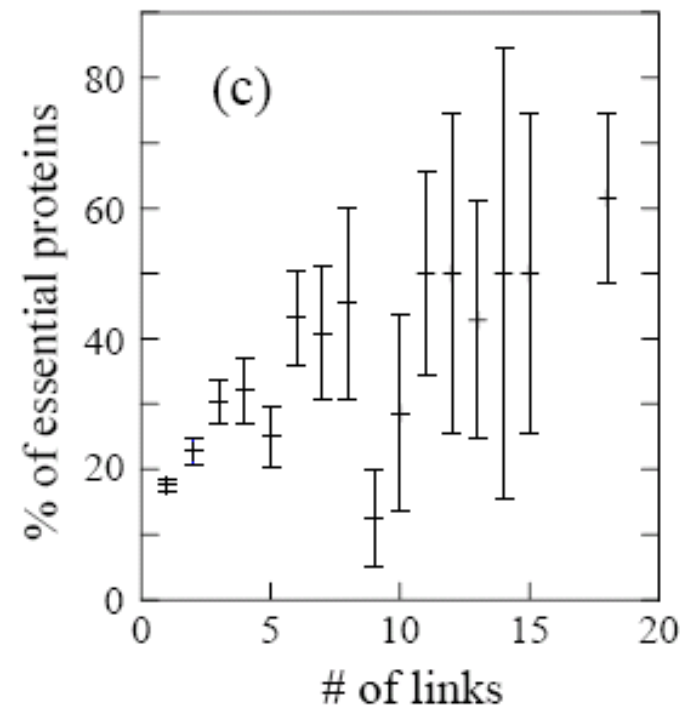


□ Hub节点主导着网络

□ 大量低连接度节点的意义呢？



red = lethal, green = non-lethal
orange = slow growth, yellow = unknown

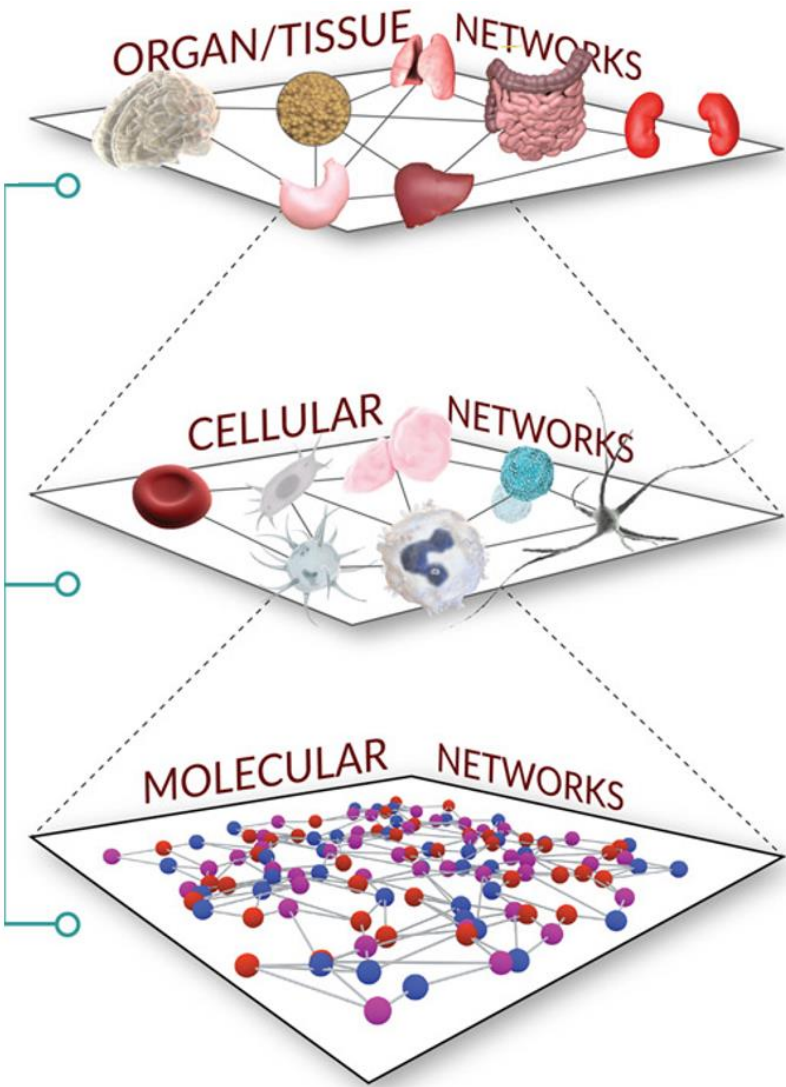


while proteins with five or less links constitute ~93% of the total number of proteins we find that only ~21% of them are essential. In contrast, only ~0.7% of the yeast proteins with known phenotypic profile have more than 15 links but single deletion of ~62% of these proves lethal.

□网络的“**模块**”(module)指网络间的节点存在着内部彼此高度连接的子节点集合，它们由许多分子相互结合形成的，有着稳定结构和功能的复合体

□模块化的网络连通更为紧密

- ✓与同样规模的随机网络相比，虽然拥有相同的节点数与边数，模块化网络的连接却更为密集
- ✓这一现象可以由聚类系数 CC 的提高表现出来



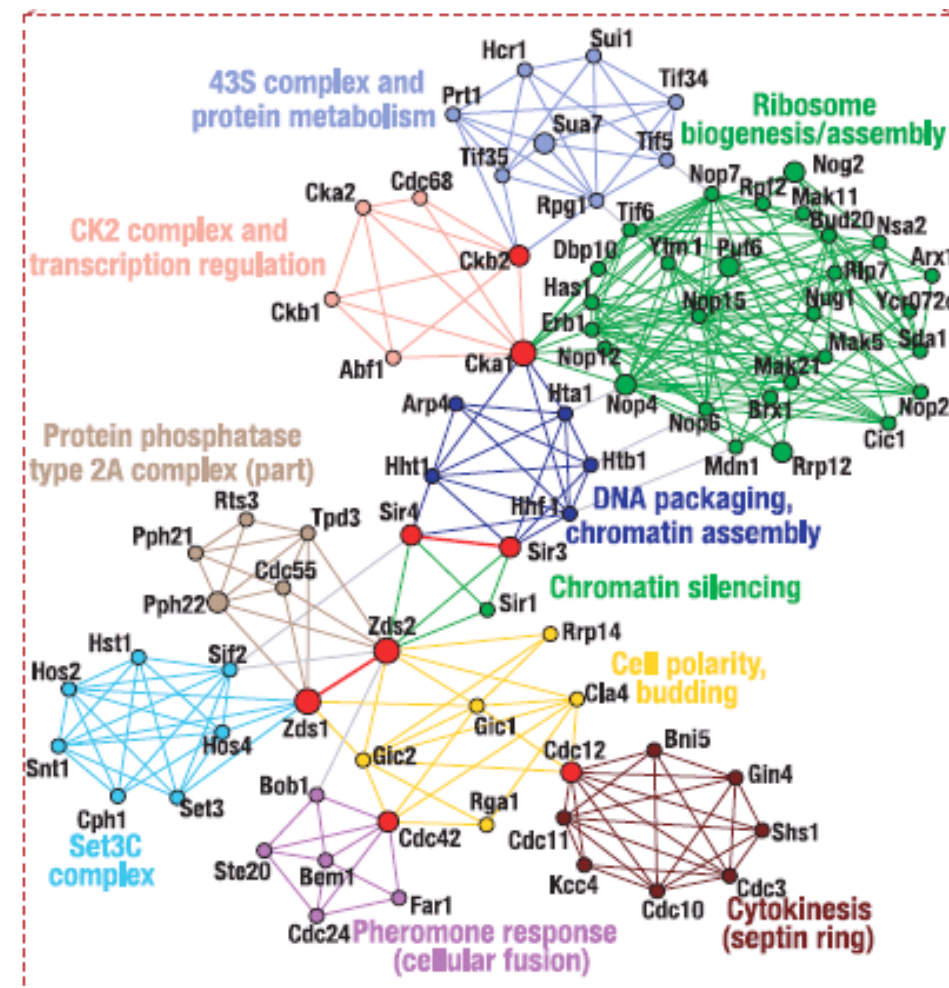
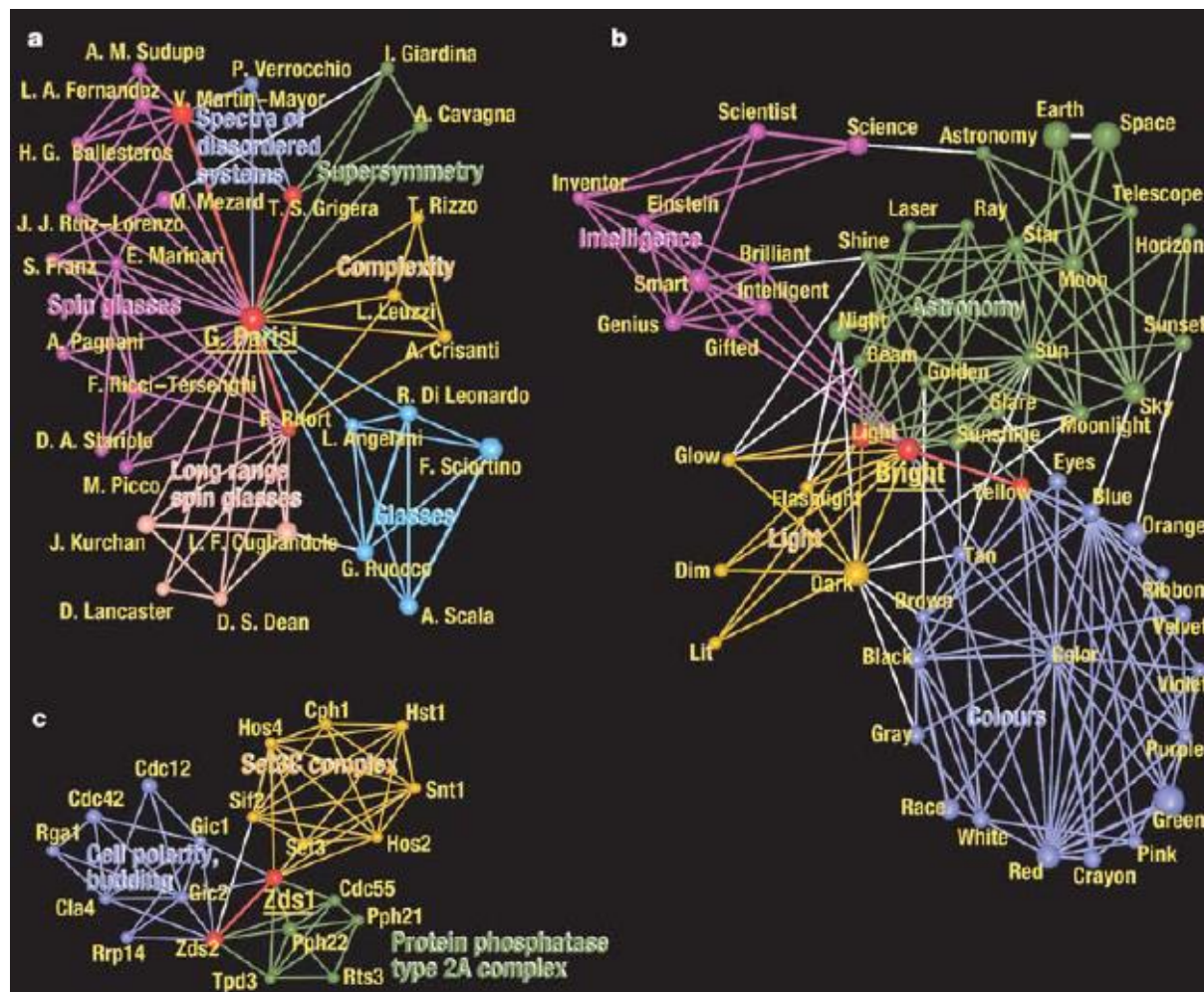
□模块化的设计使得生物系统具有以下优点：

