最优化与系统生物学
Optimization and Systems Biology

章祥荪

http://zhangroup.aporc.org

中国科学院 数学与系统科学研究院

中国运筹学会第八届全国代表大会， 南京， **10/18/2008**

最优化与系统生物学国际会议
Optimization and Systems Biology

Chinese Academy of Sciences. China
Tokyo University, Japan
Shanghai University, China

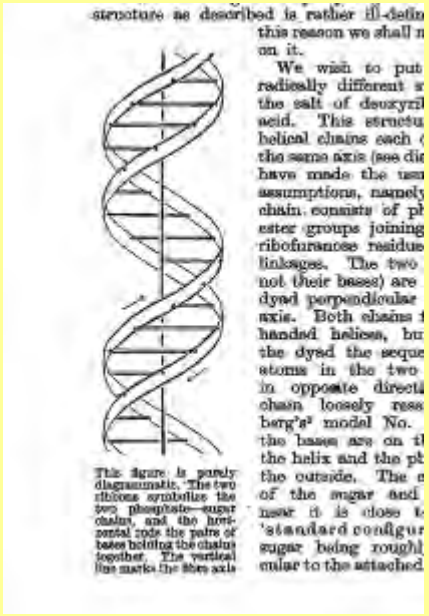
OSB2008, October 31-November 3, 丽江

报告提纲

- 研究背景
  1. 分子生物学的基本概念
  2. 什么是系统生物学？
- 最优化方法在系统生物学中的应用
  1. 基因调控网络研究
  2. 蛋白质相互作用网络研究
- 生物分子网络及一般复杂网络中的优化问题研究

二十一世纪的生命科学

《Nature》1953

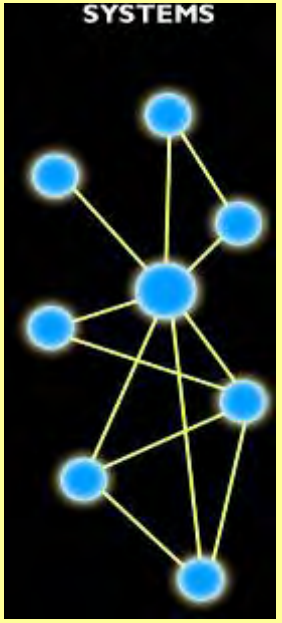


DNA双螺旋结构发现
开启分子生物学研究

《Science》2001

人类基因组计划完成
二十世纪三大科学工程之一

生物信息学



系统生物学

1950

1990

Time

数学与生物学的交叉研究

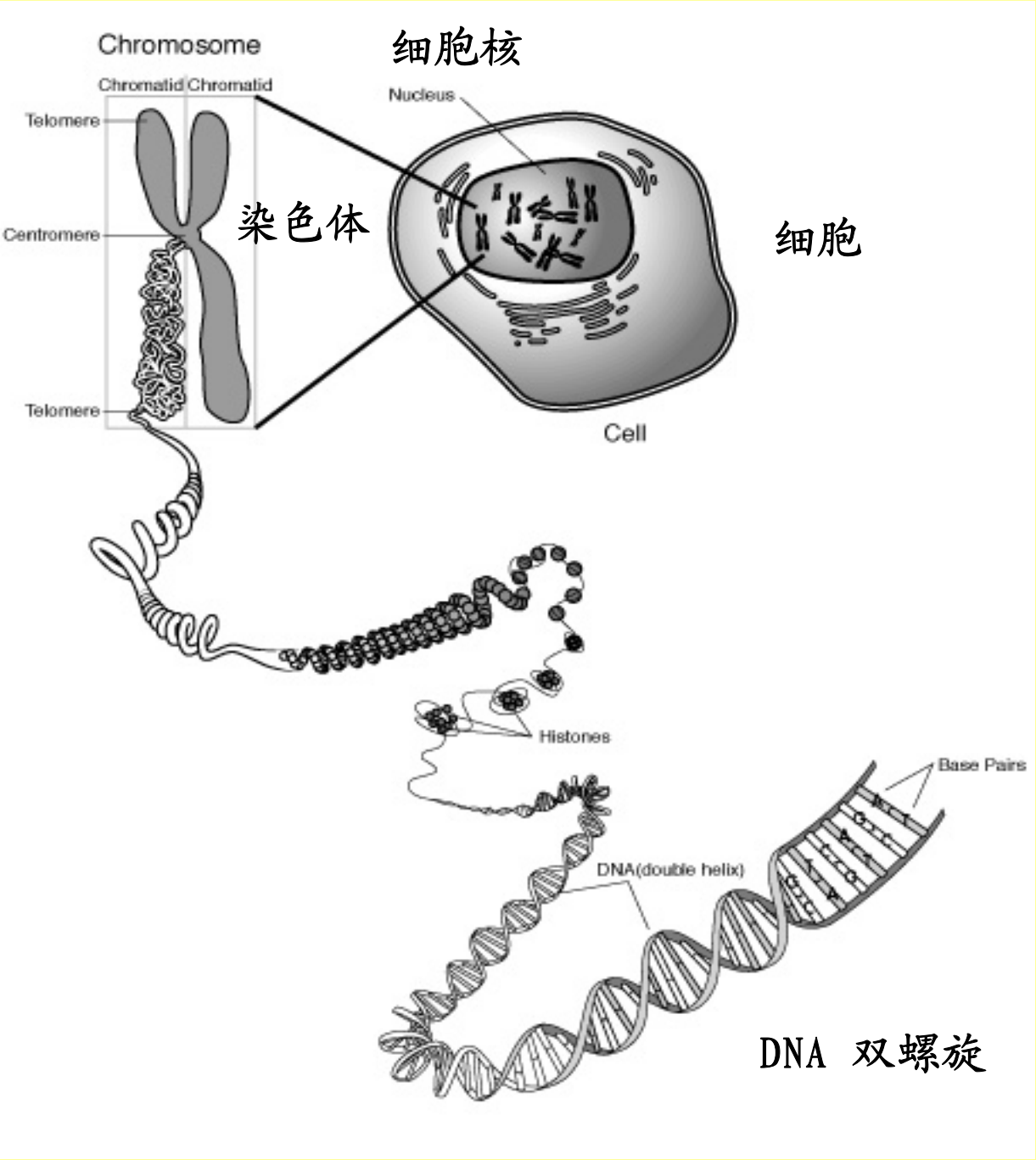
国际热点:

- 2008年9月3日，美国国家科学基金会（NSF）宣布将投资1600万美元用于建立国立数学生物学综合研究所（NIMBioS）。
- NSF为此专门启动了一项“定量的环境与整合生物学”项目。
- 美国国立卫生研究院(NIH)也设立了一项“计算生物学”的重大项目。

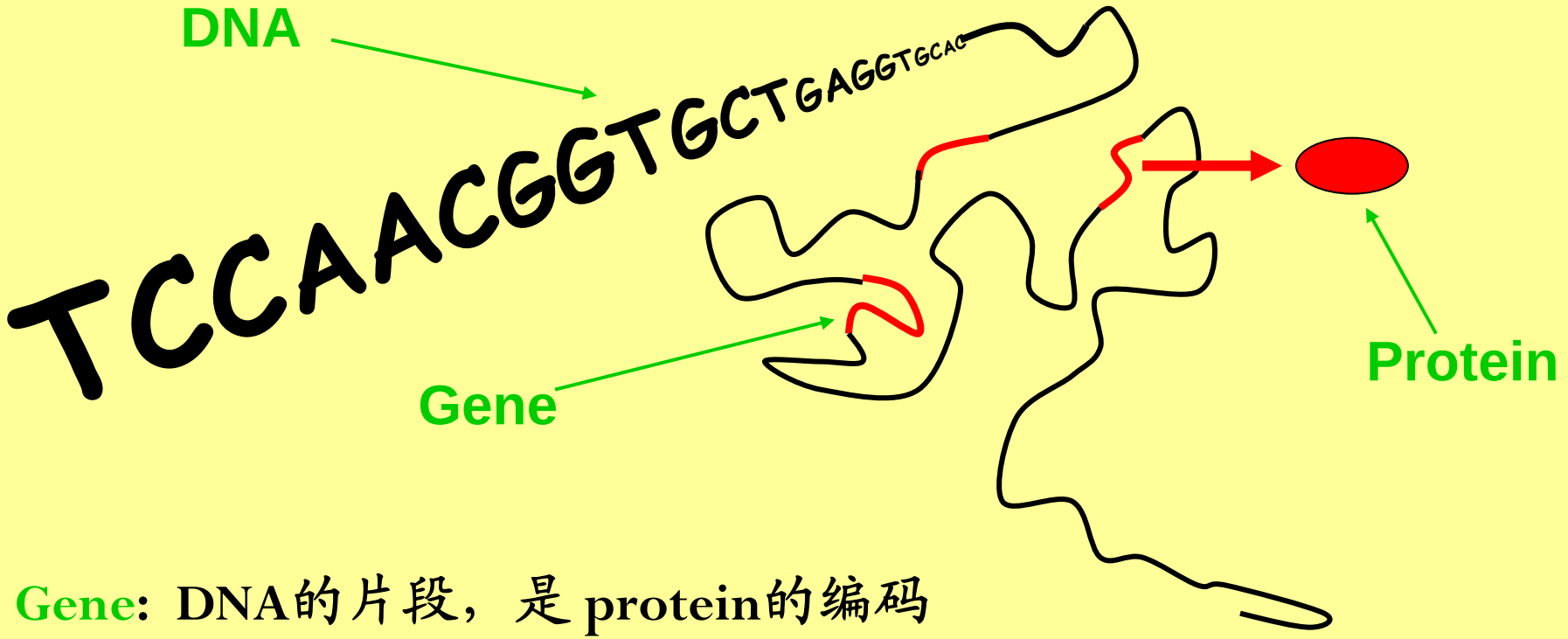
国内热点:

- 《国家中长期科学和技术发展规划纲要》指出：我国将在今后15年重点研究重要生物体系的转录组学、蛋白质组学、代谢组学、结构生物学、蛋白质生物学功能及其相互作用、蛋白质相关的计算生物学与系统生物学等。

什么是基因(Gene)和蛋白质(Protein)？


Chromosome
细胞核
Chromatid
Chromatid
Nucleus
Telomere
染色体
Centromere
细胞
Telomere
Cell
Histones
Base Pairs
DNA(double helix)
DNA 双螺旋

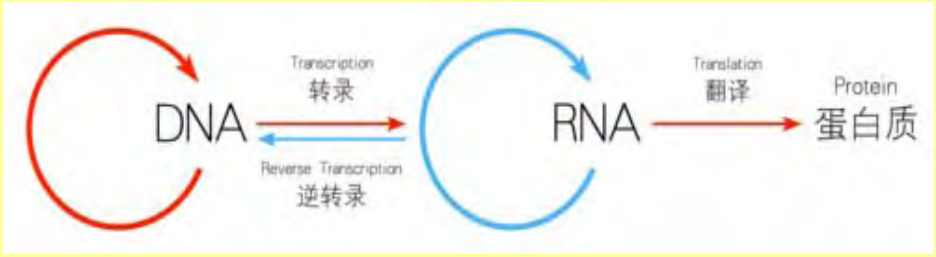
DNA, 基因(gene), 蛋白质(Protein)



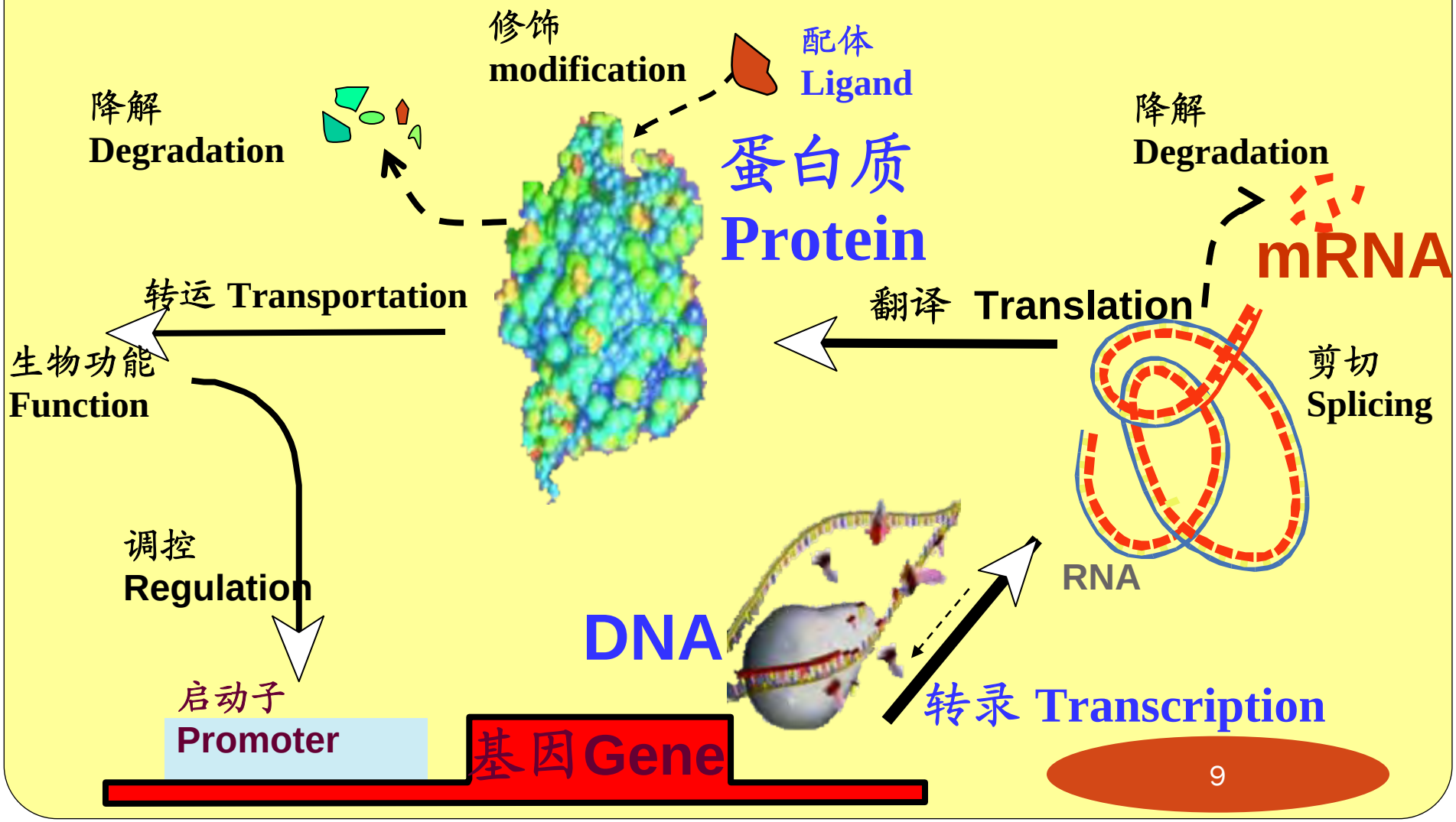
Gene: DNA的片段，是protein的编码

Protein: 实现细胞的生物功能

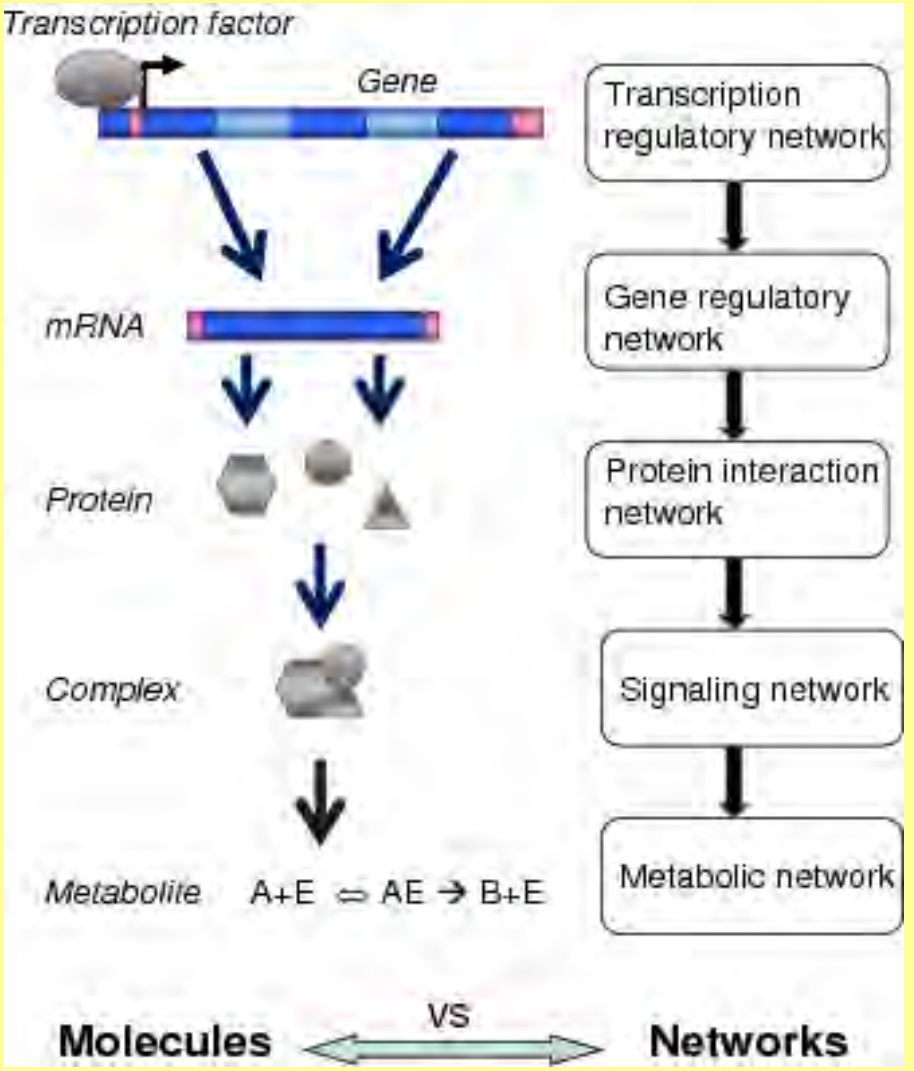
→: 分子生物学中心法则



分子生物学的中心法则(Central dogma)



- 贯穿中心法则信息流的各个层次都涉及到生物网络，如转录调控网，基因调控网，蛋白相互作用网，信号转导网，新陈代谢网等.



