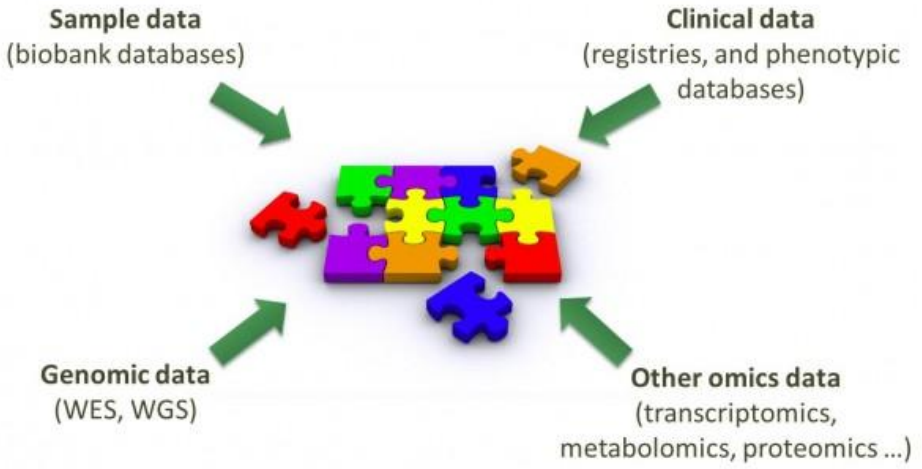
# 生物信息学：
# 组学时代的生物信息数据挖掘和理解

2020年秋



# 有关信息

- 授课教师：宁康，张礼斌，陈鹏
  - Email: ningkang@hust.edu.cn
  - Office: 华中科技大学东十一楼504室
  - Phone: 87793041, 18627968927

- 课程网页
  - http://www.microbioinformatics.org/Bioinformatics.html
  - QQ群：

# 课程安排

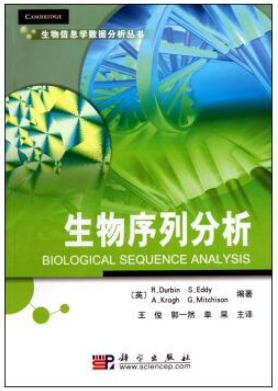
## （生物信息中的算法设计与概率统计模型）

- 生物背景和课程简介
- 生物信息学和生物数据挖掘
  - 生物数据的格式及其意义
    - 序列数据
    - 树状数据
    - 网络数据
    - 表达数据等
  - 生物数据库及其用法
  - 生物信息基本算法
    - 双序列联配
    - 多序列联配
    - 基因组组装算法
    - 基因预测和功能注释
    - 系统发育树构建
    - 蛋白质结构预测
    - 生物调控网络解析
  - 组学数据分析方法
    - 基因组变异分析
    - 基因表达和比较分析
    - 非编码RNA分析
    - 蛋白组分析
    - 宏基因组分析
  - 系统生物学与交叉科学
- 面向生物大数据挖掘的深度学习

研究对象：
生物序列，
进化树，
生物网络，
基因表达
…

方法：
生物计算与生物信息

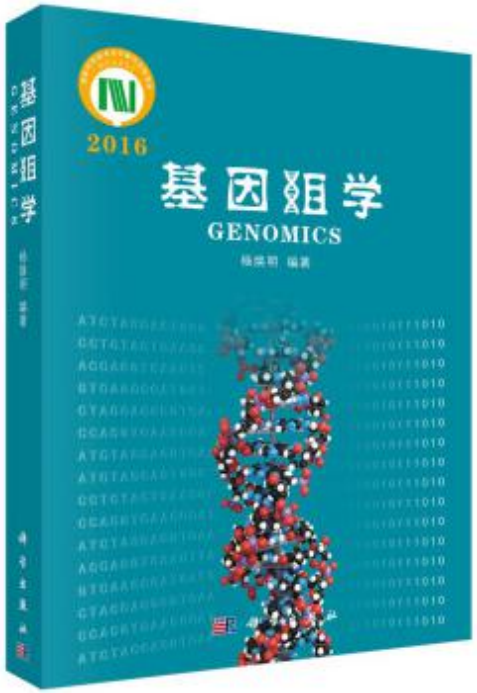
# 教材及参考书目

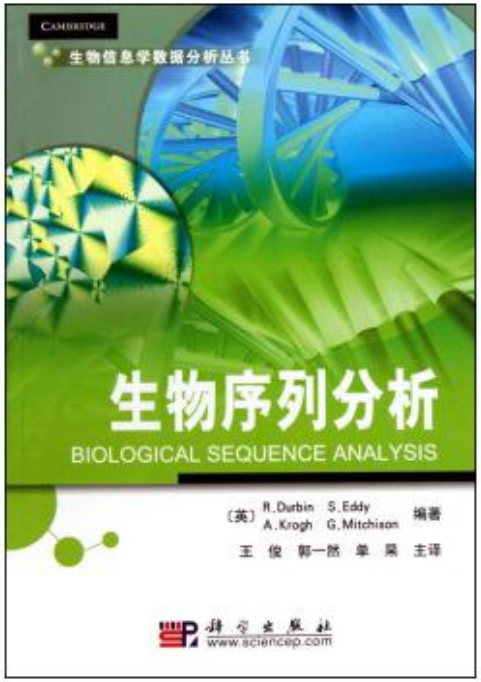


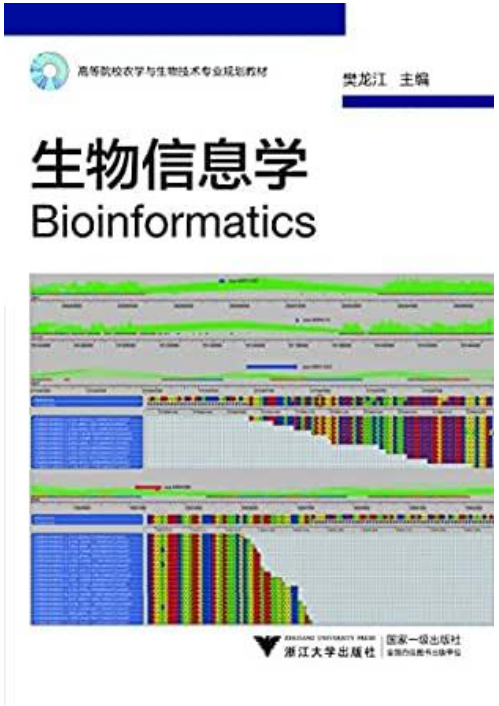

- **教学参考书：**
- 《生物序列分析》（第1版）.科学出版社. 2010年8月出版. R. Durbin等编著，王俊等主译.

- **课外文献阅读：**
- 《生物信息学》（第1版）. 浙江大学出版社. 2017年3月出版. 樊龙江主编.
- 《基因组学》（第1版）. 科学出版社. 2016年10月出版. 杨焕明主编.

# References


2016
基因组学
GENOMICS


CAMBRIDGE
生物信息学数据分析丛书
生物序列分析
BIOLOGICAL SEQUENCE ANALYSIS
R.Durbin S.Eddy
A.Krogh G.Mitchison
编著
王俊 郭一然 单渠 主译
www.sciencep.com


樊龙江 主编
生物信息学
Bioinformatics
浙江大学出版社

# Slides credits

——生物信息学研究方法概述：北京大学生物信息中心

——生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学

——神经网络与深度学习: 邱锡鹏@复旦大学

——Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Havard University

——Combinatorial Methods in  Computation Biology: Ken Sung Lab@NUS

——Deep Learning in the Life Sciences: MIT

# 第8-1章:蛋白质相互作用的实验方法和预测

- Experimental methods
- Prediction of protein-protein interactions

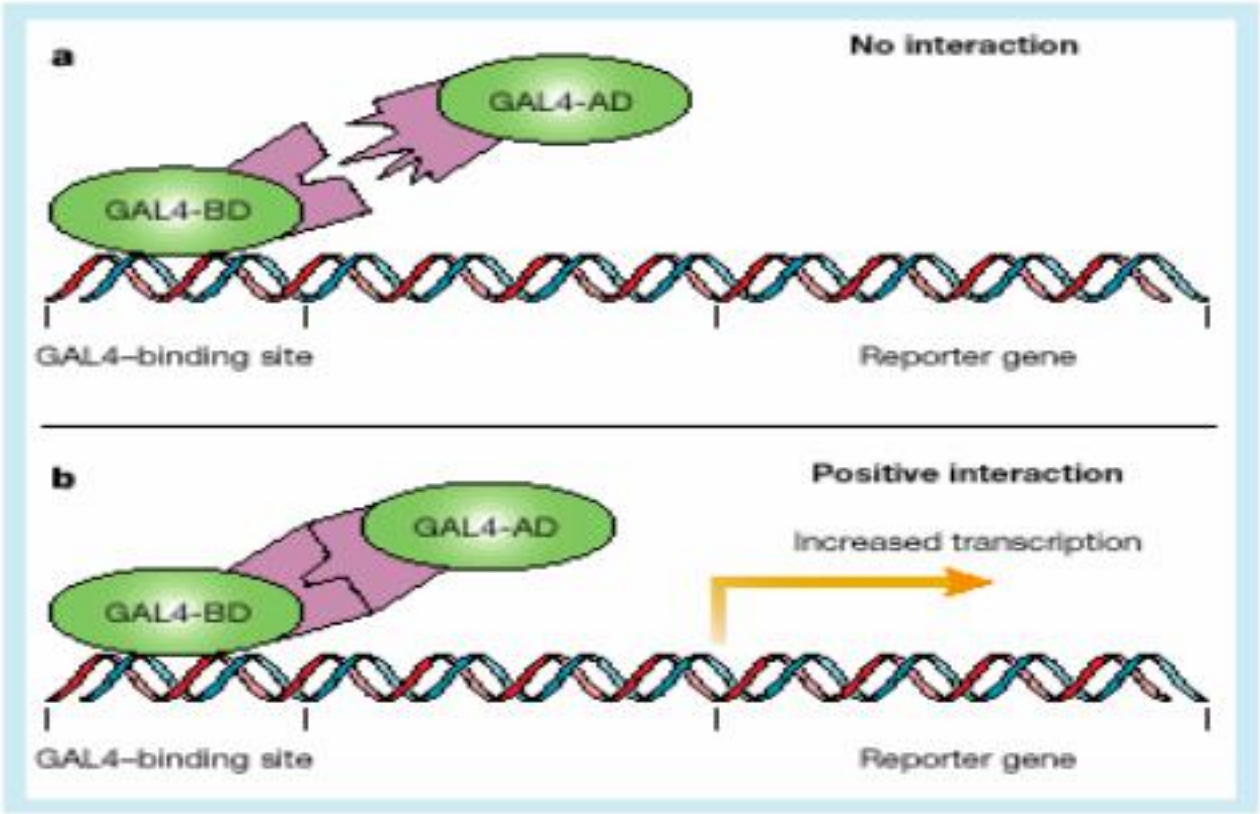
# Part I: Experimental Methods

- Physical interaction
  - Yeast two hybrid system
  - TAP-mass spectrometry
- Genetic interaction
  - SGA
  - EMAP

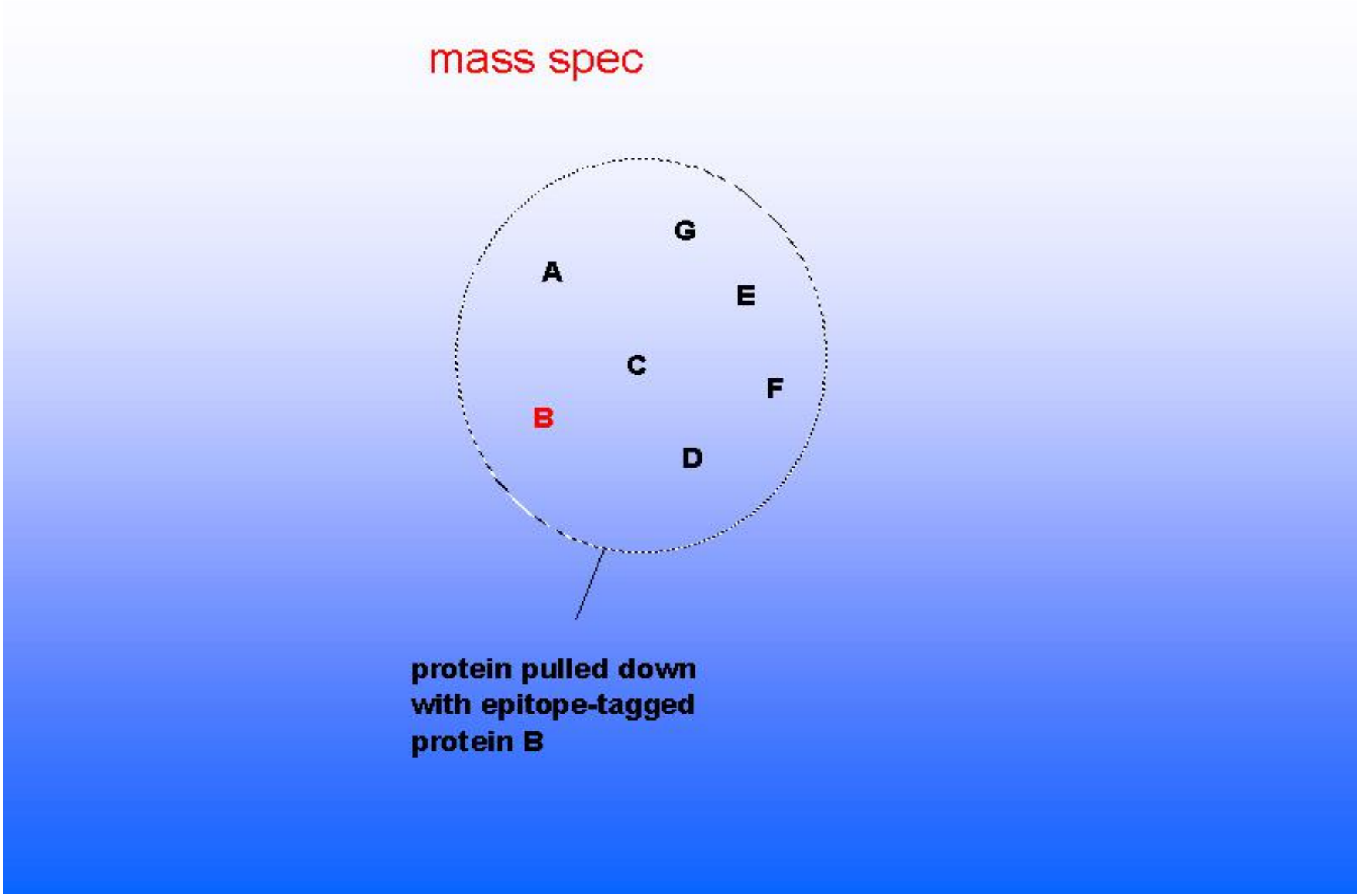
# Protein-protein interactions (Experimental methods)

- Co-immunoprecipitation.
- Two-hybrid system (Uetz et al. 2000, Ito et al. 2000, 2001).
- Purified Complex by mass spectrometry
  - TAP: Tandem affinity purification (Gavin et al. 2002).
  - HMS-PCI: high-throughput mass spectrometric protein complex identification (Ho et al. 2002).

# Mechanism of two-hybrid system




From: Nature 405, June 15, 2000, 837-846.




Gavin et al. (2002) Nature 415:141

# Mass spectrometry

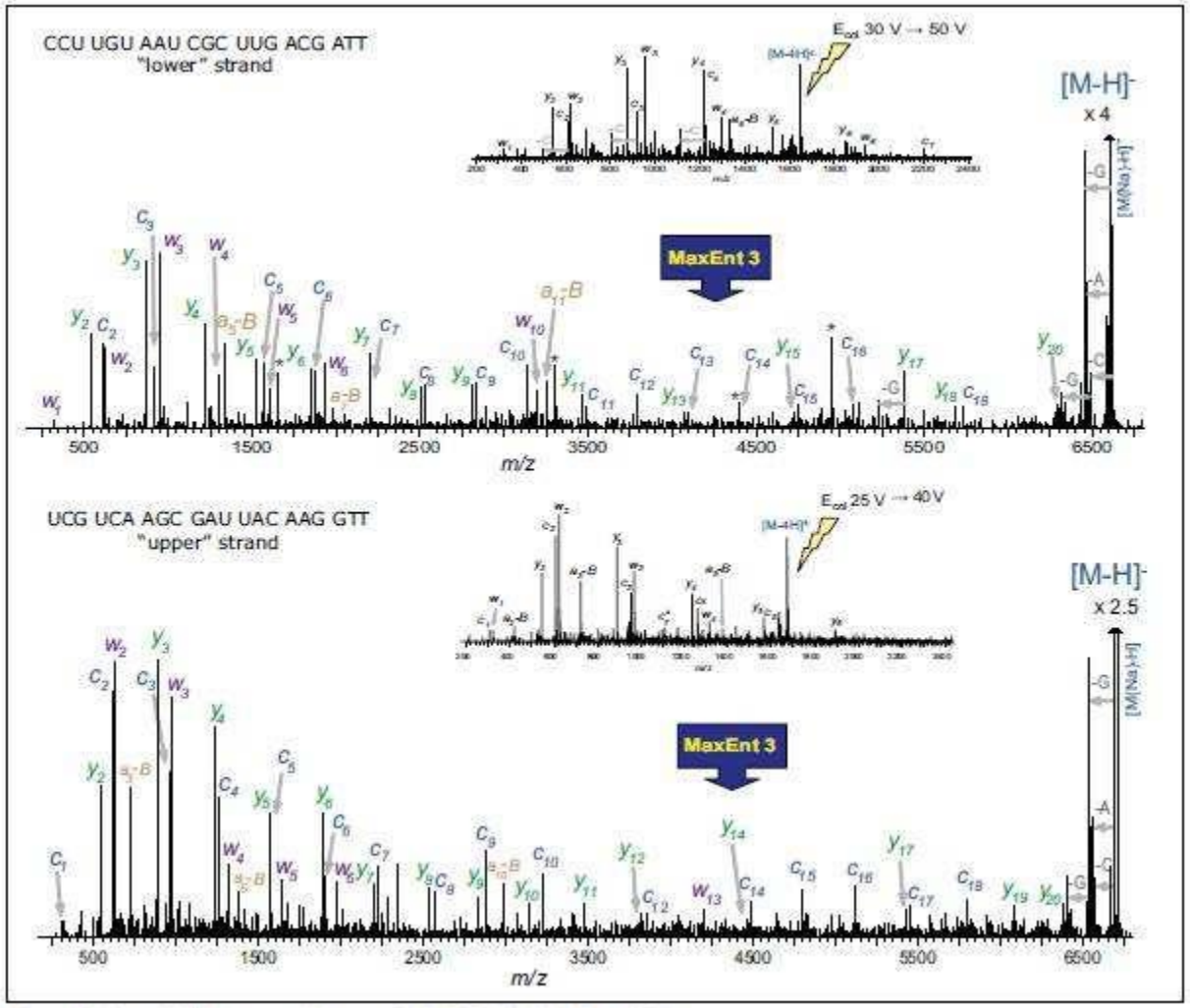
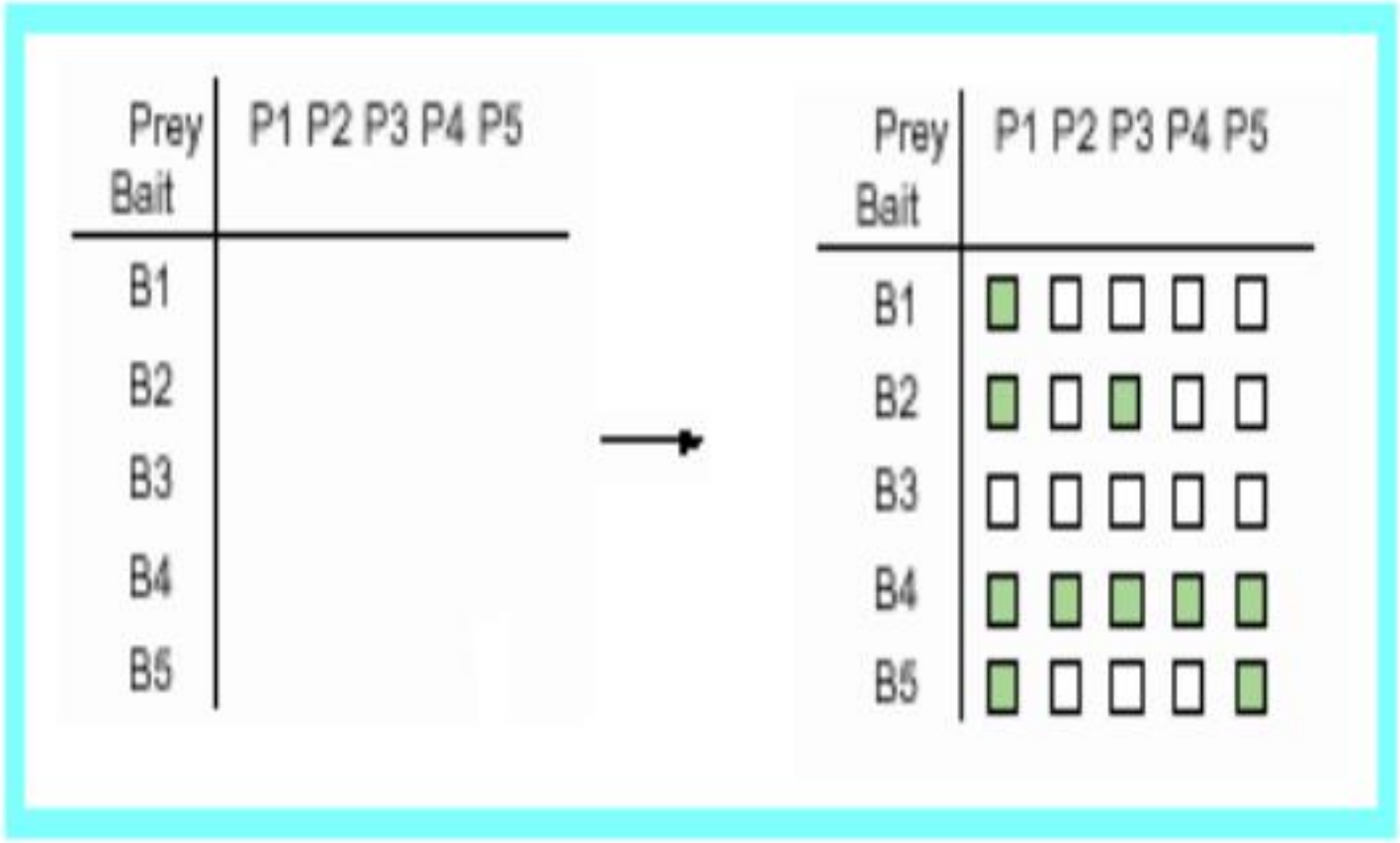


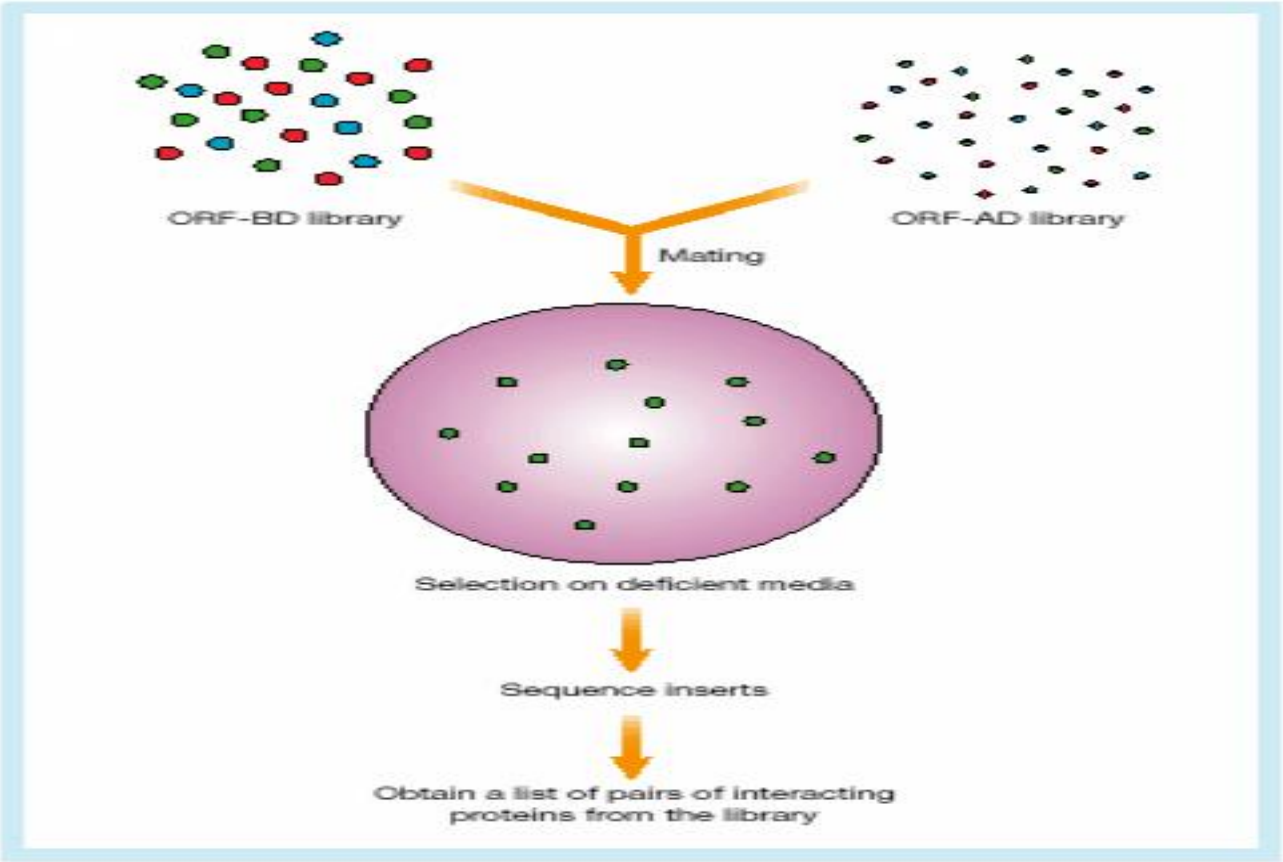

图2 21 核苷酸RNA寡核苷酸的MaxEnt3去卷积MS/MS图谱. 星号为谱波峰(去卷积的结果)

# Matrix method (two hybrid)




From: TRENDS in Genetics Vol.17, No.6, June 2001.