- ✓**提高效率**：模块化可以将复杂的功能分解成更小的、更易于管理的子功能，从而提高系统的效率
- ✓**增强鲁棒性**：模块化可以使系统更加鲁棒，因为即使其中一个模块出现故障，其他模块仍然可以正常工作
- ✓**促进进化**：模块化可以促进系统的进化，因为新的模块可以通过组合现有的模块来形成

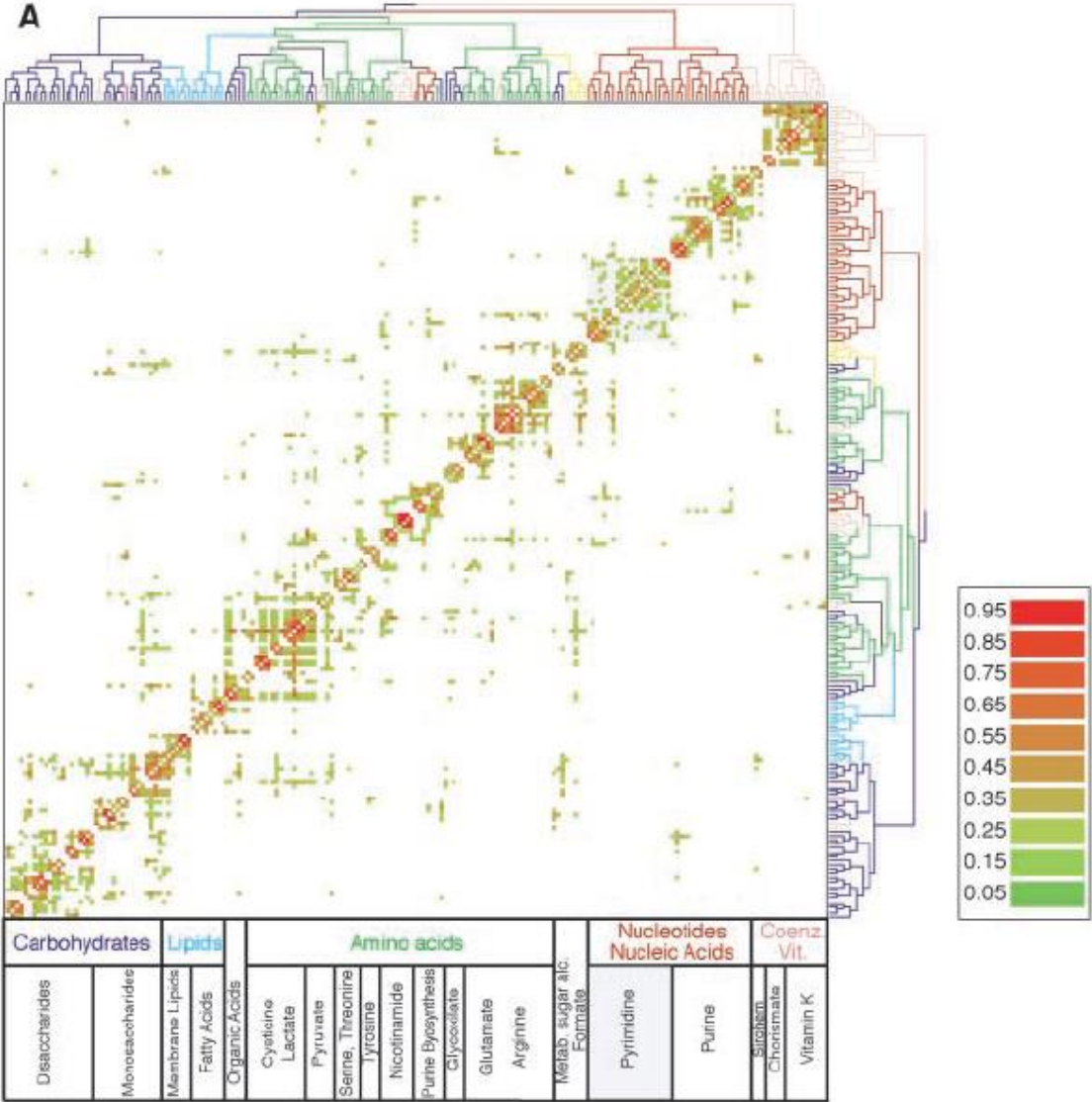
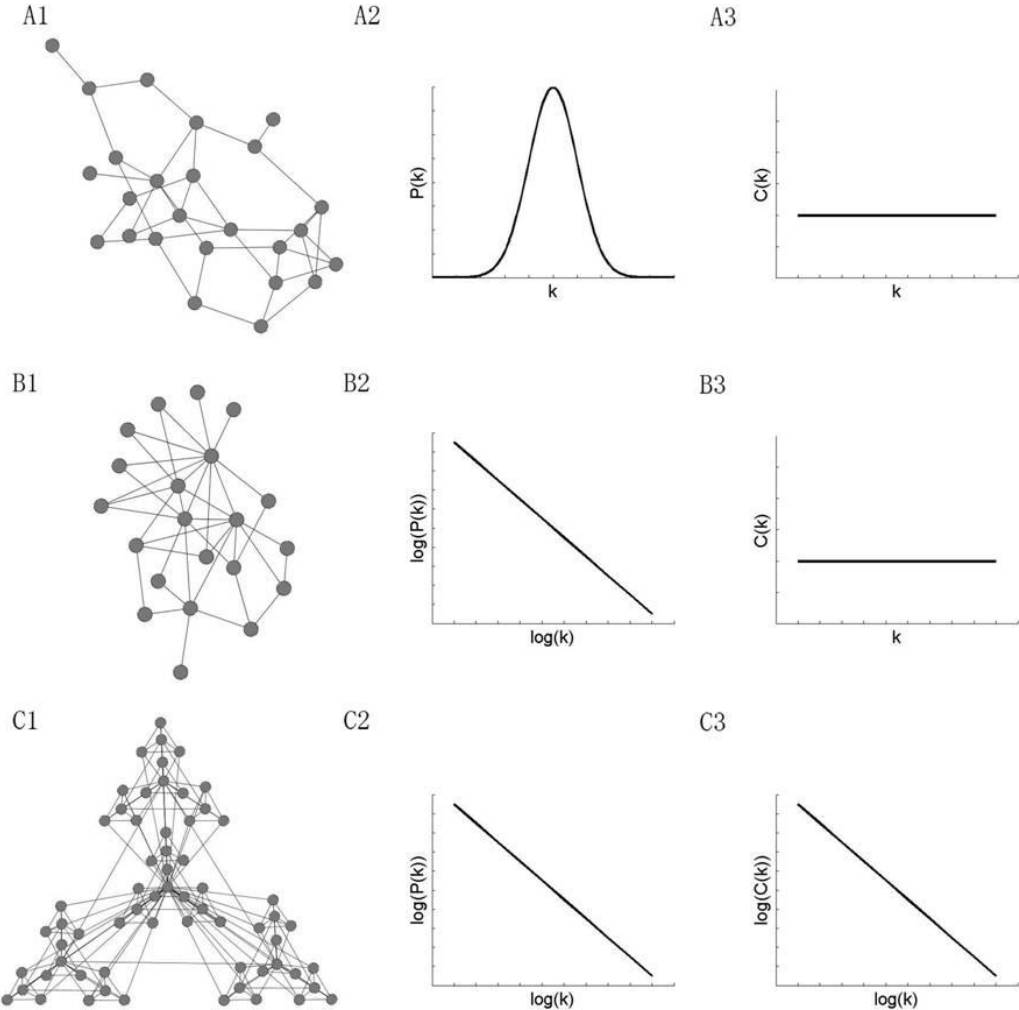
□功能模块（functional modules）的提出

- ❑ 功能模块，通过定义来看， 是“一个独立的实体， 且它的功能能和其他模块区分开来。”
- ❑ 这种区分来自于空间的独立性 (**Spatial localization**) 或化学上的独立性 (**Chemical specificity**)
- ❑ 模块之间存在链接， 这种连接性 (**Connectivity**) 允许一个不同功能之间相互影响
- ❑ 存在孤立性 (**Insulation**) ， 但允许“ the cell to carry out many diverse reactions without cross-talk that would harm the cell.”

