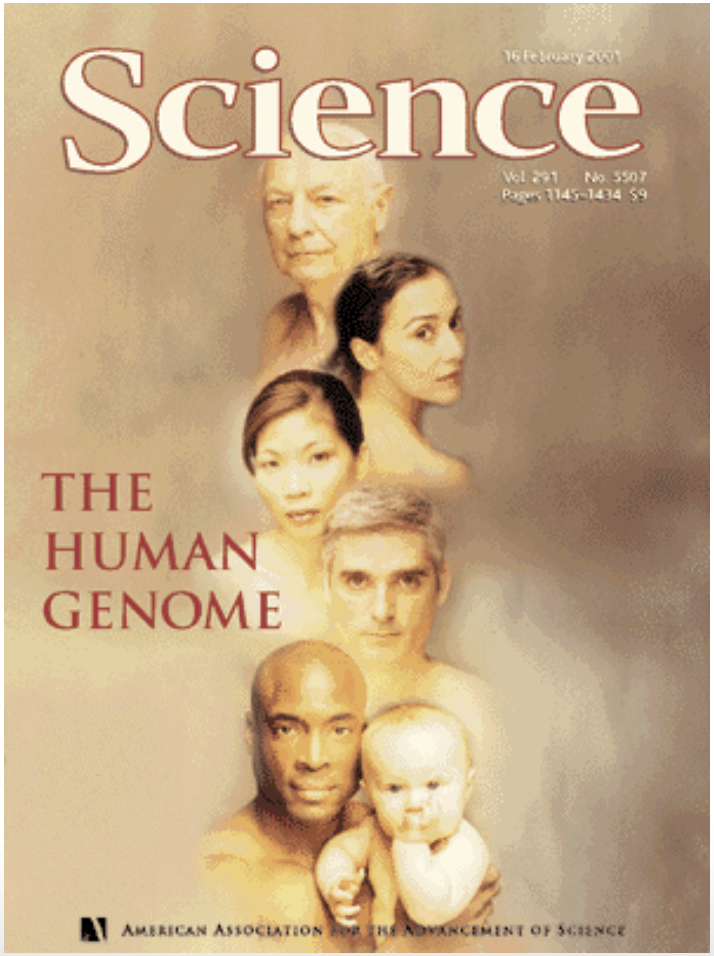
生物信息学概论

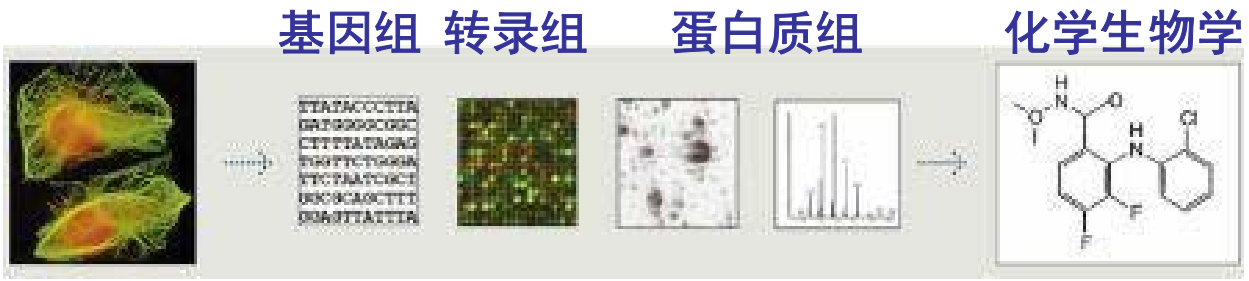
第八章 基因组分析

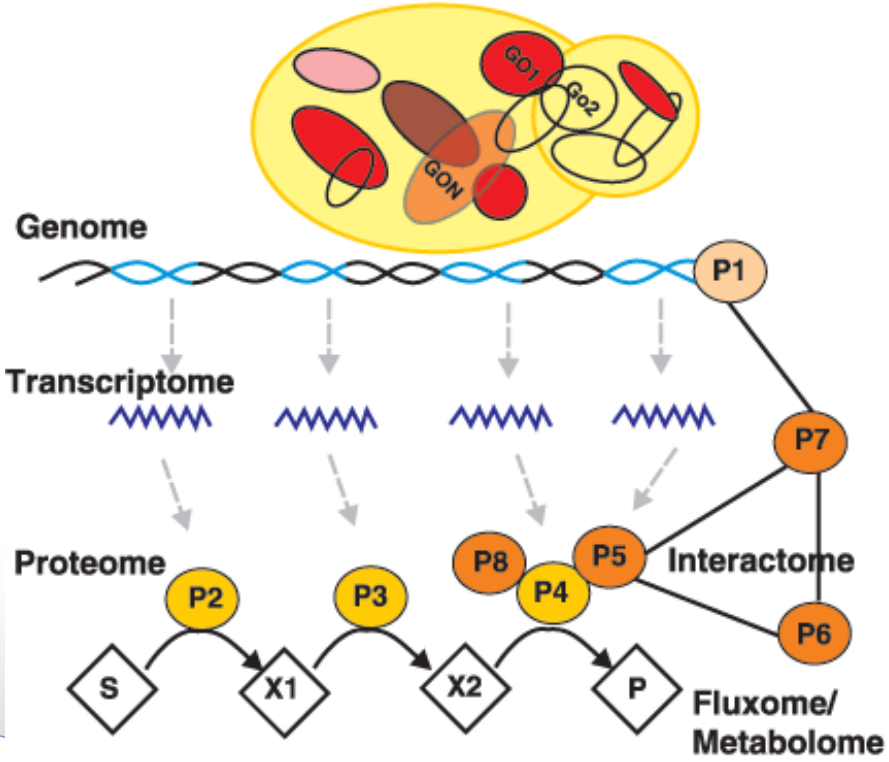
人类基因组计划


Science
16 February 2001
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
American Association for the Advancement of Science


Science
24 March 2000
Vol. 287 No. 5461
Pages 2105–2364 $8
The Drosophila Genome

基因组、转录组和蛋白质组



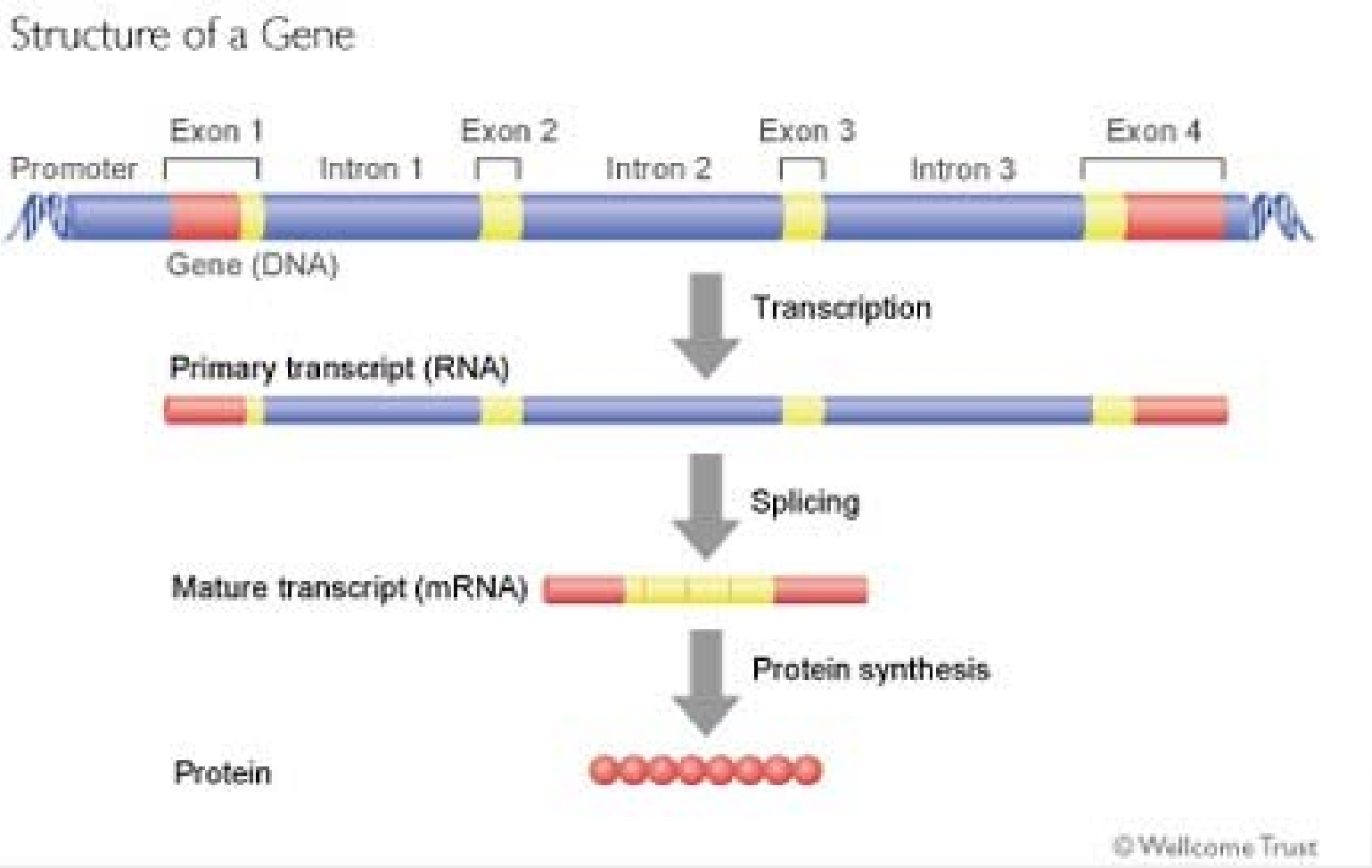


a. GO	Global functional annotation
Interaction	**Level of information**
b. TF	Transcriptional regulation
c. Protein	Flow of information through PPI
d. Complexes	Agglomerative functional modules
e. Metabolites	Metabolic network regulation

基因组的结构与内容

- 基因的结构
- mRNA：可变剪接
- 蛋白质：翻译后修饰
- 相互作用网络：基因、蛋白质、小分子之间的相互作用
- 非编码区
  - 功能元件: 转录因子结合位点；启动子…
  - 长非编码RNA、环形RNA、MicroRNA
  - RNA修饰
  - 转座子
  - 重复片段
  - 伪基因（Pseudogene）

基因的结构


Structure of a Gene
Exon 1
Exon 2
Exon 3
Exon 4
Promoter
Intron 1
Intron 2
Intron 3
Gene (DNA)
Transcription
Primary transcript (RNA)
Splicing
Mature transcript (mRNA)
Protein synthesis
Protein
© Wellcome Trust

基因组大小 & 基因数

Sequenced genomes vary from 470-30,000 genes

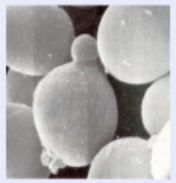
Species	Genome (Mb)	Genes	Lethal loci
Mycoplasma genitalium	0.58	470	~300
Rickettsia prowazekii	1.11	834	
Haemophilus influenzae	1.83	1,743	
Methanococcus jannaschi	1.66	1,738	
B subtilis	4.2	4,100	
E coli	4.6	4,288	1,800
S. cerevisiae	13.5	6,034	1,090
S. pombe	12.5	4,929	
A. thaliana	119	25,498	
O. sativa (rice)	466	~30,000	
D. melanogaster	165	13,601	3,100
C. elegans	97	18,424	
H. sapiens	3,300	~30,000	

Minimum gene numbers range from 500 to 30,000

500 genes
Extracellular (parasitic) bacterium

1,500 genes
Free-living bacterium

5,000 genes
Unicellular eukaryote



13,000 genes
Multicellular eukaryote

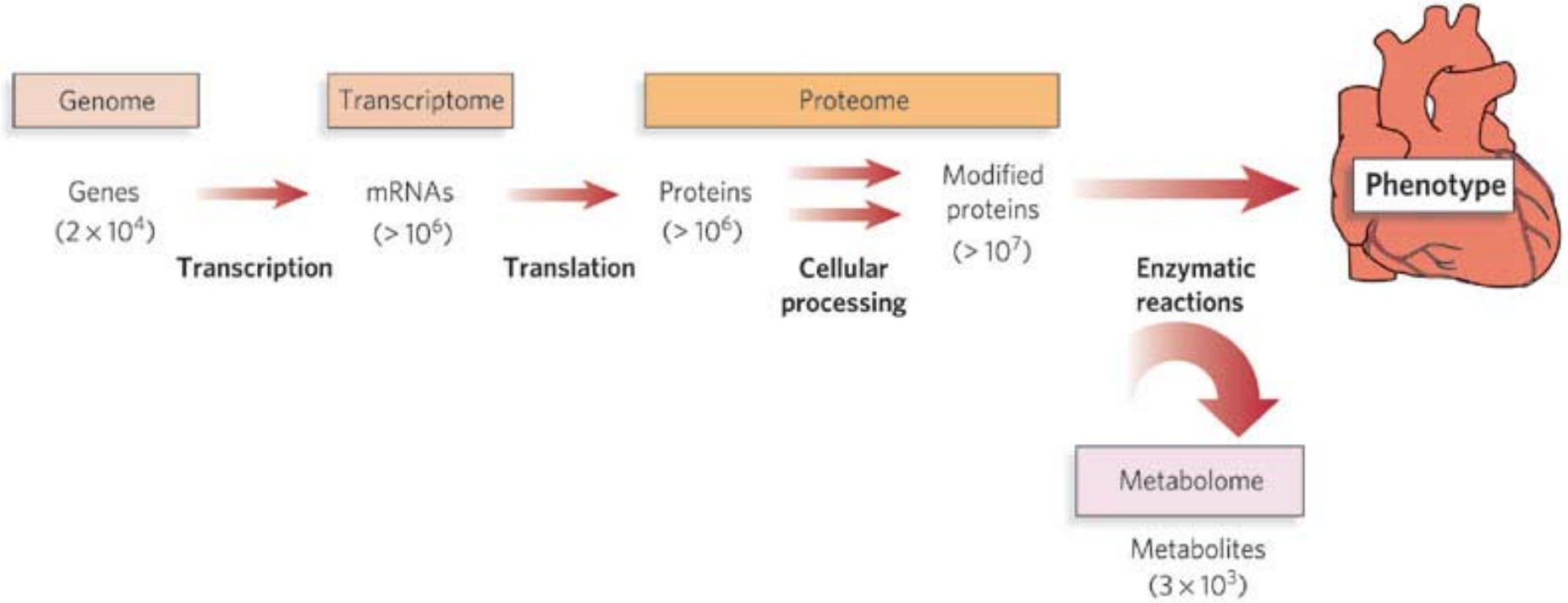
25,000 genes
Higher plants

30,000 genes
Mammals

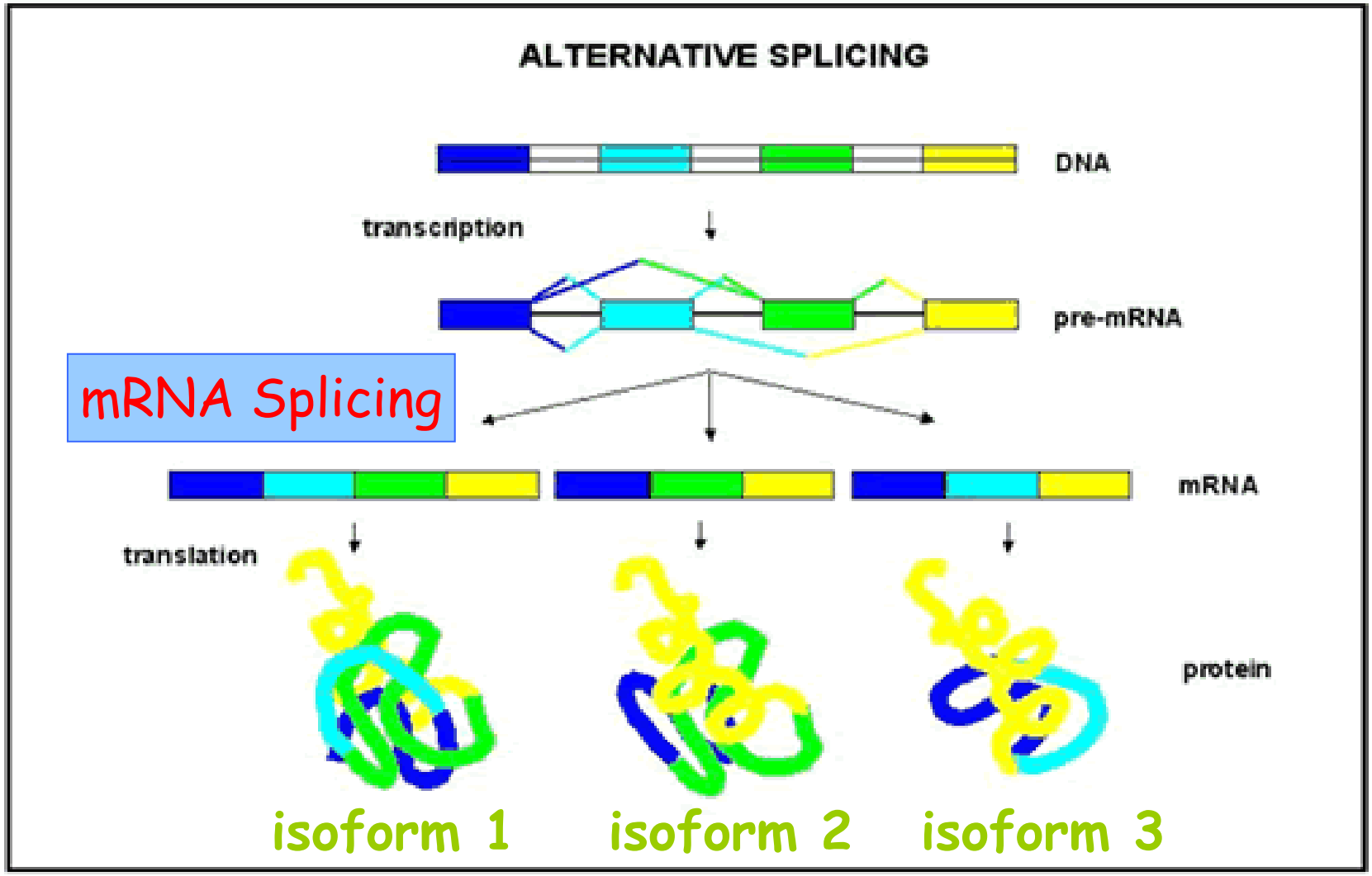
基因数量 -> 生物复杂性？

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

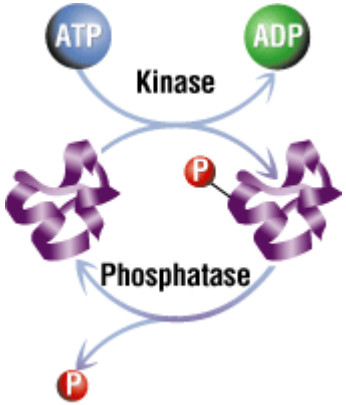
基因型到表型


Genome
Transcriptome
Proteome
Genes
(2 × 10^4)
Transcription
mRNAs
(> 10^6)
Translation
Proteins
(> 10^6)
Cellular
processing
Modified
proteins
(> 10^7)
Enzymatic
reactions
Phenotype
Metabolome
Metabolites
(3 × 10^3)

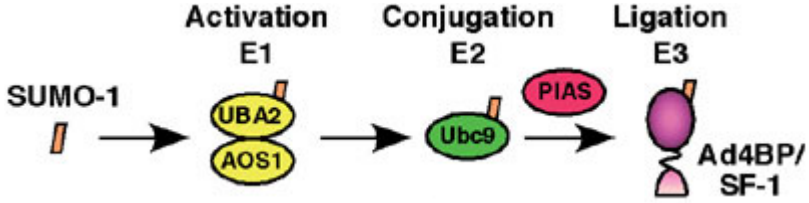
转录后层面：mRNA Splicing



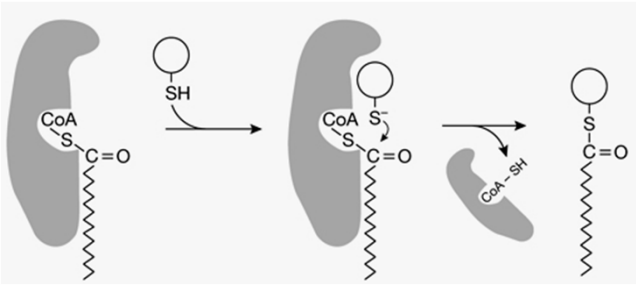
蛋白质层面：翻译后修饰



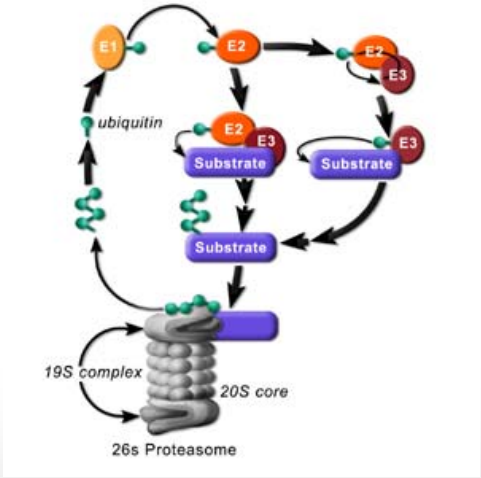
Phosphorylation



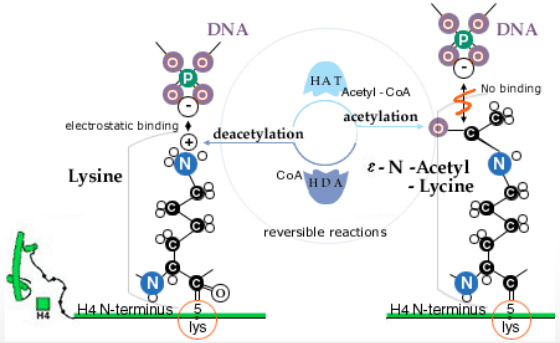
Sumoylation



Palmitoylation

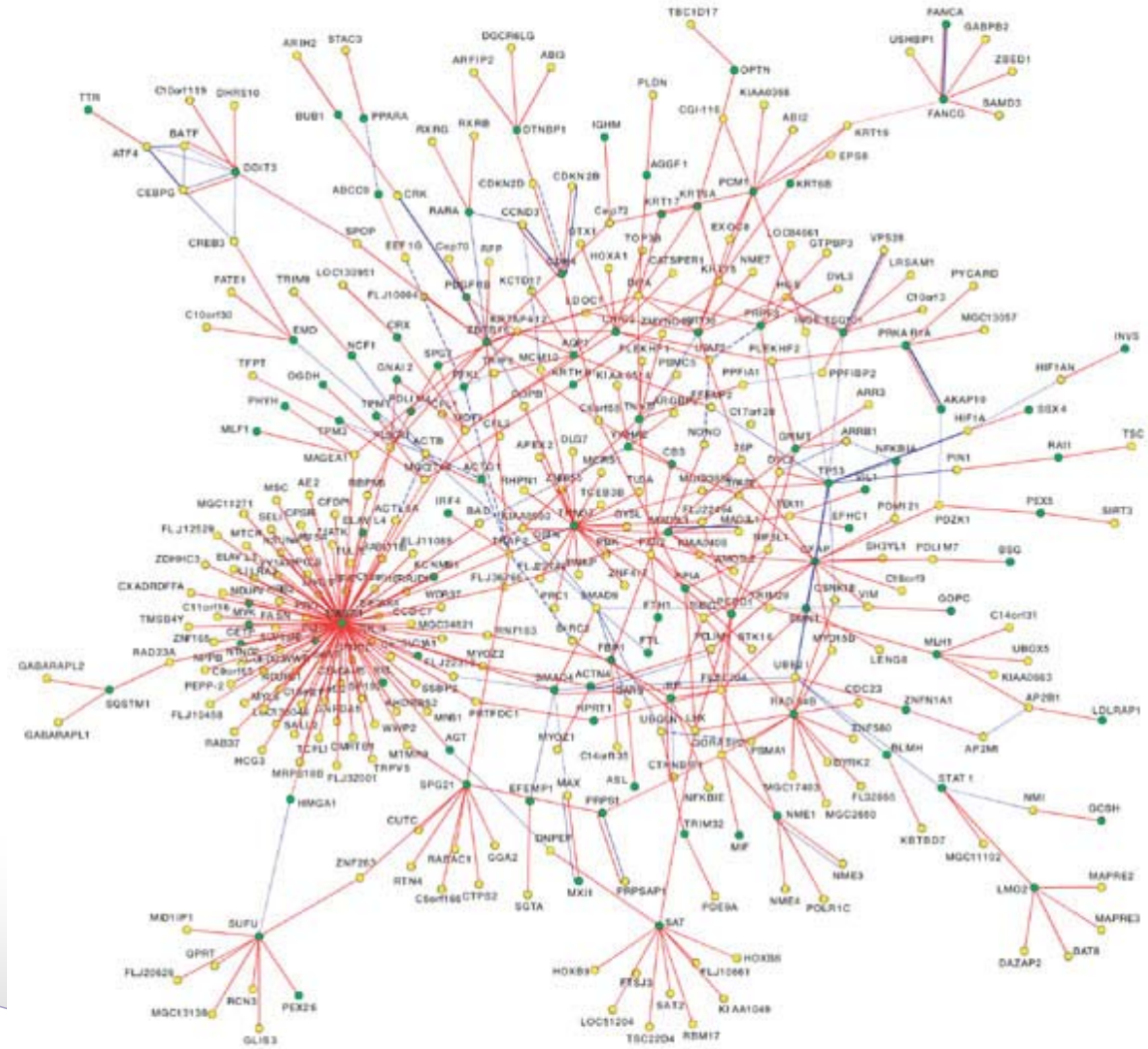


Ubiquitination



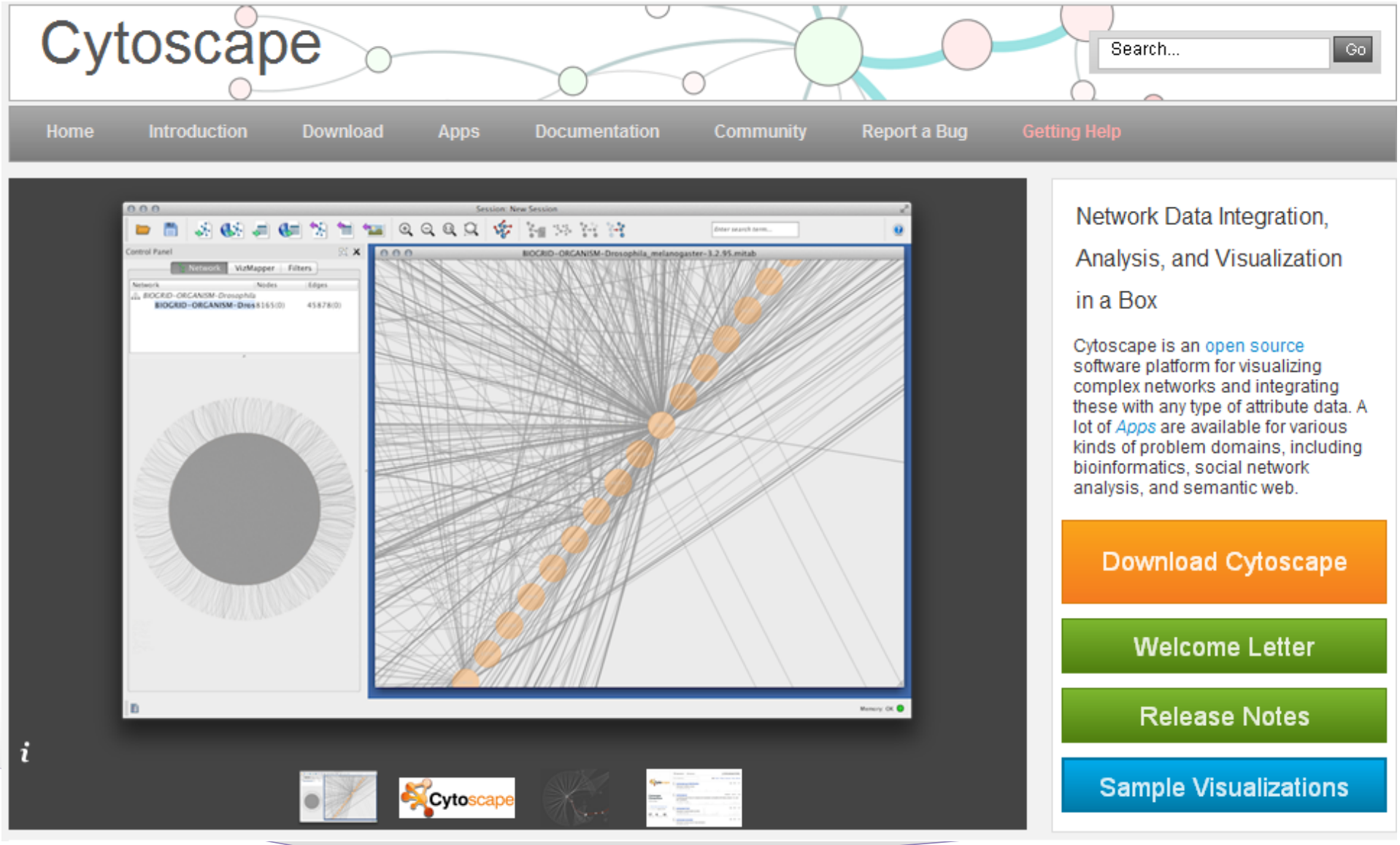
Acetylation

相互作用网络

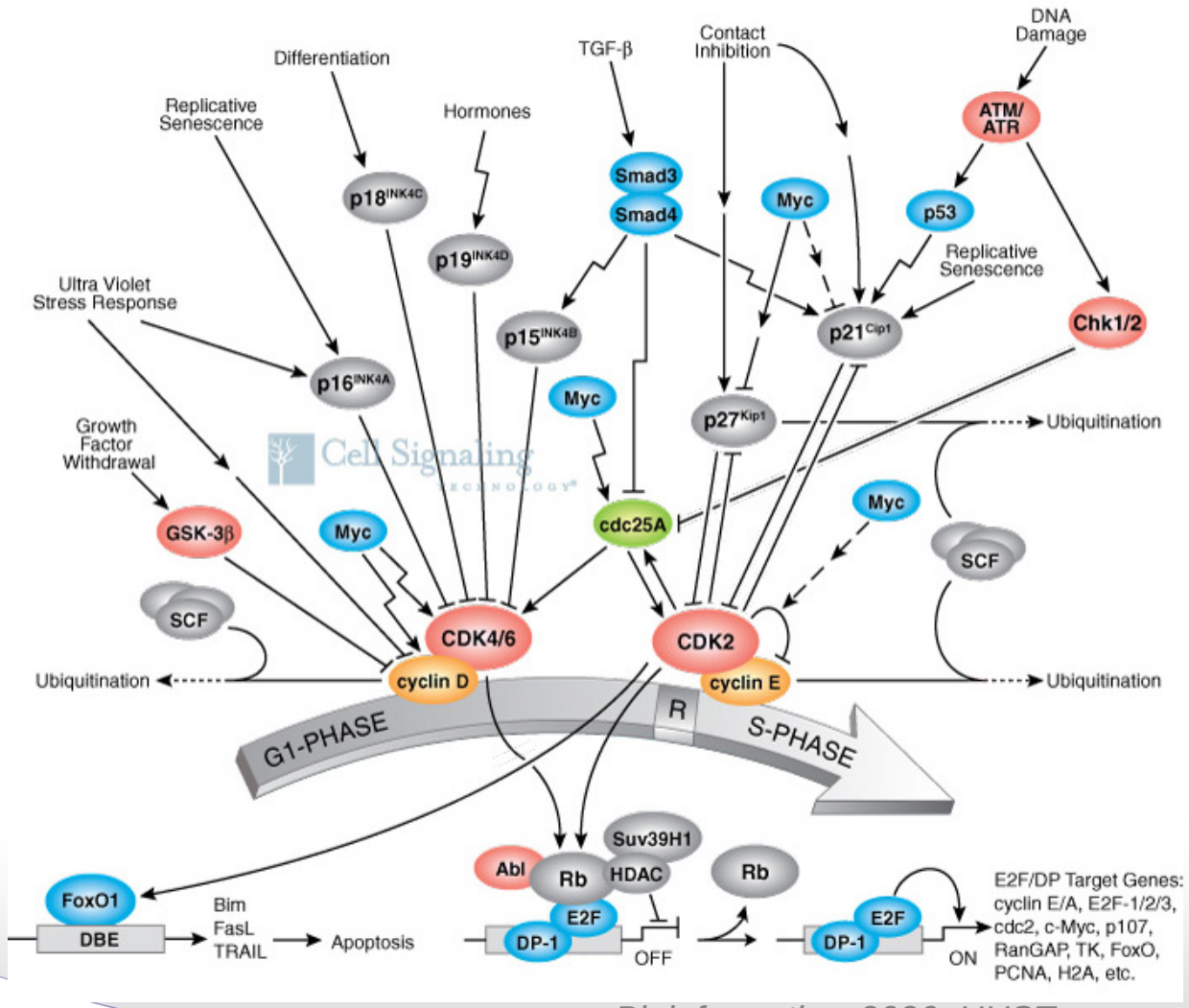


蛋白质-蛋白质相互作用网络

Cytoscape：网络构建和分析工具



细胞信号通路



G1/S检验点：

有调控方向

KEGG PATHWAY Database

KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions, and relations

KEGG2 PATHWAY BRITE MODULE DISEASE DRUG KO GENOME GENES LIGAND DBGET

Select prefix: map | Organism

Enter keywords: | Go | Help

Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps (see new maps and update history) representing our knowledge on the molecular interaction and reaction networks for:

1. **Metabolism**
   Global map Carbohydrate Energy Lipid Nucleotide Amino acid Other amino acid Glycan Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human Diseases**

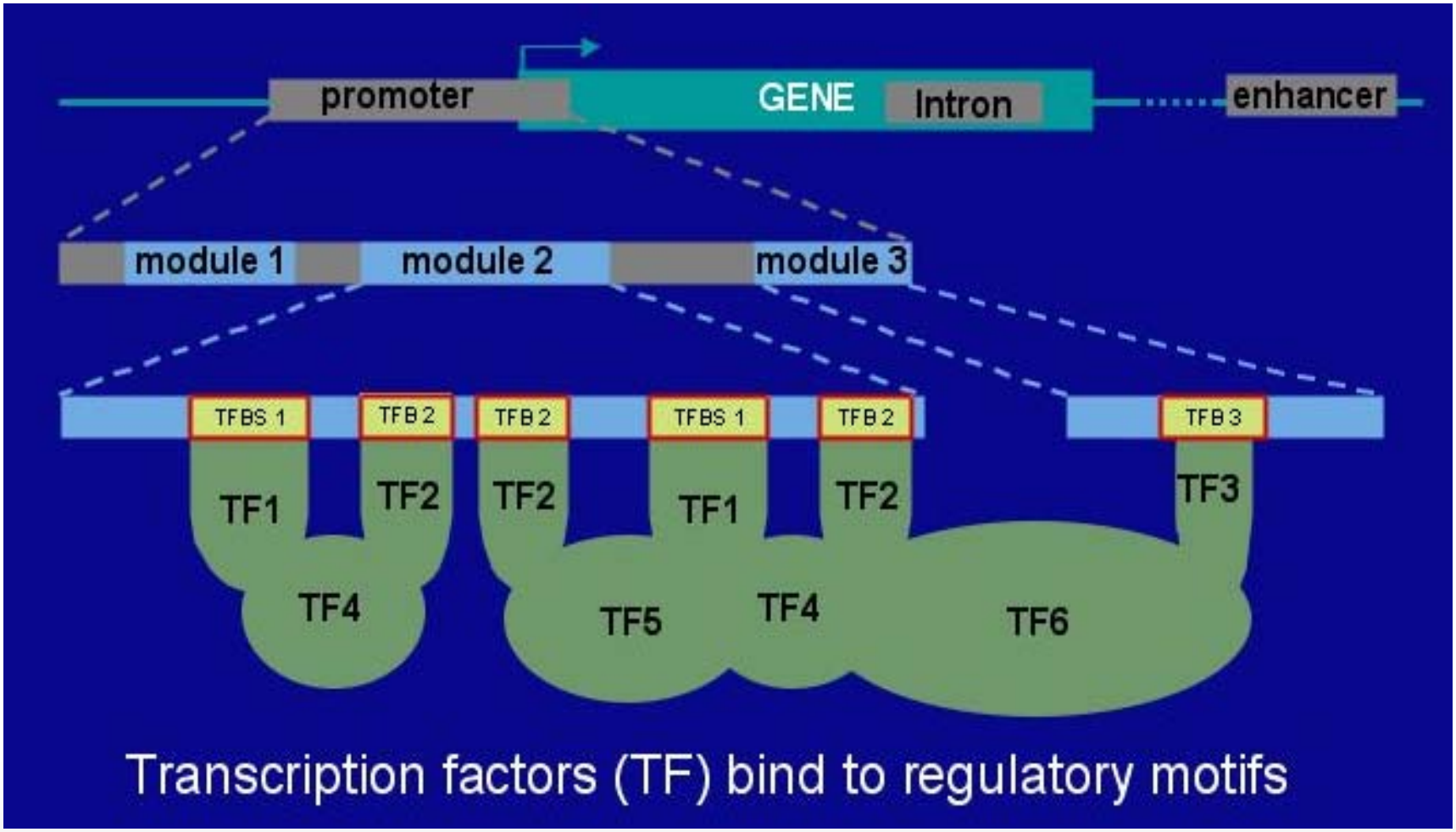
and also on the structure relationships (KEGG drug structure maps) in:

7. **Drug Development**

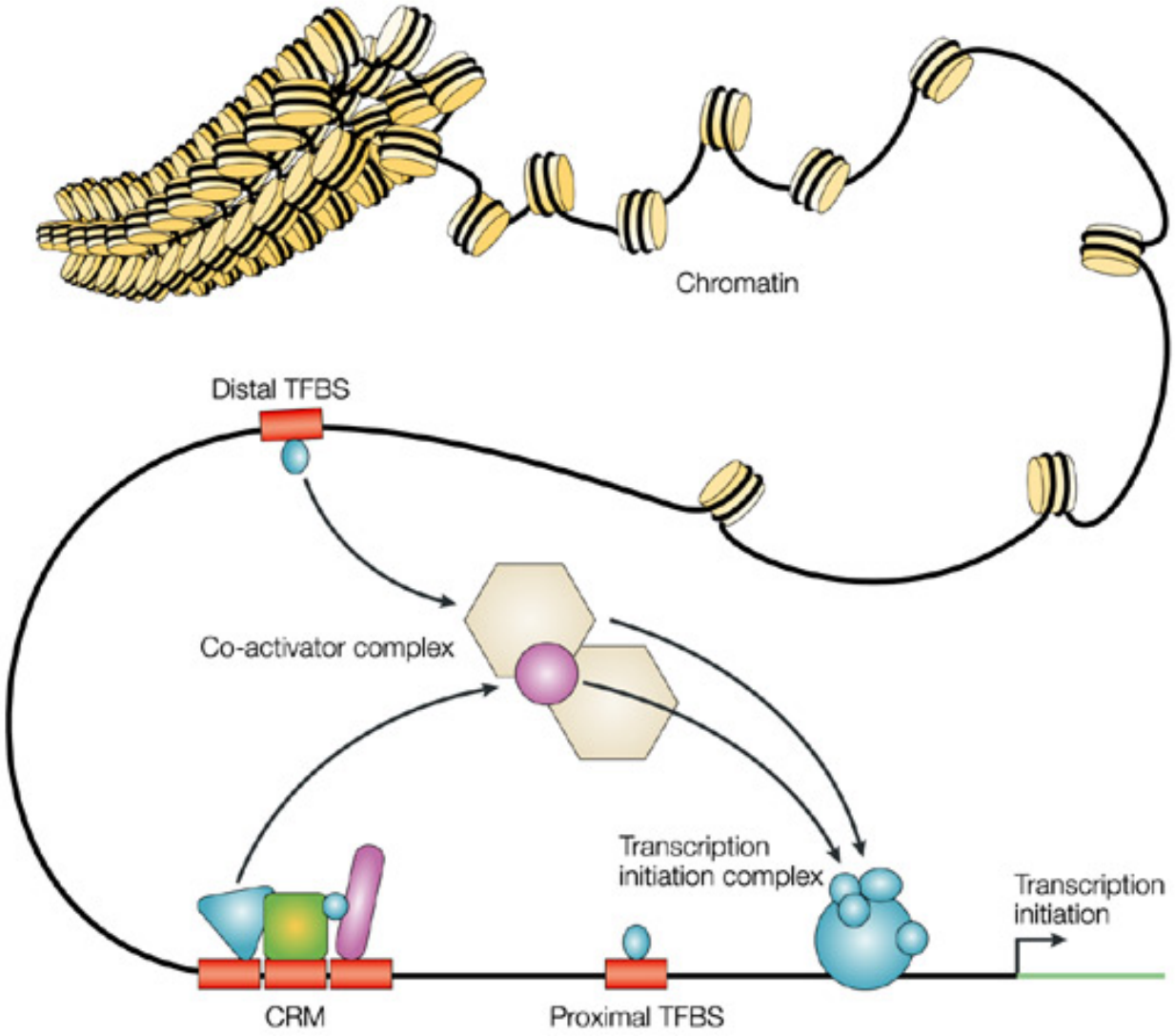
非编码区

- 功能元件: 转录因子结合位点；启动子…
- Non-coding RNA: MicroRNA
- 转座子
- 重复片段
- 伪基因（Pseudogene）

功能元件：启动子


promoter
GENE
Intron
enhancer
module 1
module 2
module 3
TFBS 1
TFB 2
TFB 2
TFBS 1
TFB 2
TFB 3
TF1
TF2
TF2
TF1
TF2
TF3
TF4
TF5
TF4
TF6
Transcription factors (TF) bind to regulatory motifs

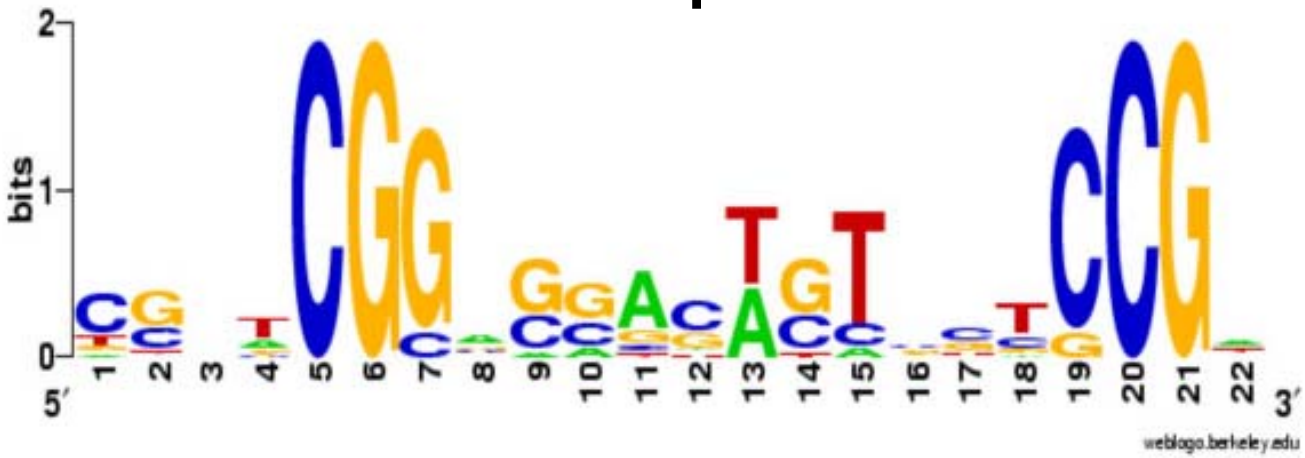
转录因子结合位点



CRM: cis-regulatory modules

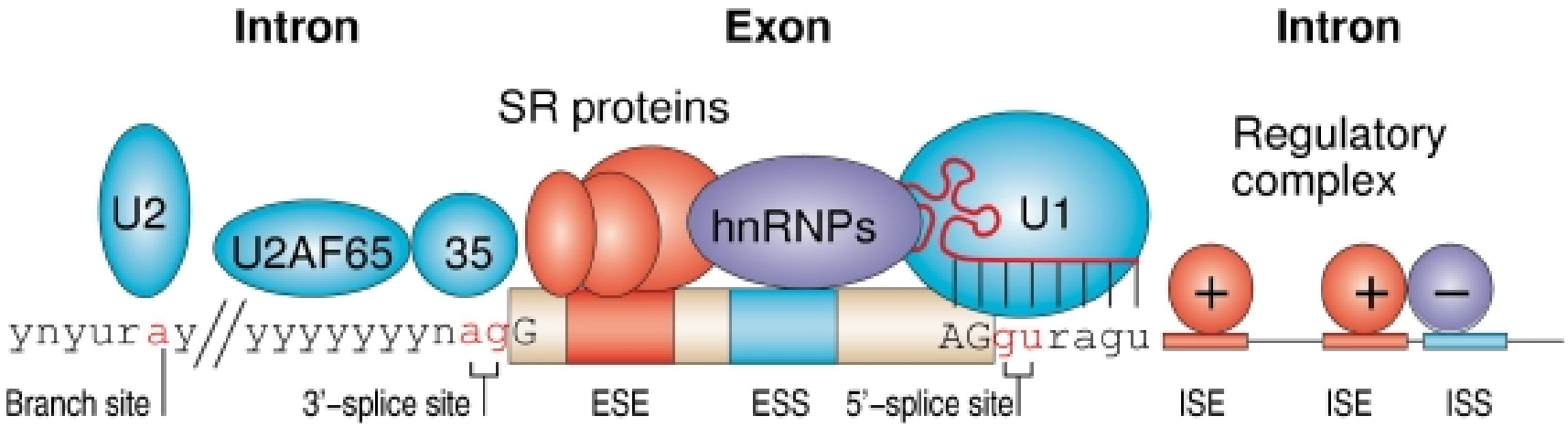
Gal4p and Kruppel

Gal4p



Kruppel

其他功能元件

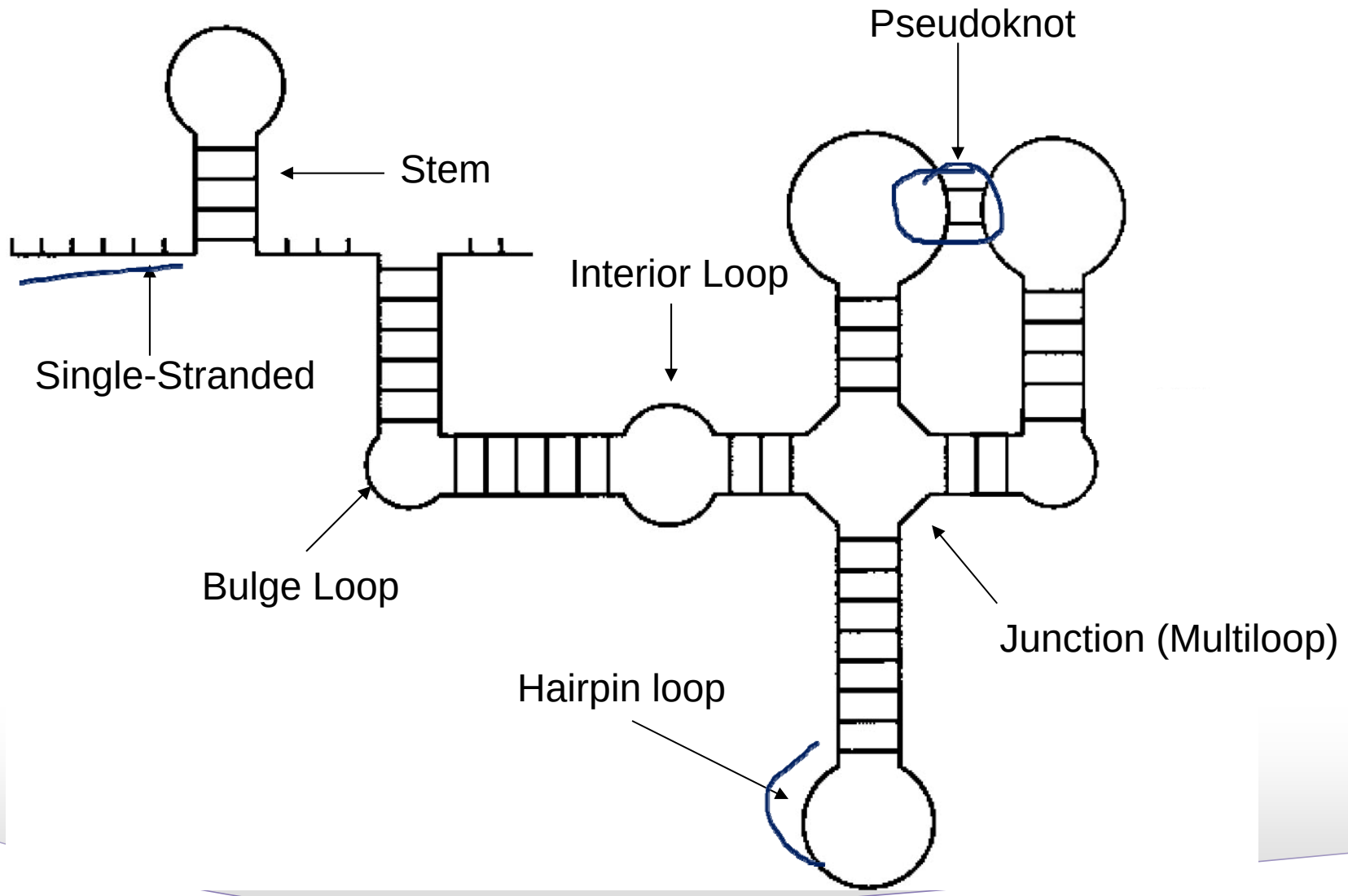


- Exon splicing enhancer (ESE) and silencer (ESS)
- Intron splicing enhancer (ISE) and silencer (ISS)

非编码RNA

- 不翻译成蛋白质，具有重要的调控功能
- 分类：
  - transfer RNA (tRNA)
  - ribosomal RNA (rRNA)
  - snoRNAs
  - microRNAs
  - siRNAs
  - piRNAs: 与piwi相互作用的RNA
  - long ncRNAs: Xist
  - circRNAs

RNA二级结构


Pseudoknot
Stem
Interior Loop
Single-Stranded
Bulge Loop
Junction (Multiloop)
Hairpin loop

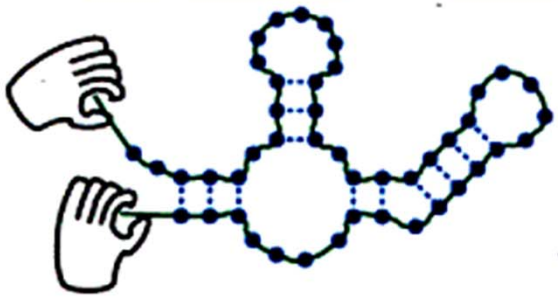
RNA二级结构的显示方式

- Grammatically correct string of parentheses

..(((.(((.......))).(((((.....)))).)).....)))

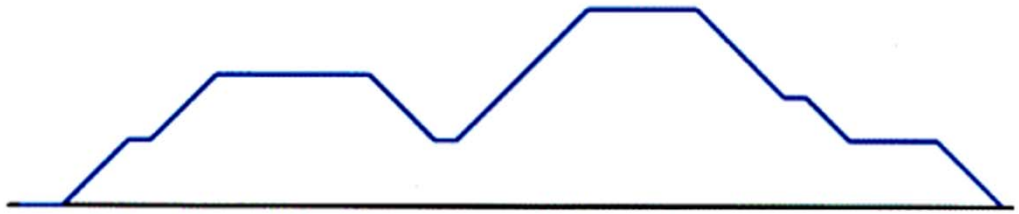
AGCTACGGAGCGATCTCCGAGCTTTCGAGAAAGCCTCTATTAGC

- Planar graph



- Arch diagram

- Mountain diagram



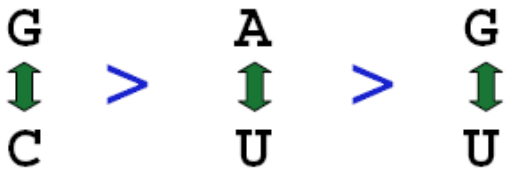
RNA的能量

- Doug Turner的能量法则

```
...CCAUUCAUAG-5'
   ||||||
5'...CGUGAGU..... 3'
```

```
          5' --> 3'
             UX
             AY
          3' <-- 5'

                X
Y      A       C       G       U
A      .       .       .    -1.30
C      .       .    -2.40      .
G      .    -2.10      .    -1.00
T   -0.90      .    -1.30      .
```

- Base pairing



- Base stacking

RNA的能量

A)

```
N p N p N p N p N p N p N
x   |   |   |   |   x   x
N p N p N p N p N p N p N
```

许多连续的配对

B)

```
N p N p N p N p N p N p N
x   |   |   x   |   |   x
N p N p N p N p N p N p N
```

存在内环 (loop)

C)

```
N p N p N p N p N p N p N
x   |   |   |   x   |   x
N p N p N p N p N p N p N
```

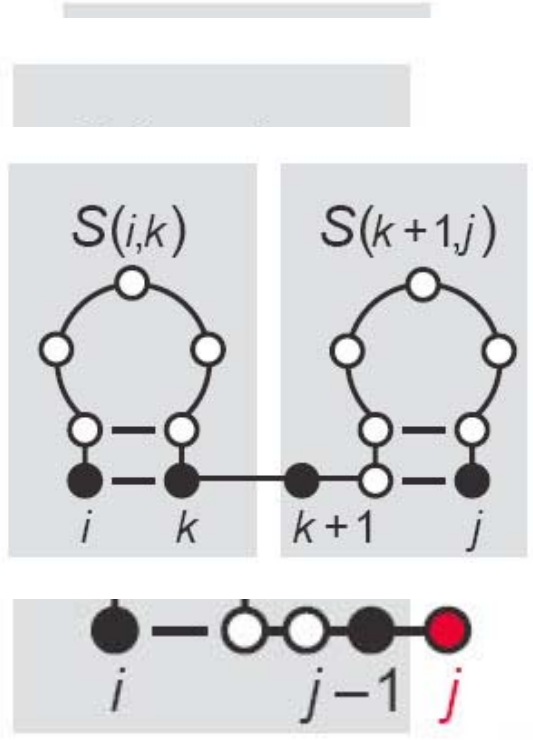
终端不稳定

- 稳定性A > B和C

动态规划算法：最大碱基配对

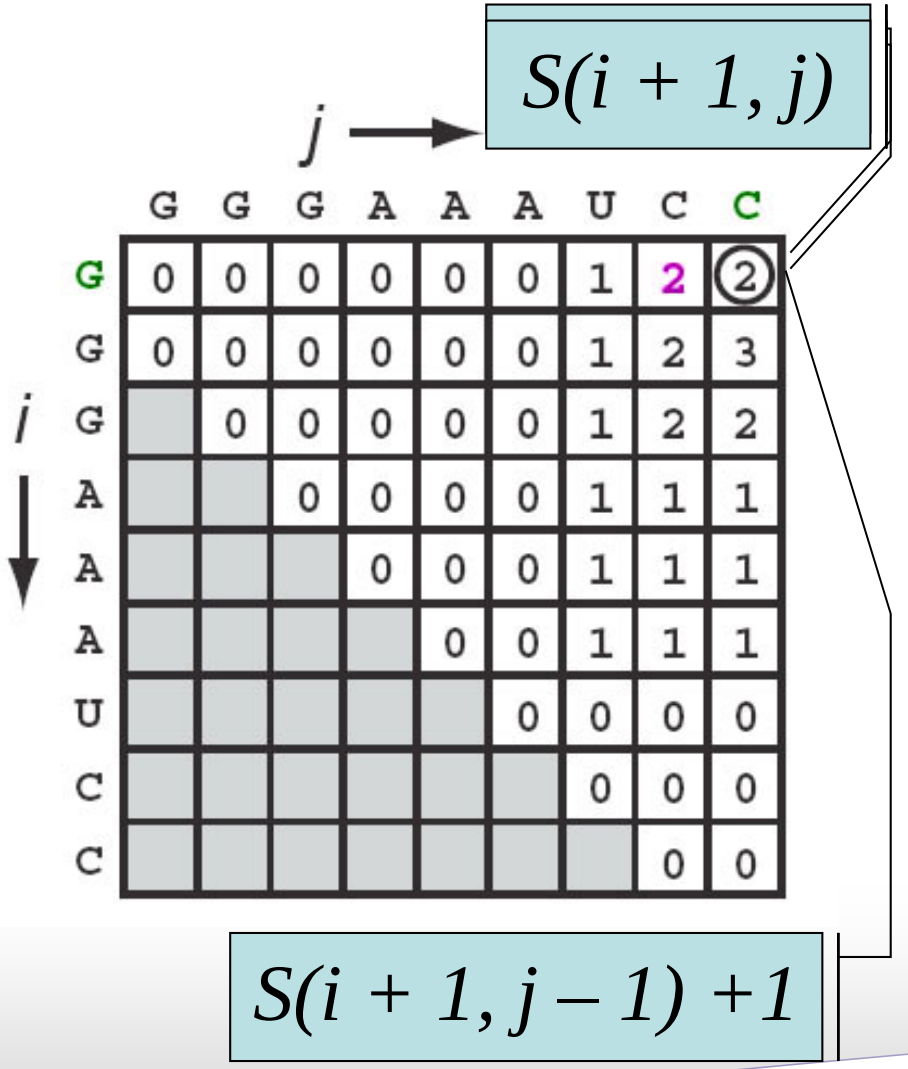
当前位置的匹配是否能保证最大的碱基匹配大化

$$S(i,j) = \max \begin{cases} S(i+1,j-1)+1 & [\text{if } i,j \text{ base pair}] \\ S(i+1,j) \\ S(i,j-1) \\ \max_{i<k<j} S(i,k) + S(k+1,j) \end{cases}$$



动态规划算法：最大碱基配对

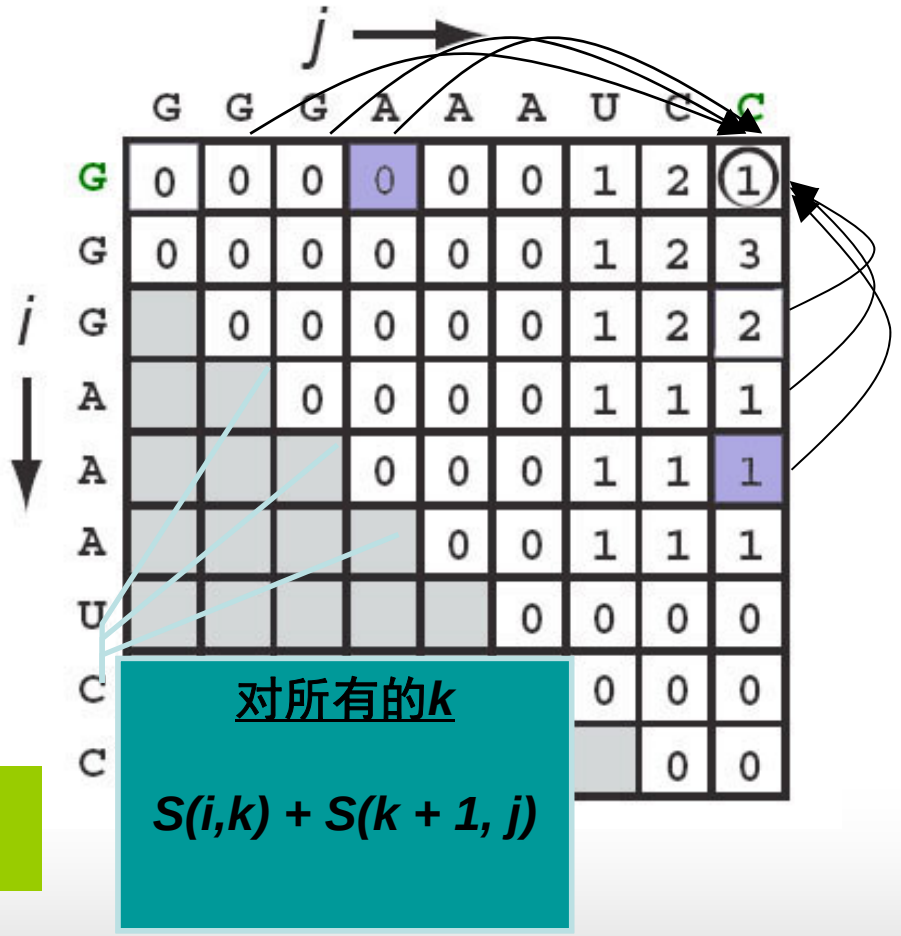
- 比对方法
  - RNA链与自身相比
  - 若存在碱基配对则增加分数
- 计算分值
- 可最后加入分歧的影响



动态规划-可能的路径

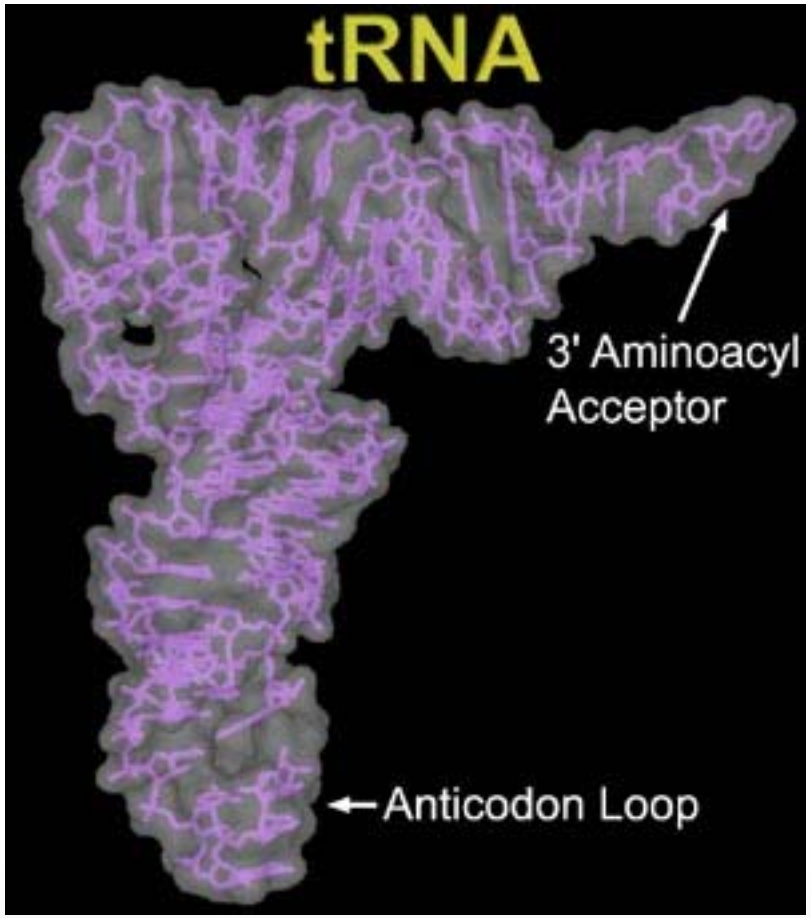
动态规划算法：最大碱基配对

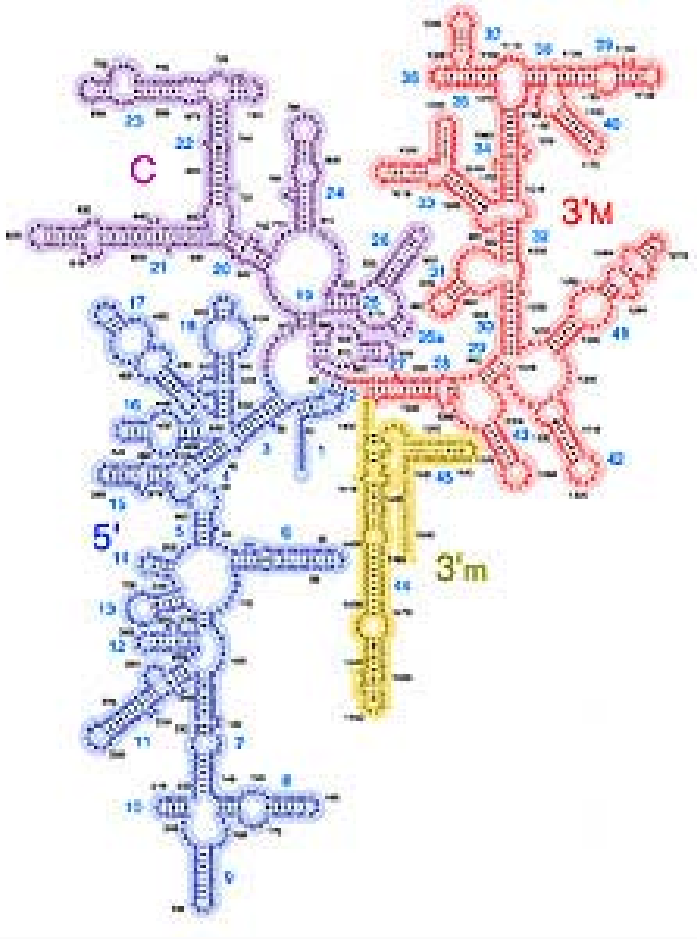
- 比对方法
  - RNA链与自身相比
  - 若存在碱基配对则增加分数
- 计算分值
- 可最后加入分歧的影响



