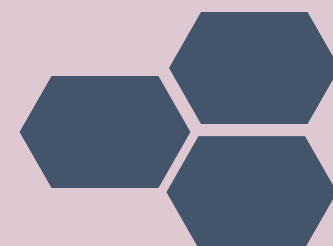


生物信息学

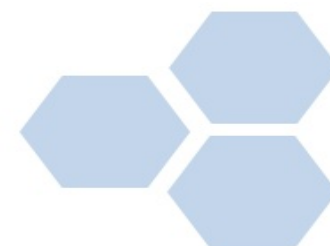


基因注释与功能分类





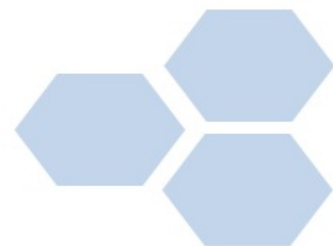
第一节 引言





背景

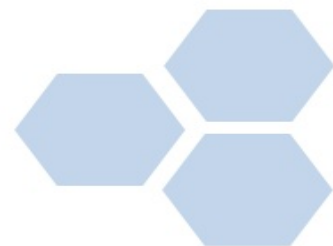
- 随着**后基因组（post-genomics）**时代研究的不断深入，基因组学的研究任务已由最开始的基因组序列识别，渐渐转移到在整体分子水平对功能进行研究。一个重要标志是**功能基因组学（functional genomics）**的不断发展。





任务

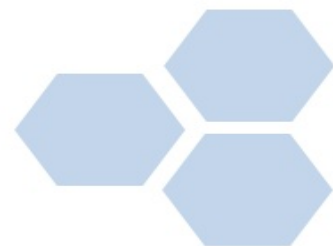
- 功能基因组学的主要任务之一是进行基因组**功能注释**（**genome annotation**），了解基因的功能，认识基因与疾病的关系，掌握基因的产物及其在生命活动中的作用等。





意义

- 快速有效的基因注释对进一步识别基因，研究基因的表达调控机制，研究基因在生物体代谢途径中的地位，分析基因、基因产物之间的相互作用关系，预测和发现蛋白质功能，揭示生命的起源和进化等具有重要的意义。

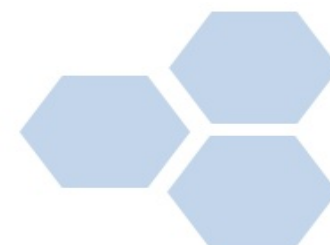




第二节

基因注释数据库

Gene Annotation Database



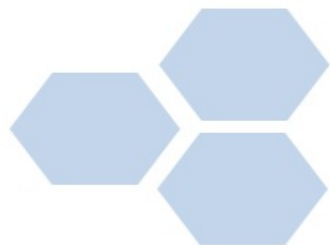
人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE



基因注释数据库产生的原因

一、研究人员已经掌握了大量的**全基因组数据**，同时关于基因、基因产物以及**生物学通路**的数据也越来越多，解释生物学实验的结果，尤其从基因组角度，需要系统的方法。

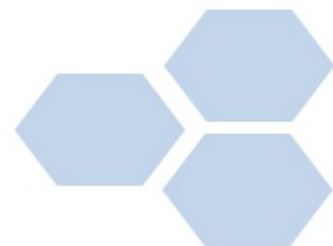
二、在基因组范围内描述蛋白质功能十分复杂，最好的工具就是计算机程序，提供结构化的标准的生物学模型，以便计算机程序进行分析，成为从整体水平系统研究基因及其产物的一项基本需求。





一、基因本体（gene ontology, GO）数据库

- 基因本体数据库是GO组织（Gene Ontology Consortium）在2000年构建的一个结构化的标准生物学模型，旨在建立基因及其产物知识的标准词汇体系，涵盖了基因的细胞组分（cellular component）、分子功能（molecular function）、生物学过程（biological process）。





GO数据库主页

Search GO data

terms and gene products

Search

Enrichment analysis (beta)

Your gene IDs here...

biological process

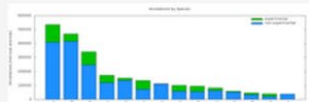
H. sapiens

Submit

Advanced options

Powered by PANTHER

Statistics



Gene Ontology Consortium



What is the Gene Ontology?

- [An introduction to the Gene Ontology](#)
- [What are annotations?](#)
- [Ten quick tips for using the Gene Ontology](#) **Important**
- [Gene Ontology tools](#)

Search

Q



Highlighted GO term

[Representing "phases" in GO biological process](#)

The GOC has recently introduced a new term [biological phase \(GO:0044848\)](#), as a direct subclass of biological process. This class represents a distinct period or stage during which biological processes can occur. [more](#)

Random FAQs

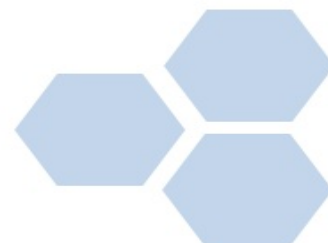
- [Which biological domains are supported by GO?](#)
- [I have a question about gene or protein nomenclature](#)
- [Who makes up the GO Consortium?](#)

