



生物信息学概论

第五章 基因组分析

人类基因组计划



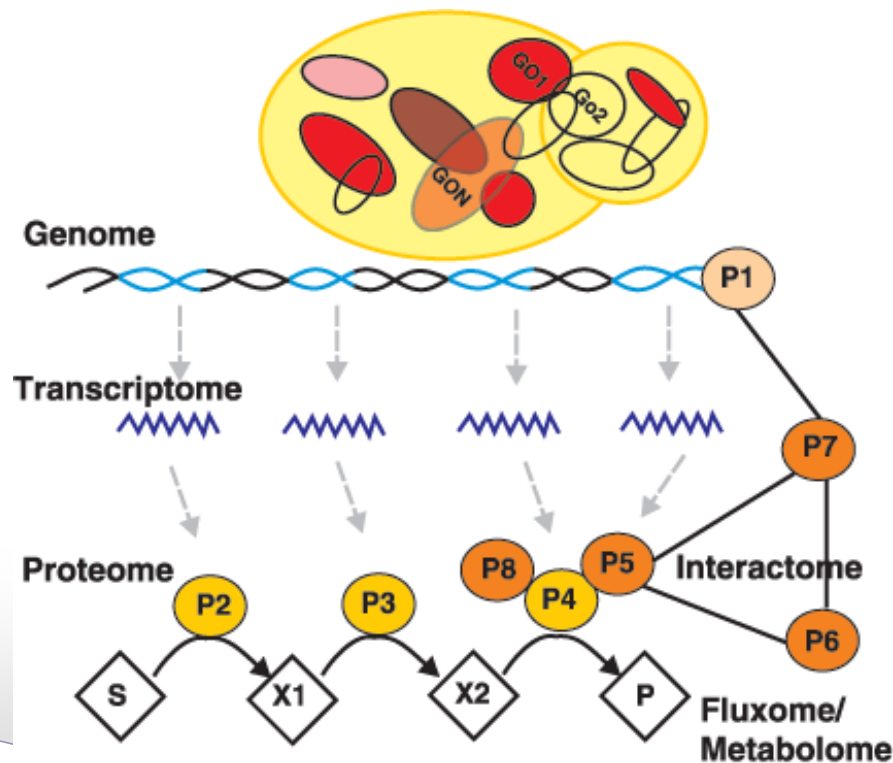
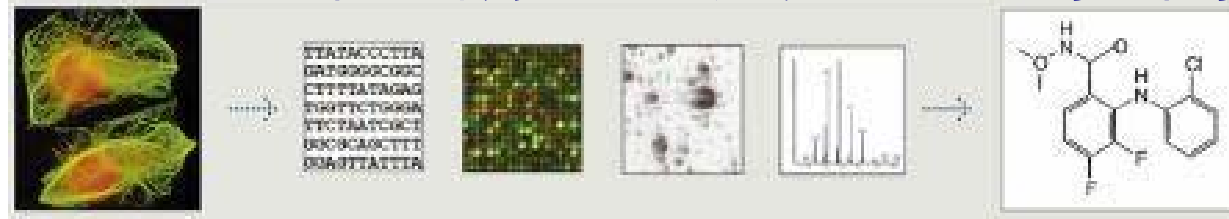
基因组、转录组和蛋白质组



基因组 转录组

蛋白质组

化学生物学



a. GO

Global functional annotation

Interaction

Level of information

b. TF

Transcriptional regulation

c. Protein

Flow of information through PPI

d. Complexes

Agglomerative functional modules

e. Metabolites

Metabolic network regulation



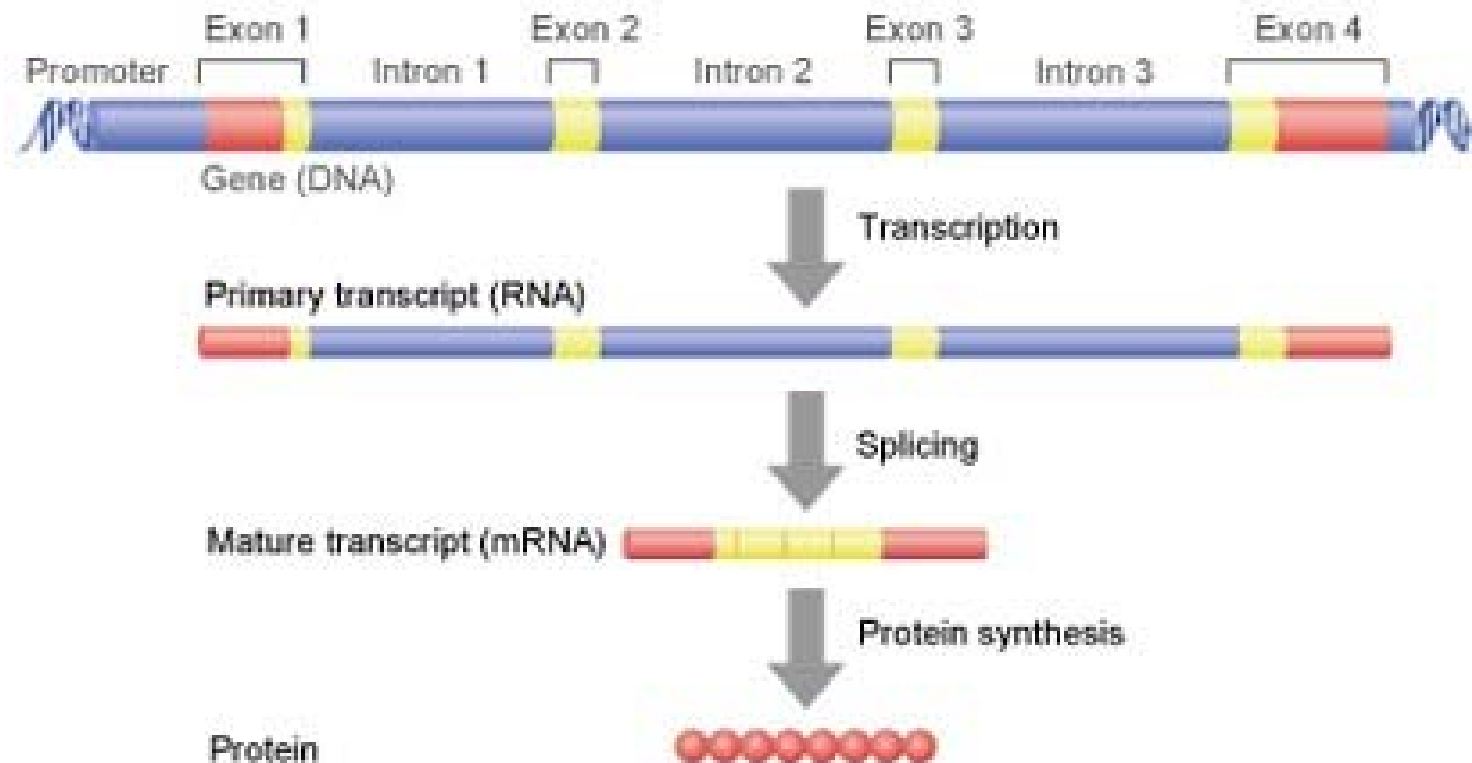
基因组的结构与内容

- ❑ 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



© Wellcome Trust

基因组大小 & 基因数



Sequenced genomes vary from 470-30,000 genes

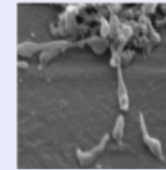
Species	Genome (Mb)	Genes	Lethal loci
<i>Mycoplasma genitalium</i>	0.58	470	~300
<i>Rickettsia prowazekii</i>	1.11	834	
<i>Haemophilus influenzae</i>	1.83	1,743	
<i>Methanococcus jannaschi</i>	1.66	1,738	
<i>B. subtilis</i>	4.2	4,100	
<i>E. coli</i>	4.6	4,288	1,800
<i>S. cerevisiae</i>	13.5	6,034	1,090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

©virtualtext www.ergito.com 2025,

Minimum gene numbers range from 500 to 30,000

500 genes

Extracellular (parasitic) bacterium



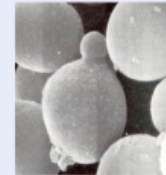
1,500 genes

Free-living bacterium



5,000 genes

Unicellular eukaryote



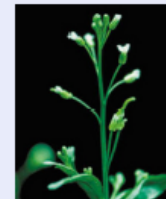
13,000 genes

Multicellular eukaryote



25,000 genes

Higher plants



30,000 genes

Mammals



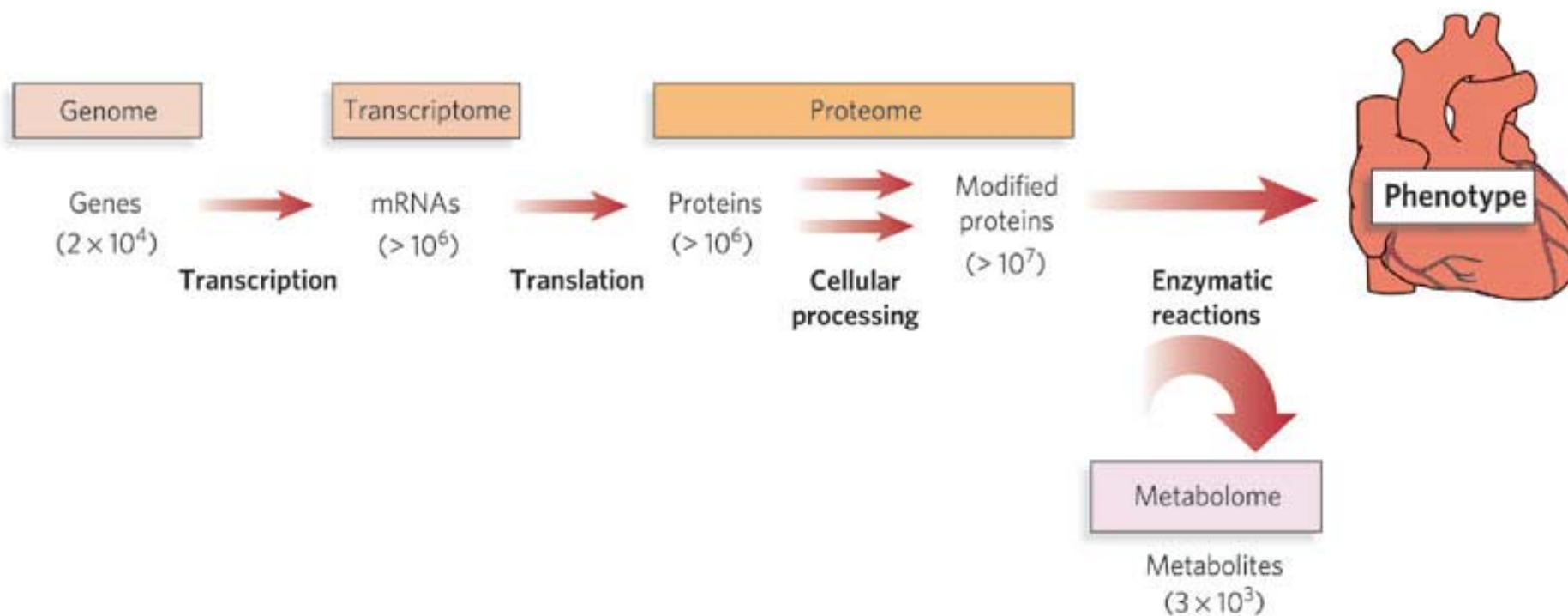
©virtualtext www.ergito.com



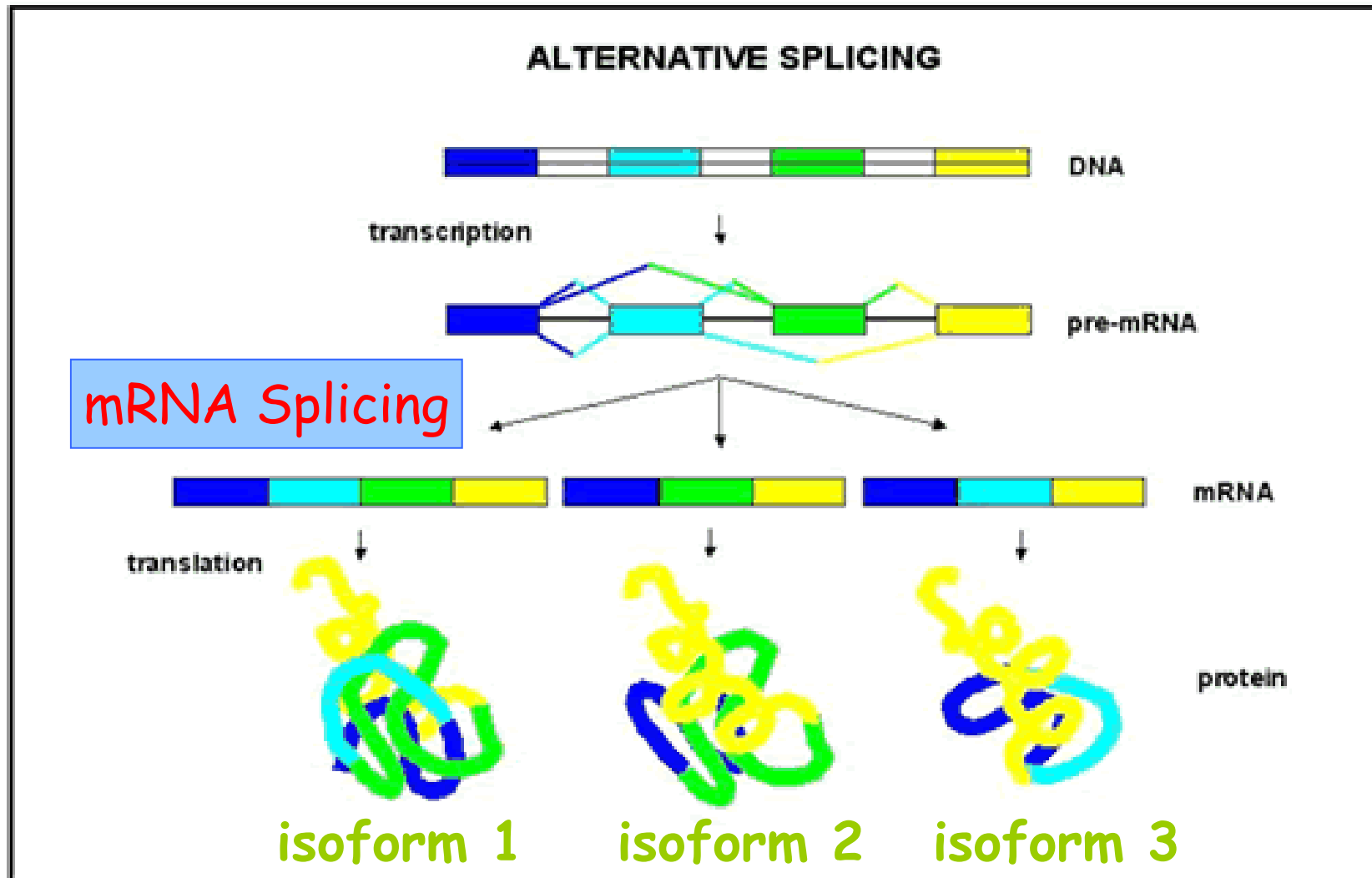
基因数量 -> 生物复杂性?

- ❑ 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剪接，产生拼接异构体
 - ✿ 蛋白质层面，蛋白质序列上一个或多个位点上发生的翻译后修饰

基因型到表型

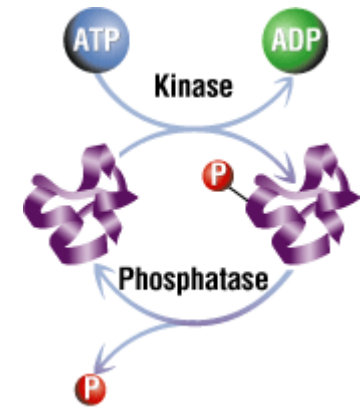


转录后层面：mRNA Splicing

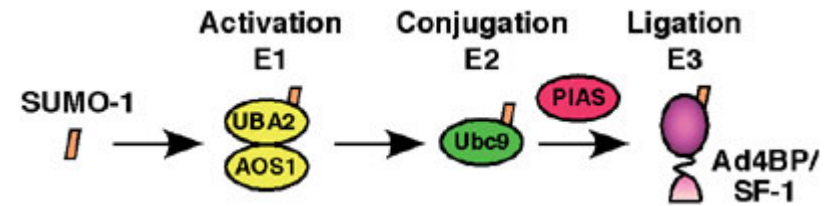




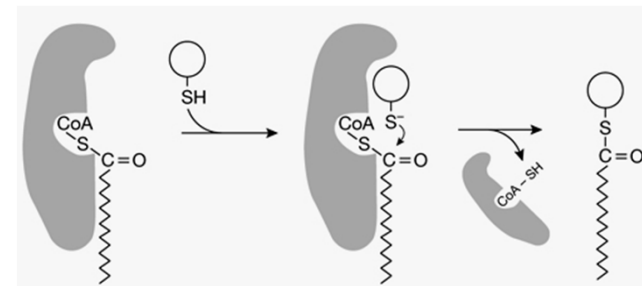
蛋白质层面：翻译后修饰



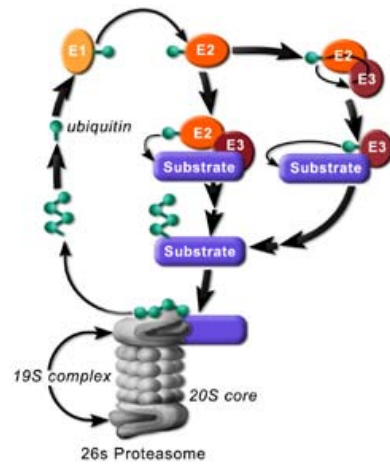
Phosphorylation



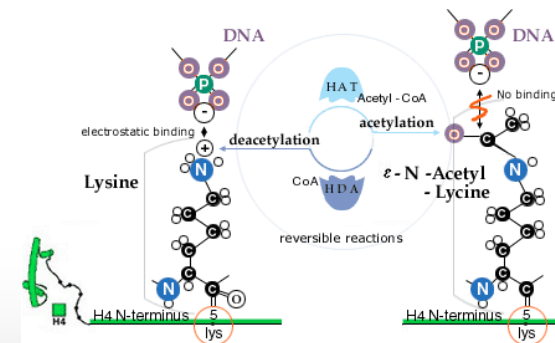
Sumoylation



Palmitoylation



Ubiquitination

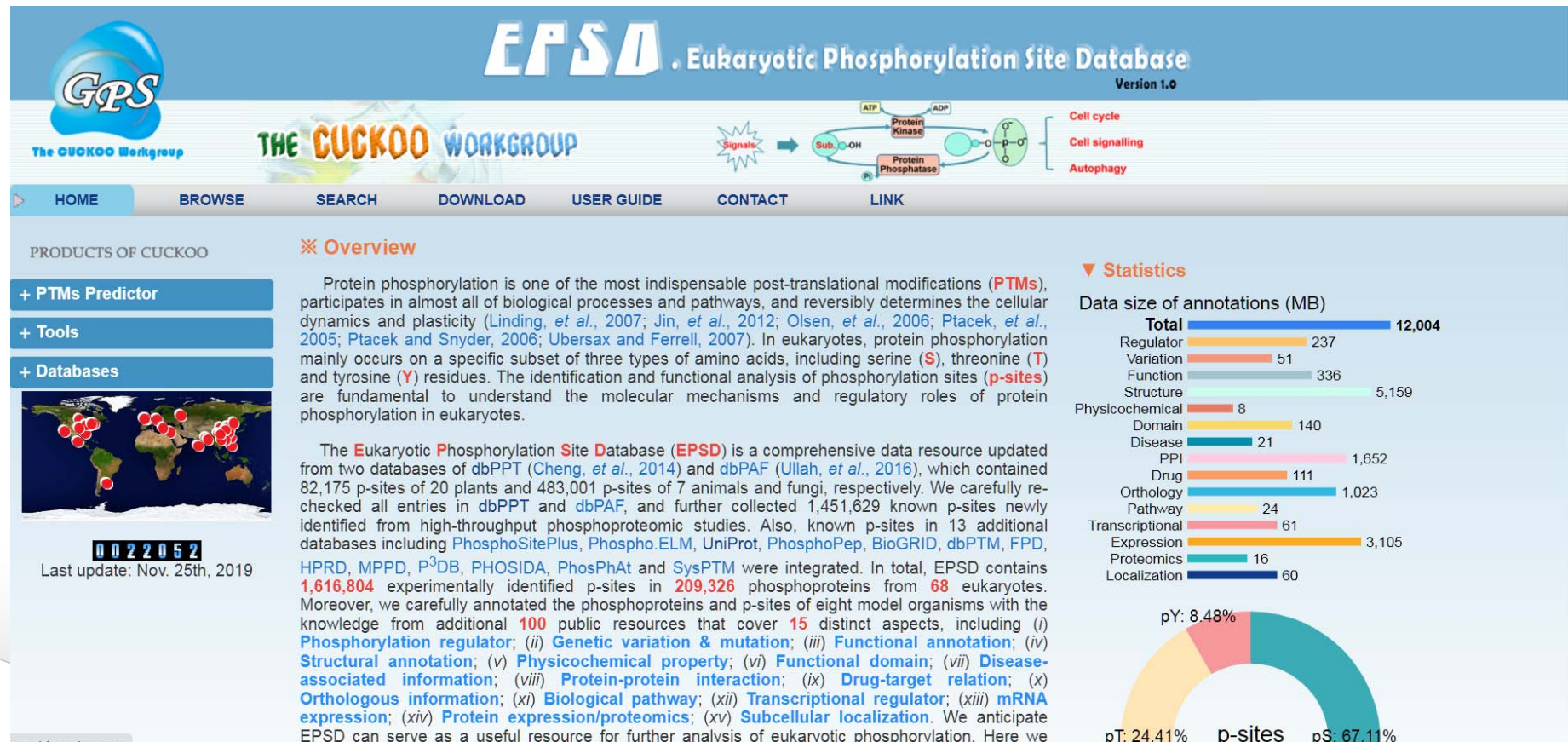


Acetylation

EPSD数据库



- ❑ <http://epsd.biocuckoo.cn/>
- ❑ 1,616,804个实验验证的磷酸化位点, 209,326个磷酸化蛋白质, 68种真核生物



GPS



☐ <http://gps.biocuckoo.cn/>

GPS Group-based Prediction System *Version 6.0*

HOME WEB SERVER CITATION USER GUIDE LINKS ARCHIVE CONTACT

PRODUCTS of euekoo

- PTMs Predictor
- Tools
- Databases

0880411
Last update: Jan. 1st, 2023

※ **GPS INTRODUCTION:**

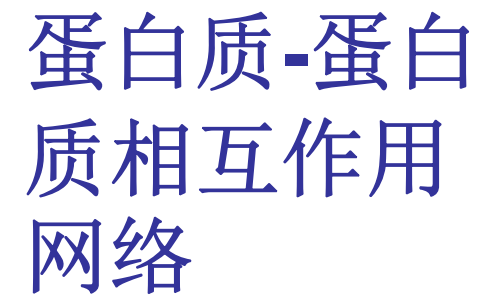
Protein phosphorylation, catalyzed by protein kinases (PKs), is one of the most important post-translational modifications (PTMs), and involved in regulating almost all of biological processes.

Here, we report an updated server, **Group-based Prediction System (GPS) 6.0**, for prediction of PK-specific phosphorylation sites (p-sites) in eukaryotes. First, we pre-trained a general model using **490,762** non-redundant p-sites in **71,407** proteins. Then, transfer learning was conducted to obtain **651** PK-specific predictors at the group, family and single PK levels, using a well-curated data set of **30,312** known site-specific kinase-substrate relations (ssKSRs) in **7059** proteins. Ten types of sequence features were extracted and integrated by 3 types of machine learning algorithms, including penalized logistic regression (PLR), deep neural network (DNN), and Light Gradient Boosting Machine (LightGBM). Using a newly collected data set of **1426** ssKSRs in 651 proteins, we compared other existing tools to GPS 6.0, which exhibited a much higher accuracy on a number of well-studied PKs. Together with the evolutionary information, GPS 6.0 could hierarchically predict PK-specific p-sites for **46,402** PKs in **184** species. For users, one or multiple protein sequences could be inputted in the FASTA format, and the output will be shown in a tabular list. Besides the basic statistics, we also integrated the knowledge from **21** public resources to annotate the prediction results, including the experimental evidence, physical interactions, sequence logos, and p-sites in sequences and 3D structures.

For the help of GPS 6.0 and the tutorial, please refer to the **USER GUIDE** page.
For the source code of GPS 6.0, please visit the **GitHub** page.

▼ **Example**

PDB: 2H6D [A] All Sites Refresh



Cytoscape: 网络构建和分析工具



Cytoscape

Search...

Home Introduction Download Apps Documentation Community Report a Bug Getting Help

Control Panel

Network

Nodes

Edges

Network: BIOGRID-ORGANISM-Drosophila

Nodes: 45878(0)

Edges: 45878(0)

Session: New Session

Enter search term...

BIOGRID-ORGANISM-Drosophila_melanogaster-3.2.95.mitab

Network Data Integration, Analysis, and Visualization in a Box

Cytoscape is an [open source](#) software platform for visualizing complex networks and integrating these with any type of attribute data. A lot of [Apps](#) are available for various kinds of problem domains, including bioinformatics, social network analysis, and semantic web.