什么是系统生物学(Systems Biology)？

一个例子

- 传统的生物学认识：物种或生命的复杂程度应该与基因数目有直接的关系。
- 线虫是一种低等动物，其基因组的基因数为1.9万多个。而人类基因组的基因总数仅仅比线虫多几千个左右。水稻基因组的基因总数在4.6万到5.5万之间，比人的基因还要多。
- 在生命从简单到复杂，从低级到高级的进化过程中，起决定作用的不是基因个体数目，而是生命系统中简单元件的相互作用或网络的复杂性。因此有必要由整体的、合成的角度诠释生命系统。

系统生物学的创始人 **Leroy Hood**，美国西雅图系统生物学研究所（全球第一个系统生物学研究所）所长:

- **Many biological problems, particularly human diseases, fall into the category of "systems problems"**

  -- Leroy Hood

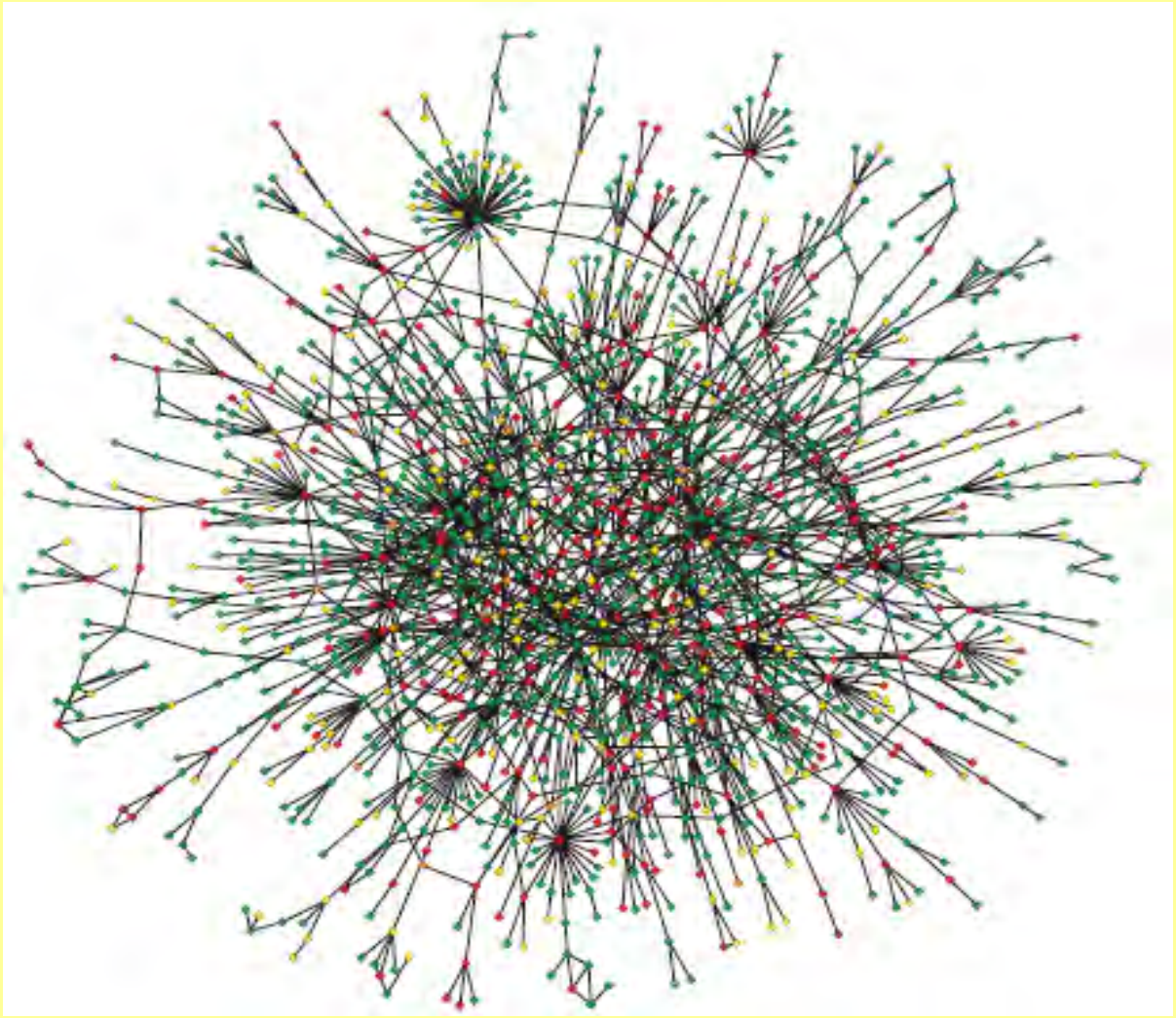
- 许多的生物问题，尤其是涉及到人类疾病方面的问题，都属于系统性的问题，不是研究清楚单个基因或者蛋白质就可以解决的。

- **Systems biology — an approach that studies biological problems through studying interrelationships of all of the elements (gene, protein, RNA, …) in a system rather than studying them one at a time.**

  -- Leroy Hood

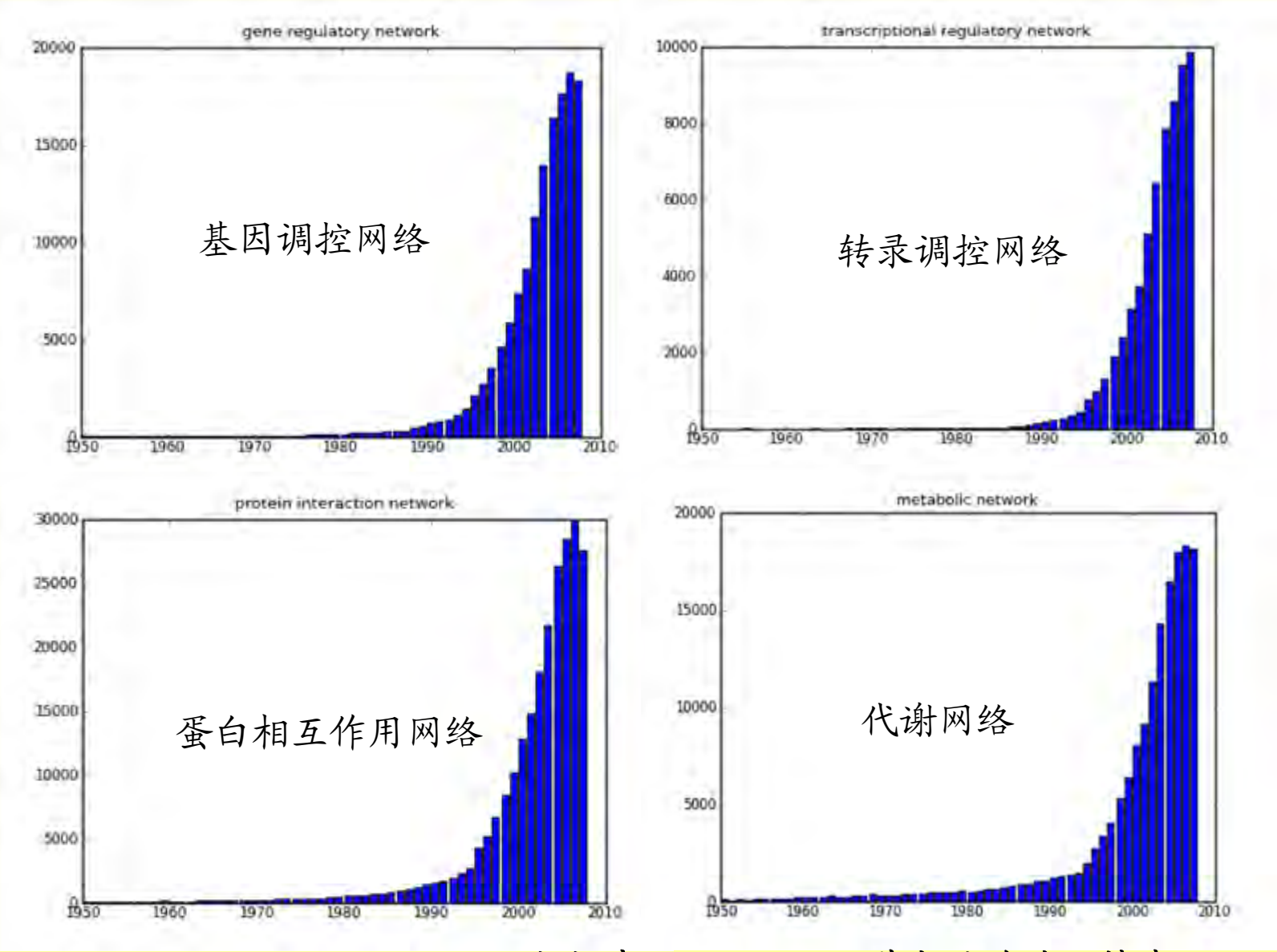
- 系统生物学——是一种研究复杂生物问题的新方法，该方法主要研究生物体中所有的要素(基因、蛋白质、RNA等)之间的相互关系，而不是针对性地研究个别基因或者蛋白质.

复杂网络的典型代表: 生物分子网络



酵母细胞中的蛋白质相互作用网络 **(A.-L. Barabási, NATURE REVIEWS GENETICS, 2004)**

生物分子网络是当前研究热点



注： 数据来源于Google 学术的关键词搜索

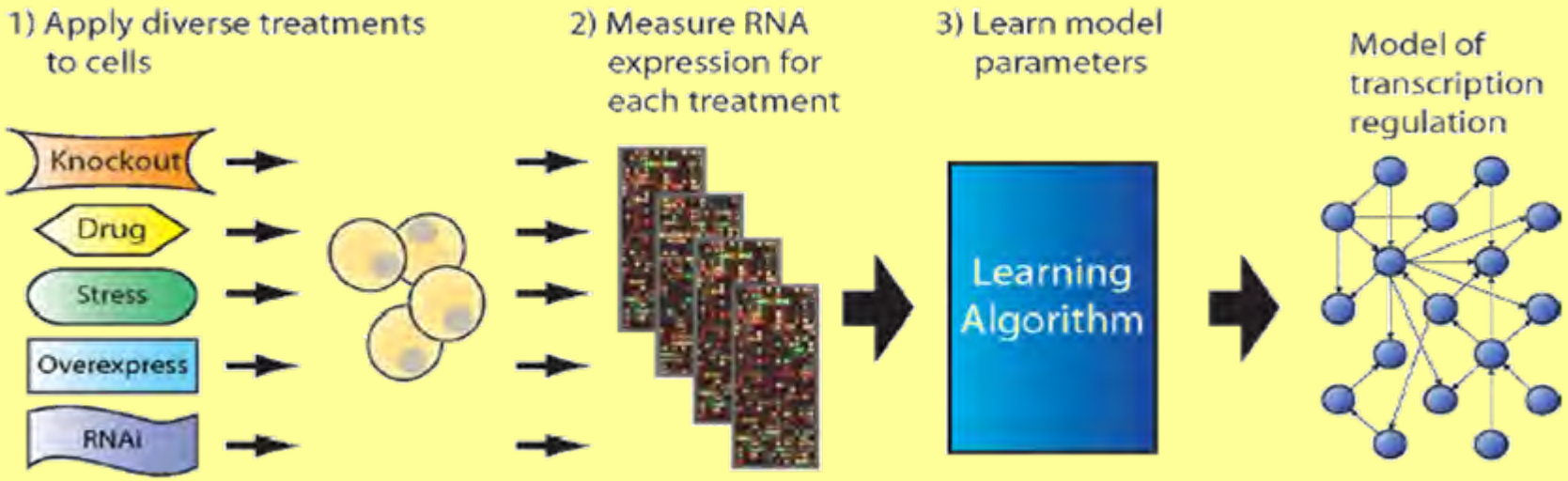
我们的研究 (Part I):
生物分子网络建模的最优化方法

我们研究的生物分子网络类型

1. 基因调控网络 (Gene regulatory network): 用于控制基因开关（基因表达量的高低）的网络.

2. 蛋白质相互作用网络 (Protein-protein interaction network): 由蛋白质之间物理相互作用形成的网络.

例一：集成数据推断基因调控网络



基因表达数据

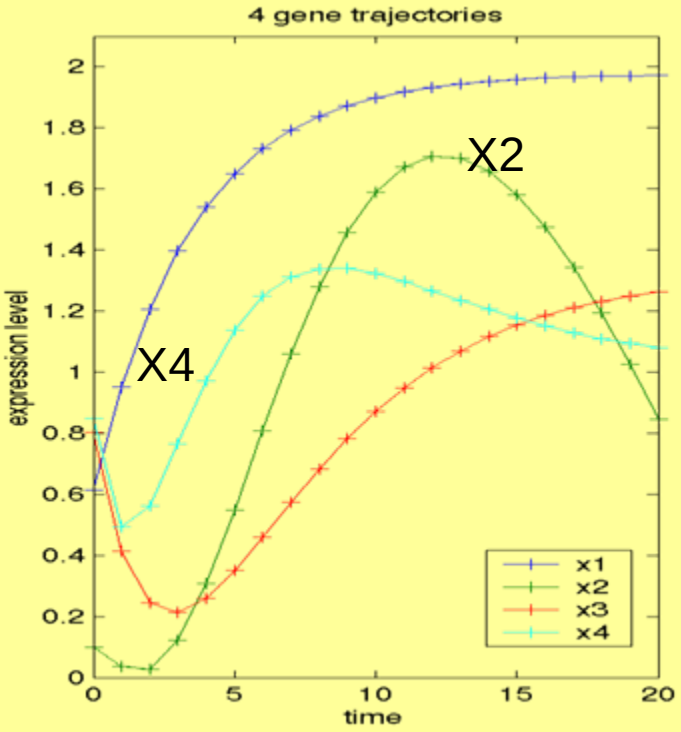
数学模型：微分方程、最优化

预测网络结构

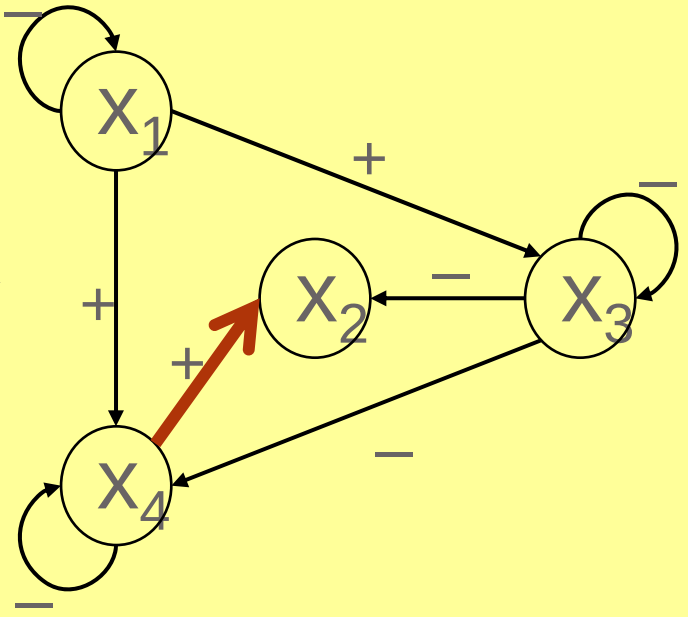
什么是基因调控网络？

- 一个有向赋权网络:
  - 节点：基因
  - 边：表示基因之间的直接或者间接调控关系，基因A对基因B的调控关系指基因A的状态可以影响基因B的状态（基因状态是一个时间序列）
  - 权：（权值大于 0 或者小于 0），调控作用可以是激活，或者是压抑

基因调控网络的推断



?