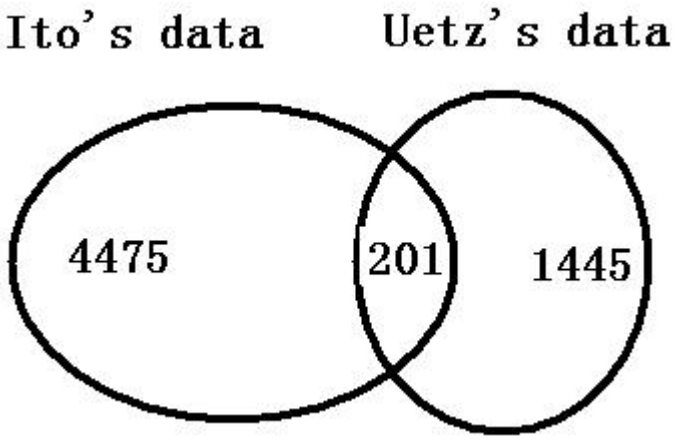
# Interaction Sequence Tags (ISTs)




From: Nature 405, June 15, 2000, 837-846.

# Two data sets from yeast two hybrid system

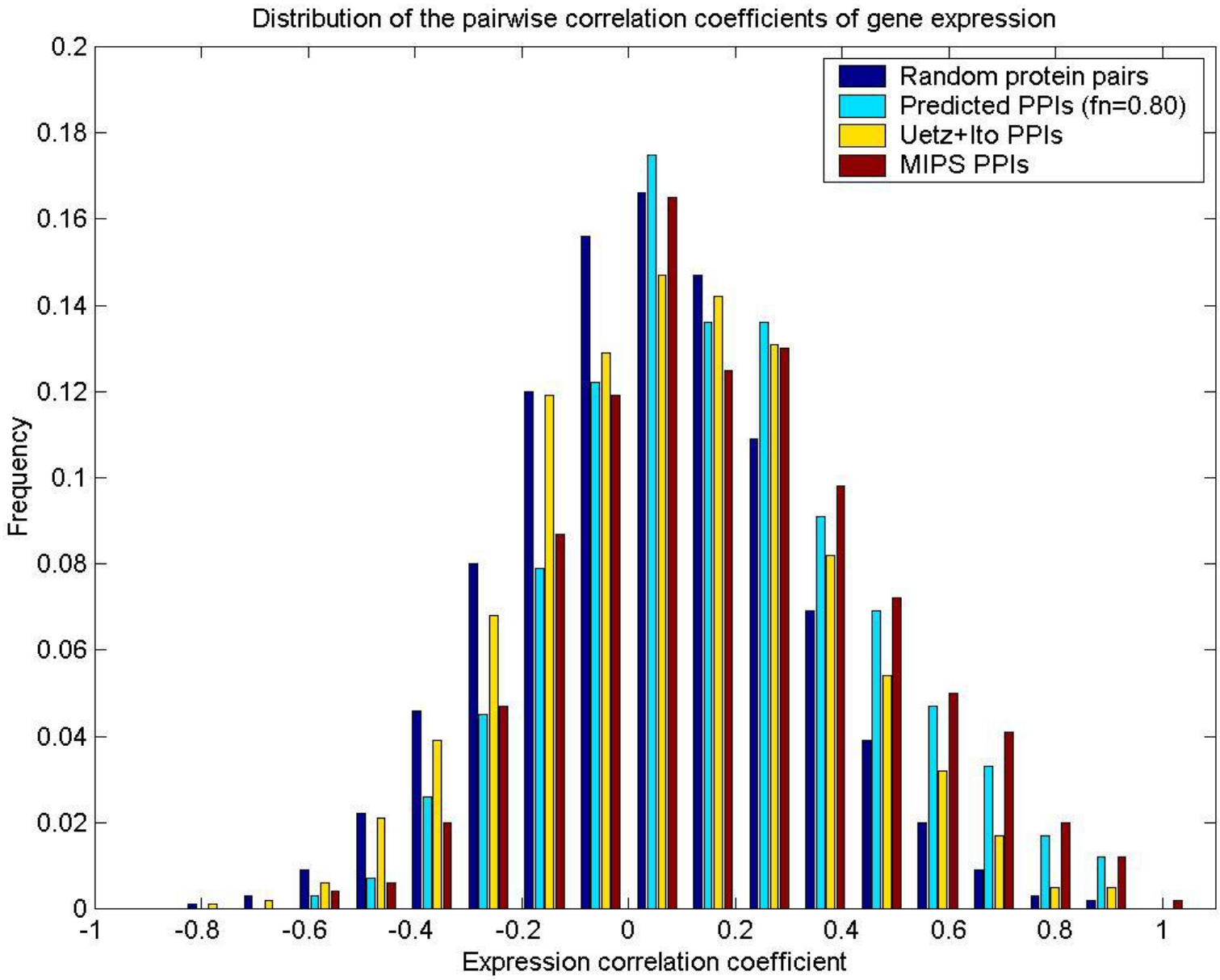
- Uetz's data (Uetz et al. 2000).
- Ito's data (Ito et al. 2000, 2001).

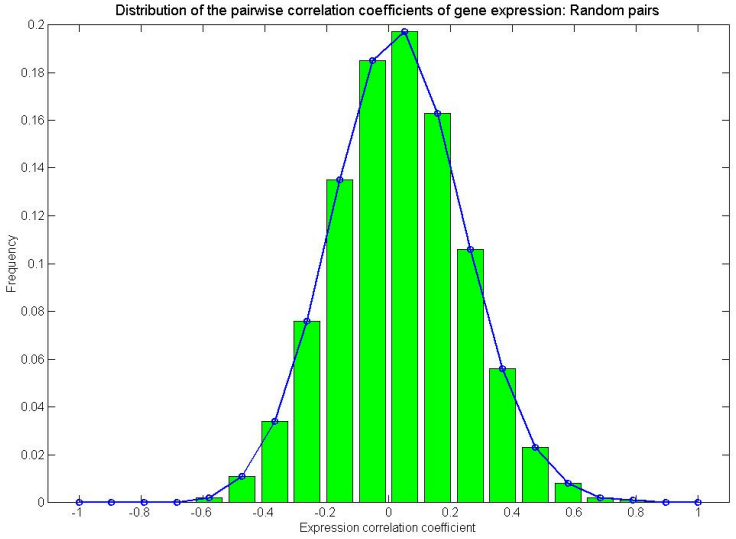


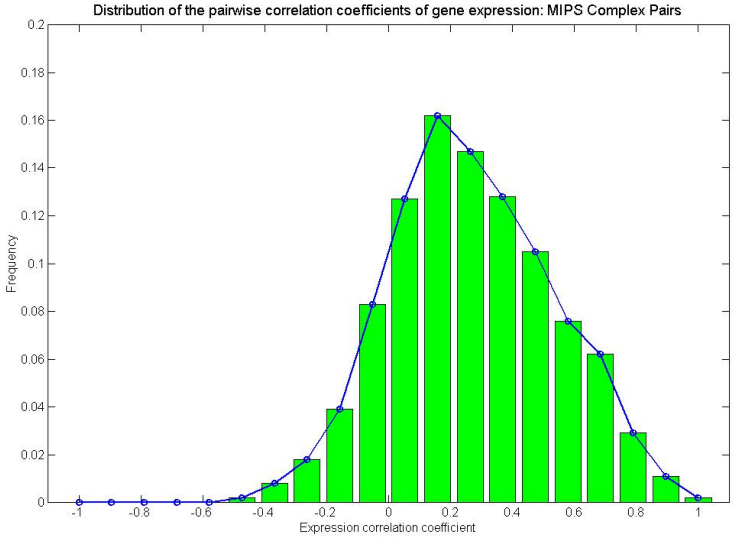
# Possible Errors in 2-hybrid system

- False positive.
  - Possible mutation during PCR-amplifying.
  - Stochastic activation of reporter gene.
- False negative.
  - Membrane protein, post-translational modification protein, those self-activating reporter genes (Removed in experiment).
  - Weak interactions.

The size of interactome for yeast (5-50/protein)


Distribution of the pairwise correlation coefficients of gene expression
Random protein pairs
Predicted PPIs (fn=0.80)
Uetz+Ito PPIs
MIPS PPIs
Frequency
Expression correlation coefficient
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1



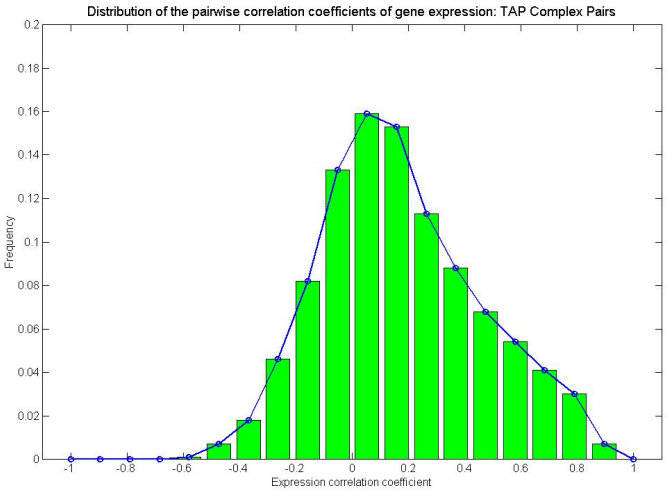




Non-interaction

$1 - \alpha$

$\alpha$

Real interaction




Observed interaction data

# MLE of the reliability

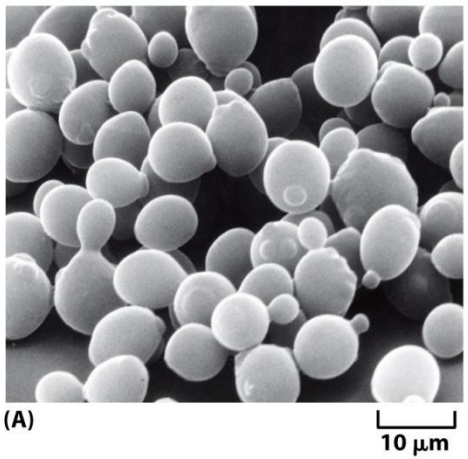
- Likelihood function

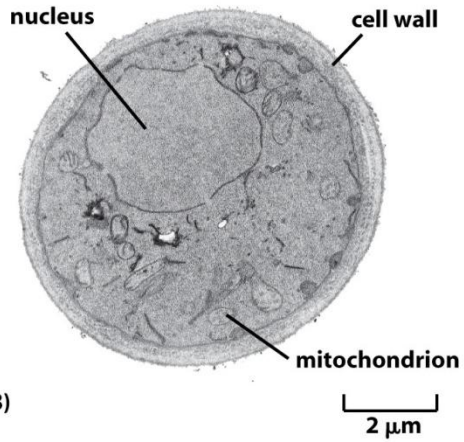
$$L(\alpha) = \prod_{k=1}^{K} (\alpha p_k + (1-\alpha) q_k)^{n_k}$$

- Precision of the estimation

$$Var(\widehat{\alpha}) = \frac{1}{\sum_{k=1}^{K} n_k \frac{(p_k - q_k)^2}{(\widehat{\alpha} p_k + (1-\widehat{\alpha}) q_k)^2}}$$

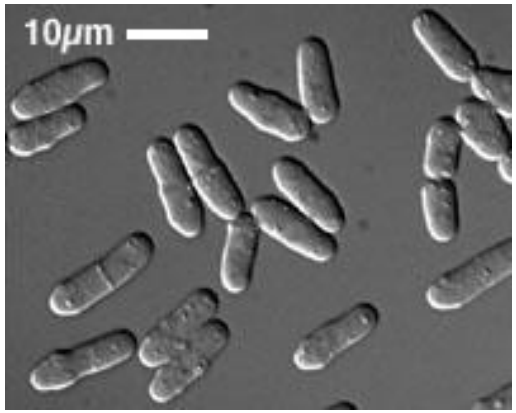
# Budding Yeast
## *Saccharomyces Cerevisiae*

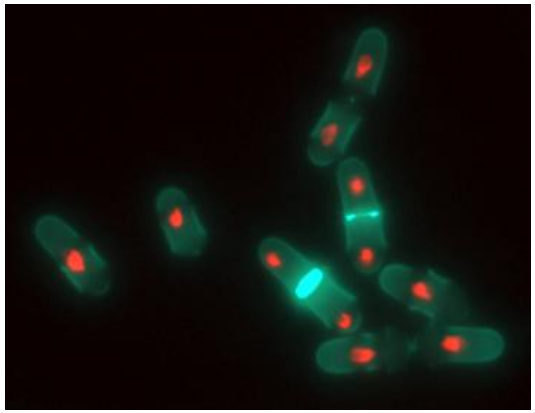


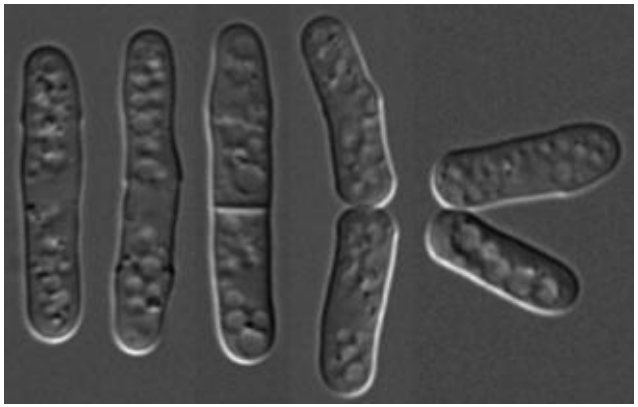




- a and α mating type,  cell cycle
- 6300 genes  (1997)
- Genome-wide single mutants analysis (2000~)

# Fission yeast
# *Schizosaccharomyces Pombe*





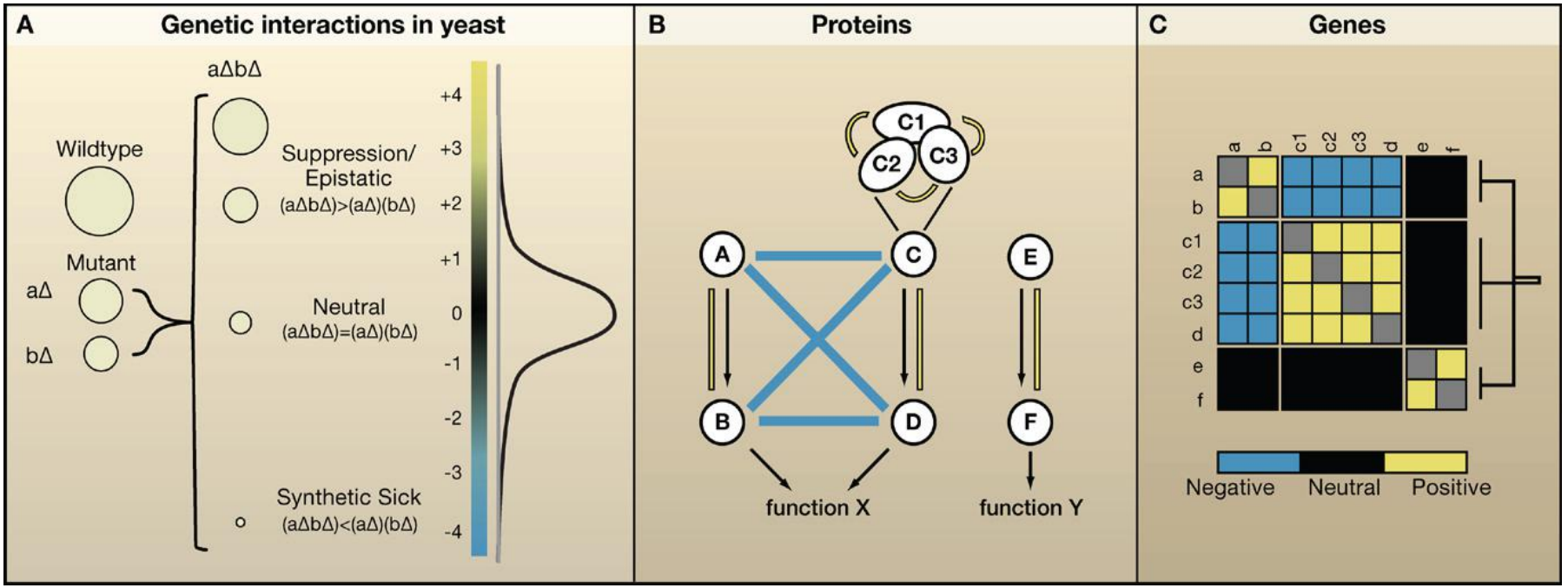
- 1000 million years separation from budding yeast;
- 13.8 Mb genome size, 4824 genes (open reading frames, OPF);
- 3 chromosomes, no genome-wide duplications; h+ and h- mating types;
- Cell cycle: 10% G1, 10% S, 70% G2 and 10% M phases.
- Genome-wide single mutants analysis (2010~)

more similar to metazoans than *S. cerevisiae*

- *cell cycle* regulation in G2/ M phase,
- gene regulation by the RNAi pathway
- the widespread presence of introns in genes

# What's Genetic Interaction

- Genetic interactions between two loci can be mapped by measuring how the phenotype of an organism lacking both genes (double mutant) differs from that expected when the phenotypes of the single mutations are combined
- Null model: F(ΔAB)= F(ΔA)* F(ΔB)



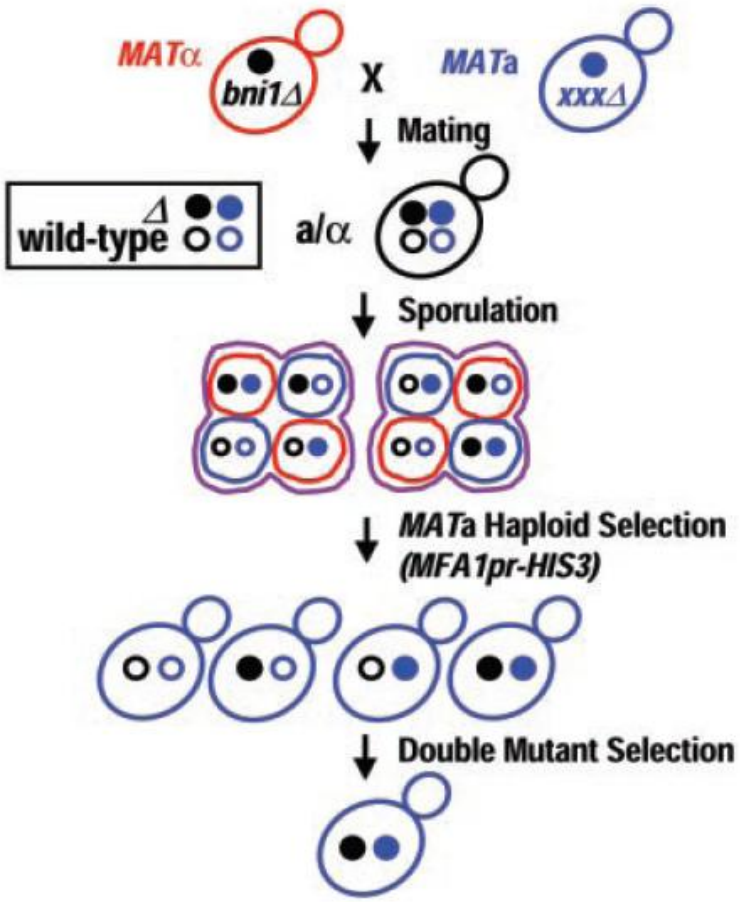

Beltrao et al. Cell 141: 739-745, 2010.

# Identification of Genetic Interactions

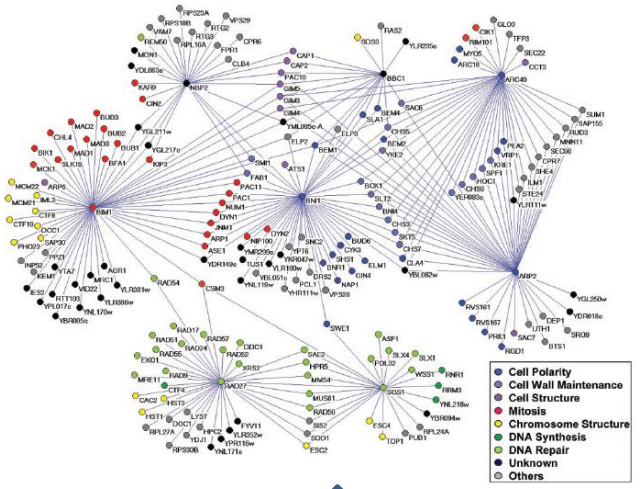
- Synthetic Gene Array (SGA) (Tong, et al. 2001)
- Diploid based Synthetic Lethality Analysis on Microarrays (dSLAM) (Pan, X., et al. 2004)

# Synthetic Gene Array (SGA)

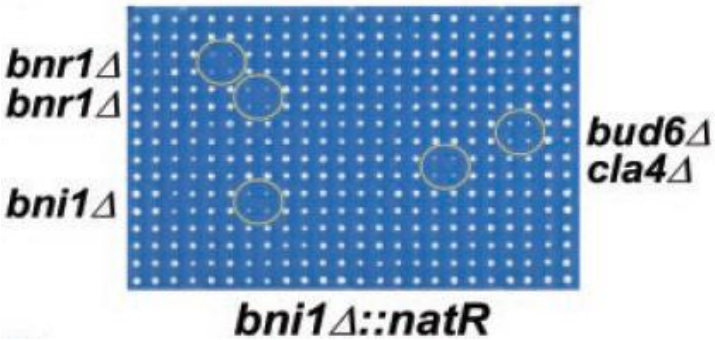
Synthetic genetic array methodology




Genetic Interaction Network




Final Screen




Amy Hin Yan Tong, et al. *Science,* 2001.

# EMAP is the Extension of SGA

- EMAP: Epistatic Miniarray Profiles (Maya Schuldiner, et al. 2005. *Cell*)
- Quantitative  measurement of phenotype (colony size)
  - Measure  both positive and negative interactions.
  - Genome-wide (DAMP for lethal genes ).