- 模块是内部紧密连接而与网络其他部分松散链接的单元
- 模块的结构不一定是严格的、固定的
- 模块的功能可以根据环境的不同发生变化
- 同一个组分，可以参与多个模块，并在不同模块中发挥不同的功能
- 模块是有层次的
 - ✓ 如有丝分裂…（纺锤体的组装、染色体的排列、细胞周期振荡器）



- 很多真实网络当中同时具备模块性、无标度性以及局部高连接性的特征，学者提出节点整合成为网络的过程类似一个循环迭代的过程，从而使网络成为一个层次性网络
 - ✓ 在此类网络中，聚类系数函数 $C(k)$ 正比于 k 的倒数
 - ✓ 大部分的真实生物网络如代谢网络、蛋白质互作网络等都是无标度网络，其网络平均聚类系数都比具有同样大小和连通度分布的随机网络更高，且聚类系数均值正比于连通度的倒数，从而表明**层次性也是生物网络的一项基本性质**



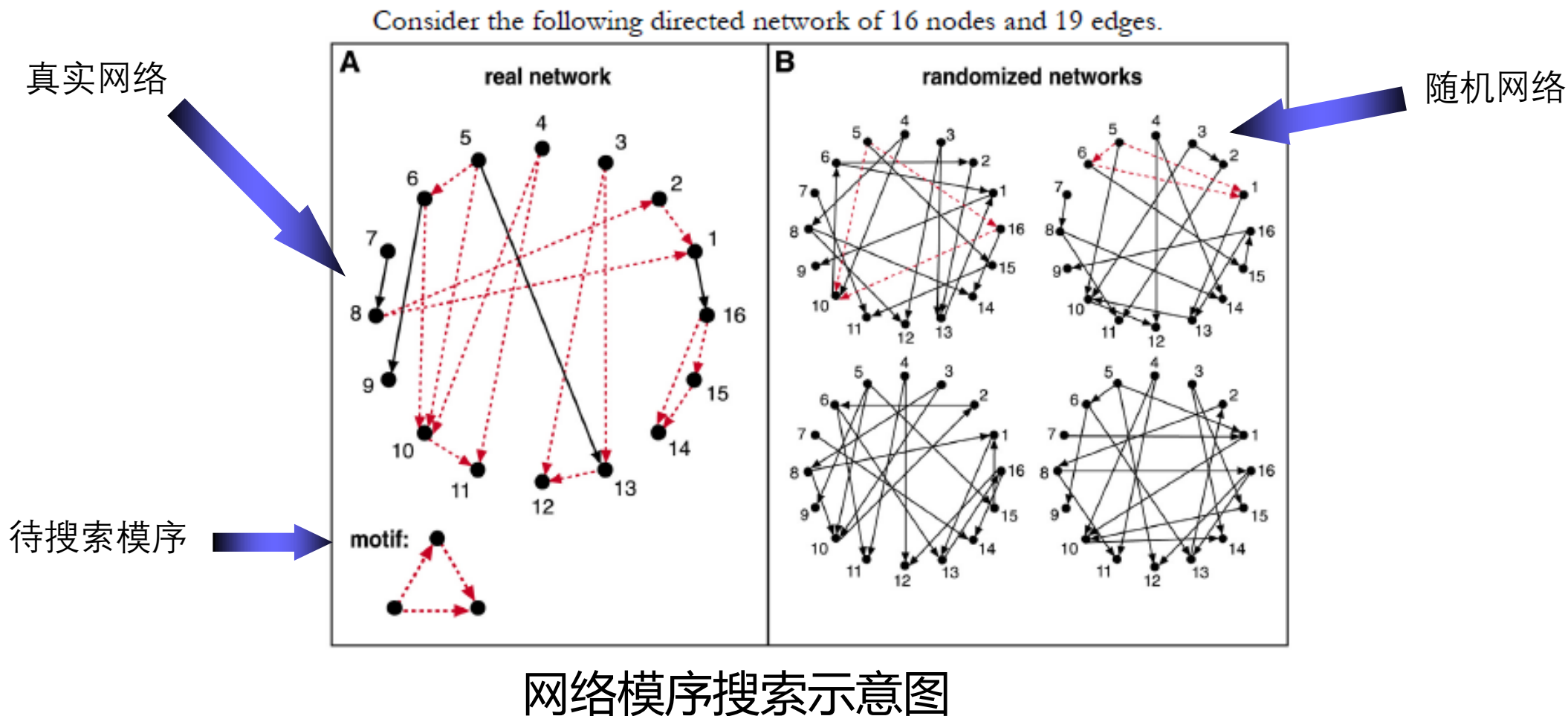
□ **网络模体(network motif)**是指网络中出现次数显著超过随机网络中出现次数的子网模式

✓ 这里子网模式是指一组节点按照特定的顺序连接而成的结构

□ 网络中不同的子网模式出现的频率并不一致

✓ 有些模式在网络中频繁出现，远远超过随机网络中期望出现的次数

□ 在某些网络中，特定出现的模体甚至是整个网络的基本组织形式，网络可以被看作是这些网络模体的组合



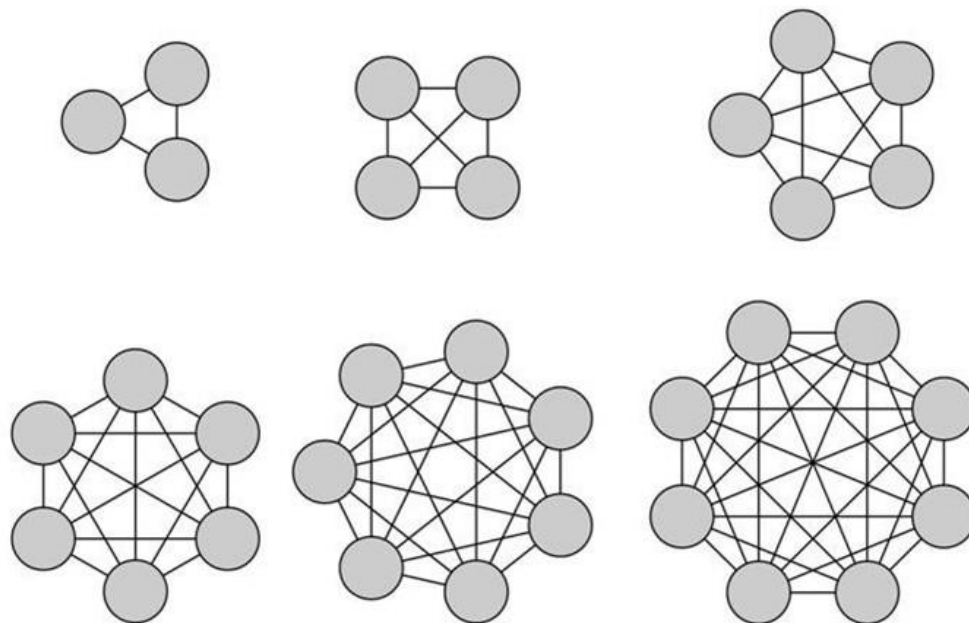
待搜索模式	→	MOTIF ID	NREAL	NREAL	NREAL	CREAL	UNIQ
真实网络中出现次数	→		STATS	ZSCORE	PVAL	[MILI]	
随机网络中出现平均次数及标准差	→						
Z值统计量	→						
显著性P值	→						

Full list of subgraphs size 3 ids:							
(Total num of different subgraphs size 3 is : 13)							
6	3	7.4+0.6	-6.93	1.000	142.86	2	
12	10	13.9+1.1	-3.44	1.000	476.19	2	
24	0	0.0+0.0	888888	0.000	0.00	0	
36	3	7.4+0.6	-6.93	1.000	142.86	1	
38	5	0.6+0.6	6.93	0.000	238.10	4	
46	0	0.0+0.0	888888	0.000	0.00	0	
74	0	0.0+0.0	888888	0.000	0.00	0	
78	0	0.0+0.0	888888	0.000	0.00	0	
98	0	0.2+0.4	-0.43	0.160	0.00	0	
102	0	0.0+0.0	888888	0.000	0.00	0	
108	0	0.0+0.0	888888	0.000	0.00	0	
110	0	0.0+0.0	888888	0.000	0.00	0	
238	0	0.0+0.0	888888	0.000	0.00	0	