[View all FAQs](#)

On the web



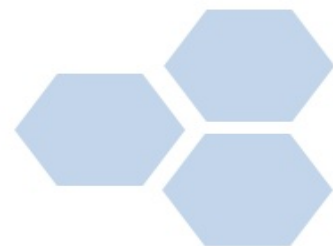
人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE





GO数据库收录的基因组数据列表

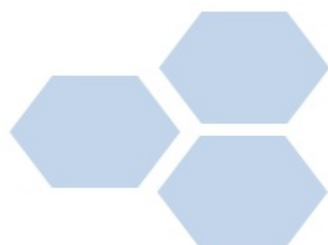
- GO数据库最初收录的基因信息来源于3个模式生物数据库：**果蝇、酵母和小鼠**，随后相继收录了更多数据，其中包括国际上主要的植物，动物和微生物基因组数据库。
- GO术语在多个合作数据库中的统一使用，促进了各类数据库对基因描述的一致性。





GO数据库收录的基因组数据列表

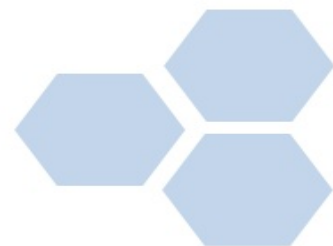
机构简称	收录的基因组数据	网站
BBOP	果蝇	http://www.berkeleybop.org
BHF-UCL	心血管基因	http://www.cardiovasculargeneontology.com
dictyBase	粘菌盘基网柄菌	http://dictybase.org
EcoliWiki	大肠杆菌	http://ecoliwiki.net
FlyBase	果蝇	http://flybase.bio.indiana.edu
GeneDB	裂殖酵母 恶性疟原虫 硕大利什曼原虫 布氏锥虫	http://www.genedb.org
GOA	UniProt 和 InterPro 注释	http://www.ebi.ac.uk/GOA
Gramene	农作物基因数据库	http://www.gramene.org
MGD and GXD	小家鼠	http://www.informatics.jax.org
RGD	褐家鼠	http://rgd.mcw.edu
Reactome	生物过程知识库	http://www.genomeknowledge.org
SGD	芽殖酵母 酿酒酵母	http://www.yeastgenome.org
TAIR	拟南芥	http://www.arabidopsis.org
IGS	基因组研究的工具和数据	http://www.igs.umaryland.edu
JCVI	若干种细菌基因组数据库	http://www.jcvi.org
WormBase	线虫	http://www.wormbase.org
ZFIN	斑马鱼	http://zfin.org





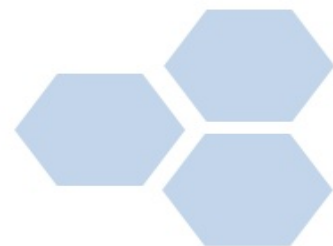
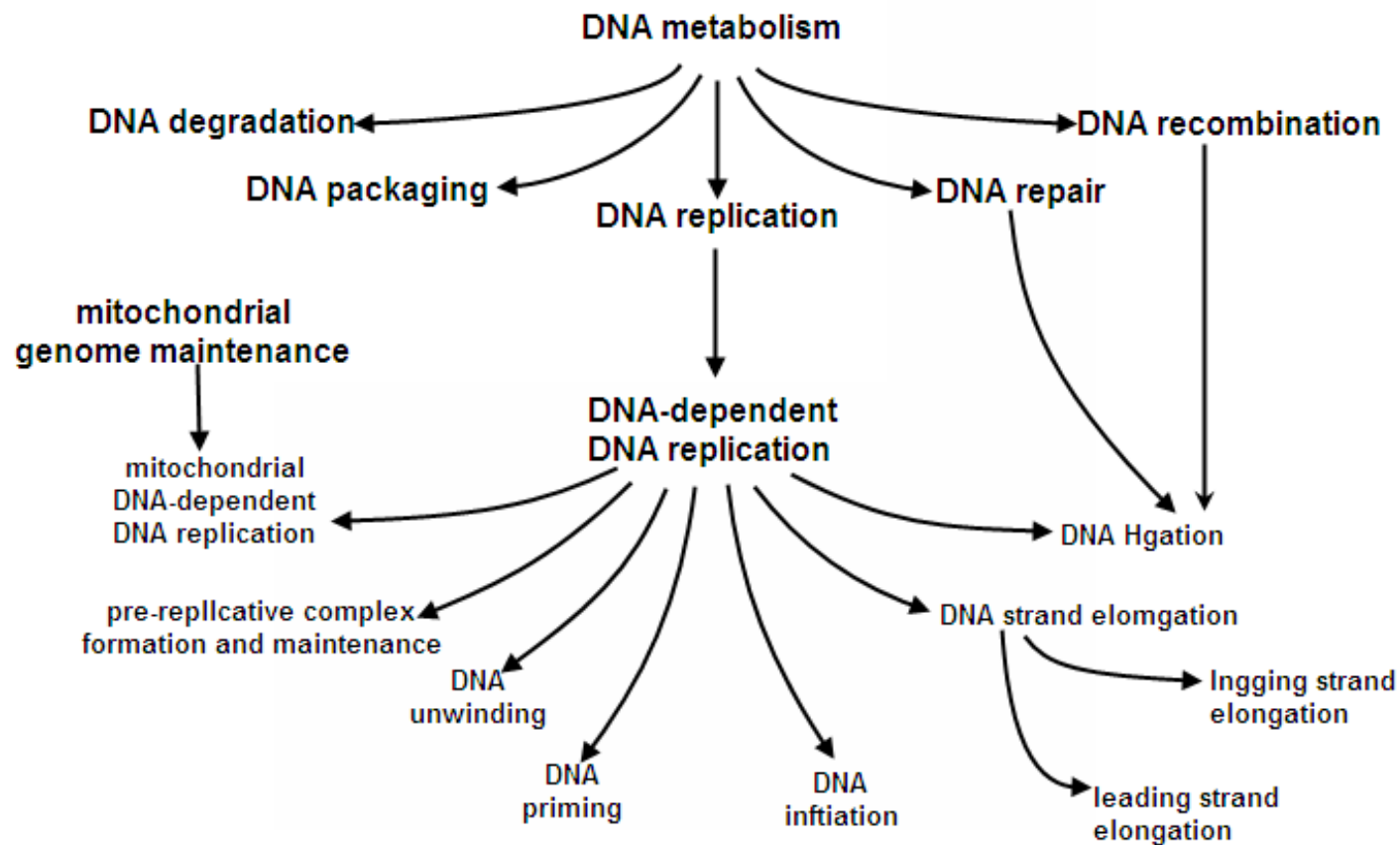
GO注释体系特点