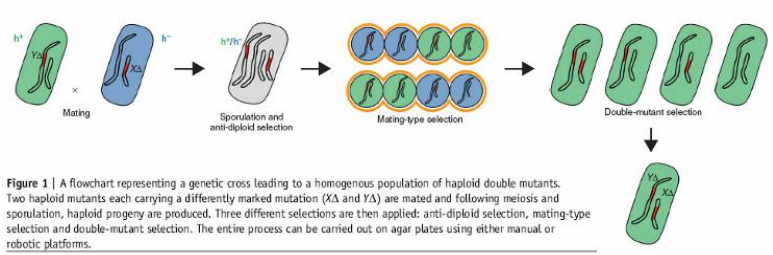
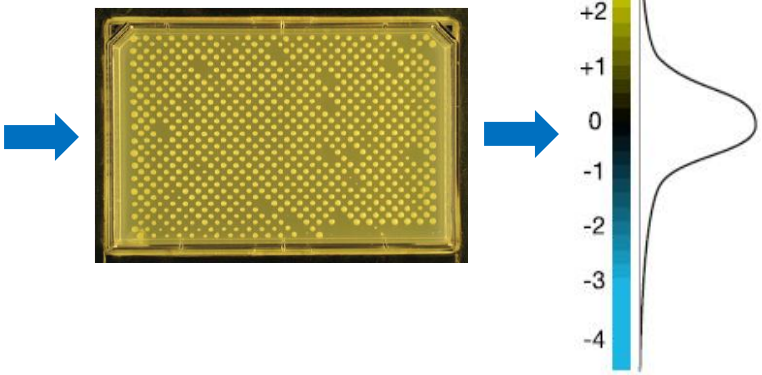



**Figure 1** | A flowchart representing a genetic cross leading to a homogenous population of haploid double mutants. Two haploid mutants each carrying a differently marked mutation (XΔ and YΔ) are mated and following meiosis and sporulation, haploid progeny are produced. Three different selections are then applied: anti-diploid selection, mating-type selection and double-mutant selection. The entire process can be carried out on agar plates using either manual or robotic platforms.



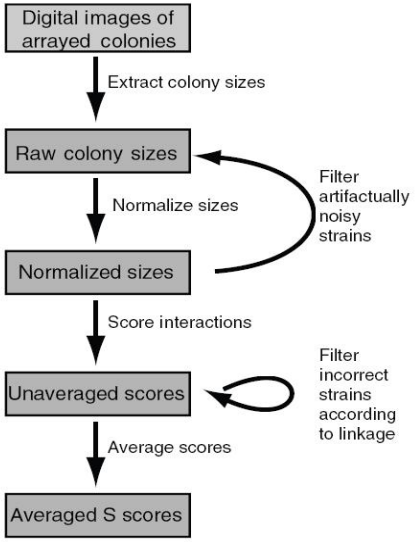

Dorothea Fiedler, et al. *Cell, 2009*

# EMAP S-score

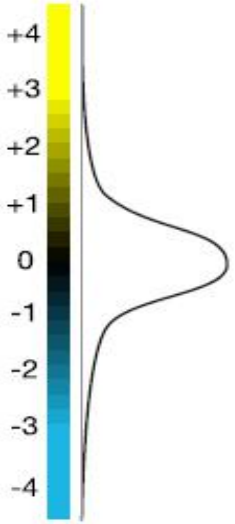
- Quantitative measure: $\epsilon = W_{ab} - W_a W_b,\ W_a = w/w_{wild}.$

| No interaction | Synthetic sick/Lethality | Synthetic alleviating |
|---|---|---|
| $\epsilon = 0$ | $\epsilon < 0$ | $\epsilon > 0$ |

  - T-Test with null hypothesis $\epsilon = 0$




$$S_0 = \frac{\mu_{ab} - w_a \mu_b}{\sqrt{n_1 S_{ab}^2 + n_2 w_a^2 S_b^2}}$$




Sean R Collins, et al. *Genome Biology 7: R63, 2006.*

# PPI databases

- MIPS: Munich Information center for Protein Sequences (http://mips.gsf.de)
- DIP: Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu)
- BIND: Biomolecular Interaction Network Database (http://www.bind.ca)
- GRID: General Repository for Interaction Datasets (http://biodata.mshri.on.ca/grid)
- MINT: Molecular Interaction Database (http://cbm.bio.uniroma2.it/mint/)

# Further Reading

- For more experimental methods and databases, please read the following review paper
  - Shoemaker BA, Panchenko AR (2007) Deciphering Protein–Protein Interactions. Part I. Experimental Techniques and Databases. PLoS Comput Biol 3(3): e42. doi:10.1371/journal.pcbi.0030042.

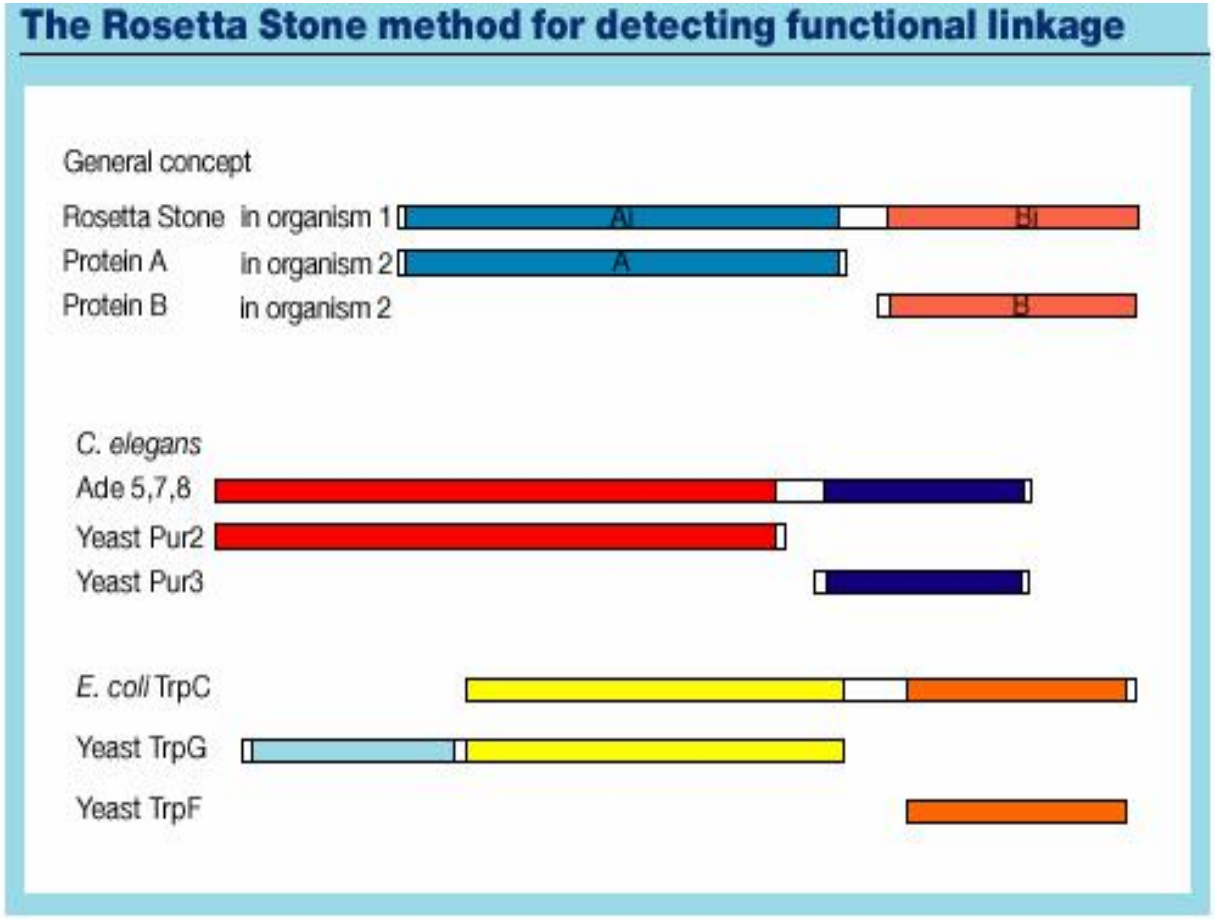
# Protein-protein interactions (Computational Methods)

- Gene fusion method (A.Enright 1999. E.Maccote 1999)
- Phylogenetic profile method (M.Pellegrini 1999, D.Eisenberg, 1999).
- Gene cluster method (R.Overbeek, 1999).
- Highly co-expressed gene pairs.

# Part II: Predicting Protein-protein Interactions

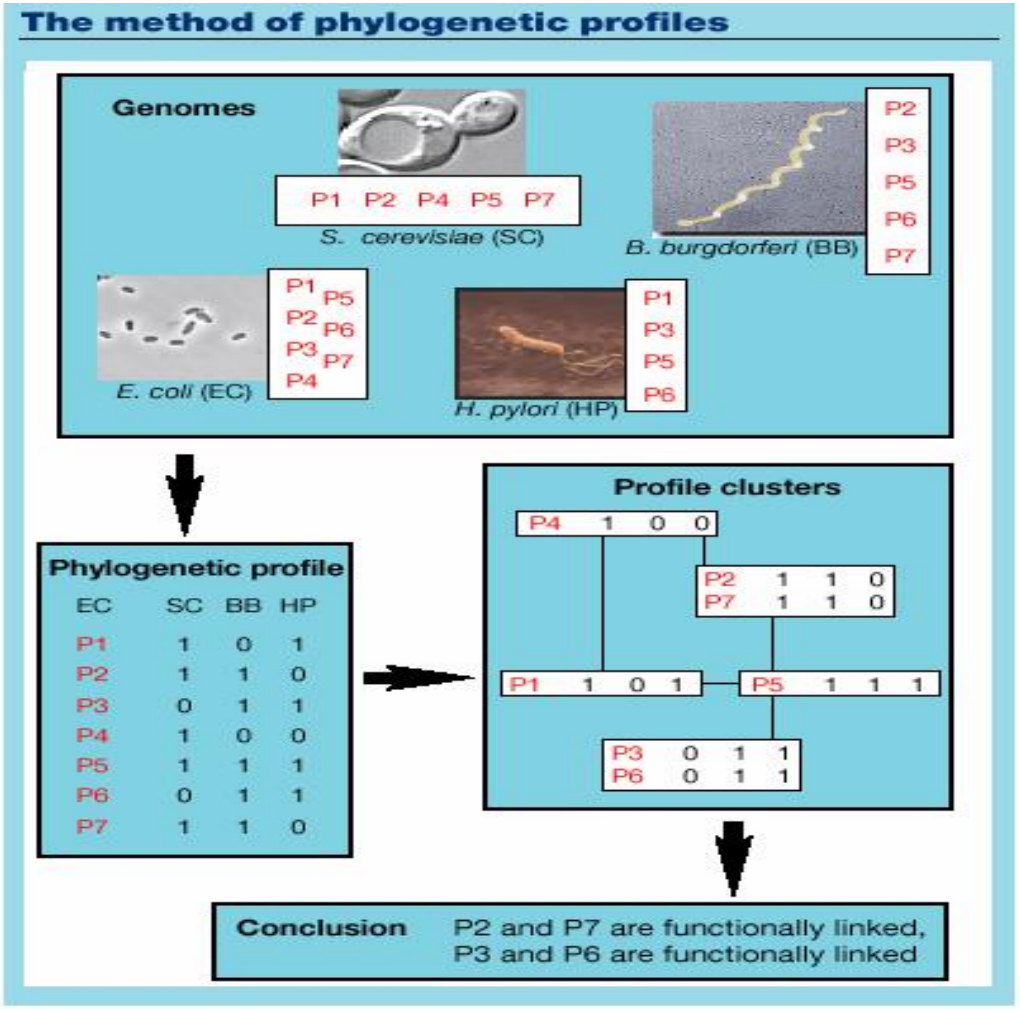
- Some computational methods
- Predicting protein-protein interaction from domains
  - Association method
  - MLE method

# Rosetta Stone Method



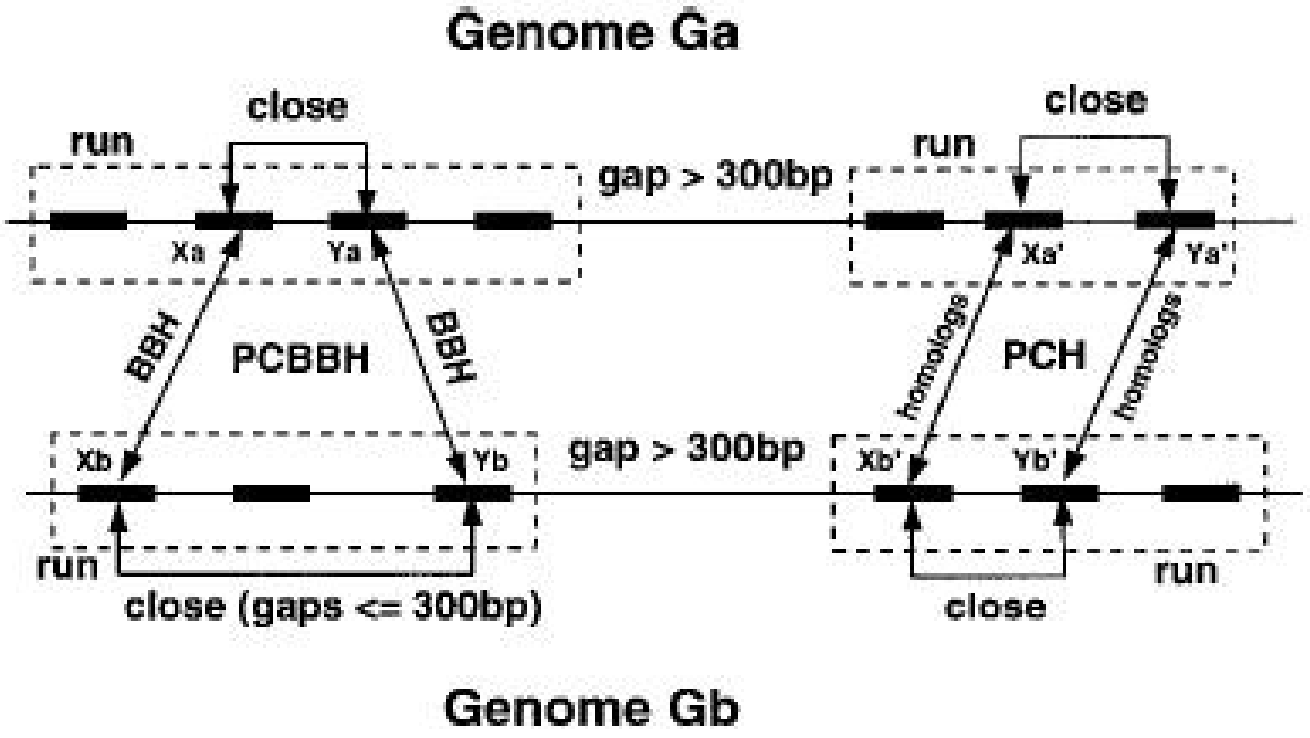

From: Nature Vol. 405, June.15, 2000, 823-826

# Phylogenetic Profiles Method



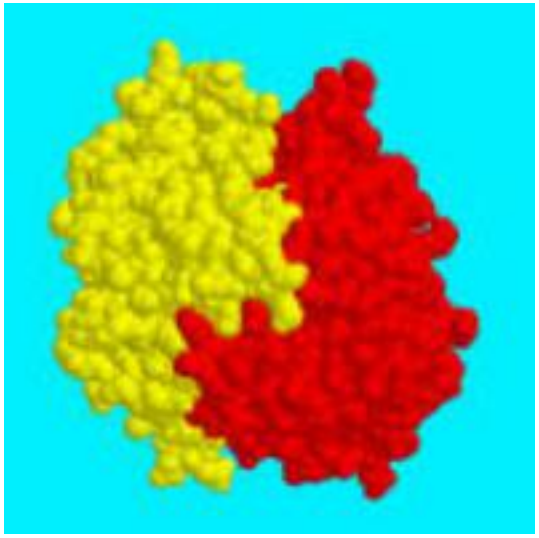

From: Nature Vol. 405, June, 15, 2000, 823-826

# Using Gene Clusters to Infer Functional Coupling




From: R.Overbeek, PNAS 96, 2896-2901, 1999.

# Structure of Proteins



# Predicting PPIs from Domains

- Domains are treated as elementary unit of function.
- Domains are responsible for the generation of interactions.
- Understanding protein-protein interaction at the domain level.

# Domain Databases

- Pfam, domain classification by HMM.
- Prodom.
- PRINTS, fingerprint information of protein sequences.
- SMART, mobile domain.
- BLOCKs, multiple alignment blocks.
- Interpro.

# SwissPfam entry for PIG1_BOVIN

## Description from Swissprot for PIG1_BOVIN :

1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase gamma 1(ec 3.1.4.11) (plc-gamma-1) (phospholipase c-gamma-1) (plc-ii)(plc-148)


PH 33-142
PI-PLC-X 321-465
SH2 550-639
SH2 668-741
SH3 794-849
PH 864-931
PI-PLC-Y 952-1070
C2 1090-1177

**Key**

signal peptide: > pfamA: > Context: > smart: > transmembrane: > low complexity: > coiled coil: > pfamB:

| Source | Domain | Start | End |
|---|---|---|---|
| Pfam | PH | 33 | 142 |
| Pfam | PI-PLC-X | 321 | 465 |

**Overlapping Domains:** Change the domain order using the ^ and v buttons. View the changes by clicking the 'Change order' button.

high priority

# Piwi

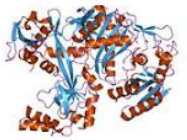
Edit Wikipedia article

**Piwi** (or **PIWI**) genes were identified as regulatory proteins responsible for stem cell and germ cell differentiation.[4] Piwi is an abbreviation of P-element Induced WImpy testis in Drosophila.[5] Piwi proteins are highly conserved RNA-binding proteins and are present in both plants and animals.[6] Piwi proteins belong to the Argonaute/Piwi family and have been classified as nuclear proteins. Studies on Drosophila have also indicated that Piwi proteins have slicer activity conferred by the presence of the Piwi domain.[7] In addition, Piwi associates with Heterochromatin protein 1, an epigenetic modifier, and piRNA-complementary sequences. These are indications of the role Piwi plays in epigenetic regulation. Piwi proteins are also thought to control the biogenesis of piRNA as many Piwi-like proteins contain slicer activity which would allow Piwi proteins to process precursor piRNA into mature piRNA.

**Contents** [hide]



**Piwi domain**



Structure of the Pyrococcus furiosus Argonaute protein.[1]

| Identifiers | |
|---|---|
| **Symbol** | Piwi |
| **Pfam** | PF02171 |
| **InterPro** | IPR003165 |
| **PROSITE** | PS50822 |
| **CDD** | cd02826 |
| **Available protein structures:** | [show] |

## Protein structure and function

The structure of several Piwi and Argonaute proteins (Ago) have been solved. Piwi proteins are RNA-binding proteins with 2 or 3 **domains**: The N-terminal **PAZ domain** binds the 3'-end of the guide RNA; the middle **MID domain** binds the 5'-phosphate of RNA; and the C-terminal **PIWI domain** acts as an RNase H endonuclease that can cleave RNA.[8][9] The small RNA partners of Ago proteins are microRNAs (miRNAs). Ago proteins utilize miRNAs to silence genes post-transcriptionally or use small-interfering RNAs (siRNAs) in both transcription and post-transcription silencing mechanisms. Piwi proteins interact with piRNAs (28–33 nucleotides) that are longer than miRNAs and siRNAs (~20 nucleotides), suggesting that their functions are distinct from those of Ago proteins.[8]

The piwi domain of an argonaute protein with bound siRNA, components of the RNA-induced silencing complex that mediates gene silencing by RNA interference.

## Human Piwi proteins

Presently there are four known human Piwi proteins—PIWI-like protein 1, PIWI-like protein 2, PIWI-like protein 3 and PIWI-like protein 4. Human Piwi proteins all contain two RNA binding domains, PAZ and Piwi. The four PIWI-like proteins have a spacious binding site within the PAZ domain which allows them to bind the bulky 2′-OCH3 at the 3′ end of piwi-interacting RNA.[10]

One of the major human homologues, whose upregulation is implicated in the formation of tumours such as seminomas, is called hiwi (for human piwi).[11]

Homologous proteins in mice have been called miwi (for mouse piwi).[12]

## Role in germline cells

PIWI proteins play a crucial role in fertility and germline development across animals and ciliates. Recently identified as a polar granule component, PIWI proteins appear to control germ cell formation so much so that in the absence of PIWI proteins there is a significant decrease in germ cell formation. Similar observations were made with the mouse homologs of PIWI, MILI, MIWI and MIWI2. These homologs are known to be present in spermatogenesis. Miwi is expressed in various stages of spermatocyte formation and spermatid elongation where Miwi2 is expressed in Sertoli cells. Mice deficient in either Mili or Miwi-2 have experienced spermatogenic stem cell arrest and those lacking Miwi-2 underwent a degradation of spermatogonia.[13] The effects of piwi proteins in human and mouse germlines seems to stem from their involvement in translation control as Piwi and the small noncoding RNA, piwi-interacting RNA (piRNA), have been known to co-fractionate polysomes. The piwi-piRNA pathway also induces heterochromatin formation at centromeres,[14] thus affecting transcription. The piwi-piRNA pathway also appears to protect the genome. First observed in Drosophila, mutant piwi-piRNA pathways led to a direct increase in dsDNA breaks in ovarian germ cells. The role of the piwi-piRNA pathway in transposon silencing may be responsible for the reduction in dsDNA breaks in germ cells.

## Role in RNA interference

The piwi domain[15] is a protein domain found in piwi proteins and a large number of related nucleic acid-binding proteins, especially those that bind and cleave RNA. The function of the domain is double stranded-RNA-guided hydrolysis of single stranded-RNA that has been determined in the argonaute family of related proteins.[1] Argonautes, the most well-studied family of nucleic-acid binding proteins, are RNase H-like enzymes that carry out the catalytic functions of the RNA-induced silencing complex (RISC). In the well-known cellular process of RNA interference, the argonaute protein in the RISC complex can bind both small interfering RNA (siRNA) generated from exogenous double-stranded RNA and microRNA (miRNA)

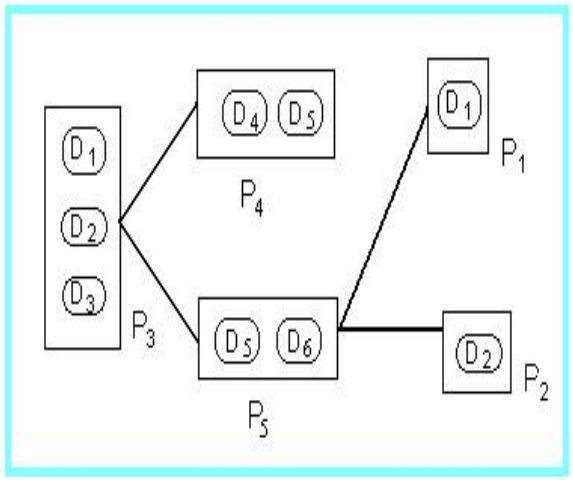

All human Piwi proteins and argonaute proteins have the same RNA binding domains, PAZ and Piwi.[2]

# Association-A simple method

$$V(D_{ij}) = \frac{\#\{\text{Interacted protein pairs contain } D_{ij}\}}{\#\{\text{All protein pairs contain } D_{ij}\}}$$

More observed PPIs for one domain pair will give higher probability of interaction for that domain pair.