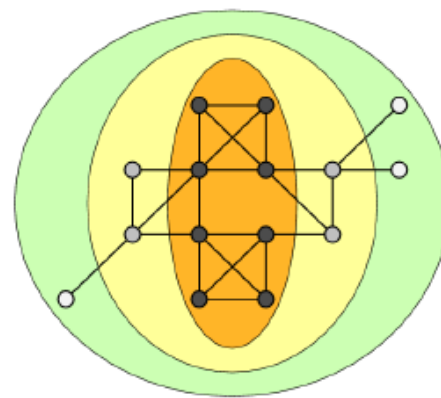
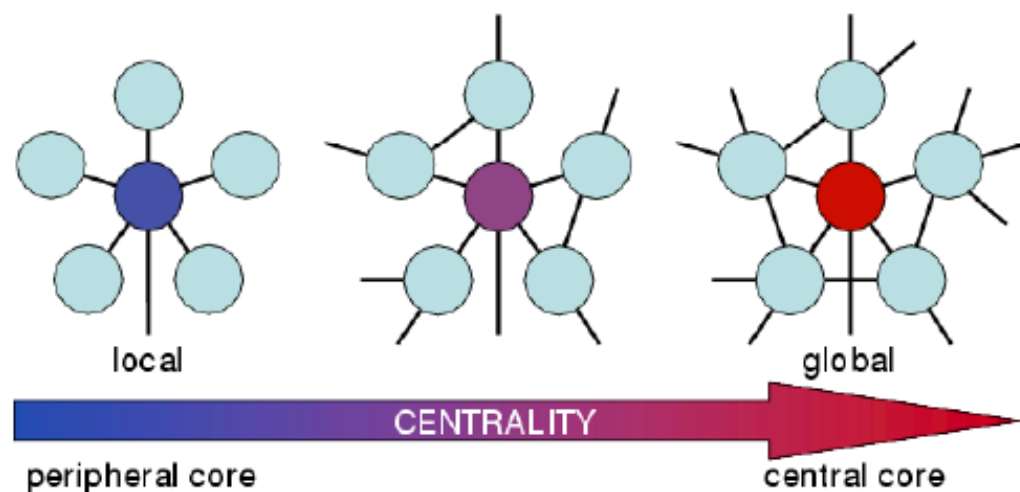
(Application total runtime was: 1.0 seconds.)							
(Real network processing runtime was: 0.0 seconds.)							
(Single Random network processing runtime was: 0.0 seconds.)							

网络模序搜索示意图

□ 早期主要是寻找完全链接的子图 (**clique**)：指任意两点都被边连接在一起的子网



□ 后期通过寻找clique或准clique代替模块

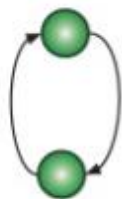


1, 2 and 3-core.

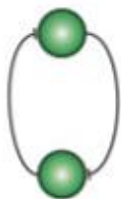
A subgraph is the k -core if all of its nodes have a degree $\geq k$

- Cores are nested
- Cores may be disconnected subgraphs
- Layer k is the intersection of cores k and $(k + 1)$