- GO通过控制注释词汇的层次结构使得研究人员能够从不同层面查询和使用基因注释信息。
- 从整体上来看GO注释系统是一个有向无环图（directed acyclic graphs）,包含三个分支,即: **生物学过程（biological process）**， **分子功能（molecular function）** 和 **细胞组分（cellular component）**。
- 注释系统中每一个结点（node）都是基因或蛋白的一种描述,结点之间保持严格的关系,即“is a”或“part of”。





GO中生物学过程的DNA代谢部分功能类示意图

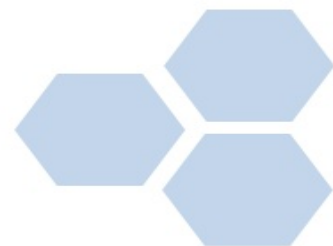




二、使用GO数据库

1. 用关键词检索GO数据库

- 检索GO数据库通常先进入**AmiGO 2.0**的首页。在GO数据库中，每条记录都有一个**数据标识号****GO:XXXXXX**和对应的术语。因此检索时需要知道待查基因的数字标识号或术语，将它们直接输入框中检索即可。如果检索的基因或蛋白质存在别名，可在检索框下勾选“**gene or proteins**”，并在检索框中输入别名检索；“**exact match**”表示是否完全匹配，可供选择。





AmiGO 2检索网页


AmiGO 2

More information on quick search [?](#)

Quick search

Search


Get Started with Grebe



Use the Grebe Search Wizard to **get started** in exploring the Gene Ontology data.

Go »


Advanced Search



Interactively **search** the Gene Ontology data for annotations, gene products, and terms using a powerful search syntax and filters.

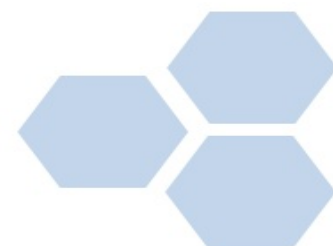
Search ▾

GOOSE



Use GOOSE to query a legacy GO database with **SQL** or edit one of the templates.

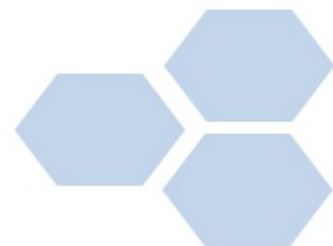
Go »





举例

- 这里以检索神经细胞分化因子6（NEUROD6）为例，选择“**Advanced Search**”下的“**Genes and gene products**”选项，在检索框中输入“**NEUROD6**”，运行后所得基因产物检索结果如图所示。





AmiGO 2检索结果示例

Free-text filtering

Your search is pinned to these filters

- + document_category: bioentity

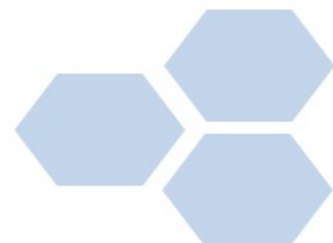
No current user filters.