# Simple Example




By assoiation method:

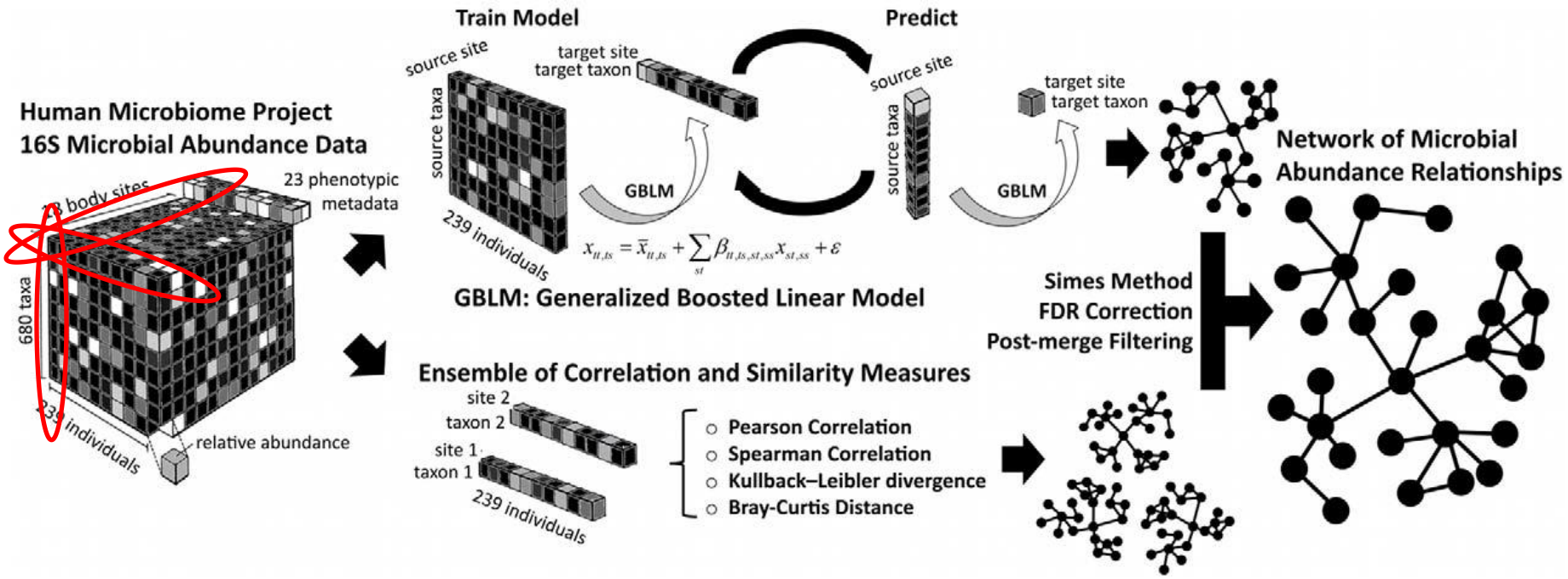
$$D_{34} = D_{35} = D_{36} = D_{26} = D_{16} = 1.0$$
$$D_{15} = D_{25} = 0.75, D_{14} = D_{24} = 0.5$$

Others are 0.0.

$D_{15}$: {P34, P35, P15}/{$P_{34}$, $P_{35}$, $P_{15}$, $P_{14}$}=0.75

# More complicated example

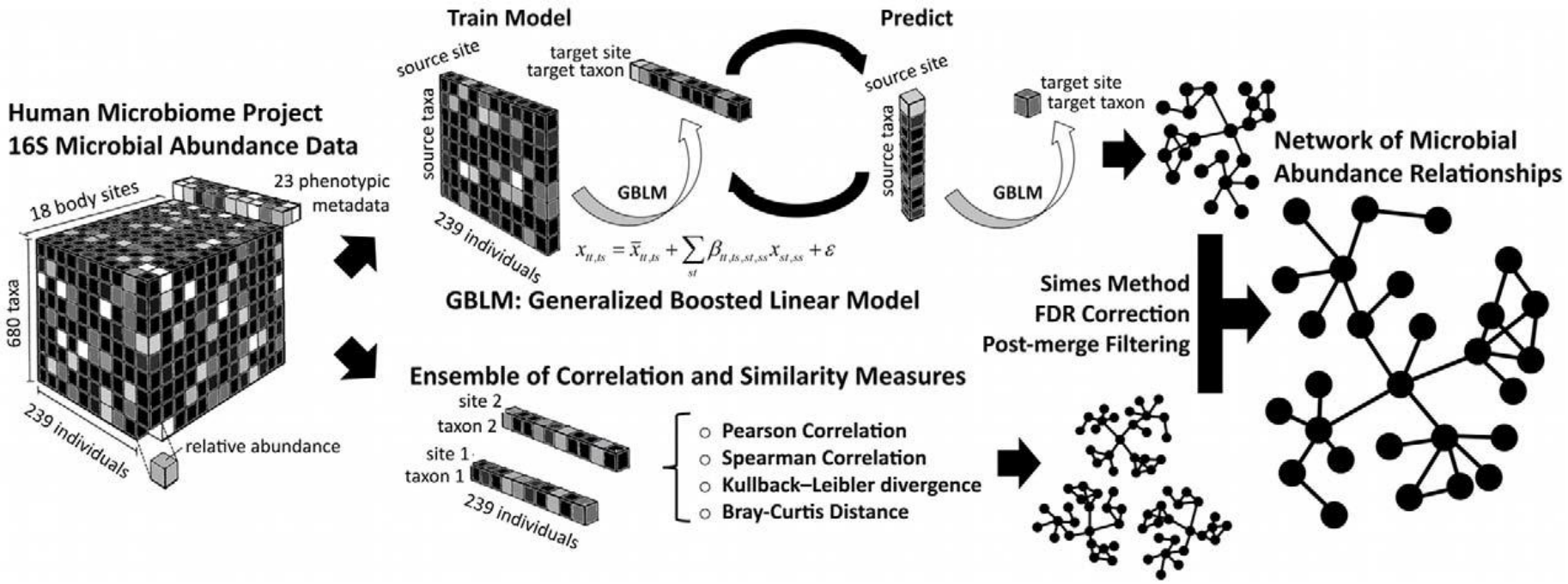
**Generalized Boosted Linear Models (GBLM)：广义线性模型**




Huttenhower, et al., *PLoS Computational Biology*, 2013

# More complicated example

**Generalized Boosted Linear Models (GBLM)：广义线性模型**




**Ensemble scoring**

Huttenhower, et al., *PLoS Computational Biology*, 2013

# More complicated example

**Generalized Boosted Linear Models (GBLM)：广义线性模型**


$$x_{tt,ts} = \bar{x}_{tt,ts} + \sum_{st} \beta_{tt,ts,st,ss} x_{st,ss}$$

$$\text{logit}(x_{tt,ts}) = \bar{x}_{tt,ts} + \sum_{st} \beta_{tt,ts,st,ss} x_{st,ss}$$

Huttenhower, et al., *PLoS Computational Biology*, 2013

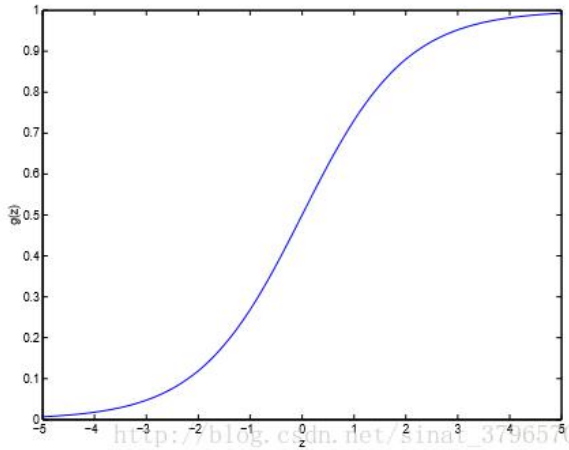
# More complicated example

**Generalized Linear Models (GLM)：广义线性模型**

广义线性模型的核心体现在：

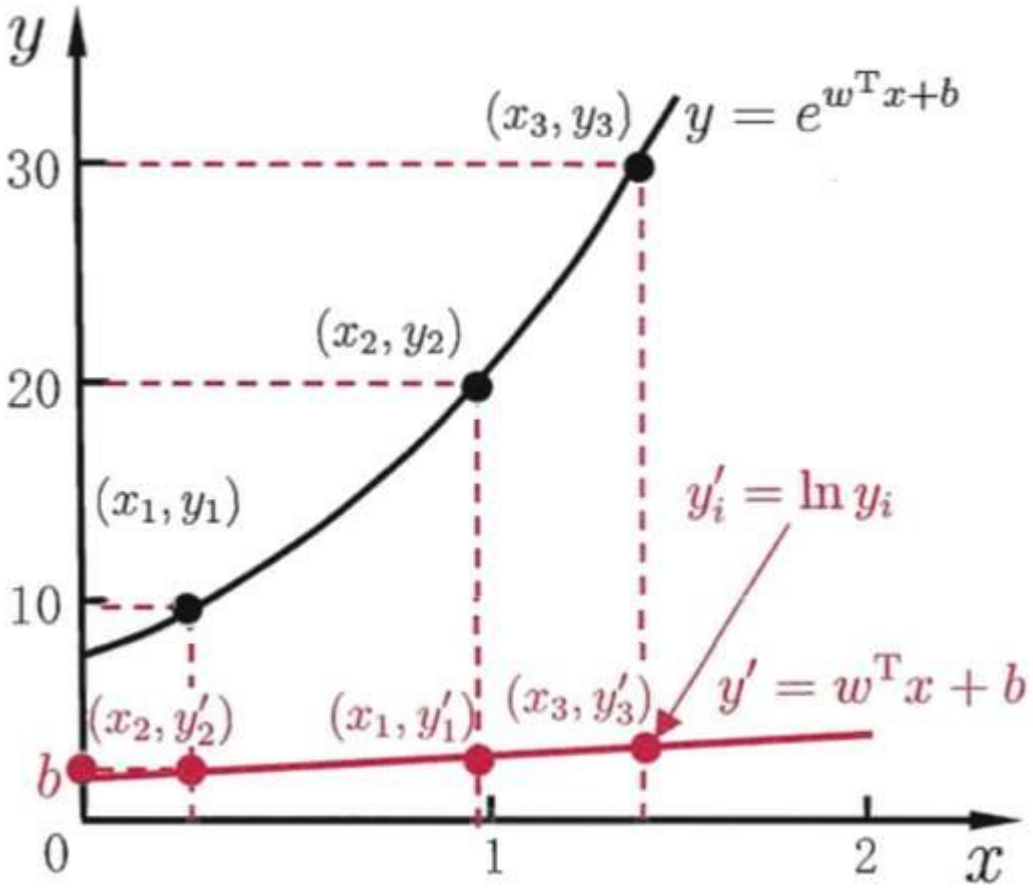
- y服从指数族分布(包括高斯分布，伯努利分布，多项式分布，泊松分布，beta分布……)，且同个样本的y必须服从同个分布

- 接着在具体分布中比较与指数分布族之间的参数关系，最重要的就是具体分布的参数(Φ)和指数分布参数(η)之间的关系



# More complicated example

**Generalized Linear Models (GLM)：广义线性模型**



# Limitation of Association Method

- For multiple-domain proteins, this method computes the value for a certain domain pair ignoring the value of other domain-domain pairs. So it's a local one.
- This method cannot deal with  possible error of the data.

# Probabilistic Model

- Domain-domain interactions are independent, which means that the event that two domains interact or not does not depend on other domains.
- Two proteins interact if and only if at least one pair of domains from the two proteins interact.

# Yeast Data

- Interactions (Uetz's and Ito's interaction data).
- Domain: Pfam (Pfam-A, Pfam-B).
- Proteins: SGD, N=6359.

# Protein Interaction Data Sources

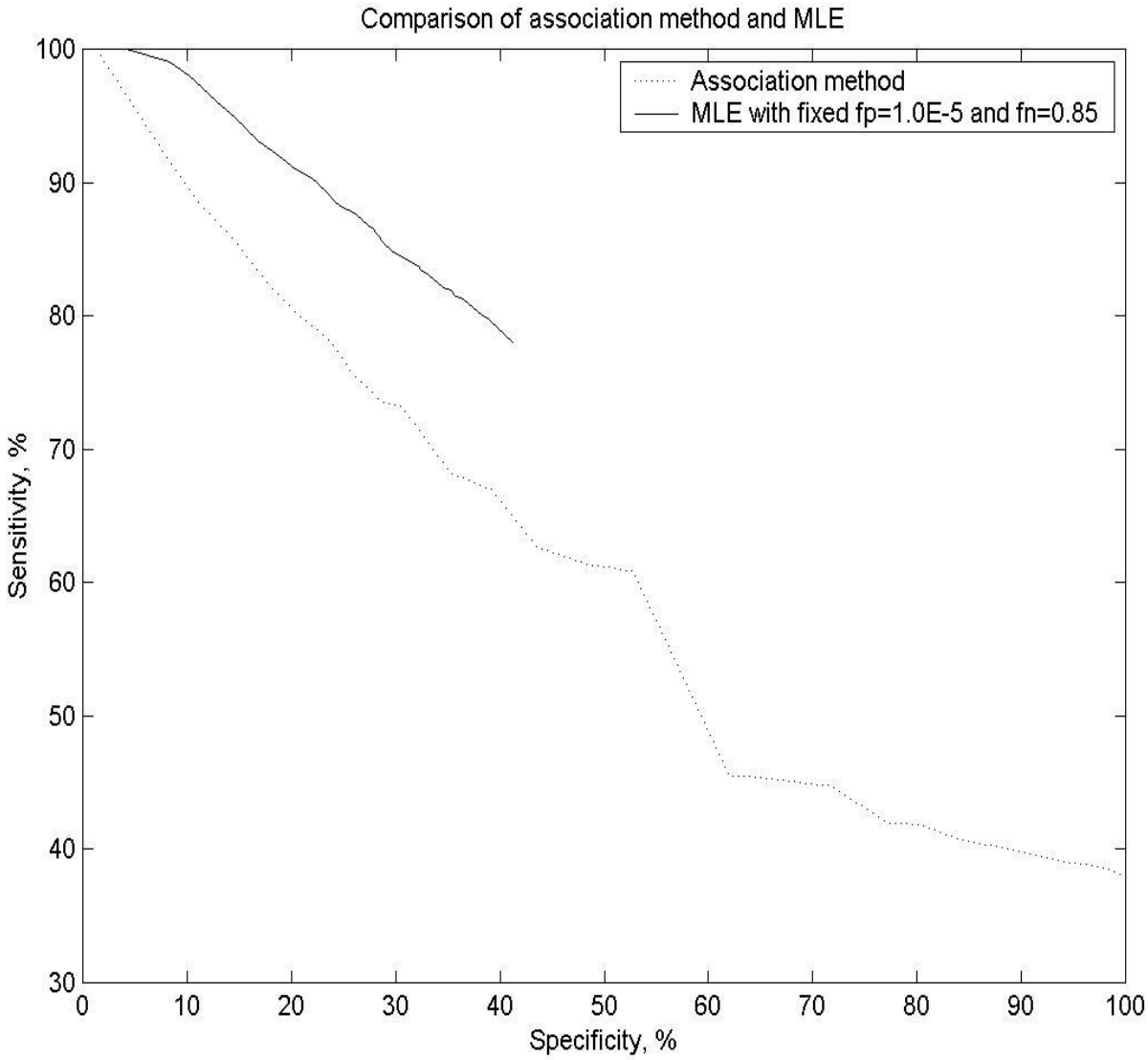
| | Proteins | Pfam domains | Super - domains | PPI |
|---|---|---|---|---|
| Uetz | 1337 | 1330 | 313 | 1445 |
| Ito | 3277 | 2776 | 909 | 4475 |
| Uetz+Ito | 3729 | 3124 | 1007 | 5719 |
| Overlap | 855 | 964 | 215 | 201 |

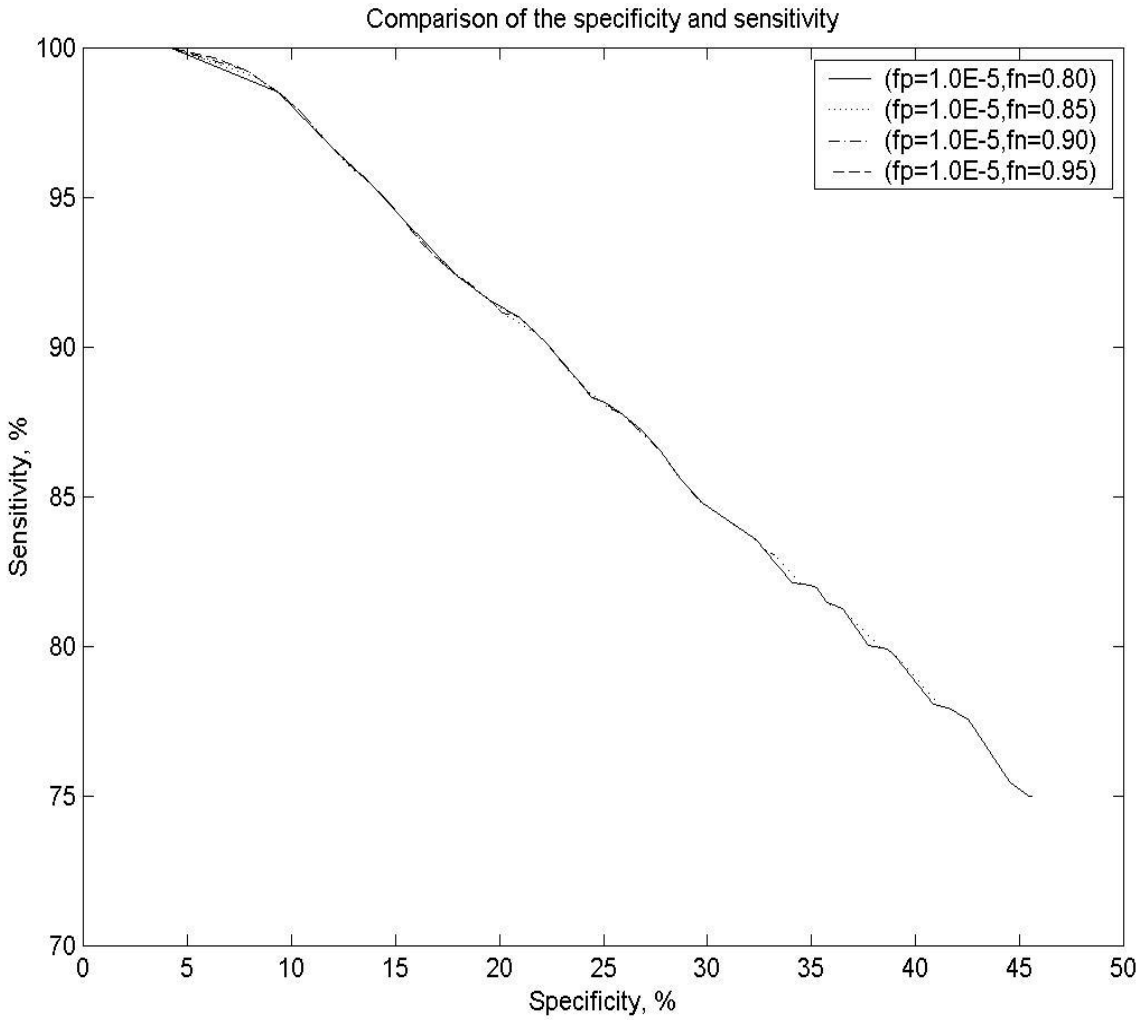
# Measure the Accuracy

- Specificity and sensitivity.
- Verification by MIPS physical interactions (as TRUE interactions).
- Relationship between protein-protein interactions and expression data.

$$SP = \frac{\text{number of matches with observation}}{\text{number of prediction}}$$

$$SN = \frac{\text{number of match with observation}}{\text{number of observation}}$$


Comparison of association method and MLE
Association method
MLE with fixed fp=1.0E-5 and fn=0.85
Sensitivity, %
Specificity, %
100
90
80
70
60
50
40
30
0
10
20
30
40
50
60
70
80
90
100


Comparison of the specificity and sensitivity
(fp=1.0E-5,fn=0.80)
(fp=1.0E-5,fn=0.85)
(fp=1.0E-5,fn=0.90)
(fp=1.0E-5,fn=0.95)
Sensitivity, %
Specificity, %
100
95
90
85
80
75
70
0
5
10
15
20
25
30
35
40
45
50

# Verification by Known PPIs

- MIPS physical interaction. (Totally 2570 PPIs, 1414 PPIs not overlapping with our training set).
- Compare with random matching.
  - Fold number
  - Larger fold number imply more reliable prediction

$$\#\,\mathrm{Fold} = \frac{\#\{\text{Our prediction matched to MIPS}\}}{\#\{\text{Expectation of random pairs matched to MIPS}\}}$$

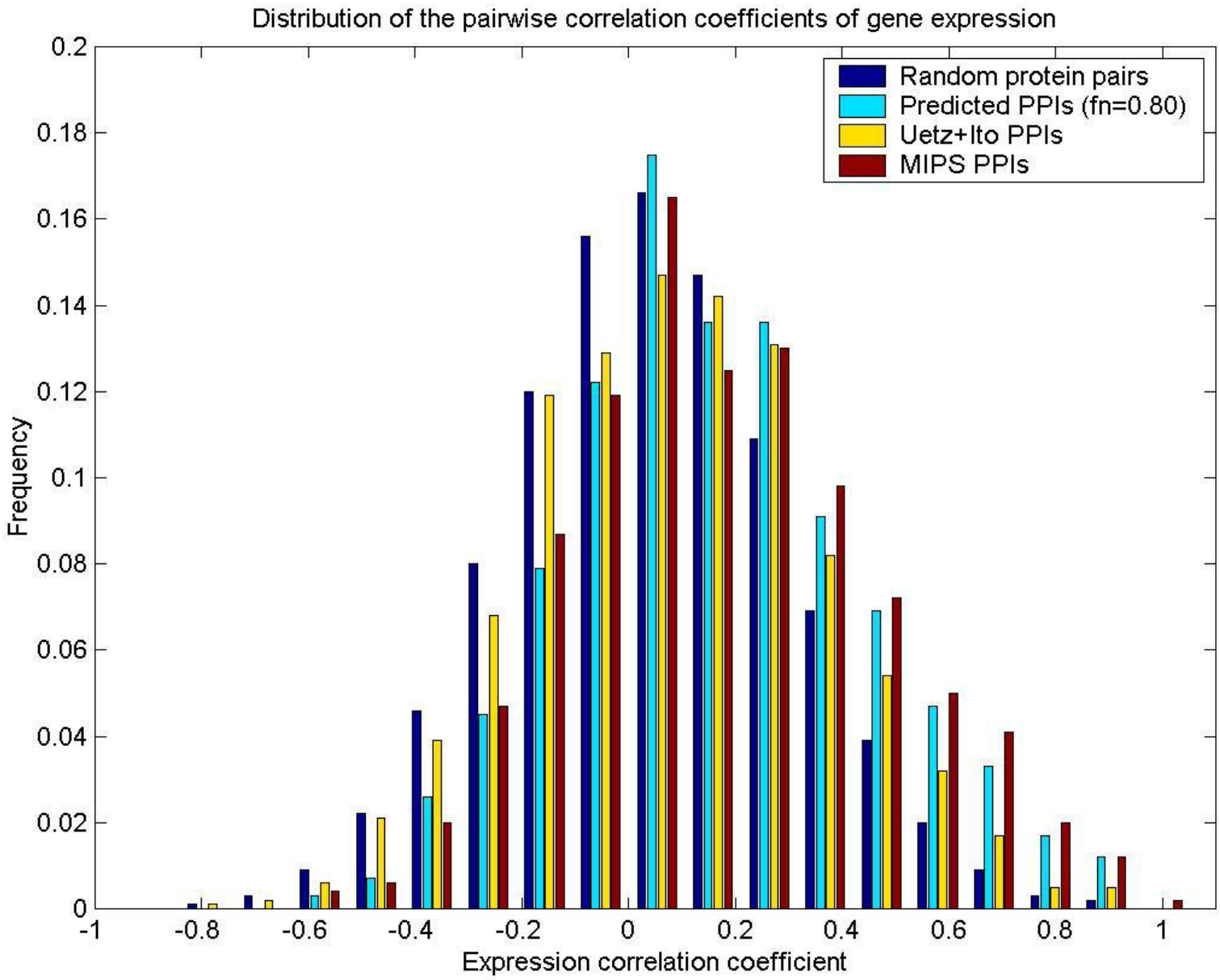
Think about FDR again...

# Matching with MIPS PPIs

| Prob | #Predict | #Train | #MIPS | | #Fold |
|---|---|---|---|---|---|
| All | 20221620 | 5719 | 2570 | 1414 | 1.00 |
| >0.00 | 136463 | 5719 | 1265 | 109 | 11.92 |
| >=0.20 | 26908 | 5238 | 1093 | 53 | 34.97 |
| >=0.40 | 19360 | 5018 | 1035 | 48 | 47.85 |
| >=0.60 | 14725 | 4775 | 971 | 47 | 67.53 |
| >=0.80 | 12734 | 4647 | 932 | 43 | 76.02 |
| >=0.975 | 10824 | 4461 | 886 | 40 | 89.88 |

# Interaction Data Correlated With Gene Expression Data

- Interacted proteins seems to have high expression correlation
  - A.Grigorieve *Nucleic Acid Res.* 29, 2001;
  - H. Ge et al. *Nature Genetics* 29, 2001;
  - R. Jansen et al. *Genome Res*.12, 2002.
- Expression data (M.Eisen, 1998); 2465 Yeast ORFs with 79 data points/ORF.
- Pearson correlation coefficient.