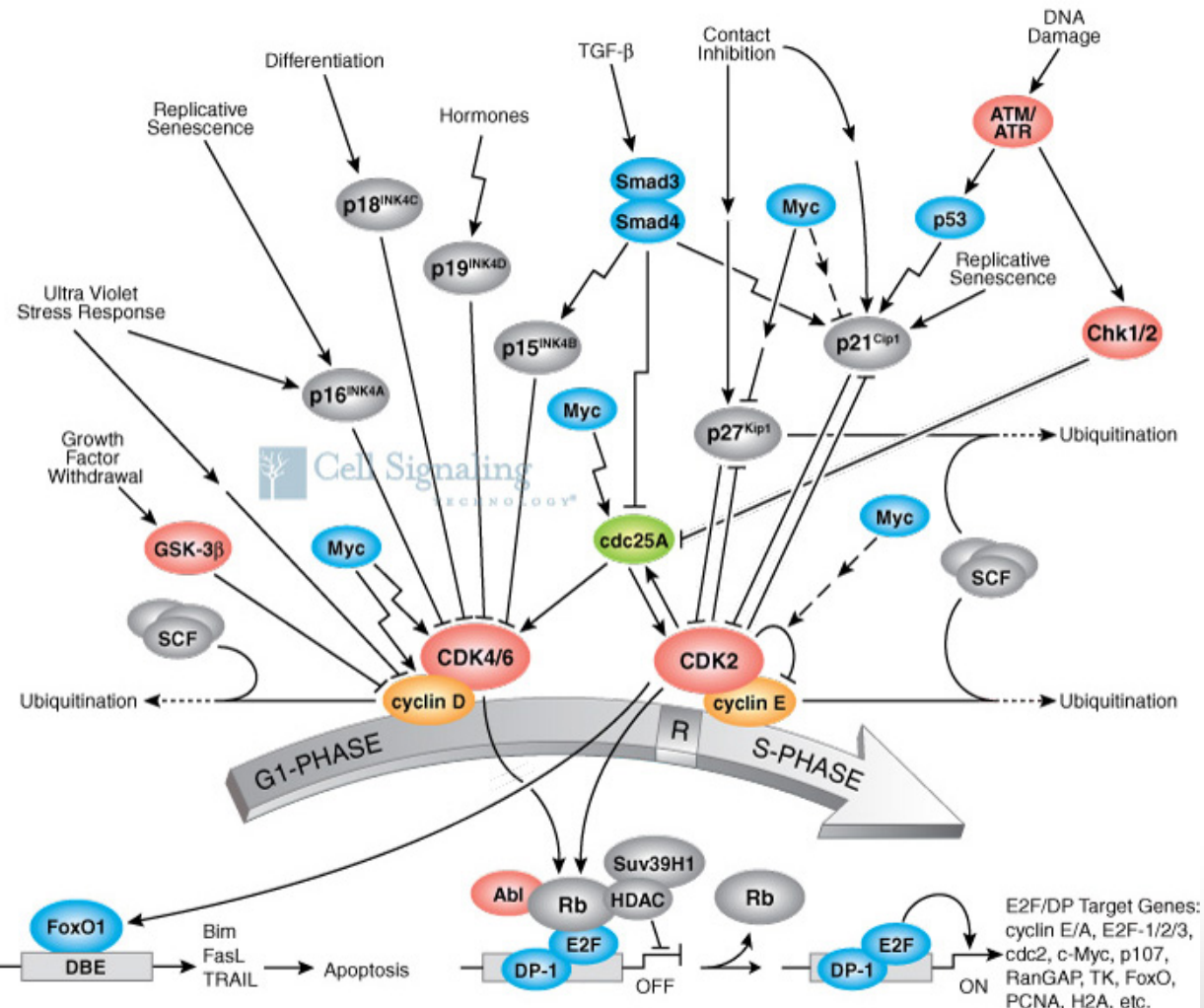
Download Cytoscape

Welcome Letter

Release Notes

Sample Visualizations

细胞信号通路



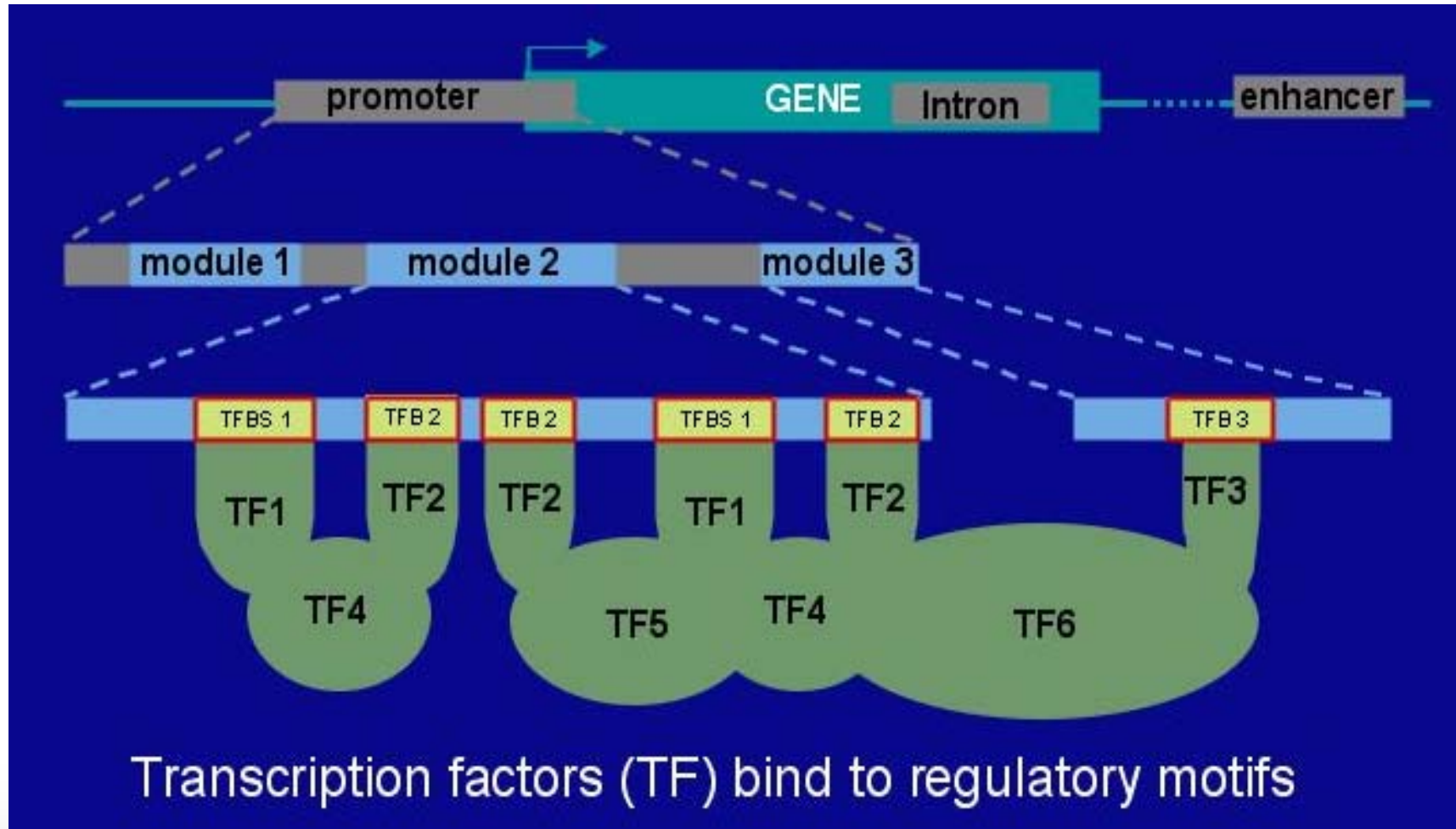
G1/S检验点:
有调控方向

非编码区

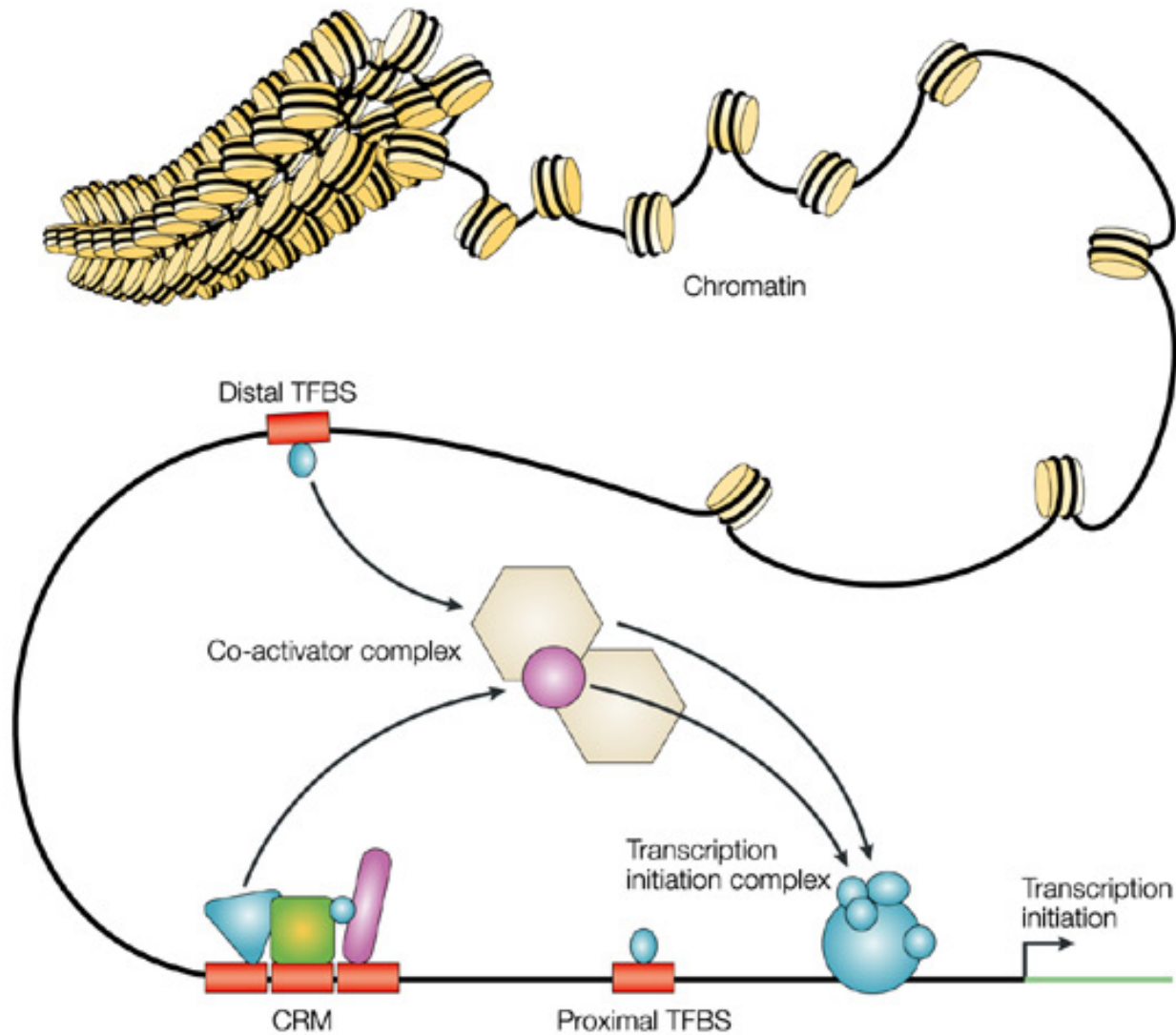


- ❑ 功能元件: 转录因子结合位点; 启动子...
- ❑ Non-coding RNA: MicroRNA
- ❑ 转座子
- ❑ 重复片段
- ❑ 伪基因 (Pseudogene)

功能元件：启动子



转录因子结合位点

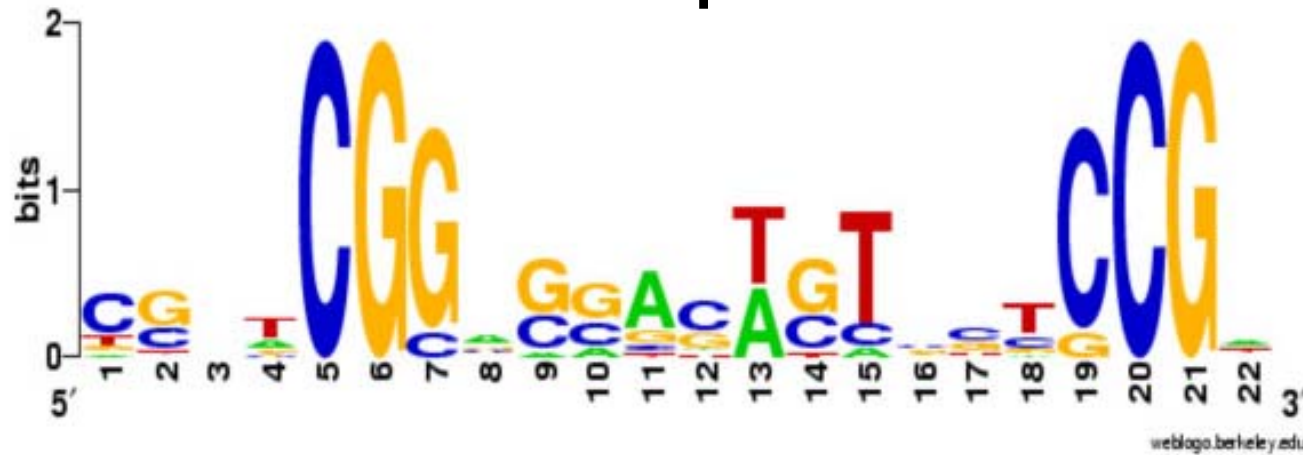


CRM: cis-regulatory modules

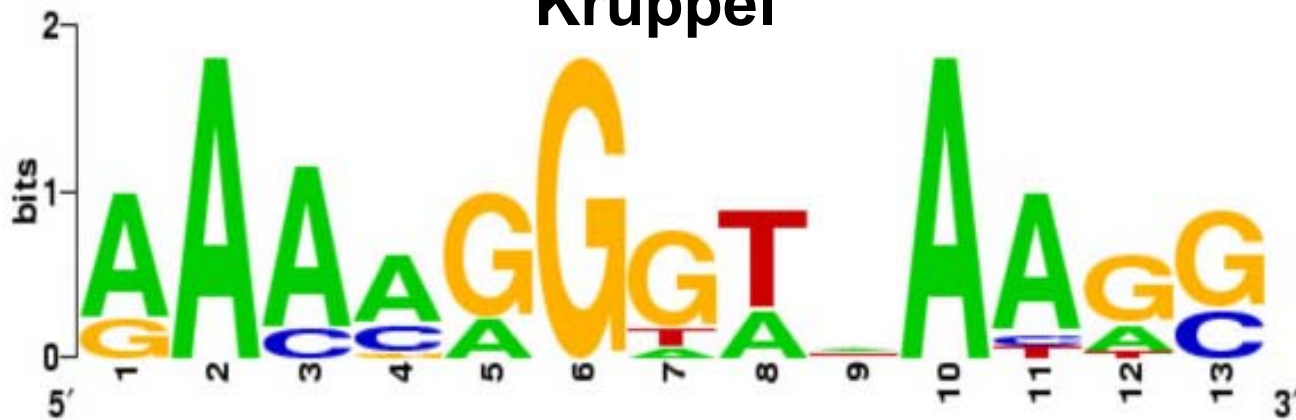
Gal4p and Kruppel



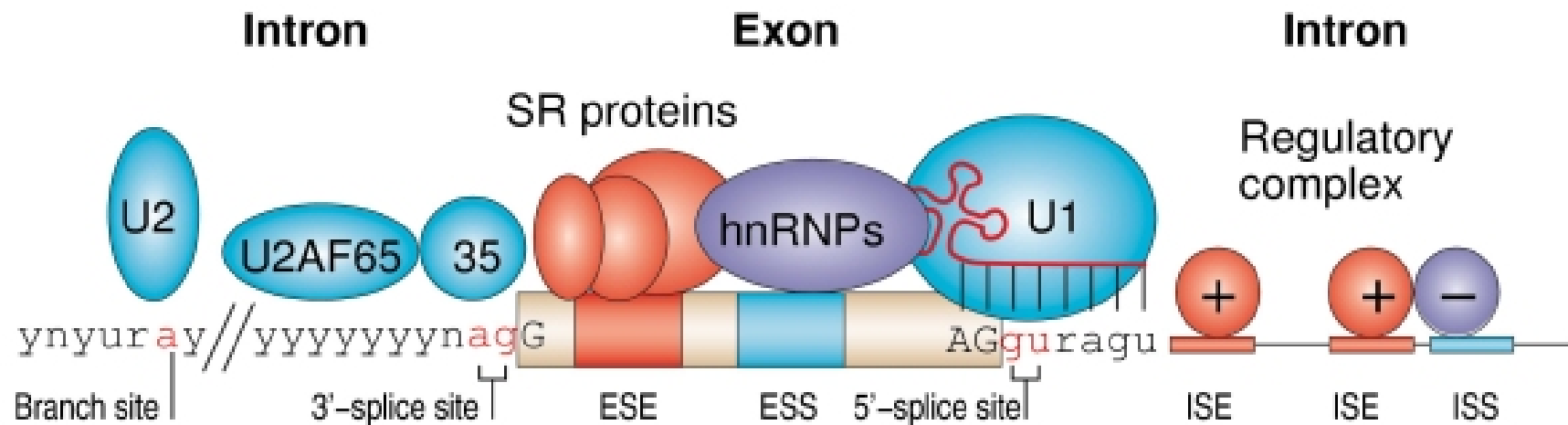
Gal4p



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

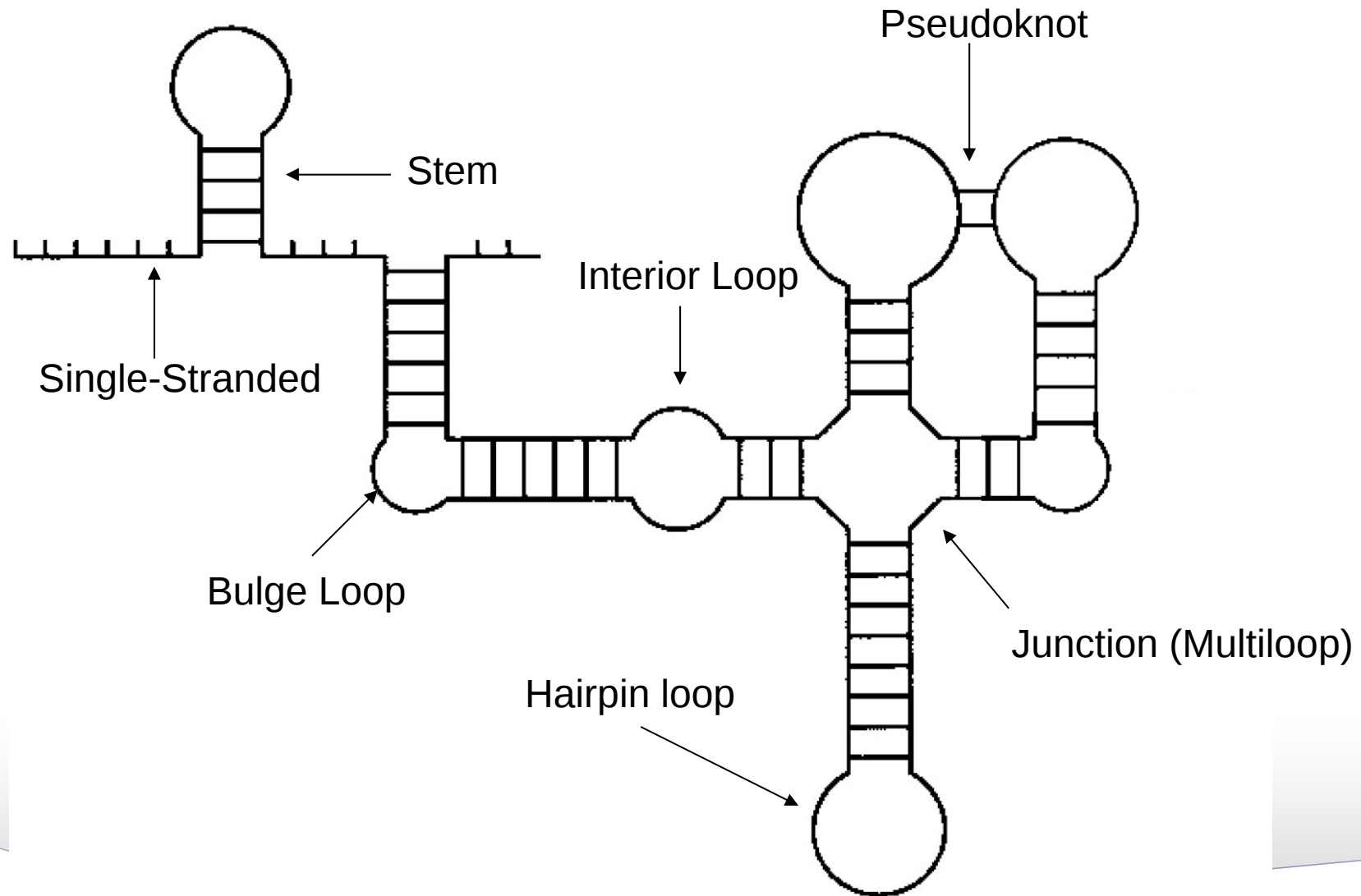
- ✿ siRNAs

- ✿ piRNAs: 与piwi相互作用的RNA

- ✿ long ncRNAs: Xist

- ✿ circRNAs

RNA二级结构



RNA二级结构的显示方式

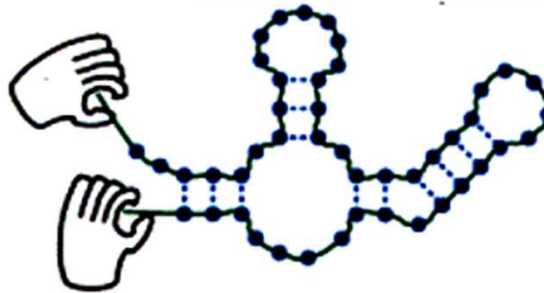


- Grammatically correct **string of parentheses**

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

AGCTACGGAGCGATCTCCGAGCTTTTCGAGAAAGCCTCTATTAGC

- **Planar graph**



- **Arch diagram**



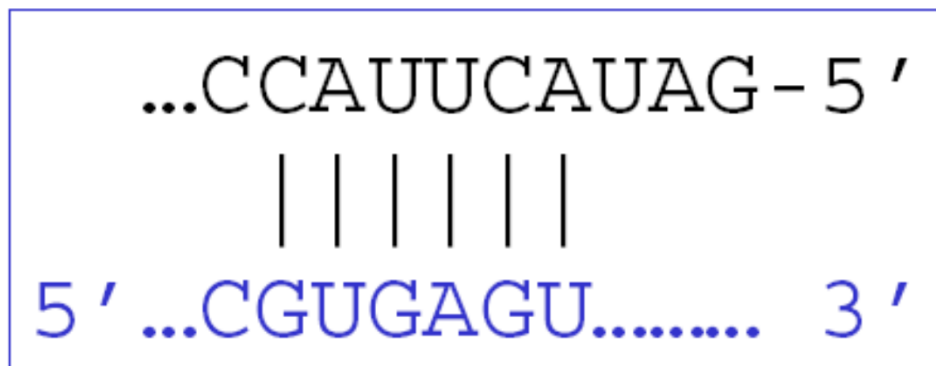
- **Mountain diagram**



RNA的能量

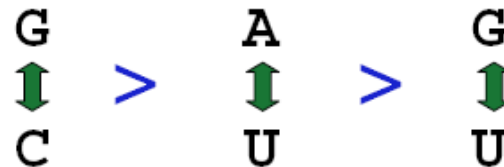


□ Doug Turner的能量法则



5' --> 3' UX AY 3' <-- 5'				
<u>X</u>				
<u>Y</u>	A	C	G	U
A	.	.	.	-1.30
C	.	.	-2.40	.
G	.	-2.10	.	-1.00
T	-0.90	.	-1.30	.