分歧：考虑所有的*k*值

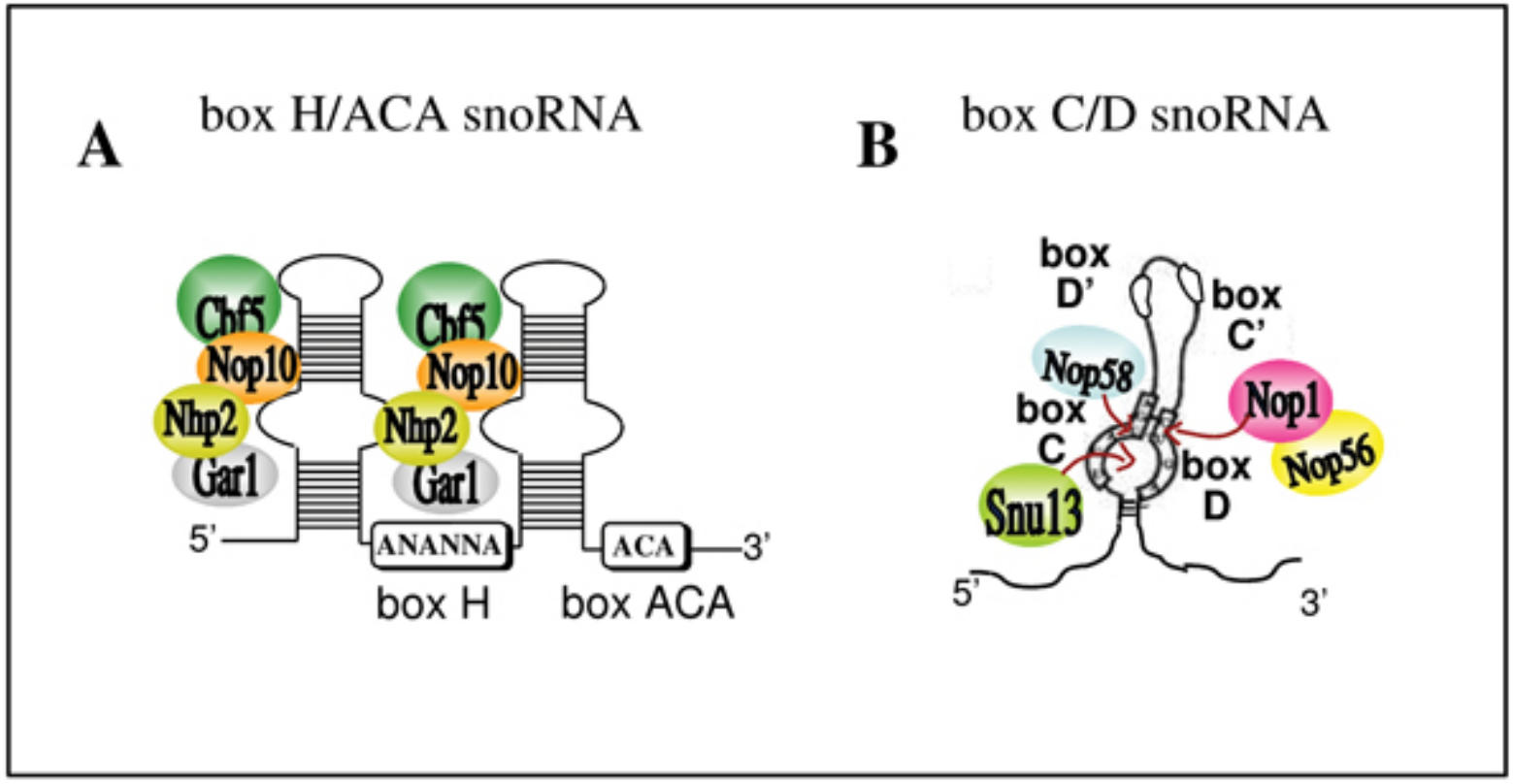
tRNA & rRNA





snoRNAs

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

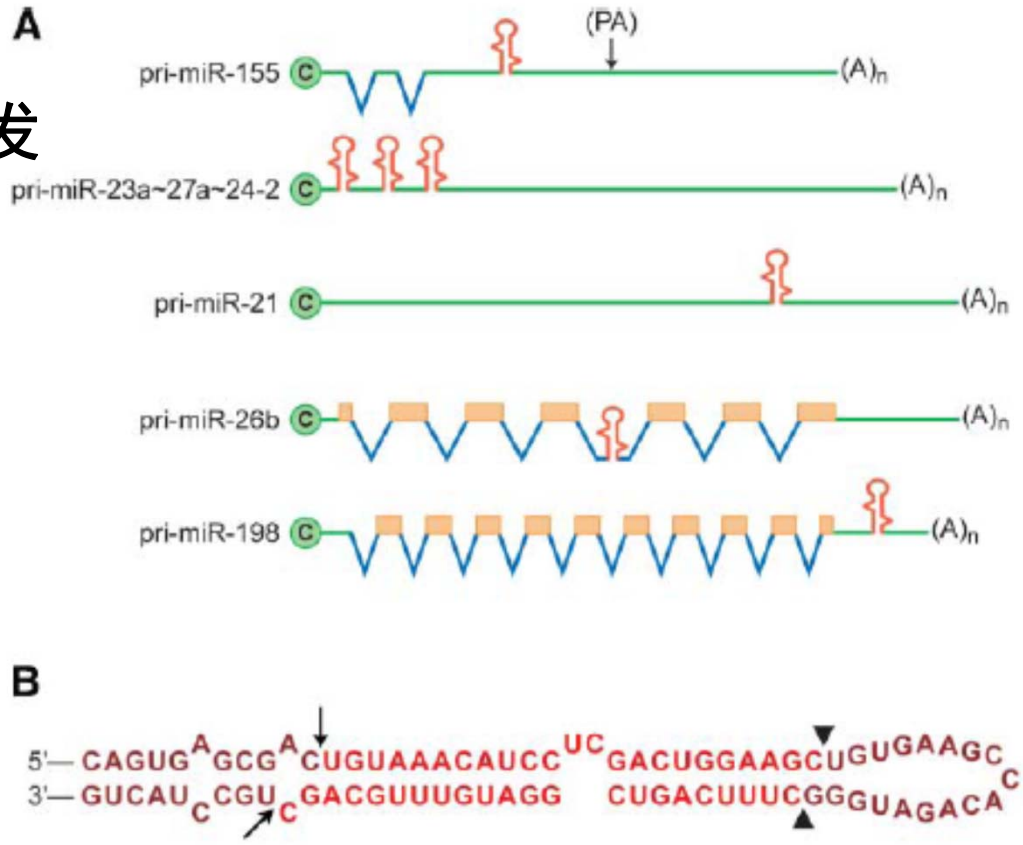


microRNA/miRNA

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

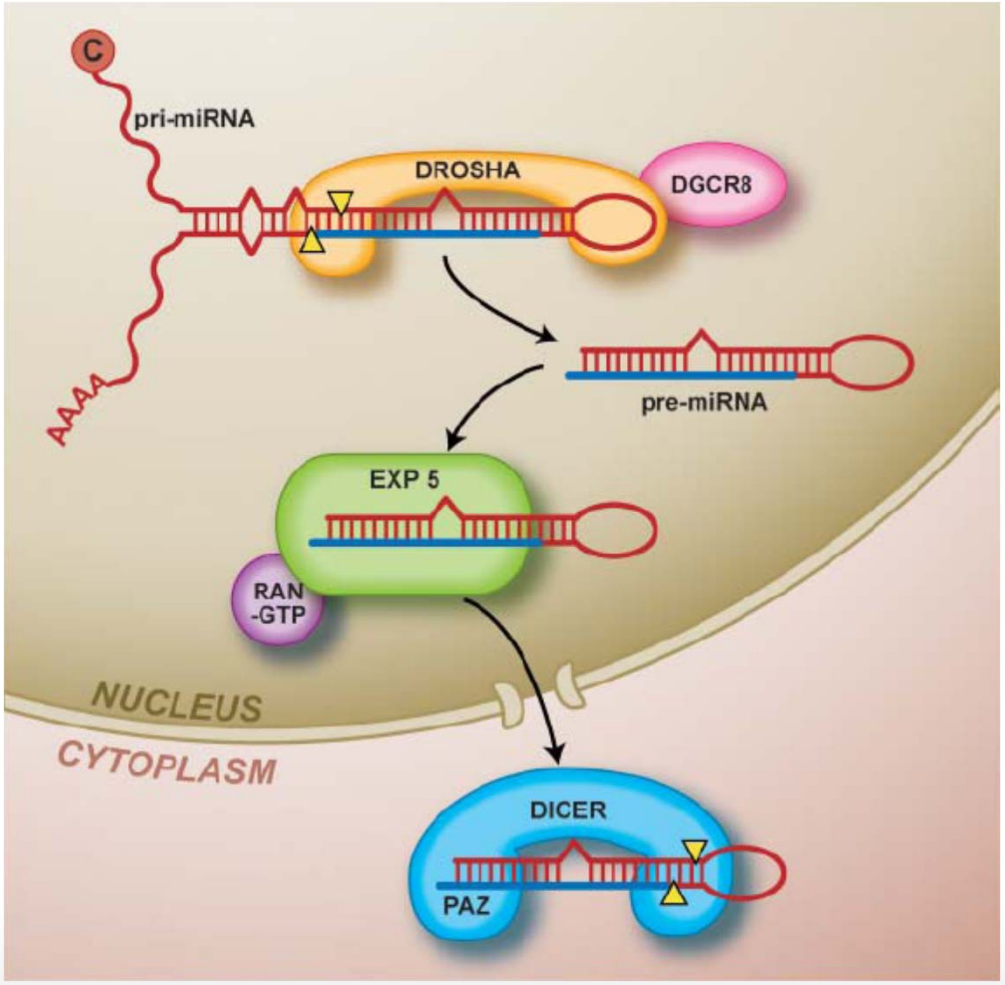
pri-miRNAs的结构

- pri-miRNAs的结构
- 人类pri-miR-30a RNA 的发夹环
  - Drosha切割位点 (箭头)
  - Dicer切割位点 (三角)



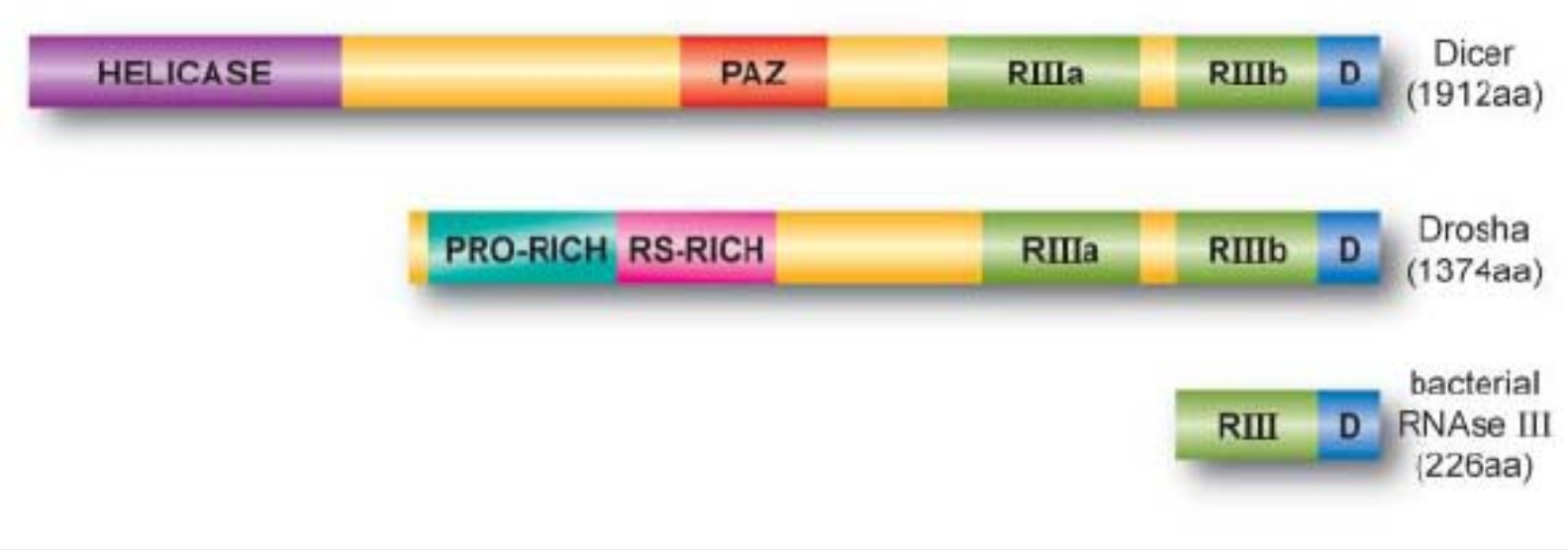
miRNA的加工

- Pri-miRNA被Drosha切割
- pre-miRNA被Dicer切割
- Exportin 5 (Exp5) 将miRNA装运到胞质中



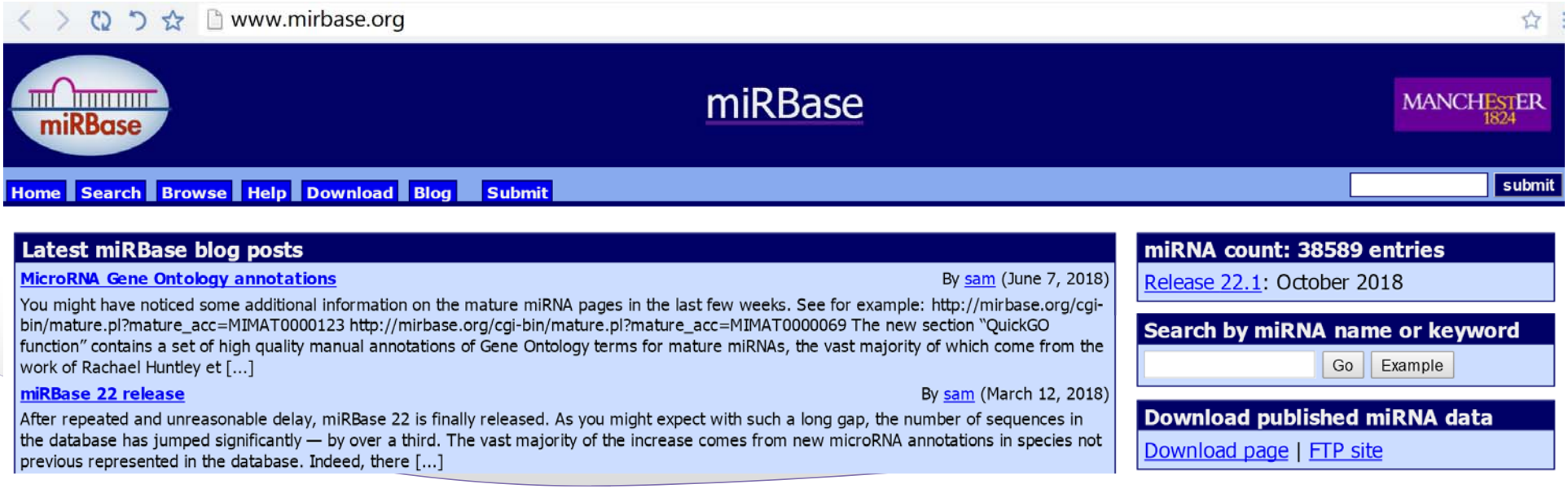
Drosha vs. Dicer

- RIII, RNase III 催化结构域; D, dsRNA 结合结构域; PRORICH, proline-rich结构域; RS-RICH, arginine/serine rich结构域
- Drosha和Dicer都具有 RNase III 和 dsRNA结合结构域



miRNAs的多样性

- miRNAs发现的数量近年来快速增长
- 2006年10月，miRBase 9.0，收录4361个具有发夹结构的前体miRNAs，能够表达出4167个miRNAs
- Release 12.0，收录8619个具有发夹结构的前体miRNAs，能够表达出8273个miRNAs

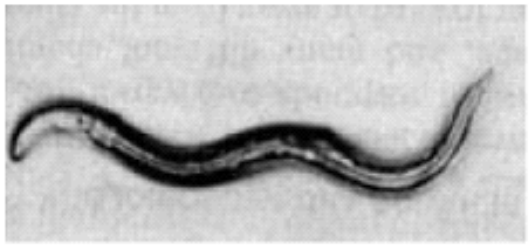


miRNAs的多样性

MicroRNAs (miRNAs)

A. thaliana/
O. sativa

拟南芥：**187**

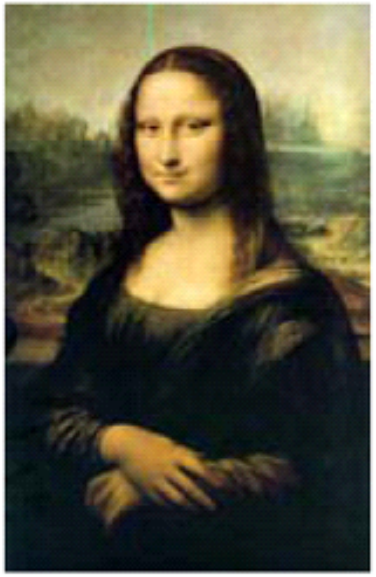


C. elegans

154

D. melanogaster

152



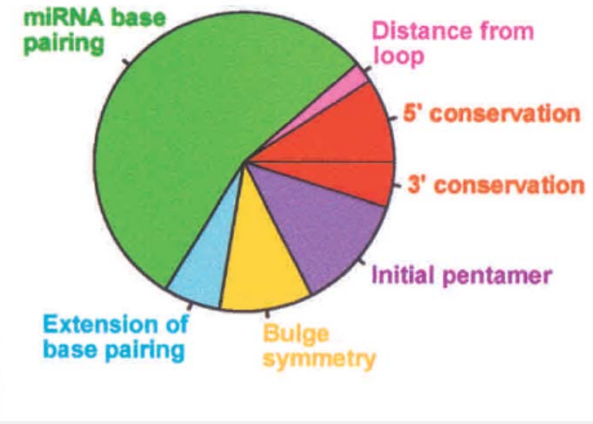
H. sapiens/
M. musculus

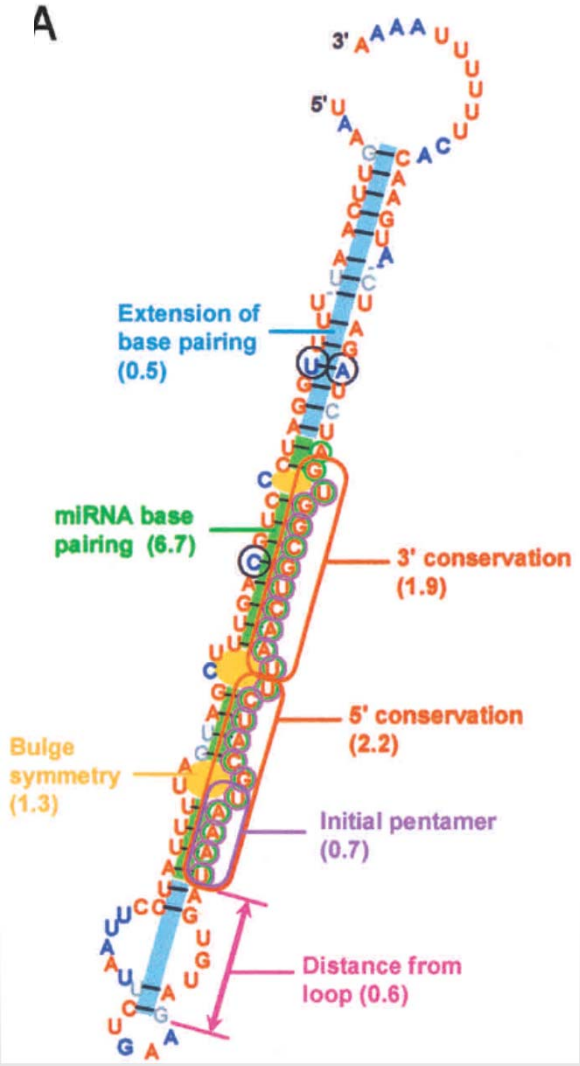
人：**695**

The miRBase Sequence Database -- Release 12.0

miRNAs的计算鉴定

- **MiRscan (Lim, *et al*. 2003)**
  - 发现*C. elegans*和*C. briggsae*之间保守的发卡结构
  - 利用50个已知的miRNA作为训练集





miRNAs在发育中的调控作用及底物

- lin-4 和let-7 miRNAs调控线虫发育的时间点

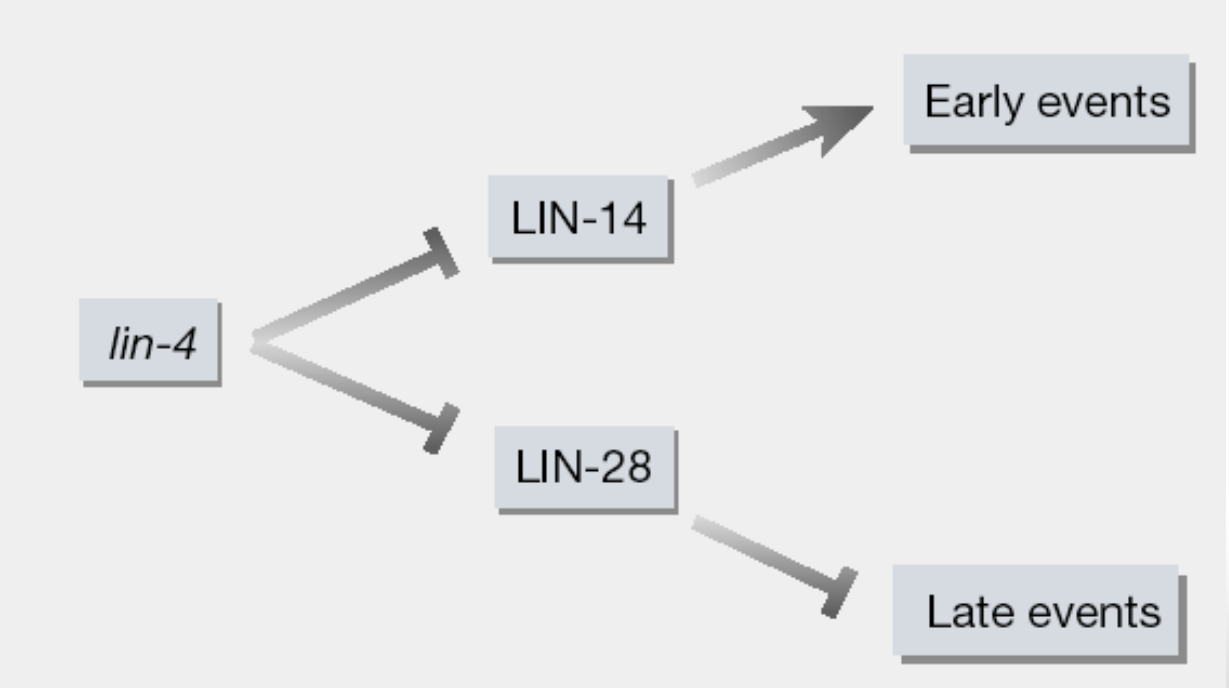
Table 1 **Animal miRNA genes with genetically assigned functions**

miRNA	Animal	Function	Targets
lin-4	Ce	developmental timing[19]	*lin-14* (refs 19, 22) *lin-28* (ref. 24)
let-7	Ce	developmental timing[20]	*lin-41* (ref. 23) *hbl-1* (refs 48, 49)
lsy-6	Ce	neuronal cell fate[29]	*cog-1* (ref. 29)
mir-273	Ce	neuronal cell fate[30]	*die-1* (ref. 30)
bantam	Dm	cell death, proliferation[26]	*hid* (ref. 26)
mir-14	Dm	cell death, fat storage[27]	caspase?
miR-181	Mm	haematopoietic cell fate[33]	?

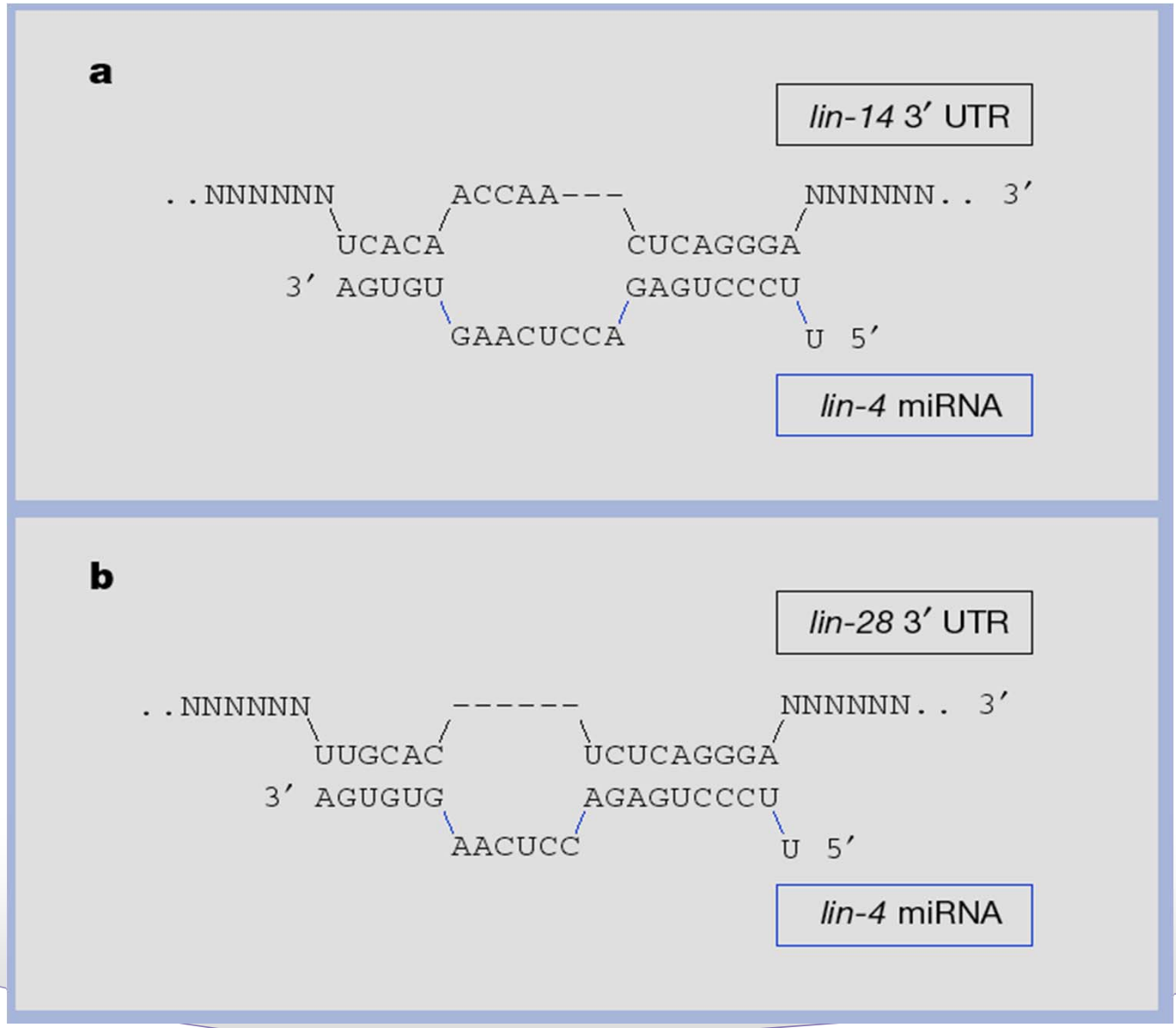
Ce, *C. elegans*; Dm, *D. melanogaster*; Mm, *M. musculus*.

lin-4在幼虫发育过程中的作用

- 在线虫的幼虫发育期，lin-4下调LIN-14和LIN-28蛋白质的浓度，从而一方面阻止后期发育，一方面促进幼虫的发育

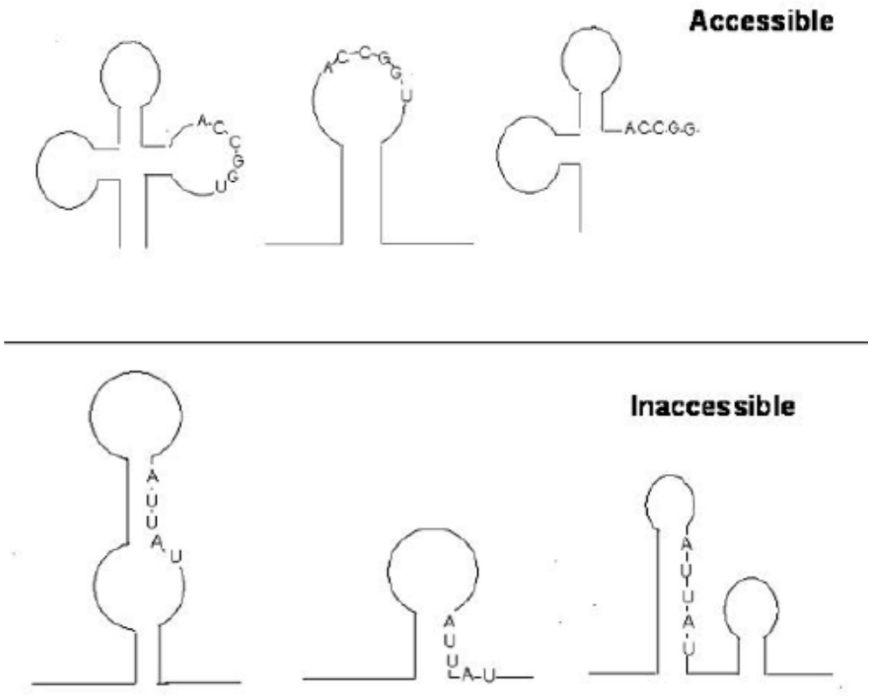


miRNA *vs.* 靶序列：不完美配对



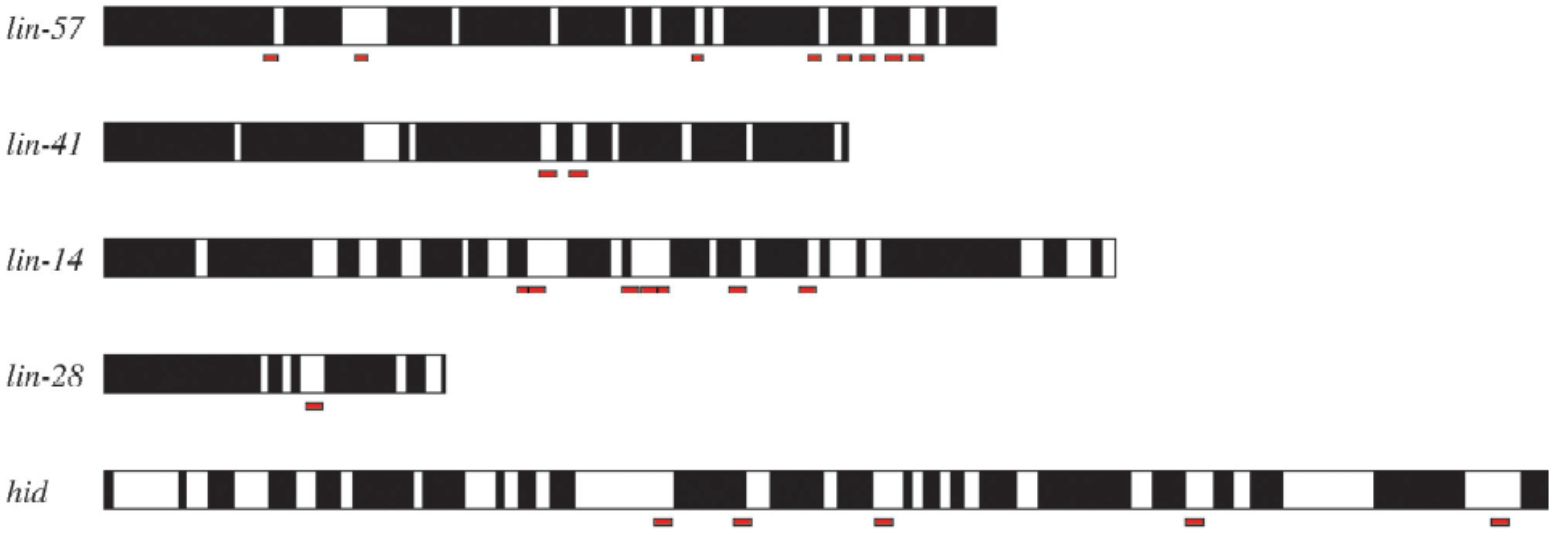
miRNA底物的计算发现

- miRNA的靶序列位于基因的3’ UTR区域，与成熟的miRNA互补
  - 植物：序列互补程度高
  - 动物：互补程度低
- 靶序列特征
  - 长度较短 (~21 nt)
  - G-U配对
  - 错配和空位 (bulges)
  - 假阳性高
- 提高准确性：考虑mRNA UTR的二级结构