Distribution of the pairwise correlation coefficients of gene expression
Random protein pairs
Predicted PPIs (fn=0.80)
Uetz+Ito PPIs
MIPS PPIs
Frequency
Expression correlation coefficient
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

# Statistics of Pairwise Correlation of Gene Expression

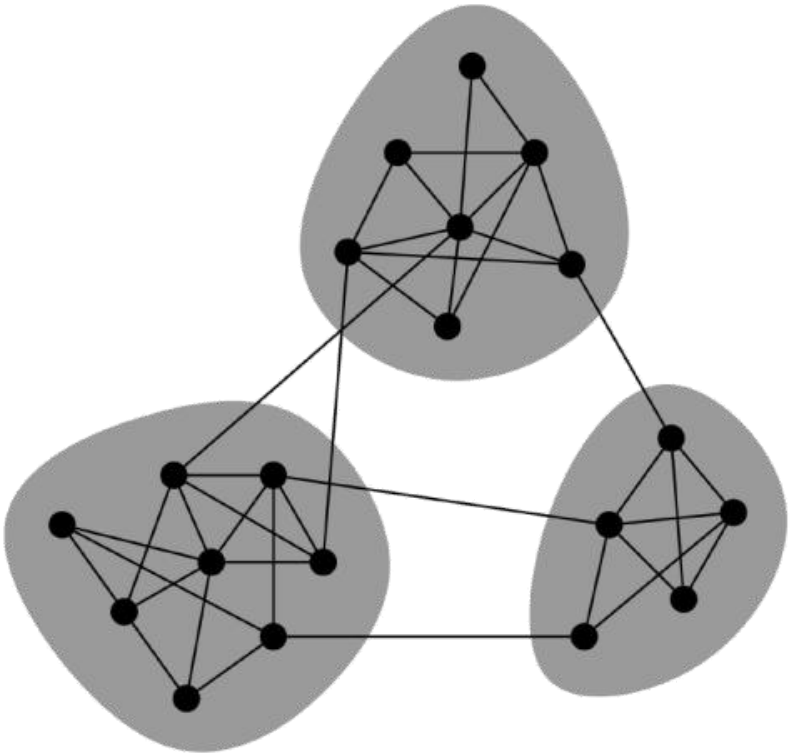
| pairs | # pairs | mean | std | T-score | p-value | $R* > 0.5$ |
|---|---|---|---|---|---|---|
| All ORFs | 3036880 | 0.0428 | 0.2473 | 0.0000 | 5.000e-01 | 3.84% |
| $\geq$0.20 | 6392 | 0.0514 | 0.2550 | 2.7984 | 2.575e-03 | 4.79% |
| $\geq$0.40 | 4433 | 0.0510 | 0.2538 | 2.2232 | 1.311e-02 | 4.96% |
| $\geq$0.60 | 3318 | 0.0598 | 0.2579 | 3.9644 | 3.715e-05 | 5.42% |
| $\geq$0.80 | 2756 | 0.0626 | 0.2622 | 4.2196 | 1.238e-05 | 5.88% |
| $\geq$0.975 | 2266 | 0.0628 | 0.2637 | 3.8482 | 6.002e-05 | 5.87% |
| Uetz+Ito | 1307 | 0.0586 | 0.2587 | 2.3213 | 1.015e-02 | 5.20% |
| MIPS | 1106 | 0.1109 | 0.2767 | 9.1619 | 2.706e-20 | 8.23% |

$$T = \frac{\mu_1 - \mu_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}}$$

# 第8-3章： Network Module

- Definition
- Module detection
- Bayesian approach
- Markov clustering algorithm

# Network Modular



# Modularity

- Suppose we are given a candidate division of the vertices into some number of groups. The modularity of this division is defined to be the fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random.

http://en.wikipedia.org/wiki/Modularity_(networks)

# Modularity

- $A_{ij}$: adjacency matrix
- $k_i$: degree
- m: total number of edges

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$$

# Modularity

- For two class problem, let $s_i$=1 of node I belongs to group1 and $s_i$=-1 if it belongs to group 2,

$$\delta(c_i, c_j) = \frac{1}{2}(s_i s_j + 1)$$

$$Q = \frac{1}{4m} \sum_{ij} S^T B S$$

$$B = (B_{ij}), B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

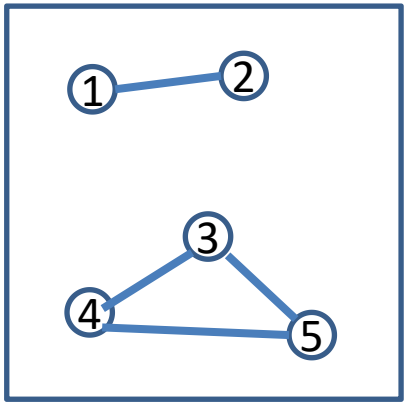
$$S = (s_1, \cdots, s_n)^T$$

# Example

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

$$m = 4, k_1 = k_2 = 1$$

$$k_3 = k_4 = k_5 = 2$$




$$B = \frac{1}{8} \begin{pmatrix} -1 & 7 & -2 & -2 & -2 \\ 7 & -1 & -2 & -2 & -2 \\ -2 & -2 & -4 & 4 & 4 \\ -2 & -2 & 4 & -4 & 4 \\ -2 & -2 & 4 & 4 & -4 \end{pmatrix}$$

# Spectrum Method

- The largest eigenvectors will gives the best grouping, positive entries corresponding to one class, and negative ones corresponding to another class.

- This can be achieved by power method

$$\lim_{k \to +\infty} = \frac{A^k e}{e^T A^k e} = w$$

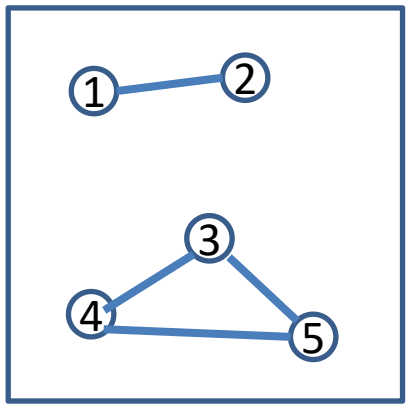
where $e=(1,1,\ldots,1)^T$

# Example

- 对于上述矩阵B, 可以计算出最大特征值为10, 对应的特征向量

$$v = (-0.55, -0.55, 0.37, 0.37, 0.37)$$

- 于是我们对节点的划分为{1,2}; {3,4,5}



# 优化方法

- 既然现在有一个衡量划分“好坏”的量Q, 那么一般的优化方法都可以使用;
  - 1. 给定初始划分
  - 2. 对于划分的某种修正, 计算Q的改变量
  - 3. 依据一定的原则考虑是否接受这种修正，重复步骤2, 直到某种收敛条件满足。
- Greedy方法
- 模拟退火方法

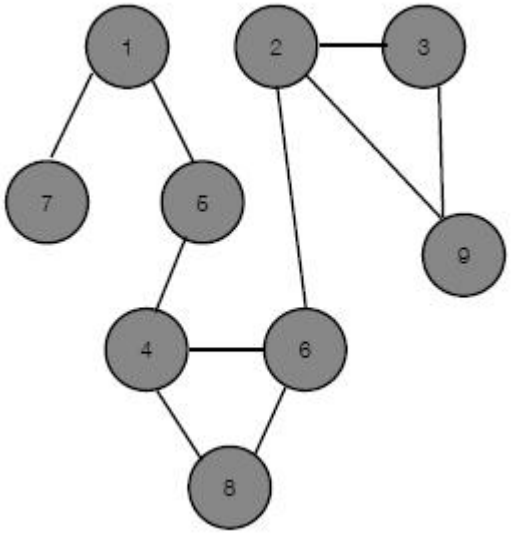
# A Bayesian Approach to Network Modularity

Slides for this part are mainly from Hofman's talk

*www.jakehofman.com/talks/**apam_20071019**.pdf*

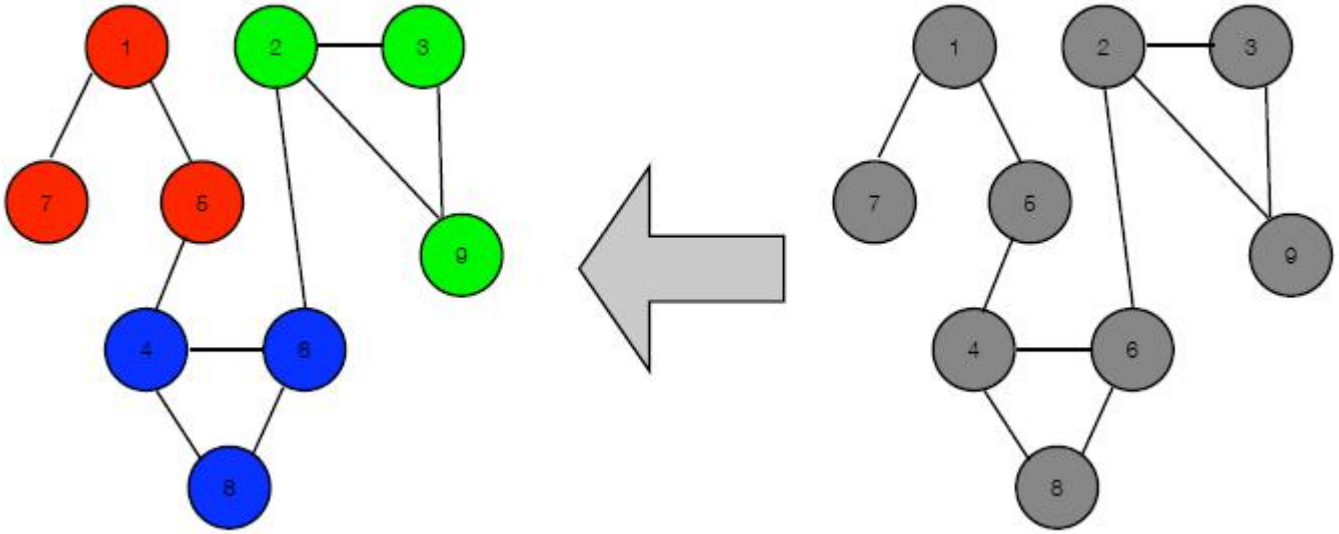
# Overview: Modular Networks

- Given a network
  - Assign nodes to modules?
  - Determine number of modules(scale/complexity)?



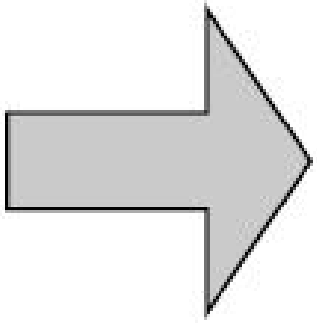
# Overview: Modular Networks

- With a generative model of modular networks, rules of probability tell us how to calculate model parameters (e.g. number of modules & assignments)



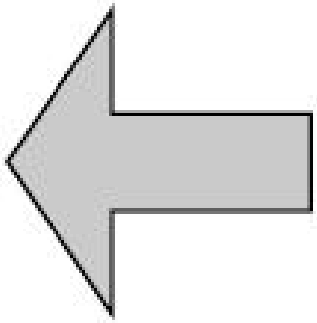
# Generative Models

Know model (parameters, assignment variables, complexity)



Generate synthetic data

Infer model (parameters, latent variables, complexity)



Observe real data

# Markov Clustering Algorithm

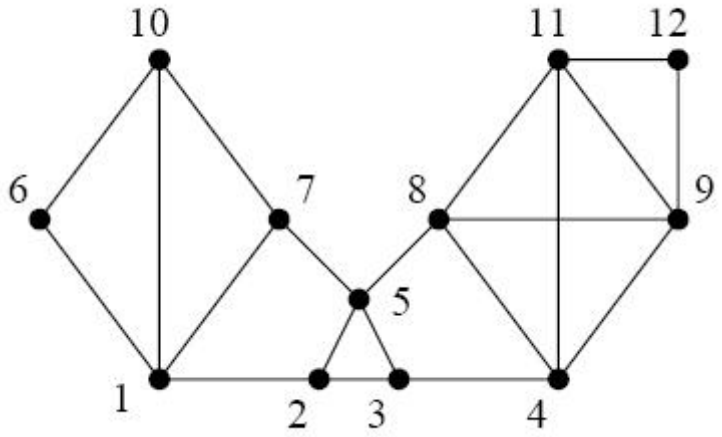
van Dongen. A cluster algorithm for graphs. Information Systems, 2000

# K-length Path

- Basic idea: dense regions in sparse graphs corresponding with regions in which the number of k-length path is relatively large.

- Random walks can also be used to detect clusters in graphs, the idea is that the more closed is a subgraph, the largest the time a random walker need to escape from it.

# K-path Clustering




$$\begin{pmatrix}
5 & 2 & 1 & 0 & 2 & \mathbf{3} & \mathbf{3} & 0 & 0 & \mathbf{4} & 0 & 0 \\
2 & 4 & \mathbf{3} & 1 & \mathbf{3} & 1 & 2 & 1 & 0 & 1 & 0 & 0 \\
1 & \mathbf{3} & 4 & 2 & \mathbf{3} & 0 & 1 & 2 & 1 & 0 & 1 & 0 \\
0 & 1 & 2 & 5 & 2 & 0 & 0 & \mathbf{4} & \mathbf{4} & 0 & \mathbf{4} & 2 \\
2 & \mathbf{3} & \mathbf{3} & 2 & 5 & 0 & 2 & 2 & 1 & 1 & 1 & 0 \\
\mathbf{3} & 1 & 0 & 0 & 0 & 3 & 2 & 0 & 0 & \mathbf{3} & 0 & 0 \\
\mathbf{3} & 2 & 1 & 0 & 2 & 2 & 4 & 1 & 0 & \mathbf{3} & 0 & 0 \\
0 & 1 & 2 & \mathbf{4} & 2 & 0 & 1 & 5 & \mathbf{4} & 0 & \mathbf{4} & 2 \\
0 & 0 & 1 & \mathbf{4} & 1 & 0 & 0 & \mathbf{4} & 5 & 0 & \mathbf{5} & \mathbf{3} \\
\mathbf{4} & 1 & 0 & 0 & 1 & \mathbf{3} & \mathbf{3} & 0 & 0 & 4 & 0 & 0 \\
0 & 0 & 1 & \mathbf{4} & 1 & 0 & 0 & \mathbf{4} & \mathbf{5} & 0 & 5 & \mathbf{3} \\
0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & \mathbf{3} & 0 & \mathbf{3} & 3
\end{pmatrix}$$

Matrix manipulation: $(N+I)^2$

# Markov Clustering

- Expansion: Through matrix manipulation (power), one obtains a matrix for a n-steps connection.

- Inflation: Enhance intercluster passages by raising the elements to a certain power and then normalize

# Markov Clustering Algorithm

- Iteratively running two operators
  - Inflation:

$$(T_r M)_{ij} = \frac{M_{ij}^r}{\sum_i M_{ij}^r}$$

Column normalization

  - Expansion:

$$\mathrm{Expand}(M) = M^k$$

# MCL Running

$$\begin{pmatrix}
0.380 & 0.087 & 0.027 & -- & 0.077 & 0.295 & 0.201 & -- & -- & 0.320 & -- & -- \\
0.047 & 0.347 & 0.210 & 0.017 & 0.150 & 0.019 & 0.066 & 0.012 & -- & 0.012 & -- & -- \\
0.014 & 0.210 & 0.347 & 0.056 & 0.150 & -- & 0.016 & 0.046 & 0.009 & -- & 0.009 & -- \\
-- & 0.027 & 0.087 & 0.302 & 0.062 & -- & -- & 0.184 & 0.143 & -- & 0.143 & 0.083 \\
0.058 & 0.210 & 0.210 & 0.056 & 0.406 & -- & 0.083 & 0.046 & 0.009 & 0.019 & 0.009 & -- \\
0.142 & 0.017 & -- & -- & -- & 0.295 & 0.083 & -- & -- & 0.184 & -- & -- \\
0.113 & 0.069 & 0.017 & -- & 0.062 & 0.097 & 0.333 & 0.012 & -- & 0.147 & -- & -- \\
-- & 0.017 & 0.069 & 0.175 & 0.049 & -- & 0.016 & 0.287 & 0.143 & -- & 0.143 & 0.083 \\
-- & -- & 0.017 & 0.175 & 0.012 & -- & -- & 0.184 & 0.288 & -- & 0.288 & 0.278 \\
0.246 & 0.017 & -- & -- & 0.019 & 0.295 & 0.201 & -- & -- & 0.320 & -- & -- \\
-- & -- & 0.017 & 0.175 & 0.012 & -- & -- & 0.184 & 0.288 & -- & 0.288 & 0.278 \\
-- & -- & -- & 0.044 & -- & -- & -- & 0.046 & 0.120 & -- & 0.120 & 0.278
\end{pmatrix}$$

$\Gamma_2 M^2$, $M$ defined in Figure 8

# MCL Running

$$\begin{pmatrix}
0.448 & 0.080 & 0.023 & -- & 0.068 & 0.426 & 0.359 & -- & -- & 0.432 & -- & -- \\
0.018 & 0.285 & 0.228 & 0.007 & 0.176 & 0.006 & 0.033 & 0.005 & -- & 0.007 & -- & -- \\
0.005 & 0.223 & 0.290 & 0.022 & 0.173 & -- & 0.010 & 0.017 & 0.003 & 0.001 & 0.003 & 0.001 \\
-- & 0.018 & 0.059 & 0.222 & 0.040 & -- & 0.001 & 0.187 & 0.139 & -- & 0.139 & 0.099 \\
0.027 & 0.312 & 0.314 & 0.028 & 0.439 & 0.005 & 0.054 & 0.022 & 0.003 & 0.010 & 0.003 & 0.001 \\
0.116 & 0.007 & 0.001 & -- & 0.004 & 0.157 & 0.085 & -- & -- & 0.131 & -- & -- \\
0.096 & 0.040 & 0.013 & -- & 0.037 & 0.083 & 0.197 & 0.001 & -- & 0.104 & -- & -- \\
-- & 0.012 & 0.042 & 0.172 & 0.029 & -- & 0.002 & 0.198 & 0.133 & -- & 0.133 & 0.096 \\
-- & 0.001 & 0.015 & 0.256 & 0.009 & -- & -- & 0.266 & 0.326 & -- & 0.326 & 0.346 \\
0.290 & 0.021 & 0.002 & -- & 0.017 & 0.323 & 0.260 & -- & -- & 0.316 & -- & -- \\
-- & 0.001 & 0.015 & 0.256 & 0.009 & -- & -- & 0.266 & 0.326 & -- & 0.326 & 0.346 \\
-- & -- & 0.001 & 0.037 & 0.001 & -- & -- & 0.039 & 0.069 & -- & 0.069 & 0.112
\end{pmatrix}$$

$$\Gamma_2(\Gamma_2 M^2 \cdot \Gamma_2 M^2)$$

# MCL Running

$$\begin{pmatrix}
0.807 & 0.040 & 0.015 & -- & 0.034 & 0.807 & 0.807 & -- & -- & 0.807 & -- & -- \\
-- & 0.090 & 0.092 & -- & 0.088 & -- & -- & -- & -- & -- & -- & -- \\
-- & 0.085 & 0.088 & -- & 0.084 & -- & -- & -- & -- & -- & -- & -- \\
-- & 0.001 & 0.001 & 0.032 & 0.001 & -- & -- & 0.032 & 0.031 & -- & 0.031 & 0.031 \\
-- & 0.777 & 0.798 & -- & 0.786 & -- & 0.001 & -- & -- & -- & -- & -- \\
0.005 & -- & -- & -- & -- & 0.005 & 0.005 & -- & -- & 0.005 & -- & -- \\
0.003 & 0.001 & -- & -- & 0.001 & 0.003 & 0.003 & -- & -- & 0.003 & -- & -- \\
-- & -- & 0.001 & 0.024 & -- & -- & -- & 0.024 & 0.024 & -- & 0.024 & 0.024 \\
-- & -- & 0.002 & 0.472 & 0.001 & -- & -- & 0.472 & 0.472 & -- & 0.472 & 0.472 \\
0.185 & 0.005 & 0.001 & -- & 0.004 & 0.185 & 0.184 & -- & -- & 0.185 & -- & -- \\
-- & -- & 0.002 & 0.472 & 0.001 & -- & -- & 0.472 & 0.472 & -- & 0.472 & 0.472 \\
-- & -- & -- & 0.001 & -- & -- & -- & 0.001 & 0.001 & -- & 0.001 & --
\end{pmatrix}$$

$(\Gamma_2 \circ Squaring)$ iterated four times on $M$

# MCL Running

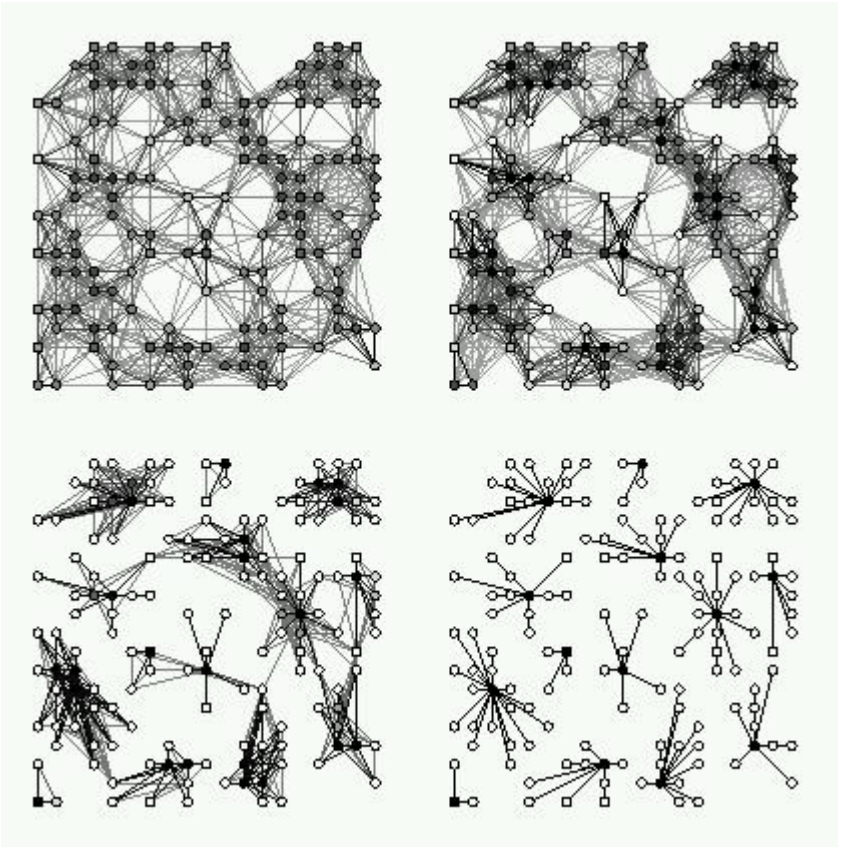
$$
\begin{pmatrix}
1.000 & -- & -- & -- & -- & 1.000 & 1.000 & -- & -- & 1.000 & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & 1.000 & 1.000 & -- & 1.000 & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & 0.500 & -- & -- & -- & 0.500 & 0.500 & -- & 0.500 & 0.500 \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & 0.500 & -- & -- & -- & 0.500 & 0.500 & -- & 0.500 & 0.500 \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & --
\end{pmatrix}
$$