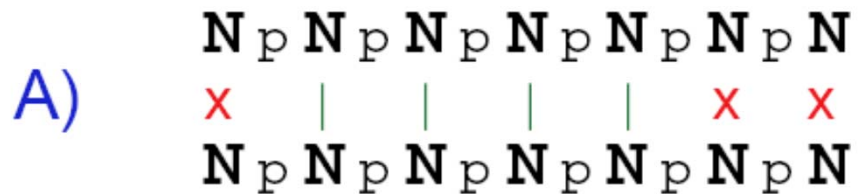
□ Base pairing



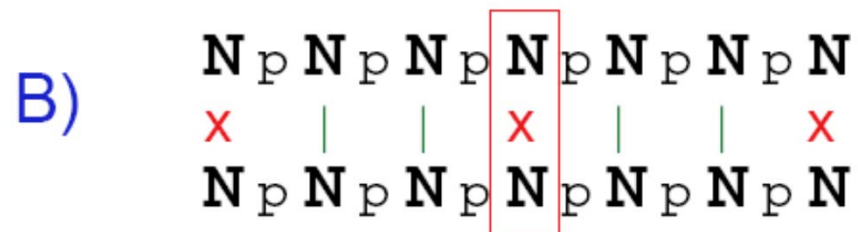
□ Base stacking



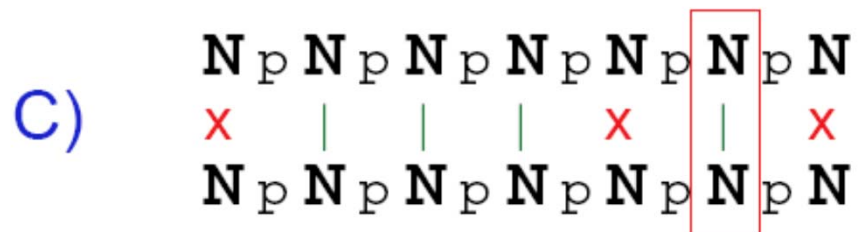
RNA的能量



许多连续的配对



存在内环 (loop)



终端不稳定

□ 稳定性A > B和C

动态规划算法：最大碱基配对

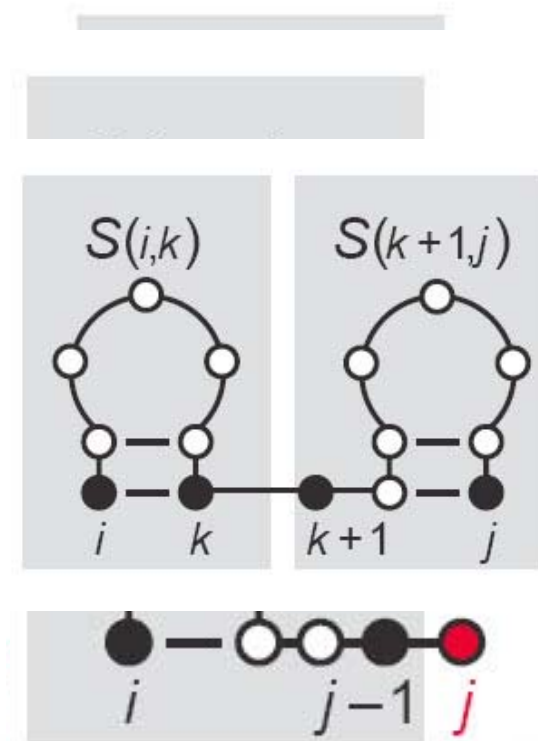


当前位置的匹配是否能
保证最大的碱基匹配

大化

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

序列匹配



动态规划算法：最大碱基配对



□ 比对方法

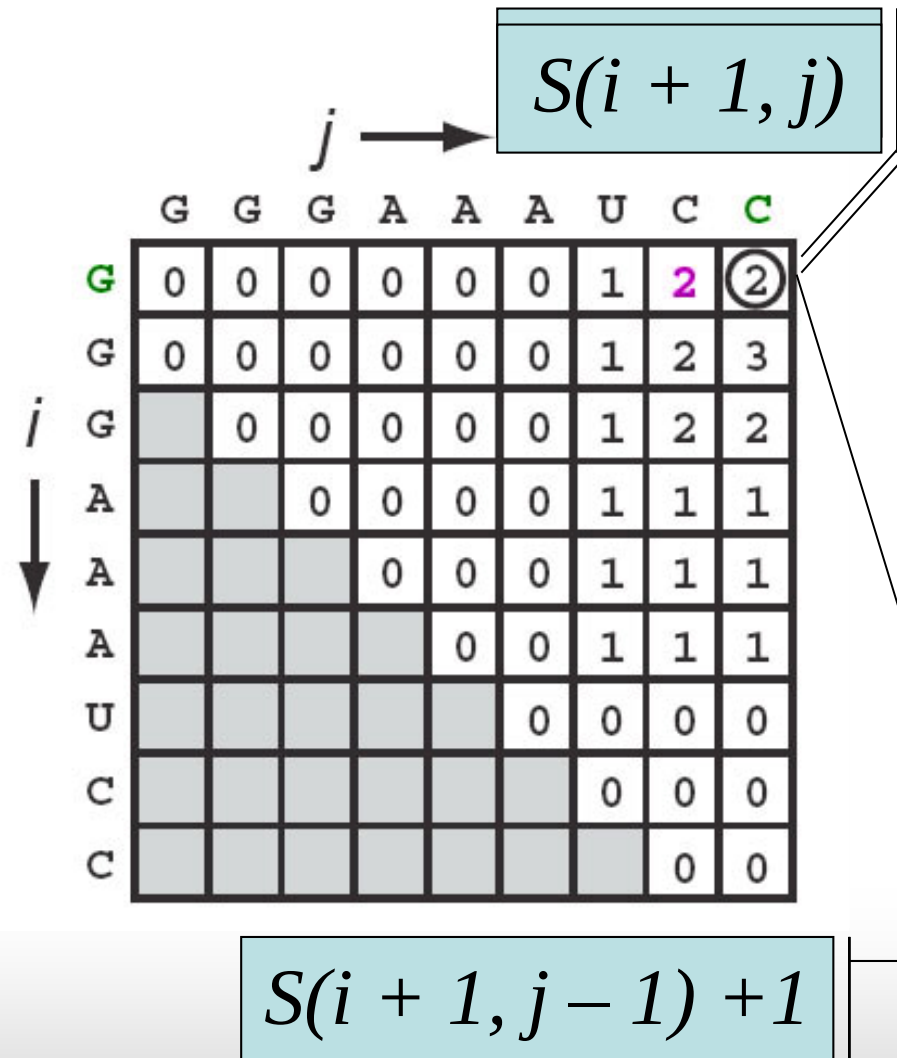
✿ RNA链与自身相比

✿ 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

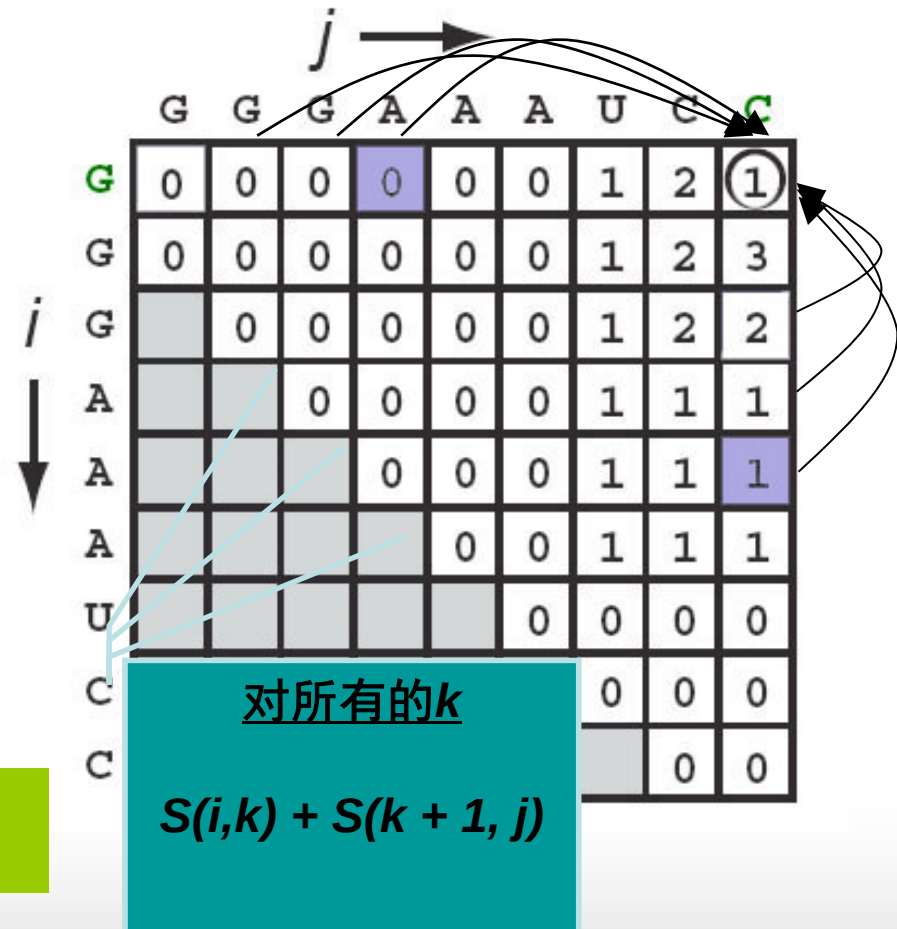
✿ RNA链与自身相比

✿ 若存在碱基配对则增加分数

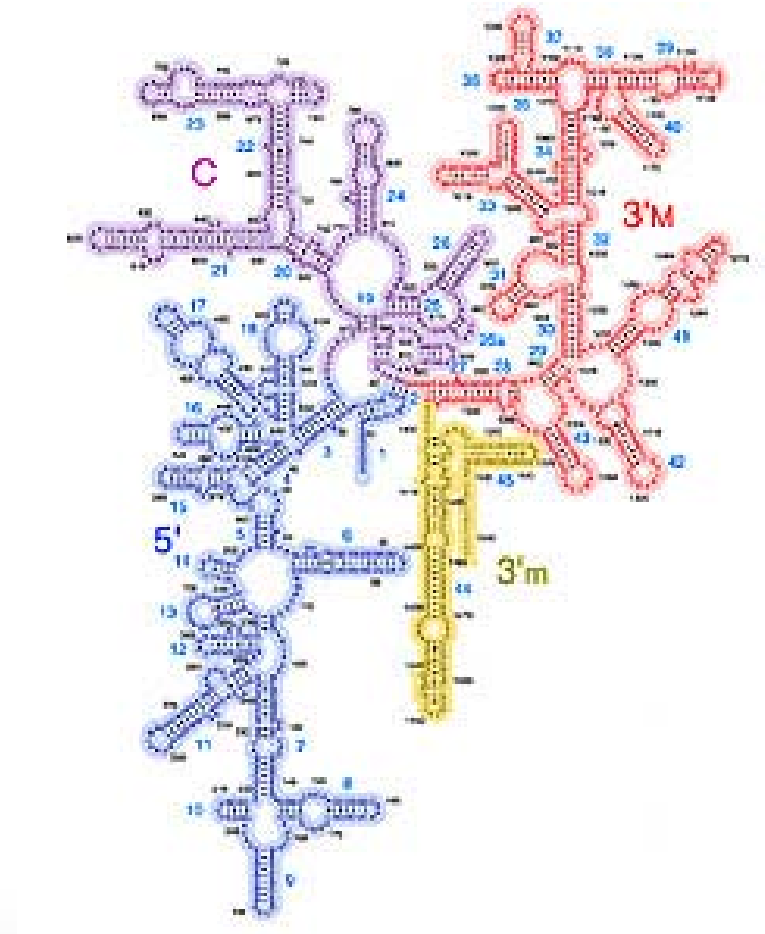
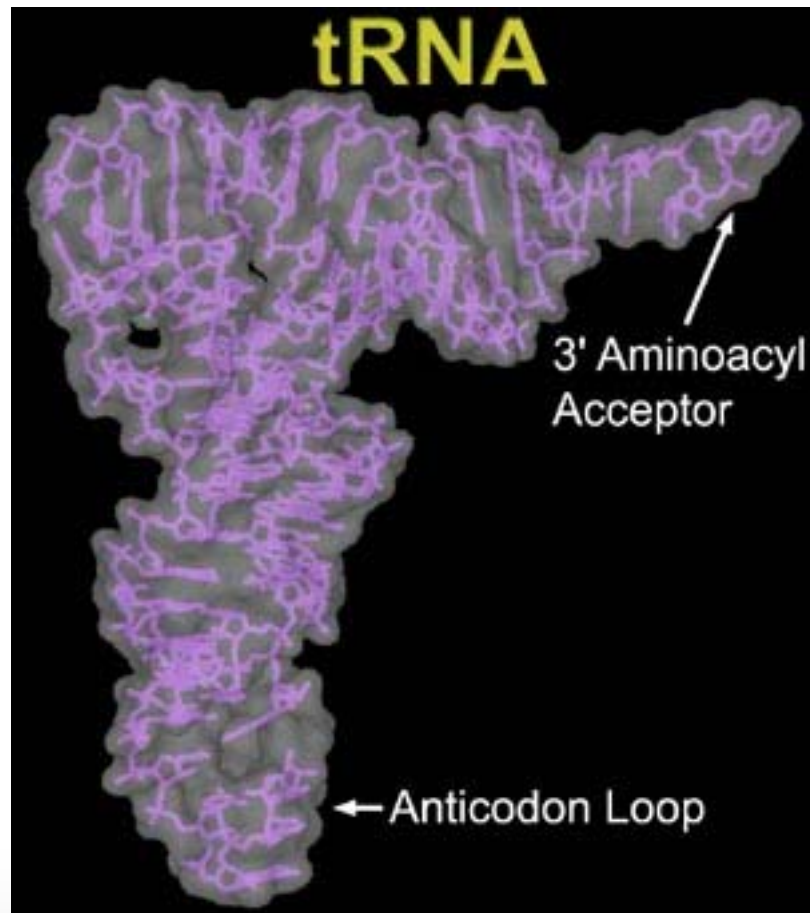
□ 计算分值

□ 可最后加入分歧的影响

分歧：考虑所有的 k 值



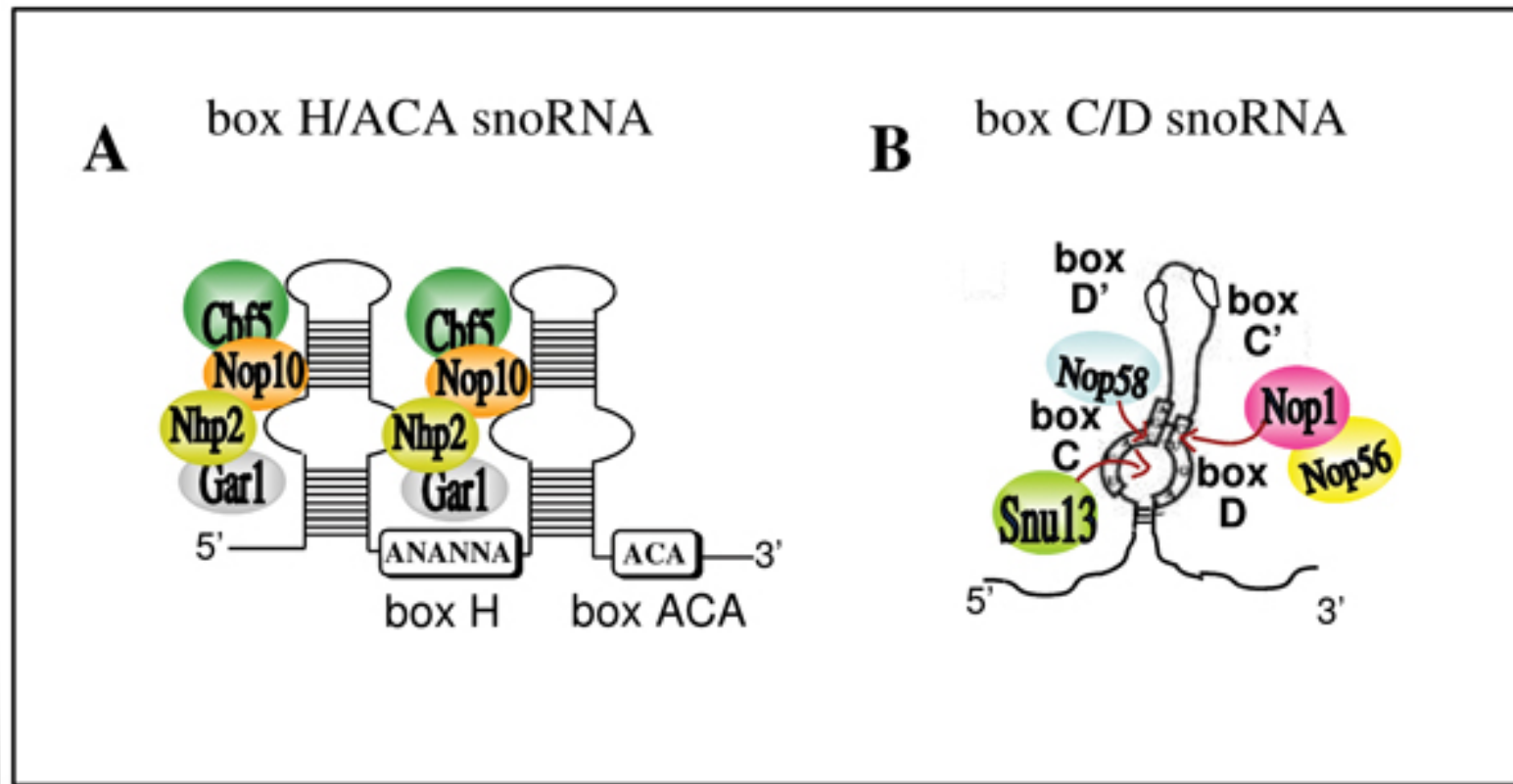
tRNA & rRNA



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

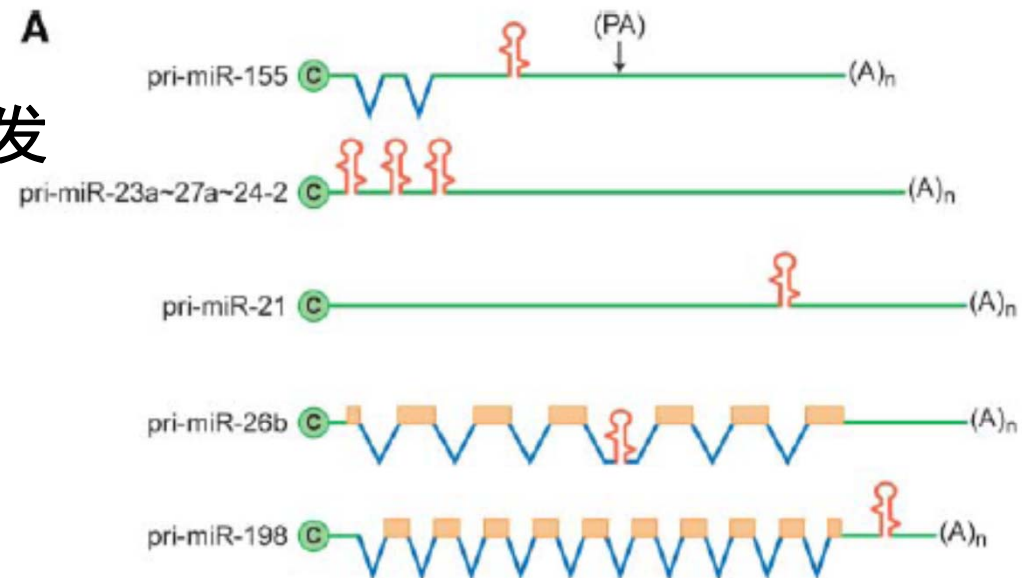
pri-miRNAs的结构



- pri-miRNAs的结构
- 人类pri-miR-30a RNA 的发育环

✿ Drosha切割位点 (箭头)

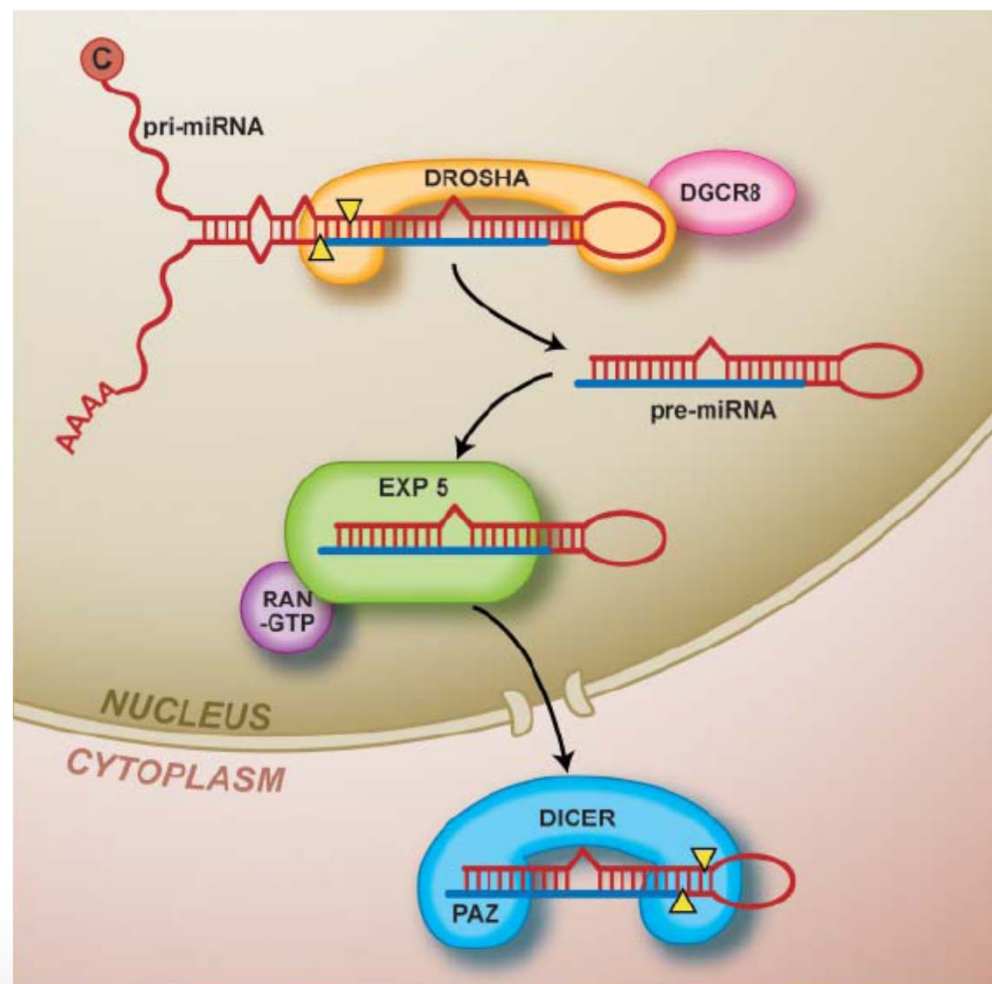
✿ Dicer切割位点 (三角)



miRNA的加工



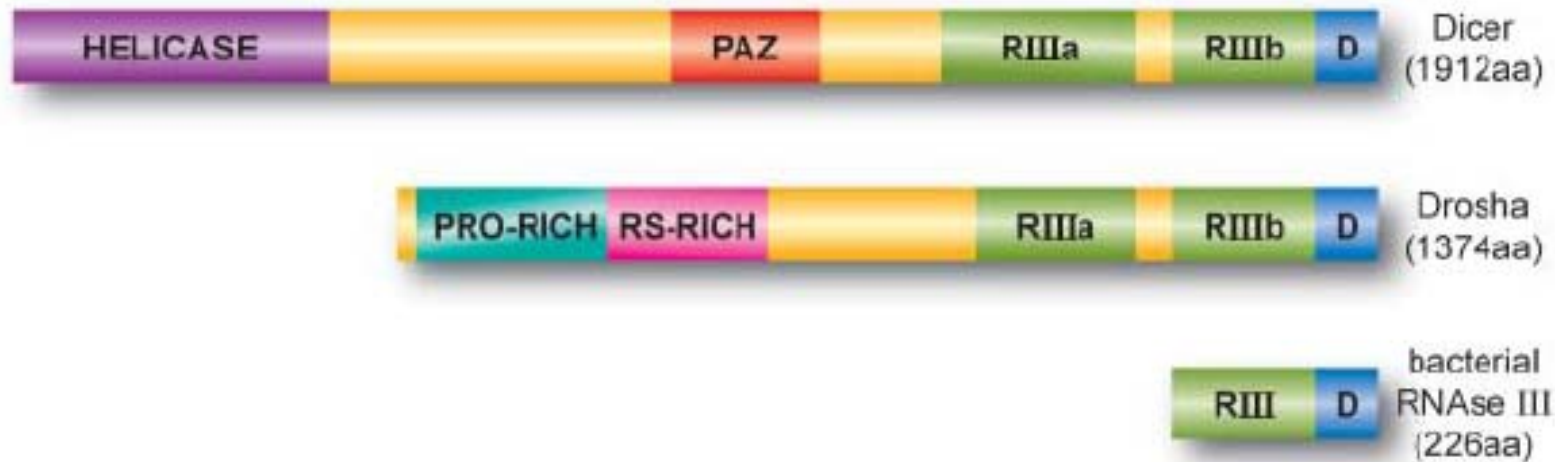
- ❑ Pri-miRNA被Drosha切割
- ❑ pre-miRNA被Dicer切割
- ❑ Exportin 5 (Exp5) 将 miRNA装运到胞质中



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

The screenshot displays the miRBase website. At the top, there's a navigation bar with links: Home, Search, Browse, Help, Download, Blog, and Submit. A search bar is located on the right side of this bar. Below the navigation bar, the main content area is divided into sections. On the left, there's a section titled "Latest miRBase blog posts" with two entries: "MicroRNA Gene Ontology annotations" by sam (June 7, 2018) and "miRBase 22 release" by sam (March 12, 2018). On the right, there's a section titled "miRNA count: 38589 entries" with a sub-section "Release 22.1: October 2018". Below this, there's a "Search by miRNA name or keyword" section with a search bar and buttons for "Go" and "Example". At the bottom right, there's a section titled "Download published miRNA data" with links for "Download page" and "FTP site".

www.mirbase.org

miRBase

MANCHESTER 1824

Home Search Browse Help Download Blog Submit

Latest miRBase blog posts

[MicroRNA Gene Ontology annotations](#) By [sam](#) (June 7, 2018)

You might have noticed some additional information on the mature miRNA pages in the last few weeks. See for example: http://mirbase.org/cgi-bin/mature.pl?mature_acc=MIMAT0000123 http://mirbase.org/cgi-bin/mature.pl?mature_acc=MIMAT0000069 The new section "QuickGO function" contains a set of high quality manual annotations of Gene Ontology terms for mature miRNAs, the vast majority of which come from the work of Rachael Huntley et [...]