miRNA靶序列的其他特性

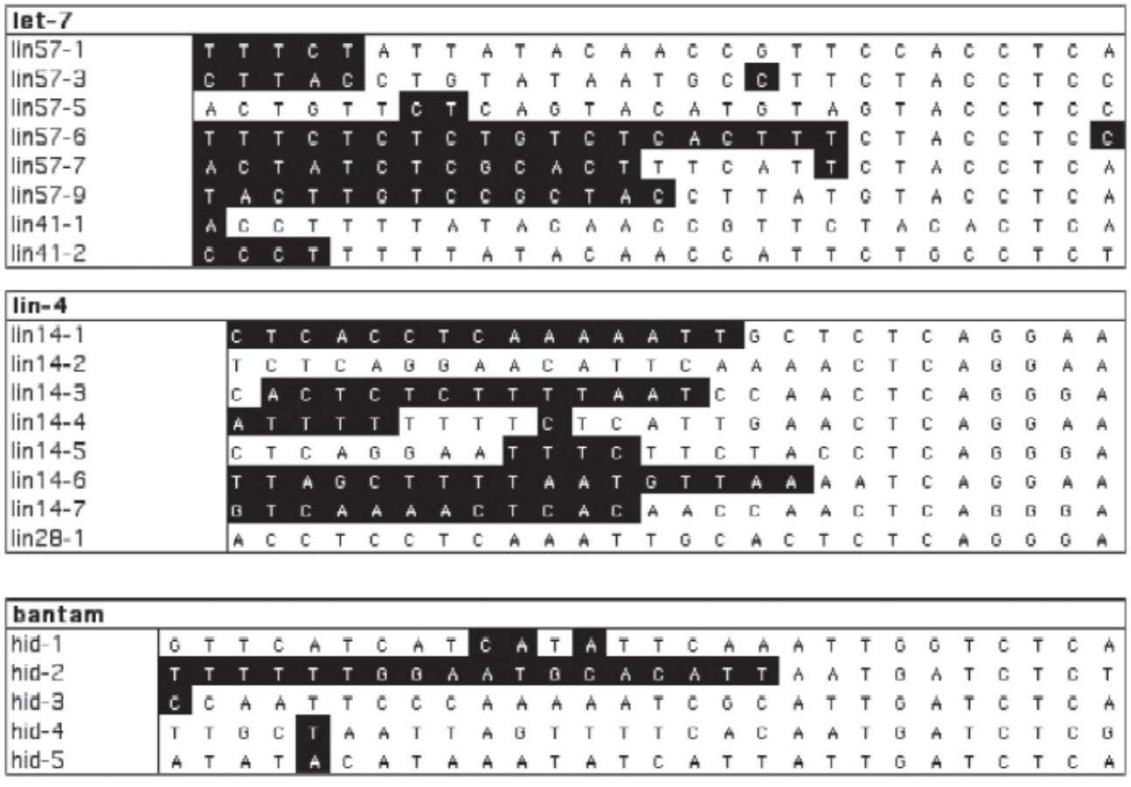
- 保守性：miRNA靶序列在不同物种中保守
- 成簇性：倾向于聚集成簇



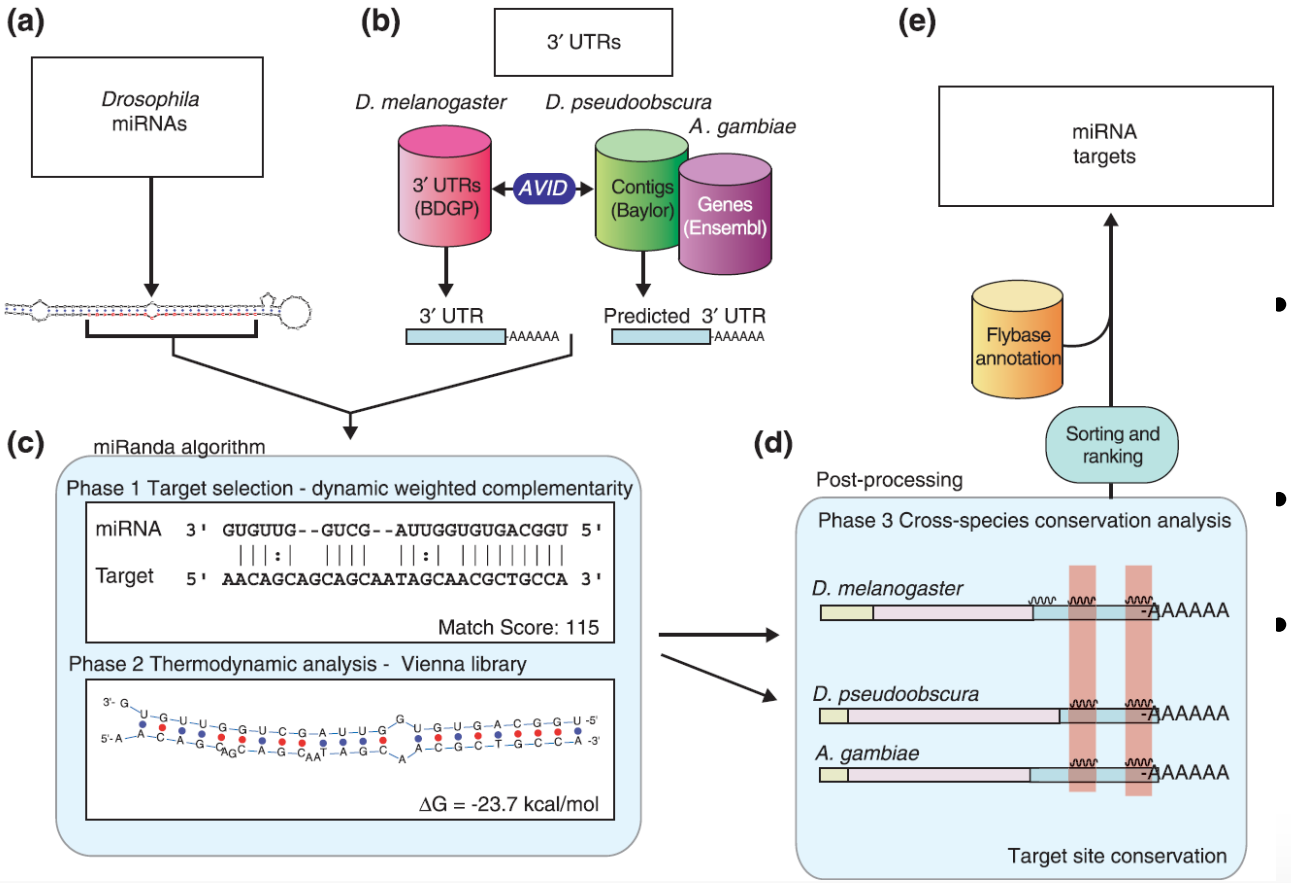
- lins：比较*C. elegans* and *C. briggsae*
- hid：比较*D. melanogaster* and *D. pseudoobscura*

miRNA靶序列的其他特性

- 靶位点的序列保守性
  - 与miRNA的5’端匹配更好



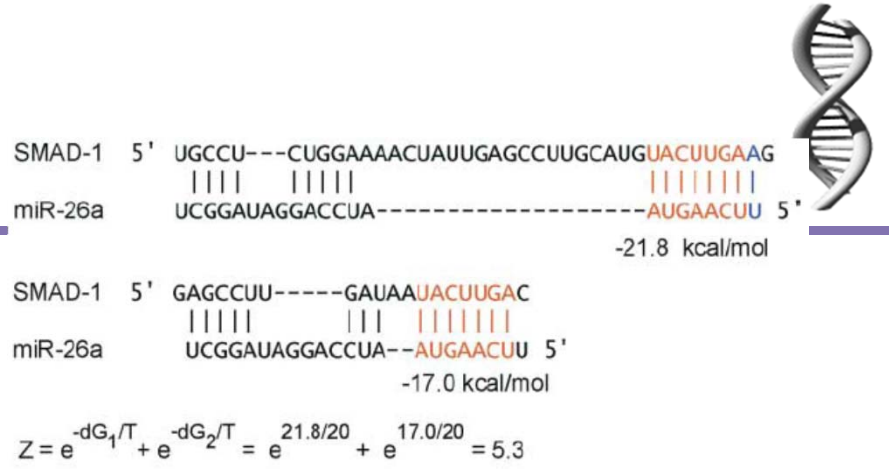
miRanda

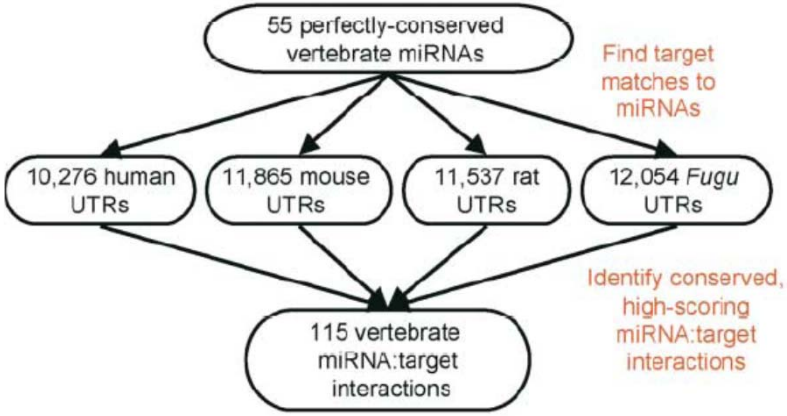


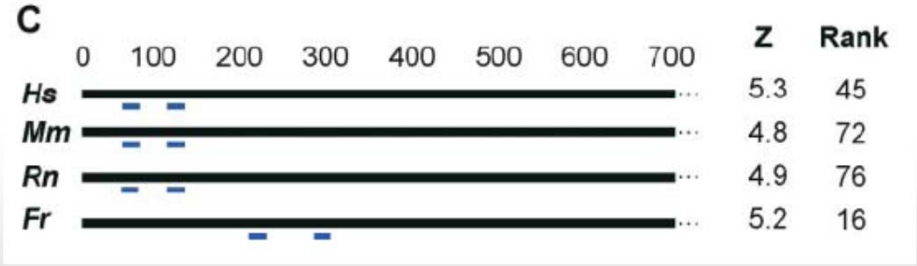
- “miRanda” Enright et al. Genome Biology 2003
- 结合上述特征，分配不同的权重
- 预测靶序列与miRNA 5’端的结合情况

Targetscan

- 给定一个在多物种中保守的miRNA及相应的直系同源UTR区域
- miRNA seed: 7nt
- 延伸每一条seed并发现最佳能量
- 计算Z值
- 排序结果，获得R_i值
- 保留$Z_i>Z_c$和$R_i<R_c$的结果

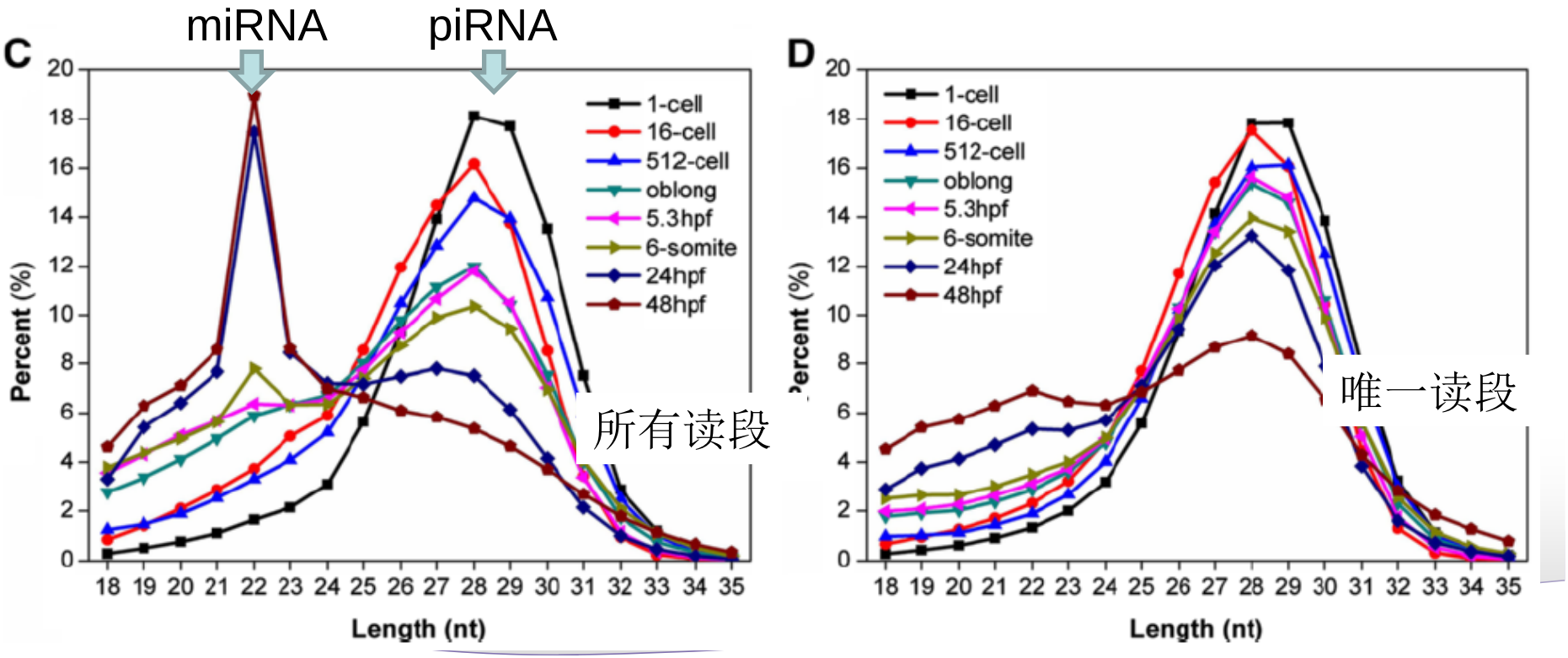






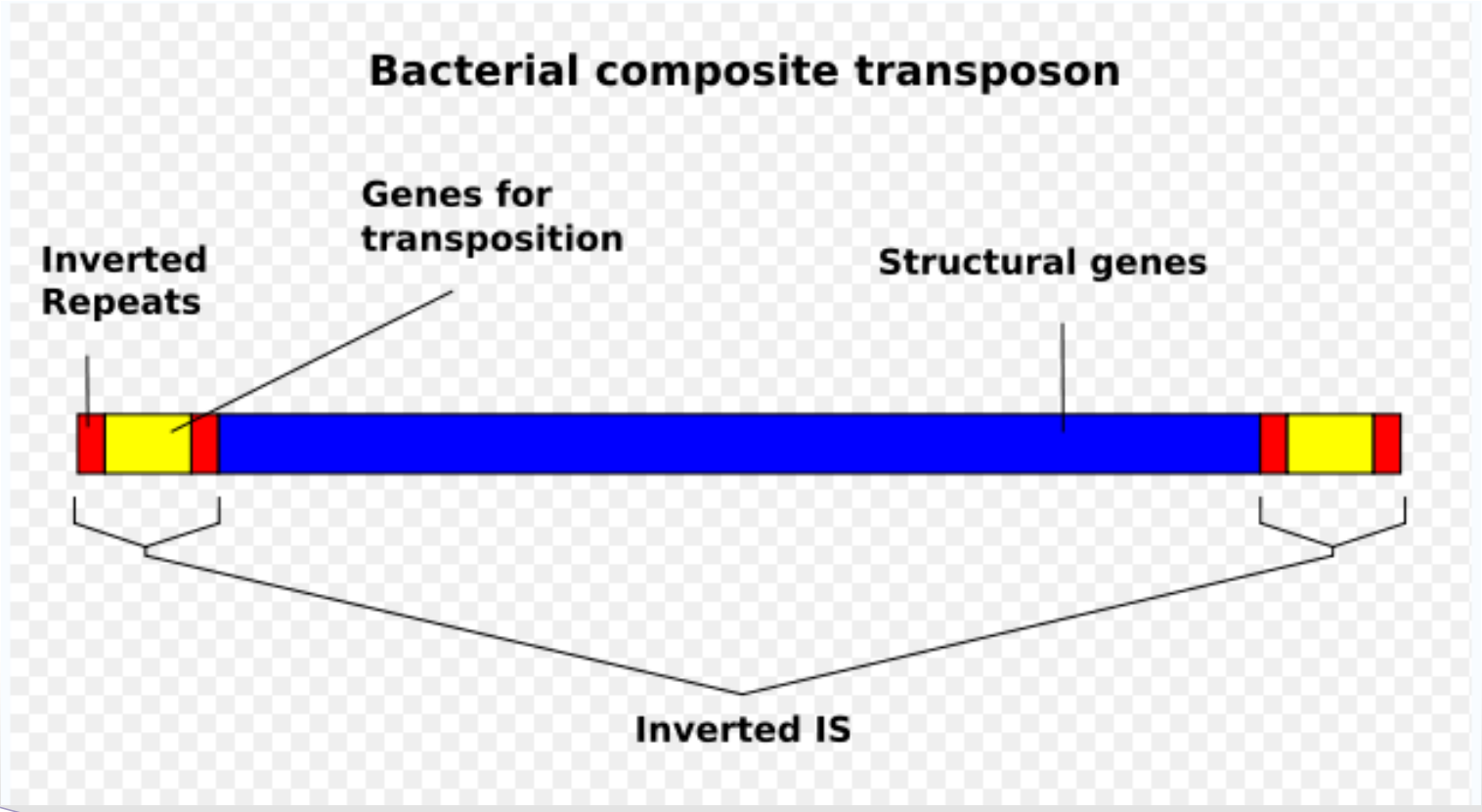
sRNA-Seq：小RNA测序

- 斑马鱼胚胎发育的8个时期
  - 1-cell (0.2 hpf), 16-cell (1.5 hpf), 512-cell (2.75 hpf), oblong (3.7 hpf), 5.3 hpf, 6-somite (12 hpf), 24 hpf and 48 hpf



Transposon

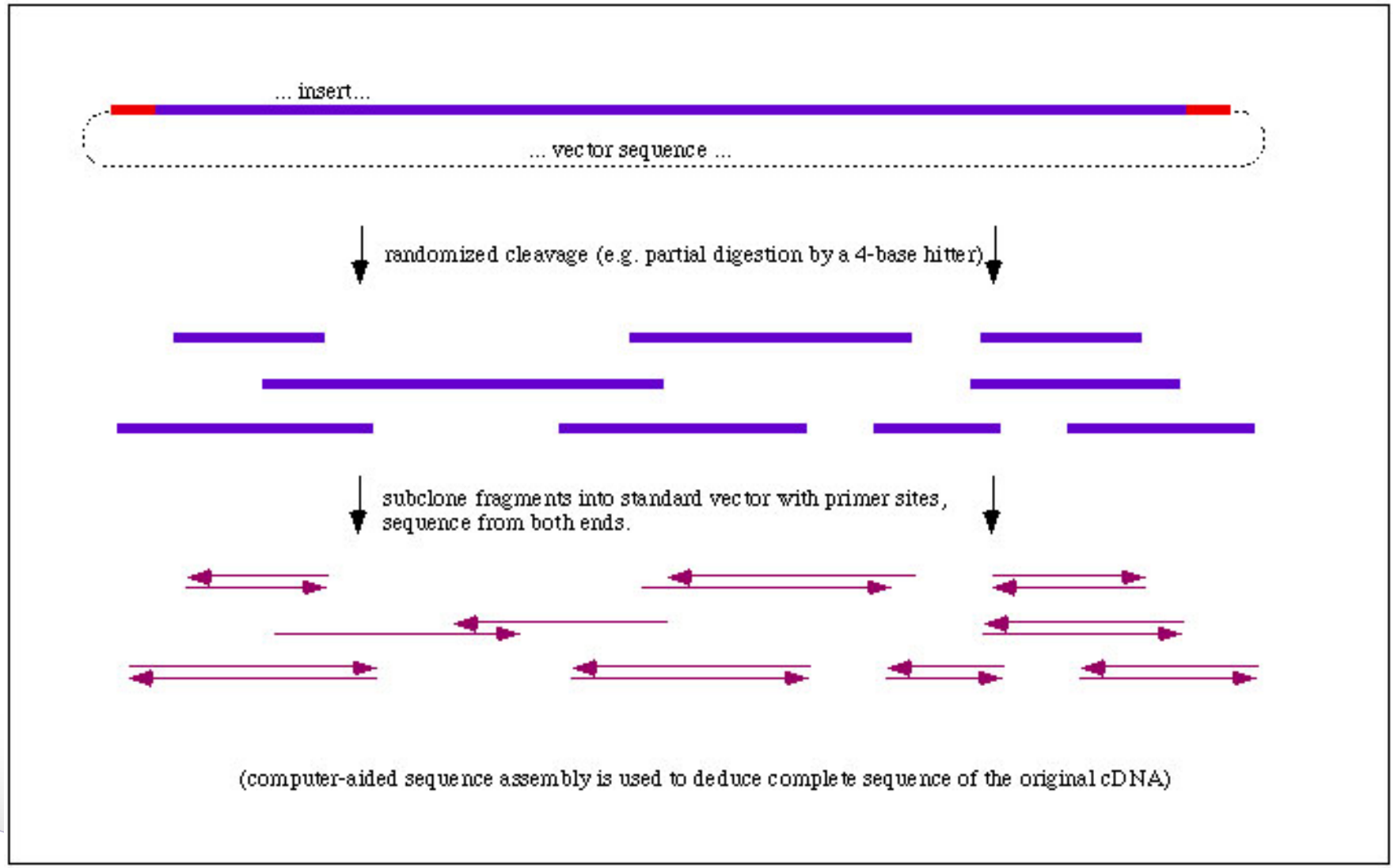
- 转座子：在基因组中能够移动位置的DNA序列



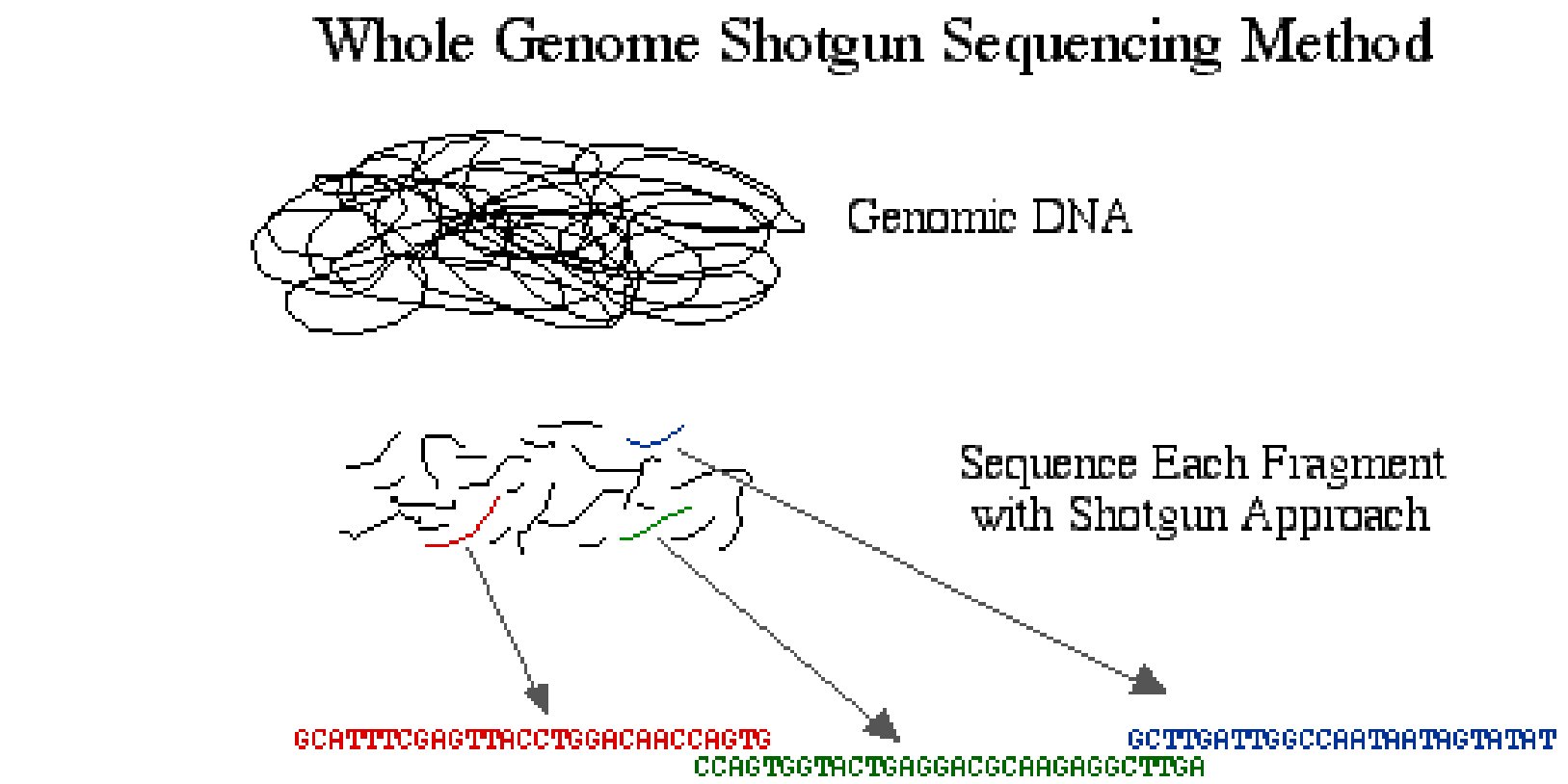
基因组注释

- 基因组序列的拼装
- 基因预测
- 可变剪切的预测
- 非编码的功能元件的预测

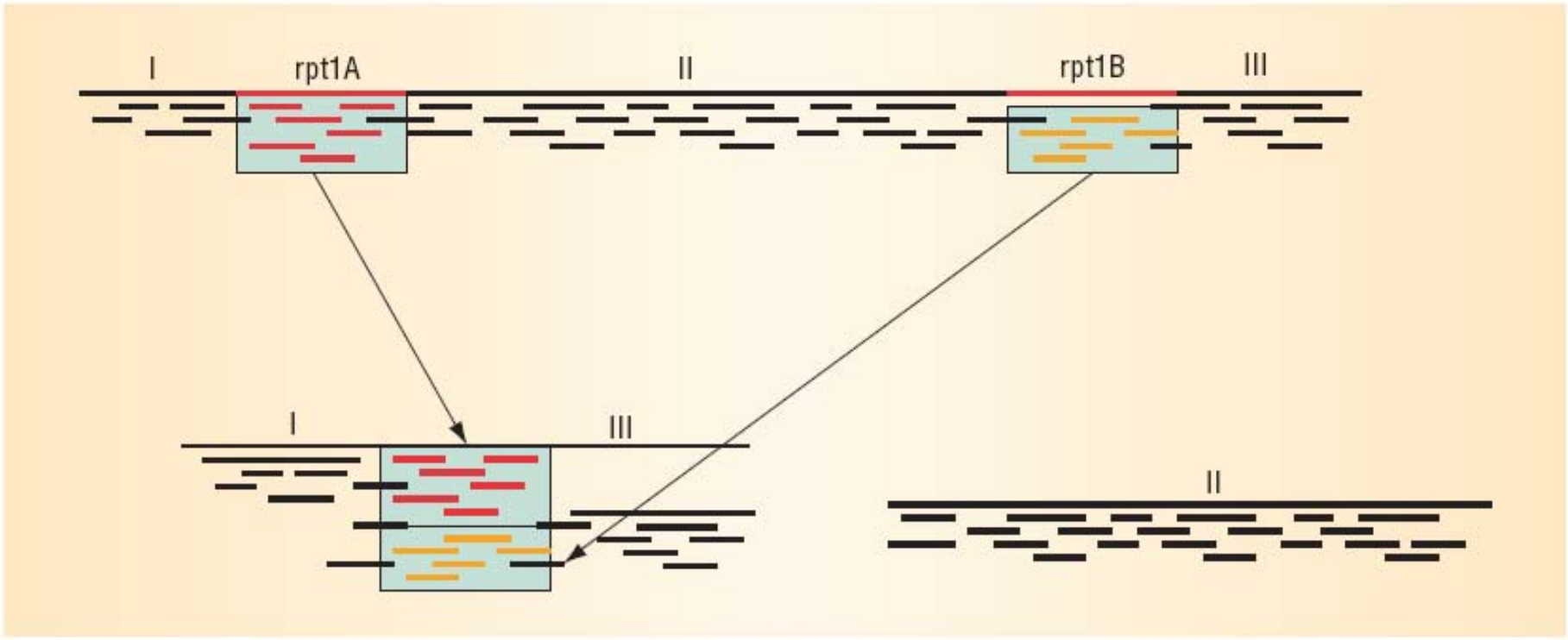
基因组测序：鸟枪法


... insert...
... vector sequence ...
randomized cleavage (e.g. partial digestion by a 4-base hitter)
subclone fragments into standard vector with primer sites, sequence from both ends.
(computer-aided sequence assembly is used to deduce complete sequence of the original cDNA)

基因组的拼装


Whole Genome Shotgun Sequencing Method
Genomic DNA
Sequence Each Fragment
with Shotgun Approach
GCATTTCGAGTTACCTGGACAACCAGTG
CCAGTGGTACTGAGGACGCAAGAGGCTTGA
GCTTGATTGGCCAATAATAGTATAT
Align Contiguous Sequences
GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT
Generate Finished Sequence

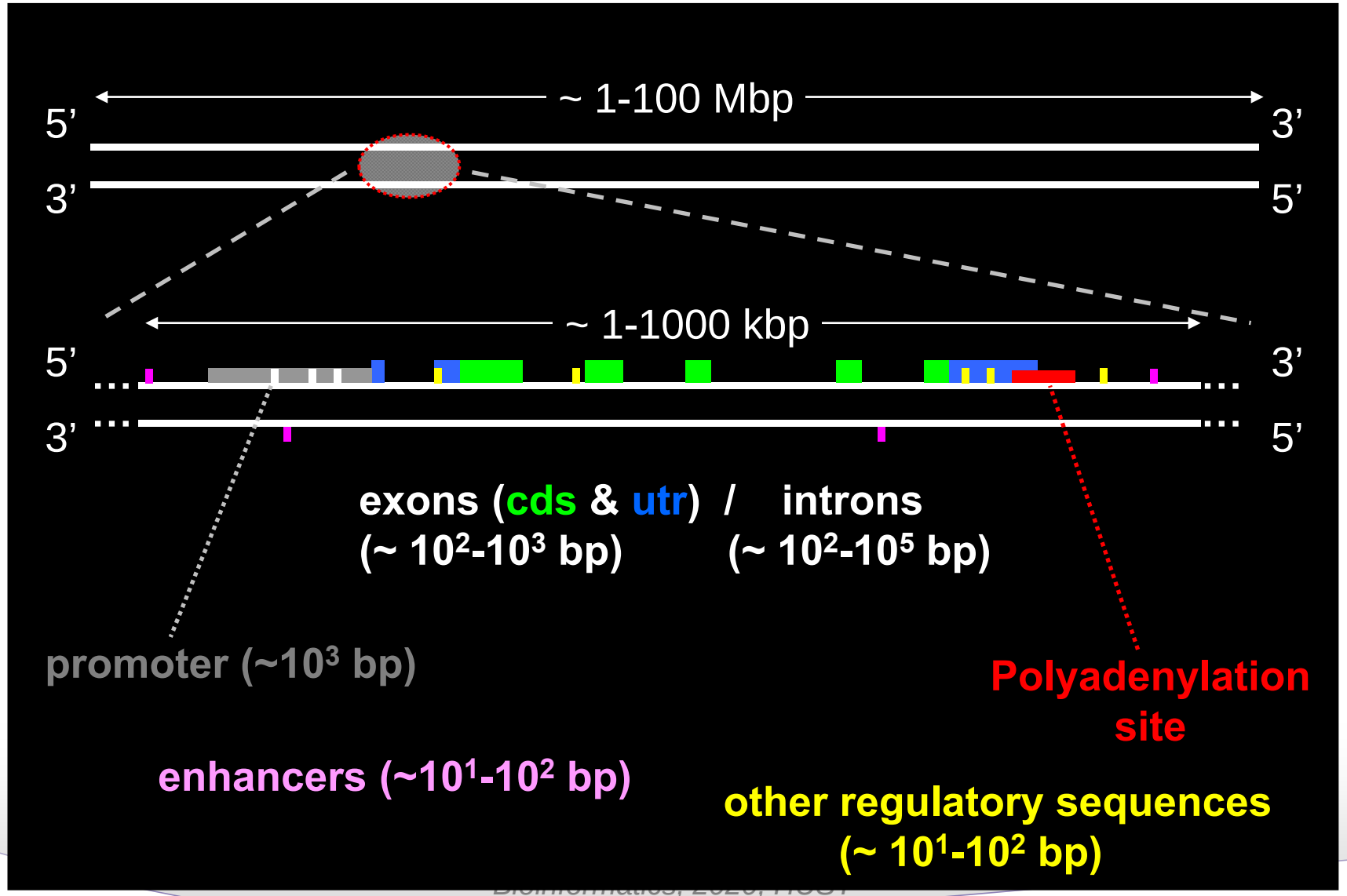
重复序列带来干扰



基因预测

- 直接的，序列高度匹配
  - 同一或近缘物种中，与EST，cDNA, 蛋白质等序列完美或近似完美的匹配
- 间接的，基于统计学的
  - 序列比对 (Homology)
  - 从头预测 (*ab initio*)
  - 以上两种方法的结合