基因表达时间序列：矩阵**X**（**nxm**维）　基因调控网络：矩阵**J**（**nxn**维）

基因的时间序列表达数据

给定 n 个基因的的 m 个时间点的芯片数据实质上就是如下一个矩阵**X**，

x_{ij} = 第 i个基因在第 j时刻芯片测量得到的表达值

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

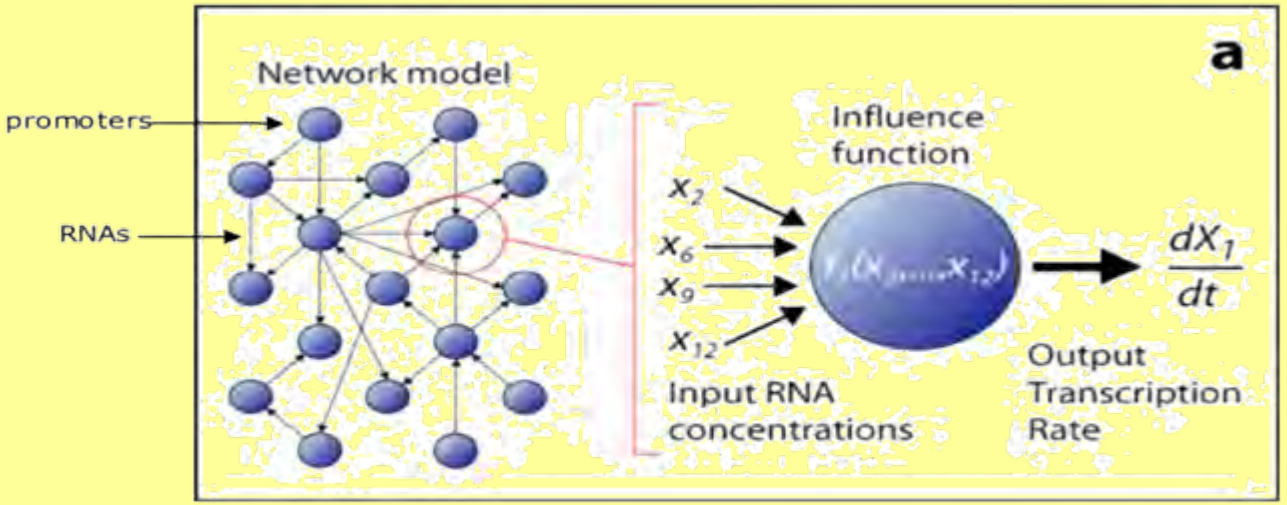
行代表第i个时刻点各个基因的表达值

列代表第 j个基因的所有时刻的表达值

基因调控的微分方程模型

通常利用线性微分方程模型来建模基因调控关系：

$$dx_i(t)/dt = a_0 + a_{i,1}\,x_1(t) + a_{i,2}\,x_2(t) + \ldots + a_{i,n}\,x_n(t)$$



$$dX_1/dt = f_7(X_2,\ldots) = a_2 X_2 + a_6 X_6 + a_9 X_9 + a_{12} X_{12}$$

例如：基因**1**的表达值的变化率是对它有调控作用的基因**2**，**6**，**9**，**13**的表达值x_2，x_6，x_9，x_{12}的线性组合

数学表达

- 反向工程（Reverse-Engineering）

$$\dot{X} = J_{n\times n} X + B + \varepsilon$$
$$X(1), \cdots, X(m) \Rightarrow J_{n\times n}$$
$$X \in \mathrm{R}^n, \qquad m << n$$

- 已知X, B, 求矩阵 J

$$J_{n\times n} X_{n\times m} = \dot{X}_{n\times m} - B_{n\times m}$$

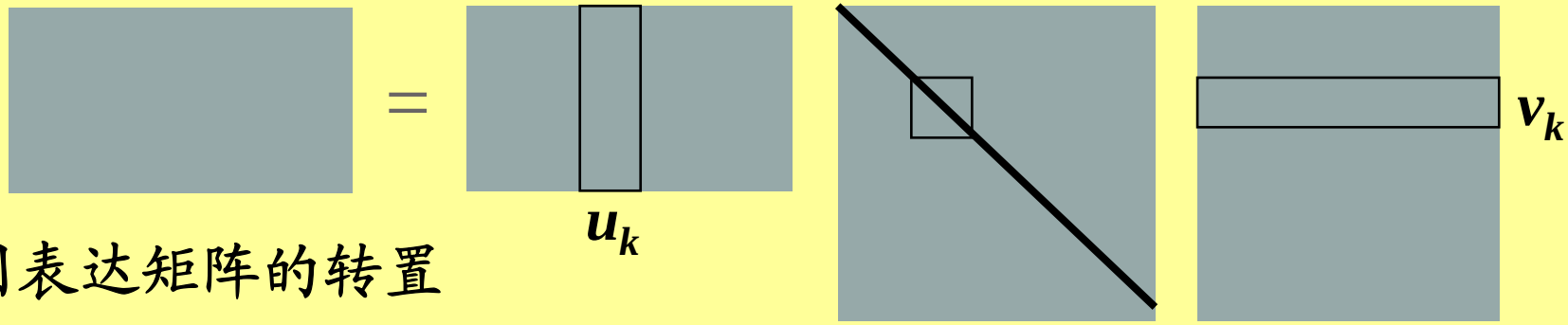
- m>n 时为最小二乘拟合问题，容易求解。但m<<n方程组有无穷多解。

维数问题

- 维数问题（Dimension Problem）：生物实验提供的数据的特点是时间点个数<<变量的个数。
- 例如在酵母中，最多可测量的时间点 m(约为20)<<酵母中基因的个数 n(约为6000) 。
- 上述网络推断问题从数学上是不定的，即有无穷多个网络结构可以拟合出实验观测到的数据。
- 关键问题：**利用优化技术克服维数问题探求基因调控网络的拟最优结构**。

奇异值分解（SVD）

SVD分解: $X^T_{m\times n} = U_{m\times n}E_{n\times n}V^T_{n\times n} \quad (m << n)$



基因表达矩阵的转置

奇异值

特解: $\hat{J} = (\dot{X} - B)UE^{-1}V^T$

通解表达

- SVD 解是最小二乘意义下的特解

$$\hat{J} = \arg\min \|JX + B - \dot{X}\|_2$$

- 通解表达 (General solution)

$$J = \hat{J} + YV^T$$

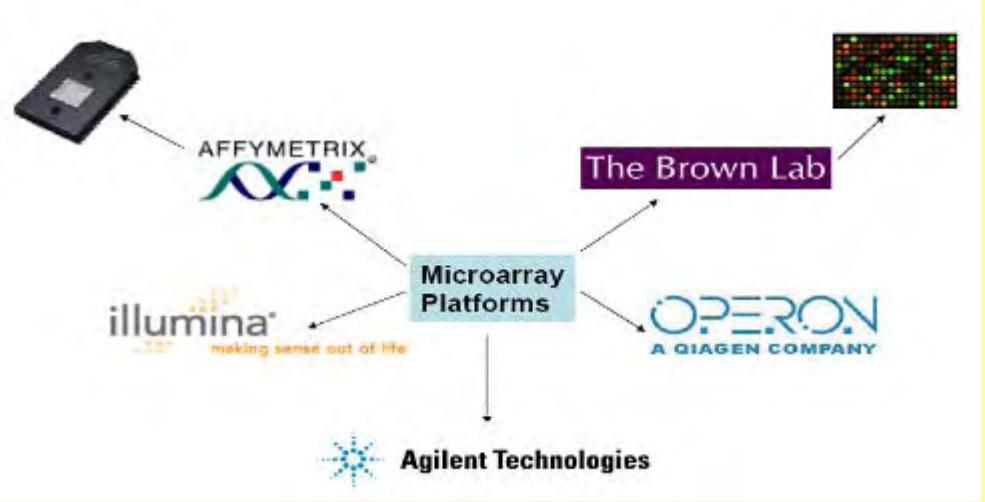
- Y 作为优化变量可以寻找最能解释生物数据的的通解

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1l} & 0.0 & \cdots & 0.0 \\ y_{21} & y_{22} & \cdots & y_{2l} & 0.0 & \cdots & 0.0 \\ \cdots & & \cdots & & & & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nl} & 0.0 & \cdots & 0.0 \end{bmatrix}$$

l 为基因表达矩阵中非零奇异值的个数

虽然每个时间序列只有有限的几个时间点，但是却有成千上万的时间序列

•众多的基因芯片平台来测量基因表达



•公开数据库中的基因表达数据每年翻三番

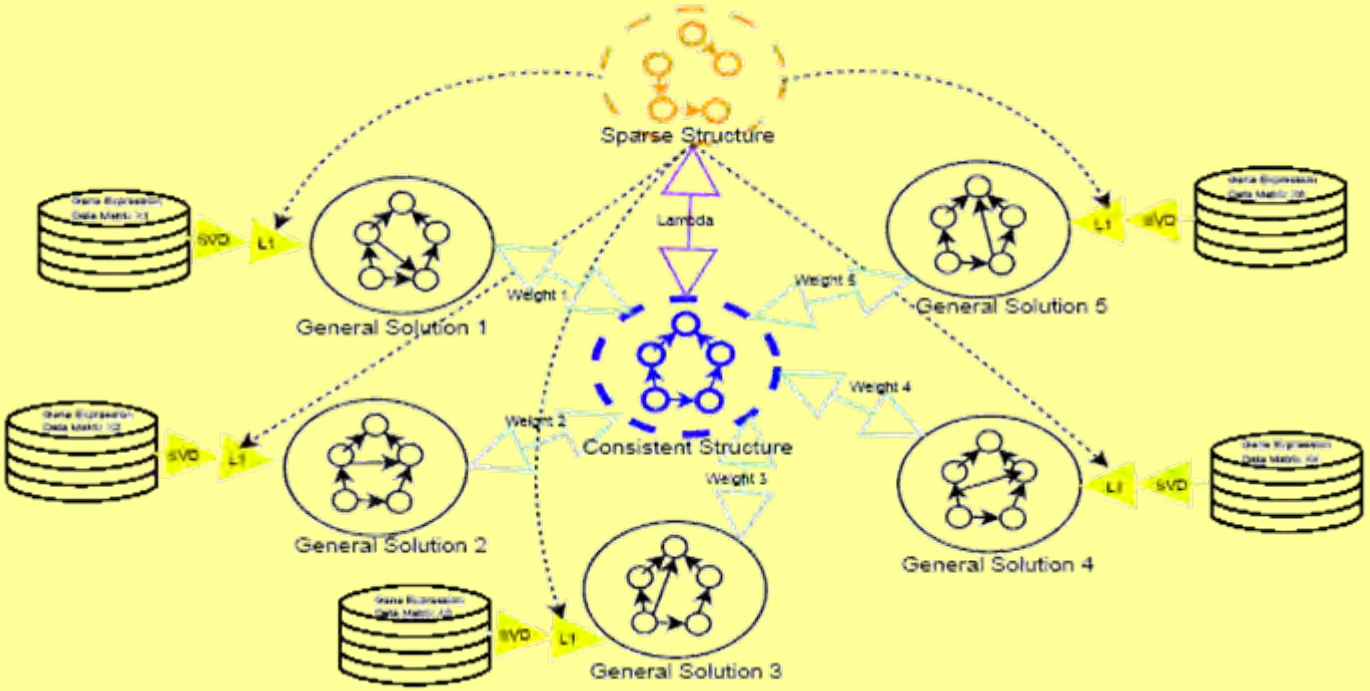
- NCBI Gene Expression Omnibus

GEO Gene Expression Omnibus 137231 experiments

- EBI Array Express

EMBL-EBI 55228 experiments

工作一：集成多个基因表达数据集



Yong Wang, Trupti Joshi, Xiang-Sun Zhang, Dong Xu, and Luonan Chen. Inferring gene regulatory networks from multiple microarray datasets, ***Bioinformatics***, 22, 2413-2420, 2006.

最优化模型

每个单数据集可以得到一个通解表达，代表与这个数据集相容的所有网络结构

集成多个数据集的目标是构建一个与各个数据集尽量相容的聚合的**(aggregate)**网络结构

$$\min_{Y,J} \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{n} [\omega^k |J_{ij} - J_{ij}^k| + \lambda |J_{ij}|]$$