[miRBase 22 release](#) By [sam](#) (March 12, 2018)

After repeated and unreasonable delay, miRBase 22 is finally released. As you might expect with such a long gap, the number of sequences in the database has jumped significantly — by over a third. The vast majority of the increase comes from new microRNA annotations in species not previous represented in the database. Indeed, there [...]

miRNA count: 38589 entries

[Release 22.1](#): October 2018

Search by miRNA name or keyword

Go Example

Download published miRNA data

[Download page](#) | [FTP site](#)

miRNAs的多样性

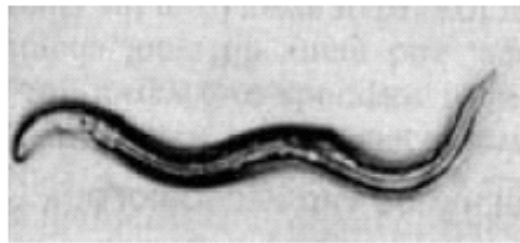


MicroRNAs (miRNAs)



A. thaliana/
O. sativa

拟南芥: **187**



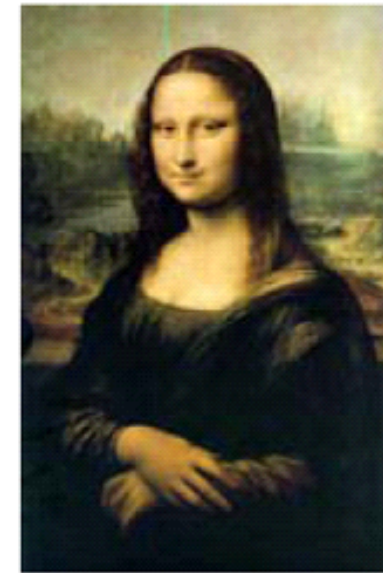
C. elegans

154



D. melanogaster

152



H. sapiens/
M. musculus

人: **695**

The miRBase Sequence Database -- Release 12.0

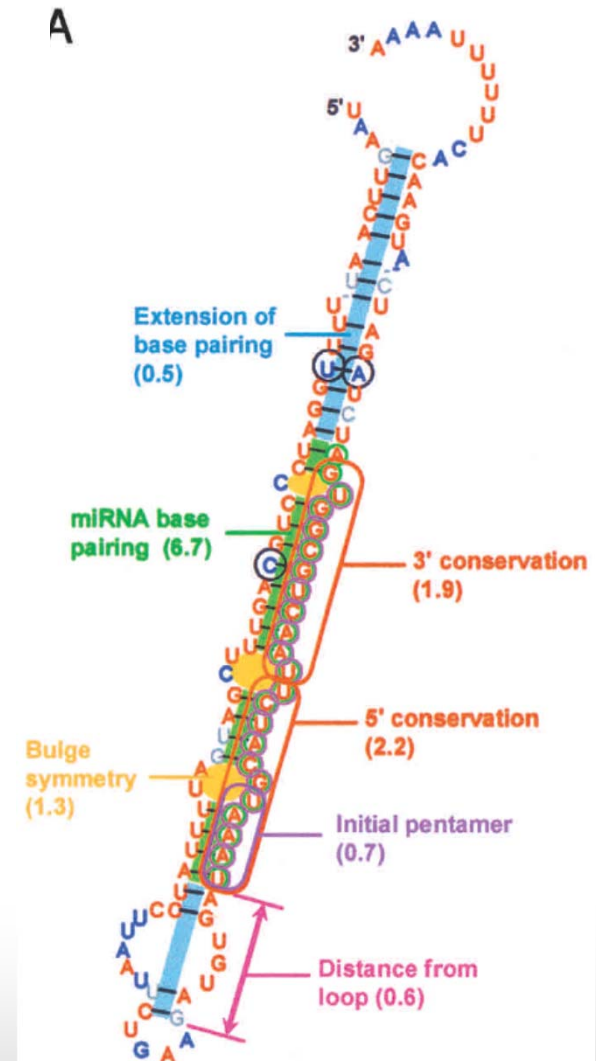
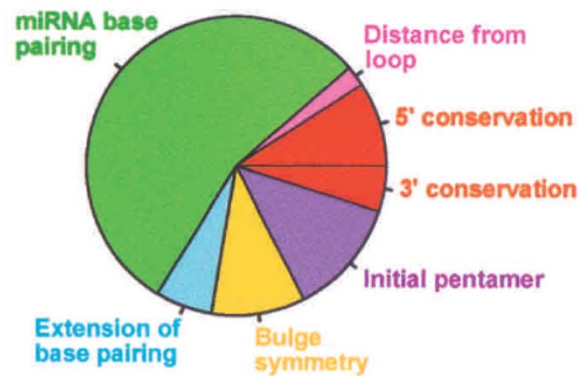
Bioinformatics, 2025, HUST

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**

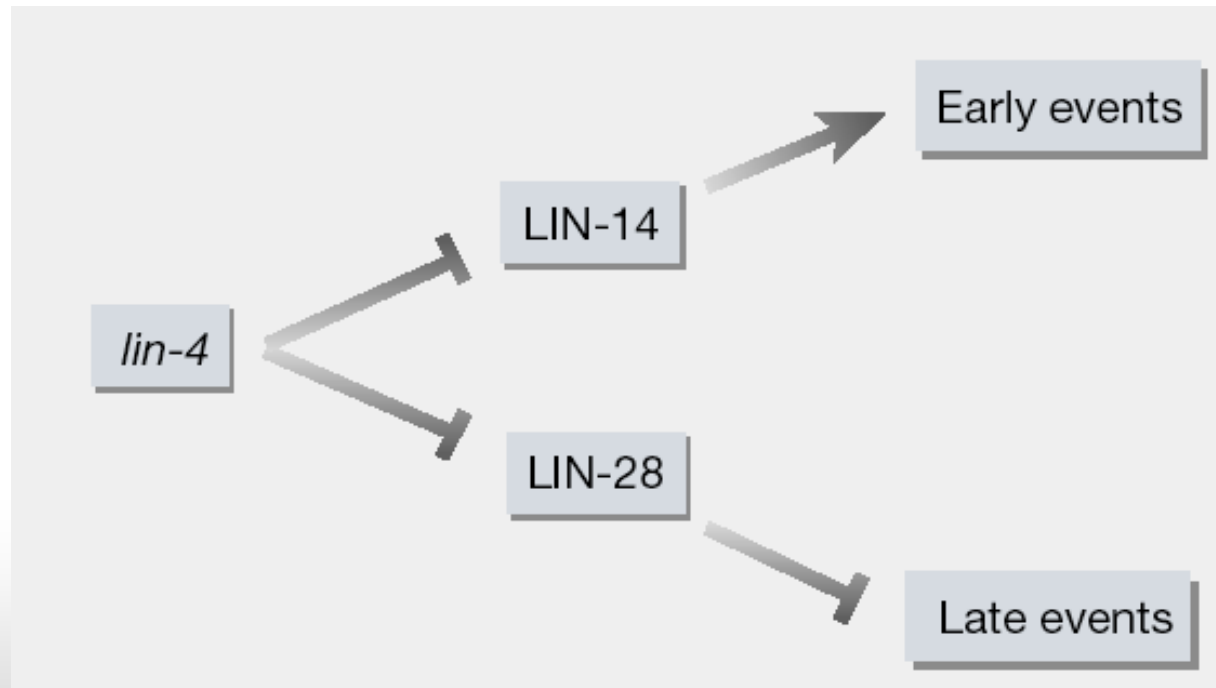
miRNA	Animal	Function	Targets
<i>lin-4</i>	Ce	developmental timing ¹⁹	<i>lin-14</i> (refs 19, 22) <i>lin-28</i> (ref. 24)
<i>let-7</i>	Ce	developmental timing ²⁰	<i>lin-41</i> (ref. 23) <i>hbl-1</i> (refs 48, 49)
<i>lisy-6</i>	Ce	neuronal cell fate ²⁹	<i>cog-1</i> (ref. 29)
<i>mir-273</i>	Ce	neuronal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

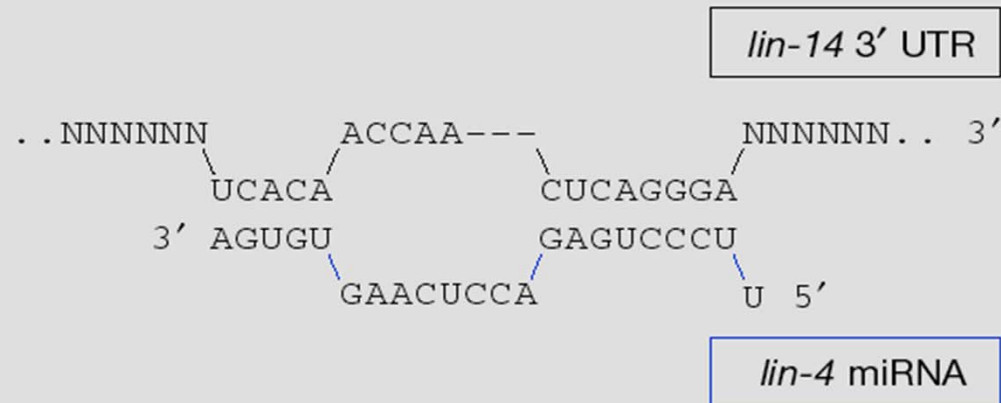
- 在线虫的幼虫发育期，lin-4下调LIN-14和LIN-28蛋白质的浓度，从而一方面阻止后期发育，一方面促进幼虫的发育



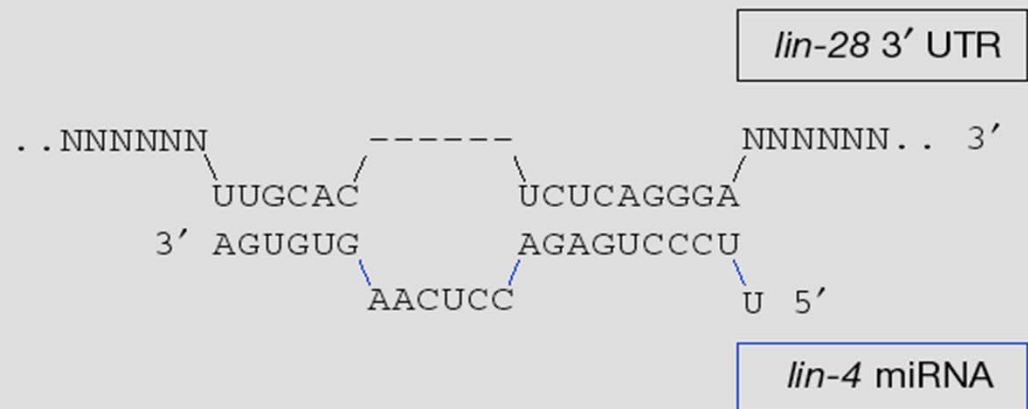
miRNA vs. 靶序列: 不完美配对



a



b

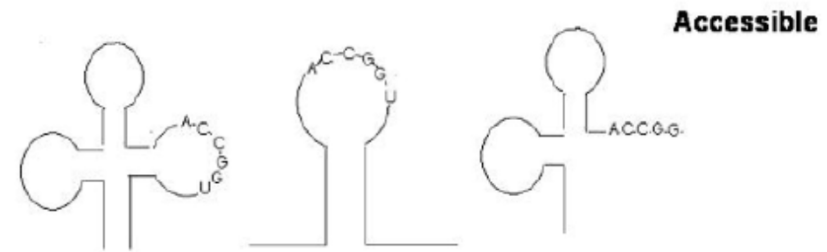


miRNA底物的计算发现



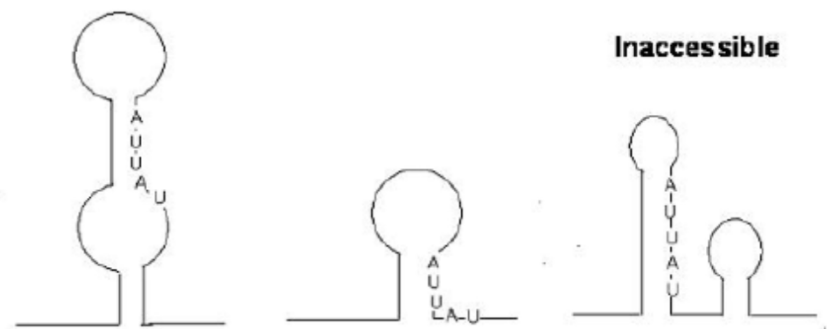
- miRNA的靶序列位于基因的3' UTR区域，与成熟的miRNA互补

- ✿ 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

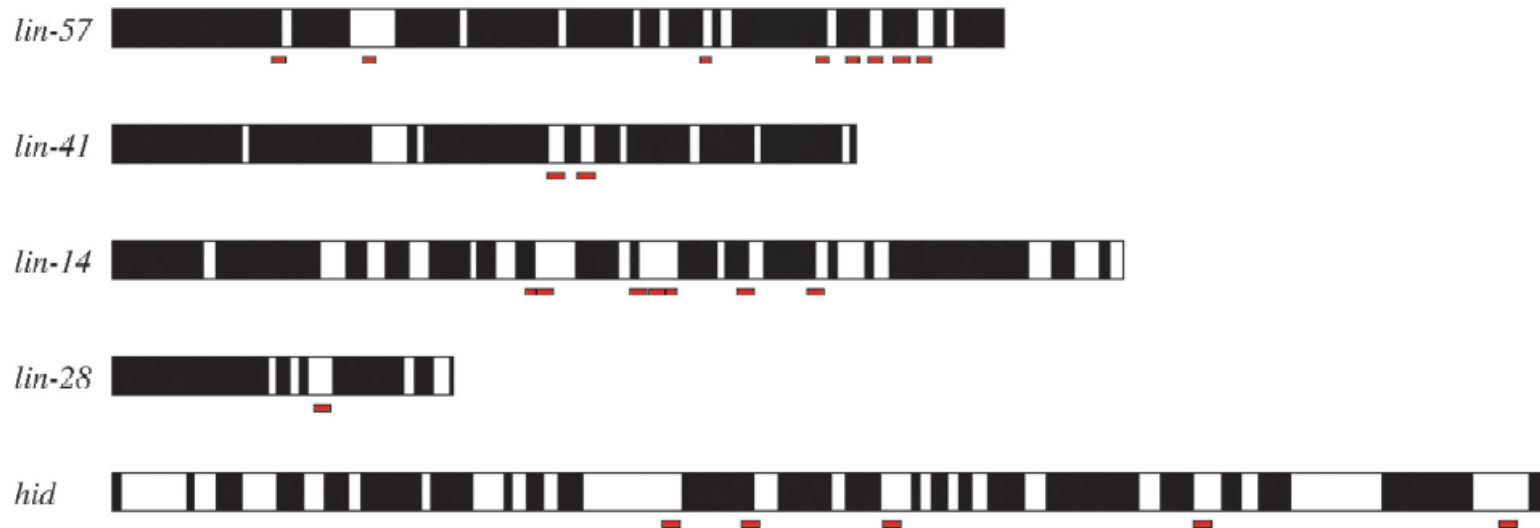


- 提高准确性：考虑mRNA UTR的二级结构

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

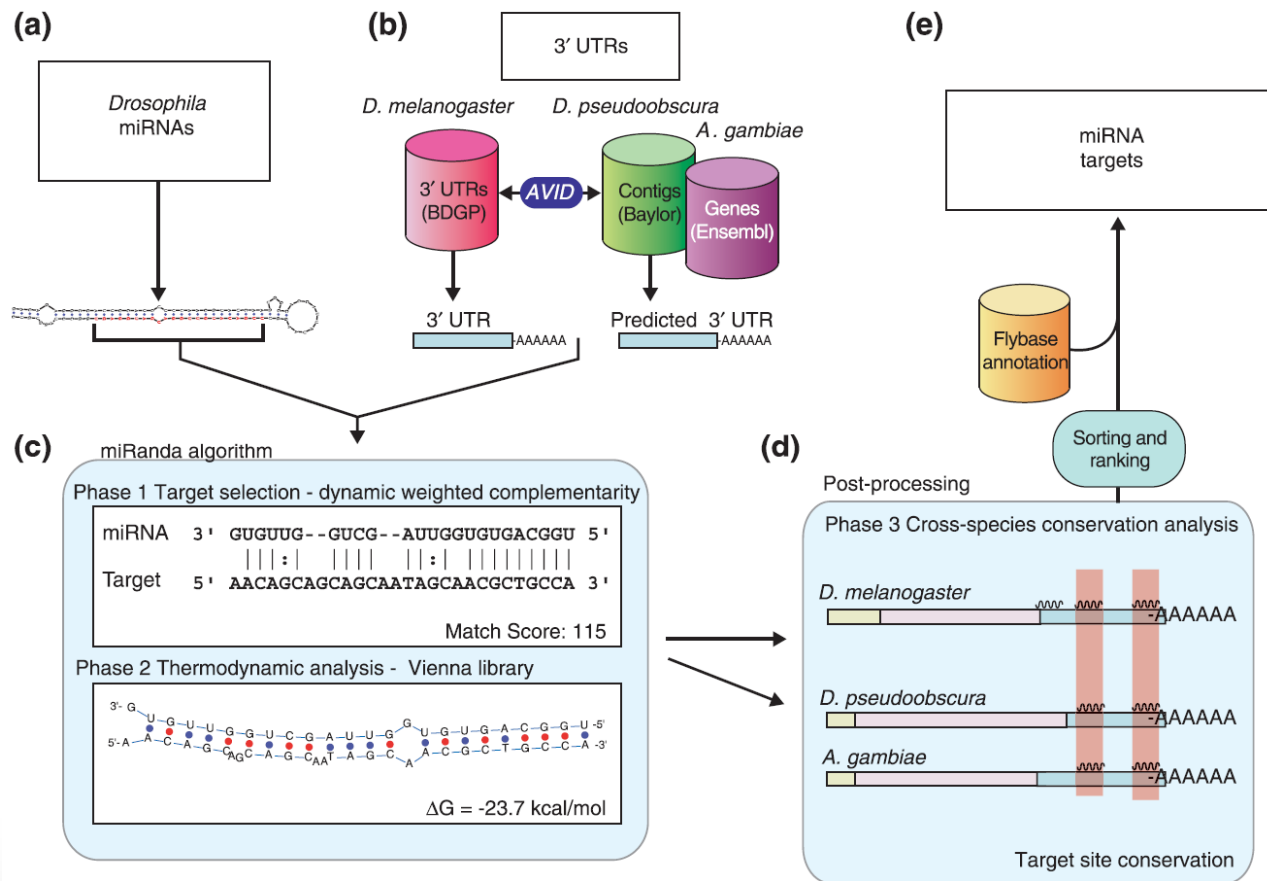
✿ 与miRNA的5'端匹配更好

let-7	
lin57-1	T T T C T A T T A T A C A A C C G T T C C A C C T C A
lin57-3	C T T A C C T G T A T A A T G C C T T C T A C C T C C
lin57-5	A C T G T T C T C A G T A C A T G T A G T A C C T C C
lin57-6	T T T C T C T C T G T C T C A C T T T C T A C C T C C
lin57-7	A C T A T C T C G C A C T T T C A T T C T A C C T C A
lin57-9	T A C T T G T C C G C T A C C T T A T G T A C C T C A
lin41-1	A C G T T T T A T A C A A C C G T T C T A C A C T C A
lin41-2	C C C T T T T A T A C A A C C A T T C T G C C T C T

lin-4	
lin14-1	C T C A C C T C A A A A A T T G C T C T C A G G A A
lin14-2	T C T C A G G A A C A T T C A A A A C T C A G G A A
lin14-3	C A C T C T C T T T T A A T C C A A C T C A G G G A
lin14-4	A T T T T T T T T C T C A T T G A A C T C A G G A A
lin14-5	C T C A G G A A T T T C T T T C T A C C T C A G G G A
lin14-6	T T A G C T T T T A A T G T T A A A A T C A G G A A
lin14-7	G T C A A A A C T C A C A A C C A A C T C A G G G A
lin28-1	A C C T C C T C A A A T T G C A C T C T C A G G G A

bantam	
hid-1	G T T C A T C A T C A T A T T C A A A T T G G T C T C A
hid-2	T T T T T G G A A T G C A C A T T A A T G A T C T C T
hid-3	C C A A T T C C C A A A A A T C G C A T T G A T C T C A
hid-4	T T G C T A A T T A G T T T T C A C A A T G A T C T C G
hid-5	A T A T A C A T A A A T A T C A T T A T T G A T C T C A

miRanda

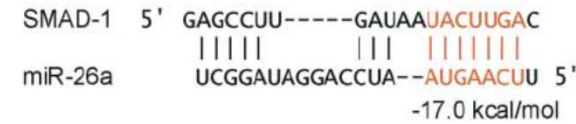
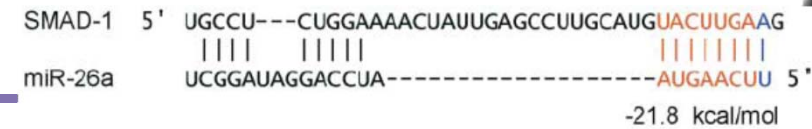


- “miRanda” Enright et al. Genome Biology 2003
- 结合上述特征，分配不同的权重
- 预测靶序列与miRNA 5'端的结合情况

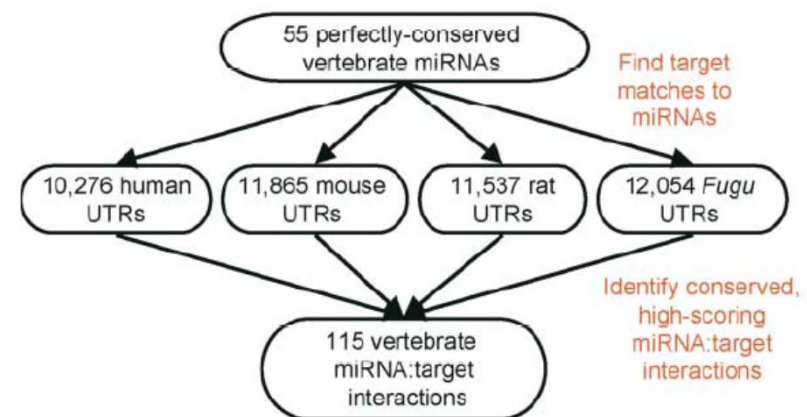
Targetscan



- 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



$$Z = e^{-dG_1/T} + e^{-dG_2/T} = e^{21.8/20} + e^{17.0/20} = 5.3$$



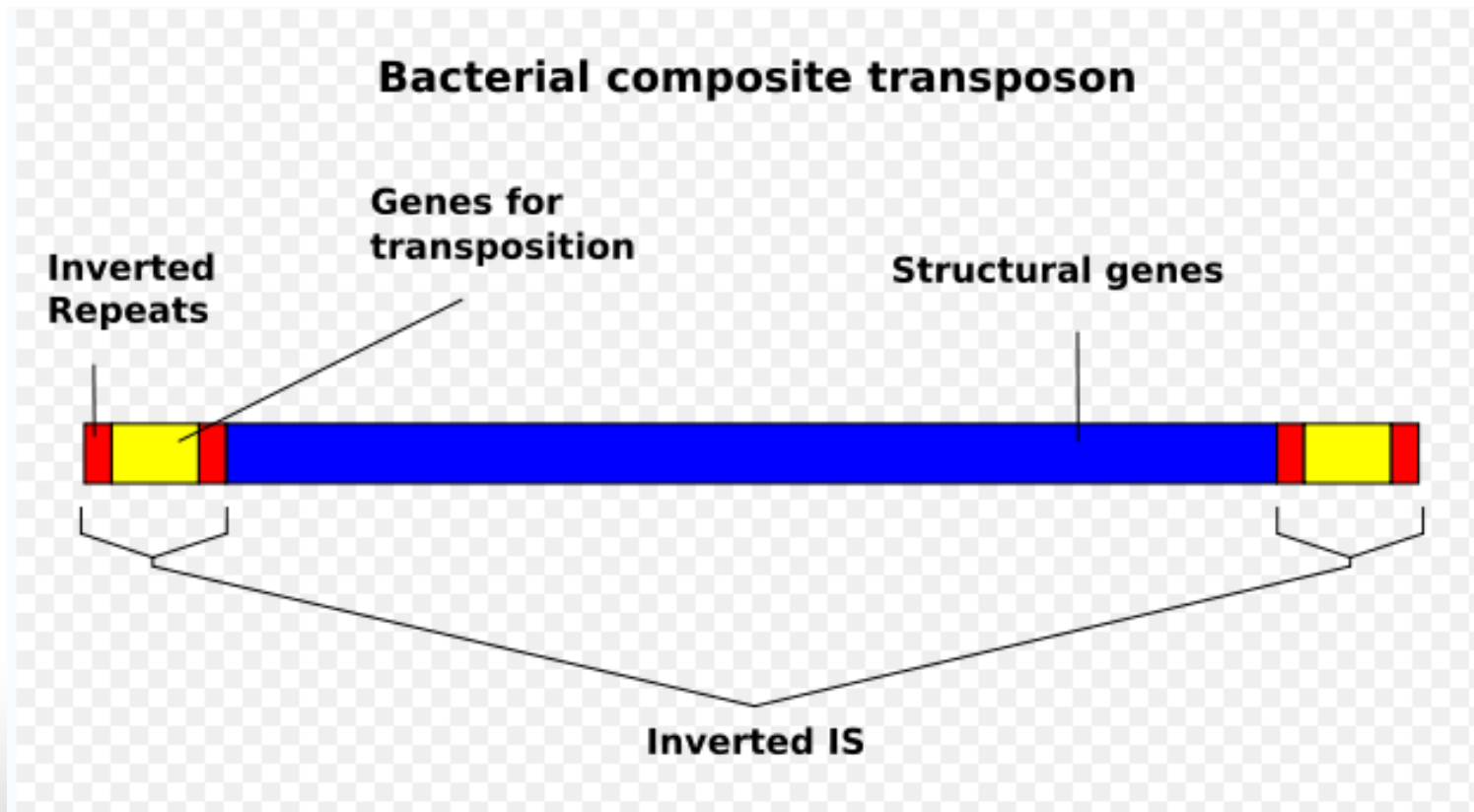
C

	0	100	200	300	400	500	600	700	Z	Rank
Hs		—							5.3	45
Mm		—							4.8	72
Rn		—							4.9	76
Fr				—					5.2	16

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
- 可变剪接的预测
- 非编码的功能元件的预测

基因组测序：鸟枪法



(computer-aided sequence assembly is used to deduce complete sequence of the original cDNA)