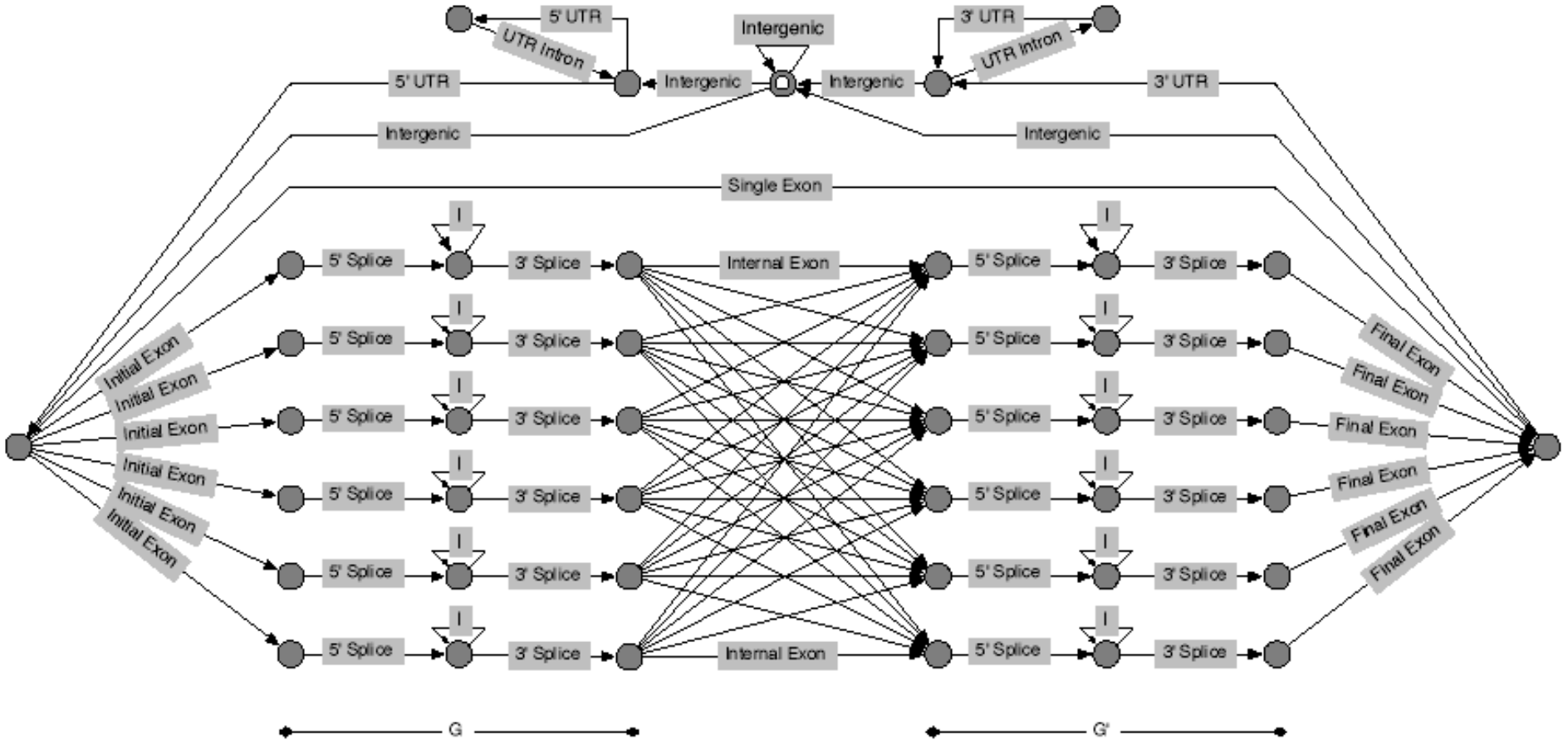
真核生物的基因结构


~ 1-100 Mbp
5’
3’
3’
5’
~ 1-1000 kbp
5’
3’
3’
5’
exons (cds & utr) / introns
(~ 10²-10³ bp) (~ 10²-10⁵ bp)
promoter (~10³ bp)
Polyadenylation site
enhancers (~10¹-10² bp)
other regulatory sequences
(~ 10¹-10² bp)

基因的其他特征

- ORF (Open Reading Frame): 从AUG开始，至stop codon终止
- Codon Usage: CAI
- …

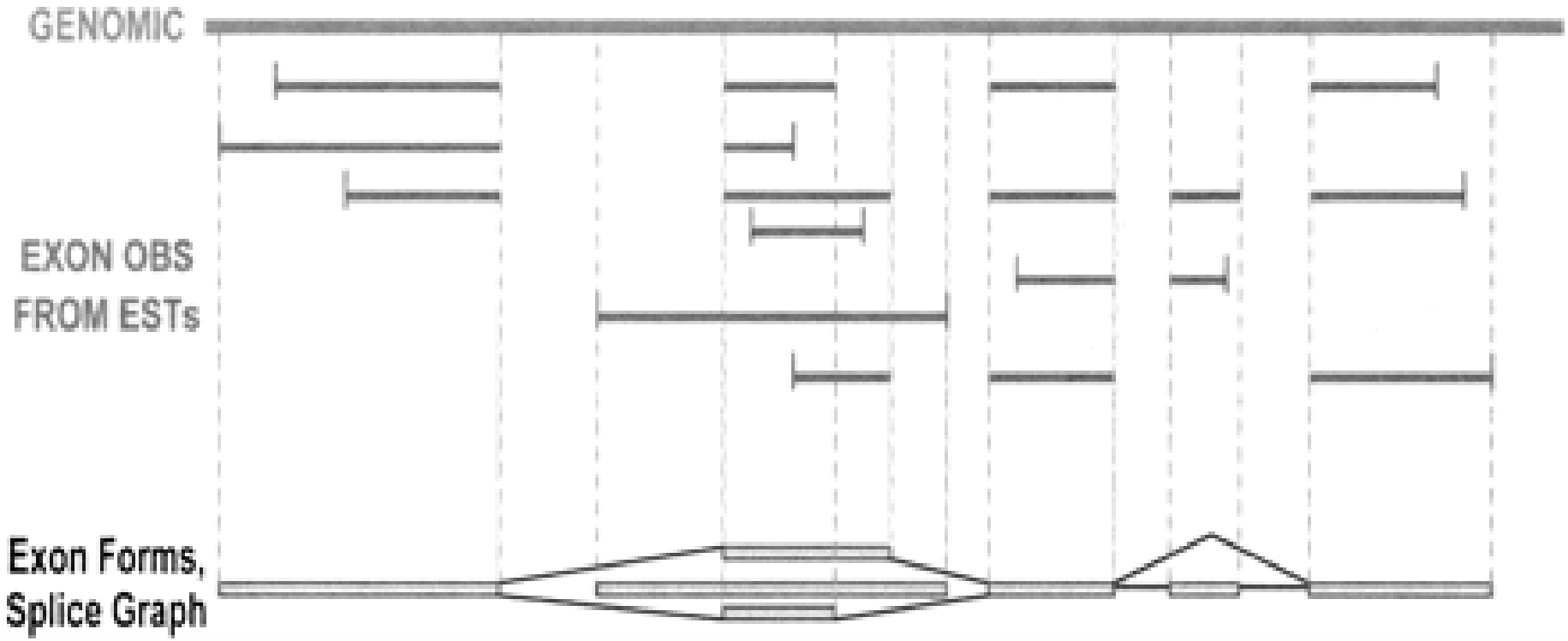
HMM model for Gene Prediction (Genie)



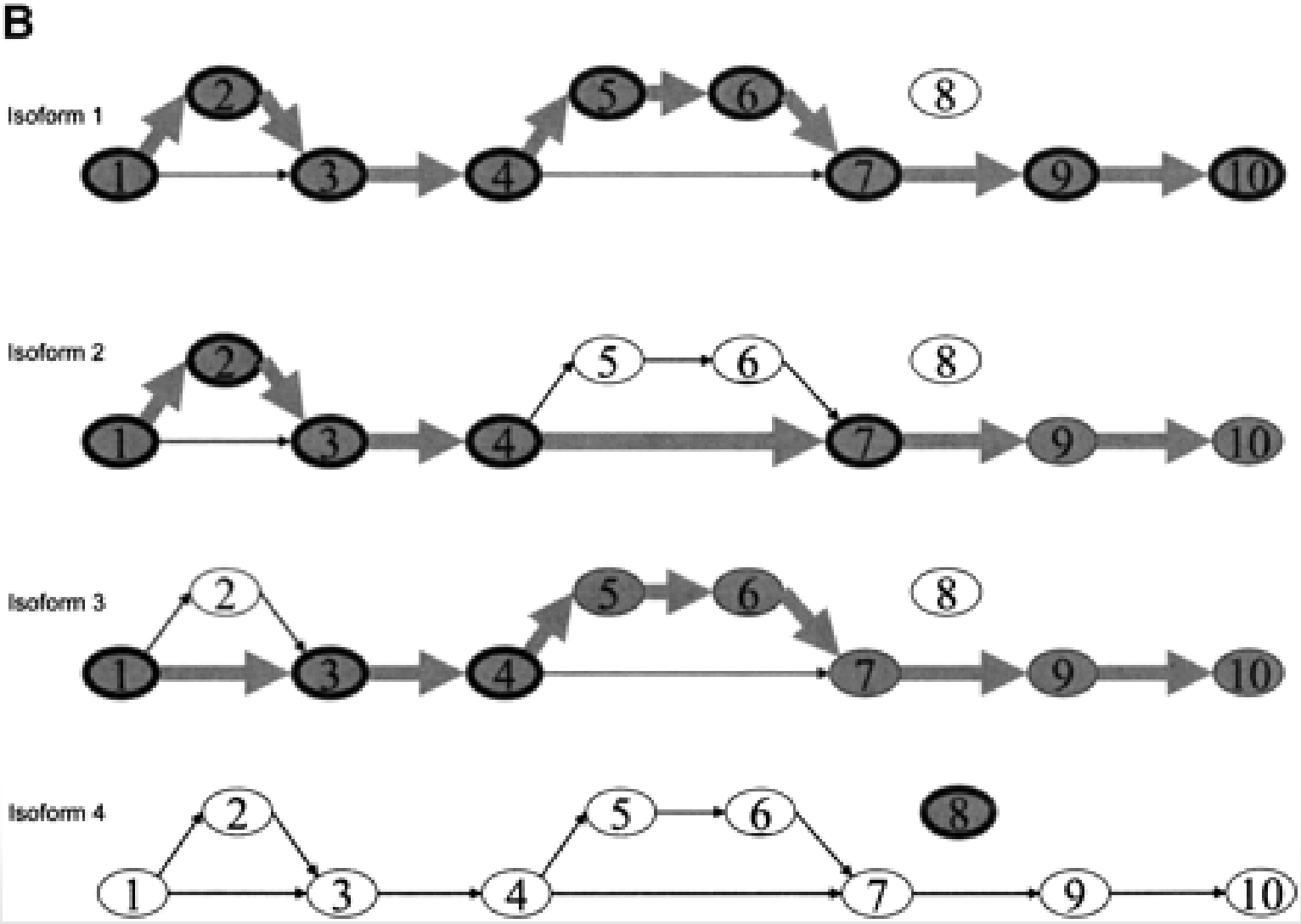
Kulp, D., PhD Thesis, UCSC 2003

可变剪接的预测

- 将EST, cDNA序列比对到基因组上



部分有向图算法（POA）



基因/蛋白质的功能预测

- 一级序列的比较：相似的序列具有相似的功能
- 保守的功能结构域：保守的功能
- 三级结构的比较：相似的结构具有相似的功能
- 蛋白质相互作用的预测

一级序列的比较

- 同源序列的鉴定：不同物种中的直系、旁系同源物的预测
- 主要工具：BLAST

保守的功能结构域

- 保守的功能结构域：保守的功能
- 常用工具：

工具	网址
Interpro	http://www.ebi.ac.uk/interpro/
Pfam	http://pfam.sanger.ac.uk/
SMART	http://smart.embl.de/
PROSITE	http://www.expasy.org/prosite/
ProDom	http://prodom.prabi.fr/prodom/current/html/home.php
CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

保守的功能结构域

- Pfam: 手工收集结构域，蛋白质家族分类，计算速度较慢，准确性较高
- SMART：根据蛋白质三级结构获得结构域信息，准确性较高
- InterPro：整合多个功能结构域的数据库，灵敏度高，可能有假阳性

例：Nek2

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP

Pfam keyword search Go

Protein: *NEK2_HUMAN* (P51955)

1 architecture | 1 sequence | 0 interactions | 1 species | 0 structures

- Summary
- Features
- Sequence
- Interactions
- Structures
- TreeFam

Jump to... enter ID/acc Go

Summary

NEK2_HUMAN

This is the summary of UniProt entry NEK2_HUMAN (P51955).

Description:	Serine/threonine-protein kinase Nek2 (EC 2.7.11.1) (NimA-relatedprotein kinase 2) (NimA-like protein kinase 1) (HSPK 21).
Source organism:	Homo sapiens (Human). (NCBI taxonomy ID 9606) View Pfam genome data.
Length:	445 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains. Note that some domains may be obscured by other, overlapping domains. This is noted in the table where applicable.

Source	Domain	Start	End
PfamA	Pkinase	8	271
PfamB	Pfam-B_22130	376	445

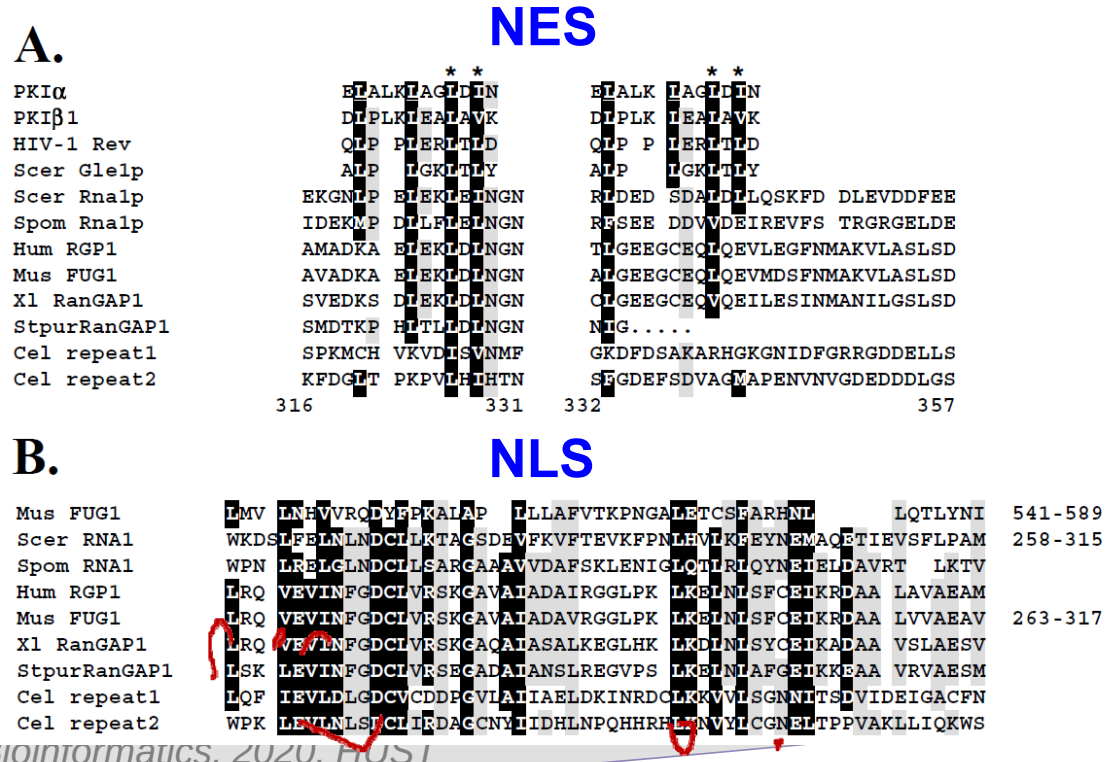
细胞亚定位分析

- 信号肽 (signal peptides)
  - Nuclear export sequences (NES)
  - Nuclear localization sequences (NLS)

The Nobel Prize in Physiology or Medicine 1999

Günter Blobel
Prize share: 1/1

The Nobel Prize in Physiology or Medicine 1999 was awarded to Günter Blobel *"for the discovery that proteins have intrinsic signals that govern their transport and localization in the cell".*



TargetP

- 根据蛋白质序列N端氨基酸组成预测细胞亚定位
  - 叶绿体转运肽 (cTP)、线粒体定位肽 (mTP)和分泌通路的信号肽 (SP)

TargetP 1.1 Server

TargetP 1.1 predicts the subcellular location of eukaryotic proteins. The location assignment is based on the predicted presence of any of the N-terminal presequences: chloroplast transit peptide (**cTP**), mitochondrial targeting peptide (**mTP**) or secretory pathway signal peptide (**SP**).

For the sequences predicted to contain an N-terminal presequence a potential cleavage site can also be predicted.

NOTE 1: TargetP uses ChloroP and SignalP to predict cleavage sites for **cTP** and **SP**, respectively.

NOTE 2: The method has been tested on *A. thaliana* and *H. sapiens* sets; see the results.

***New:* the paper about using TargetP and other protein subcellular localization prediction methods:**

Locating proteins in the cell using TargetP, SignalP, and related tools
Olof Emanuelsson, S??ren Brunak, Gunnar von Heijne, Henrik Nielsen
Nature Protocols **2**, 953-971 (2007).

is now again available for download - please click here.

Instructions | Output format | Article abstract

SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:

Submit a file in FASTA format directly from your local disk:
选择文件 未选择任何文件

Organism group
- Non-plant
- Plant

Prediction scope
- Perform cleavage site predictions

Cutoffs
- no cutoffs; winner-takes-all (default)
- specificity >**0.95** (predefined set of cutoffs that yielded this specificity on the TargetP test sets)
- specificity >**0.90** (predefined set of cutoffs that yielded this specificity on the TargetP test sets)
- define your own cutoffs (0.00 - 1.00): **cTP**: 0.00 **mTP**: 0.00 **SP**: 0.00 **other**: 0.00

Submit | Clear fields

神经网络算法

SubLoc

- 首次将SVM算法应用于细胞亚定位预测
- 中国最早发表在国际专业期刊的生物信息学工作之一

BIOINFORMATICS

Vol. 17 no. 8 2001
Pages 721–728

Support vector machine approach for protein subcellular localization prediction

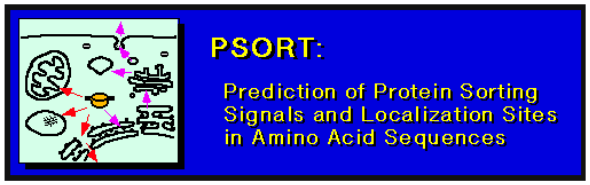
Sujun Hua and Zhirong Sun*

Institute of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, People's Republic of China

Received on December 12, 2000; revised on March 28, 2001; accepted on April 24, 2001

PSORT

❒ *k*-nearest neighbor算法



PSORT WWW Server

PSORT is a computer program for the prediction of protein localization sites in cells. It receives the information of an amino acid sequence and its source orgin, *e.g.*, Gram-negative bacteria, a sequence by applying the stored rules for various sequence features of known protein sorting signals. Finally, it reports the possiblity for the input protein to be localized at each candidate si

PSORT is mirrored at Tokyo, Okazaki, and Peking

- December 1, 1998, Official release of the PSORT II package
- June 1, 1999, K. Nakai moved to Univ. Tokyo
- October 13, 1999, The Web server has been moved from Osaka to Tokyo
- March 11, 2001, Introduction of iPSORT
- September 23, 2001, New mirror site at Peking University
- December 22, 2001, Distribution of caml-iPSORT
- January 18, 2003, Replacing the training data for PSORT II at Peking
- February 22, 2003, Rebuilding the PSORT II server at Tokyo
- April 16, 2003, Minor update of the top page
- November 9, 2003, Minor updates of several pages
- May 27, 2005, Link to WoLF PSORT; update some links
- January 5, 2007, Modification of the link to WolfPSORT

CONTENTS

WoLF PSORT (an update of PSORT II for fungi/animal/plant sequences)

WoLF PSORT Prediction

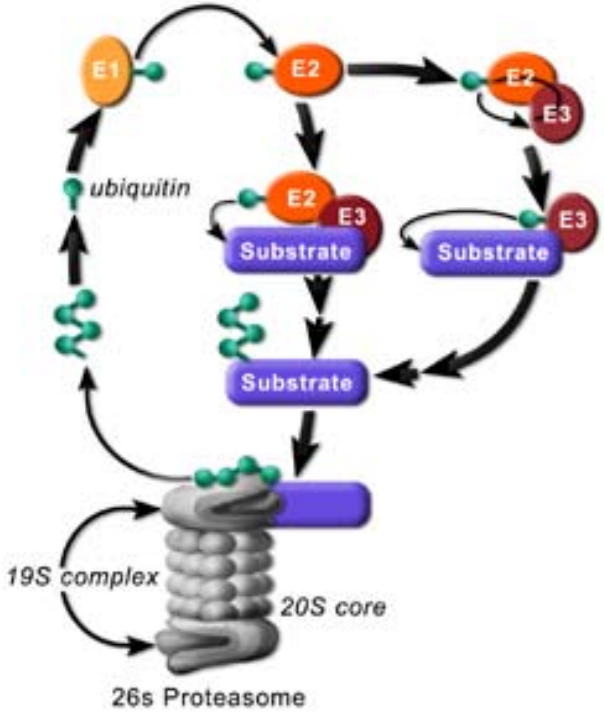
PSORT II (Recommended for animal/yeast sequences)

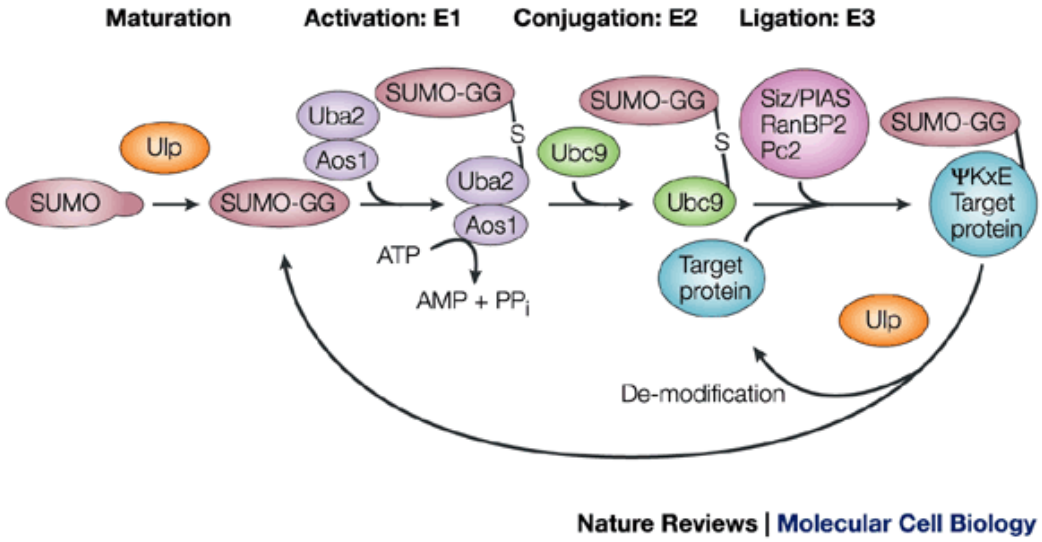
PSORT II Users' Manual
PSORT II Prediction

三级结构的比较

- Ubiquitin: 泛素，主要负责蛋白质的降解
- SUMO：小的类泛素蛋白质，基因转录 & 信号通路
- 催化反应通路的分子机制相似
- 序列相似性：不显著！

Ubiquitin *vs*. SUMO

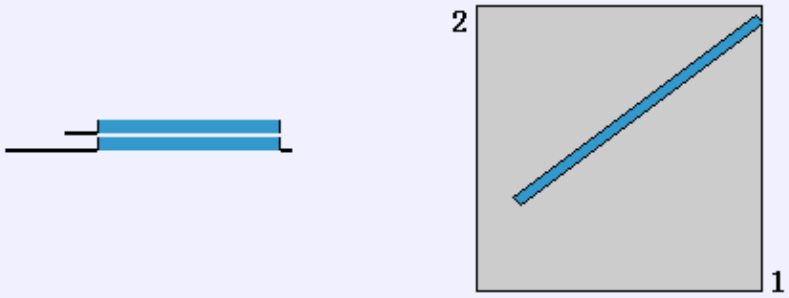




序列相似性：~20%

Sequence 1: sp|P62988|UBIQ_HUMAN Ubiquitin OS=Homo sapiens GN=RPS27A PE=1 SV=1
Length = 76 (1 .. 76)

Sequence 2: sp|P63165|SUMO1_HUMAN Small ubiquitin-related modifier 1 OS=Homo sapiens GN=SUMO1 PE=1 SV=1
Length = 101 (1 .. 101)



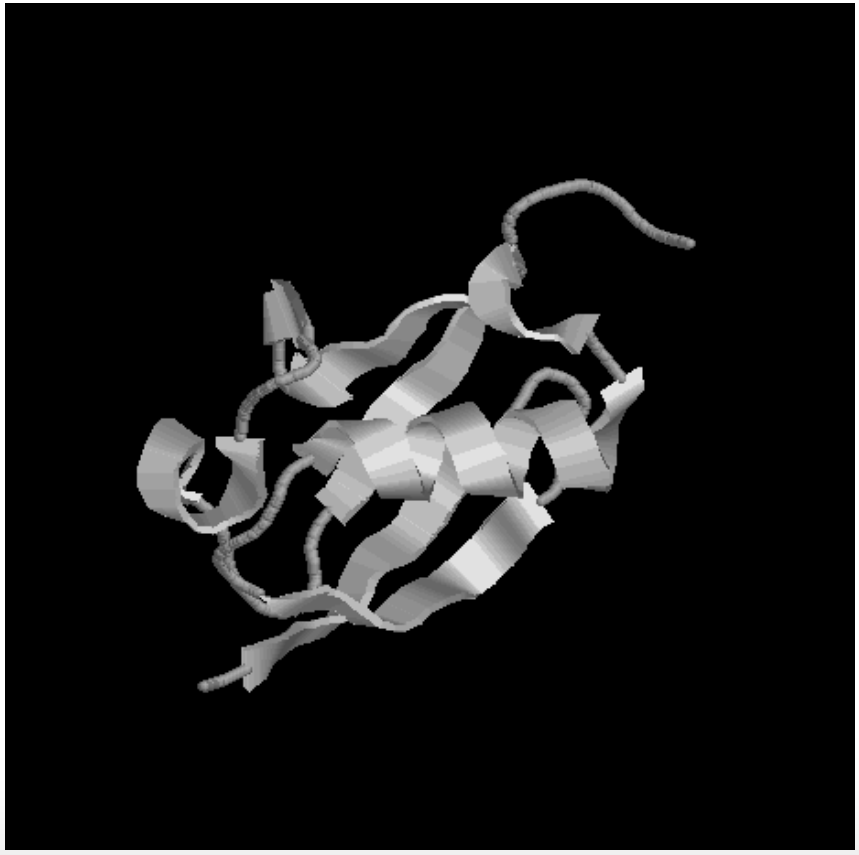
NOTE:Bitscore and expect value are calculated based on the size of the nr database.

Score = 32.7 bits (73), Expect = 8.3
Identities = 13/64 (20%), Positives = 33/64 (51%), Gaps = 0/64 (0%)

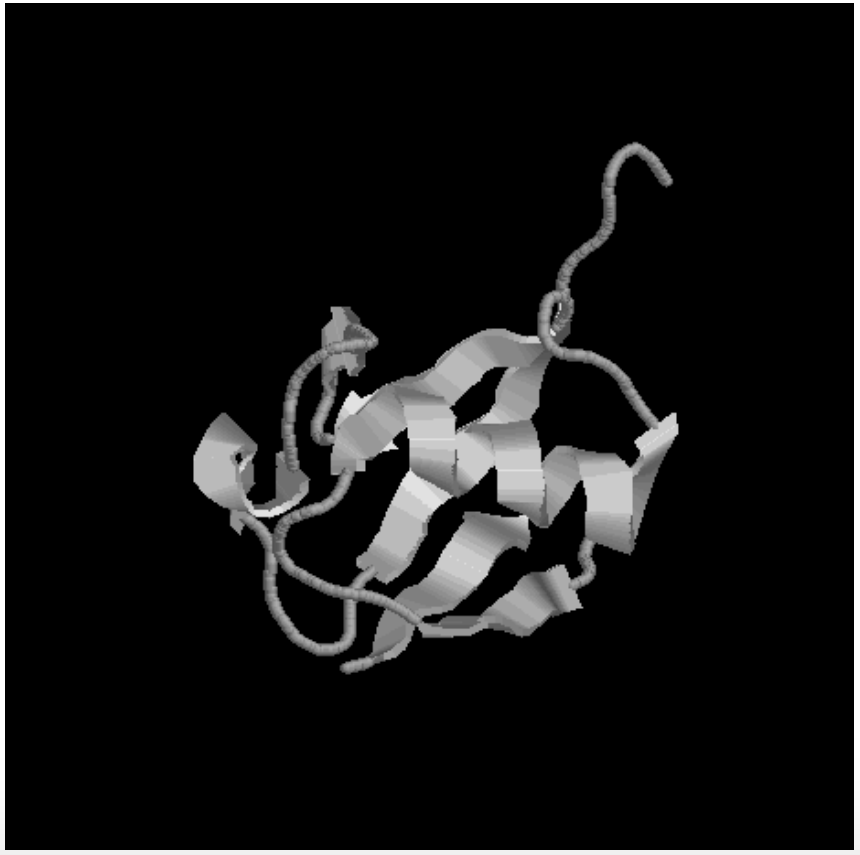
```
Query  13  ITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR  72
           I  +V+ +  ++ +K     ++G+P +  R +F G+++ D  T  +  +++E  + +
Sbjct  34  IHFKVKMTTHLKKLKESYCQRQGVPMNSLRFLFEGQRIADNHTPKELGMEEEDVIEVYQE  93

Query  73  LRGG  76
             GG
Sbjct  94  QTGG  97
```

结构相似性

Ubiquitin

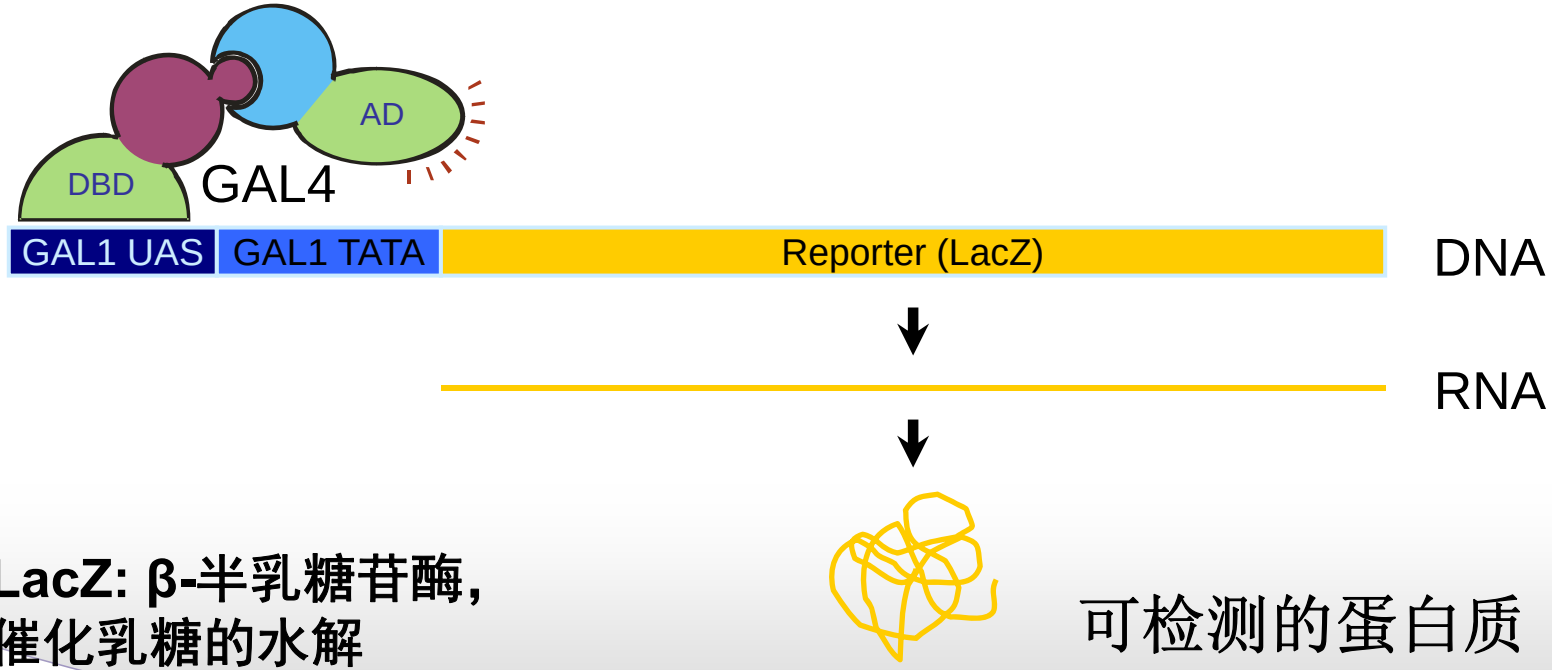


SUMO



蛋白质-蛋白质相互作用

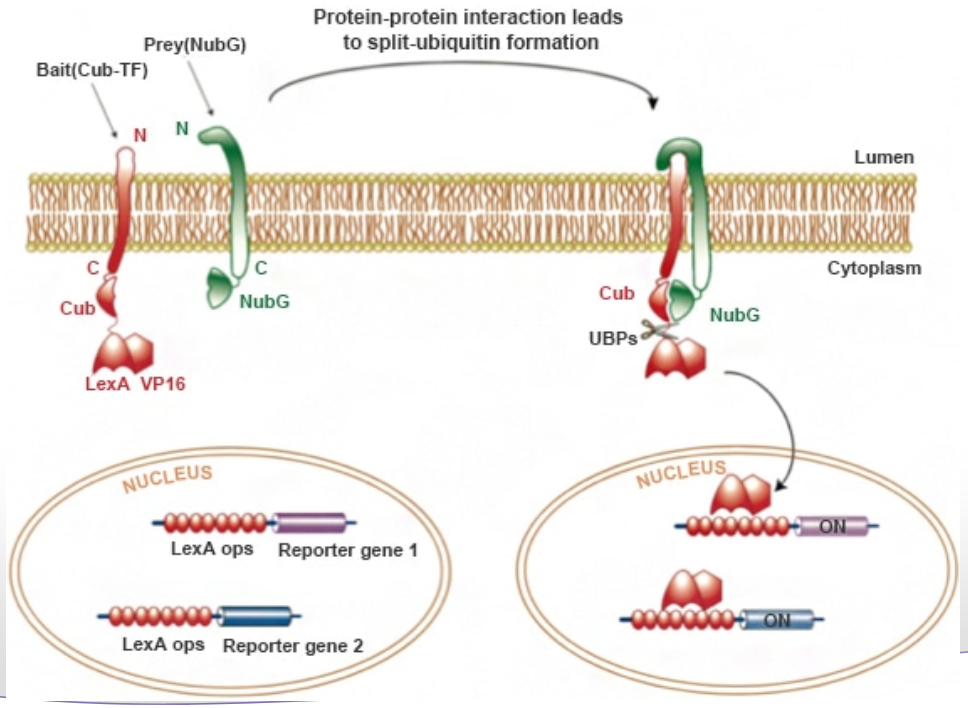
- 酵母双杂交，Yeast two-hybrid, Y2H
  - GAL4-DBD-bait (hybrid 1)
  - GAL4-AD-prey (hybrid 2; single or library)
  - bait-prey相互作用: GAL4激活报告基因