$M_{mcl}^{\infty}$

# A Heuristic for MCL

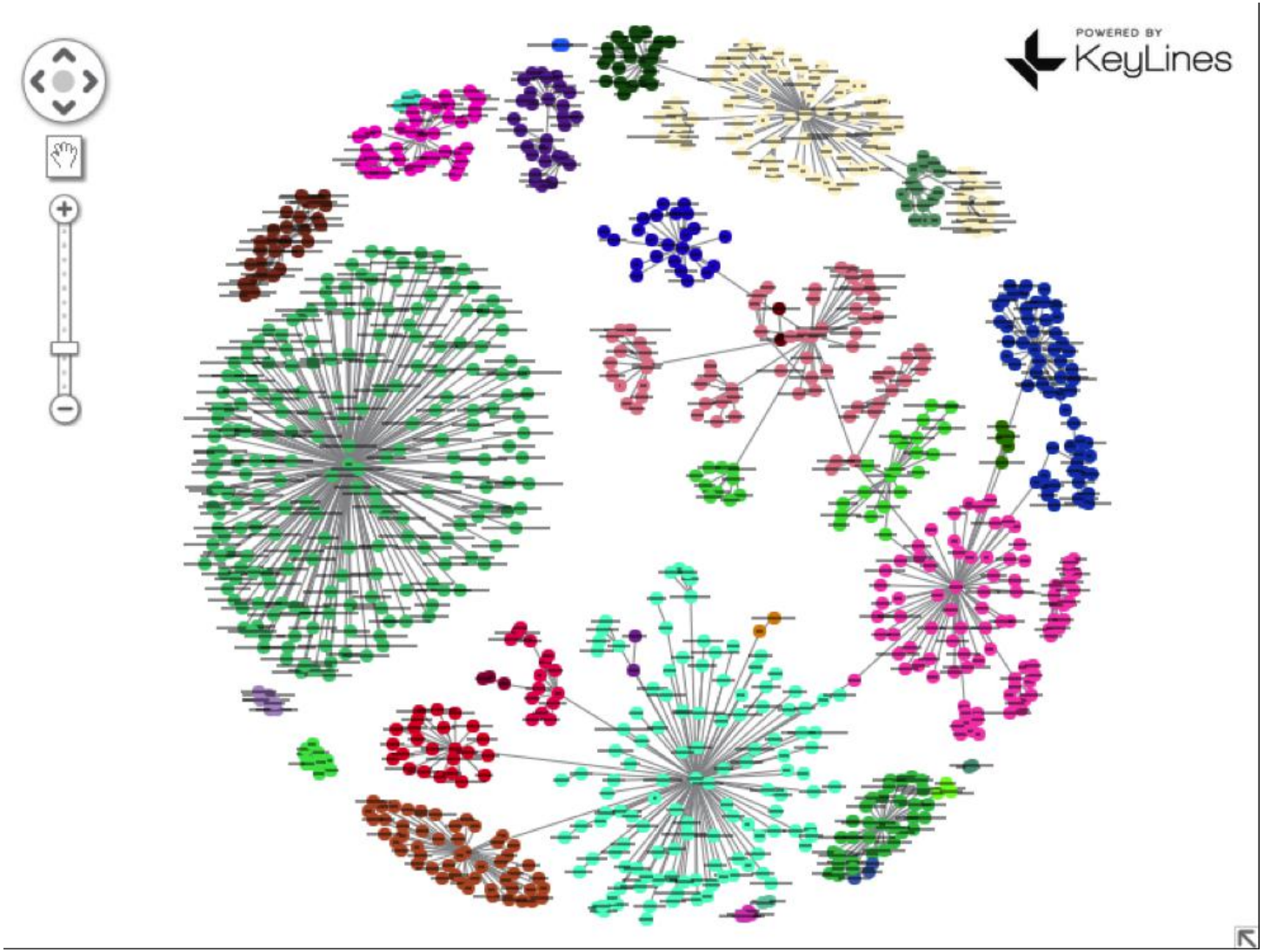
We take a random walk on the graph described by the similarity matrix

After each step we weaken the links between distant nodes and strengthen the links between nearby nodes

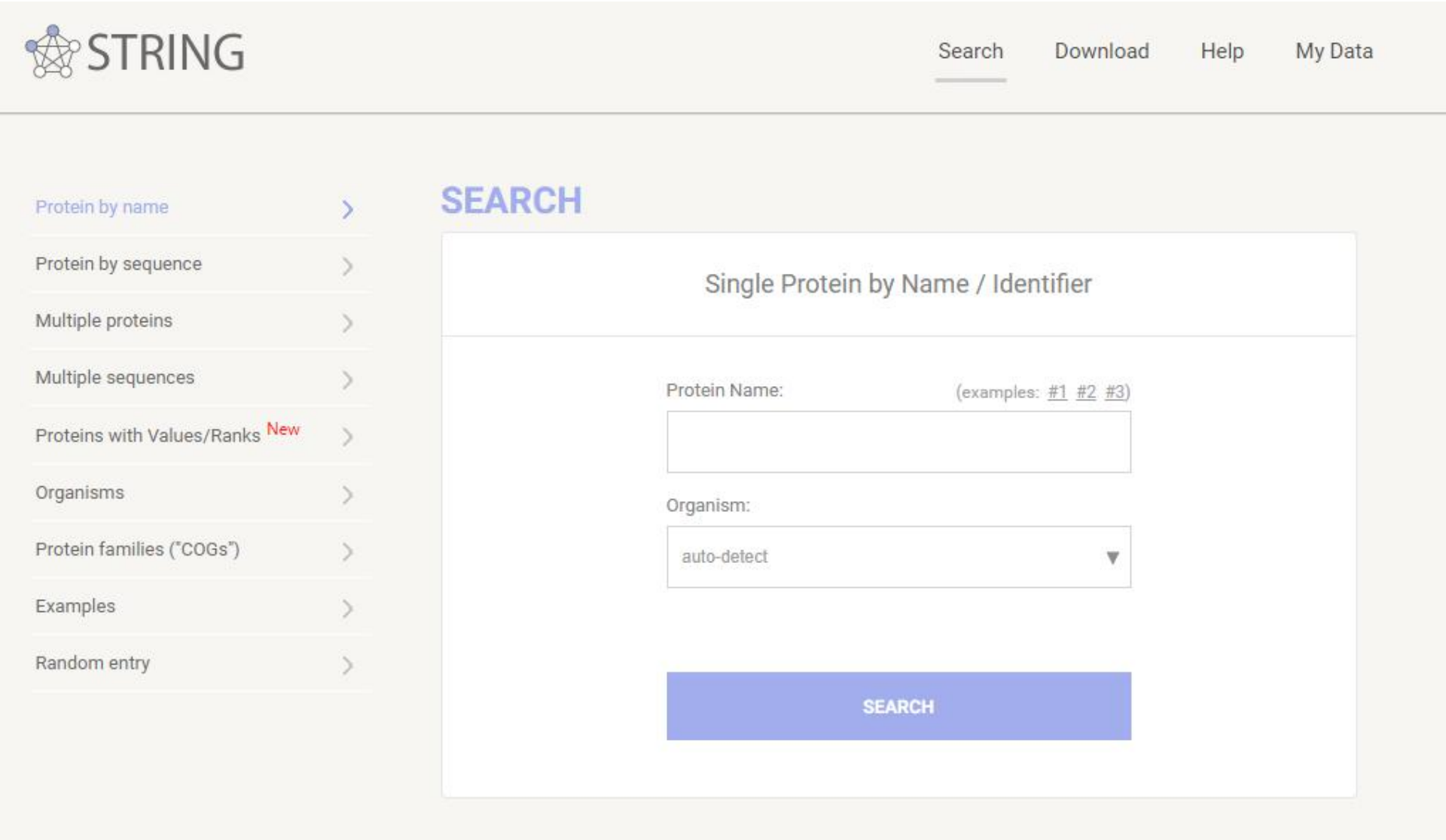


Graphic from van Dongen, 2000

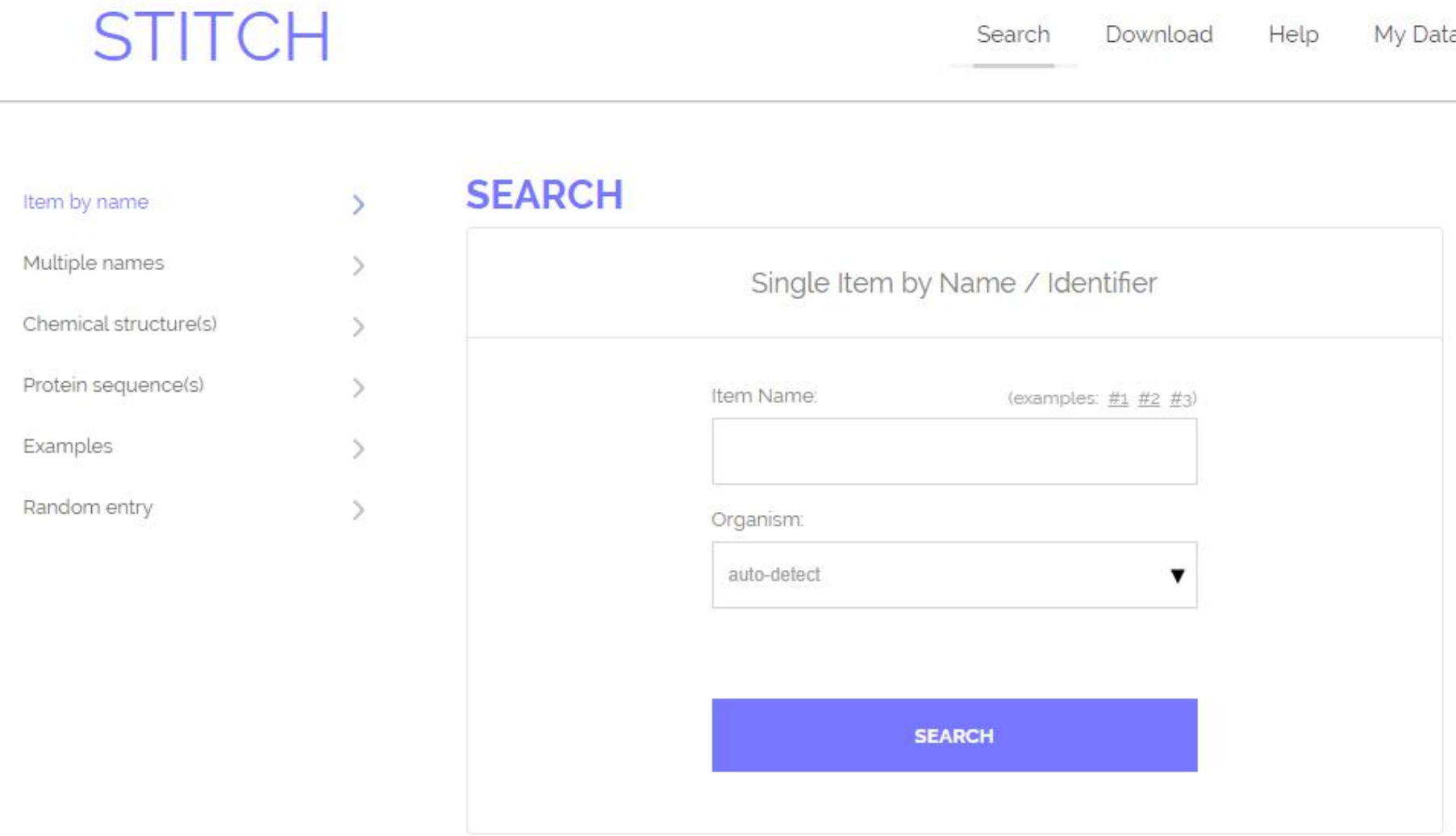
# Clustering examples



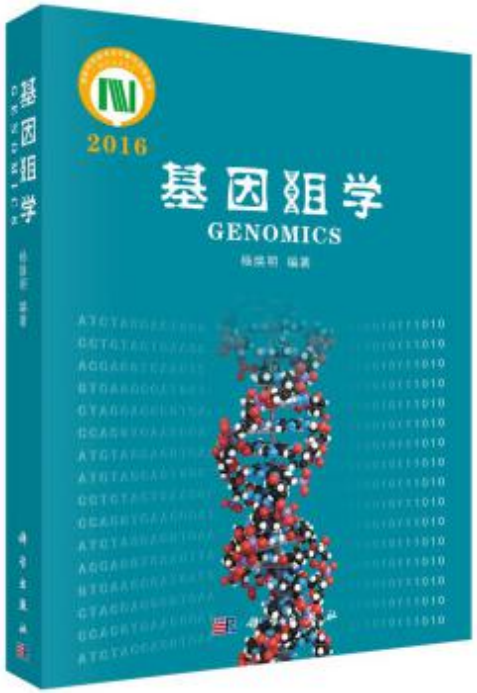
# STRING

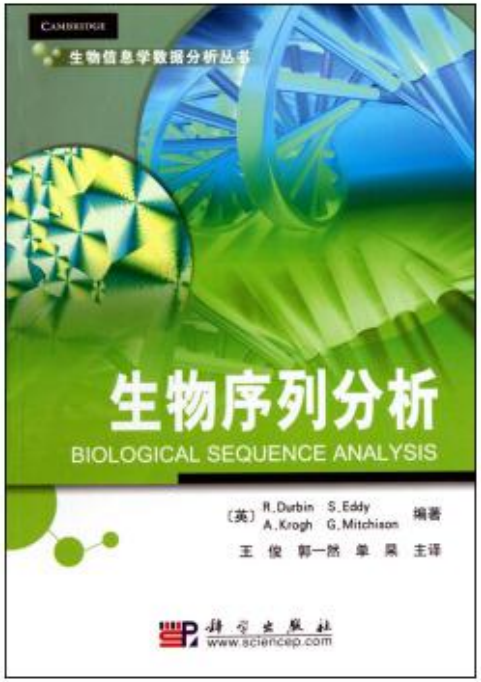

STRING
Search
Download
Help
My Data
Protein by name
Protein by sequence
Multiple proteins
Multiple sequences
Proteins with Values/Ranks New
Organisms
Protein families ("COGs")
Examples
Random entry
SEARCH
Single Protein by Name / Identifier
Protein Name:
(examples: #1 #2 #3)
Organism:
auto-detect
SEARCH

# STITCH


STITCH
Search
Download
Help
My Data
Item by name
Multiple names
Chemical structure(s)
Protein sequence(s)
Examples
Random entry
SEARCH
Single Item by Name / Identifier
Item Name:
(examples: #1 #2 #3)
Organism:
auto-detect
SEARCH

# References

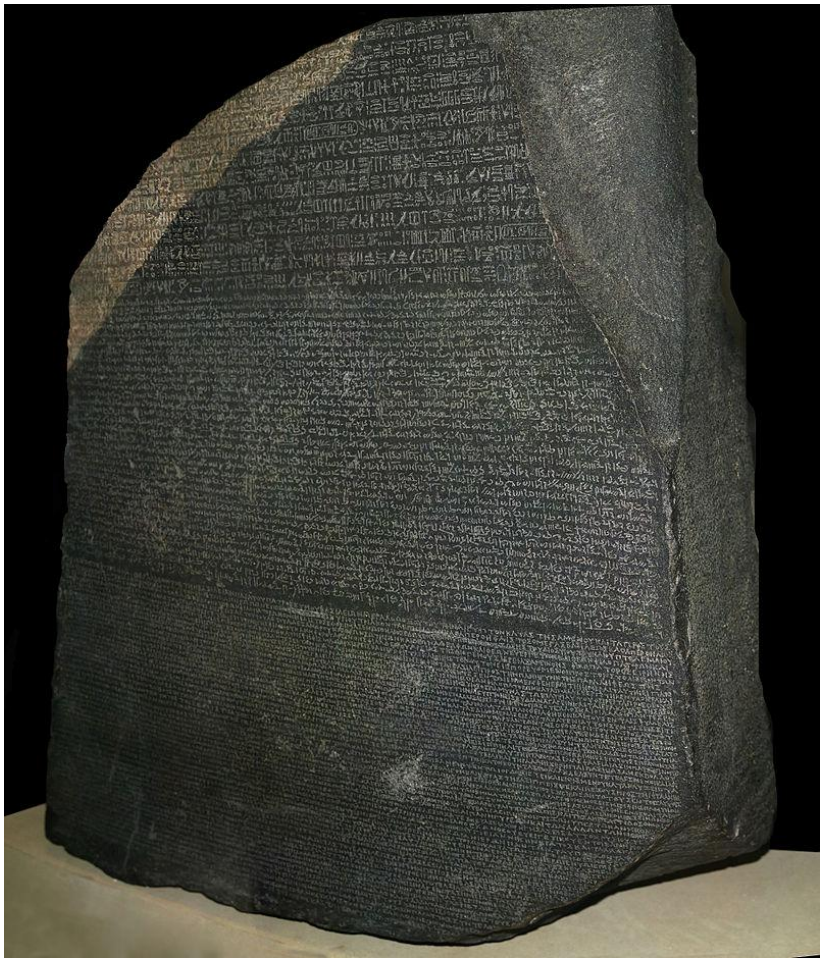

2016
基因组学
GENOMICS


CAMBRIDGE
生物信息学数据分析丛书
生物序列分析
BIOLOGICAL SEQUENCE ANALYSIS
R.Durbin S.Eddy
A.Krogh G.Mitchison
编著
王俊 郭一然 单昊 主译
www.sciencep.com

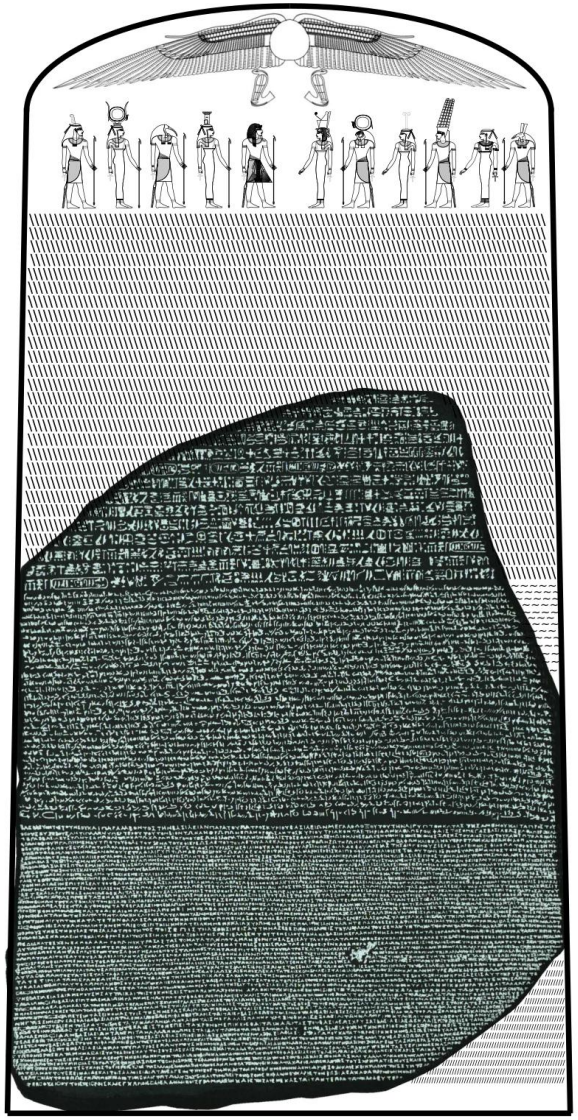

生物统计学
(第五版)