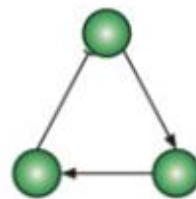
有向网络中模体



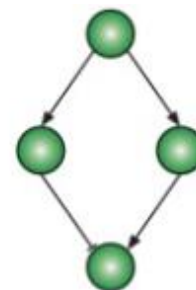
Two-node positive-feedback loop



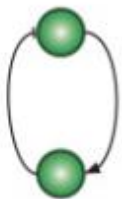
Two-node double-negative-feedback loop



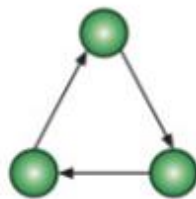
Three-node negative-feedback loop



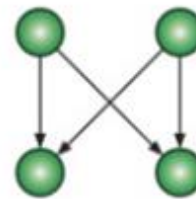
Four-node incoherent feedforward motif



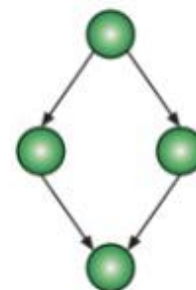
Two-node negative-feedback loop



Three-node positive-feedback loop

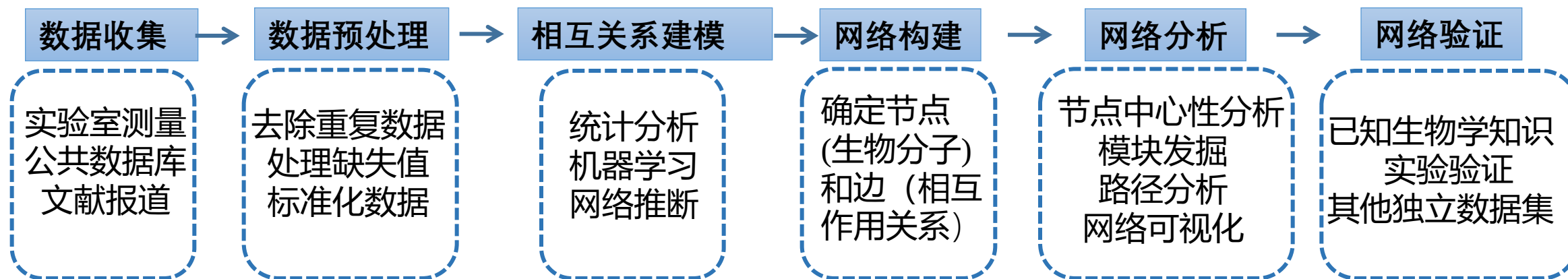


Four-node bi-fan motif

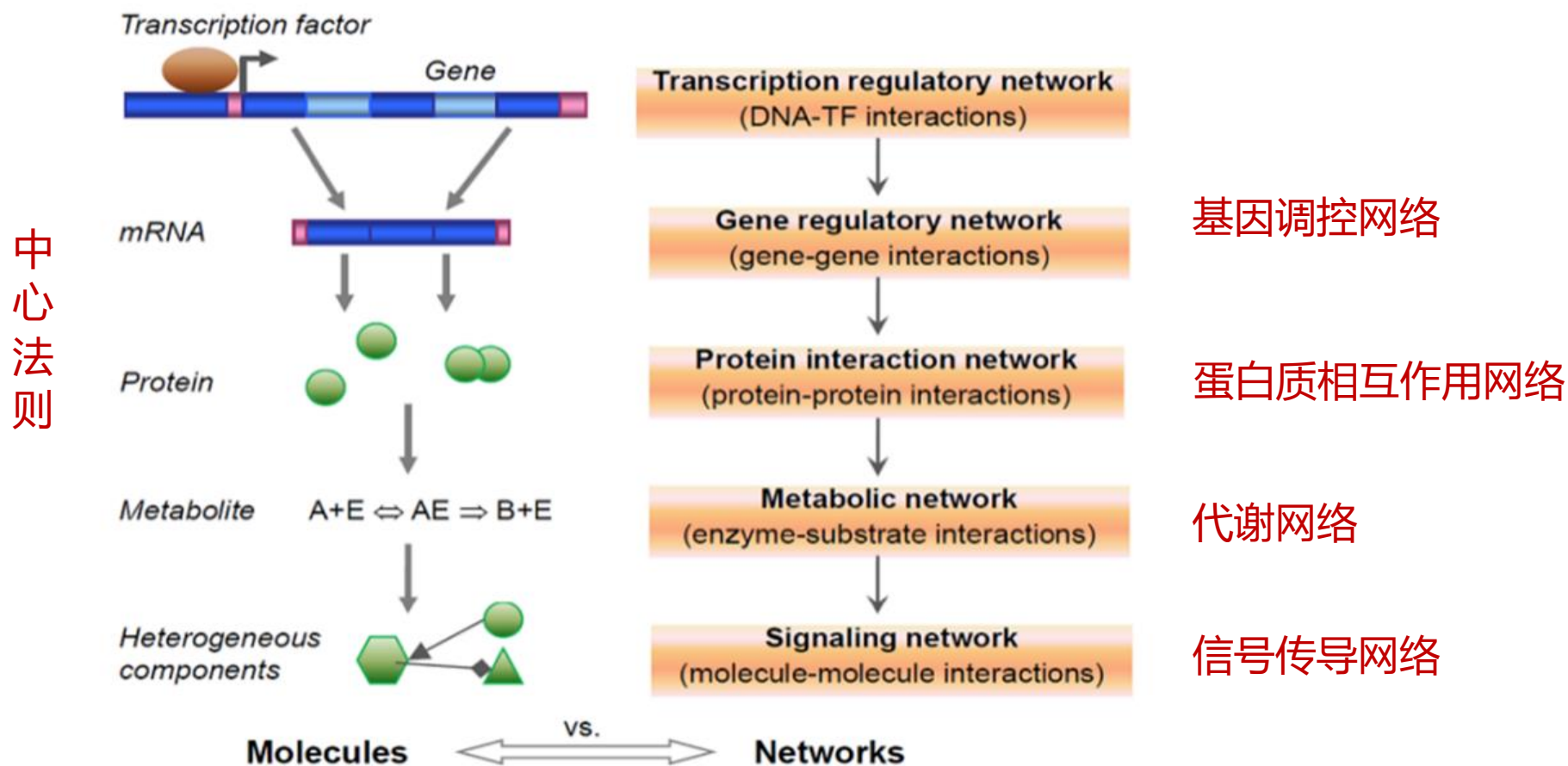


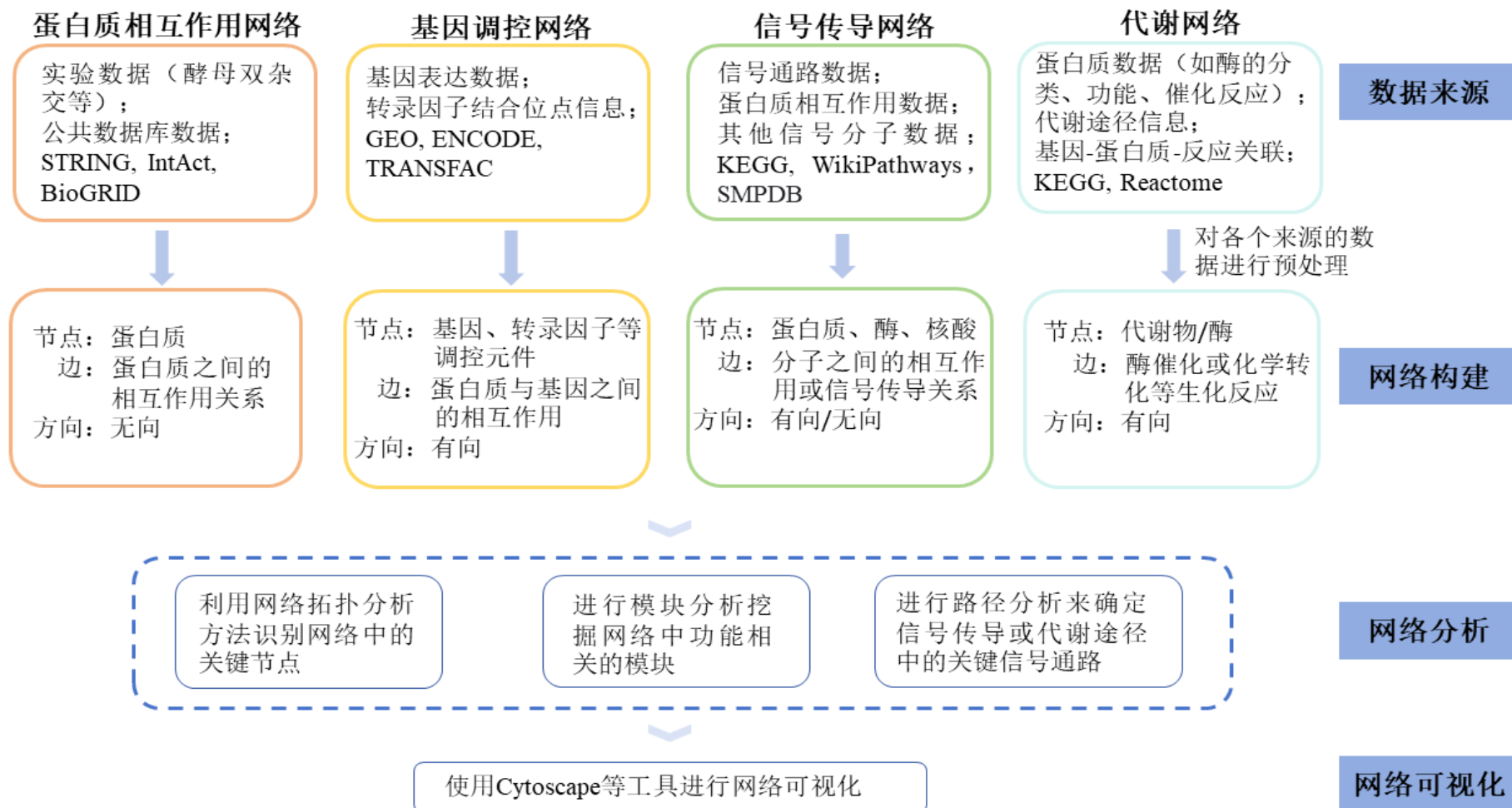
Four-node coherent feedforward motif

- CytoScape软件
 - ✓ <http://www.cytoscape.org>
- CFinder软件
 - ✓ <http://cfinder.org>
- mfinder软件和MAVisto软件
 - ✓ mfinder: <http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html>
 - ✓ MAVisto: <http://mavisto.ipk-gatersleben.de/>
- BGL软件及 Matlab bgl软件
 - ✓ BGL: <http://www.boost.org/doc/libs/release/libs/graph>
 - ✓ MatlabBGL: http://www.stanford.edu/~dgleich/programs/matlab_bgl/
- PathwayStudio软件
- GeneGO软件及数据库



“组件 → 网络 → 体外模型 → 表型” 的系统生物学新范式





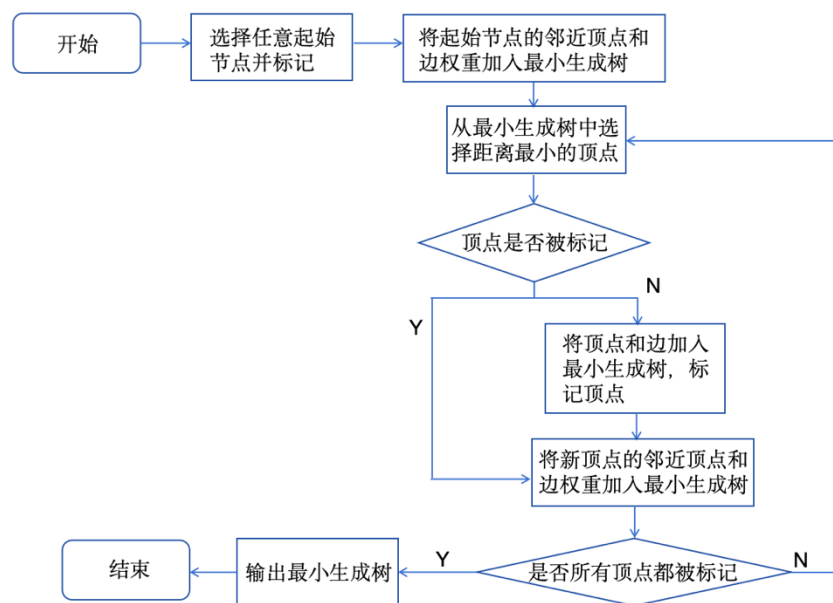
数据类别	来源	描述	挑战
实验数据	酵母双杂交； 免疫共沉淀； 质谱检测	直接从实验中产生的数据，如蛋白质相互作用实验	实验产生的误差； 标识符不一致；数据冗余（同一蛋白质在多个数据库重复报道）；数据选择（来源广泛，质量和类型有差异）；映射外部信息时不准确不完整
有实验证据支持的公共数据库	IntAct, MINT, STRING, UniHI, APID	从实验中收集整理获得的蛋白质相互作用数据	
元数据库和预测数据库	IMEx	利用其他数据库通过算法预测蛋白质的相互作用	

数据库名称	网址
HPRD	http://www.hprd.org/
BioGRID	http://www.thebiogrid.org/
DIP	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
STRING	https://string-db.org/

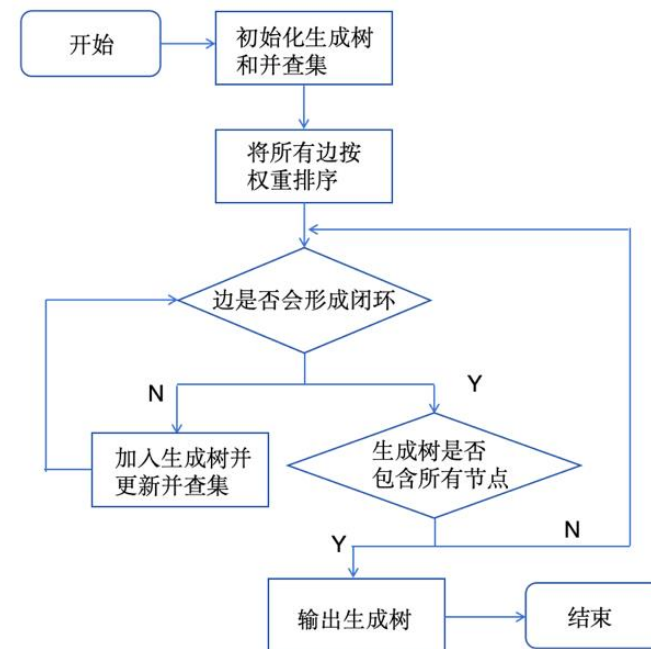
□最小生成树算法

- ✓假设有一个由10个蛋白质构成的网络，每个蛋白质与其他蛋白质的互作有不同的可信度得分。使用最小生成树算法，可以找出连接这10个蛋白质的9条最可信的边，从而获得一个核心互作网络
- ✓**Prim算法**是一种贪心算法。它以一个顶点作为起始点开始构建最小生成树，然后逐步扩展最小生成树
- ✓**Kruskal算法**是一种基于边的贪心算法。它按照边的权值递增的顺序选择边，并且保证所选的边不会构成回路，直到生成树中含有 $n-1$ 条边为止

□最小生成树算法



Prim 算法流程



Kruskal 算法流程

□连通性算法

- ✓ 连通性算法专注于识别网络中高度相互关联的节点群体，即连通子图
- ✓ 在蛋白质网络中，这有助于找出功能相关的蛋白群集
 - 在同一个包含10个蛋白质的网络中，连通性算法可能会识别出其中3个高度互相连接的蛋白质，这些蛋白质可能在某一特定生物过程（如凋亡）中起重要作用
- ✓ 连通性算法主要包括**DFS（深度优先搜索）**和**BFS（广度优先搜索）**
 - 这两者都是常用的图遍历算法