- Source
- Type
- PANTHER family
- Taxon
- Direct annotation
- Inferred annotation

Found entities

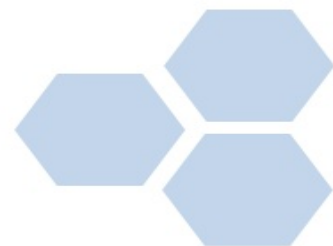
Total: 6; showing 1-6 Results count 10

<input type="checkbox"/>	Acc	Name	Taxon	PANTHER family	Type	Source	Direct annotation	Synonyms
<input type="checkbox"/>	Neurod6	neurogenic differentiation 6	Mus musculus	basic helix-loop-helix protein neurogenin-related pthr19290	protein	MGI	dentate gyrus development regulation of transcription, DNA-templated more...	Atoh2 bHLHa2 Math-2 Math2 Nex Nex1m
<input type="checkbox"/>	Neurod6	neuronal differentiation 6	Rattus norvegicus	basic helix-loop-helix protein neurogenin-related pthr19290	gene	RGD	dentate gyrus development RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription more...	
<input type="checkbox"/>	NEUROD6	Neurogenic differentiation factor	Gallus gallus	basic helix-loop-helix protein neurogenin-related pthr19290	protein	UniProtKB	dentate gyrus development RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription more...	E1C3F4_CHICK NEUROD6





- 检索得到的六个记录分别是不同物种中的神经源性分化因子6，点击物种为人类“**Homo sapiens**”的“**NEUROD6**”记录，得到结果如图所示，显示了该基因的基本信息，包括类型、物种、名称来源等信息。



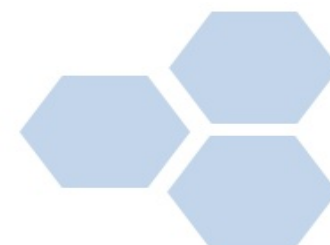


AmiGO 2基因描述示例1

Neurogenic differentiation factor 6

Gene Product Information

Symbol NEUROD6
Name(s) Neurogenic differentiation factor 6
Type protein
Taxon Homo sapiens
Synonyms NDF6_HUMAN
NEUROD6
ATOH2
BHLHA2
My051
Database UniProtKB, [Q96NK8](#)
Related [Link](#) to all direct and indirect **annotations** to NEUROD6.
[Link](#) to all direct and indirect **annotations download** (limited to first 10,000) for NEUROD6.



- 检索下方还显示了该基因产物的关联 (**gene product associations**) 图，要查看该基因的分子功能，可点击“**direct annotation**”中的记录查看，如点击“**protein dimerization activity**”的结果如图所示。

protein dimerization activity

Term Information

Accession GO:0046983

Name protein dimerization activity

Ontology molecular_function

Synonyms None

Definition The formation of a protein dimer, a macromolecular structure consists of two noncovalently associated identical or nonidentical subunits. Source: [ISBN:0198506732](#)

Comment None

History See term [history for GO:0046983](#) at QuickGO

Subset gosubset_prok

Community Unavailable

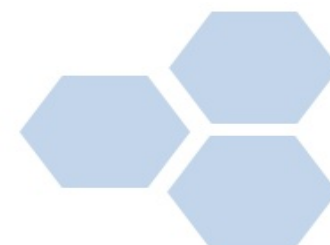
Related [Link](#) to all **genes and gene products** associated to protein dimerization activity.

[Link](#) to all direct and indirect **annotations** to protein dimerization activity.

[Link](#) to all direct and indirect **annotations download** (limited to first 10,000) for protein dimerization activity.

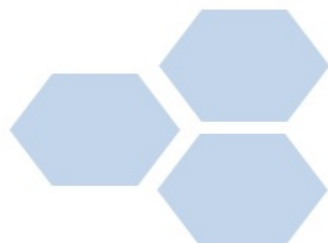
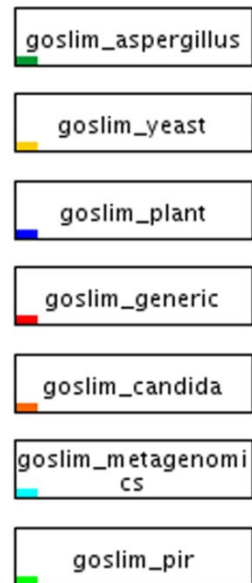
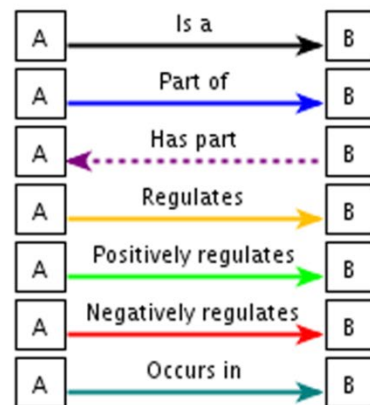
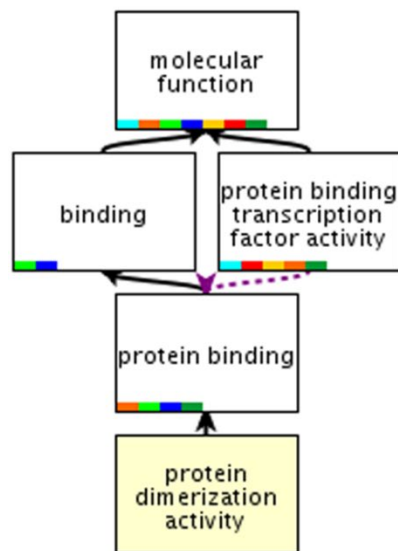