LacZ: β-半乳糖苷酶，
催化乳糖的水解

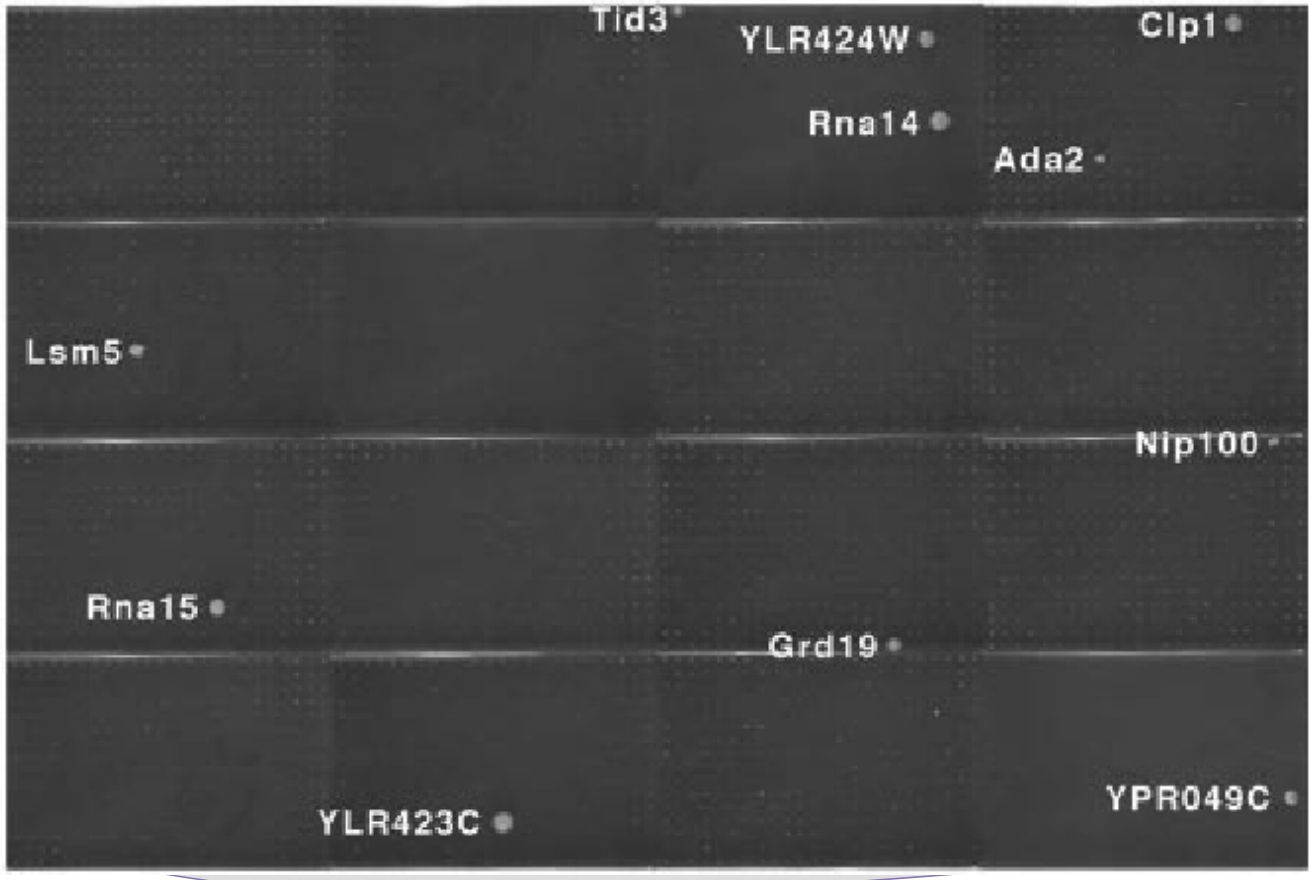
Split-ubiquitin yeast two-hybrid

- 传统酵母双杂交的缺点：只能筛选可溶性蛋白质
- 膜蛋白的相互作用筛选
  - 两个膜蛋白分别与泛素分子的C端片段 (“Cub”, 35–76) 和N端片段 ("Nub", 1–34) 连接
  - Cub与转录因子融合，并且可被泛素蛋白酶切割
  - bait–prey相互作用，泛素分子被切割，转录因子核内诱导报告基因表达



酵母蛋白质-蛋白质相互作用

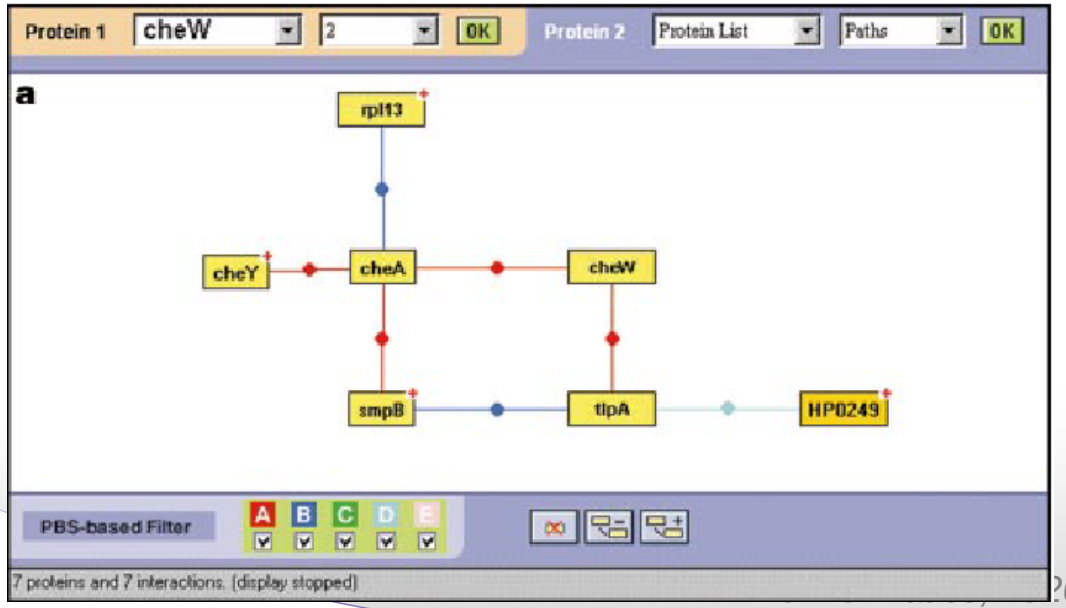
- 2000年，6,000 * 6,000 ORFs
- 1,004个蛋白质，957对相互作用

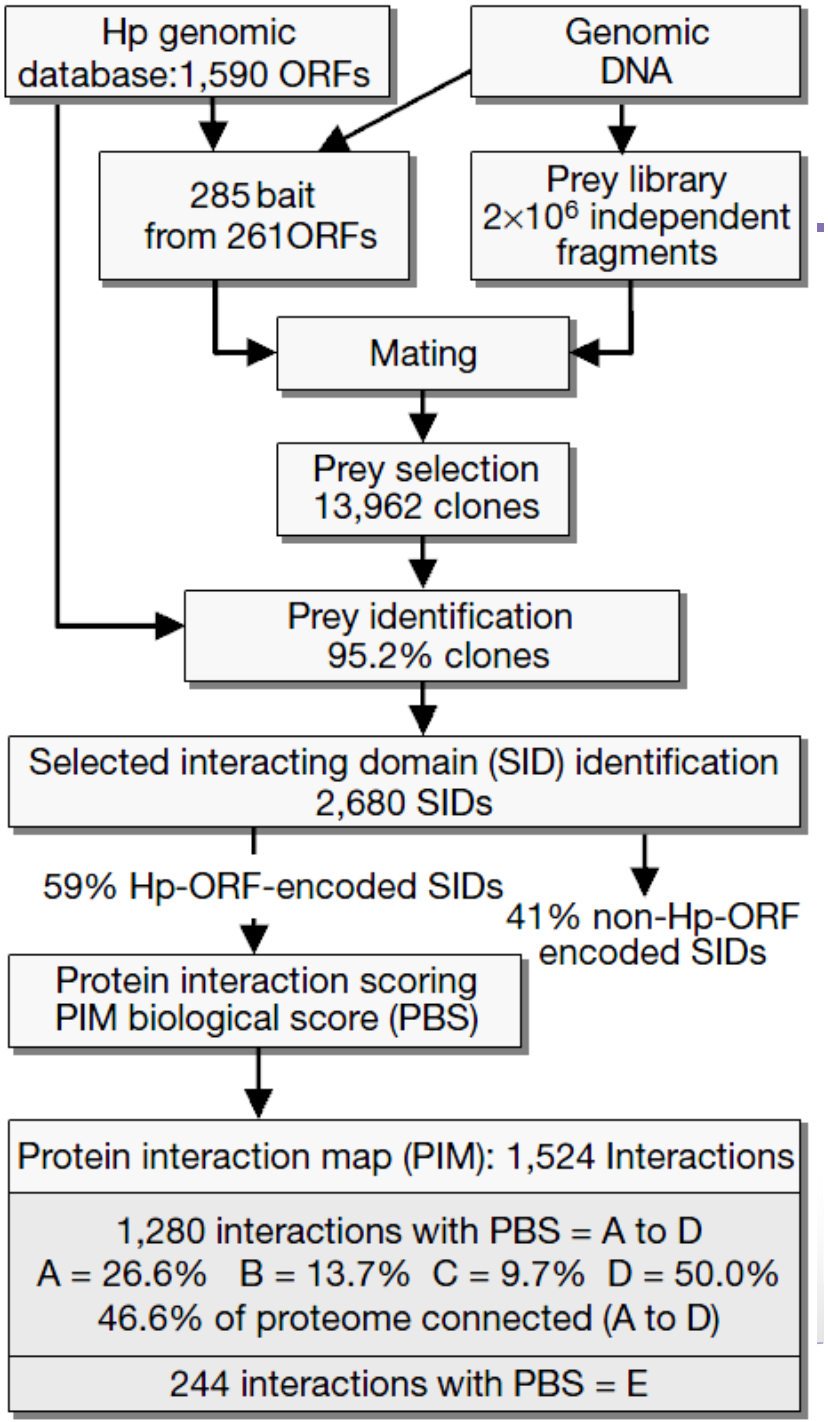


Pcf11

幽门螺旋杆菌

- 2001年，261个蛋白质 * 基因组片段
- ~1,200对相互作用

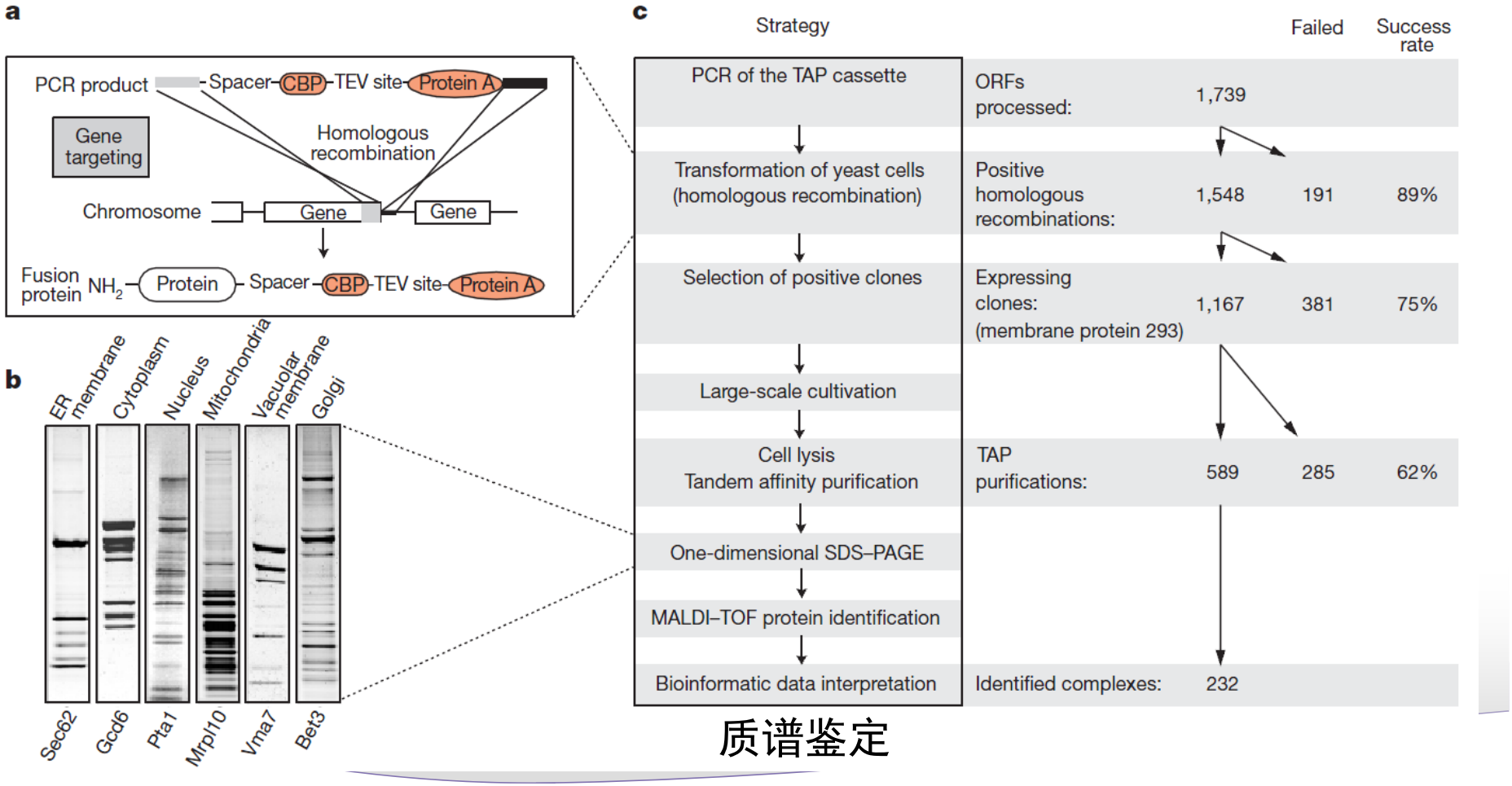




酵母蛋白质复合物

❒ 2002年，酵母蛋白质复合物

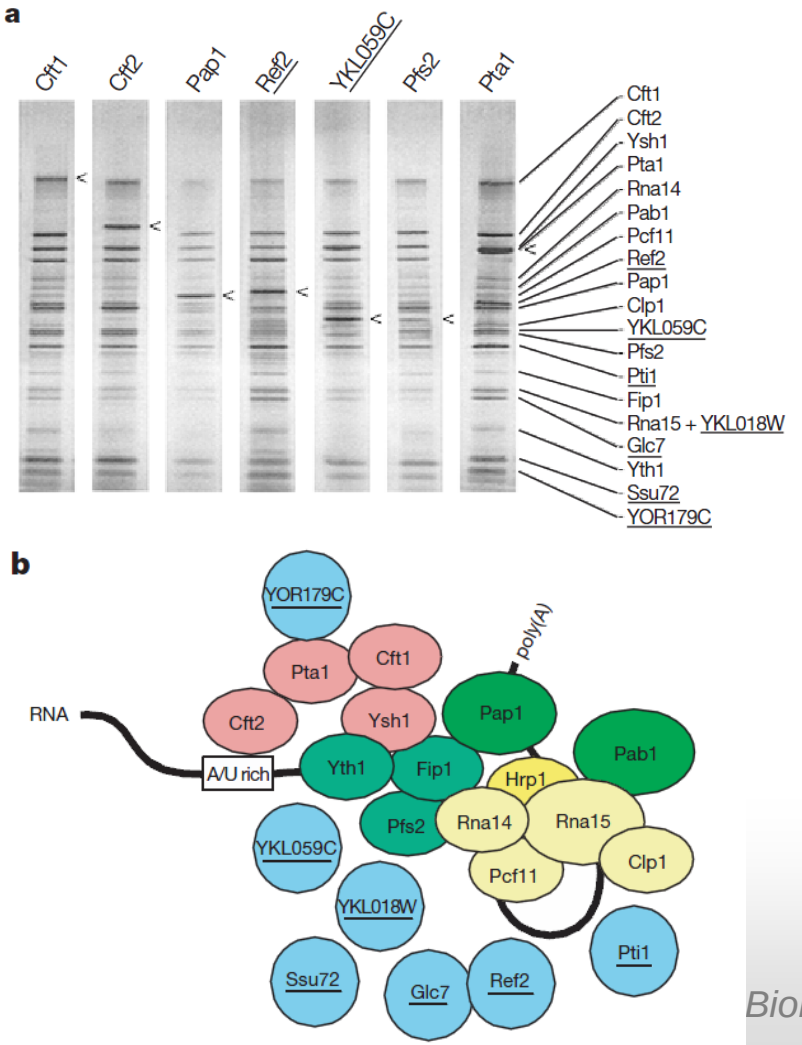
❁ 纯化589个蛋白质，获得232个复合物



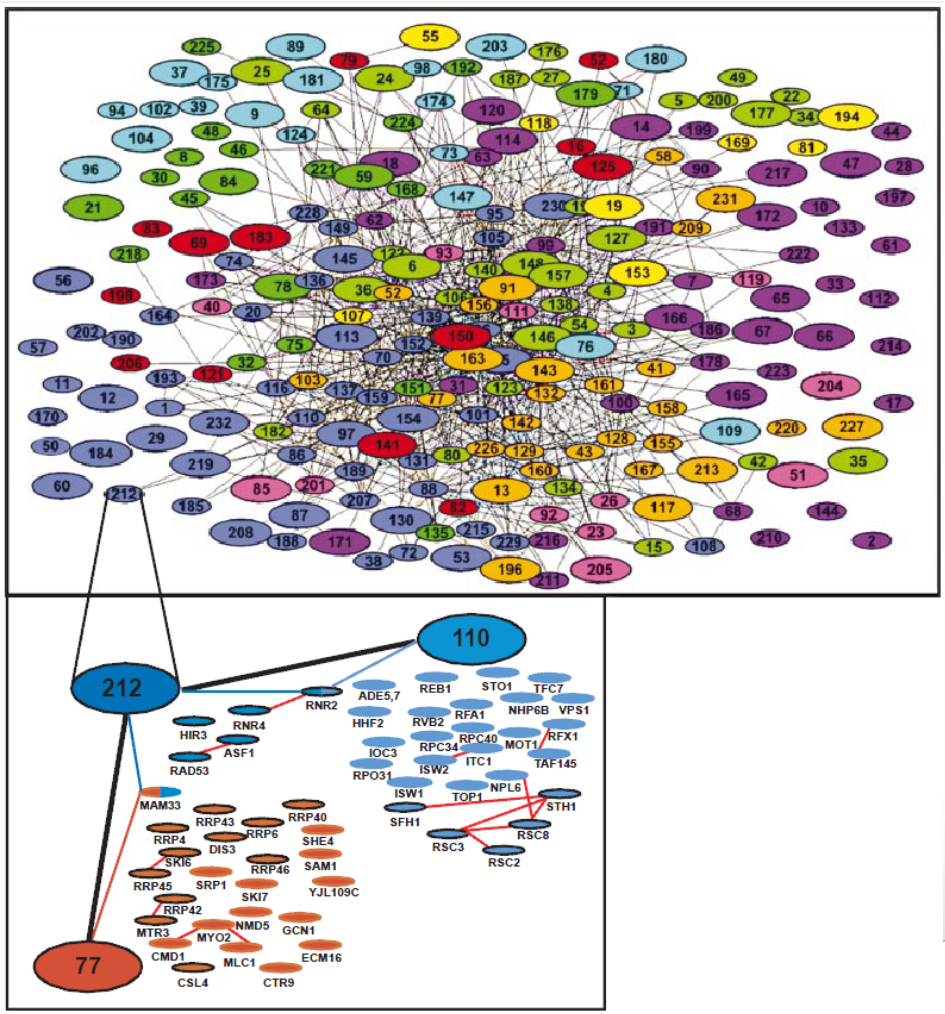
质谱鉴定

酵母蛋白质复合物

❒ 多腺苷酸化的分子机器



❒ 酵母相互作用网络

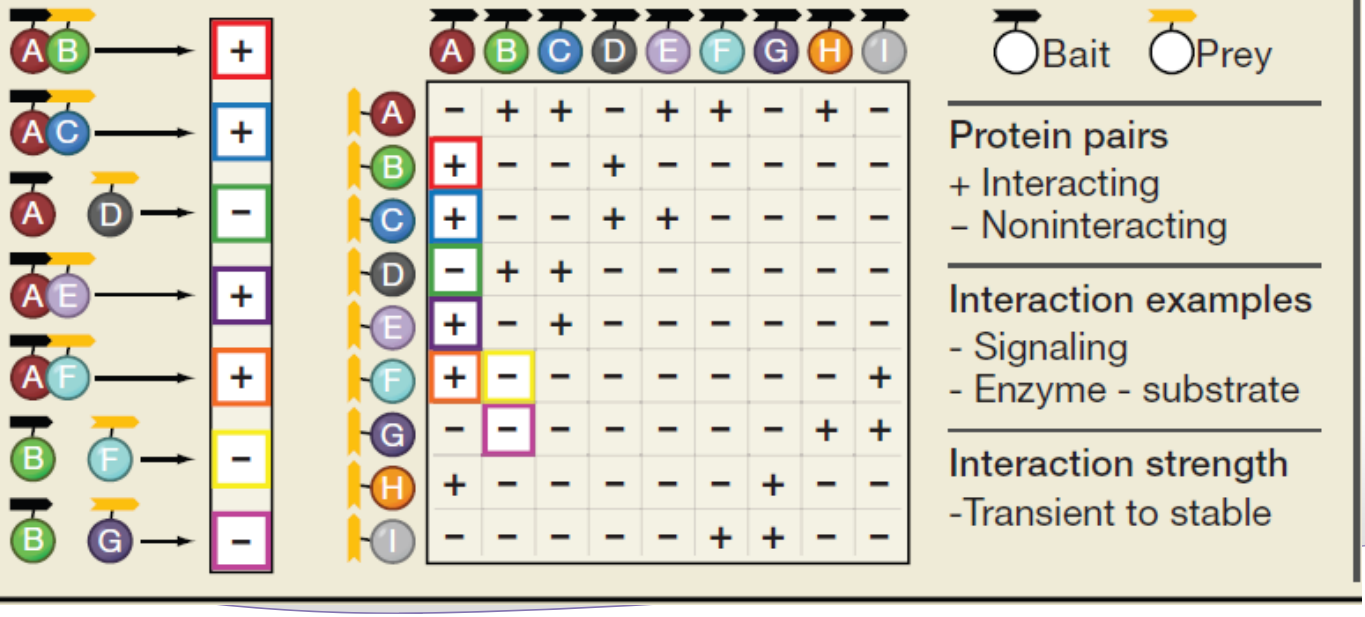


蛋白质-蛋白质相互作用

❒ 酵母双杂交

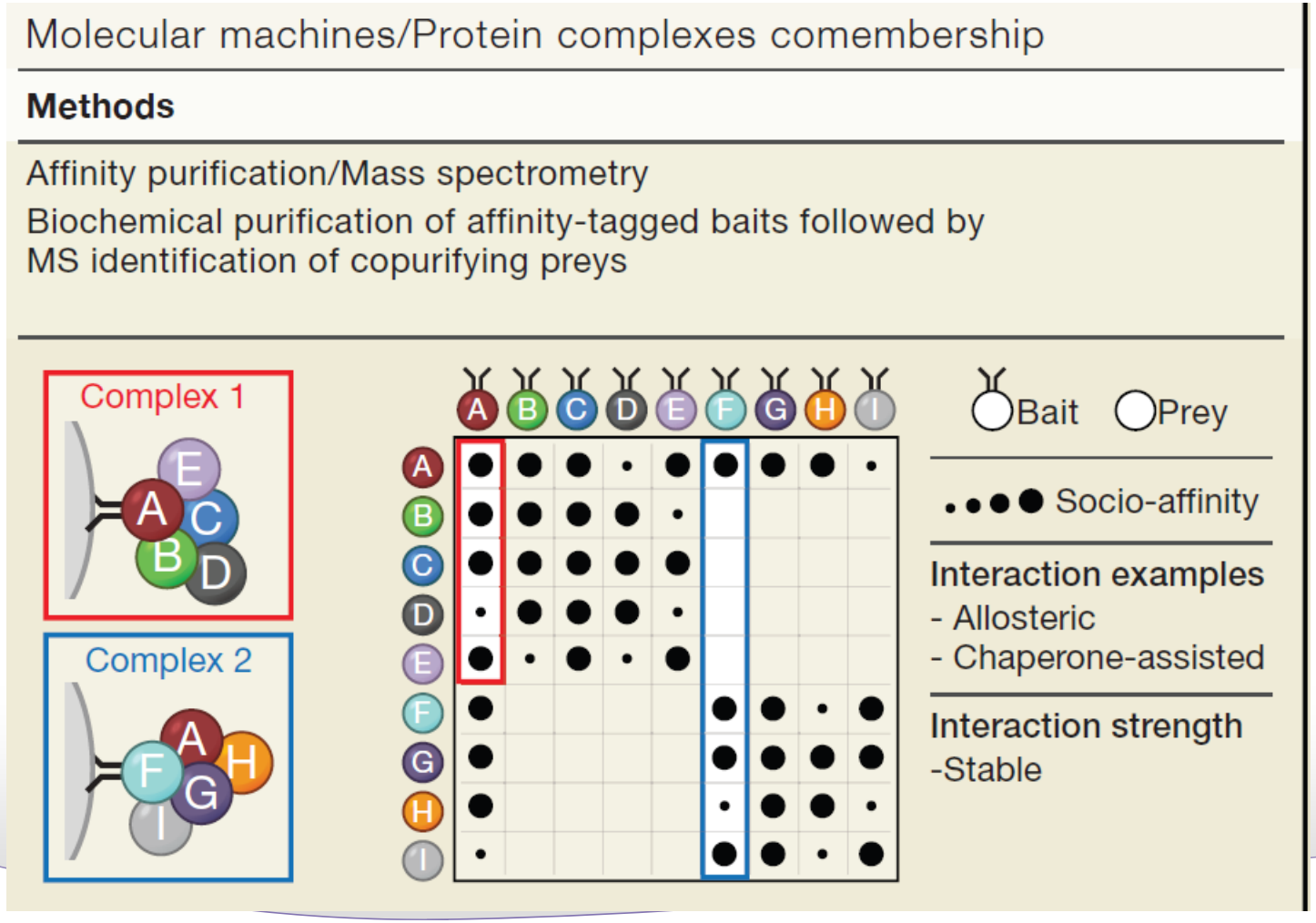
Binary interactions

Methods	Split proteins	Assay/Readout
Yeast two-hybrid Protein fragment complementation assay	Transcription factor, ubiquitin Dehydrofolate reductase GFP or YFP	Transcription Antibiotic resistence Fluorescence



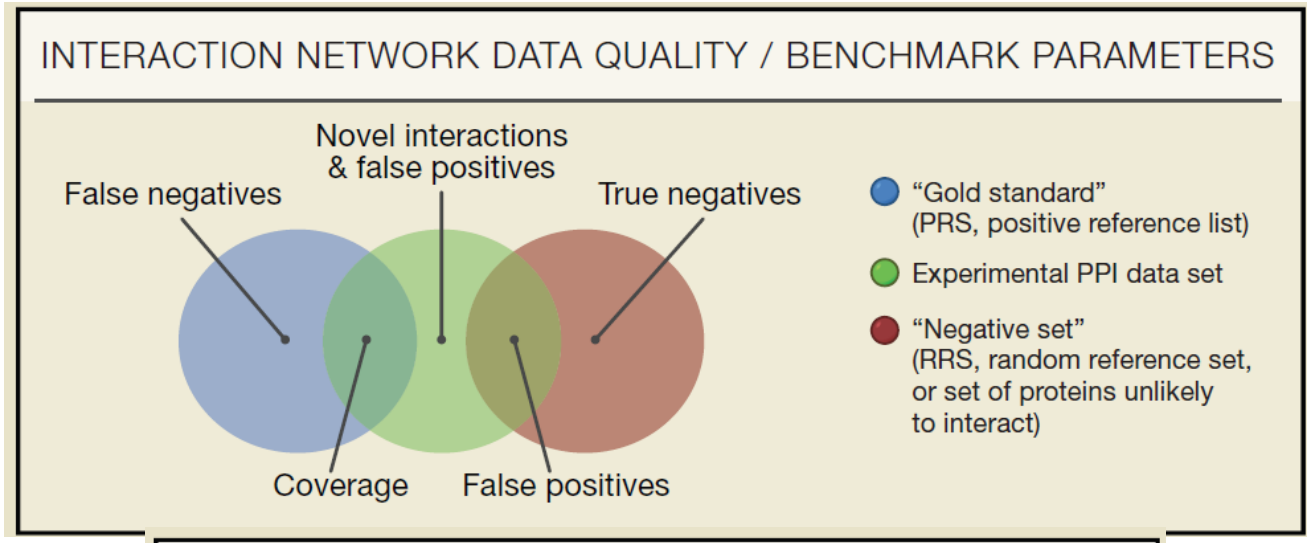
蛋白质-蛋白质相互作用

❒ 亲和纯化/质谱



相互作用网络

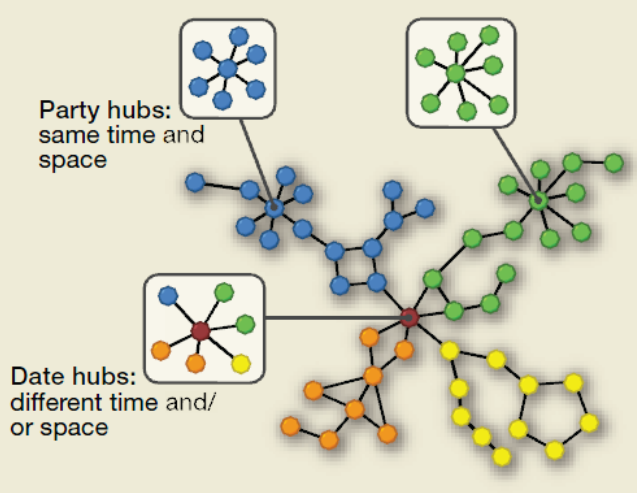
- 相互作用的数据质量，网络拓扑结构



蛋白质-蛋白质相互作用

❒ 网络分析：Hub节点、拓扑结构、参数

NETWORK COMPONENTS



Hub: node with high degree

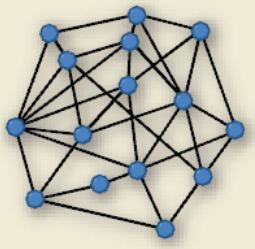
Edge: link between two nodes (interaction)

Node (protein)

Expression profiles and/or localization

NETWORK TOPOLOGIES

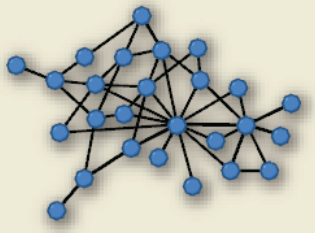
Random network



-Degrees follow Poisson (or peaked) distribution

-Vulnerability to failure

Scale-free network
(Biological/cellular networks)

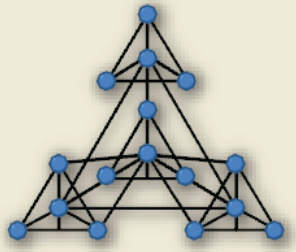


-Degrees follow power-law distributions

-Robustness against random failure

-Vulnerability to targeted attacks

Hierarchical network
(Many types of real networks)



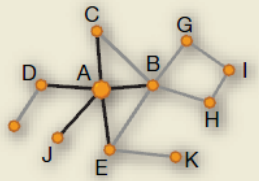
-Degrees follow power-law distributions

-Account for modularity, local clustering, and scale-free topology

-High clustering coefficient (C)

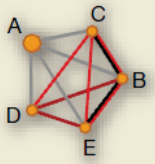
NETWORK MEASURES

Degree/ connectivity (k)



k_A=Nb of edges through A=5

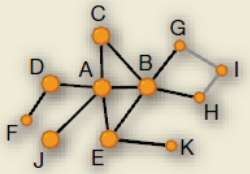
Clustering coefficient/ interconnectivity (C)



$$C_A = \frac{\text{Actual links between A's neighbors (black)}}{\text{Possible links between A's neighbors (orange)}}$$

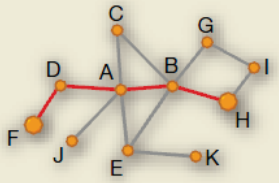
$C_A=n_A/[k_A(k_A-1)/2]$
$=2/[4x(4-1)/2]=0.333$

Assortativity/average nearest neighbor's connectivity (NC)



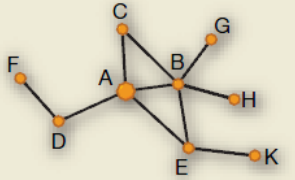
$NC_A=(k_B+k_C+k_D+k_E+k_J)/5$
$=(5+2+2+3+1)/5=2.6$

Shortest path (SP) between two nodes



SP_{FH}=(F,D,A,B,H)=4

Betweenness/ centrality (B)



B_4=Fraction of SPs passing through A
=0.446

Betweenness centrality

❒ 介数中心性

Betweenness centrality

The **betweenness centrality** [4] $C_b(n)$ of a node n is computed as follows:

$C_b(n) = \sum_{s \neq n \neq t} (\sigma_{st}(n) / \sigma_{st})$,

where s and t are nodes in the network different from n, σ_{st} denotes the number of shortest paths from s to t, and $\sigma_{st}(n)$ is the number of shortest paths from s to t that n lies on.