□模块化算法

- ✓通过检测网络中的高密度区域来识别模块或社群
- ✓最初主要用于社交网络分析，后来被应用到蛋白质互作网络和其他生物网络
- ✓模块通常对应于特定的生物过程或信号传导途径

□构建转录调控网络要考虑的关键因素

✓ 如何表征调控关系？

- 转录调控是布尔的还是连续的？
- 转录应该被视为随机过程还是确定性过程？

✓ 什么构成了显著的调控？

- 许多细胞外的信号可以影响基因的表达水平
- 哪些信号是实际生理上显著的？

✓ 文献中实验数据的问题

- 在不同条件下进行的实验
- 大部分实验聚焦于研究清楚的局部调控

□ 基于实验数据的方法

✓ 染色质免疫共沉淀 (ChIP-seq)

- ChIP-seq (Chromatin Immunoprecipitation followed by sequencing) 是一种强大的实验技术，用于鉴定特定转录因子或其他DNA结合蛋白在基因组上的结合位点
- 通过分析ChIP-seq数据，可以直接识别转录因子与目标基因的直接调控关系

✓ 表达敲除/过表达实验

- 利用基因敲除或过表达技术，通过观测特定基因的表达变化对其他基因表达的影响，推断出基因间的调控关系
- 这种方法可以通过RNA干扰 (RNAi)、CRISPR/Cas9等技术实现

□ 基于数据整合的方法

✓ 共表达网络 (Co-expression Networks)

- 利用基因表达数据 (如mRNA测序、微阵列数据)，通过分析基因共表达模式来推断基因间的调控关系
- 常用的共表达网络构建算法包括加权基因共表达网络分析 (Weighted Gene Co-expression Network Analysis, WGCNA) 和相关分析 (如皮尔逊相关系数)

✓ 数据整合方法

- 通过整合多种类型的数据，如基因表达、ChIP-seq、表观遗传数据和DNA序列特征，构建更可靠和全面的基因调控网络。贝叶斯网络和机器学习算法常用于这种数据整合任务

□基于统计和计算方法的方法

✓ 贝叶斯网络 (Bayesian Networks)

- 贝叶斯网络是一种基于概率图模型的方法，常用于推断基因调控网络。通过贝叶斯网络，可以捕捉因果关系和基因间的调控关系，特别是处理高维数据时具有优势

✓ 动态贝叶斯网络 (Dynamic Bayesian Networks, DBNs)

- 动态贝叶斯网络是贝叶斯网络在时间序列数据上的延伸，适用于处理时间动态变化的基因表达数据，能够捕捉基因调控网络中的时序信息

✓ 互信息 (Mutual Information)

- 互信息度量两个随机变量之间的依赖关系，而非简单的线性相关。ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) 和 CLR (Context Likelihood of Relatedness) 方法使用互信息来推断基因调控网络

□基于机器学习和深度学习的方法

✓基于特征工程的机器学习方法

- 利用多个基因表达数据集，提取特征并输入到机器学习算法（如支持向量机、随机森林、逻辑回归）中，预测基因间的调控关系

✓深度学习方法

- 深度学习，尤其是图神经网络（Graph Neural Networks, GNNs），在基因调控网络构建中越来越受到关注。这些方法能够从复杂的高维基因表达数据中学习并捕捉到复杂的非线性关系，推断基因调控网络

□基于物理和化学模型的方法

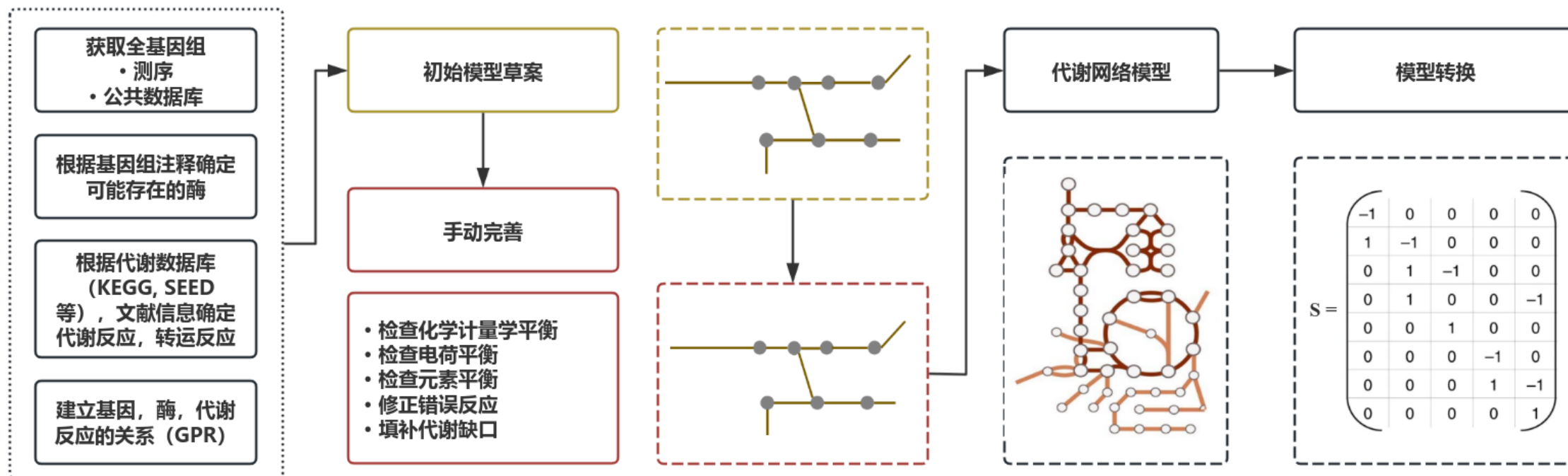
✓ODE模型（Ordinary Differential Equations）

- 常微分方程（ODEs）模型利用基因表达的数据，建立基因产物浓度随时间变化的数学模型。这种方法可以捕捉基因调控机制和动力学特征

□ 代谢网络构建的常用方法

- ✓ 基于代谢途径数据库 (KEGG、Reactome、MetaCyc)
- ✓ 基于代谢物共现关系
- ✓ 基于相关性分析
- ✓ 机器学习方法
- ✓ **全基因组规模代谢网络模型**

□全基因组规模代谢网络模型构建的流程



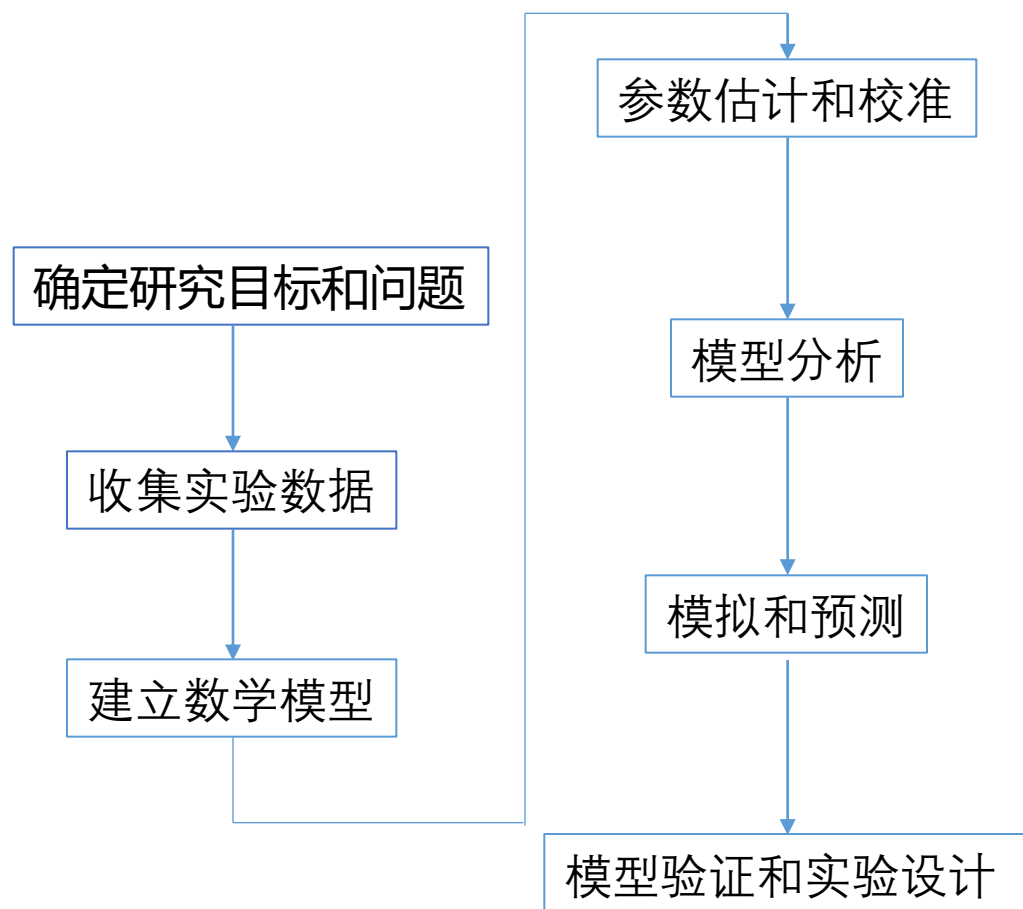
□ 专门的模型描述语言或工具有助于生物网络的数据整合与标准化

- ✓ 系统生物学标记语言 (SBML)
 - ✓ 系统生物学图形注释 (SBGN)
 - ✓ 生物通路交换 (BioPAX)、CellML、FieldML等
 - ✓ 开源库SBSCl, 功能丰富, 可供研究者方便快捷地进行系统生物学建模
- 大多是基于可扩展标记语言 (XML) 的, 并使用了MathML技术