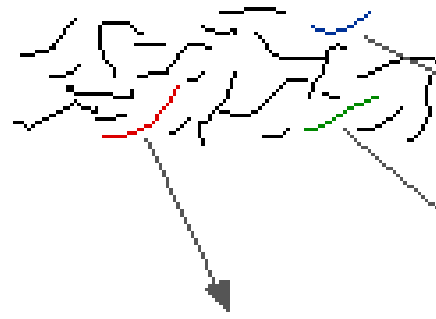
基因组的拼装



Whole Genome Shotgun Sequencing Method



Genomic DNA



Sequence Each Fragment
with Shotgun Approach

GCATTTCGAGTTACCTGGACAACCACTG

CCAGTGGTACTGAGGACGCCAAGAGGCTTGA

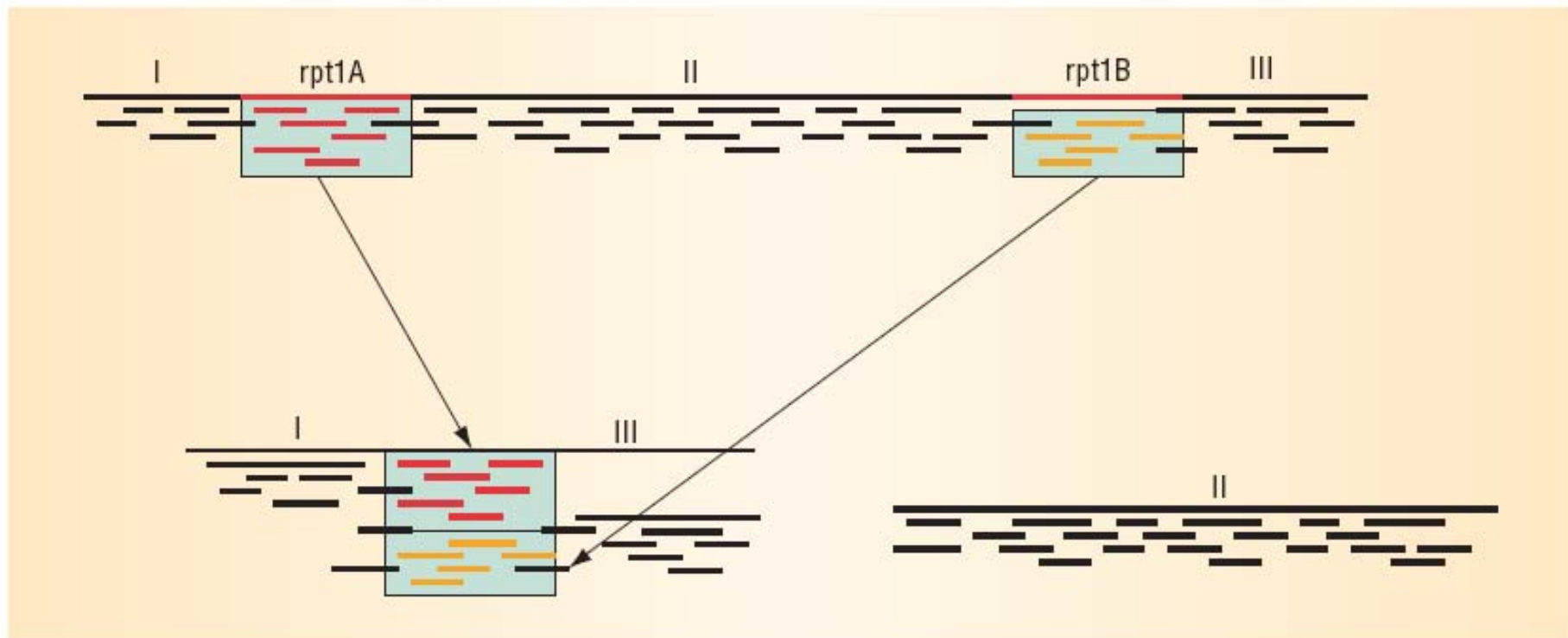
GCTTGATTTGGCCAATAATAGTATAT

Align Contiguous Sequences

GCATTTCGAGTTACCTGGACAACCACTGGTACTGAGGACGCCAAGAGGCTTGATTTGGCCAATAATAGTATAT

Generate Finished Sequence

重复序列带来干扰

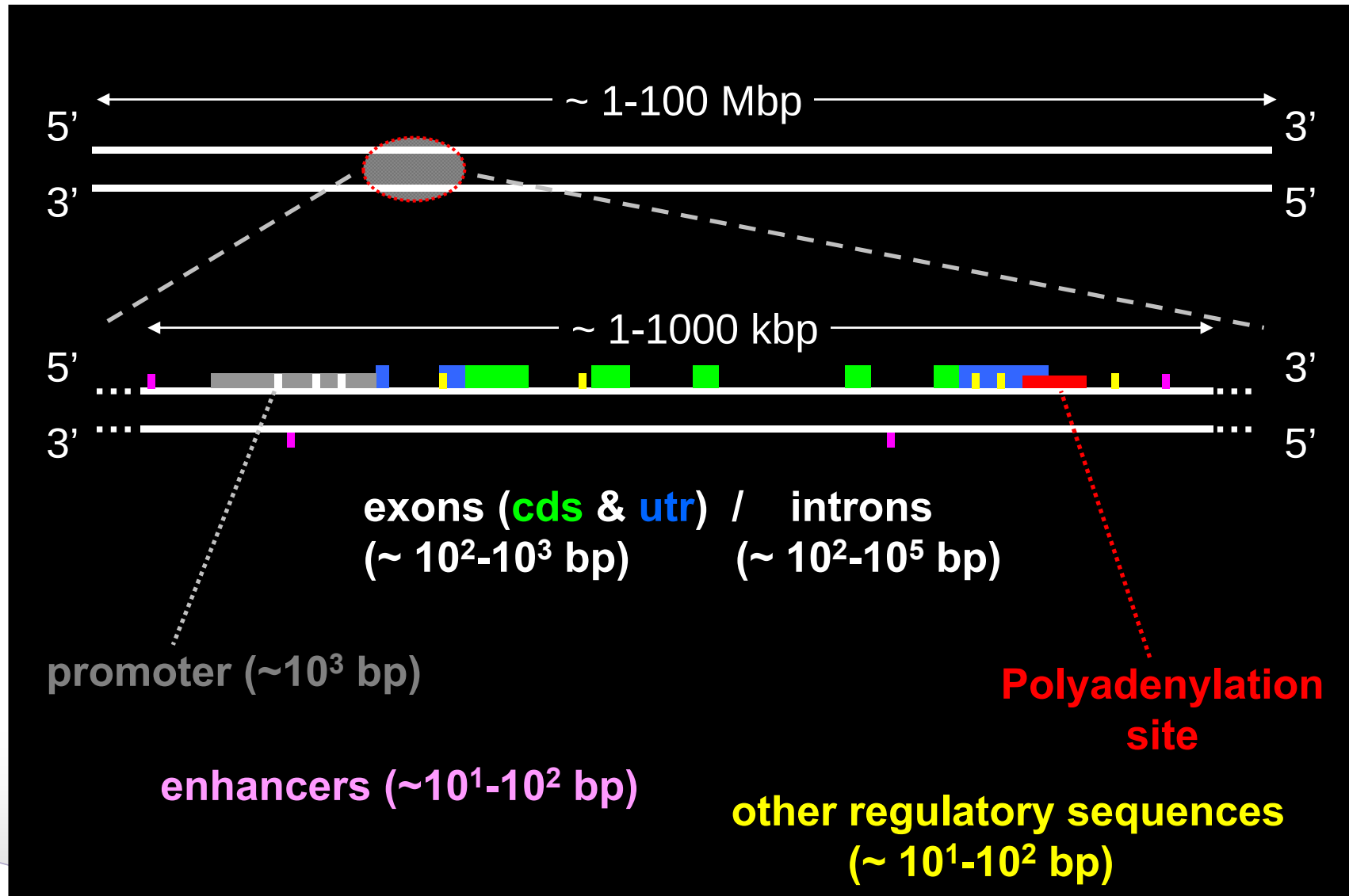


基因预测



- ❑ 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

真核生物基因结构

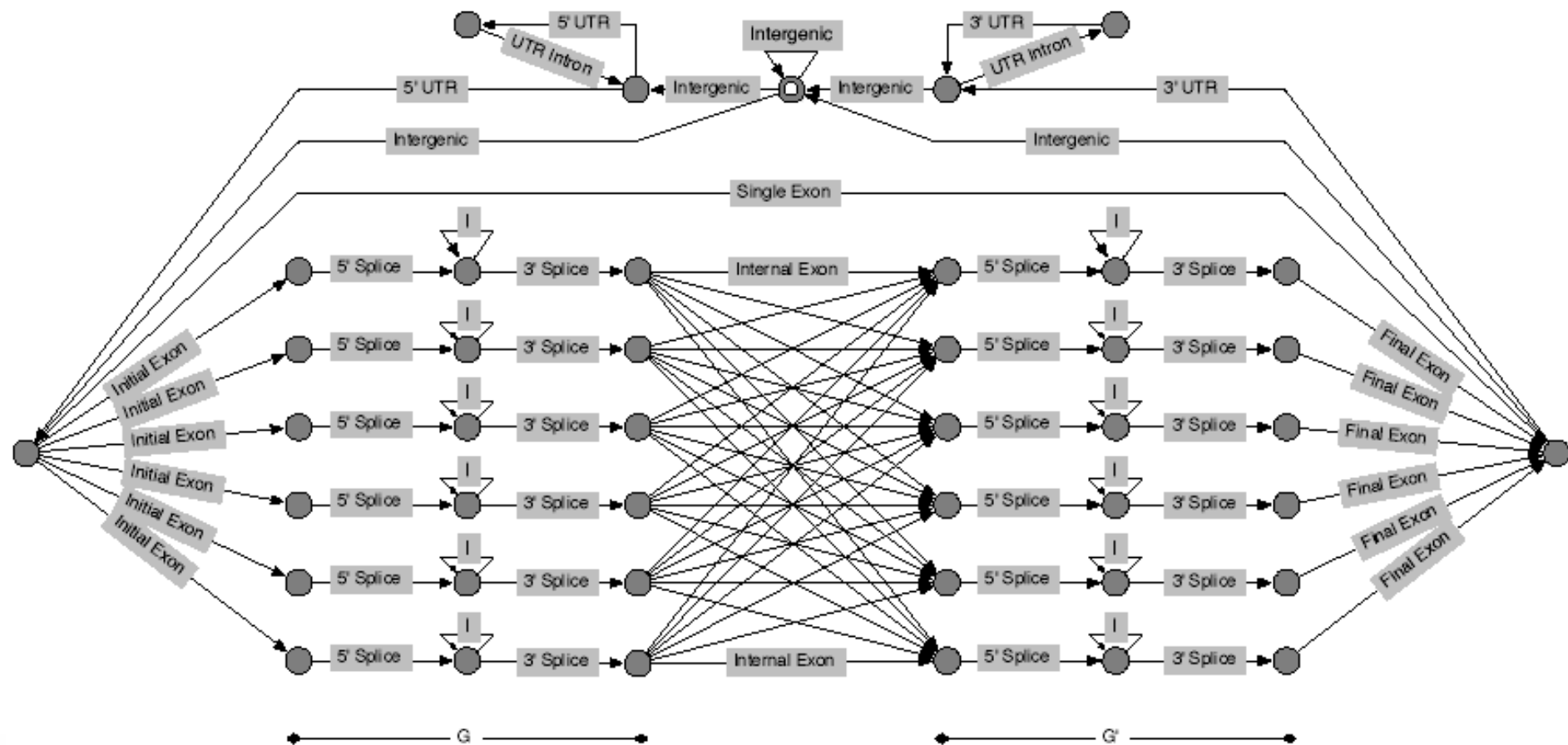




基因的其他特征

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

HMM model for Gene Prediction (Genie)

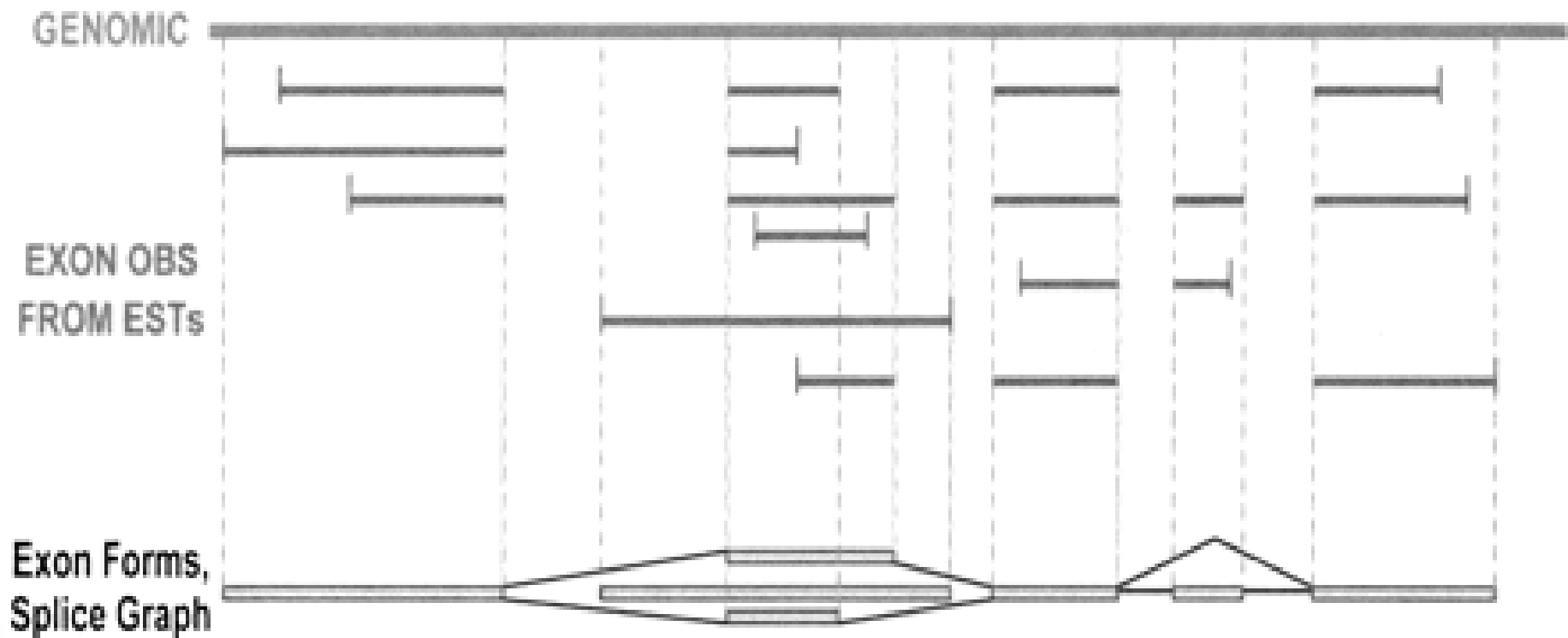


Kulp, D., PhD Thesis, UCSC 2003

可变剪接的预测



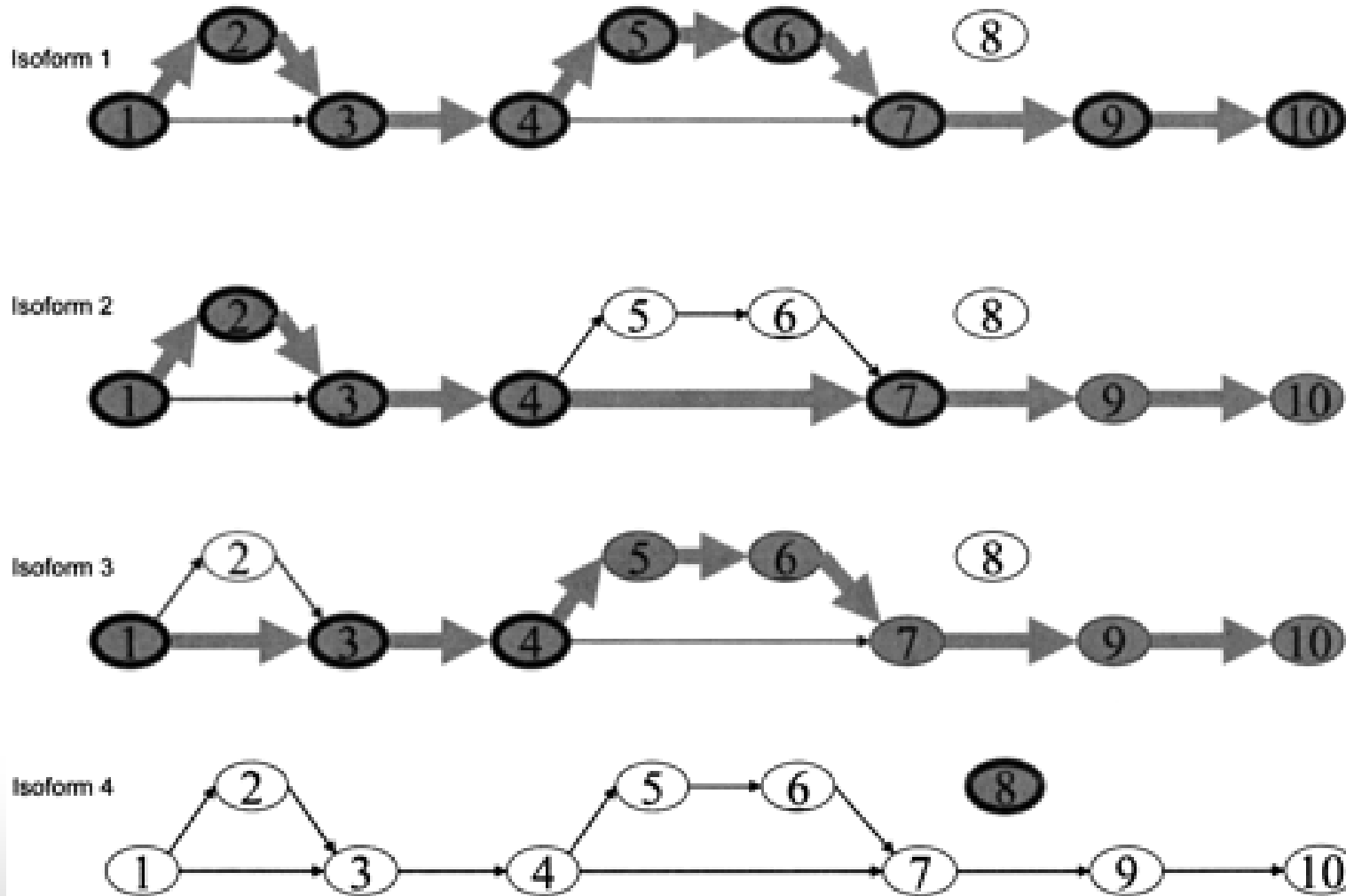
□ 将EST, cDNA序列比对到基因组上



部分有向图算法 (POA)



B



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



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

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-related protein 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".

A.

PKI α
PKI β 1
HIV-1 Rev
Scer Glelp
Scer Rnalp
Spom Rnalp
Hum RGP1
Mus FUG1
Xl RanGAP1
StpurRanGAP1
Cel repeat1
Cel repeat2

NES

ETALKIAGTIN	ETALKIAGTIN
DIPLKIEALAVK	DIPLKIEALAVK
QLP FLERITLD	QLP FLERITLD
ALP LGKLTLY	ALP LGKLTLY
EKGNLP ELEKLEINGN	RIDED SDALDLLQSKFD DLEVDDFEE
IDEKMP DLLELEINGN	RESEE DDVVDEIREVFS TRGRGELDE
AMADKA ELEKLEINGN	TIGEEGCEQLQEVLEGFNMAKVLASLSD
AVADKA ELEKLEINGN	ATGEEGCEQLQEVMSFNMAKVLASLSD
SVEDKS DLEKLEINGN	CTGEEGCEQVQEILESINMANILGSLSD
SMDTKP HTTLEINGN	NIG.....
SPKMCH VKVDISVNMF	GKDFDSAKARHGKGNIDFGRRGDELLS
KFDGLT PKPVITHTN	STGDEFSDVAGVAPENVNVGDEDDLG

316 331 332 357

B.

Mus FUG1
Scer RNA1
Spom RNA1
Hum RGP1
Mus FUG1
Xl RanGAP1
StpurRanGAP1
Cel repeat1
Cel repeat2

NLS

LMV LNHVVRQDYFKAAP	LLAFVTKPNGALETCSFARHNL	LQTLYNI	541-589
WKDSLIFELNLNDCLLKTAGSDEVFKVFTEVKFPNLHVLFKEYNEMACETIEVSFLPAM			258-315
WPN LRELCLNDCLLSARGAAAVDAFSKLENIGLQTLRTQYNEIELDAVRT	LKTV		
LRQ VEVINFGDCIVRSKCAVAIADAIRGGLPK	LKELNLSFCETIKRDAALAVAEAM		
LRQ VEVINFGDCIVRSKCAVAIADAVRGGLPK	LKELNLSFCETIKRDAALVVAEAV		
LRQ VEVINFGDCIVRSKCAQAIASALKEGLHK	LKELNLSYCEIKADAAVSLAESV		263-317
LSK LEVINFGDCIVRESCADAIANSIREGVPS	LKELNLSFCETIKRDAALVVAEAM		
LQF IEVLDLGDCVCDDEGLATIAELDKINRDCLKKVVLGGNNITSTVIDEIGACFN			
WPK LEVLNLSDCILRDACCNVITDHLNPQHRRHLKNVYLCGNELTPPVAKLLIQKWS			

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, Søren 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:

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: mTP: SP: other:

神经网络算法



- 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**

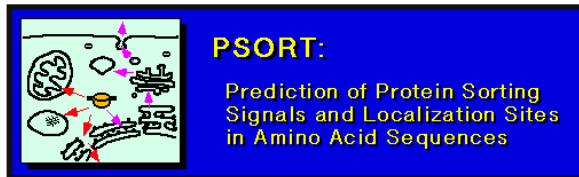
Institute of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, People's Republic of China

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 origin, *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 possibility 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](#)

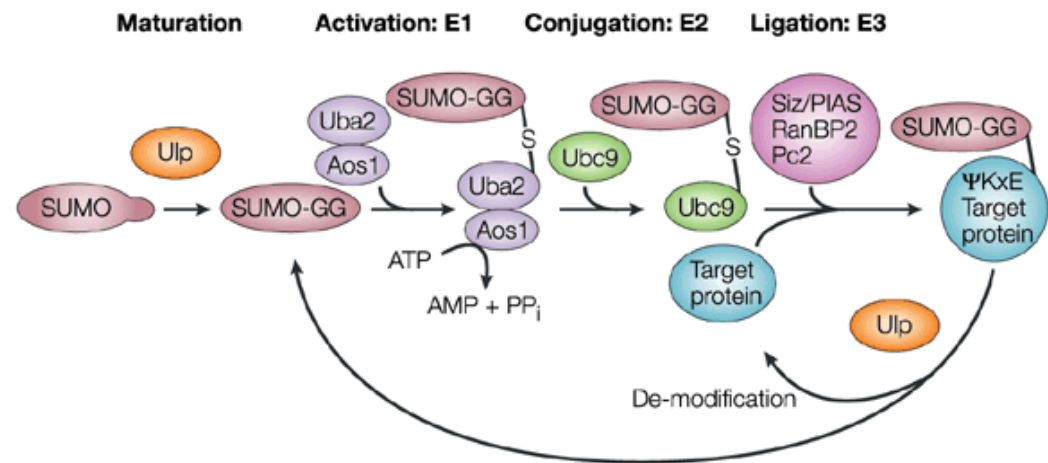
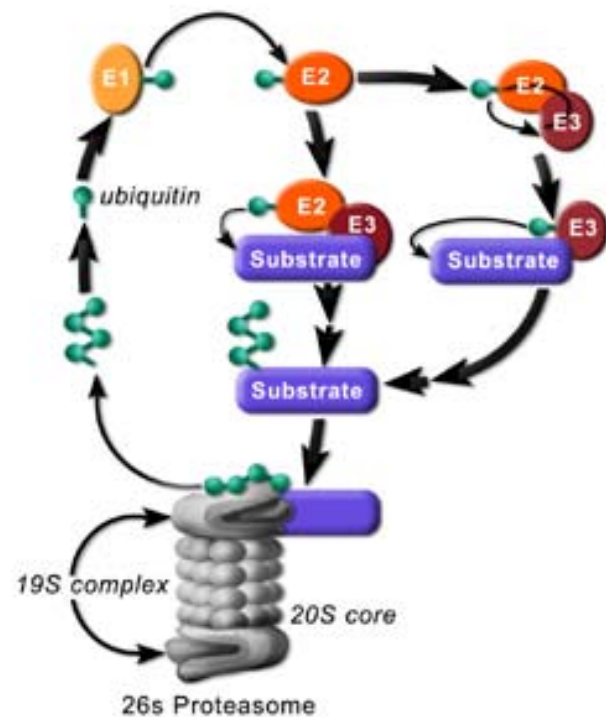
[PSORT II Prediction](#)