- 此外，还列举了该功能的详细注释，包括
“Associations”、“Graph Views”、“Inferred Tree View”、“Ancestors and Children”和
“Mappings”等。如点击可视化视图 “Graph Views”就可清晰地显示该分子功能构成的复杂功能网状结构，既有上下隶属关系，也存在平行关系。





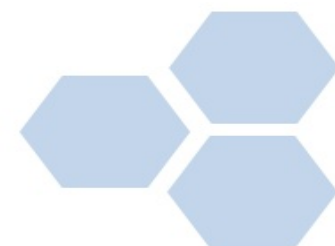
AmiGO 2查询 结果图形视图





2. 用序列检索GO数据库

- 在**AmiGO 1.8** 版本中，对于未知基因名的序列，还可以用序列直接检索GO 数据库。点击AmiGO 1.8首页上方的“**BLAST**”。
- 界面风格类似于其他数据库BLAST搜索的网页，在检索框中输入氨基酸或核酸序列，网页能自动识别并相应地做**BLASTP**或**BLASTX**和数据库中的序列比对。
- 这里以检索**RPIA**基因的序列为例，如图所示。





AmiGO 1.8 BLAST序列检索网页

BLAST Search

The sequence search is performed using either BLASTP or BLASTX (from the [WU-BLAST](#) package), depending on the type of the input sequence.

BLAST Query

Enter your query [?](#)

Enter a UniProt accession **or** upload a text file of queries **or** paste in FASTA sequence(s)

UniProt accession:

Text file (maximum file size 500K): [浏览...](#)

FASTA sequence(s):

Sequences should be separated with an empty line.

```
CCCTGCAAGGAGCAGAGTGTGTTACCTTGAGTCTCCAGCCCCAGCCAA
GGTGGACGTACCTCTCCAGGAGCCCTTTGCCCTTAATGTATCTCTGCTGGA
CAACTTGTGGTGGGGGTGGGGGGAAGAGTGGGAGGGGGAGTTAAATCCA
GTCTTATGAAGTATTGTTATTAAATGTCTTTTTAAAAAGAGAAATATAAA
CATATATTTTACTATTAAATATTCACTTTTTTAAATGAAGTAGAACTT
GAGTTCATGTTTTATATGAAATATTTACCAAAAAAAAAAAATGAGGTAAA
CTGTATTTAAACCTTTGACTTGAGTCTGCTGGTAAAGCTTCTGAATATT
GAGTTTGTGAGAAATAAAAATCAAACTTCITTAAGCTGATAAGTGAG
GGGCCCAACAGAGTGATCTCCTGATGCTTACTGGAACTTTGTTTACT
TGTCTGCTACCTCTGATTTGTTTTAGTTAGTTTTTATTGTGAGCACAC
ATAGTACCTAGTTACATCTTAAGATCAGTTTTATAAACTGTGGAGTGA
GCGGTATGTTATGGAATGACTTGGAAATGTAAGCTGTAGGGAGAAAATGT
TGTACACTTTTGTAAAGATCTGGGGGTTTCTTCATATTCTGCTGTTGG
AAGCAGTTGACCAAGAAATGCTTGCCAGTACTGCCAAGCACTGCTGTGAA
ATGTGAAGTACTTTGTTTTTTTAAATGATTTTTCTTTTGTATTATA
ATATTTTTCTCTGTTCTTTTGTATTACTTGCAATGGTTTGGCTCAGAG
TCCTTACCTCTTTATATTGTTTGCAGSTTTAAATAAACAGTGTGGTGCC
ATTTTG
```

Maximum number of sequences: 100

Maximum total length of sequence: 3,000,000 residues

[提交查询内容](#)

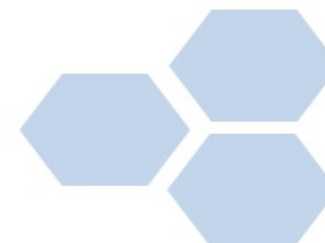
BLAST settings [?](#)

Expect threshold

Maximum number of alignments

BLAST filter: ☒ On ☐ Off

[提交查询内容](#)



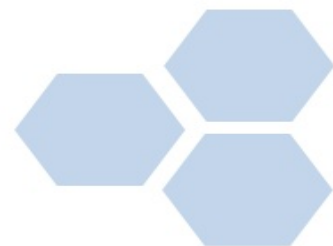
人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE



三、京都基因与基因组百科全书

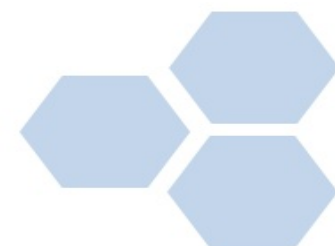
1. 简介

- 京都基因与基因组百科全书 (Kyoto encyclopedia of genes and genomes, KEGG) 是系统分析基因功能、基因组信息的数据库，它整合了基因组学、生物化学以及系统功能组学的信息，有助于研究者把基因及表达信息作为一个整体网络进行研究。