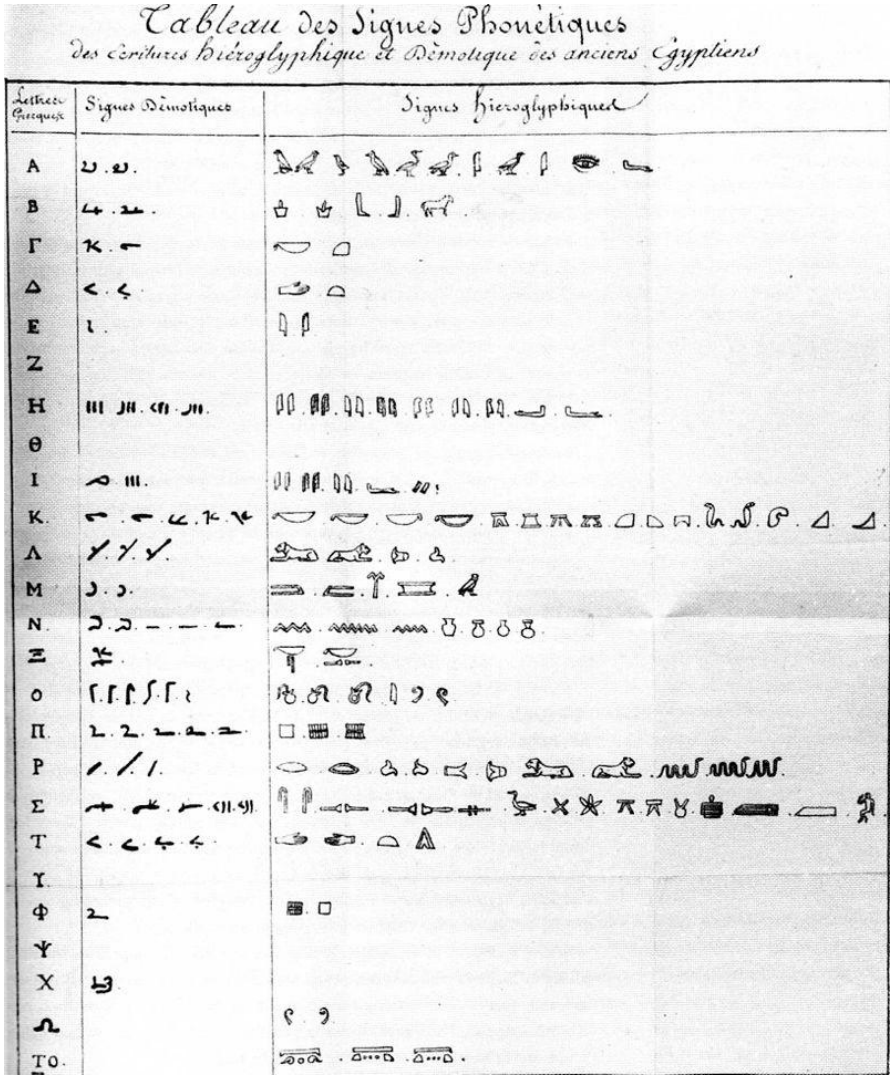
# 补充知识

# Rosetta stone

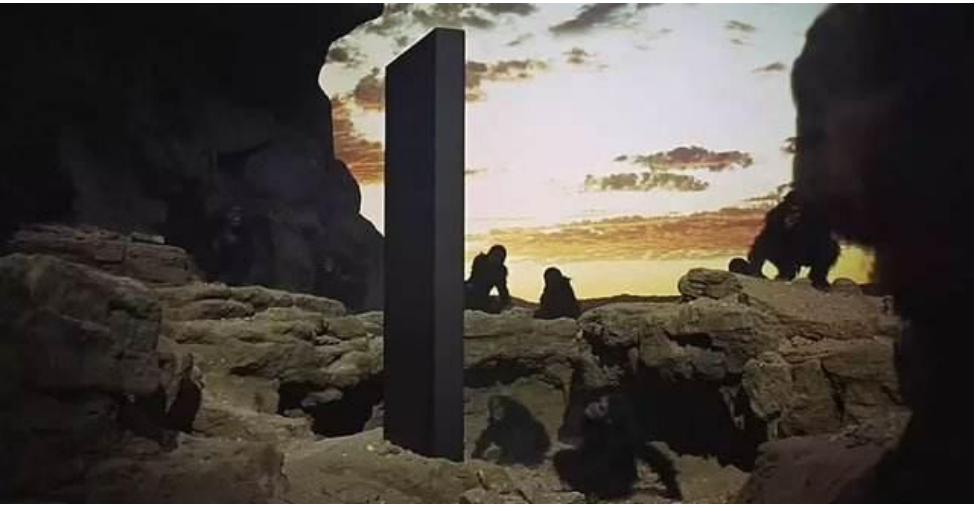


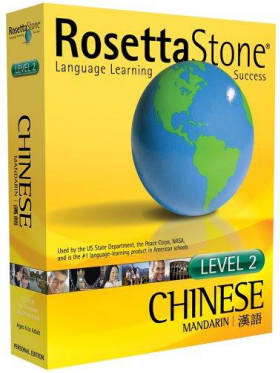
# Rosetta stone





# Rosetta stone

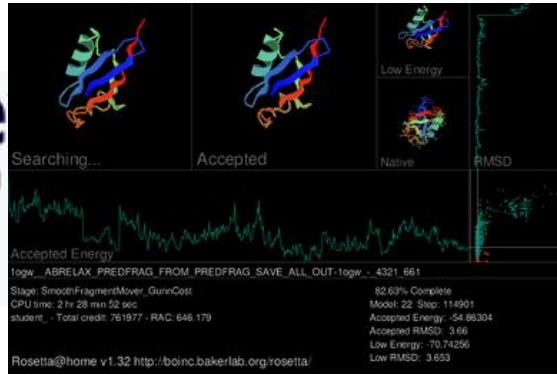




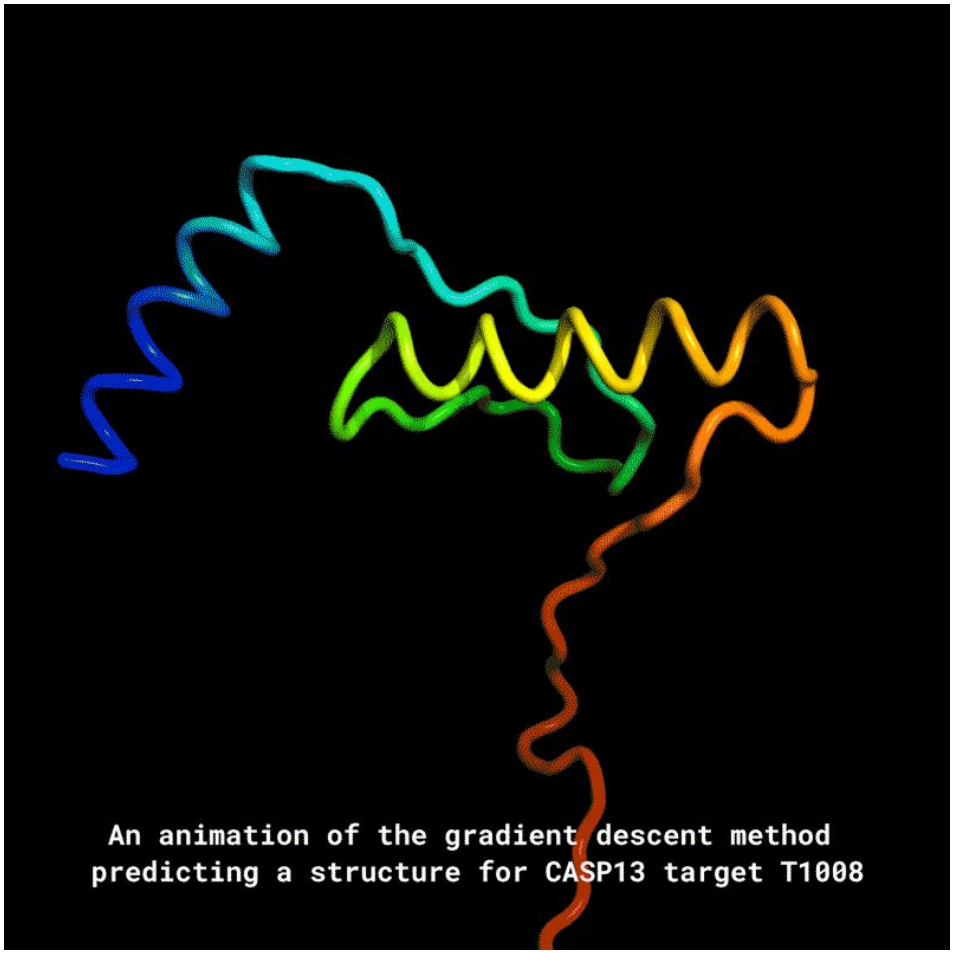
Rosetta@home

Protein Folding, Design, and Docking



# CASP

## (Critical Assessment of Techniques for Protein Structure Prediction)



# CASP

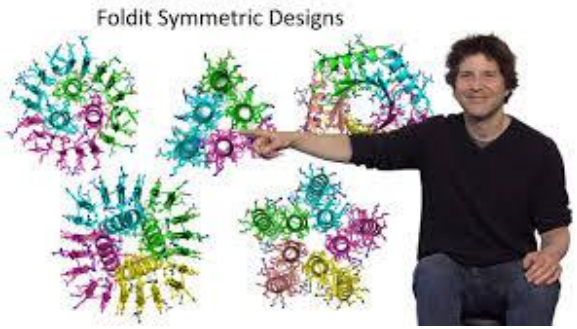
## (Critical Assessment of Techniques for Protein Structure Prediction)

Rosetta Commons

The hub for Rosetta modeling software

I-TASSER

Protein Structure & Function Predictions



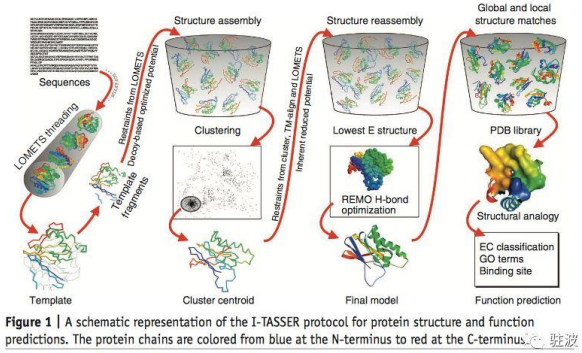
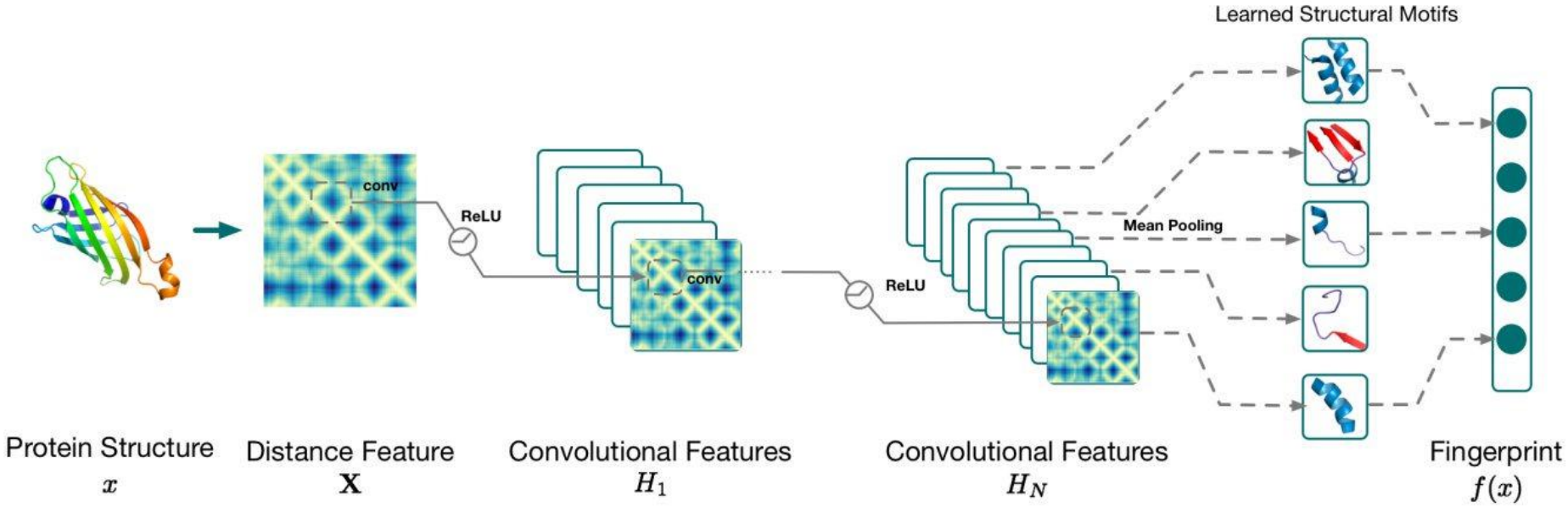




**Figure 1** | A schematic representation of the I-TASSER protocol for protein structure and function predictions. The protein chains are colored from blue at the N-terminus to red at the C-terminus.

# CASP

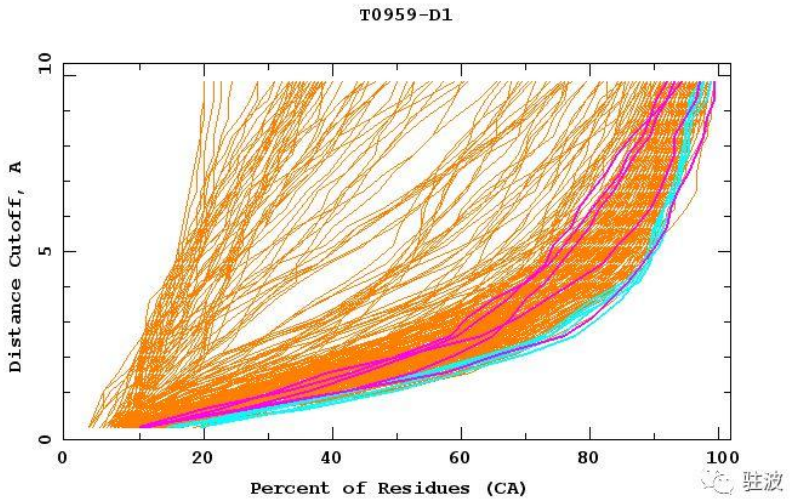
## (Critical Assessment of Techniques for Protein Structure Prediction)

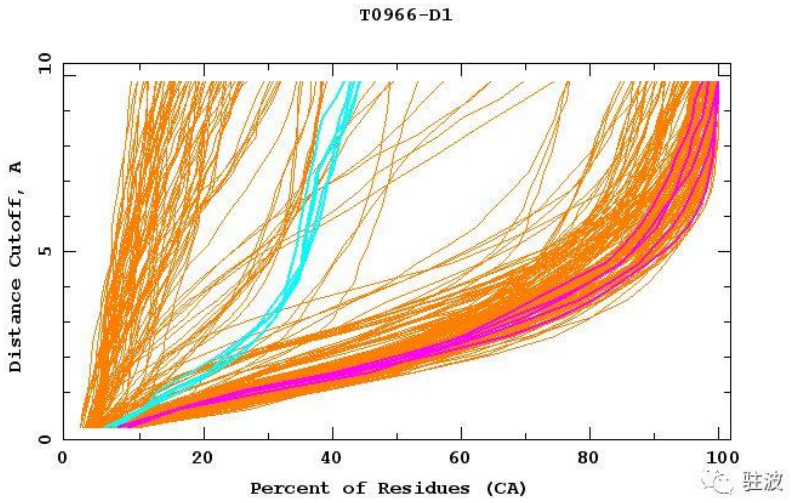
## DeepFold



# CASP

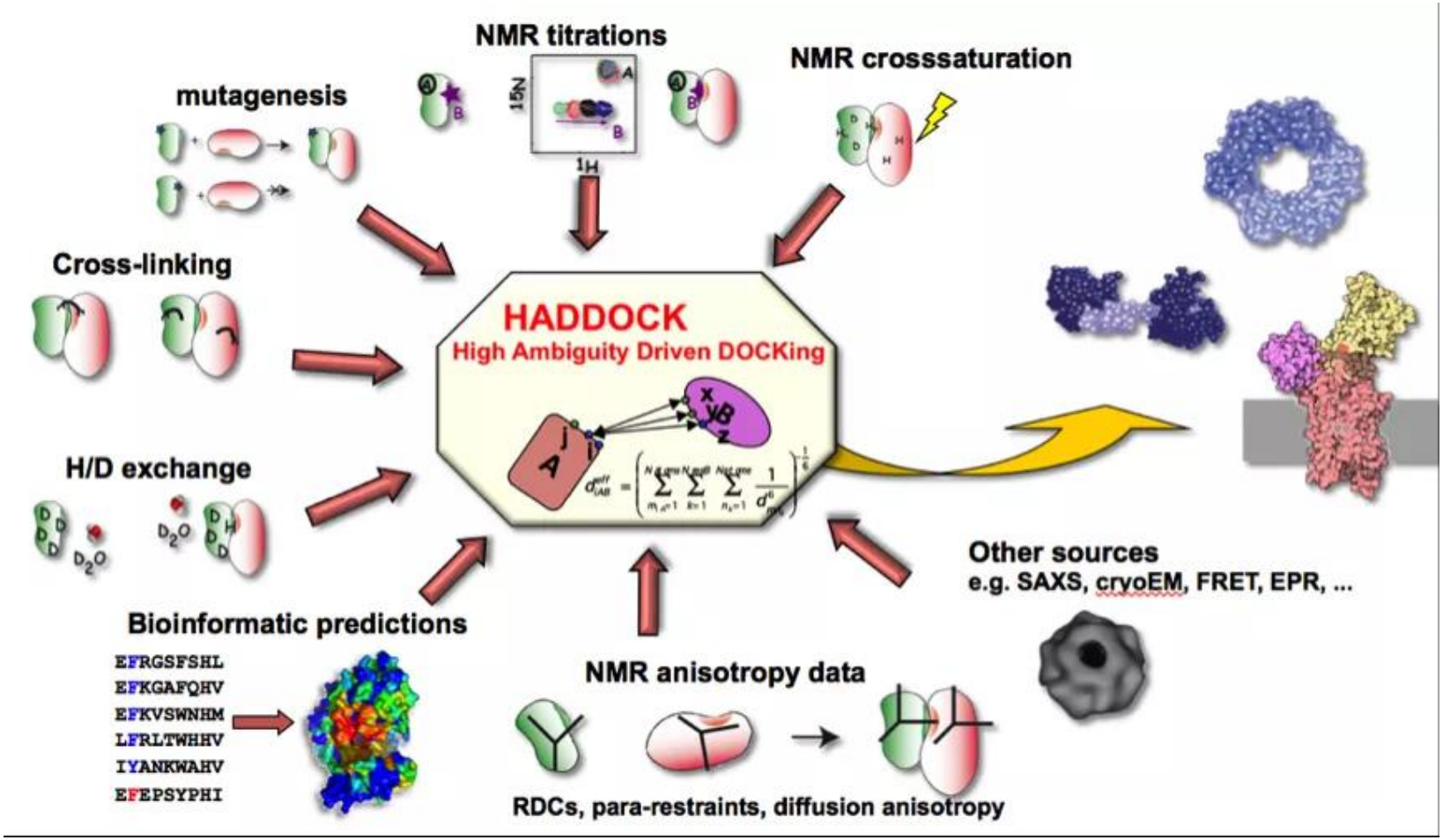
## (Critical Assessment of Techniques for Protein Structure Prediction)



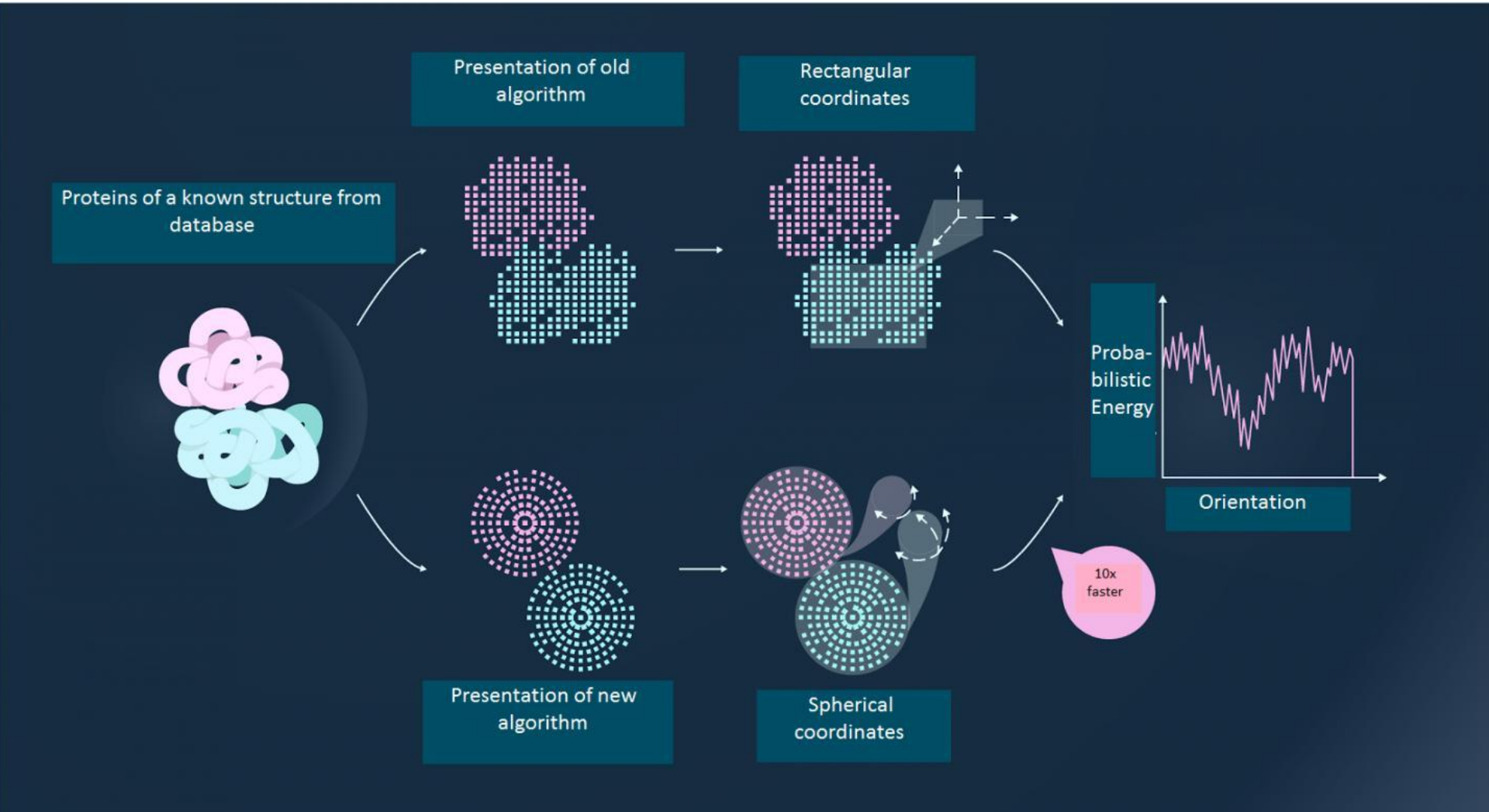



# CAPRI

## (Critical Assessment of PRediction of Interactions)

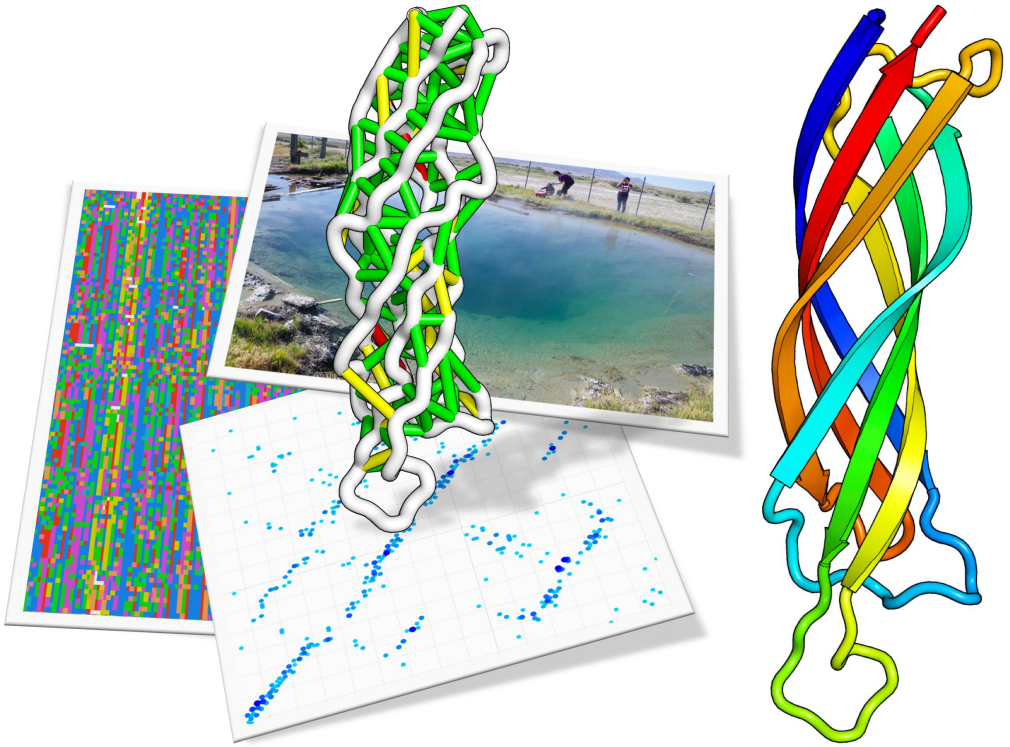


# CAPRI

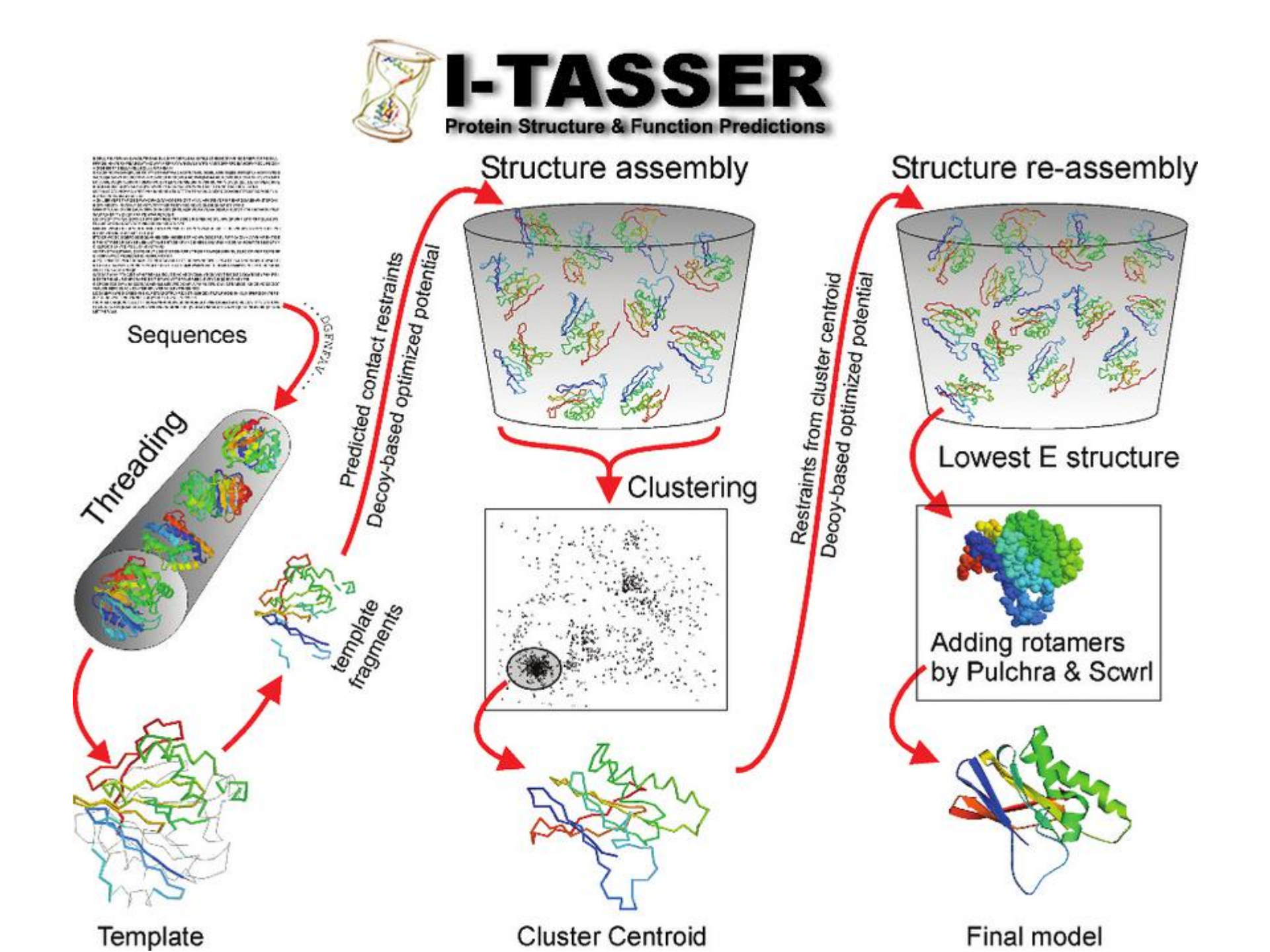

Presentation of old algorithm
Rectangular coordinates
Proteins of a known structure from database
Proba-bilistic Energy
Orientation
10x faster
Presentation of new algorithm
Spherical coordinates