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



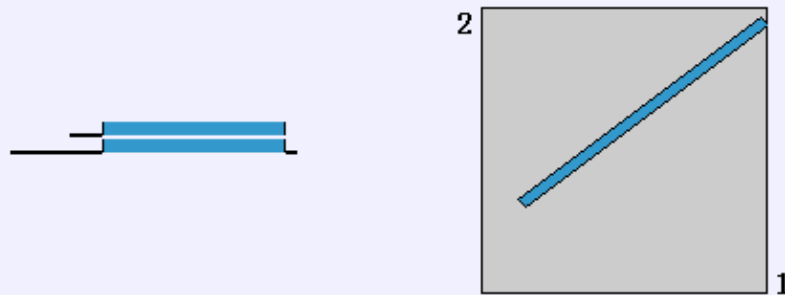
Nature Reviews | Molecular Cell Biology

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 IHFKVKMTHTLKKLKESYCRQGVPMMNSLRFLEGGQRIADNHTPKELGMEEEDVIEWYQE 93

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

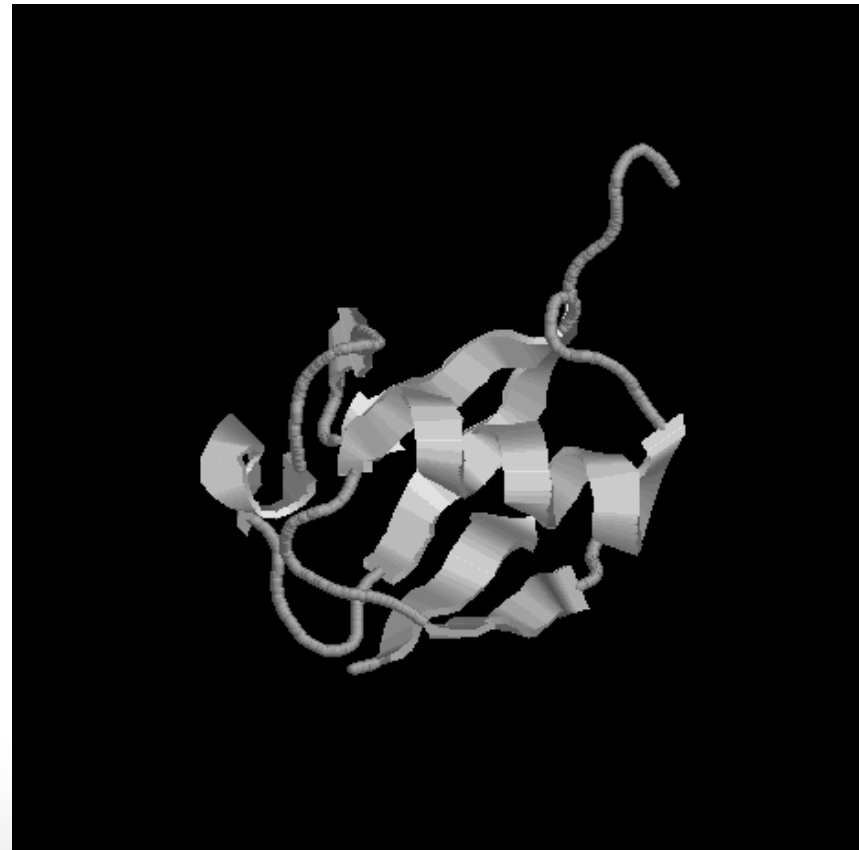
结构相似性



Ubiquitin



SUMO



蛋白质-蛋白质相互作用

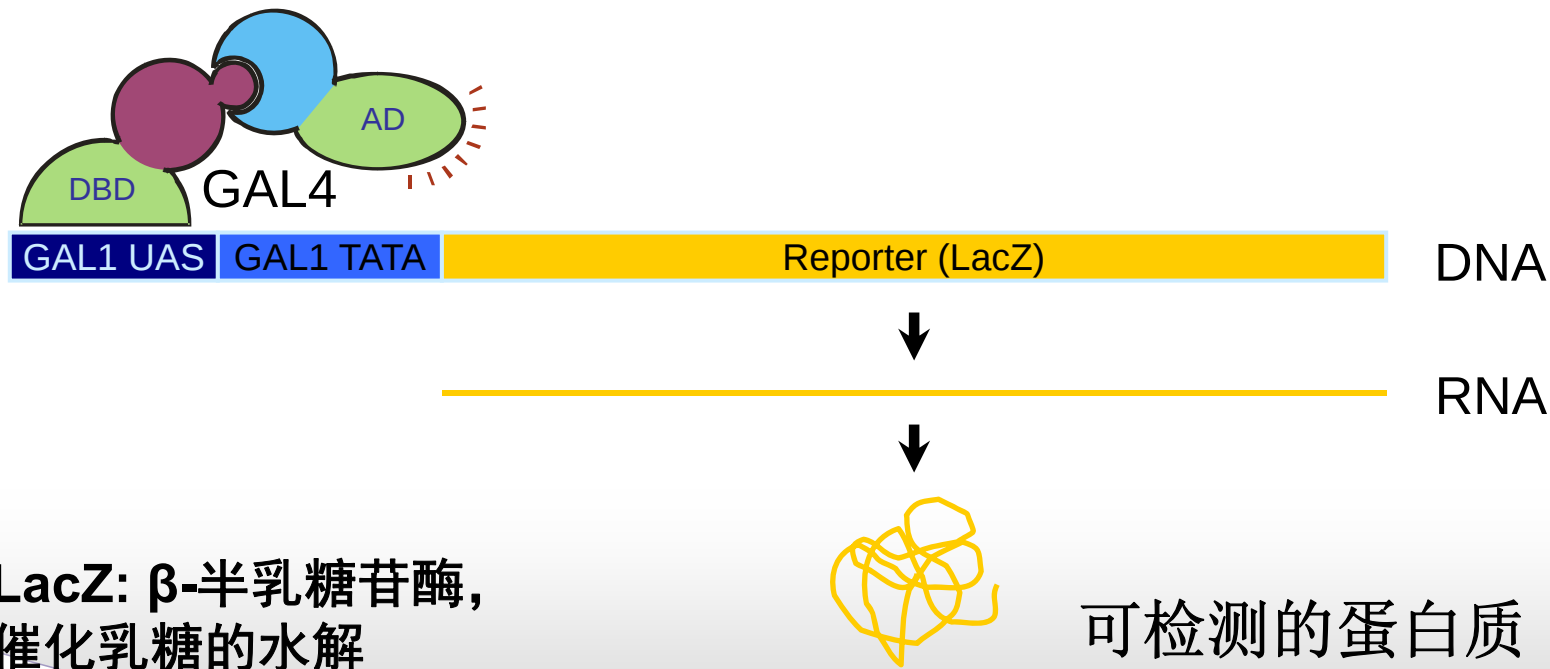


❑ 酵母双杂交, Yeast two-hybrid, Y2H

✿ GAL4-DBD-bait (hybrid 1)

✿ GAL4-AD-prey (hybrid 2; single or library)

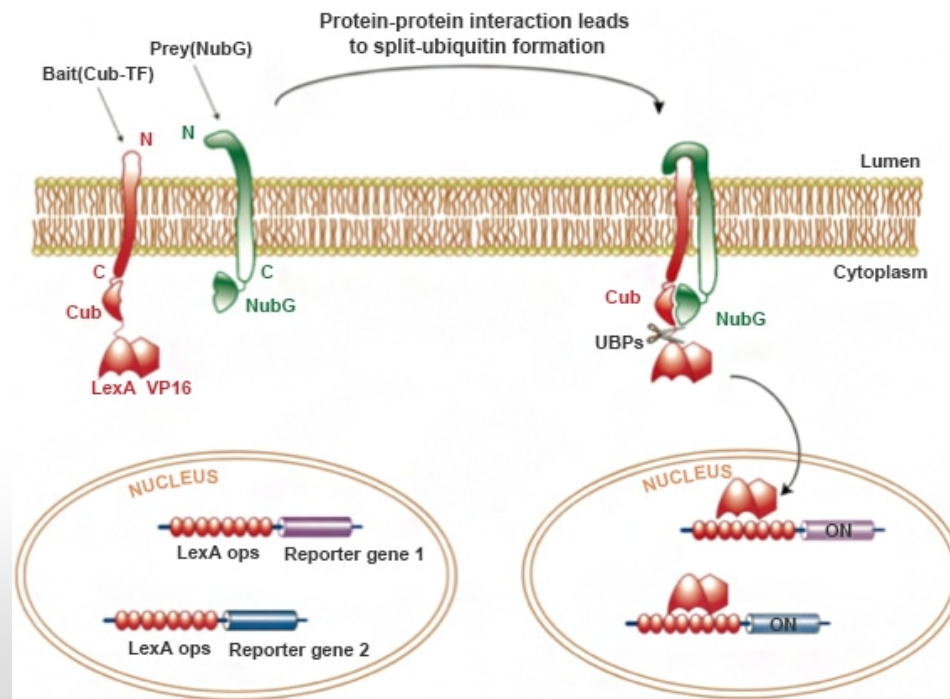
✿ bait-prey相互作用: GAL4激活报告基因



Split-ubiquitin yeast two-hybrid



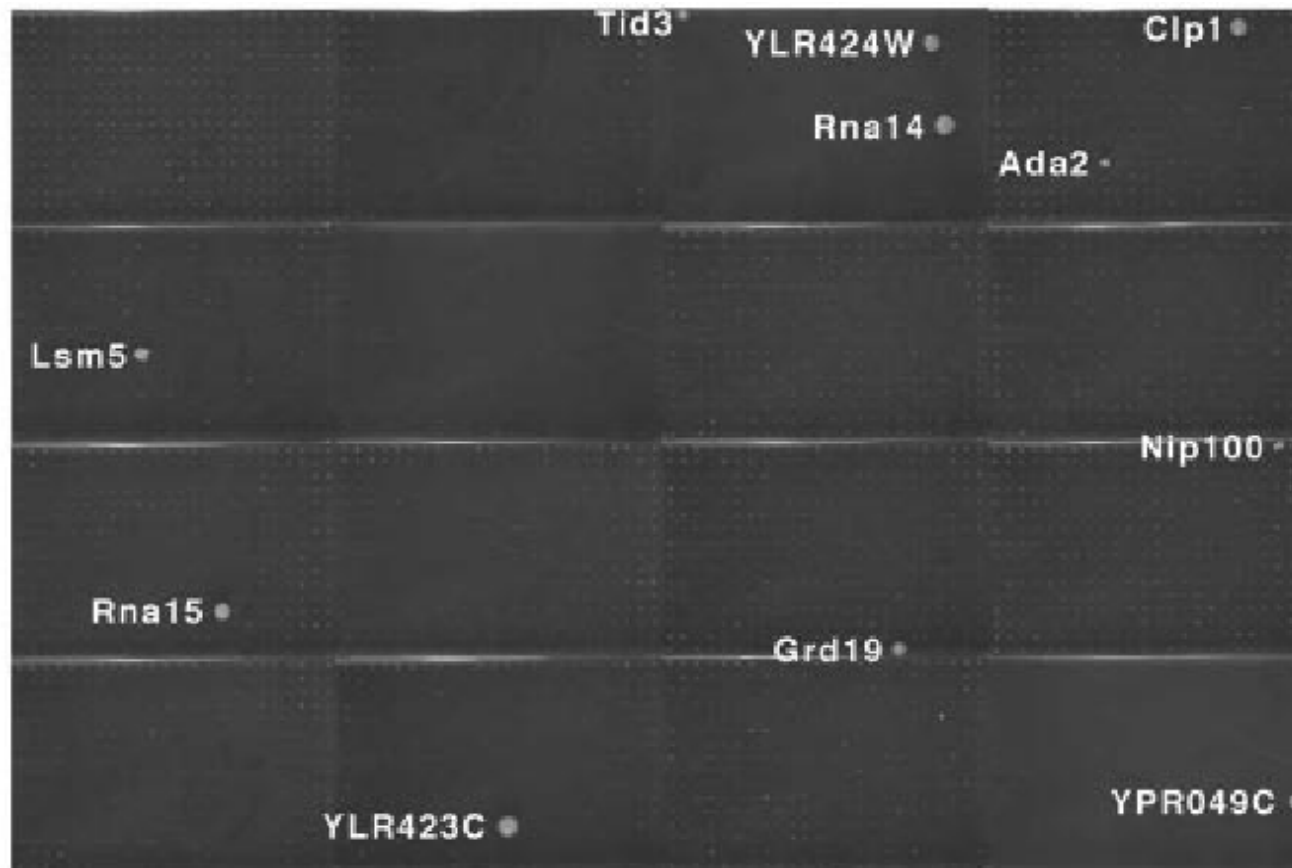
- ❑ 传统酵母双杂交的缺点：只能筛选可溶性蛋白质
- ❑ 膜蛋白的相互作用筛选
 - ✿ 两个膜蛋白分别与泛素分子的C端片段 (“Cub”, 35–76) 和N端片段 (“Nub”, 1–34) 连接
 - ✿ Cub与转录因子融合，并且可被泛素蛋白酶切割
 - ✿ bait-prey相互作用，泛素分子被切割，转录因子核内诱导报告基因表达



酵母蛋白质-蛋白质相互作用



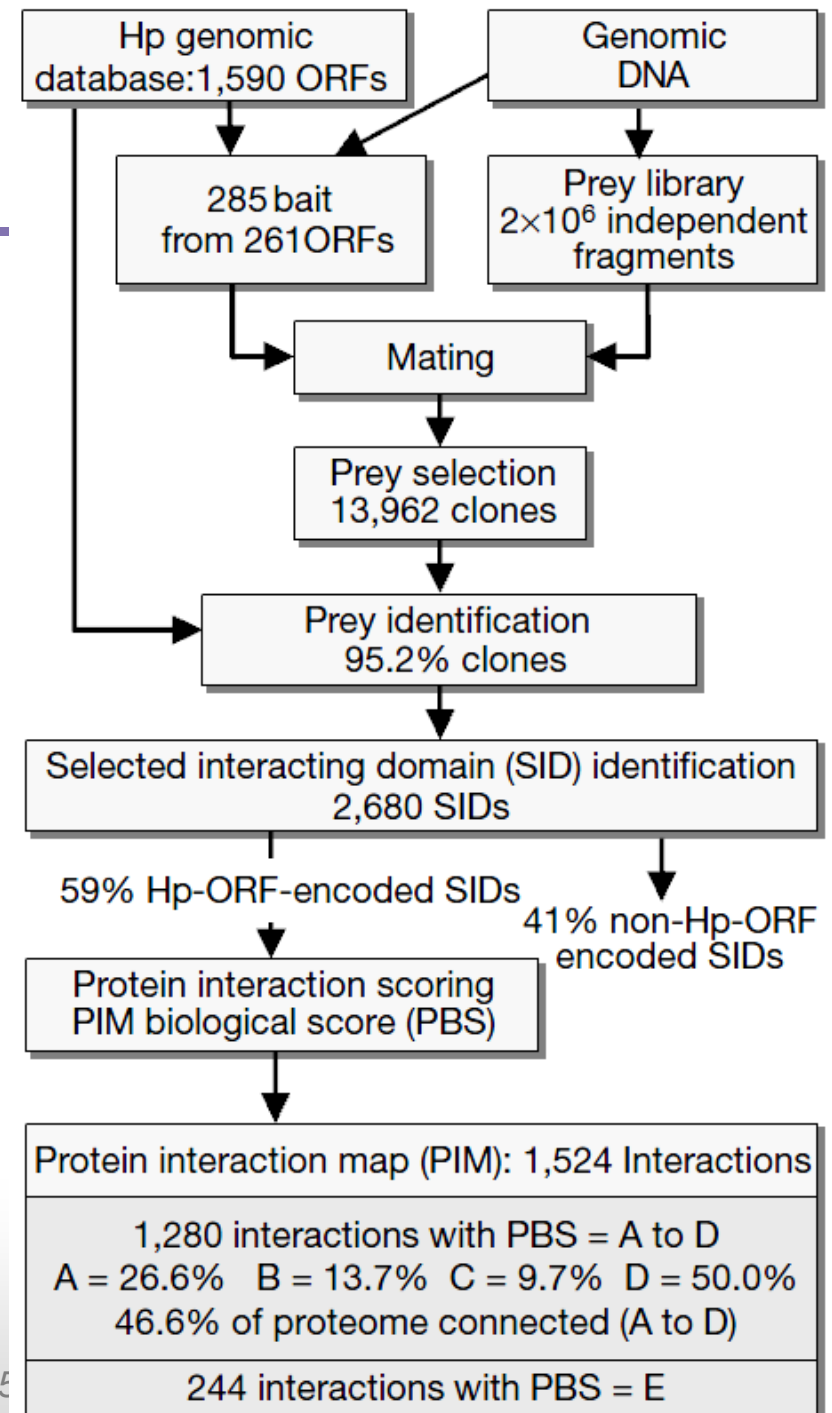
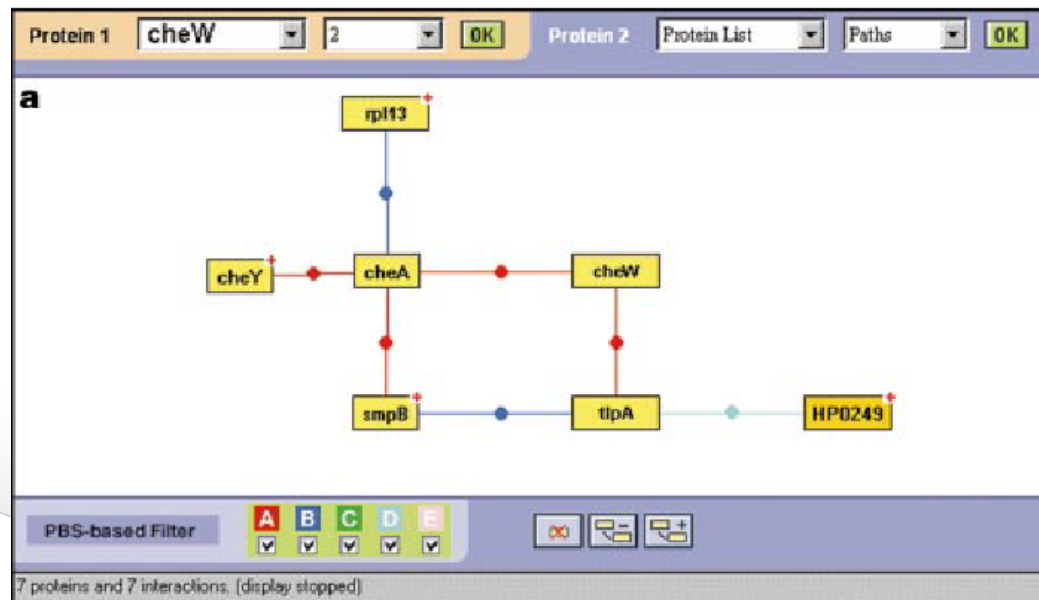
- ❑ 2000年, 6,000 * 6,000 ORFs
- ❑ 1,004个蛋白质, 957对相互作用



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

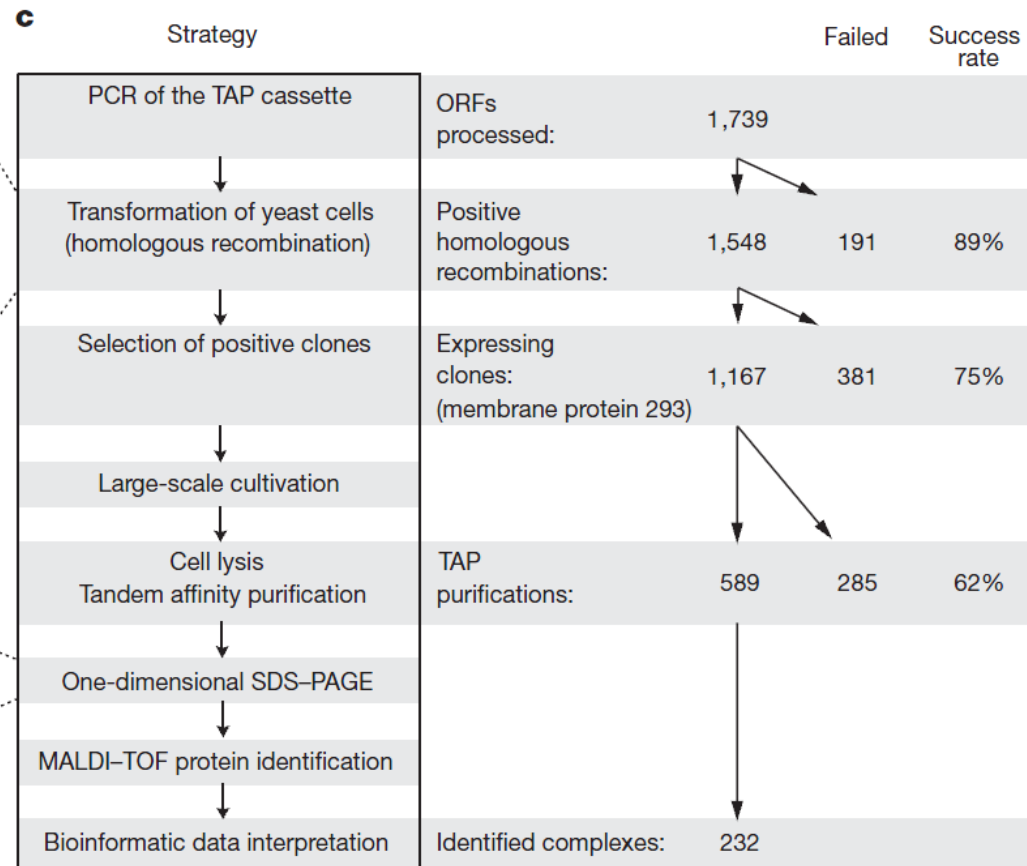
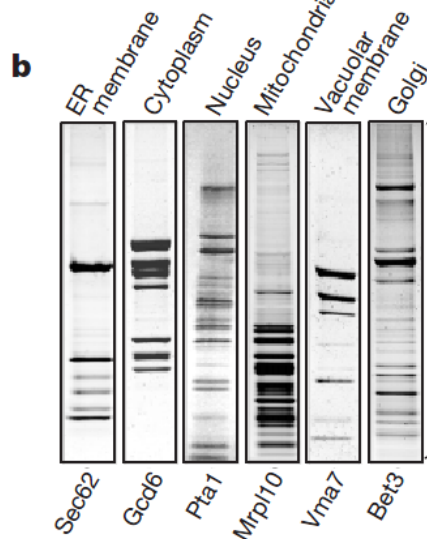
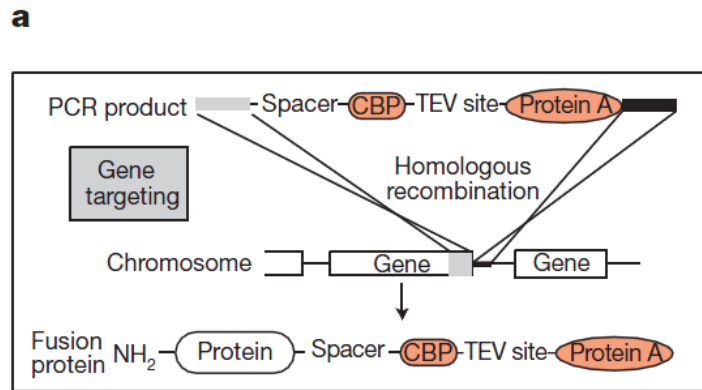


酵母蛋白质复合物



2002年，酵母蛋白质复合物

❁ 纯化589个蛋白质，获得232个复合物

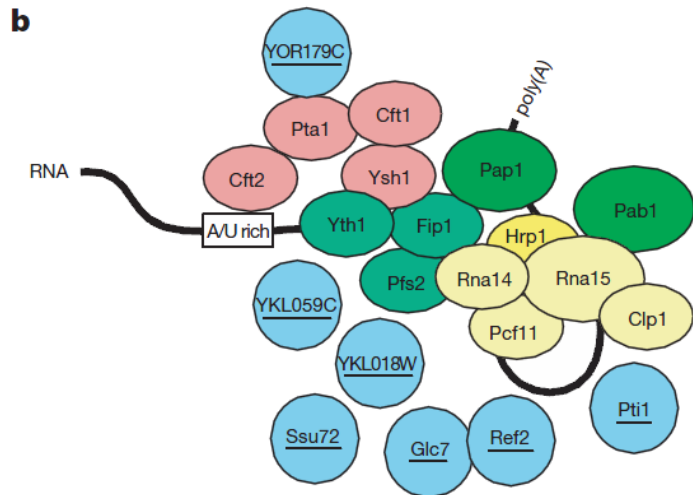
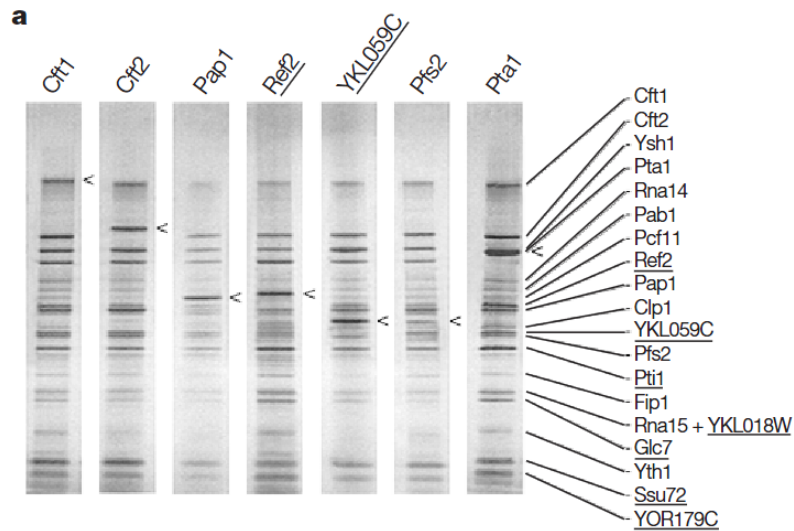


质谱鉴定

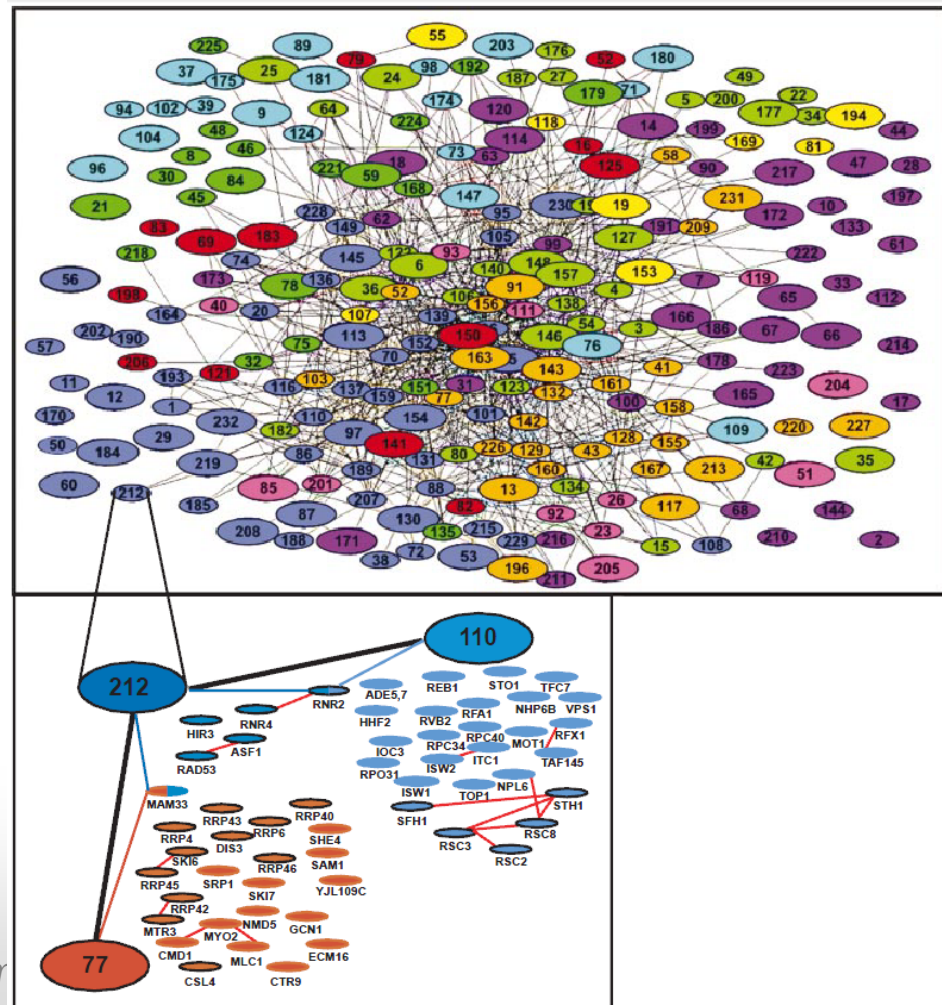
酵母蛋白质复合物



多腺苷酸化的分子机器



酵母相互作用网络



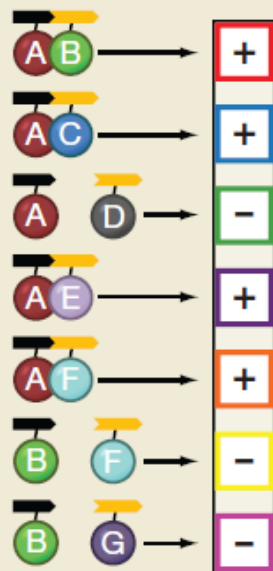


蛋白质-蛋白质相互作用

□ 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 resistance Fluorescence



Binary interactions between protein fragments:

- A-B: +
- A-C: +
- A-D: -
- A-E: +
- A-F: +
- B-F: -
- B-G: -



Protein pairs

+ Interacting
- Noninteracting

Interaction examples

- Signaling
- Enzyme - substrate

Interaction strength

- Transient to stable

蛋白质-蛋白质相互作用



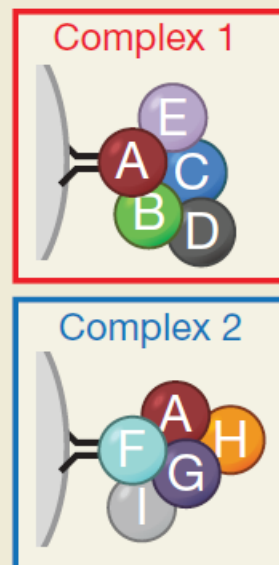
□ 亲和纯化/质谱

Molecular machines/Protein complexes comembership

Methods

Affinity purification/Mass spectrometry

Biochemical purification of affinity-tagged baits followed by MS identification of copurifying preys



	A	B	C	D	E	F	G	H	I
A	●	●	●	●	●	●	●	●	●
B	●	●	●	●	●	●	●	●	●
C	●	●	●	●	●	●	●	●	●
D	●	●	●	●	●	●	●	●	●
E	●	●	●	●	●	●	●	●	●
F	●	●	●	●	●	●	●	●	●
G	●	●	●	●	●	●	●	●	●
H	●	●	●	●	●	●	●	●	●
I	●	●	●	●	●	●	●	●	●

○ Bait ○ Prey

●●●● Socio-affinity

Interaction examples

- Allosteric
- Chaperone-assisted

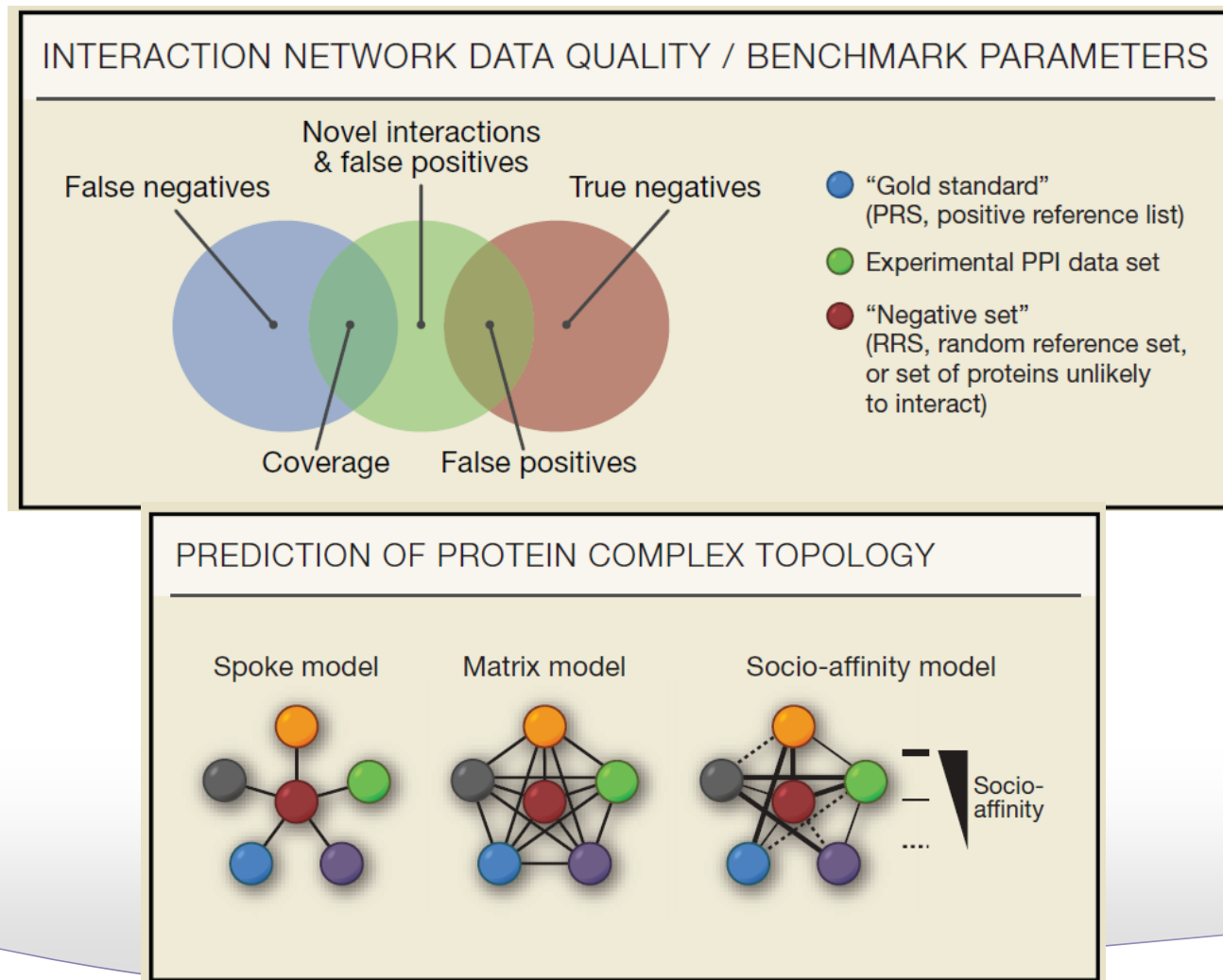
Interaction strength

-Stable



相互作用网络

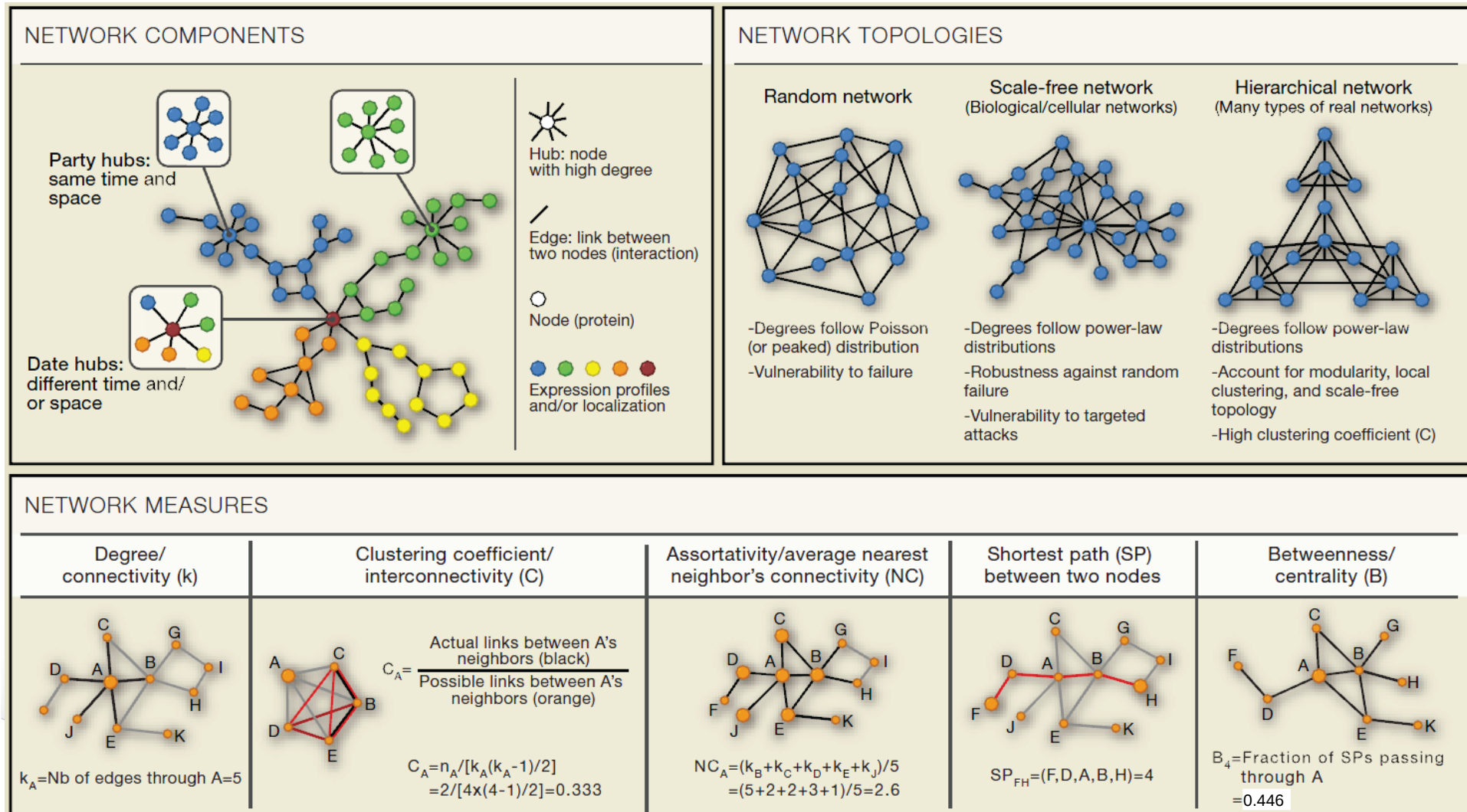
❑ 相互作用的数据质量，网络拓扑结构



蛋白质-蛋白质相互作用



网络分析：Hub节点、拓扑结构、参数



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 : $(N-1)(N-2)/2$, 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:

$$C_b(b) = ((\sigma_{ac}(b) / \sigma_{ac}) + (\sigma_{ad}(b) / \sigma_{ad}) + (\sigma_{ae}(b) / \sigma_{ae}) + (\sigma_{cd}(b) / \sigma_{cd}) + (\sigma_{ce}(b) / \sigma_{ce}) + (\sigma_{de}(b) / \sigma_{de})) / 6 \\ = ((1 / 1) + (1 / 1) + (2 / 2) + (1 / 2) + 0 + 0) / 6 = 3.5 / 6 \approx 0.583$$

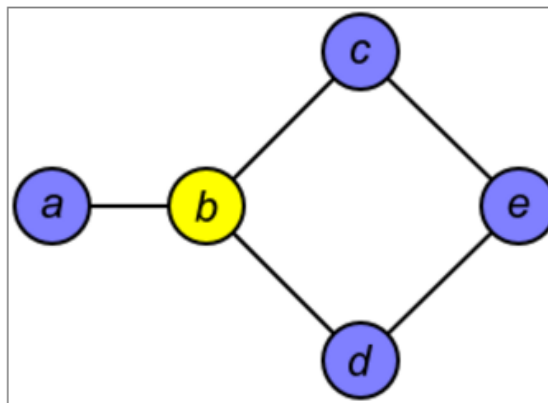


Figure 7 Example network with five nodes and five edges.

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



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

[Jobs](#)
[Help](#)
[News](#)
[Register](#)
[Statistics](#)
[Satellites](#)
[SEARCH](#)
[SUBMIT](#)
[Software](#)
[Services](#)
[Articles](#)
[Links](#)
[Files](#)
[MIF](#)

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 maintenance. 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 (<i>registration required</i>).
HELP	A short description of the DIP database.

MINT



□ 2002年, Molecular INTeraction

[Home](#) [Search](#) [Statistics](#) [Download](#) [Administration](#) [Contacts/Links](#) [About MINT](#)

Statistics
interactions: **125464**
articles: **5941**
proteins: **25530**
organisms: **611**



Molecular Genetics
Group

Bionformatics
Resources



Welcome to MINT, the Molecular INTERaction database.
MINT focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators.

PLEASE UPDATE YOUR BOOKMARK

The full MINT dataset can be freely **downloaded**.

[You are browsing a beta version of the new MINT interface.
Please, consider that some features might not work properly]

MINT has signed the **IMEx agreement** to share curation efforts and supports the Protein Standard Initiative (PSI) recommendation.



Starting September 2013, MINT uses the IntAct database infrastructure to limit the duplication of efforts and to optimise future software development. Data maintenance and release, **MINT PSICQUIC** and **IMEx services** are under the responsibility of the IntAct team, while curation effort will be carried by both groups. Data manually curated by the MINT curators can now also be accessed from the **IntAct homepage** at the EBI.

Other resources:
MENTHA: <http://mentha.uniroma2.it/>
VirusMENTHA: <http://virusmentha.uniroma2.it/>
SIGNOR: <http://signor.uniroma2.it/>

Please, in any articles making use of the data extracted from MINT, refer to:
MINT, the molecular interaction database: 2012 update.
Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardoza AP, Santonico E, Castagnoli L, Cesareni G. Nucleic Acids Res. 2012 Jan;40(Database issue):D857-61. doi: 10.1093/nar/gkr930. Epub 2011 Nov 16.

Search proteins in MINT:






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. [IntAct](#) [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 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 Q

Advanced Search

Search Tips

Featured Datasets

By Gene

By Publication

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**.

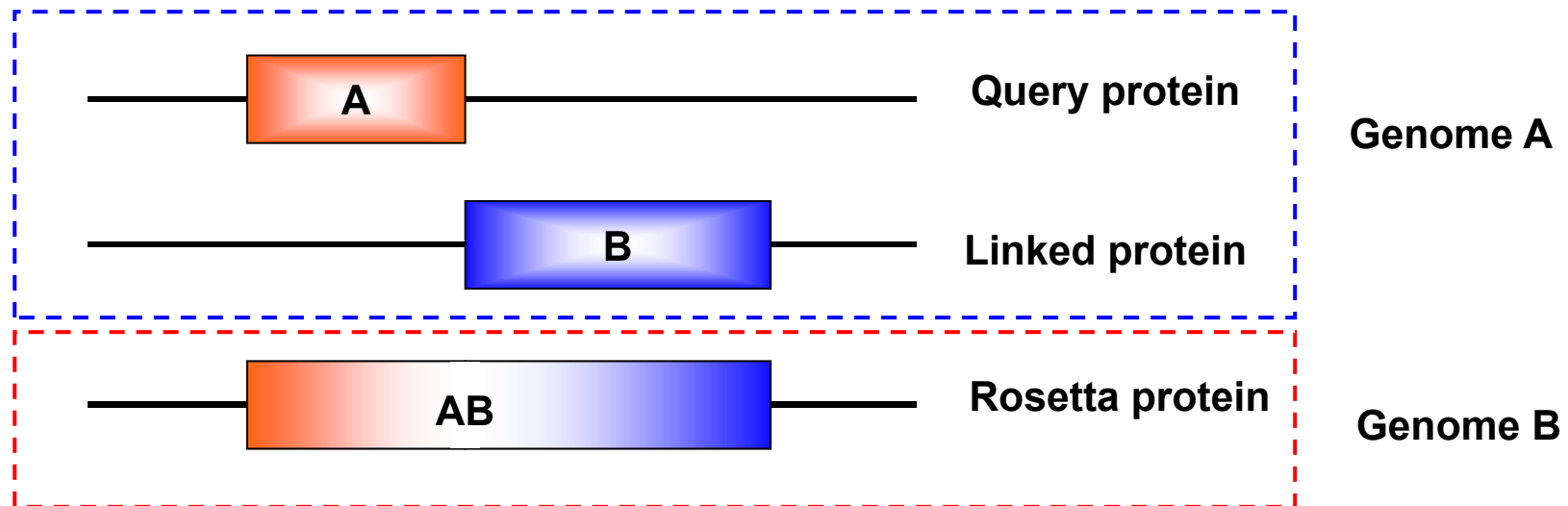
Posted: February 1, 2017 - 2:18 am

LATEST UPDATES

[Tweets by @biogrid](#)

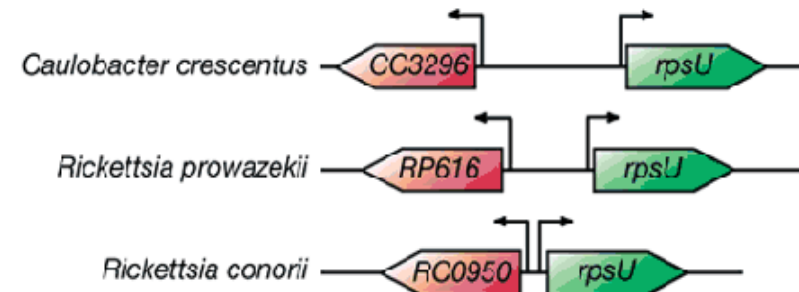
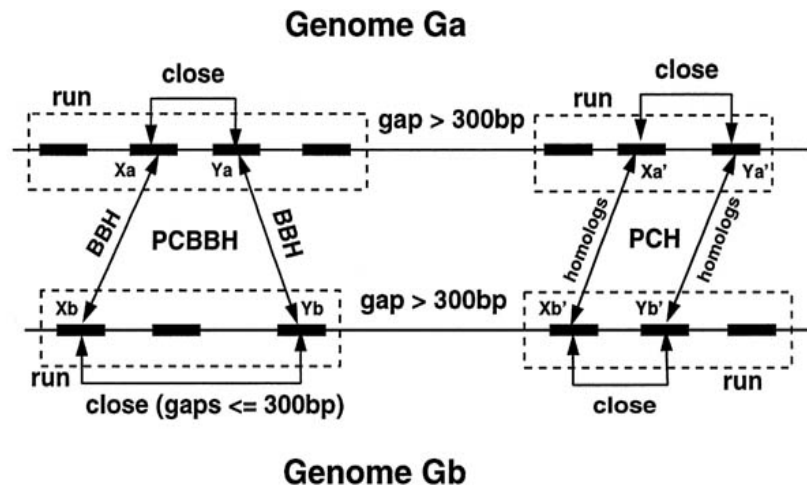
Bioinformatics, 2025, HUST

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

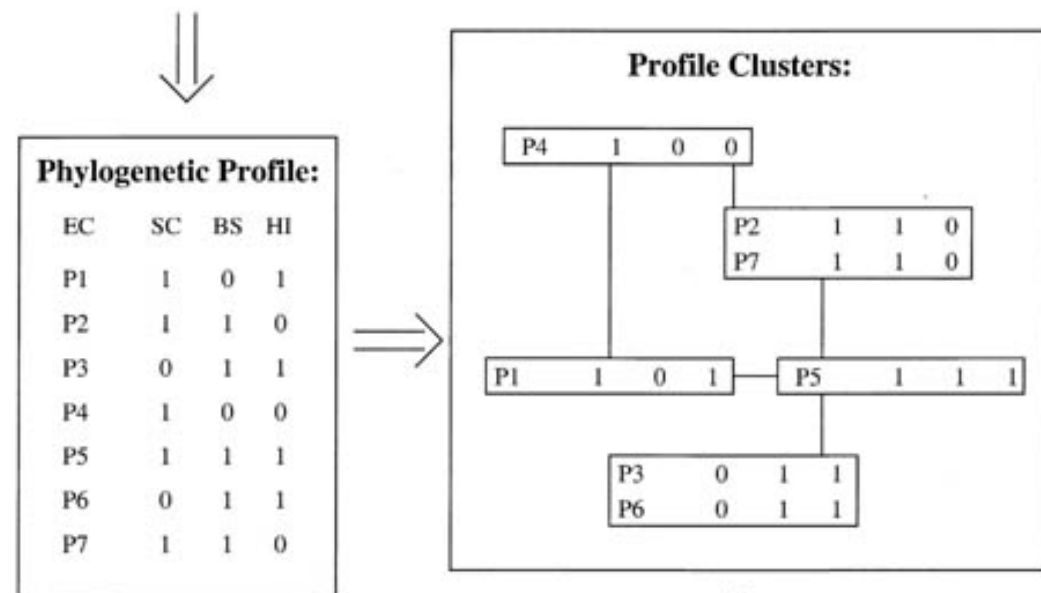
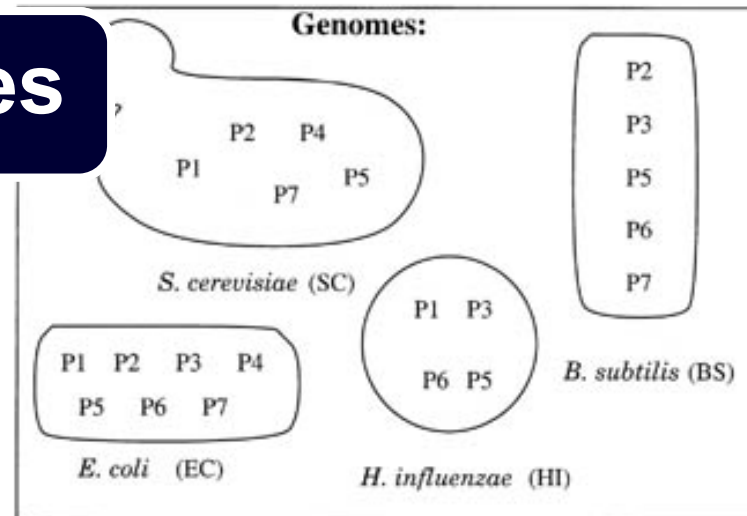


Gene order pairs

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

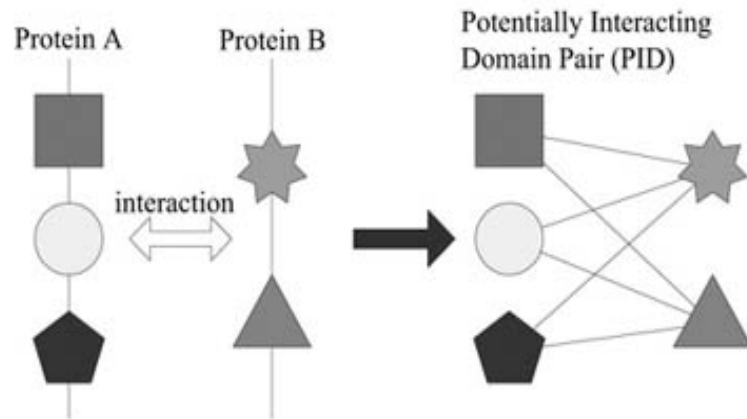


Pellegrini M et al., PNAS, 1999, 96:4285-4288;
Huynen MA et al., PNAS, 1998, 95:5849-5856

Bioi

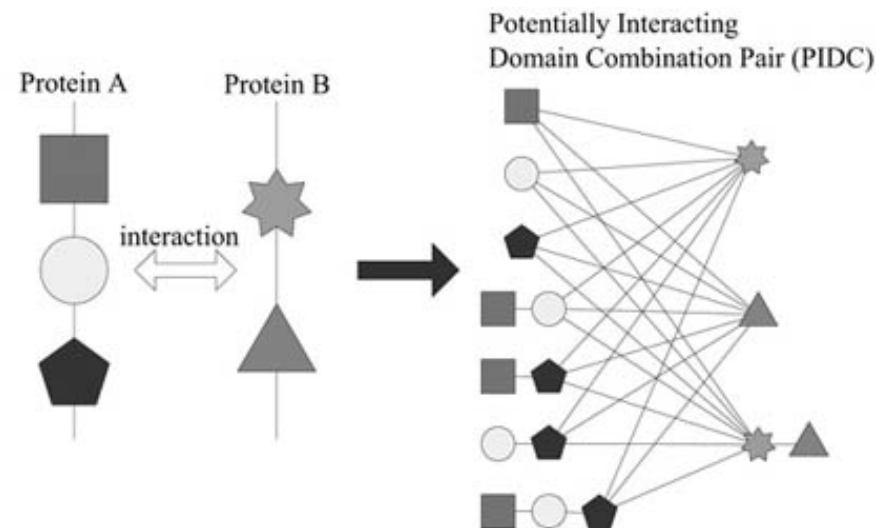
Conclusion: P2 and P7 are functionally linked,
P3 and P6 are functionally linked

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 $P_i \times P_j$.