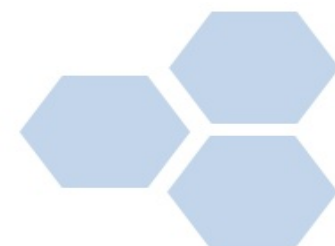
- **KEGG**提供的整合代谢途径查询十分出色，包括碳水化合物、核苷酸、氨基酸等代谢及有机物的生物降解，不仅提供了所有可能的代谢途径，还对催化各步反应的酶进行了全面的注解，包含其氨基酸序列、到**PDB**数据库的链接等。此外，**KEGG**还提供基于**Java**的图形工具访问基因组图谱、比较基因组图谱和操作表达图谱，以及其他序列比较、图形比较和通路计算的工具。因此，**KEGG**数据库是进行生物体内代谢分析、代谢网络分析等研究的强有力工具之一。





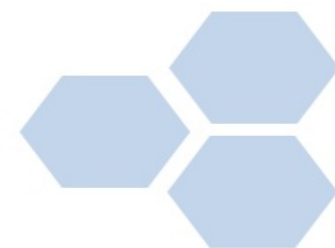
KEGG存储内容

- KEGG目前共包含了19个子数据库，它们被分类成系统信息、基因组信息和化学信息三个类别。



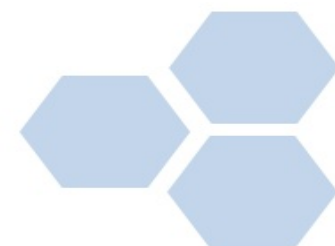


- 基因组信息存储在**GENES**数据库里，包括全部完整的基因组序列和部分测序的基因组序列，并伴有实时更新的基因相关功能的注释。
- **KEGG**中化学信息的6个数据库被称为**KEGG LIGAND**数据库，包含化学物质、酶分子、酶化反应等信息。**KEGG BRITE**数据库是一个包含多个生物学对象的基于功能进行等级划分的本体论数据库，它包括分子、细胞、物种、疾病、药物、以及它们之间的关系。





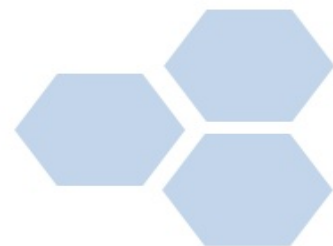
- 一些小的通路模块被存储在MODULE数据库中，该数据库还存储了其他的一些相关功能的模块以及化合物信息。
- **KEGG DRUG**数据库存储了目前在日本所有非处方药和美国的大部分处方药品。
- **KEGG DISEASE**是一个存储疾病基因、通路、药物、以及疾病诊断标记等信息的新型数据库。





KEGG数据库的注释与检索

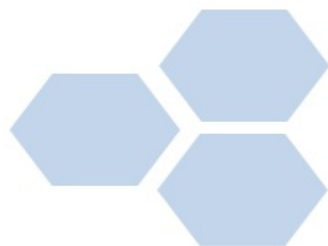
- **KEGG**通常被看作是生物系统的计算机表示，它囊括了生物系统中的各个对象与对象之间的关系。在分子层面、细胞层面、组织层面都可以对数据库进行检索。每个数据库中的检索条目按照一定规律被赋予一个检索号，也就是**ID**。表中列出了**KEGG**的13个核心数据库的检索号。





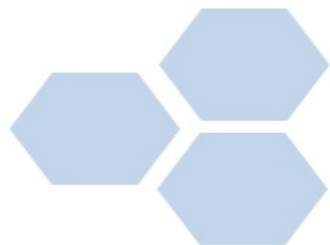
KEGG的13个核心数据库的检索号

Release	Database	Object Identifier
1995	KEGG PATHWAY	map number
	KEGG GENES	locus_tag / GeneID
	KEGG ENZYME	EC number
	KEGG COMPOUND	C number
2000	KEGG GENOME	organism code / T number
2001	KEGG REACTION	R number
2002	KEGG ORTHOLOGY	K number
2003	KEGG GLYCAN	G number
2004	KEGG RPAIR	RP number
2005	KEGG BRITE	br number
	KEGG DRUG	D number
2007	KEGG MODULE	M number
2008	KEGG DISEASE	H number
2009	KEGG PLANT	
Future releases	KEGG MEDICUS	Integrate KEGG DISEASE, KEGG DRUG, and various aspects of human body systems