目标 1：聚合的网络同各个数据集产生的网络之间尽量相容

目标 2：所得的生物网络结构是稀疏的

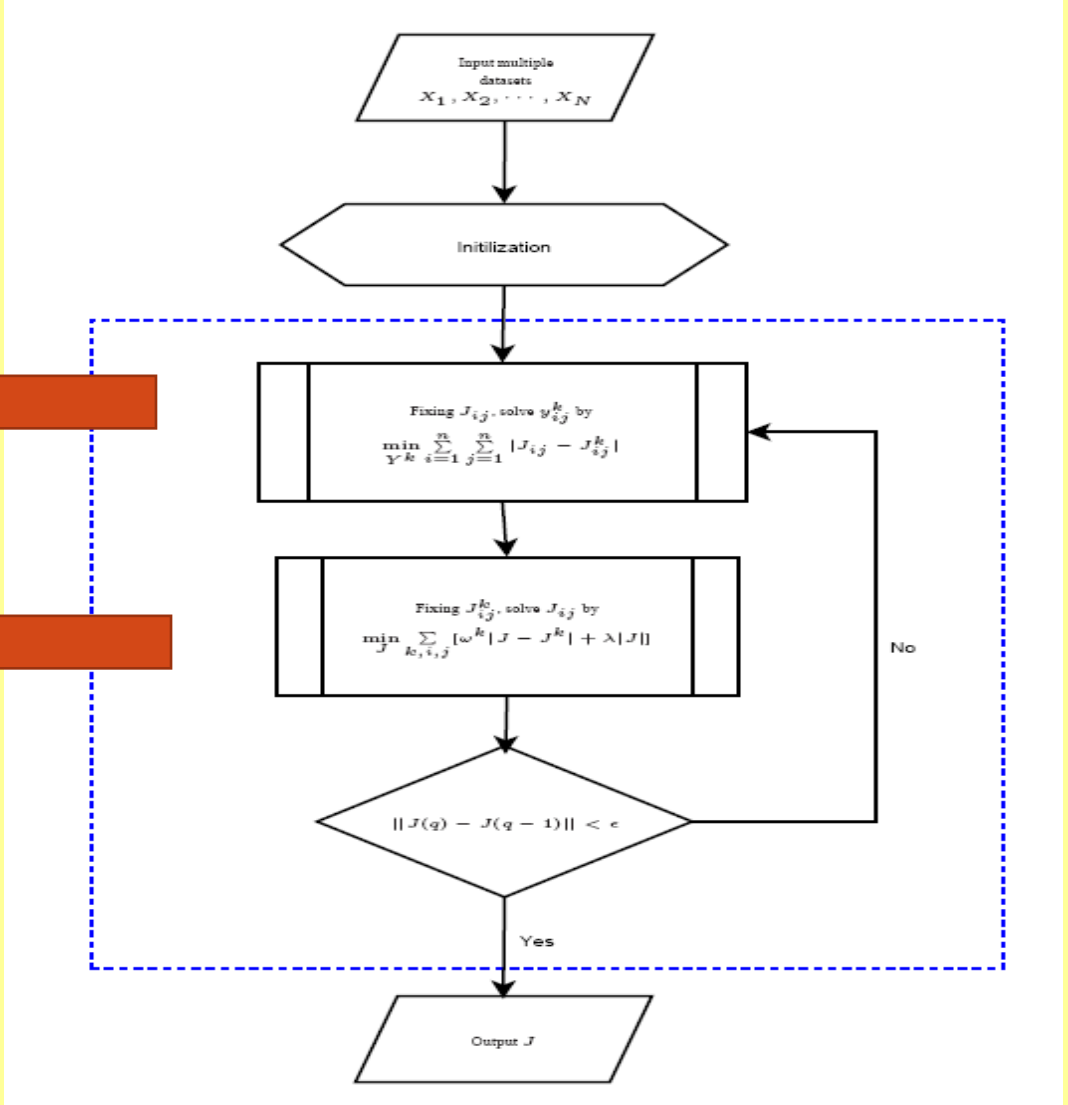
多目标转化为单目标：引入一个参数来权衡这两个优化目标

分解算法

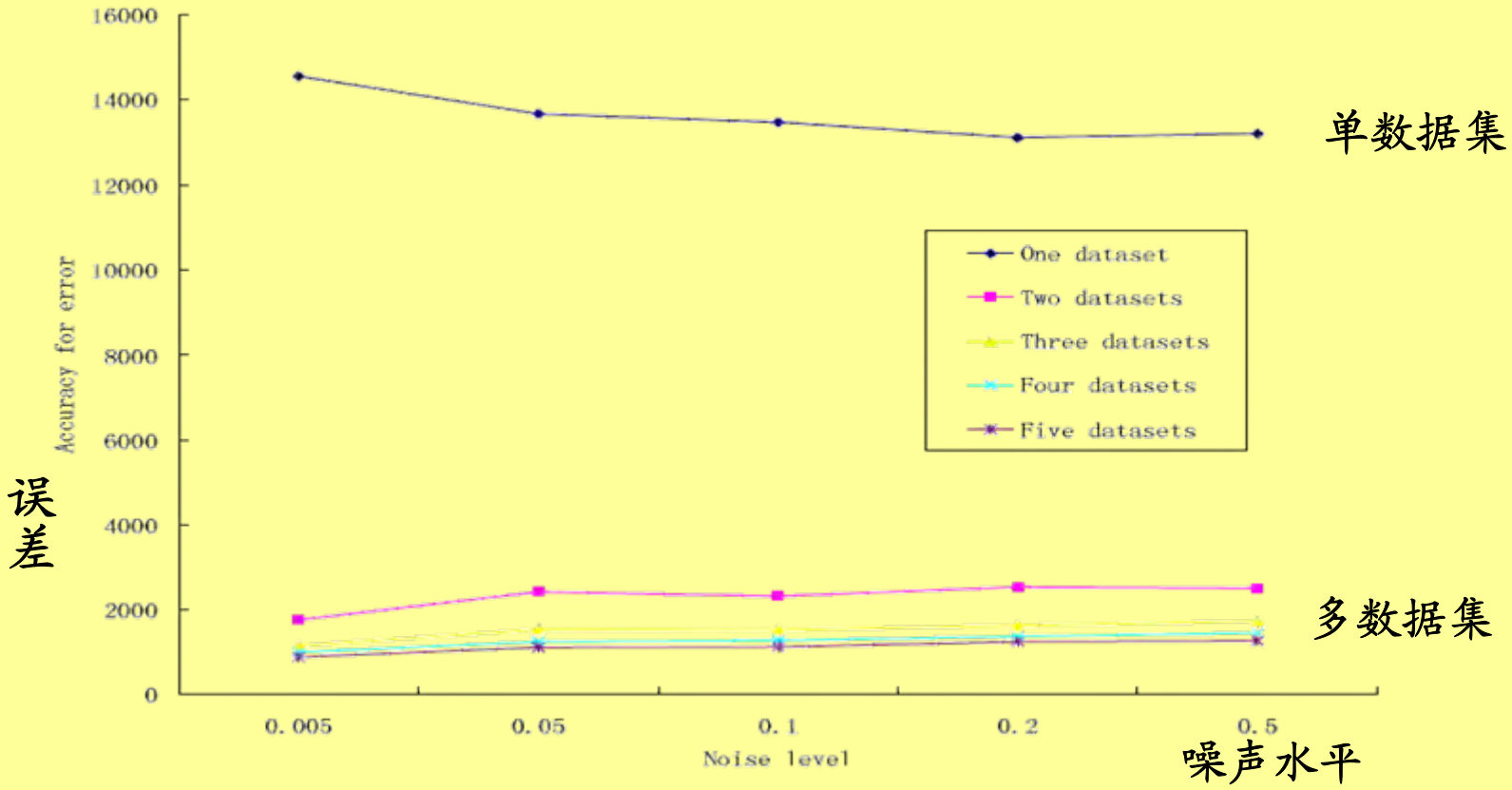
上述问题可以转化为大规模的线性规划问题

固定J，求Y

固定Y，求J



多个数据集显著提高精度



工作二：集成先验信息的线性规划模型

有很多对网络结构推断有价值的的先验信息，例如从数据库或文献中得到的基因间调控数据，这些信息可通过添加线性规划的约束来提高所得到的聚合网络的精度。

$$\min_{Y^1,Y^2,\ldots,Y^N,L} \quad \sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n+s}\omega_k \left| L_{ij} - L_{ij}^k \right| + \lambda \sum_{(i,j)\in\{(i,j)|K_{ij}=0 \text{ or } U_{ij}=0\}} \left| L_{ij} \right|$$

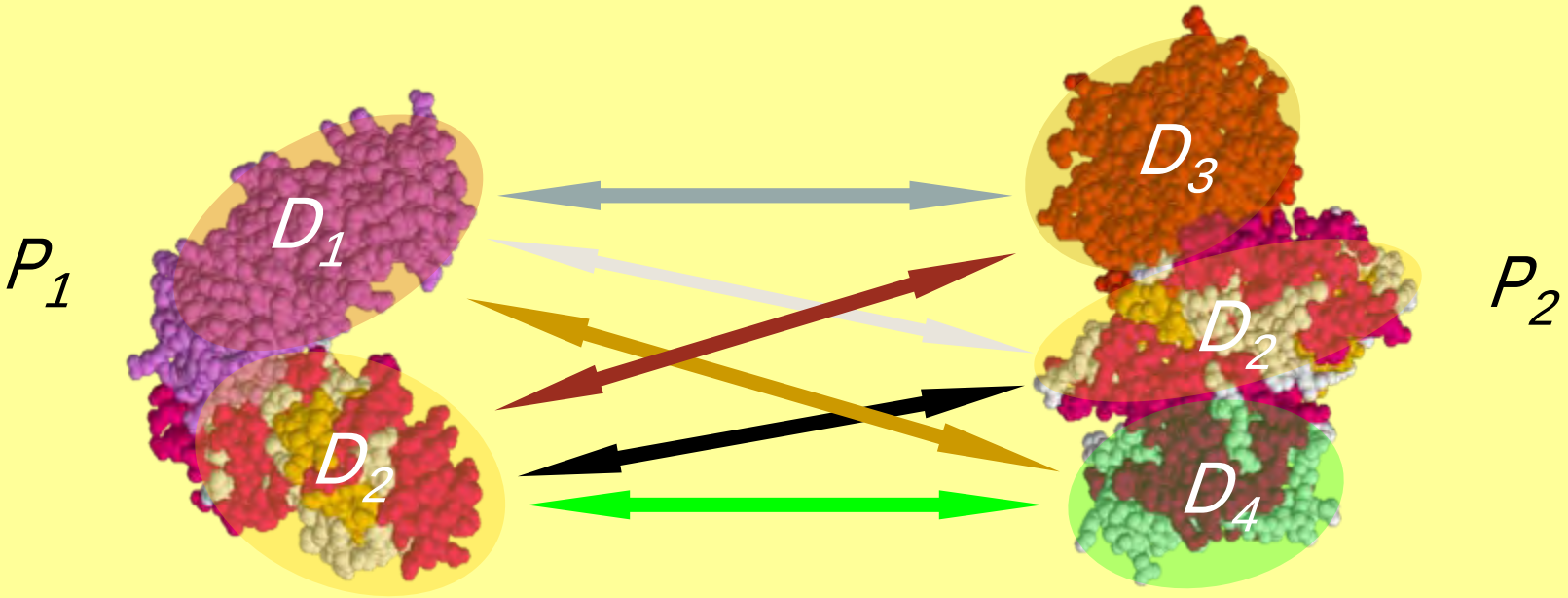
$$s.t. \qquad L_{ij} > 0 \quad \text{if} \quad K_{ij} > 0 \qquad i,j \in \{1,2,\ldots,n\}$$

$$L_{ij} < 0 \quad \text{if} \quad K_{ij} < 0 \qquad i,j \in \{1,2,\ldots,n\}$$

$$L_{ij} = 0 \quad \text{if} \quad E_{ij} = 0 \qquad i,j \in \{1,2,\ldots,n\}$$

Yong Wang, Trupti Joshi, Xiang-Sun Zhang, Dong Xu, and Luonan Chen. Supervised Inference of Gene Regulatory Networks by Linear Programming, *Lecture Notes of Bioinformatics*, Vol. 4115, pp. 551-561, Springer-Verlag, 2006.

例二：蛋白质相互作用网络推断的最节俭模型



已知蛋白质相互作用

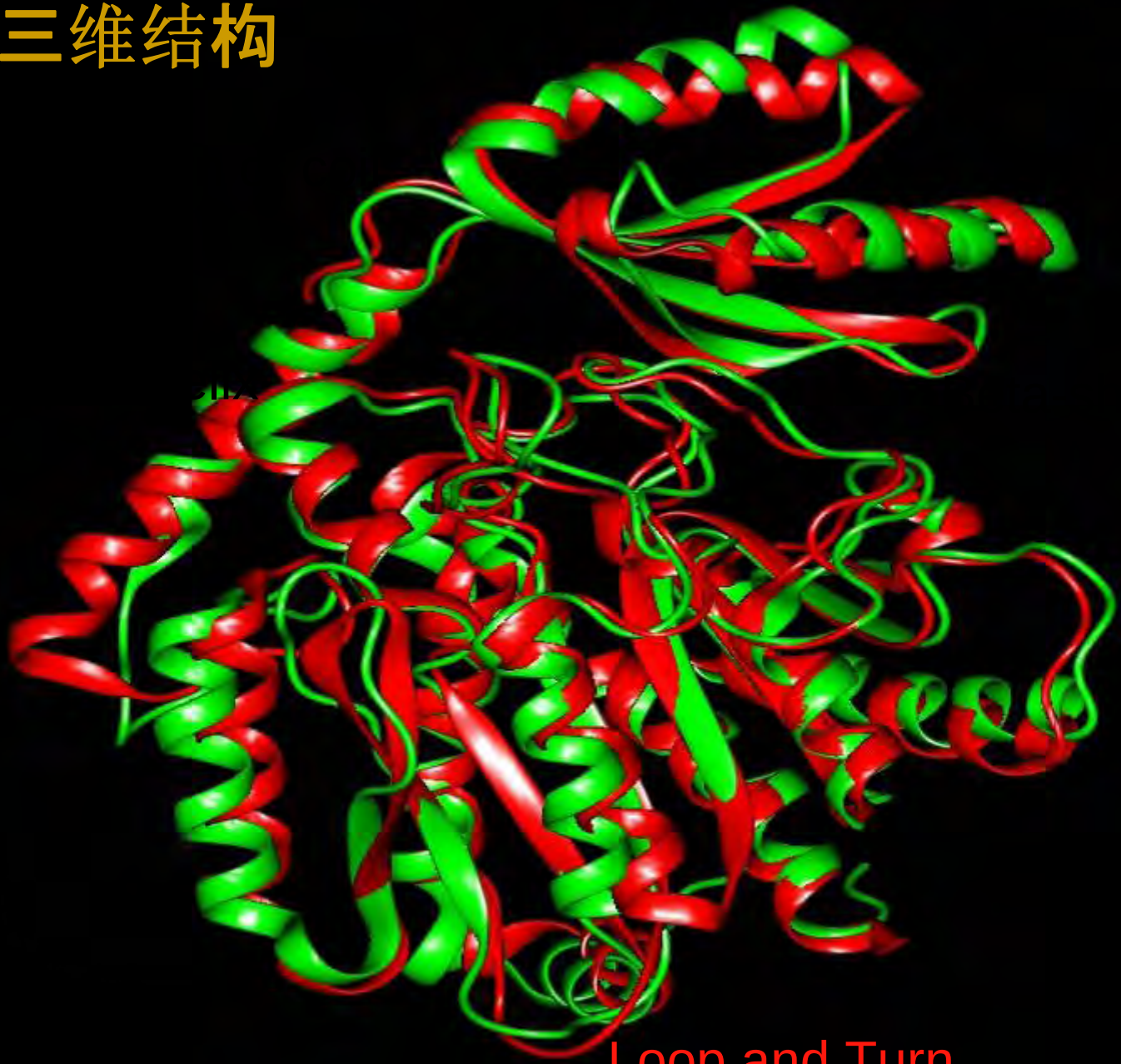
优化模型

预测未知蛋白质相互作用

什么是蛋白质相互作用网络？

- 图论语言：
  - 节点：蛋白质
  - 边：蛋白质之间的直接物理相互作用
  - 无向图
- 生物学语言：
  - 蛋白质有三维空间结构

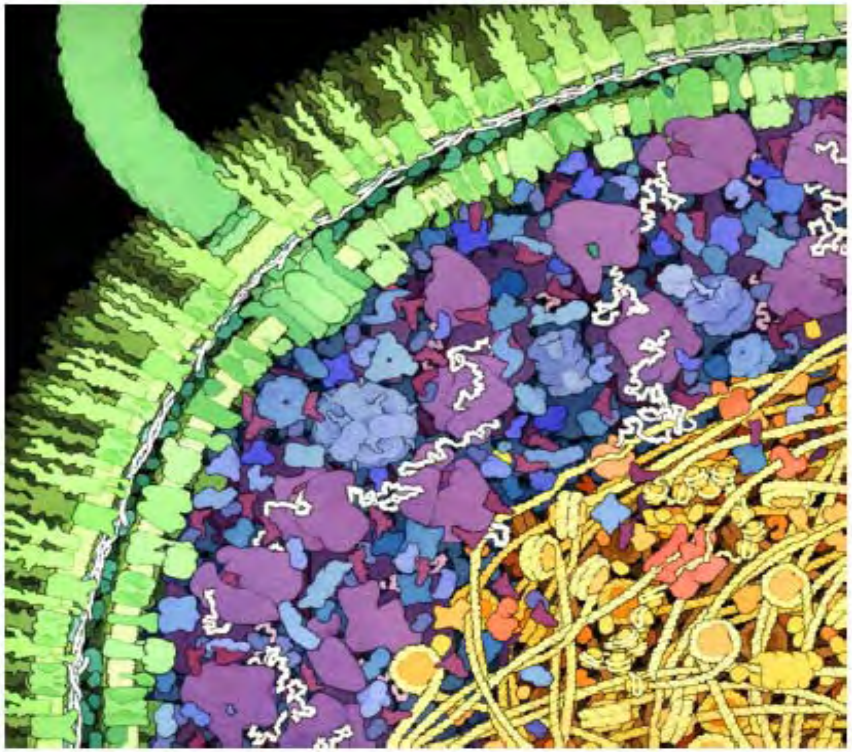
蛋白质三维结构



什么是蛋白质相互作用网络？

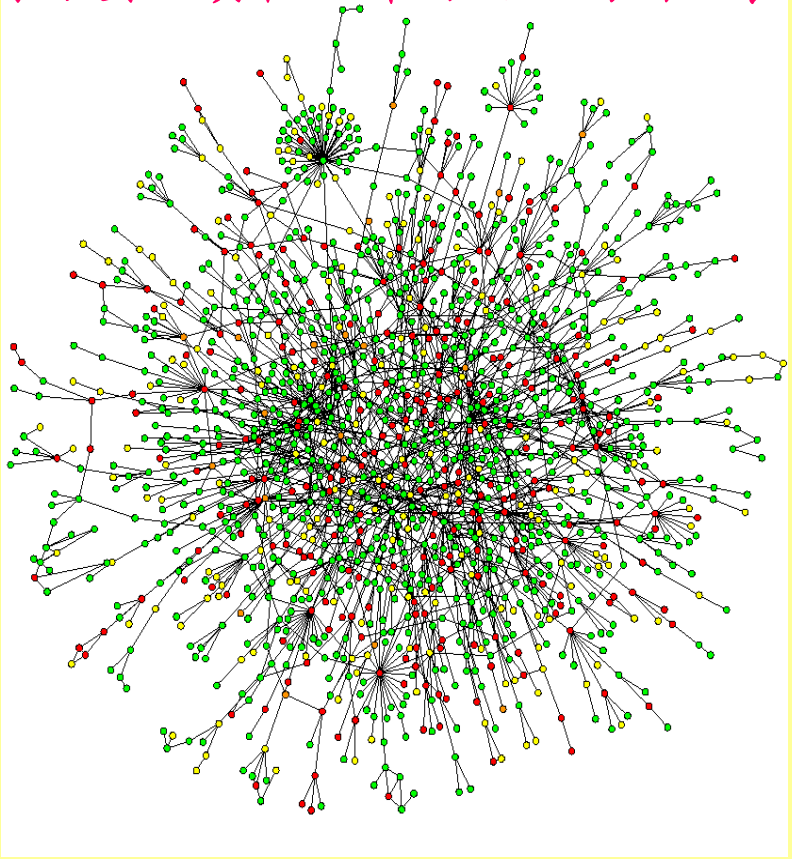
- 图论语言：
  - 节点：蛋白质
  - 边：蛋白质之间的直接物理相互作用
  - 无向图
- 生物学语言：
  - 蛋白质有三维空间结构
  - 蛋白质A与蛋白质B的相互作用、两个蛋白质的三维结构在空间中通过化学键结合