$P(I_{mn} = 1)$: 训练



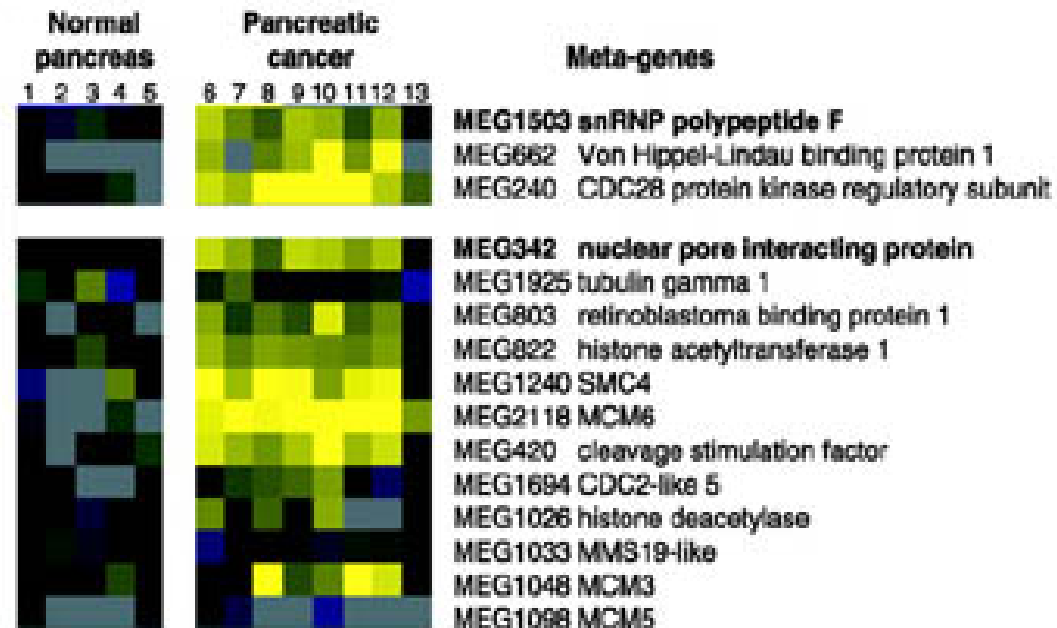
The $P(I_{mn} = 1)$ could be obtained from training non-redundant PPI data set. And the equation for calculating the two *probabilities* could be proposed as:

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

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

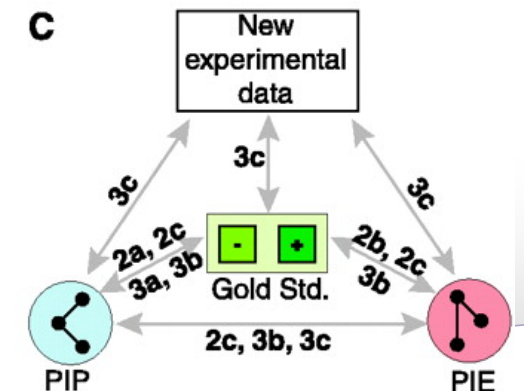
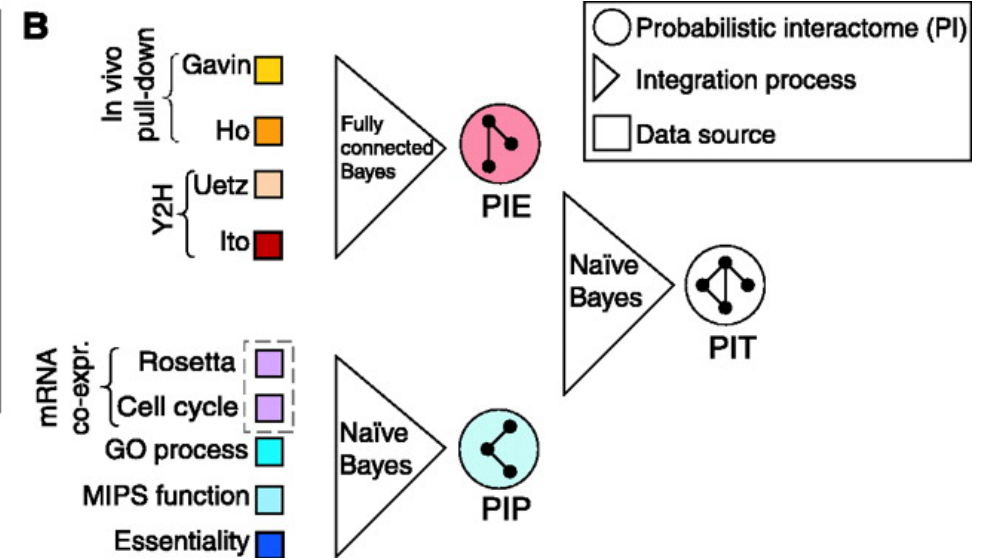
方法整合：贝叶斯算法



□ 2003年，Mark Gerstein

A

Data type	Dataset	# protein pairs	Used for ...
Experimental interaction data	In-vivo pull-down	Gavin et al.	Integration of experimental interaction data (PIE)
		Ho et al.	
	Yeast two-hybrid	Uetz et al.	
		Ito et al.	
Other genomic features	mRNA Expression	Rosetta compendium	De novo prediction (PIP)
		Cell cycle	
	Biological function	GO biological process	
		MIPS function	
	Essentiality		
Gold standards	Positives	Proteins in the same MIPS complex	Training & testing
	Negatives	Proteins separated by localization	



STRING: 方法的整合



[Home](#) · [Download](#) · [Help/Info](#)



STRING - Proteins and their Interactions

search by name

search by protein sequence

multiple names

multiple sequences

protein name: (examples: #1 #2 #3)

(STRING understands a variety of protein names and accessions; you can also try a [random entry](#))

organism:

interactors wanted:
☐ COGs ☒ Proteins

please enter your protein of interest...

What it does ...

STRING is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:

Genomic Context


High-throughput Experiments


(Conserved) Coexpression


Previous Knowledge


STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently contains 1,513,782 proteins in 373 species.

[More Info ...](#)

[Funding / Support ...](#)

[Acknowledgements ...](#)

STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) is developed at [EMBL](#), [SIB](#) and [Unizh](#).
STRING references: [von Mering et.al. 2007](#) / [2005](#) / [2003](#) / [Snel et.al. 2000](#).
Miscellaneous: [Access Statistics](#), [Robot Access Guide](#), [Medusa Network Viewer](#), [Supported Browsers](#).

What's New? This is version 7.1 of STRING. For the latest developments, check out our new [Blog](#) ...

New Sister Project: check out [STITCH](#) - built on STRING and serving networks of proteins with their associated small molecules!

New Partner: the Swiss Institute of Bioinformatics ([SIB](#)) has become a new STRING partner!

Previous Releases: Trying to reproduce an earlier finding? Confused? Refer to our [old releases](#).

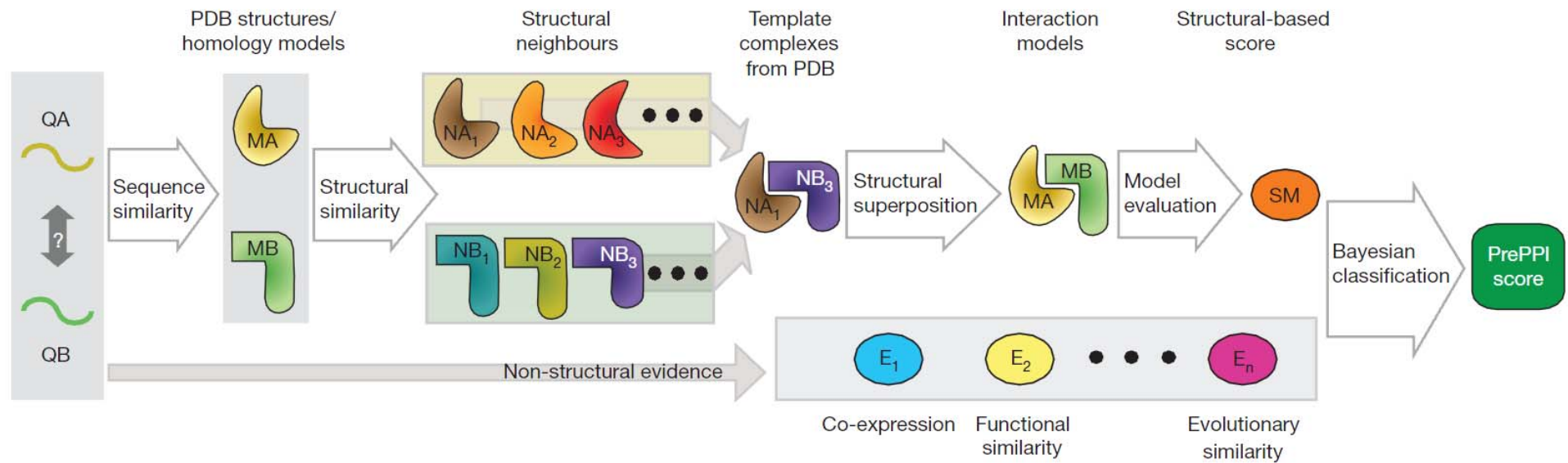
PrePPI



□ 2012年, 张强锋

✿ >30,000酵母PPI

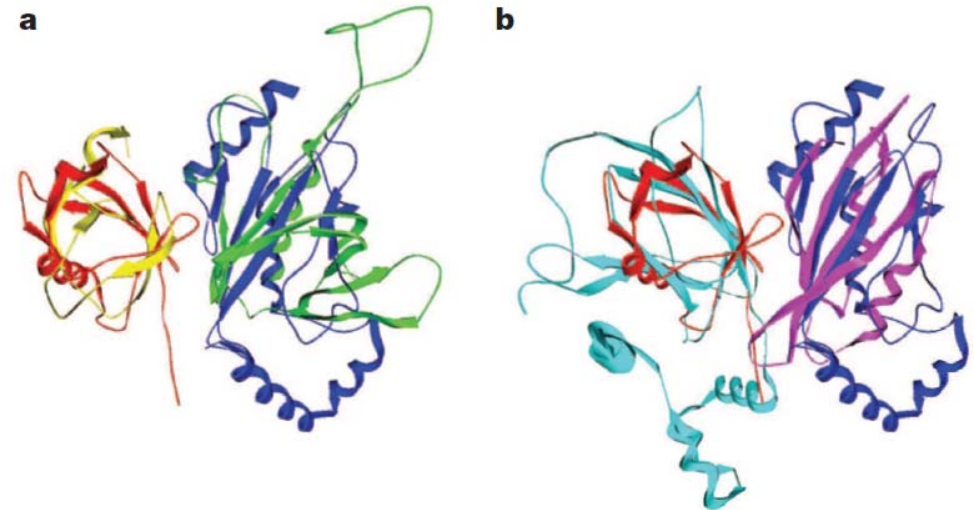
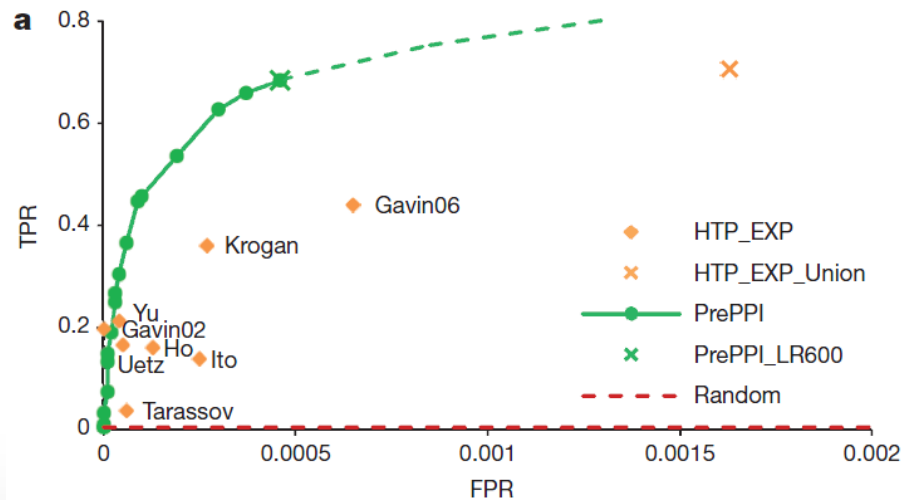
✿ >300,000人类PPI



PrePPI



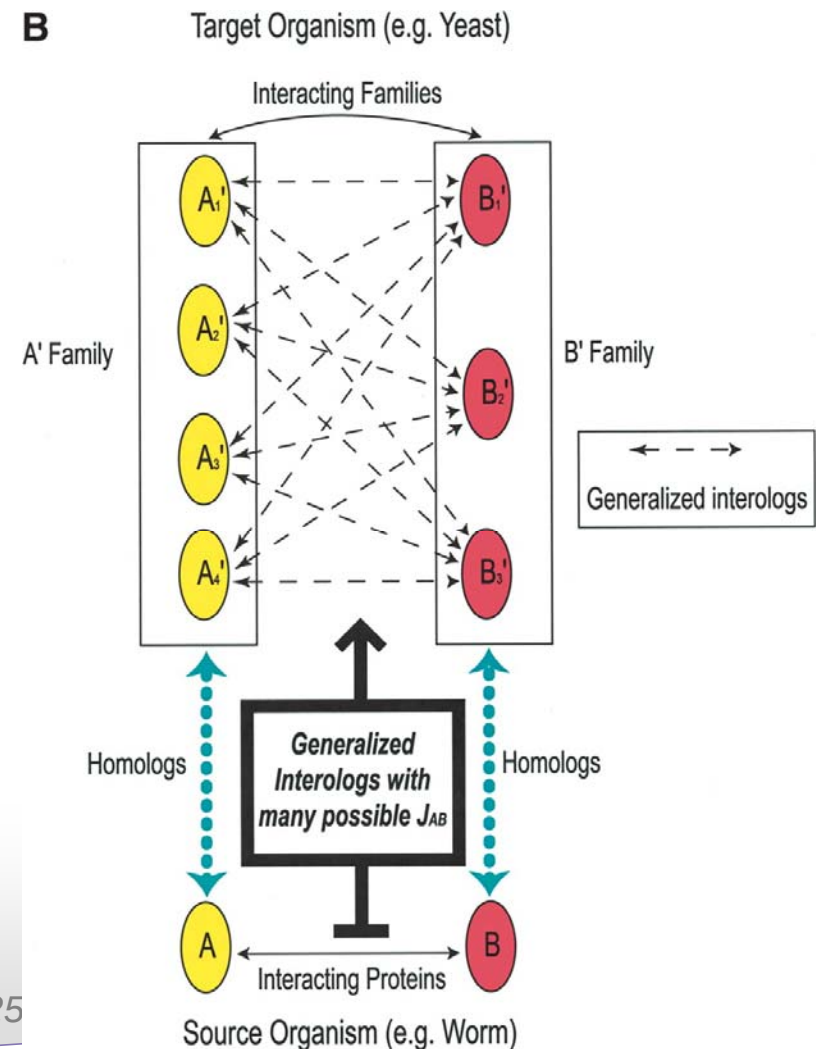
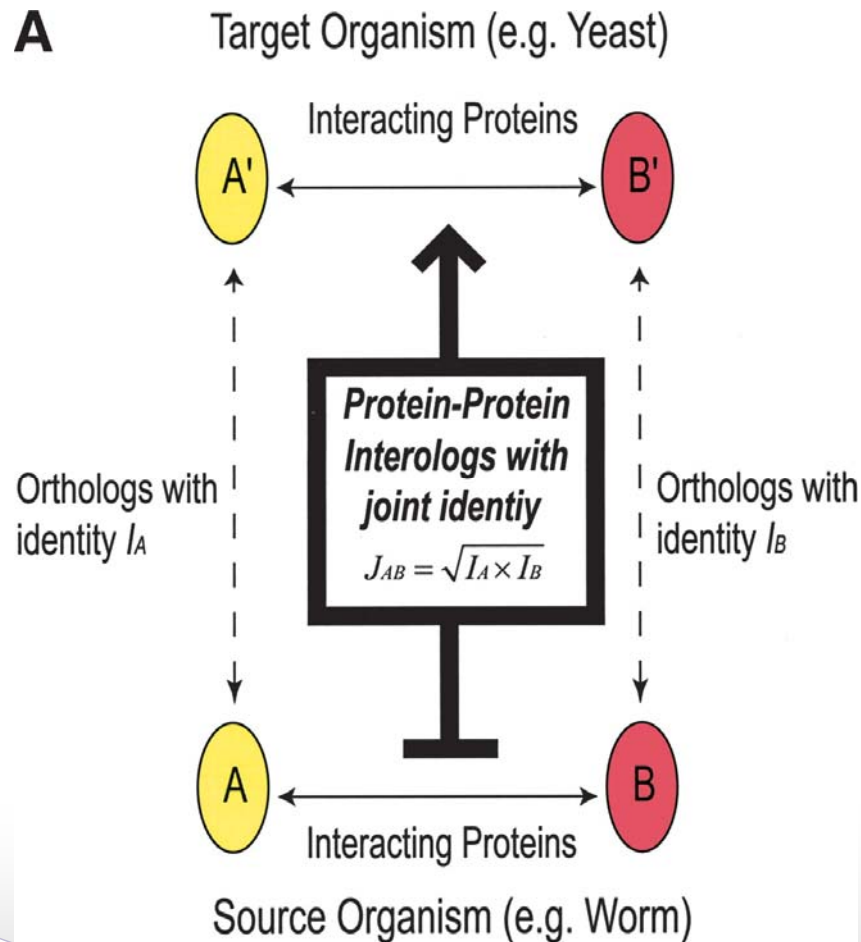
- 准确性比高通量实验高
- 根据三级结构能够预测蛋白质相互作用界面
 - ✿ PRKD1 and PRKCE
 - ✿ EEF1D and VHL



Interolog



□ 2004年, 于海源、Mark Gerstein



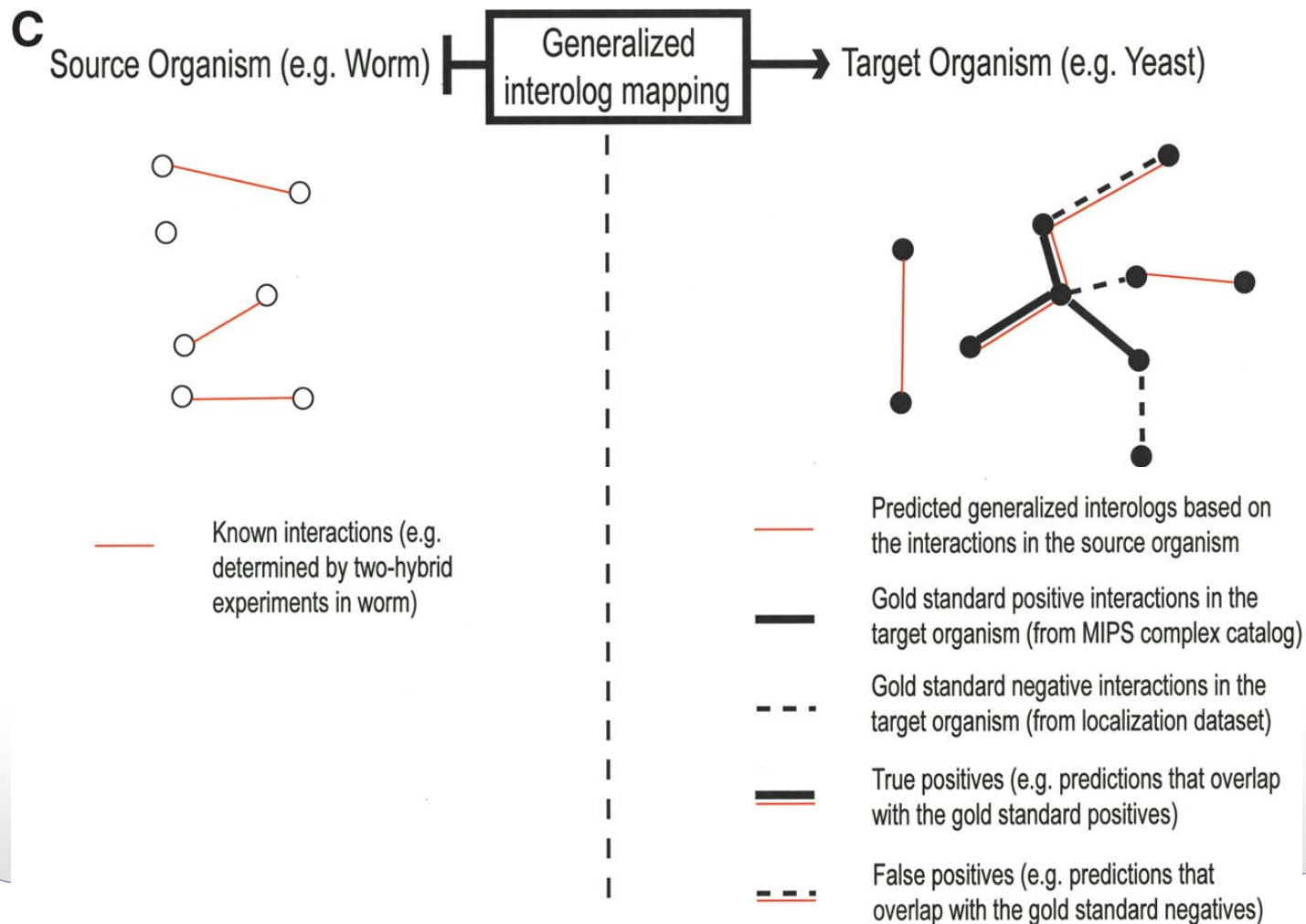
Haiyuan Yu et al. Genome Res. 2004;14:1107-1118

Bioinformatics, 2025

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™



InBio Map™

ABOUT

SEARCH

DOWNLOADS

DATA SOURCES


from data to biology



Search InBio Map™

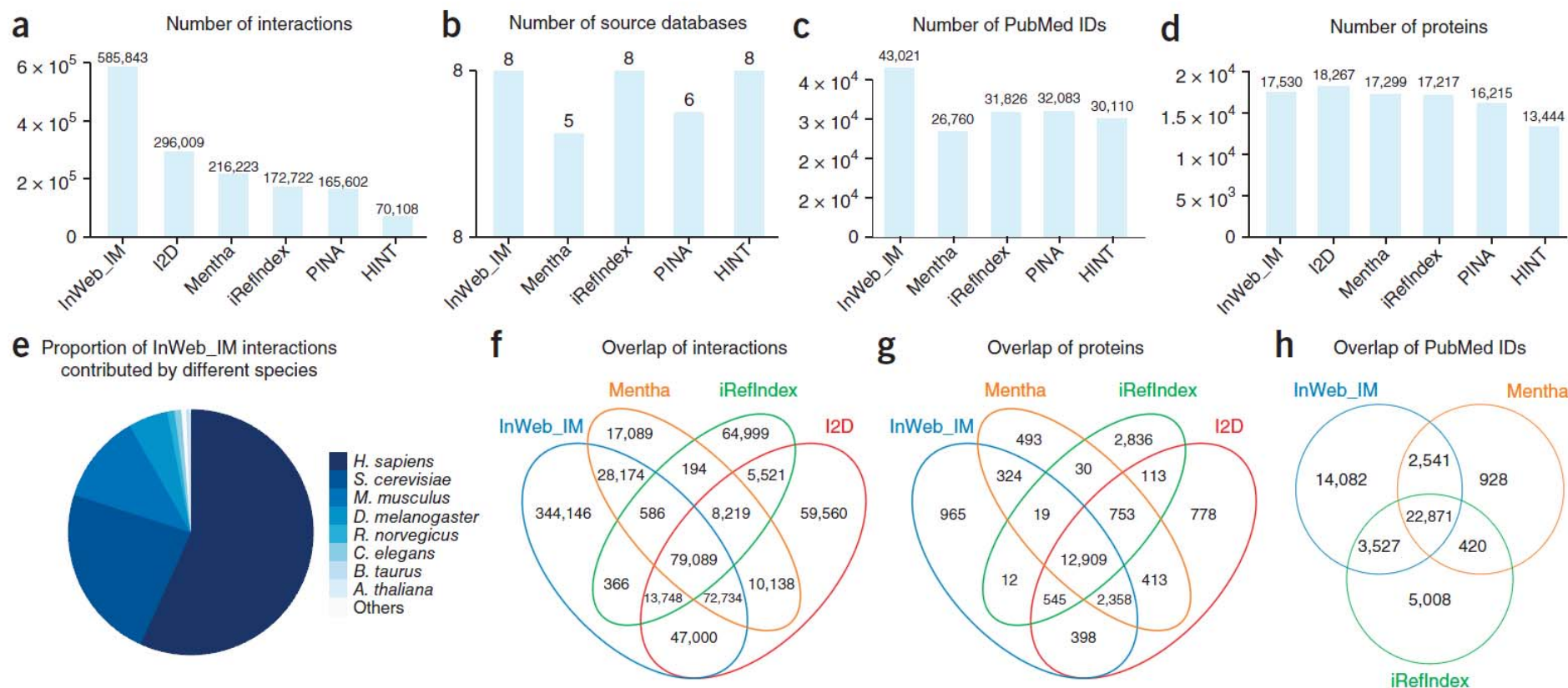
Examples: [INS_HUMAN](#) / [P01308](#) / [ENSG00000254647](#) / [INS](#) / [INS_HUMAN, IL6_HUMAN, INSR_HUMAN, PIR](#)

[Advanced search...](#)

InWeb_IM



实验 + Interolog





❑ 实验验证、Interolog、预测的PPI

Integrated Interactions Database

version 2016-03

tissue specific PPI networks across species

[Search By Proteins](#)[Search By PPIs](#)[Statistics](#)[About](#)[Contact](#)[Download](#)[FpClass](#)

1. Enter protein, gene, or dataset IDs:

Maximum number of IDs: 100
Accepted IDs: Symbol (e.g., TP53), UniProt (e.g., P04637), Entrez Gene (e.g., 7157)

2. Select species:

human
mouse
rat
fly
worm

3. Select tissues:

any
adipose tissue
adrenal gland
amygdala
bone

Some tissues are not available for some species

Find interaction partners supported by:

- ☒ Experimental evidence
- ☒ Orthologous interaction evidence
- ☒ Computational predictions
- ☐ Retrieve only interactions among query proteins
- ☐ Include interactions among partners of query proteins

Options for searching across species:

- ☒ Search using orthologs of your proteins
- ☐ Return only interactions conserved across all selected species

Options for searching across tissues:

- ☐ Return only interactions present in all selected tissues
- ☒ Required evidence: gene OR protein expression
- ☐ Required evidence: gene AND protein expression