Betweenness centrality is computed only for networks that do not contain multiple edges. The betweenness value for each node n is normalized by dividing by the number of node pairs excluding n: $(N\text{-}1)(N\text{-}2)/2$, where N is the total number of nodes in the connected component that n belongs to. Thus, the betweenness centrality of each node is a number between 0 and 1.

For example, the betweenness centrality of node b in Figure 7 is computed as follows:

$C_b(b) = ((\sigma_{ac}(b) / \sigma_{ac}) + (\sigma_{ad}(b) / \sigma_{ad}) + (\sigma_{ae}(b) / \sigma_{ae}) + (\sigma_{cd}(b) / \sigma_{cd}) + (\sigma_{ce}(b) / \sigma_{ce}) + (\sigma_{de}(b) / \sigma_{de})) / 6$

$= ((1 / 1) + (1 / 1) + (2 / 2) + (1 / 2) + 0 + 0) / 6 = 3.5 / 6 \approx 0.583$

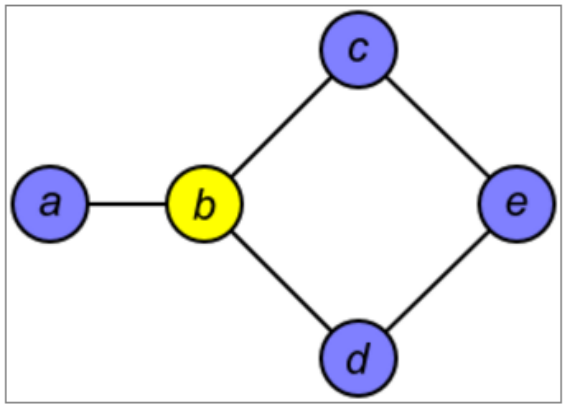


Figure 7 *Example network with five nodes and five edges.*

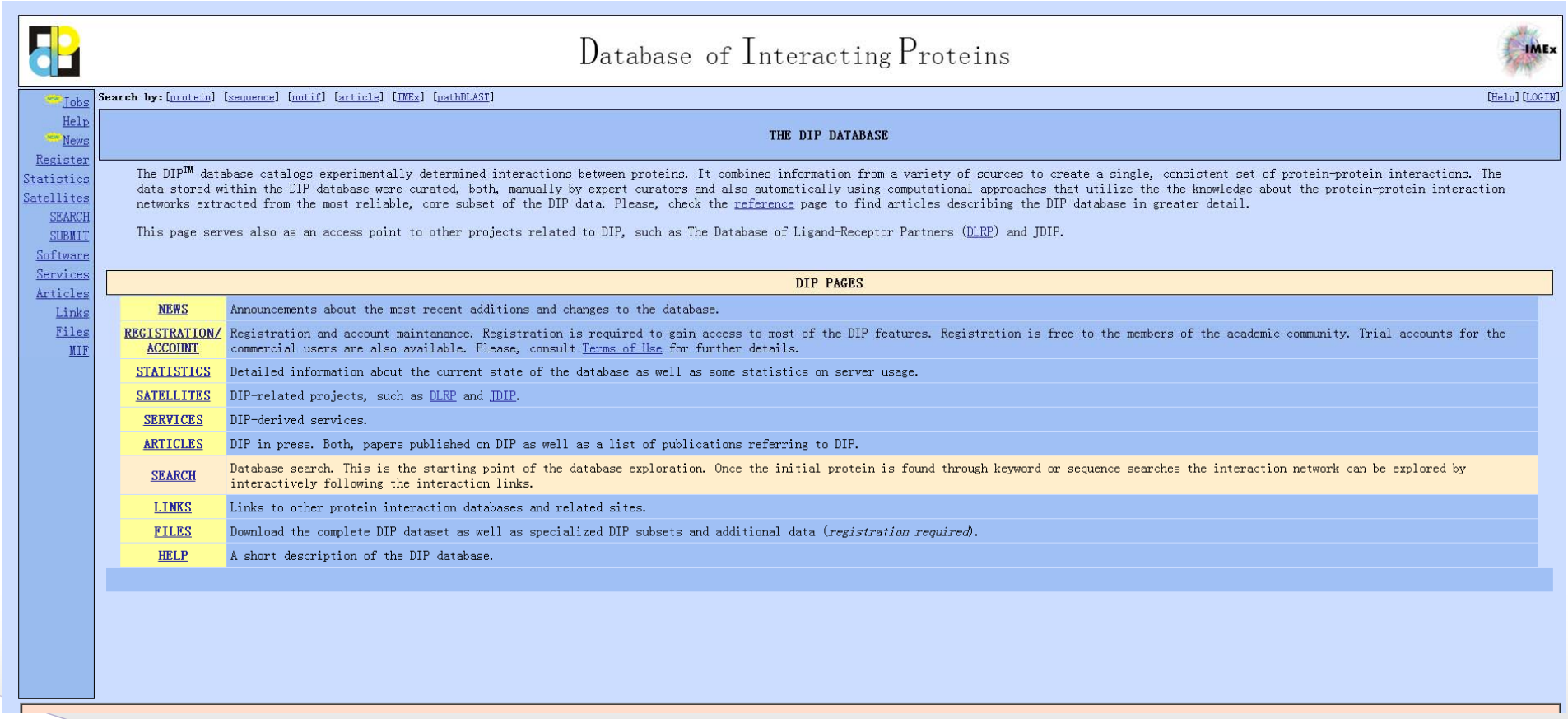
http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/index.html#nodeBetween

蛋白质相互作用的预测

- Biocuration & Literature mining
- 基因组信息 (genomic context method)
  - Gene fusion and fission
  - Conservation of gene order/bidirectional pairs
  - Phylogenetic profile
  - 关联序列特征 (Correlated sequence signatures)
  - mRNA co-expression
- Interolog: 直系同源的相互作用

DIP

- 2000年，David Eisenberg研究组



Database of Interacting Proteins

Jobs
Help
News
Register
Statistics
Satellites
SEARCH
SUBMIT
Software
Services
Articles
Links
Files
MIF

Search by: [protein] [sequence] [motif] [article] [IMEx] [pathBLAST]

[Help] [LOGIN]

THE DIP DATABASE

The DIP™ database catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated, both, manually by expert curators and also automatically using computational approaches that utilize the the knowledge about the protein-protein interaction networks extracted from the most reliable, core subset of the DIP data. Please, check the reference page to find articles describing the DIP database in greater detail.

This page serves also as an access point to other projects related to DIP, such as The Database of Ligand-Receptor Partners (DLRP) and JDIP.

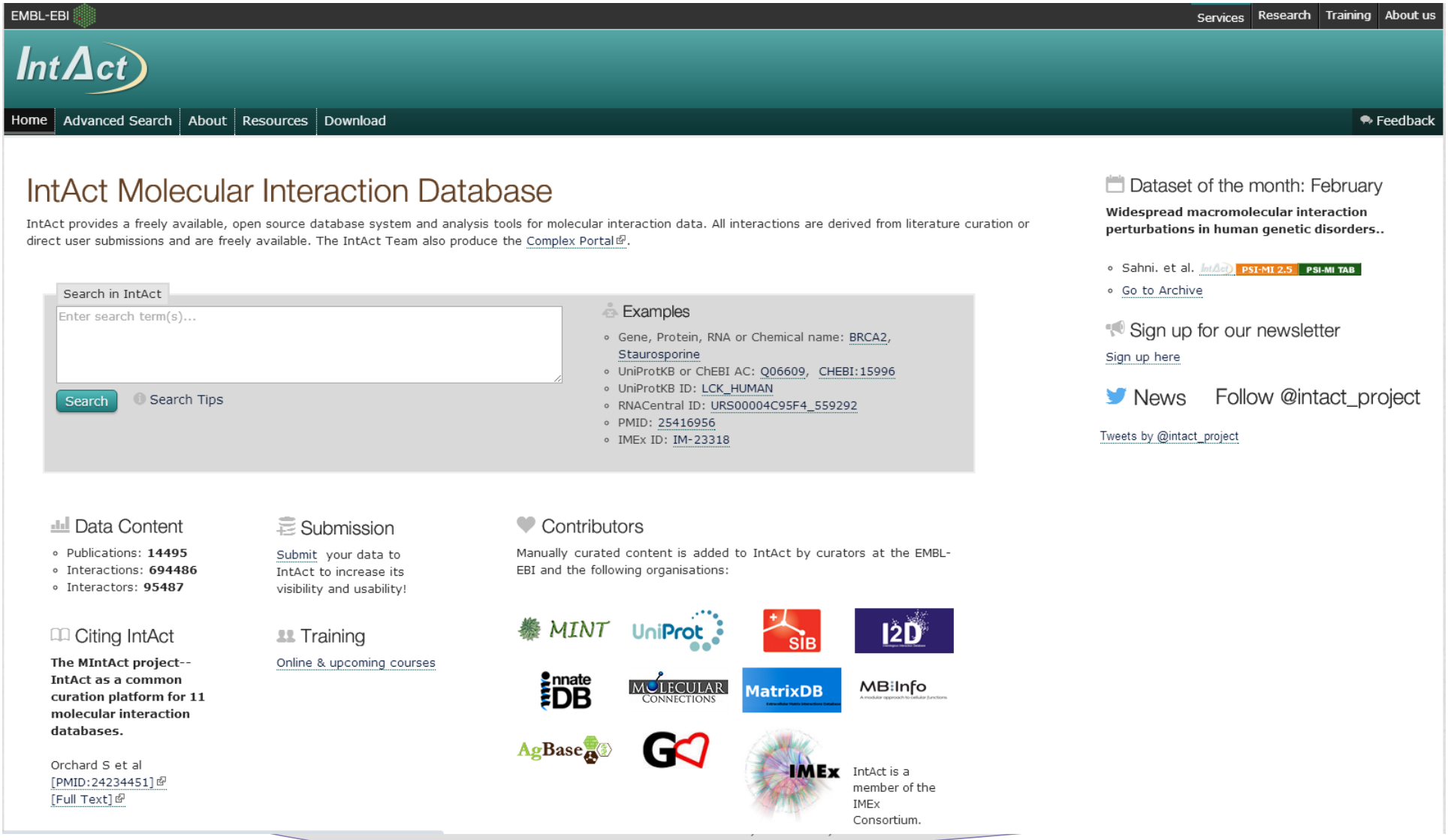
DIP PAGES

NEWS	Announcements about the most recent additions and changes to the database.
REGISTRATION/ ACCOUNT	Registration and account maintanance. Registration is required to gain access to most of the DIP features. Registration is free to the members of the academic community. Trial accounts for the commercial users are also available. Please, consult Terms of Use for further details.
STATISTICS	Detailed information about the current state of the database as well as some statistics on server usage.
SATELLITES	DIP-related projects, such as DLRP and JDIP.
SERVICES	DIP-derived services.
ARTICLES	DIP in press. Both, papers published on DIP as well as a list of publications referring to DIP.
SEARCH	Database search. This is the starting point of the database exploration. Once the initial protein is found through keyword or sequence searches the interaction network can be explored by interactively following the interaction links.
LINKS	Links to other protein interaction databases and related sites.
FILES	Download the complete DIP dataset as well as specialized DIP subsets and additional data (*registration required*).
HELP	A short description of the DIP database.

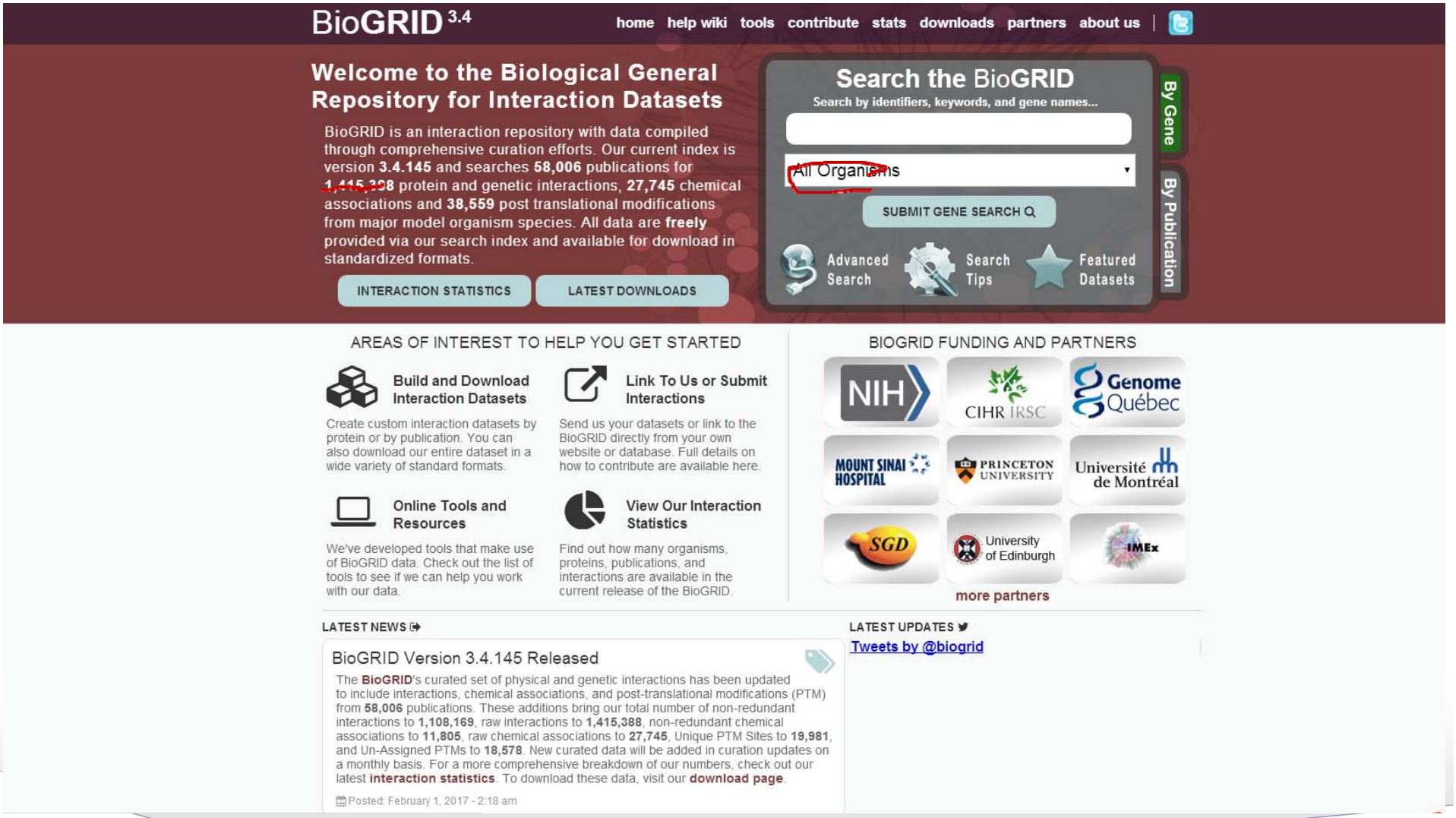
MINT

❒ 2002年，Molecular INTeraction

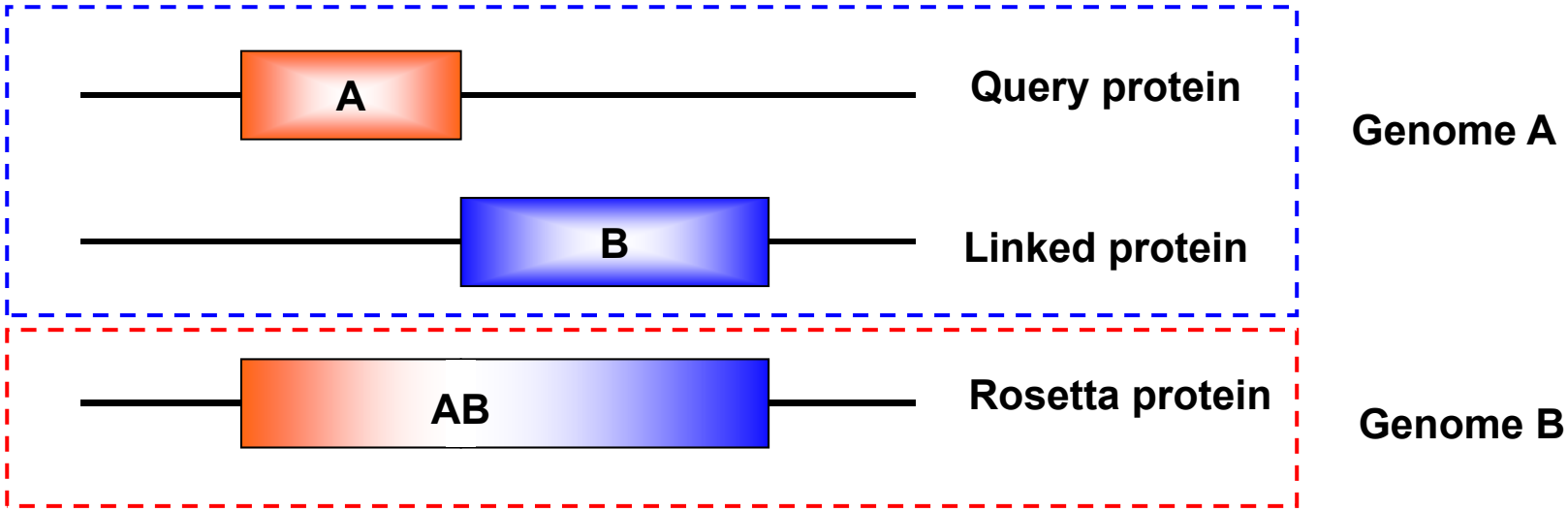
IntAct


EMBL-EBI
Services
Research
Training
About us
IntAct
Home
Advanced Search
About
Resources
Download
Feedback
IntAct Molecular Interaction Database
IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available. The IntAct Team also produce the Complex Portal.
Search in IntAct
Enter search term(s)...
Search
Search Tips
Examples
Gene, Protein, RNA or Chemical name: BRCA2, Staurosporine
UniProtKB or ChEBI AC: Q06609, CHEBI:15996
UniProtKB ID: LCK_HUMAN
RNACentral ID: URS00004C95F4_559292
PMID: 25416956
IMEx ID: IM-23318
Dataset of the month: February
Widespread macromolecular interaction perturbations in human genetic disorders..
Sahni. et al.
PSI-MI 2.5
PSI-MI TAB
Go to Archive
Sign up for our newsletter
Sign up here
News
Follow @intact_project
Tweets by @intact_project
Data Content
Publications: 14495
Interactions: 694486
Interactors: 95487
Submission
Submit your data to IntAct to increase its visibility and usability!
Contributors
Manually curated content is added to IntAct by curators at the EMBL-EBI and the following organisations:
Citing IntAct
The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.
Orchard S et al
[PMID:24234451]
[Full Text]
Training
Online & upcoming courses
IntAct is a member of the IMEx Consortium.

BioGRID

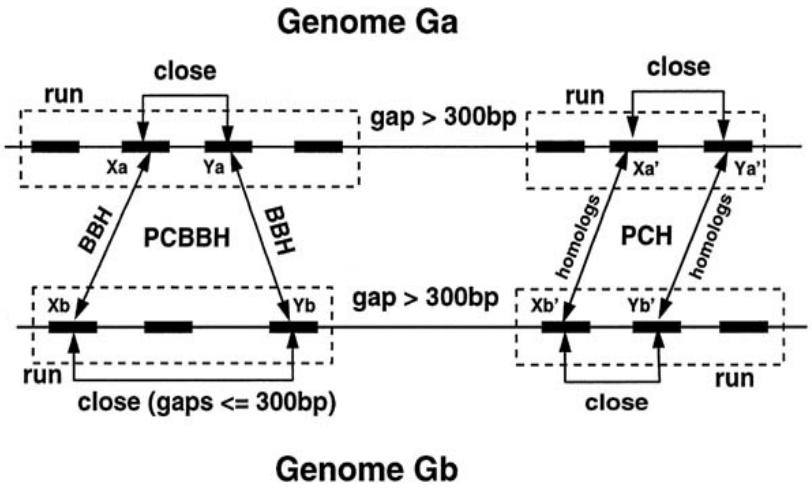

BioGRID 3.4
home help wiki tools contribute stats downloads partners about us
Welcome to the Biological General Repository for Interaction Datasets
BioGRID is an interaction repository with data compiled through comprehensive curation efforts. Our current index is version 3.4.145 and searches 58,006 publications for 1,415,388 protein and genetic interactions, 27,745 chemical associations and 38,559 post translational modifications from major model organism species. All data are freely provided via our search index and available for download in standardized formats.
INTERACTION STATISTICS
LATEST DOWNLOADS
Search the BioGRID
Search by identifiers, keywords, and gene names...
All Organisms
SUBMIT GENE SEARCH
By Gene
By Publication
Advanced Search
Search Tips
Featured Datasets
AREAS OF INTEREST TO HELP YOU GET STARTED
Build and Download Interaction Datasets
Create custom interaction datasets by protein or by publication. You can also download our entire dataset in a wide variety of standard formats.
Link To Us or Submit Interactions
Send us your datasets or link to the BioGRID directly from your own website or database. Full details on how to contribute are available here.
Online Tools and Resources
We've developed tools that make use of BioGRID data. Check out the list of tools to see if we can help you work with our data.
View Our Interaction Statistics
Find out how many organisms, proteins, publications, and interactions are available in the current release of the BioGRID.
BIOGRID FUNDING AND PARTNERS
NIH
CIHR IRSC
Genome Québec
MOUNT SINAI HOSPITAL
PRINCETON UNIVERSITY
Université de Montréal
SGD
University of Edinburgh
IMEx
more partners
LATEST NEWS
BioGRID Version 3.4.145 Released
The BioGRID's curated set of physical and genetic interactions has been updated to include interactions, chemical associations, and post-translational modifications (PTM) from 58,006 publications. These additions bring our total number of non-redundant interactions to 1,108,169, raw interactions to 1,415,388, non-redundant chemical associations to 11,805, raw chemical associations to 27,745, Unique PTM Sites to 19,981, and Un-Assigned PTMs to 18,578. New curated data will be added in curation updates on a monthly basis. For a more comprehensive breakdown of our numbers, check out our latest interaction statistics. To download these data, visit our download page.
Posted: February 1, 2017 - 2:18 am
LATEST UPDATES
Tweets by @biogrid

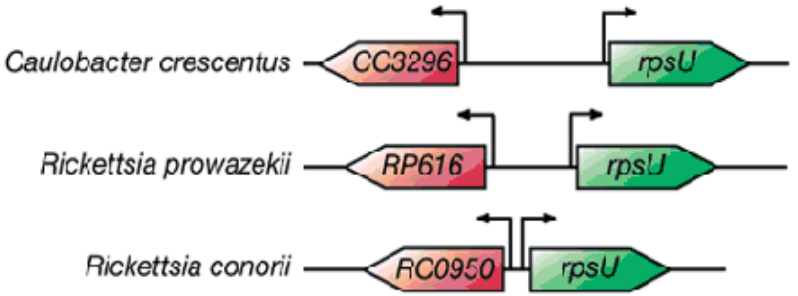
Gene fusion/fission: Rosetta Stone



Marcotte EM *et al.*, ***Science*** 1999, 285:751-753;
Enright AJ *et al.*, ***Nature***, 1999, 402:86-90

Conservation of gene order/ bidirectional pairs



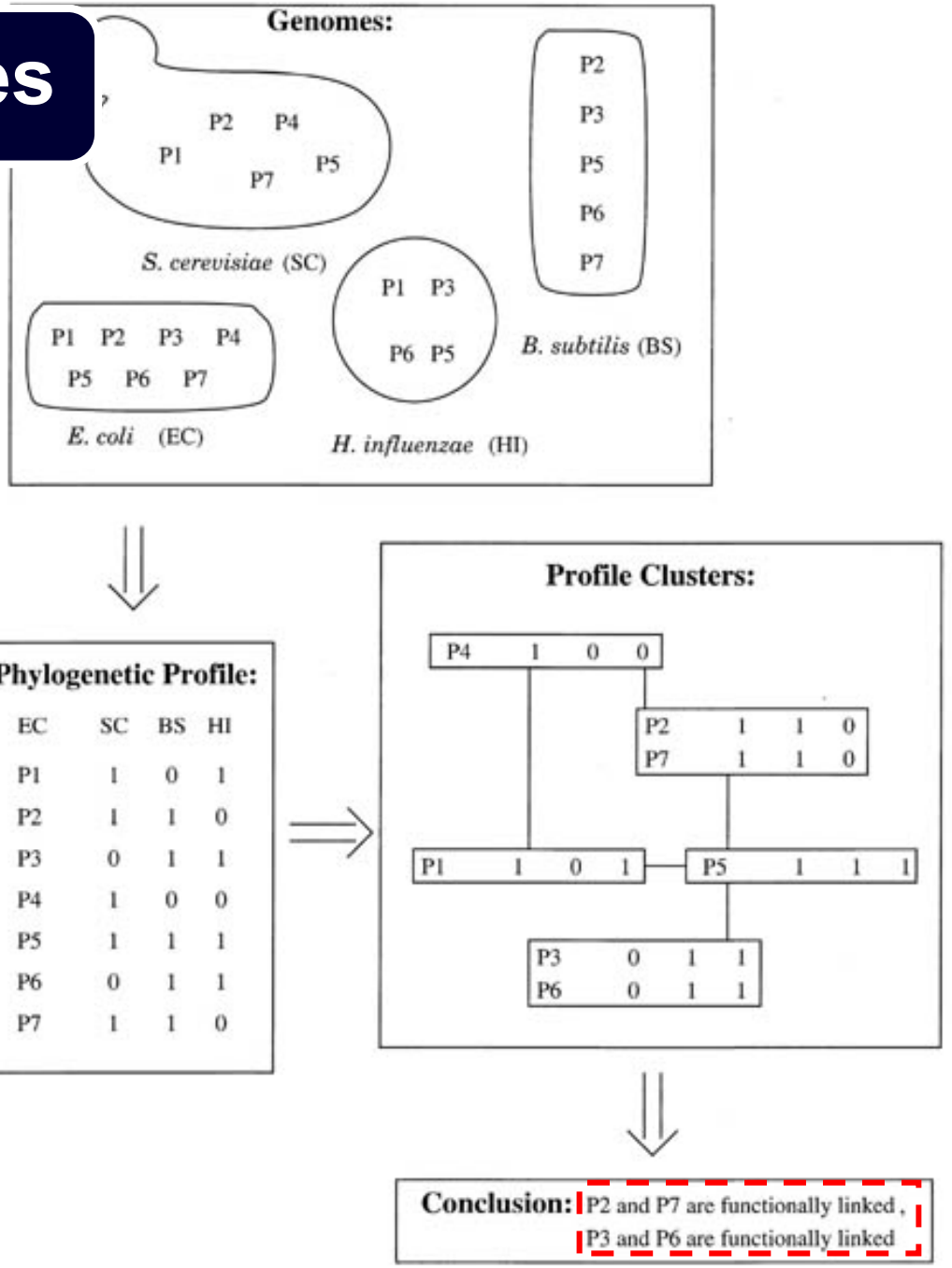


Gene order pairs

Bidirectional transcribed gene pairs

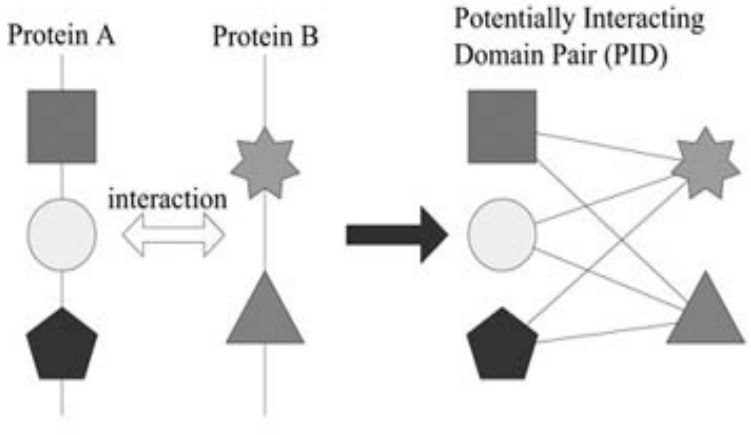
Dandekar T *et al.*, ***TIBS***, 1998, 23:324-328;
Overbeek R *et al.*, ***PNAS***, 1999, 96:2896-2901;
Korbel JO *et al.*, ***NBT***, 2004, 22:911-917

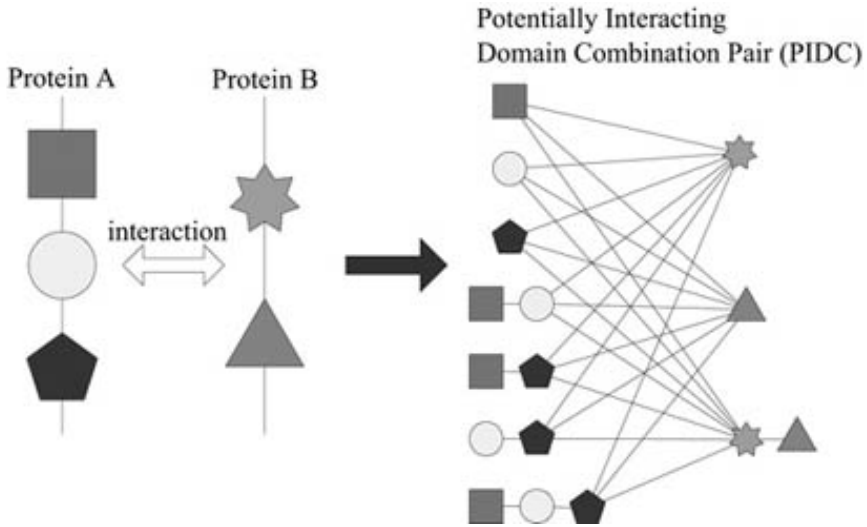
Phylogenetic profiles



Pellegrini M et al., PNAS, 1999, 96:4285-4288;
Huynen MA et al., PNAS, 1998, 95:5849-5856

Correlated sequence signatures





PID model

This model is computationally faster and more convenient

PIDC model

PID模型：最大似然性法

Two proteins P_i, P_j have *m, n* Interpro domain (I), then the probability of P_i and P_j to be interacting pair is shown below:

$$P(PPI_{ij}=1)=1-\prod_{(I_m,I_n)\in(P_i\times P_j)}(1-P(I_{mn}=1))$$

PPI_{ij}=1: Protein P_i interacts with Protein P_j.