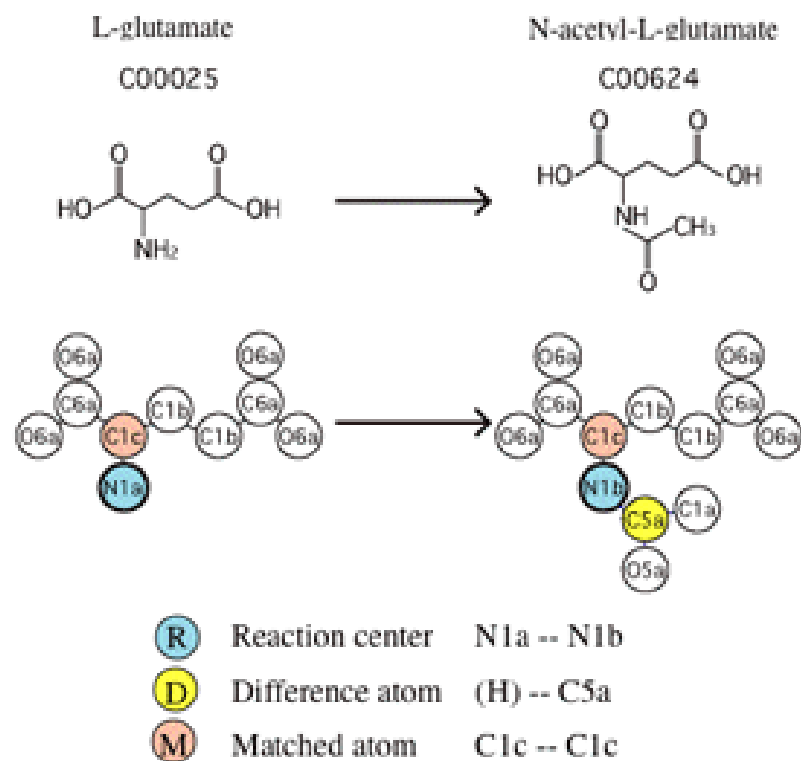
- 另外一种化学注释的方法是以小分子化学结构的生物学意义为特征来实现的。
- 在KEGG数据库中，酶与酶之间的反应信息以及相关的化学结构信息分别存储在KEGG REACTION数据库和KEGG REPAIR数据库中。
- 每个化合物的化学结构都被转化为RDM（atom type changes at R:reaction center D: difference atom M: matched atom）模式。



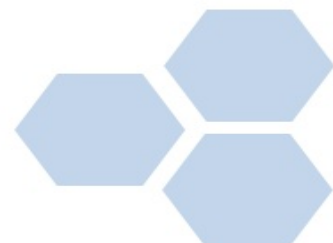


KEGG 数据库存储的RDM模式

RDM Pattern



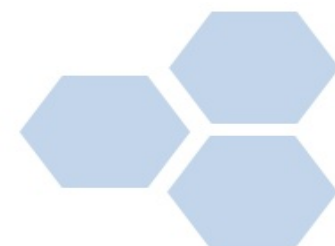
(Example) RDM pattern for A04458





KEGG数据库的注释与检索

- 下面以人类编码葡萄糖磷酸变位酶的基因“**PGM1**”为例：首先进入KEGG首页，在首页顶端的输入框中输入人类葡萄糖磷酸变位酶基因名称“**PGM1**”





KEGG查询首页



KEGG



Search

[Help](#)

[» Japanese](#)

KEGG Home

[Release notes](#)
[Current statistics](#)
[Plea from KEGG](#)

KEGG Database

[KEGG overview](#)
[Searching KEGG](#)
[KEGG mapping](#)
[Color codes](#)

KEGG Objects

[Pathway maps](#)
[Brite hierarchies](#)

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (See [Release notes](#) for new and updated features).

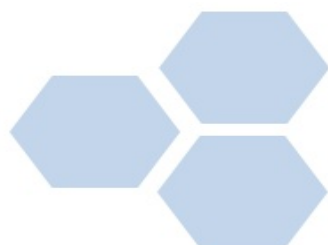
Please see: [Renewed plea to support KEGG](#)

New service

BlastKOALA for genome/metagenome annotation is now available. [more ...](#)

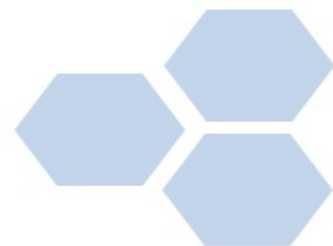


人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE





- 点击搜索按钮“**GO**”进入查询结果页面，该页面会列出针对基因“**PGM1**”在KEGG数据库中的搜索结果，除人类外，包含“**PGM1**”基因的物种条目也会被列出。





查询结果



Search

KEGG



for

PGM1

Go

Clear

Database: KEGG - Search term: PGM1

KEGG GENES

[hsa:5236](#)

PGM1, CDG1T, GSD14; phosphoglucomutase 1 (EC:5.4.2.2); K01835 phosphoglucomutase [EC:5.4.2.2]

[ptr:456908](#)

PGM1; phosphoglucomutase 1; K01835 phosphoglucomutase [EC:5.4.2.2]

[pps:100977295](#)

PGM1; phosphoglucomutase 1; K01835 phosphoglucomutase [EC:5.4.2.2]

[ggo:101128874](#)

PGM1; phosphoglucomutase-1 isoform 1; K01835 phosphoglucomutase [EC:5.4.2.2]

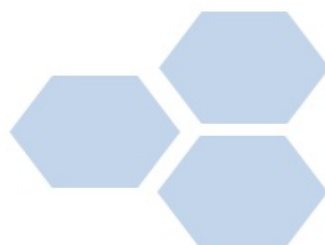
[pon:100438793](#)

PGM1; phosphoglucomutase 1; K01835 phosphoglucomutase [EC:5.4.2.2]

[... » display all](#)