蛋白质相互作用网络



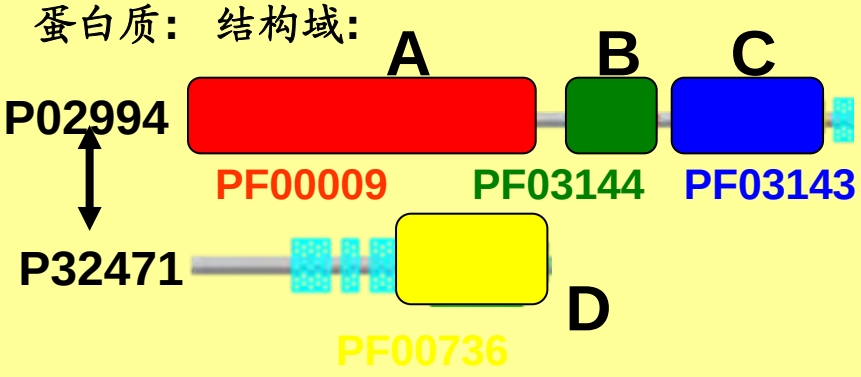
细胞内真实的蛋白质相互作用网络

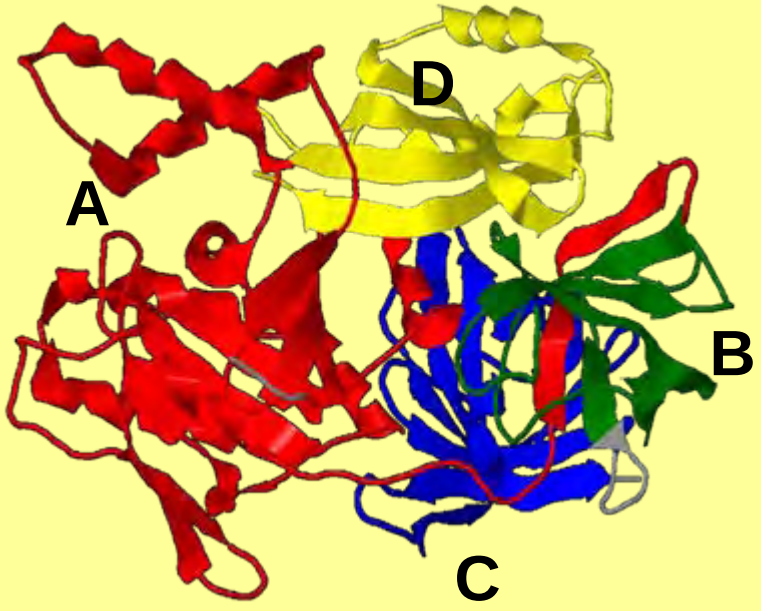
酵母蛋白质相互作用网络的图论表示



蛋白质结构域 (Domain)

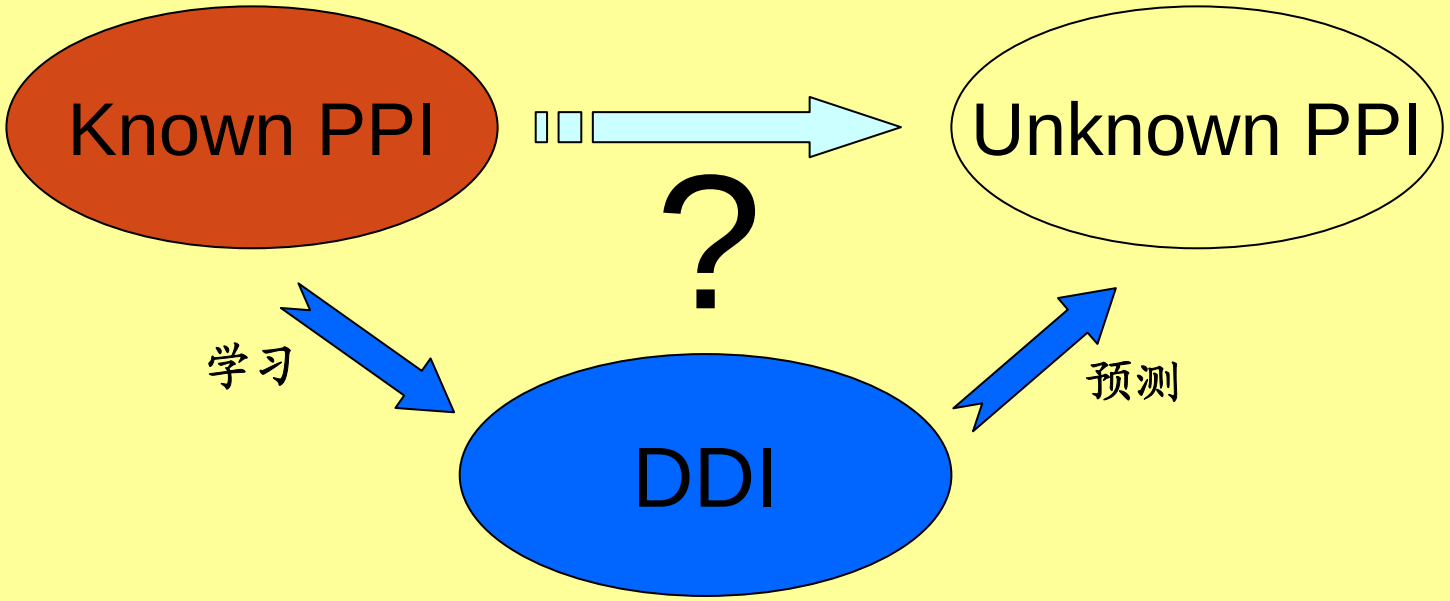
- 结构域 (Domain): 蛋白质的一个组成部分，它的空间结构可独立于其他氨基酸折叠而成，具有自己的功能。一个蛋白质中的结构域组合在一起确定该蛋白质的总的功能。





基于结构域的蛋白相互作用预测

- 基本假设: 两个蛋白质相互作用 (PPI)，至少各有一个结构域相互作用(domain-domain interaction, DDI)



蛋白质相互作用网络预测的概率模型

- 基本假定:
  - 每一 DDI 是相互独立的
  - 两个蛋白质相互作用，至少有一个 DDI

- 于是，知道了DDI的概率，可以推出PPI的概率

$$\Pr(p_{ij}=1)=1-\prod_{D_{mn}\in P_{ij}}(1-\Pr(d_{mn}=1))$$

工作一：蛋白质相互作用预测的概率模型

- 关联概率方法 (An association probabilistic method):

$$\lambda_{mn} = \frac{\sum_{\substack{D_{mn} \in P_{ij} \\ P_{ij} \in \mathcal{P}_{train}}} \left(1 - (1 - \rho_{ij})^{\frac{1}{|P_{ij}|}}\right)}{\sum_{\substack{D_{mn} \in P_{ij} \\ P_{ij} \in \mathcal{P}_{train}}} 1}$$

Luonan Chen, Ling-Yun Wu, Yong Wang, and Xiang-Sun Zhang. Inferring Protein Interactions from Experimental Data by Association Probabilistic Method. *Proteins: Structure, Function, and Bioinformatics*, Vol. 62, pp. 833-837, 2006.

工作二：蛋白质相互作用预测的优化模型——最节俭模型（*Parsimony Model*）

基本思想：

- 两个蛋白质相互作用， 在各自的domain组中仅有少数几对 domain对相互作用
- 对于已出现在多对PPI中的某一对DDI，一旦出现在另一蛋白质对的domain对中，很有可能成为这一对蛋白质相接的媒介
- 我们用尽可能少的DDI对来“解释”被观察到的PPI对.

Xiang-Sun Zhang, Rui-Sheng Wang, Lin-Yun Wu, ShihuaZhang, Luonan Chen, Inferring Protein-Protein Interactions by Combinatorial Models, *IFMBE Proceedings*, vol.14, pp.181-184,World Conference on Medical Physics and Biomedical Engineering,Seoul, Korea, Springer-Verlag, 2006

整数线性规划

- **Parsimony Method (PM): 用以下的整数线性规划 (ILP) 来描述:**

$$\min_{d_{mn}} \sum_{D_{mn} \in \mathcal{D}} d_{mn}$$

$$\text{s.t.} \sum_{D_{mn} \in P_{ij}} d_{mn} \geq 1 \quad \text{for } P_{ij} \in \mathcal{P}$$

$$d_{mn} \in \{0, 1\} \quad \text{for all } m, n.$$

- **所得到的解称为节俭的DDI组**

考虑实验数据中的噪声

- **sd (searching degree)称为有效信息利用度，是指所有被观察到的PPI对中具有有效信息的利用部分的比例。以上的ILP可以改进为**

$$\min_{d_{mn},e_{ij}} \sum_{D_{mn}\in\mathcal{D}} d_{mn}$$

$$\text{s.t.} \sum_{D_{mn}\in P_{ij}} d_{mn} + e_{ij} \geq 1 \quad \text{for } P_{ij} \in \mathcal{P}$$

$$\sum_{P_{ij}\in\mathcal{P}} e_{ij} \leq (1 - sd) \cdot |\mathcal{P}|$$

$$d_{mn} \in \{0, 1\} \quad \text{for all } m, n$$

$$e_{ij} \in \{0, 1\} \quad \text{for all } P_{ij} \in \mathcal{P}.$$

- **改进后的模型称为 ILP_sd.**

考虑不相互作用数据

- 我们还可以利用已知的负例子集（a gold-standard negative data set）N，即由这些蛋白在细胞中的定位信息可知它们不会相接，按假定它们的domain也不会相接。此时模型为

$$\min_{d_{mn}} \sum_{D_{mn}\in\mathcal{D}} d_{mn} + M \sum_{D_{mn}\in N_{ij}} d_{mn}$$

$$\text{s.t.} \sum_{D_{mn}\in P_{ij}} d_{mn} \geq 1 \quad \text{for } P_{ij} \in \mathcal{P}$$

$$d_{mn} \in \{0,1\} \quad \text{for all } m,n$$

- 记这一模型为 ILP_neg.

模型的计算复杂性分析

- 基本模型 ILP 是图论中有名的 “Hitting Set problem”，是NP-hard problem.
- 我们直接解这些问题的线性松弛问题，分别记为 LP, LP_sd, LP_neg
- LP_sd 和 LP_neg 还可以结合成为 LP_neg_sd:

$$\min_{d_{mn},e_{ij}} \sum_{D_{mn}\in\mathcal{D}} d_{mn} + M \sum_{D_{mn}\in N_{ij}} d_{mn}$$

$$\text{s.t.} \sum_{D_{mn}\in P_{ij}} d_{mn} + e_{ij} \geq 1 \quad \text{for } P_{ij} \in \mathcal{P}$$

$$d_{mn} \in \{0,1\} \quad \text{for all } m,n$$

$$\sum_{P_{ij}\in\mathcal{P}} e_{ij} \leq (1-sd)\cdot|\mathcal{P}|$$

$$e_{ij} \in \{0,1\} \quad \text{for all } P_{ij} \in \mathcal{P}.$$

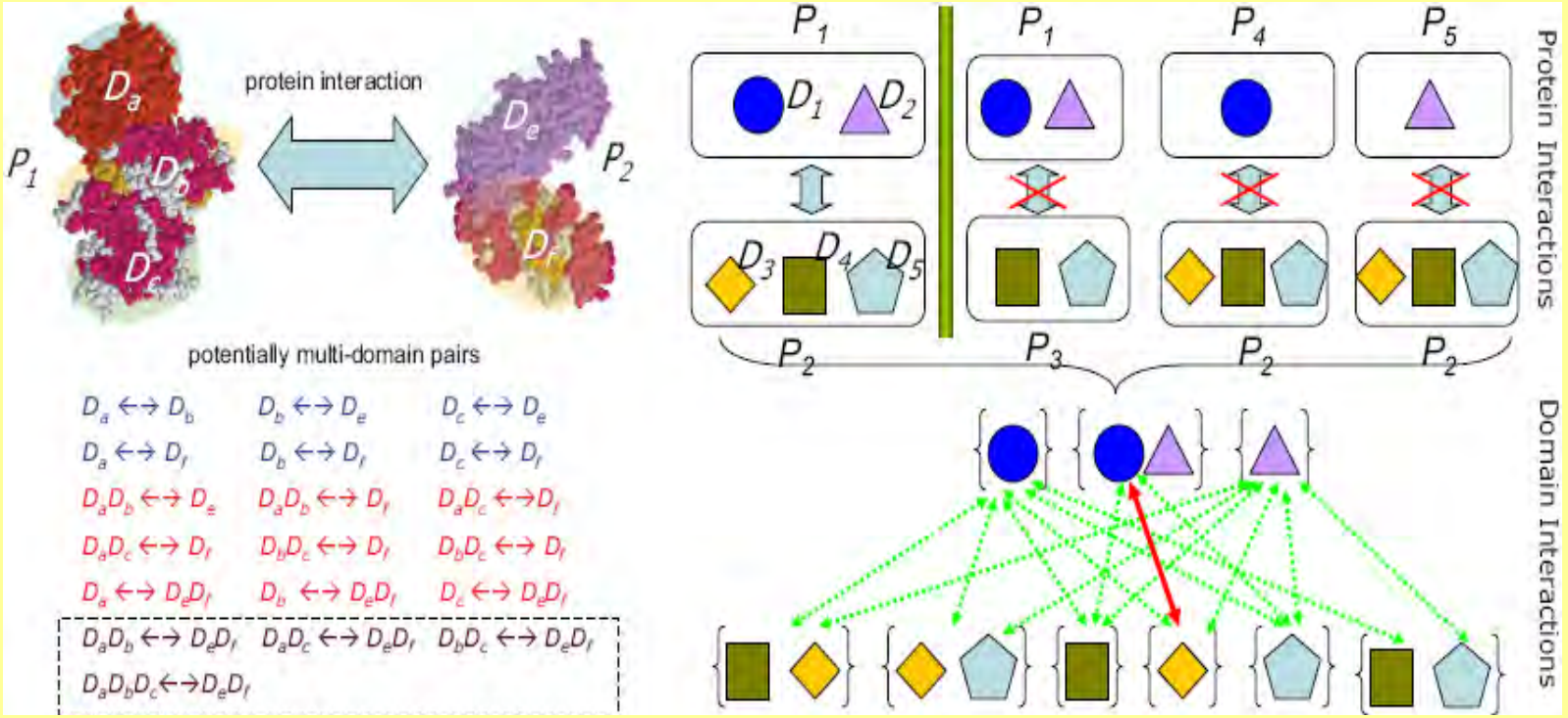
蛋白质相互作用的预测

得到一组节俭的 DDI 组以后，我们用以下公式来预测蛋白质对 Pij 是否相互作用：

$$\text{If} \sum_{D_{mn} \in P_{ij}} d_{mn} \geq 1, \text{then } p_{ij} = 1, \text{else } p_{ij} = 0$$

此处 ρ 可以设为 1 或任意小于 1 的正数.