- SBML是一种用于描述生物化学和细胞生物学模型的标准文件格式，是一种机器可读的表示方法，用于在计算机系统中存储和交换生物系统模型的信息
- SBML的主要目标是提供一种通用的、可扩展的格式，用于表示生物化学反应、代谢途径、细胞信号传导通路等生物系统模型。
- 它不仅用于存储模型的数学公式，还包括模型的结构、参数、初始条件、物质单位等信息

软件名	简介	下载地址
CellDesigner	是一个用于绘制生物通路图的工具，使用SBML格式来表示生物化学模型，可用于构建和可视化代谢途径、信号传导网络等。	https://www.celldesigner.org/download.html
SBMLToolbox	是一个MATLAB工具箱，用于创建、修改和分析SBML格式的生物化学模型，适用于需要在MATLAB环境中进行系统生物学建模和分析的用户。	https://sourceforge.net/projects/sbml/files/SBMLToolbox/
PySB	是一个用于构建和模拟生物系统动力学模型的Python库，专注于定量模型的构建，可用于研究信号传导、代谢途径等。	https://pysb.org/download
BioNetGen	是一个用于建模和模拟生物网络动力学的工具。允许用户以规则为基础的方式描述分子相互作用，并生成相应的动力学模型。	https://bionetgen.readthedocs.io/en/latest/install.html
BioModels	是一个用来存储和交流生物模型的数据库，所有的模型都是在科学文章中发表过的，用户可以查询和模拟已有的生物模型，也可以递交自己的模型。	https://www.ebi.ac.uk/biomodels/
COBRA	是一种用于建立、模拟和分析代谢网络的方法和软件，有助于理解生物体内代谢过程的调控和功能。通常使用线性规划、整数规划等数学工具来解决约束条件下的模拟和优化问题。COBRA软件主要包括COBRAPy（Python中的实现）、COBRA Toolbox（MATLAB中的实现）等。	https://opencobra.github.io/cobratoolbox/stable/installation.html
yEd	是一个流程图和图形绘制工具，也可以用于绘制生物网络，可视化生物网络的结构和关系	https://www.yworks.com/products/yed/download

- 生物系统是极其复杂的非线性、非平衡系统，需要在高通量实验技术获得的海量生物信息基础上建立物理和数学模型，最终通过建模与实验相结合的手段来定量阐述生命现象的本质规律
- 生物系统建模就是将生物系统的各个组成部分（如基因、蛋白质、代谢物等）及其相互作用转化为数学方程或模型，从而模拟和预测系统的行为
- 生物体内的生物化学反应具有显著的随机性，宜采用随机性模型进行定量描述



根据实验数据，对模型中的参数进行估计和校准，使用最小二乘法或贝叶斯推断等参数估计方法，通过拟合模型与实验观测数据的吻合程度比较来实现模型校准。

利用数学分析方法，对建立的模型进行分析，包括稳定性分析、灵敏度分析、稳态分析、周期分析等，这些分析可以帮助我们理解系统的稳定性、响应特性、关键因素等。

使用数值计算方法，对建立的模型进行模拟和预测。通过数值求解模型的方程，可以模拟生物网络在不同条件下的行为和响应，进一步验证模型的有效性。

将模拟结果与实验观测进行比较，验证模型的准确性和可靠性。根据模型的预测，设计实验来验证模型的预测结果，并进一步完善和改进模型。

□布尔网络模型

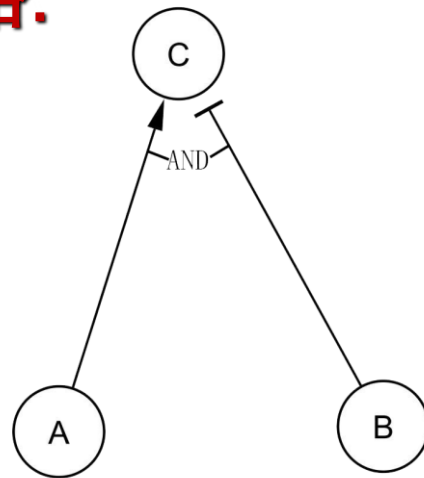
- ✓ 每个节点代表一个基因，或者代表一个环境刺激
- ✓ 每条有向边代表基因之间的相互作用关系
- ✓ 各节点的值或者是“1”，或者是“0”，分别表示“高水平”和“低水平”

布尔表达式：

$$A \cap (\neg B) \rightarrow C$$



布尔网络：



□ 布尔网络模型

□ 当选定初值为 $A=0$, $B=0$, $C=0$ 时

由接线图可知下一步各点状态:

$A'=B \text{ AND } C=0$, $B'=\text{NOT}(C)=1$, $C'=A=0$

迭代, $A''=B' \text{ AND } C'=0$, $B''=\text{NOT}(C')=1$, $C''=A=0$

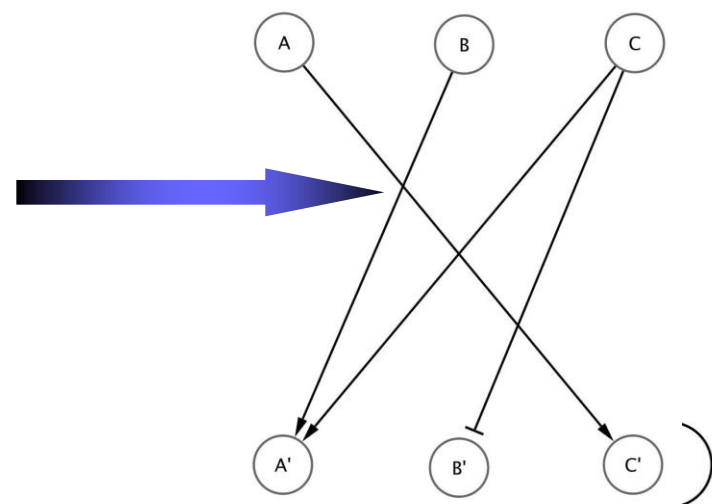
结果稳定, 在迭代中不会改变

□ 当选定初值为 $A=1$, $B=0$, $C=0$ 时, 会有怎样的变化?

布尔网络



接线图



□线性组合模型

- ✓ 是一种连续网络模型，在这种模型中，一个基因的表达值是若干个其他基因表达值的加权和。表示形式为：

$$X_i(t+1) = \sum_j w_{ij} X_j(t)$$

- ✓ 其中， $X_i(t+1)$ 是基因*i*在*t*+1时刻的表达水平， $X_j(t)$ 是基因*j*在*t*时刻的表达水平，而 w_{ij} 代表基因*j*的表达水平对基因*i*的影响
- ✓ 给定一系列基因表达水平，即给定每个基因的时间序列 $\{X_i(t)\}$ ，就可以利用最小二乘法或者多重分析法求解整个系统的差分方程组，从而确定方程中的所有参数，即确定 w_{ij}

□加权矩阵模型

- ✓ 加权矩阵模型与线性组合模型相似，在该模型中，一个基因的表达值是其他基因表达值的函数
- ✓ 基因 i 的总调控输入 $r_i(t)$ 为：

$$r_i = \sum_j w_{ij} u_j(t)$$

- ✓ 其中 $u_j(t)$ 是 t 时刻基因 j 的状态，若 w_{ij} 为正值，则基因 j 激活基因 i 的表达，而负值表示基因 j 抑制基因 i 的表达，0 表示基因 j 对基因 i 没有作用
- ✓ 则基因 i 在第 $t+1$ 时刻的状态可以通过 Sigmoid 函数计算得到

$$u_i(t+1) = \frac{1}{1 + e^{-[\alpha_i r_i(t) + \beta_i]}}$$

□代谢通量分析 (Metabolic Flux Analysis, MFA)

- MFA是一种基于实验数据的通量分析方法，用于预测代谢网络中各个反应的通量分布
- 该方法通过测定代谢产物中的标记分布来推测通量，在不需知道代谢途径中各种酶动力学特征的情况下，可以得到关于代谢的特征信息
- MFA根据质量守恒定律建立一系列的代谢平衡方程，反映了在代谢网络中每个节点物质的输入和输出是相等的，目标是确定代谢网络中每个代谢反应的通量
- 由于代谢反应速率可能与浓度呈现非线性关系，所以一般是通过求解非线性方程来得到结果