I_{mn}= 1: Interpro annotation I_m and I_n are interacting functional domain.

m, n: Numbers of Interpro annotations in P_i and P_j, respectively.

$(I_{m'}, I_{n'}) \in (P_i \times P_j)$: Interpro pair ($I_m$, I_n) is included in protein pair $Pi\times Pj$.

P(I_{mn} = 1): 训练

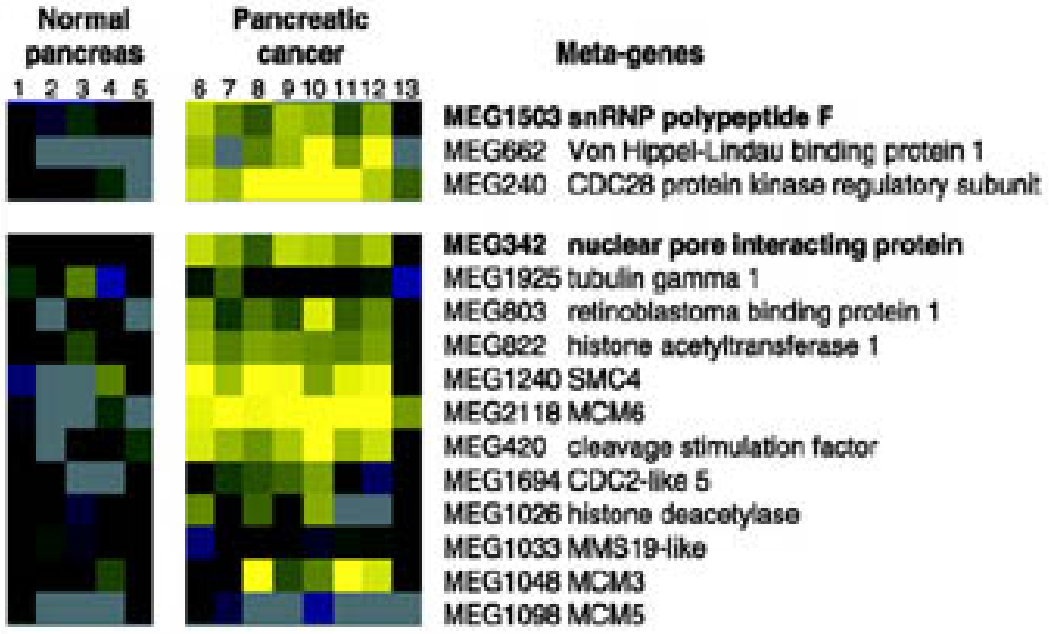
The $P(I_{mn}=1)$ could be obtained from training non-redundant PPI data set. And the equation for calculating the two *probabilities* could be proposed as:

$$P(I_{mn}=1)=\frac{Int_{mn}}{N_{mn}}$$

Int_{mn} : Number of PPIs that include (I_k, I_l);

N_{mn}: Number of all potential PPIs that include (I_k, I_l).

Conserved co-expression



Stuart JM *et al.*, ***Science***, 2003, 302:249-255;
von Mering C *et al.*, ***NAR***, 2005, 33:D433-D437

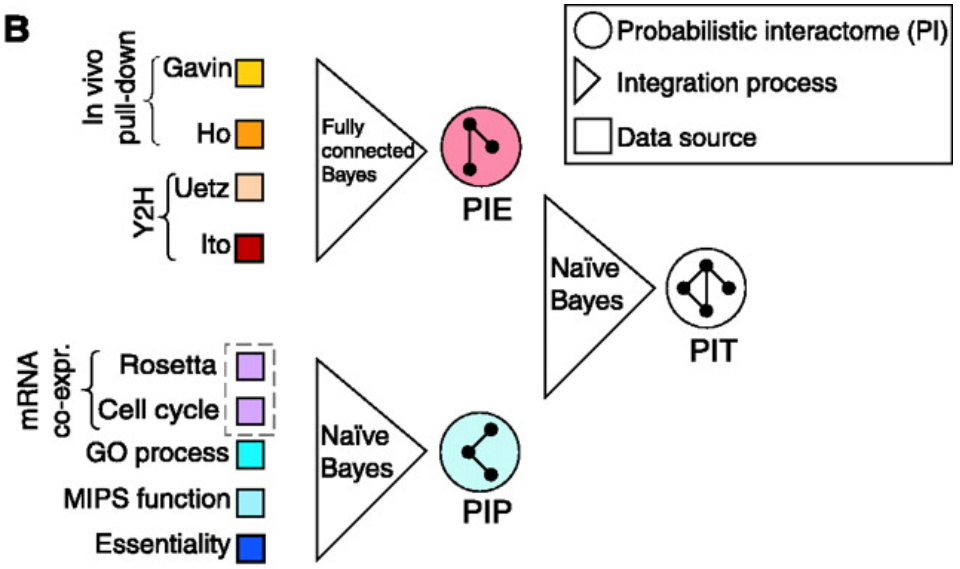
方法整合：贝叶斯算法

❒ 2003年，Mark Gerstein

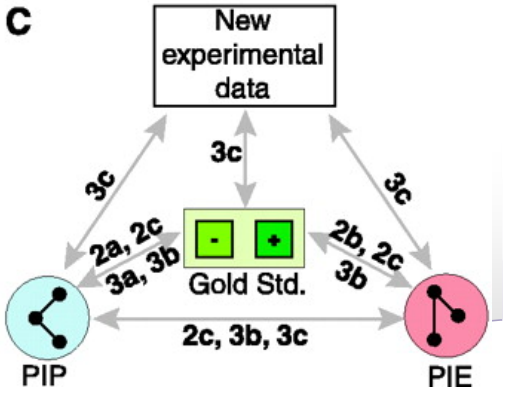
A

Data type	Dataset		# protein pairs	Used for ...
Experimental interaction data	In-vivo pull-down	Gavin et al.	31,304	Integration of experimental interaction data (PIE)
		Ho et al.	25,333	
	Yeast two-hybrid	Uetz et al.	981	
		Ito et al.	4,393	
Other genomic features	mRNA Expression	Rosetta compendium	19,334,806	De novo prediction (PIP)
		Cell cycle	17,467,005	
	Biological function	GO biological process	3,146,286	
		MIPS function	6,161,805	
	Essentiality		8,130,528	
Gold standards	Positives	Proteins in the same MIPS complex	8,250	Training & testing
	Negatives	Proteins separated by localization	2,708,746	

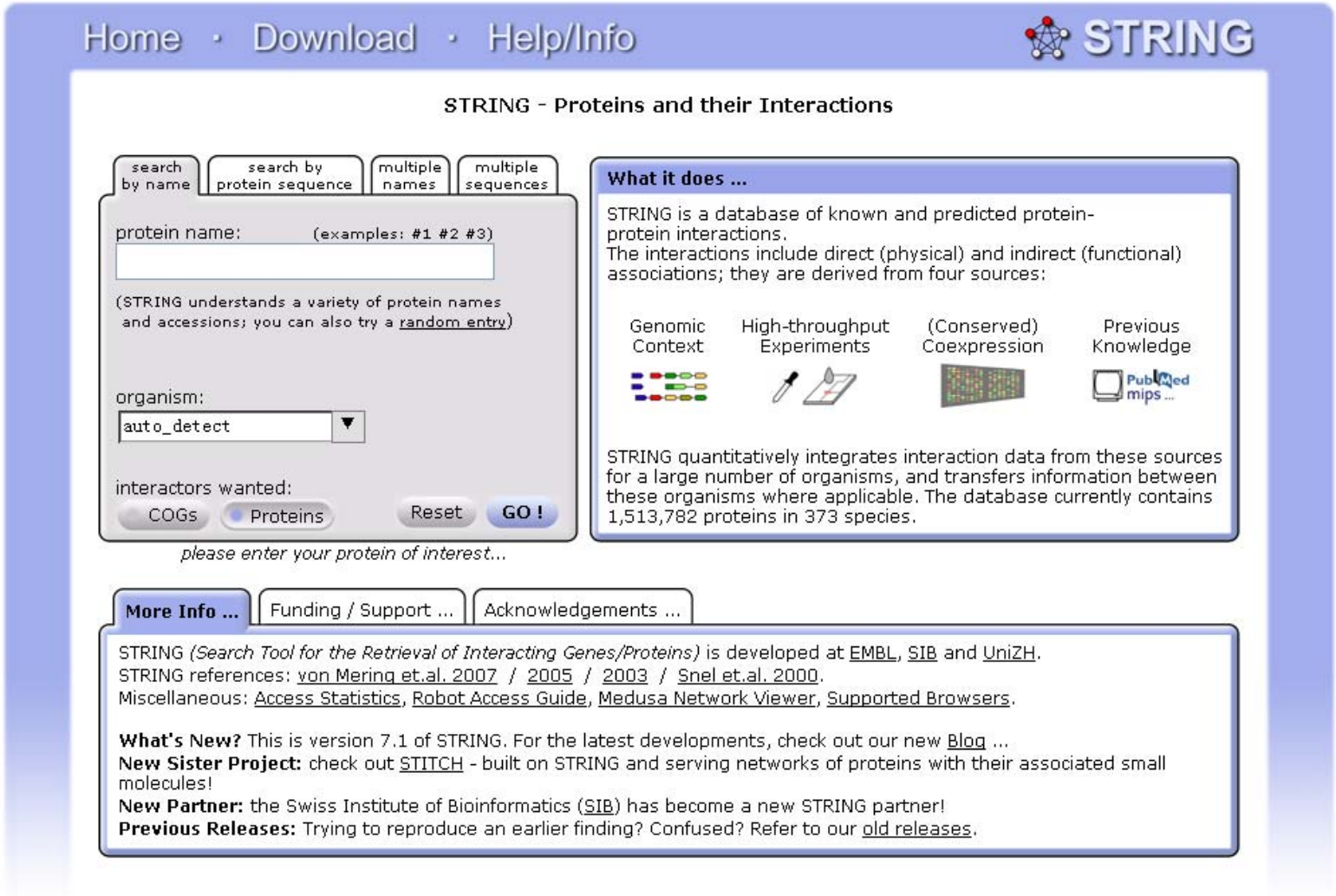
B



C

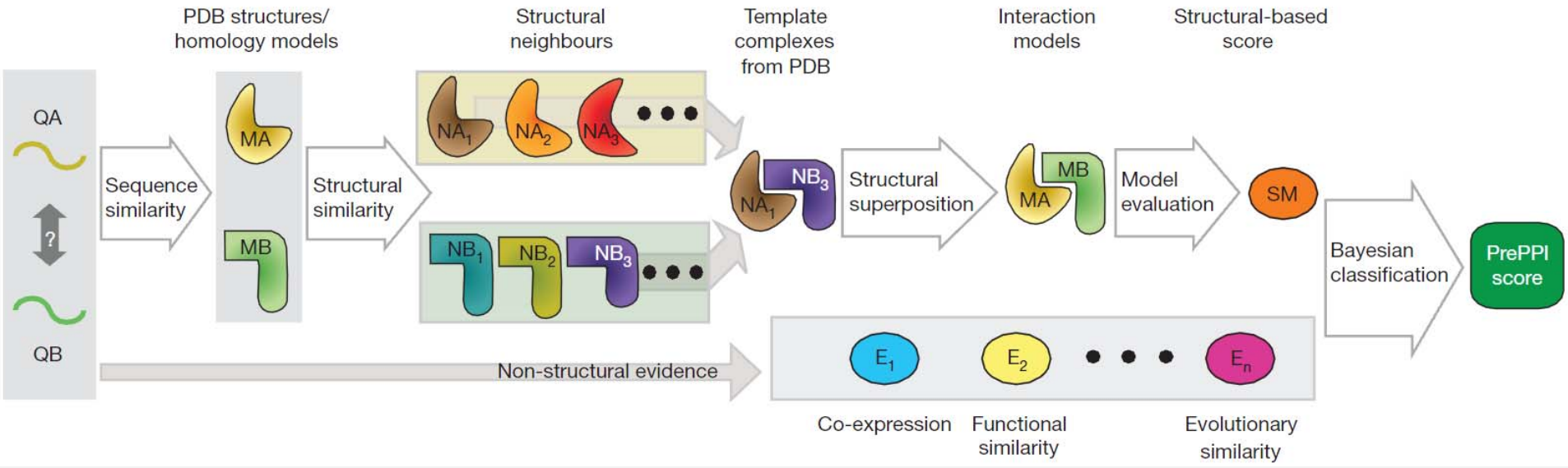


STRING: 方法的整合



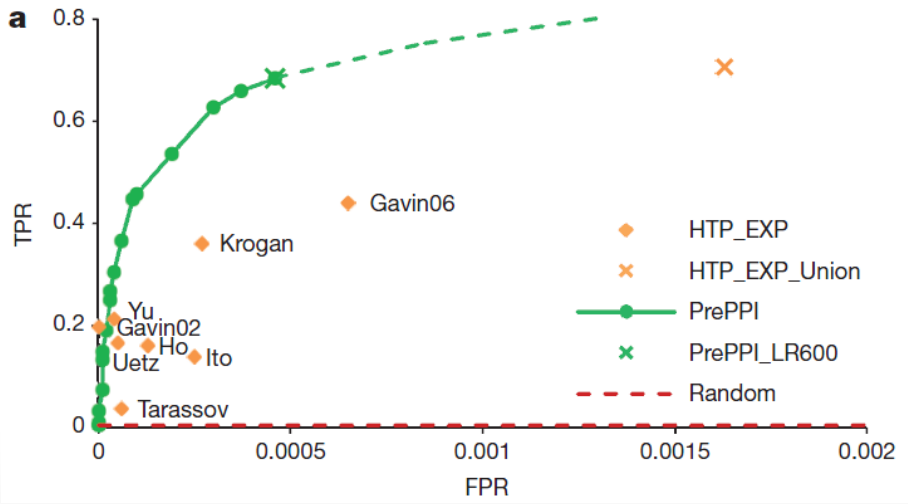
PrePPI

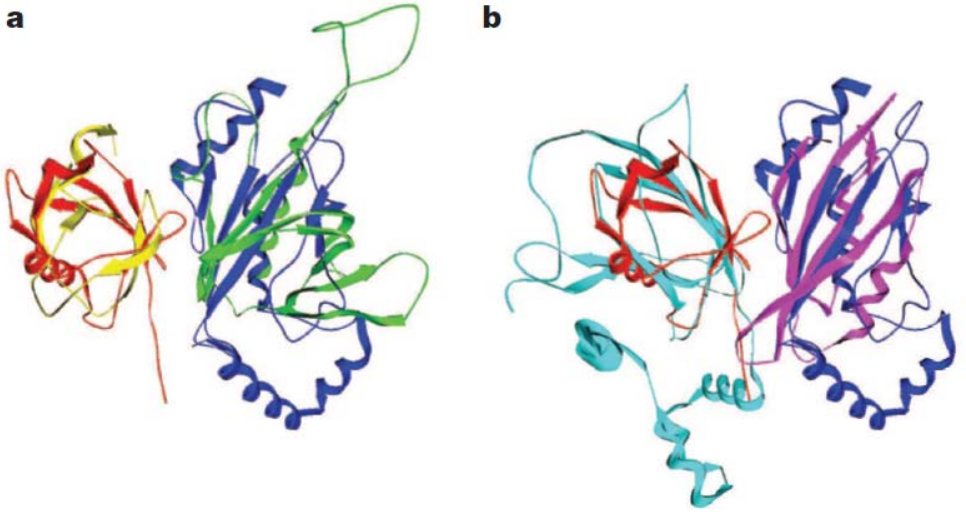
- 2012年，张强锋
  - >30,000酵母PPI
  - >300,000人类PPI



PrePPI

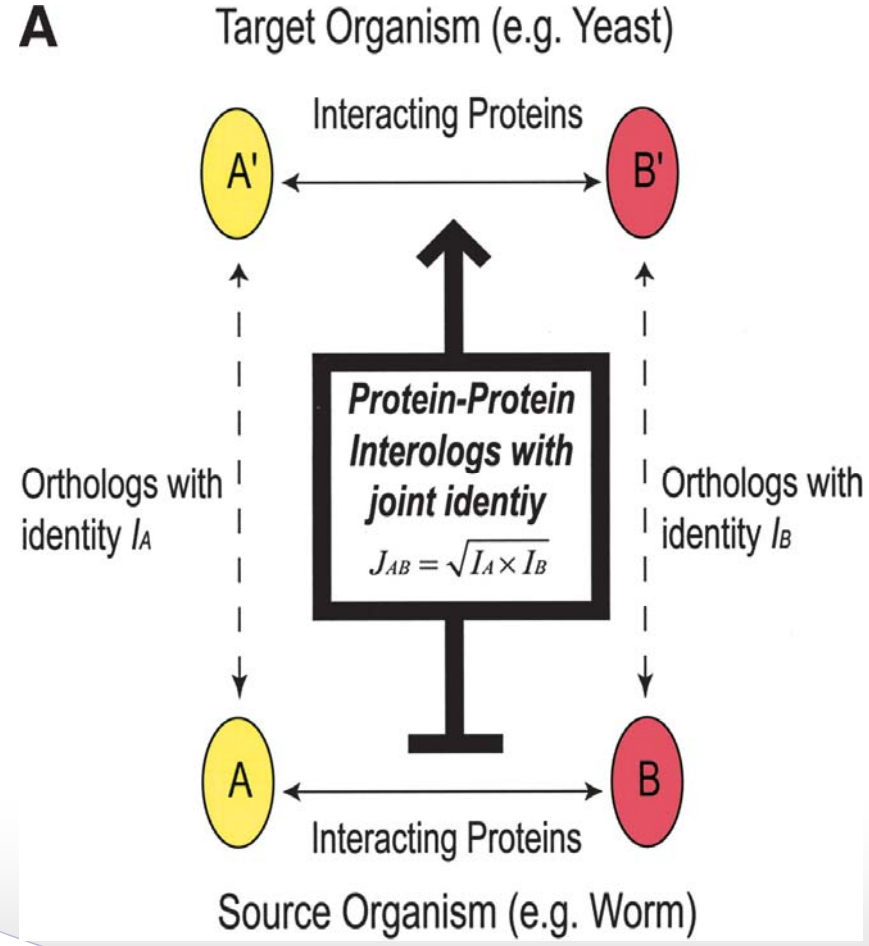
- 准确性比高通量实验高
- 根据三级结构能够预测蛋白质相互作用界面
  - PRKD1 and PRKCE
  - EEF1D and VHL

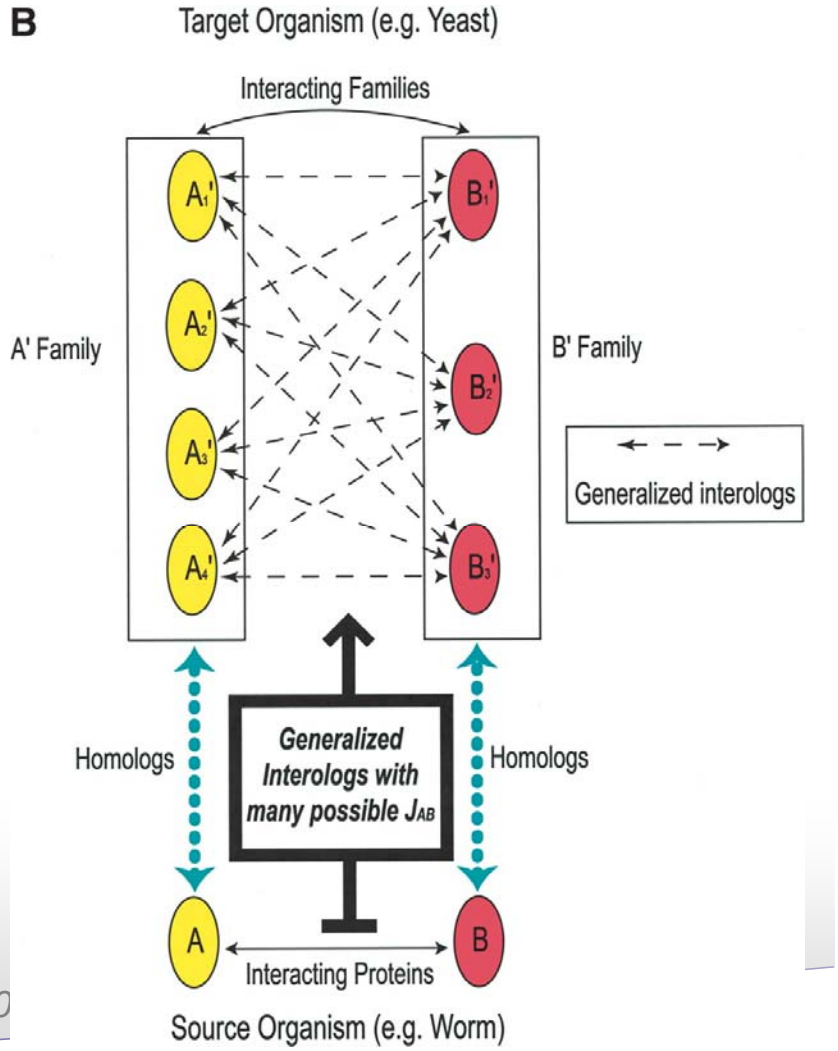




Interolog

- 2004年，于海源、Mark Gerstein

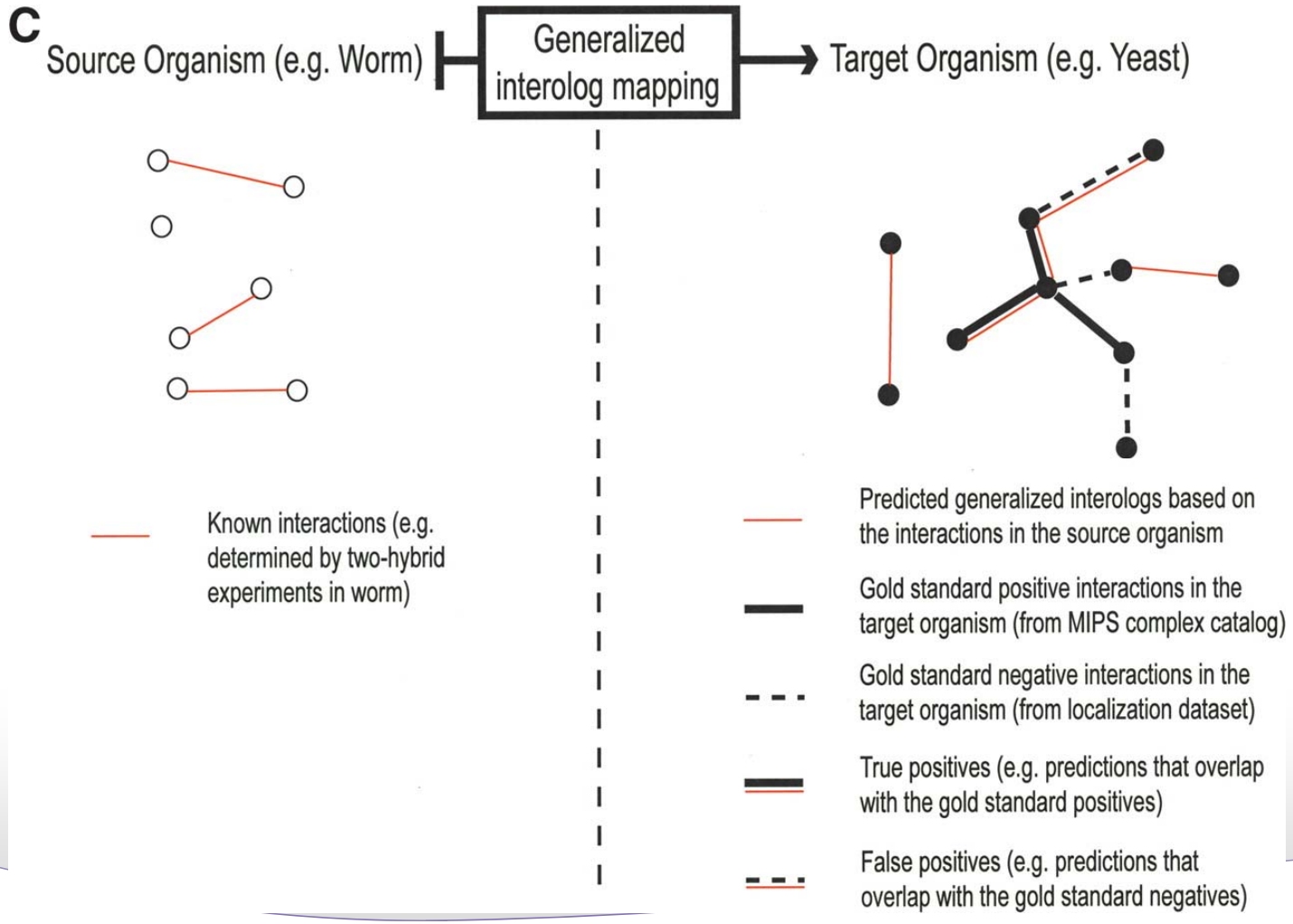




Haiyuan Yu et al. Genome Res. 2004;14:1107-1118

相互作用数据整合

❒ 实验 + Interolog

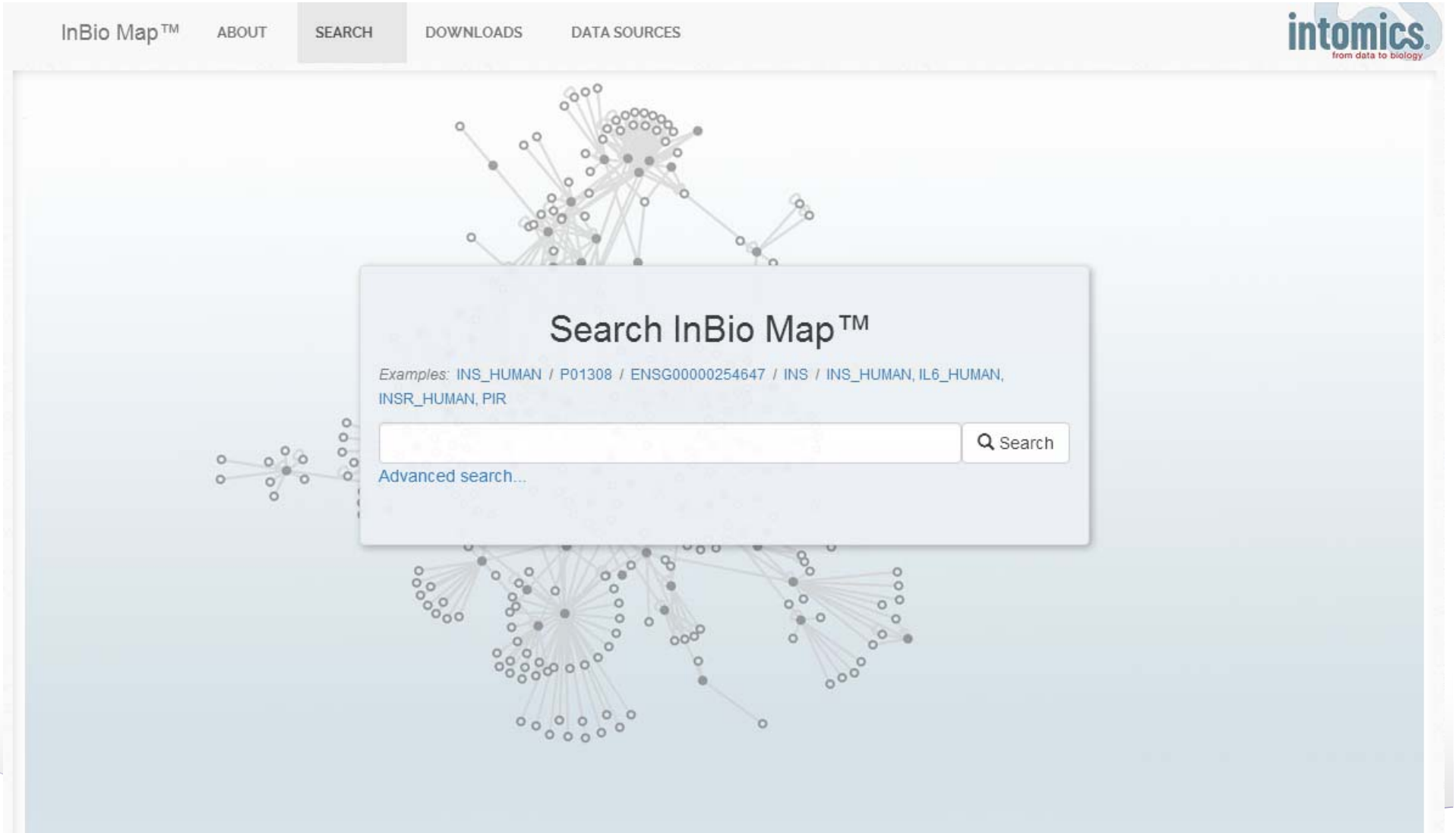


PPI数据整合

- 三种PPI数据：实验验证、Interolog、预测的PPI

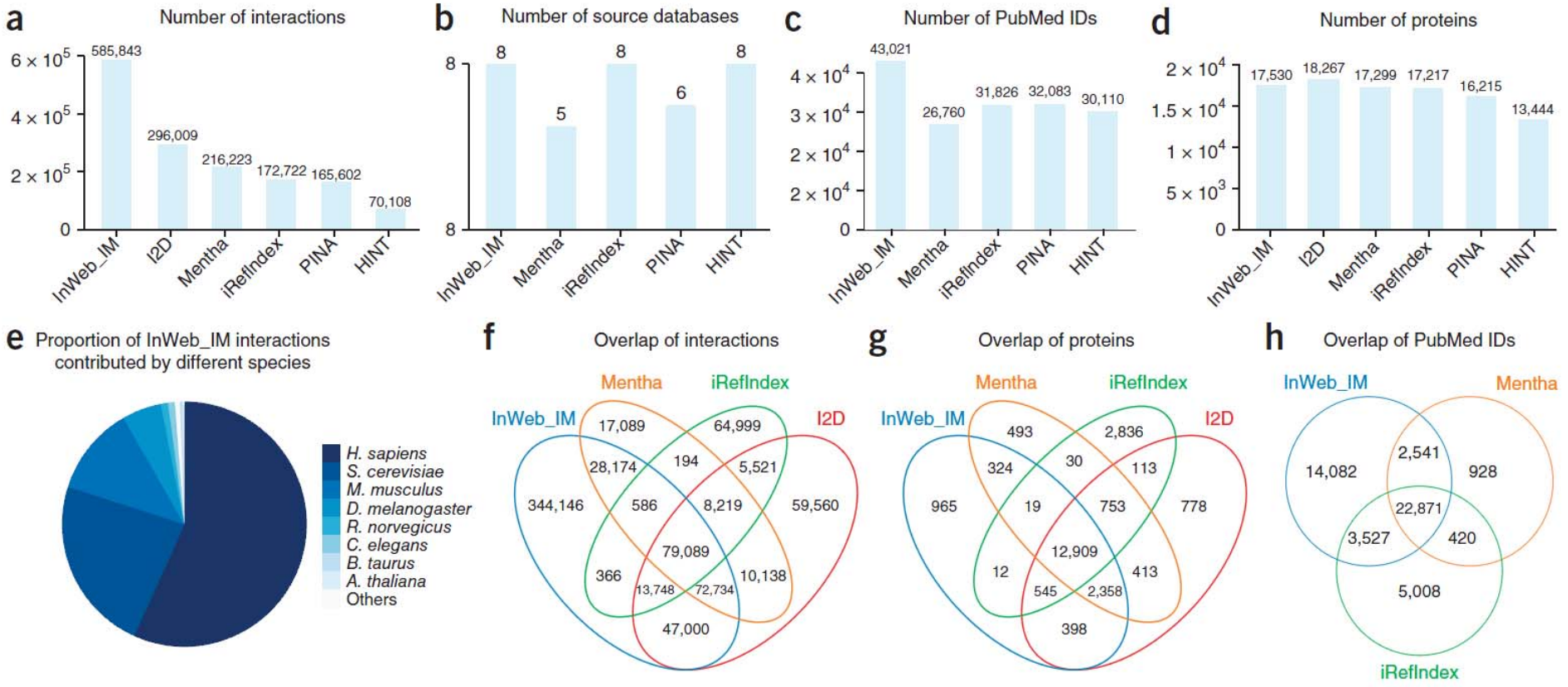
Database	Web Link	Proteins	PPIs[a]	Resource[b]	Last updated
iRefIndex	http://irefindex.org	25,306[c]	199,395	12	2015/4/20
PINA	http://cbg.garvan.unsw.edu.au/pina	17,109	166,776	6	2014/5/21
HINT	http://hint.yulab.org	17,777	277,670	8	N/A[d]
Mentha	http://mentha.uniroma2.it	18,245	259,599	5	2016/12/25
InWeb_IM	http://www.intomics.com/inbio/map; http://www.lagelab.org/resources/	17,653	625,641	8	2016/9/12
IID	http://ophid.utoronto.ca/iid	18,215	911,446	7	2016/3/1

InWeb_IM/InBio Map™



InWeb_IM

❒ 实验 + Interolog



IID

- 实验验证、Interolog、预测的PPI