工作三：扩展到多个结构域的相互作用



基本思想：多结构域之间的合作作为新的变量考虑

考虑多结构域相互作用的线性规划模型

$$\frac{\Pr(o_{ij}^k = 1) - fp^k}{1 - fn^k - fp^k} = 1 - \prod_{D_{m,n} \in P_{ij}^k} (1 - \Pr(d_{m,n} = 1)) \cdot \prod_{D_{mr,n} \in P_{ij}^k} (1 - \Pr(d_{mr,n} = 1)) \cdot \prod_{D_{m,nr} \in P_{ij}^k} (1 - \Pr(d_{m,nr} = 1))$$

$$\min_{\varepsilon,\ x} \quad \sum_{P_{ij}^k} |\varepsilon_{ij}^k|$$

$$\text{s.t.} \sum_{D_{m,n} \in P_{ij}^k} x_{m,n} + \sum_{D_{mr,n} \in P_{ij}^k} x_{mr,n} + \sum_{D_{m,nr} \in P_{ij}^k} x_{m,nr} = \beta_{ij}^k - \varepsilon_{ij}^k \text{ for all } P_{ij}^k$$

$$x_{m,n} \le 0, x_{mr,n} \le 0, x_{m,nr} \le 0,$$

$$i, j = 1, \ldots, N_k,\ k = 1, \cdots, K$$

Rui-Sheng Wang, Yong Wang, Ling-Yun wu, Xiang-Sun Zhang and Luonan Chen. Analysis on multi-domain cooperation for predicting protein-protein interactions. *BMC Bioinformatics*, 8:391, 2007.

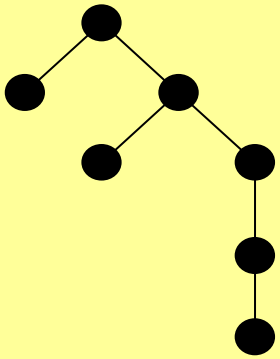
例三：生物分子网络比对

- 不同物种的生物分子相互作用网络中存在相同或相似的子图，即功能或遗传保守区域。这些保守区域在遗传学和生物医学研究中具有重要作用。

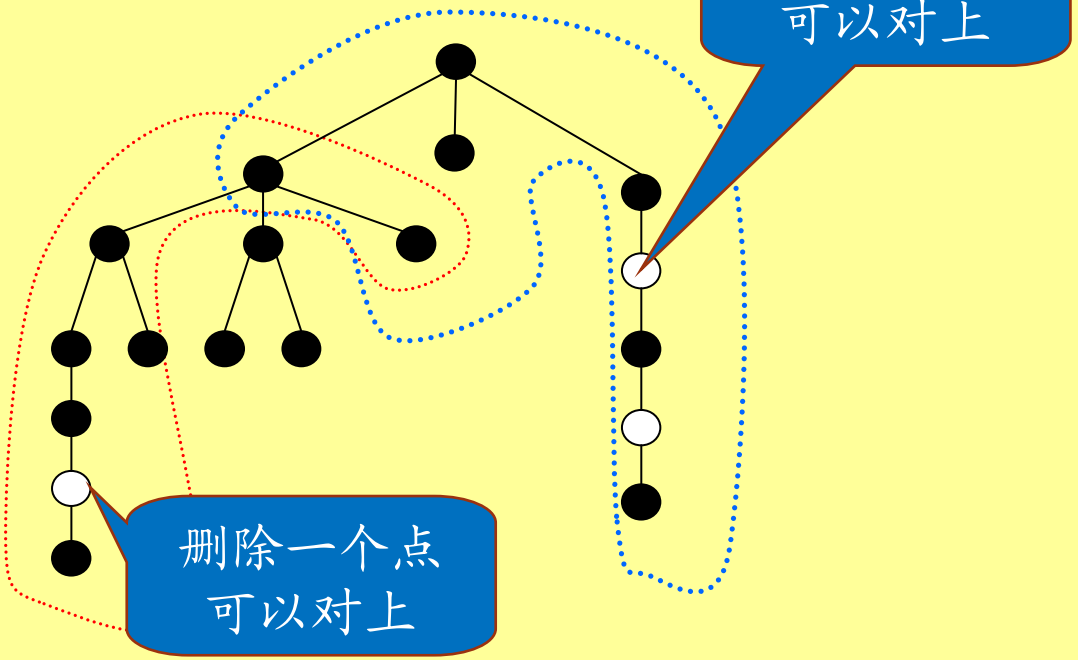
- “保守”意味着两个网络中的子图包含着执行相同或相似功能的蛋白质，并且它们之间的连接也相似。(点相似，点之间的连接也保持)

简单网络比对示意图

网络A

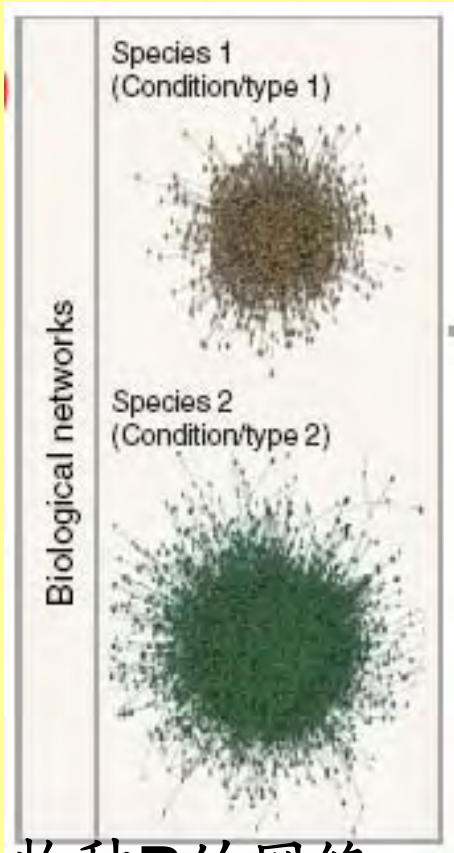


网络B



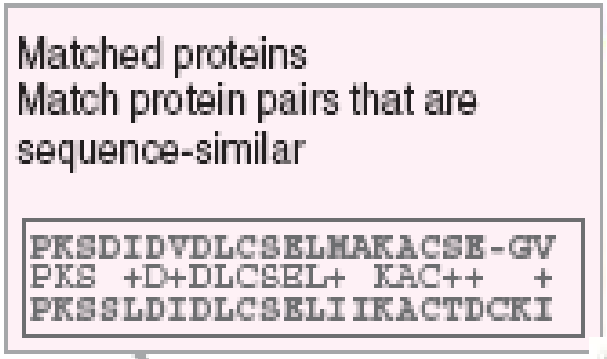
那种比对更好?

物种A的网络

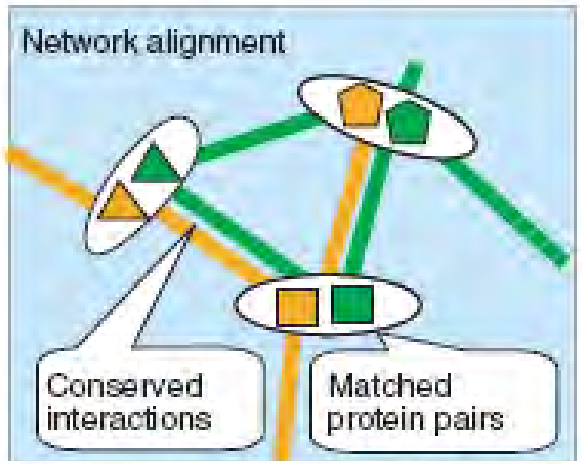


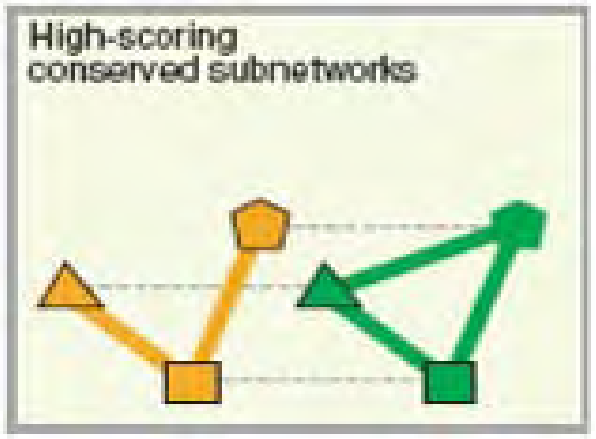
物种B的网络

节点的相似性



相似的子图





- 图论中经典的子图同构问题是NP 难问题
- 生物网络比对问题更难，因需要考虑节点相似性，并允许插入删除节点

我们的网络比对方法——整数二次规划

给定两个网络$G_1=(V_1, E_1)$, $G_2=(V_2,E_2)$,不妨设

$$V_1 = \{v_1^1, v_2^1, \ldots, v_m^1\},$$
$$V_2 = \{v_1^2, v_2^2, \ldots, v_n^2\},$$

两个网络的邻接矩阵分别为

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{pmatrix} \qquad B = \begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ b_{21} & b_{22} & \ldots & b_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ b_{n1} & b_{n2} & \ldots & b_{nn} \end{pmatrix}$$

其中

$$a_{ij} = \begin{cases} 1, & if\ (v_i^1, v_j^1) \in E_1 \\ 0, & otherwise \end{cases} \qquad b_{ij} = \begin{cases} 1, & if\ (v_i^2, v_j^2) \in E_2 \\ 0, & otherwise \end{cases}$$

两个网络的结点之间的相似性

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{pmatrix}$$

其中 s_{ij}表示第一个网络中的v_i^1与第二个网络中的v_j^2之间的相似性。节点相似性的确定方法：

（1）用两个蛋白或基因之间的序列相似性进行量化；

（2）根据两个蛋白或基因的同源相似性进行量化；

定义变量

$$x_{ij} = \begin{cases} 1 & \text{if } v_i^1 \in V_1 \text{ matches } v_j^2 \in V_2, \\ 0 & \text{otherwise} \end{cases}.$$

则网络比对问题可以用如下整数二次规划模型表示

$$\max_X \quad f(G_1, G_2) = \lambda \sum_{i=1}^{m} \sum_{j=1}^{n} s_{ij} x_{ij} + (1-\lambda) \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{l=1}^{n} a_{ik} b_{jl} x_{ij} x_{kl}$$

$$s.t. \begin{cases} \sum_{j=1}^{n} x_{ij} \leq 1 & i = 1, 2, \cdots m \\ \sum_{i=1}^{m} x_{ij} \leq 1 & j = 1, 2, \cdots n \\ x_{ij} = 0, 1 & i = 1, 2, \cdots m; j = 1, 2, \cdots, n \end{cases}$$

模型解释

目标函数：第一项对应比对结果的点相似度，第二项对应比对结果中的边的相似度。

目标函数中的参数λ是调整点和边相似重要性的参数。

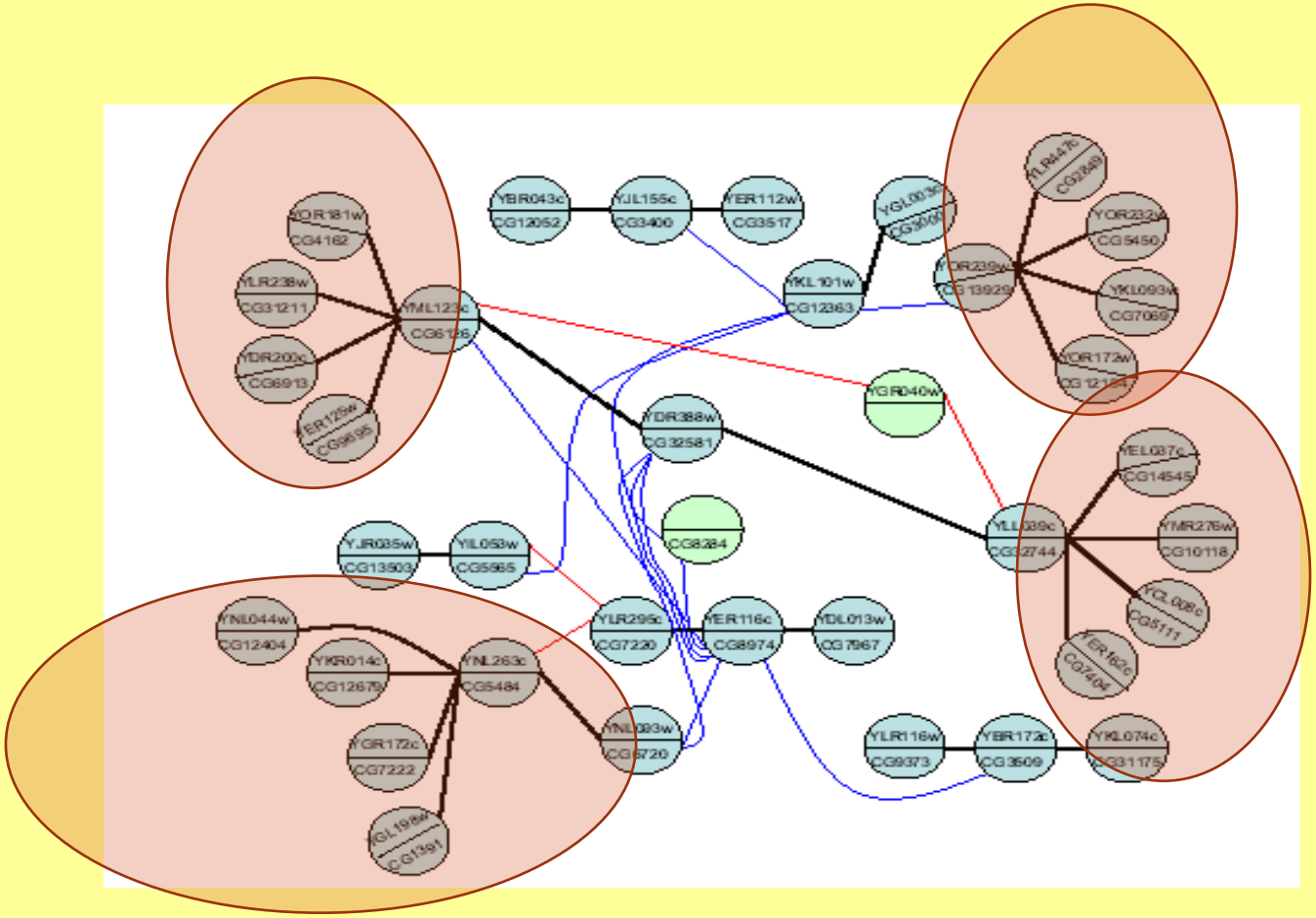
约束条件：一个网络中的每一个点最多只能与另一个网络中的一个点对应。

注：由于模型约束条件系数矩阵是一个全单模矩阵，所以可以直接使用松弛问题求解该整数二次规划模型，一般情况下，得到的解都是整数解。这里有一个生物模型数据结构的问题！

- **Zhenping Li, Shihua Zhang, Yong Wang, Xiang-Sun Zhang and Luonan Chen. Alignment of molecular networks by integer quadratic programming. *Bioinformatics*， Vol. 23, no.13, pp. 1631–1639, 2007.**

- **Zhenping Li, Yong Wang, Shihua Zhang, Xiang-Sun Zhang, Luonan Chen. Alignment of protein interaction networks by integer quadratic programming. 28th *IEEE EMBS* Annual International Conference.**