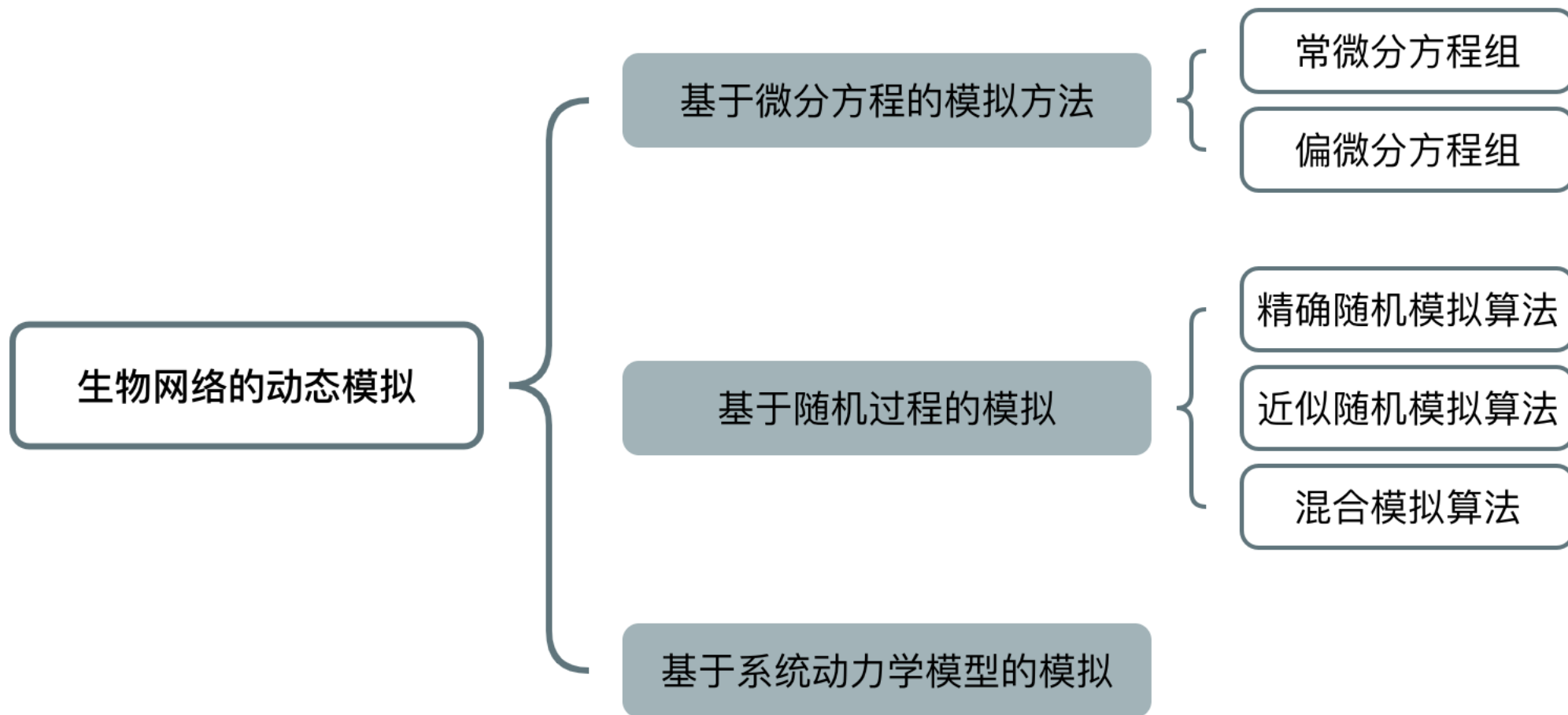
□基于约束的优化方法（Constraint-based Optimization）

□**通量平衡分析（Flux balance analysis, FBA）**是最重要的基于约束的分析方法，该方法将代谢网络表示为一组线性方程和不等式约束

- ✓ 其中约束包括代谢反应速率、代谢物之间的转化关系、代谢物的摄取和产出速率等
- ✓ 通过最大化或最小化特定目标函数（如生物体生长速率或关键代谢物产量），采用线性规划可以计算出在满足约束条件下的最优代谢通量分布

□**通量变异性分析（Flux Variability Analysis, FVA）**，用于评估在特定条件下代谢通量的最大和最小的可能值，可以反映代谢网络中通量的可变性

□**几何通量平衡分析（Geometric FBA）**是FBA的一种改进，不但找到最优解，还考虑了解空间的几何结构，代谢通量作为需要被优化的变量，可找到网络中的冗余路径和代谢通量的多样性



□基于微分方程的模拟方法-常微分方程

□常微分方程组是一组描述多个变量随时间变化的方程的集合，这些方程描述了系统中各个变量之间的相互关系和随时间的演化，被广泛用于描述生物系统中的动态过程

数值方法

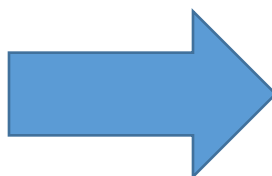
Euler 法

Runge-Kutta法

解析方法

分离变量法

变换法



在**基因调控网络**中，可以描述基因的转录和翻译过程，通过模拟基因表达变化，可以预测基因调控、信号传递等过程的响应

在**代谢网络**中，可以描述化学反应、底物转化以及代谢产物生成，有助于预测代谢产物的积累、底物的消耗，找到关键限速步骤

在**信号传导网络**中，可以描述信号分子在细胞内的传递过程，理解细胞对不同刺激的响应，以及信号通路中的调控机制

□基于微分方程的模拟方法–偏微分方程

□偏微分方程在生物网络研究中可以用于描述动态系统中随空间和时间变化的行为，与常微分方程相比区别在于是否依赖于一个变量，依赖于多个变量的微分方程称为偏微分方程（PDE）

- ✓ 在反应-扩散系统中，偏微分方程可以描述生物网络中反应和扩散的耦合行为，如基因表达在细胞中的空间分布和调控等
- ✓ 在生物膜传输体系中，偏微分方程可以用于模拟生物膜上物质的传输过程，如细胞膜上的离子通道等

□基于随机过程的模拟

□随机过程模拟方法考虑了生物网络中的随机性和噪声效应，对于研究生物系统中的噪声传播、稳定性和可变性具有重要意义

✓ 例如，蒙特卡洛方法和随机微分方程模型可以用于描述生物网络中的随机变化和随机相互作用

□随机模拟算法可以分为：

✓ 精确随机模拟算法

✓ 近似随机模拟算法

✓ 混合模拟算法

□精确随机模拟算法

- ✓ 以一个生化反应系统为例，该系统由N种化学物质 $S_1, \dots, S_N$ 在混合良好的环境中通过M种反应 $R_1, \dots, R_M$ 相互作用
- ✓ 该假设又称**空间均质性假设**，即反应空间中的所有分子都均匀分布，在空间上不可区分。这里所讨论的所有随机算法只有在这个假设下才准确
- ✓ 这类方法通过从联合反应概率密度函数中**随机采样**下一个要触发的反应 R_μ 以及其触发时间 τ ，来明确地模拟每个反应事件
- ✓ 采样步骤不同的精确随机方法有：直接法（DM）、第一反应法（FRM）、下一反应法（NRM）和基于拒绝的随机模拟算法（RSSA）

□近似随机模拟算法

- ✓ 精确随机模拟方法通过逐一模拟每个系统事件，提供最准确的模型动力学实现
 - 但是，当应用于大规模生物系统时，这种方法可能会变得非常缓慢，因为每个时间点可能触发数千个反应
- ✓ 近似算法通过减少运行时间来提高模拟效率，但代价是降低精度
- ✓ 除了考虑空间均匀性假设外，近似方法还引入了一种新的近似水平，即**假设当一组反应事件的确切触发时间阈值足够接近时，它们可以在同一瞬间被触发**
- ✓ 对于近似方法，较低的迭代次数显著降低了生成随机数对模拟策略整体性能的影响
- ✓ 近似算法包括 τ 跳跃法、化学Langevin法等

□混合模拟算法

- ✓ 混合模拟算法结合并充分利用了精确和近似模拟算法的特点，其思想是根据每个模型反应的某些特征对系统进行分区，然后应用不同的特定模拟方法，以更好地适应不同的分区特征
- ✓ HRSSA是一种经典的混合模拟算法
- ✓ HRSSA将系统反应分为慢反应和快反应两个子系统

□混合模拟算法

系统动力学模型是一种用于描述和分析动态系统的模型方法。系统动力学模型的一般形式通常由一组关联的微分方程或差分方程组成。对于连续时间系统，模型可以表示为：

$$\frac{dx}{dt} = f(x, u, t)$$

对于离散时间系统，模型可以表示为

$$x(t+1) = f(x(t), u(t), t)$$

系统动力学模型通常以状态空间形式表示，其中状态方程和输出方程分别为：

$$\frac{dx}{dt} = Ax + Bu$$

$$y = Cx + Du$$

- 系统生物学不同于以往仅仅关心个别基因和蛋白质的分子生物学，它着眼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
- 系统生物学已成为生物信息学最新前沿领域之一，并已从全基因组水平的生物网络模型发展到全细胞模型（whole cell model），对细胞中所有的基因、产物以及已知的代谢相互作用进行重建
- 未来可望能建立一个完整的多细胞生物的系统生物学模型
- 系统生物学的研究在不同领域已得到了应用，如药物研发（网络药理学）、疾病诊断和治疗、合成生物学等

1. **版权声明：**本PPT及其所有内容（以下简称“本PPT”）仅用于教育和教学用途，版权归属于本PPT作者。
2. **使用要求：**任何使用本PPT的行为均须遵守以下条件：
 - 1) **致谢和标注：**若部分或全部使用本PPT的内容，请在使用内容的适当位置标注出处，并致谢本PPT作者。
 - 2) **修改和再分发：**未经作者书面许可，不得对本PPT进行修改或再分发。
3. **禁止商业化使用：**严禁将本PPT用于任何形式的商业化用途，包括但不限于：
 - 1) 通过网络或其他途径进行付费使用或分发；
 - 2) 在商业培训、广告或其他商业活动中使用本PPT的内容。
4. **法律责任：**任何违反上述条款的行为，作者保留追究法律责任的权利，包括但不限于：
 - 1) 要求停止侵权行为；
 - 2) 追究侵权使用者的经济赔偿责任。
5. **其他规定：**
 - 1) 本使用条款的解释权归本PPT作者所有。
 - 2) 作者保留随时更新本使用条款的权利，更新后的条款将即时生效。

谢谢大家!

廖明帜
西北农林科技大学