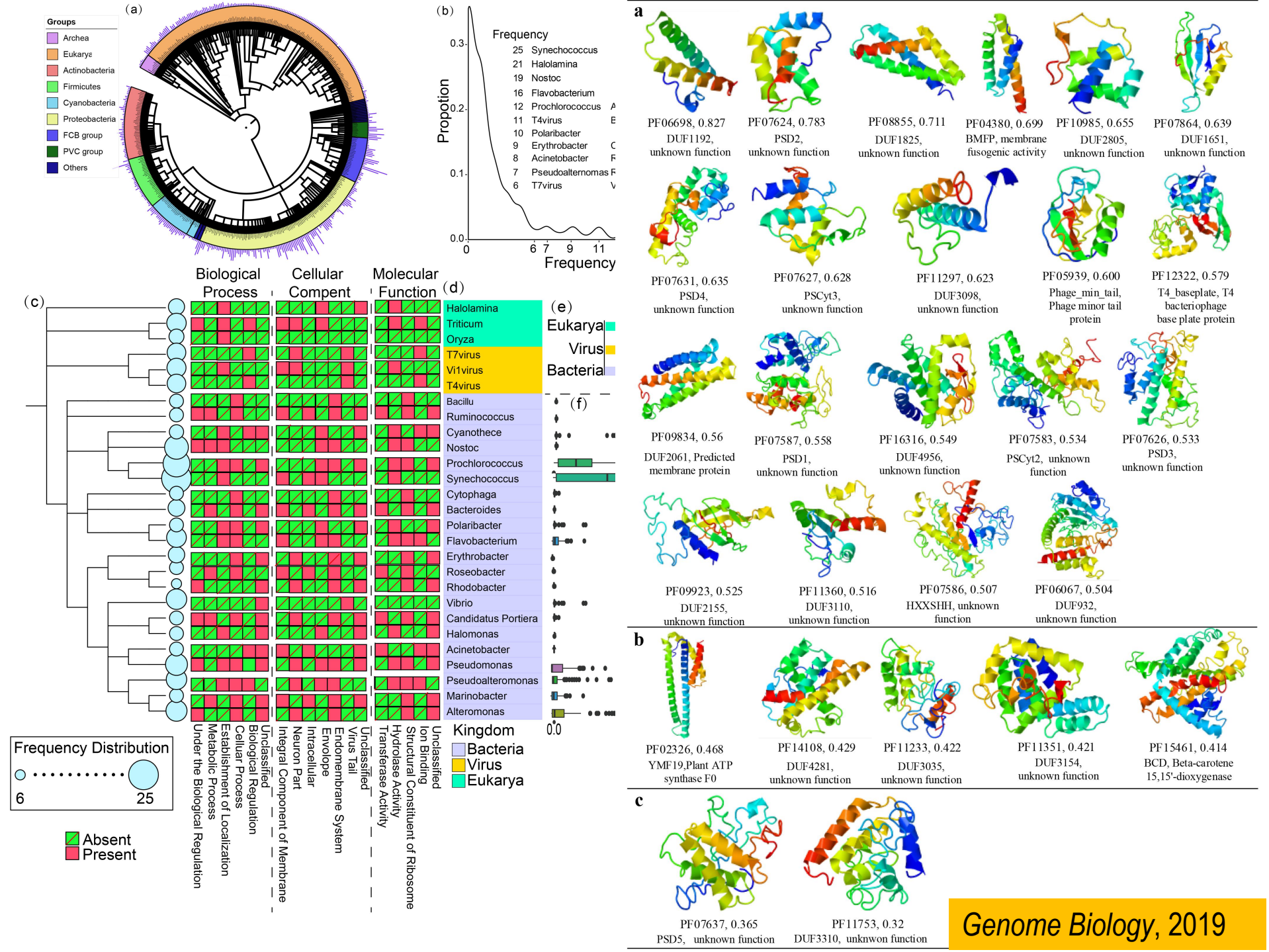
Rosetta Commons

The hub for Rosetta modeling software

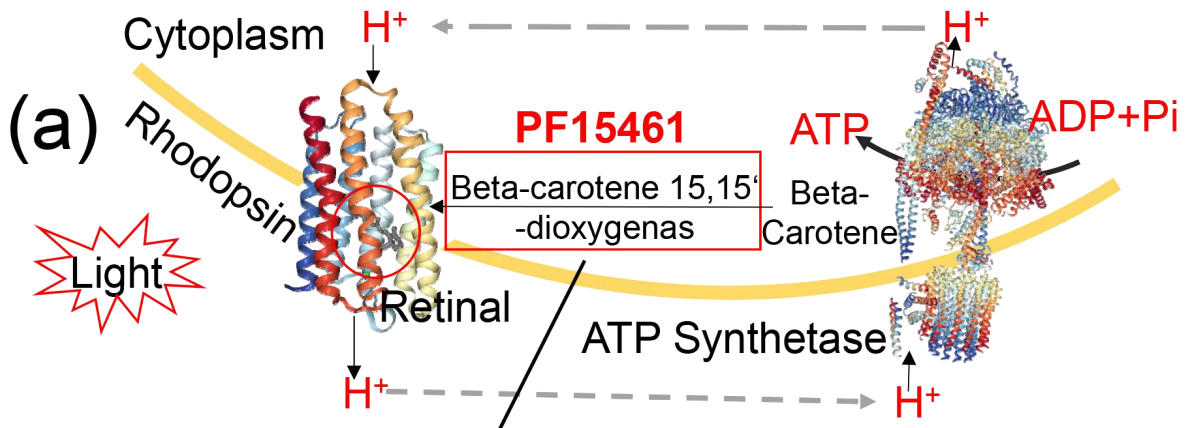


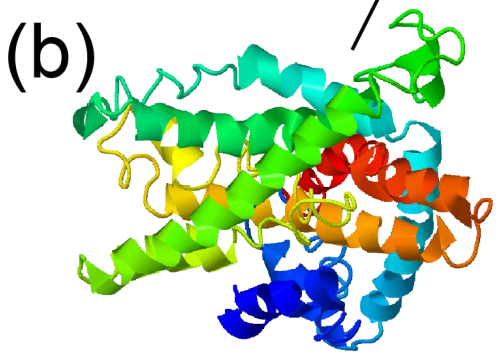
Top: Researchers gathering samples from Great Boiling Spring in Nevada. Left: a snapshot of aligned metagenomic sequences. Each row is a different sequence (the different colors are the different amino acid groups). Each position (or column) is compared to all other positions to detect patterns of co-evolution. Bottom: the strength of the top co-evolving residues is shown as blue dots, these are also shown as colored lines on the structure above. The goal is to make a structure that makes as many of these contacts as possible. Right: a cartoon of the protein structure predicted. The protein domain shown is from Pfam DUF3794, this domain is part of a Spore coat assembly protein SafA. (Image of Great Boiling Spring by Brian Hedlund, UNLV. Protein structure and composite image by Sergey Ovchinnikov, UW)


I-TASSER
Protein Structure & Function Predictions
Sequences
Threading
template fragments
Predicted contact restraints
Decoy-based optimized potential
Structure assembly
Clustering
Restraints from cluster centroid
Decoy-based optimized potential
Structure re-assembly
Lowest E structure
Adding rotamers by Pulchra & Scwrl
Template
Cluster Centroid
Final model




*Genome Biology*, 2019

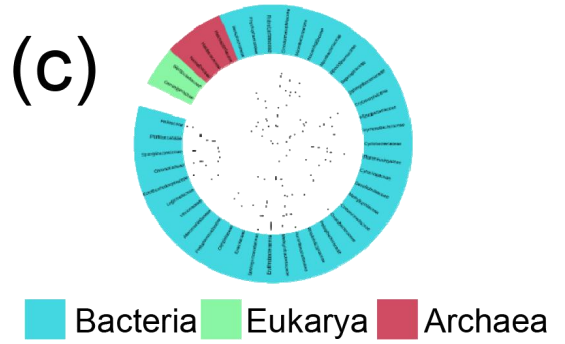


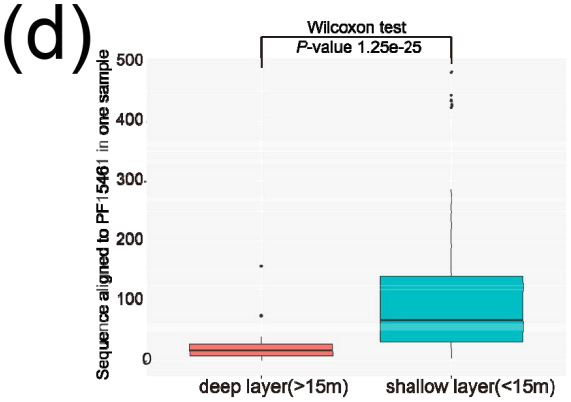




**Predicted structure of PF15461**

318 Amino Acid
Beta-carotene 15,15'-dioxygenase
369 sequence→14,353 sequence

Predicted Function:
Cellular Component: Respiratory Chain
Biological Process:
Single-organism Metabolic Process
Molecular Function:Oxidoreductase Activity

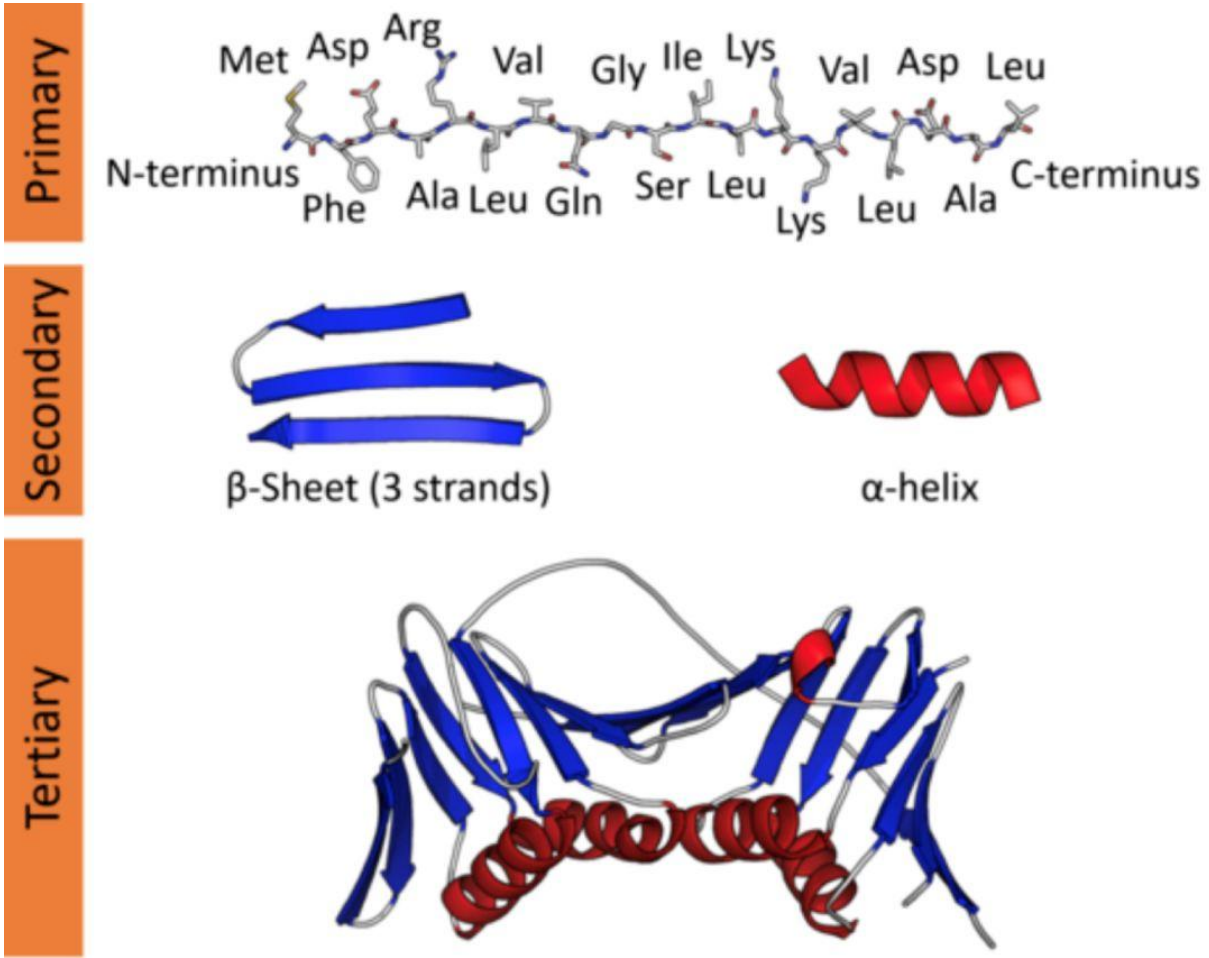



*Genome Biology*, 2019

# Protein 3D structure

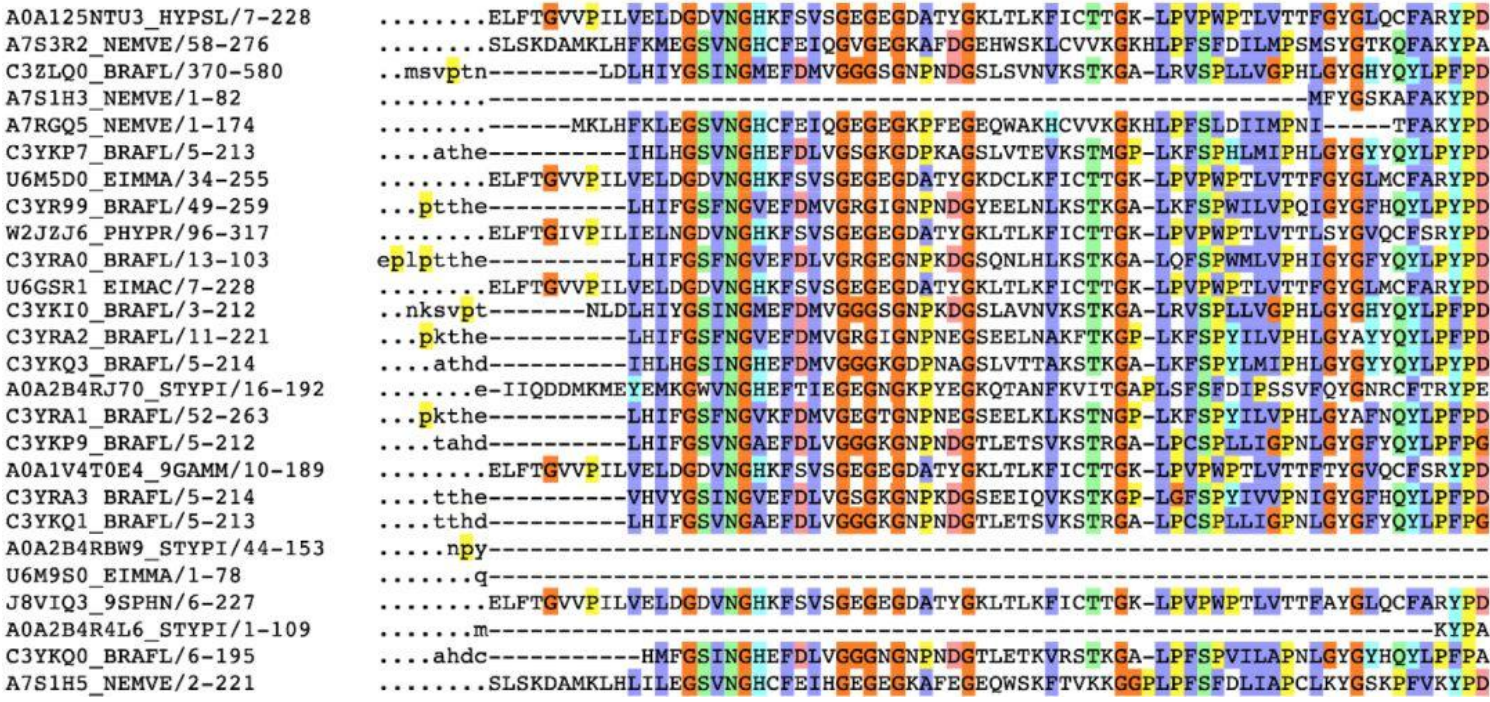
基本原理



# Protein 3D structure

## 结构和序列

### 其中颜色渐变表示序列中从头到尾的索引



# Protein 3D structure

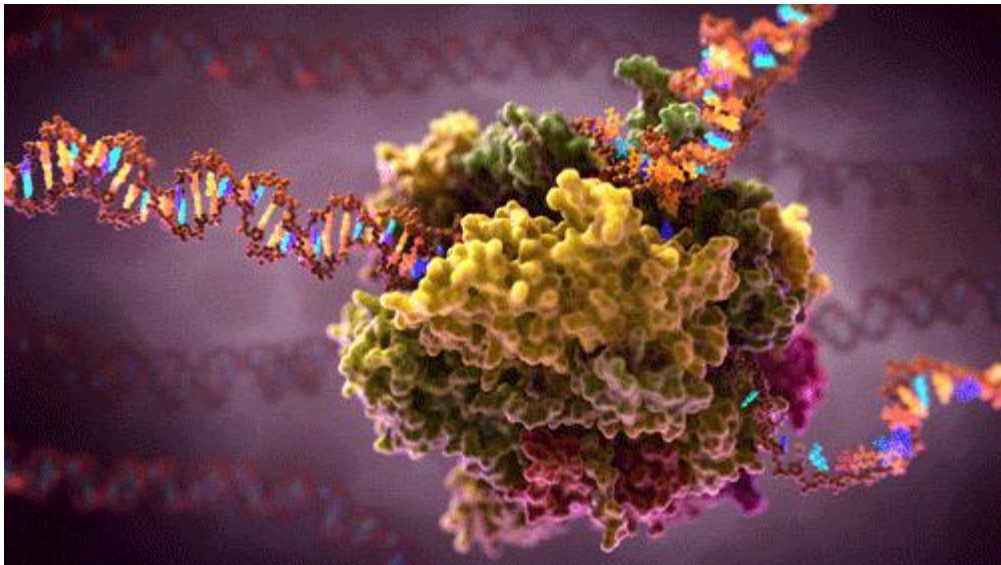
大数据问题

# Protein 3D structure

眼见为实

Overview

Ribosome



# Protein 3D structure

眼见为实

ATP

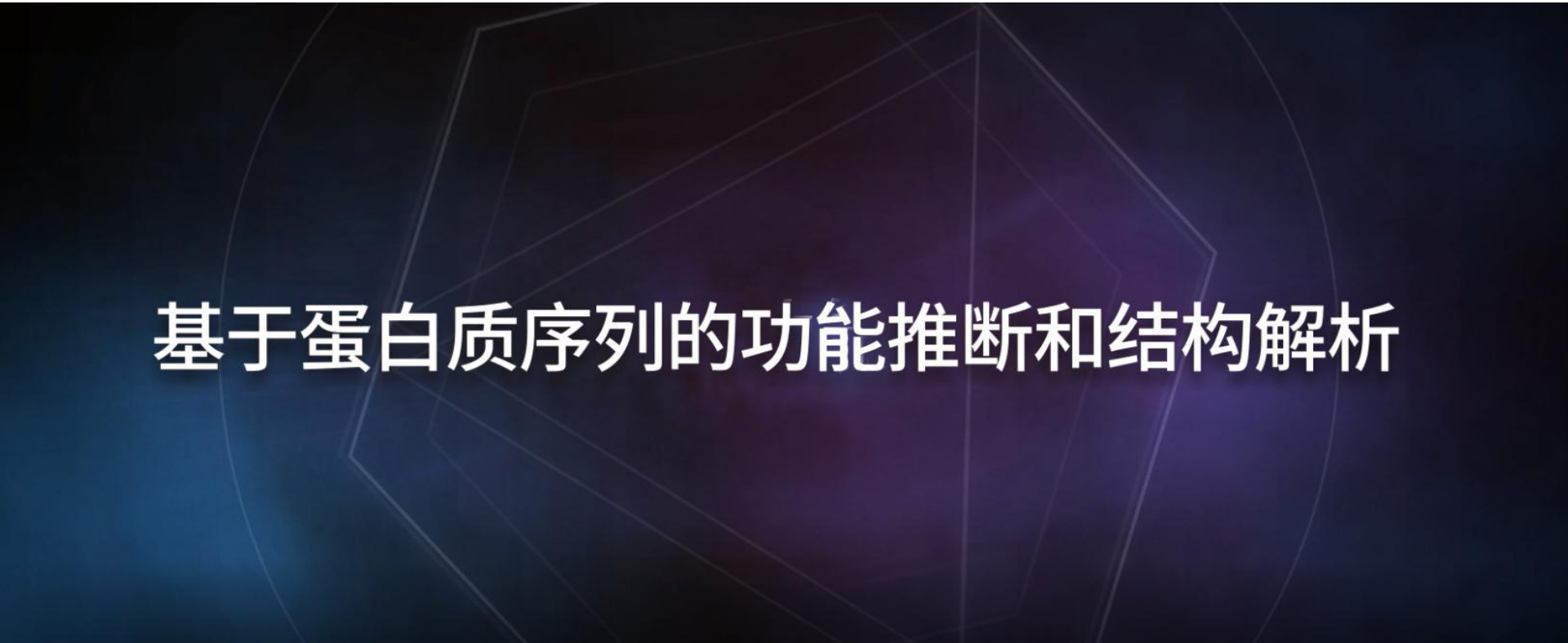
Transportation

# Protein 3D structure

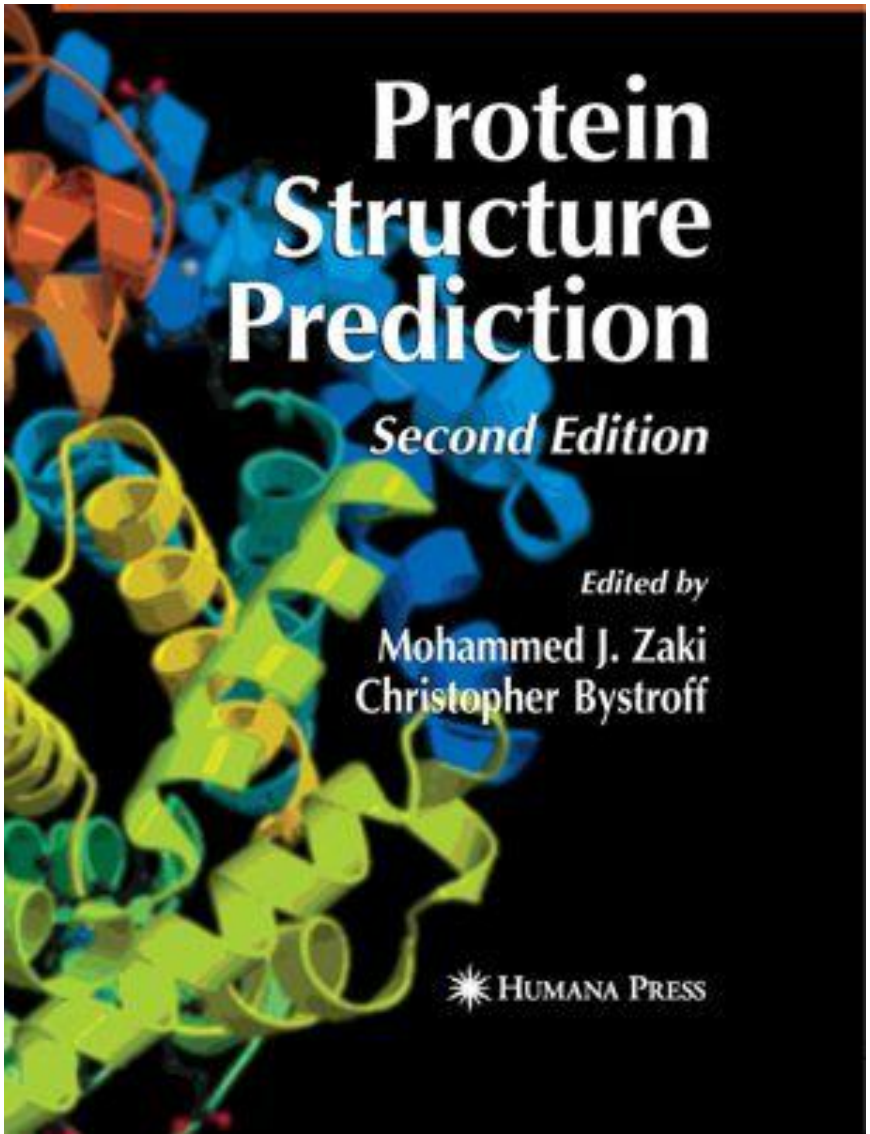
眼见为实

# Protein 3D structure

眼见为实



# References


Protein
Structure
Prediction
Second Edition
Edited by
Mohammed J. Zaki
Christopher Bystroff
HUMANA PRESS