酵母和果蝇蛋白相互作用网络比对结果



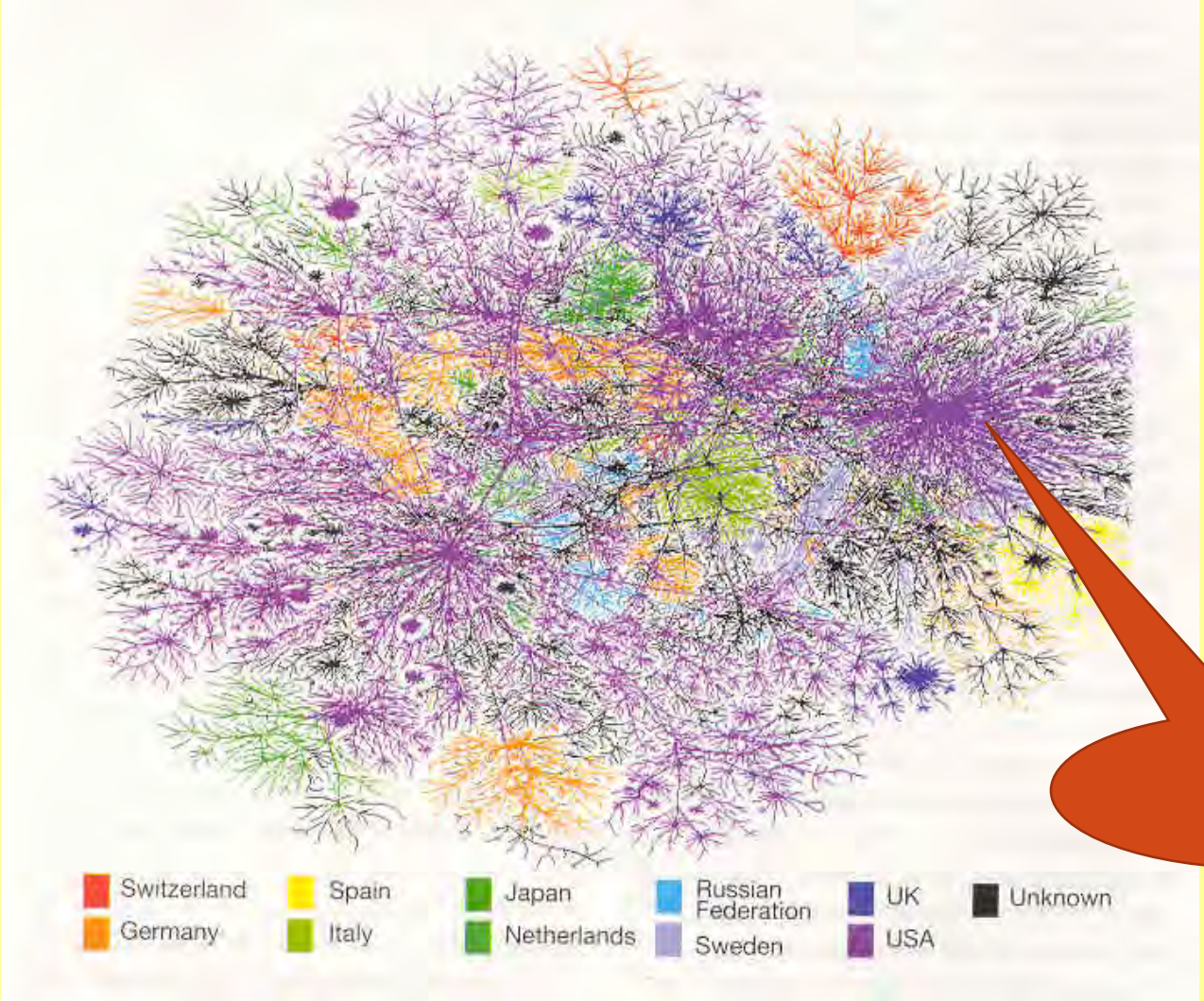
找出进化中保守的模块

我们的研究 (Part II):
一般复杂网络性质的最优化研究

一般复杂网络

- 许多复杂系统可以用网络进行表示
- 社会网络：科学合作网、食物网络和运输网络等
- 科学技术网络：因特网, 万维网, 软件相关网络等
- 生物分子网络：蛋白质相互作用网络, 基因调控网络, 新陈代谢网络等
- 复杂网络的一些拓扑性质：
  - 小世界(Small world)
  - 无尺度（Scale free)
  - 聚类特性 (Clustering)
  - ……

例如：世界IP地址网络


Switzerland
Germany
Spain
Italy
Japan
Netherlands
Russian Federation
Sweden
UK
USA
Unknown
美国

复杂网络的模块化性质

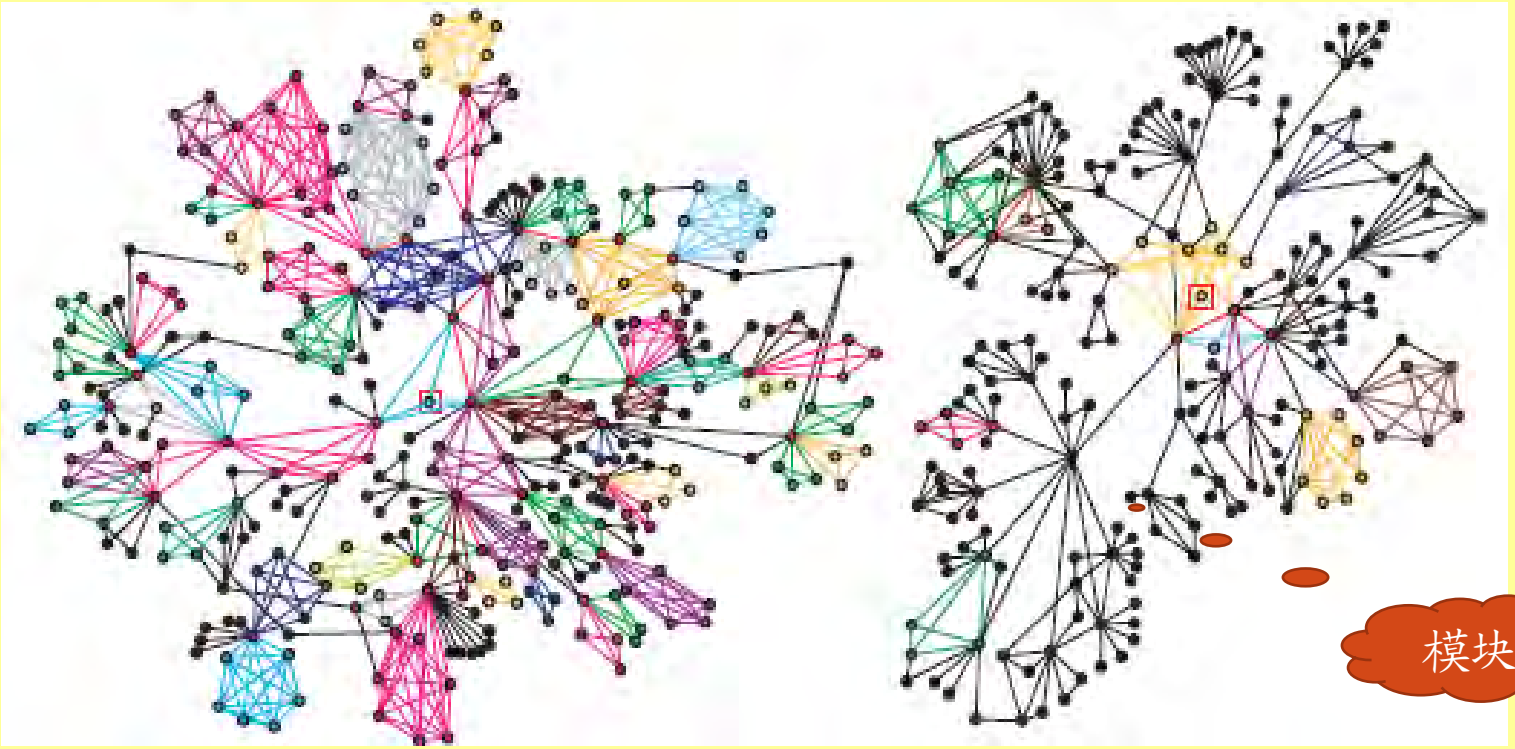
- 复杂网络中存在模块或者社区结构 (Module or Community structure)
- 模块或者社区定义为网络中内部连接稠密，与外部连接稀疏的节点的集合 (Filippo Radicchi et. al. *PNAS*, Vol.101, No.9, *2658-2663*, 2004).
- 数学表述:

$$\sum_{i \in V} k_i^{\text{in}}(V) > \sum_{i \in V} k_i^{\text{out}}(V)$$

其中V是子图，K是顶点的度。即子图 V 是模块的条件是模块内顶点的内部连边的度值之和大于模块内顶点的外部连边的度值之和。

PNAS --- Proc. Natl. Acad. Sci. USA 美国科学院院刊

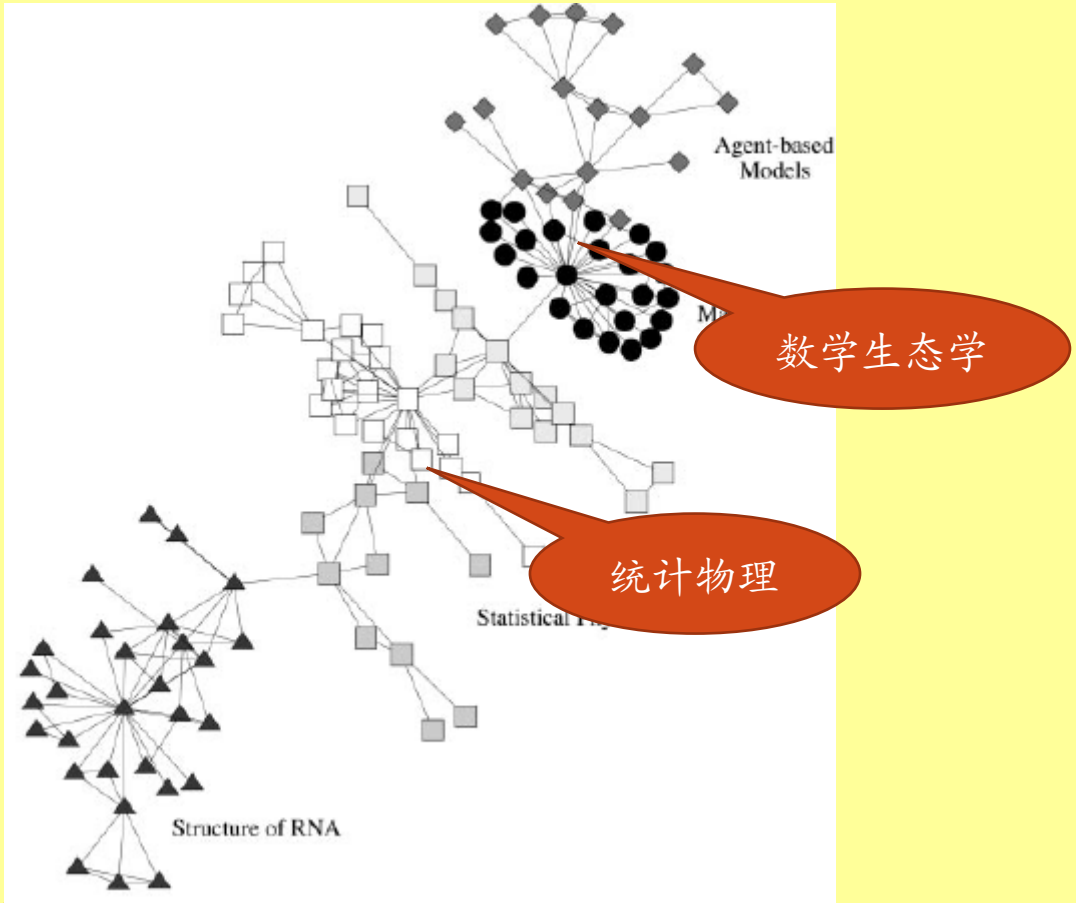
复杂网络的模块性质



发表文章合作网　　电话网络

模块划分的重要性

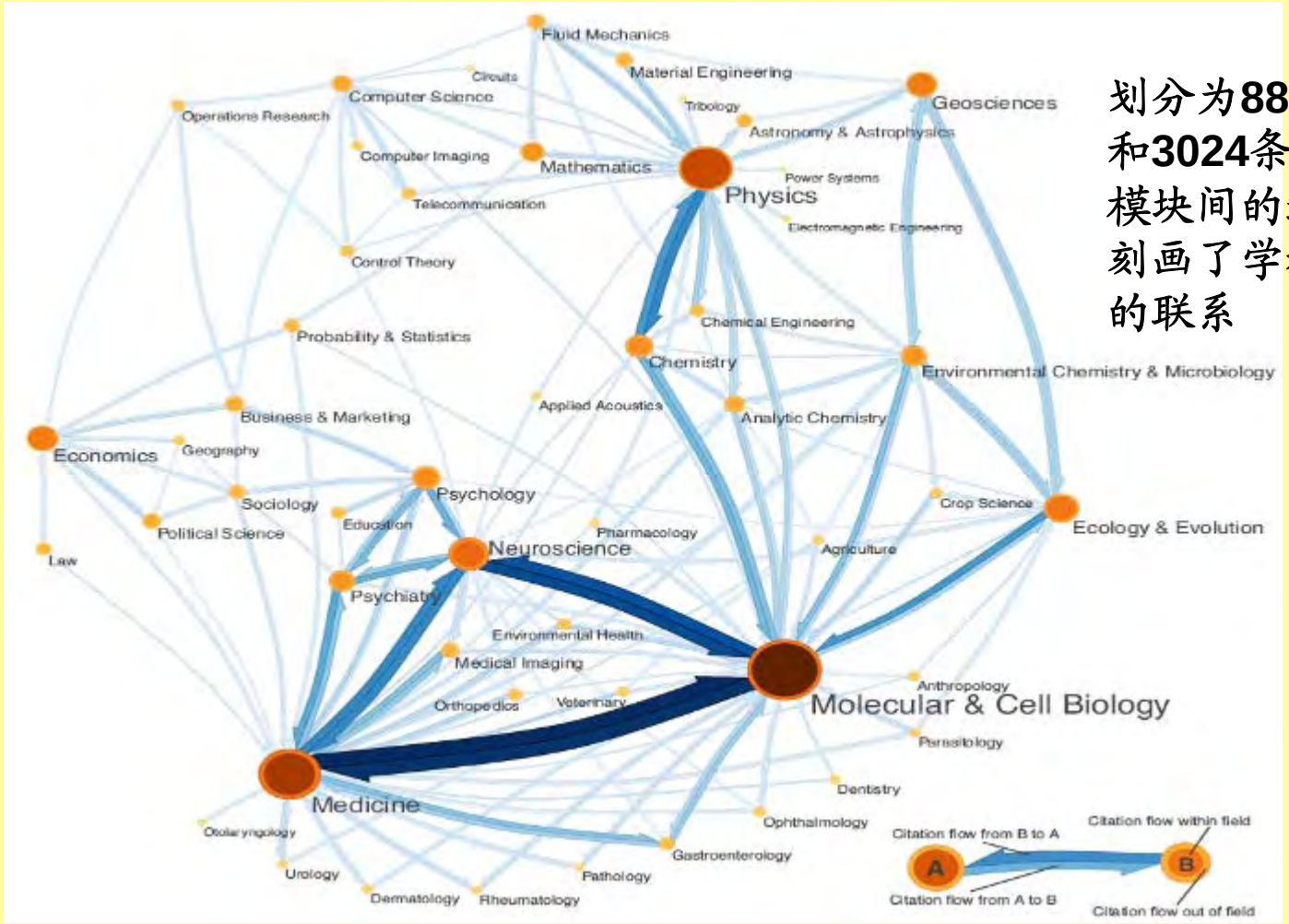
- 许多复杂网络共有的性质。
- 研究模块结构有助于研究整个网络的结构和功能



圣塔菲研究所的科学家合作网：模块代表从事相似领域研究的科学家集合

Martin Rosvall, Carl T. Bergstrom, PNAS, vol. 105, no.4. 1118-1123, 2007

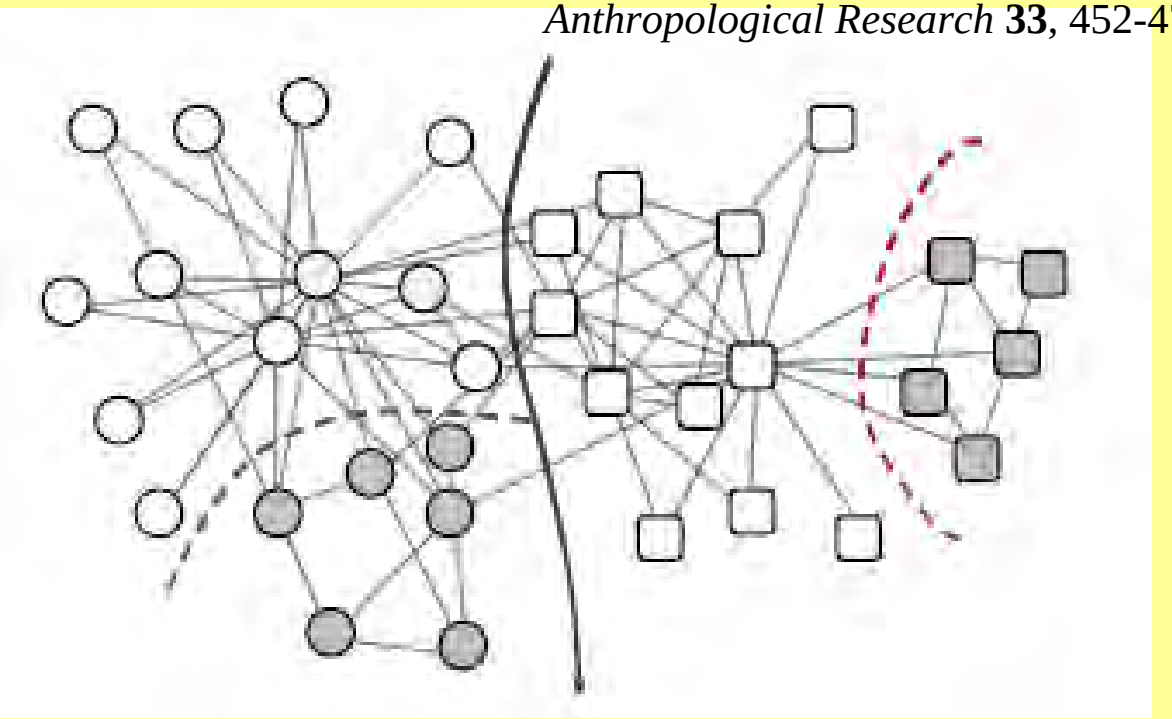
自然科学论文引用网络：6128期刊，约600万次引用，



划分为88个模块和3024条模块间的连接，刻画了学科之间的联系

一个社会网络的例子

W. W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* **33**, 452-473 1977



- 1970年美国大学里的一个空手道俱乐部关系网络：节点是其34名成员，边是他们两年间的友谊关系，边数为78。俱乐部里的矛盾导致其分裂为两个小的俱乐部。问题是能否用网络的模块结构来重现这个过程？
- 它是模块探测研究中的经典例子。

工作一：模块探测方法研究

- 有了模块的定义，网络的模块探测问题就是对网络进行划分，使得划分得到的一系列非空的模块满足基本模块定义。
- 我们发展了一些基于图论和矩阵谱分解的模块探测算法

Shihua Zhang, Rui-Sheng Wang, and Xiang-Sun Zhang. Identification of Overlapping Community Structure in Complex Networks Using Fuzzy c-means Clustering. *Physica A*, 2007, 374, 483–490.

Shihua Zhang, Rui-Sheng Wang and Xiang-Sun Zhang. Uncovering fuzzy community structure in complex networks. *Physical Review E*, 76, 046103, 2007

Rui-Sheng Wang, Shihua Zhang, Yong Wang, Xiang-Sun Zhang, Luonan Chen. Clustering complex networks and biological networks by Nonnegative Matrix Factorization with various similarity measures. *Neurocomputing*, DOI: 10.1016/j.neucom.2007.12.043

衡量网络模块化的指标Q值

- Newman 和 Girvan (*Physical Review E*, 2004) 提出一种衡量网络社区结构的指标 Q 值

$$Q(P_k) = \sum_{c=1}^{k} \left[\frac{L(V_c, V_c)}{L(V, V)} - \left(\frac{L(V_c, V)}{L(V, V)} \right)^2 \right]$$

指标Q的问题 (Resolution limit)

Fortunato and Barthélemy, *PNAS*, 2007

- 利用Q划分网络的计算步骤:

  1. 固定要分成的模块数k, 将网络N分成k块N_1, …, N_k, 使$Q_k = Q(N_1) + \ldots + Q(N_k)$最大

  2. 对$k = 1, \ldots, n$, 求k^*使Q_{k^*}最大

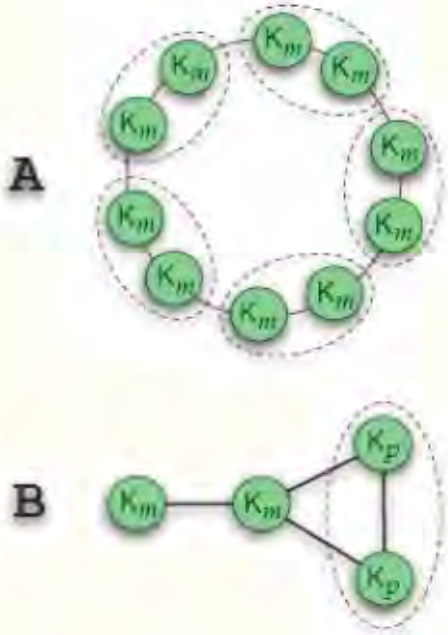
- 目前很大一部分模块探测的方法集中于利用各种启发式算法来极大化Q值，例如模拟退火、遗传算法等(Newman, *PNAS*, 2006; Guimera, Nature, 2005).

- Q值依赖于网络的规模，及网络的总边数.

- 无法正确识别一些明显的模块，例如一个完全子图

极端例子：ring of cliques

Problems, or not?



Modules indistingisable via
Optimization of modularity

$$l_S < 2l_R^{min} = \sqrt{2L}.$$

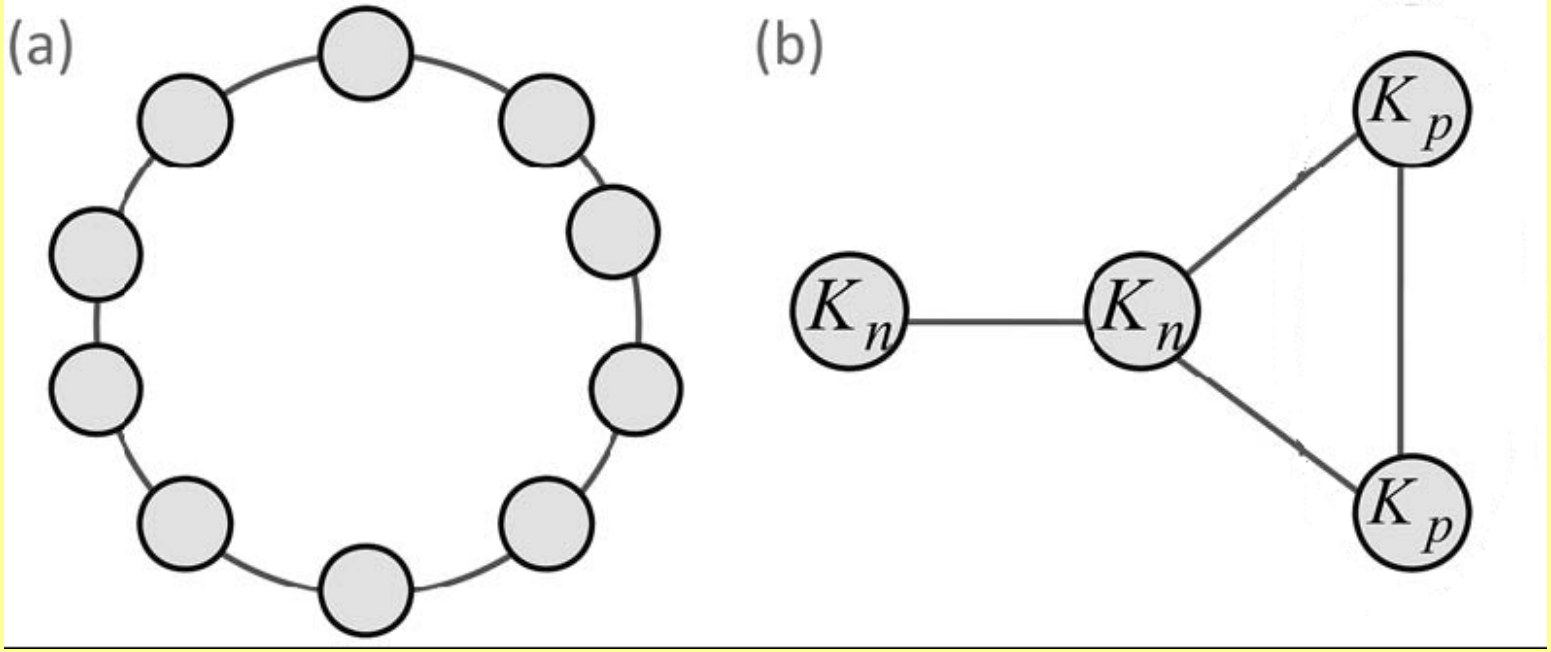
Fortunato & Barthelemy, Proc. Natl. Acad. Sci. USA 104 (1), 36-41 (2007)

提出新的模块化指标D值

- 模块化密度函数 D:

$$D(P_k) = \sum_{c=1}^{k} \frac{L(V_c, V_c) - L(V_c, \overline{V}_c)}{|V_c|}$$

Zhenping Li, Shihua Zhang, Rui-Sheng Wang, Xiang-Sun Zhang, Luonan Chen, Quantitative function for community detection. *Physical Review E*, 77, 036109, 2008



D值克服了Q值存在的 resolution limit 问题

结果

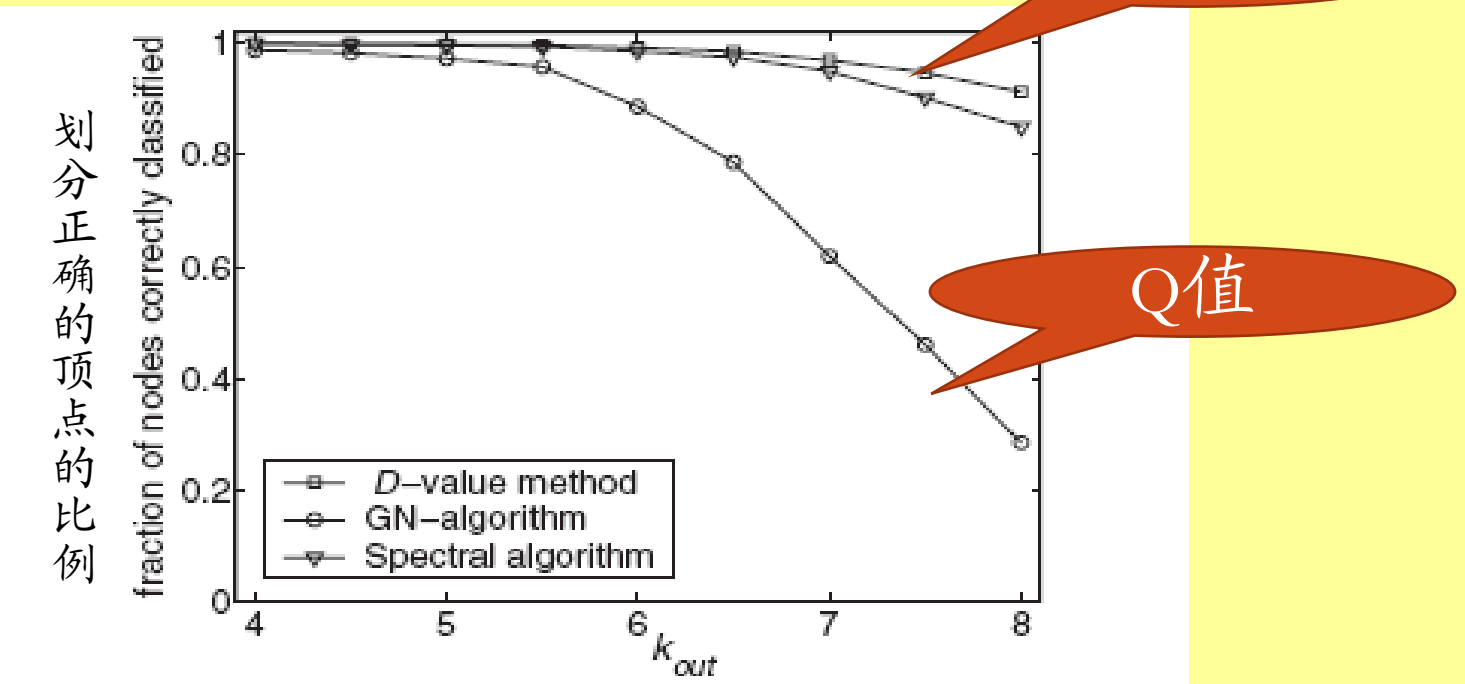


FIG. 2. Test of various methods on computer-generated networks with known community structures. It is a plot of the fraction of nodes correctly classified with respect to k_{out}. Each point is an average over 100 realizations of the networks.

工作二: 网络模块划分的优化模型建立

- Q 的整数规划模型

$$\begin{aligned} \max \quad & \sum_{j=1}^{k} \left[\frac{\sum_{s,t \in V} e_{st} x_{sj} x_{tj}}{\sum_{(s,t) \in E} e_{st}} - \left(\frac{\sum_{s,t \in V} e_{st} x_{sj}}{\sum_{(s,t) \in E} e_{st}} \right)^2 \right] \\ s.t. \quad & \sum_{j=1}^{k} x_{ij} = 1, \quad i = 1, \cdots, n \\ & x_{ij} \in \{0, 1\}, \ i = 1, \cdots, n, \ j = 1, \cdots, k \end{aligned}$$

Xiang-Sun Zhang and Rui-Sheng Wang, Optimization analysis of modularity measures for network community detection, *OSB 2008*.

工作二: 网络模块划分的优化模型建立

- D 的整数规划模型

$$\begin{aligned}
\max \quad & \sum_{j=1}^{k} \left[\frac{\sum_{s,t\in V} e_{st} x_{sj} x_{tj}}{\sum_{t\in V} x_{tj}} - \frac{\sum_{s,t\in V} e_{st} x_{sj} (1-x_{tj})}{\sum_{t\in V} x_{tj}} \right] \\
s.t. \quad & \sum_{j=1}^{k} x_{ij} = 1, \ i = 1, \cdots, n, \\
& x_{ij} \in \{0, \ 1\}, \ i = 1, \cdots, n, \ j = 1, \cdots, k
\end{aligned}$$

Xiang-Sun Zhang and Rui-Sheng Wang, Optimization analysis of modularity measures for network community detection, *OSB 2008*.

离散凸规划的引入

- Q值和D值的最优化模型都是非线性整数规划
- 目标函数的凸性和凹性无法解析得到
- 对两个具有特殊结构的网络进行分析
- 引入离散凸规划（变量是离散的，可以嵌入一个连续的凸规划）的概念进行分析,得到解析解

致谢

- 研究团队 **ZHANGroup**

张菊亮	北京交通大学
张继红	北京外国语大学
*李珍萍	北京物资学院
吴凌云	数学与系统科学研究院
*王勇	数学与系统科学研究院
*王瑞省	中国人民大学
张世华	数学与系统科学研究院
刘治平	大阪产业大学
金光旭	上海大学
张忠元	中央财经大学

博士后&博士生：杨志霞，徐守军，邱宇青，王林，王吉光，任仙文，曲积彬

- 资助：
  - 国家自然基金委员会，中国科技部，中国科学院基础局，数学与系统科学研究院，日本科学技术振兴协会

谢谢大家！

- 欢迎访问 ZHANGroup

http://zhangroup.aporc.org