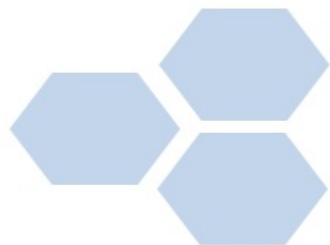


人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE





- 其中排在第一位的是人类基因“**PGM1**”的相关信息，点击该条目进入到详细信息页面。
- 该页面以表格的形式列出了该基因有关的详细信息，包括基因编号，基因的详细定义，所编码的酶的编号，基因所在通路，以及序列的编码信息。同时，在页面的右侧还提供了该基因在其他分子生物学数据库的链接，如OMIM、NCBI、GenBank等。





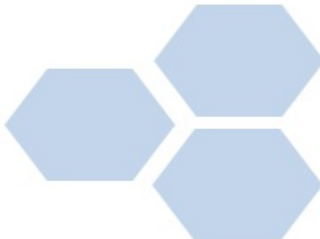
详细信息页面



Homo sapiens (human): 5236

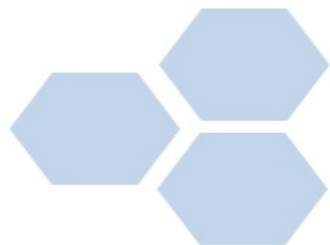
Help

Entry	5236	CDS	T01001
Gene name	PGM1, CDG1T, GSD14		
Definition	phosphoglucomutase 1 (EC:5.4.2.2)		
Orthology	K01835 phosphoglucomutase [EC:5.4.2.2]		
Organism	hsa Homo sapiens (human)		
Pathway	hsa00010 Glycolysis / Gluconeogenesis hsa00030 Pentose phosphate pathway hsa00052 Galactose metabolism hsa00230 Purine metabolism hsa00500 Starch and sucrose metabolism hsa00520 Amino sugar and nucleotide sugar metabolism hsa01100 Metabolic pathways		
Module	hsa_M00549 Nucleotide sugar biosynthesis, glucose => UDP-glucose		
Disease	H00069 Glycogen storage diseases (GSD)		
Brite	KEGG Orthology (KO) [BR: hsa00001] Metabolism Carbohydrate metabolism 00010 Glycolysis / Gluconeogenesis 5236 (PGM1) 00030 Pentose phosphate pathway 5236 (PGM1) 00052 Galactose metabolism 5236 (PGM1) 00500 Starch and sucrose metabolism 5236 (PGM1) 00520 Amino sugar and nucleotide sugar metabolism 5236 (PGM1) Nucleotide metabolism 00230 Purine metabolism 5236 (PGM1) Enzymes [BR: hsa01000] 5. Isomerases 5.4 Intramolecular transferases 5.4.2 Phosphotransferases (phosphomutases) 5.4.2.2 phosphoglucomutase (alpha-D-glucose-1,6-bisphosphate-dependent) 5236 (PGM1) BRITE hierarchy		





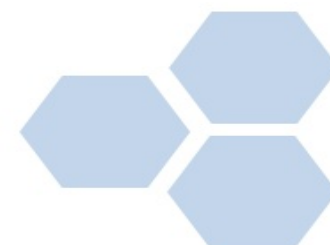
- 通过点击相应的链接，我们可以进入该基因相应信息的页面。在pathway这一栏中列出了该基因所在的生物学通路，点击编号为hsa00010（糖酵解/糖异生通路）的通路，进入到该通路的相应页面。该编号为hsa00010的通路页面以简单的几何图形显示出了糖酵解/糖异生相关生物过程。图中红色的方框即为基因“PGM1”所编码的酶，以此就可以通过该酶所在位置以及通路的拓扑结构来综合分析基因。







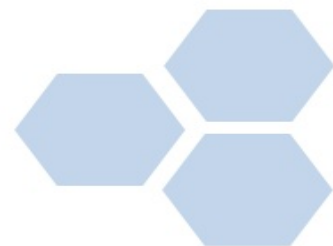
- 此外，可以通过页面顶部的下拉列表框来选择该通路在其他物种中的信息，也可以通过该列表框的选择来查看相关的基因、酶、反应、化合物等相关通路信息。





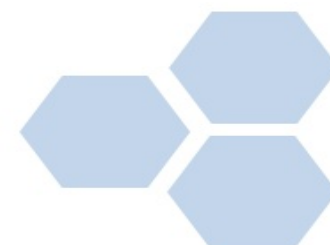
KEGG数据库的改进与更新

- **KEGG PATHWAY**还存储了一些人类疾病通路数据，这些疾病通路被分为六个子类：癌症、免疫系统疾病、神经退行性疾病、循环系统疾病、代谢障碍、传染病
- 循环系统疾病。



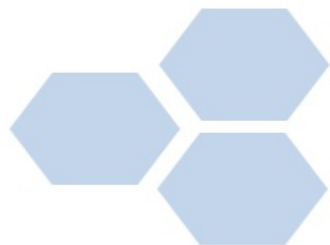


- **KEGG DRUG**数据库也在不断地完善，其中的药物数据几乎涵盖了日本的所有非处方药和美国的大部分处方药品。**DRUG** 是一个以存储结构为基础的数据库，每条记录都包含唯一的化学结构以及该药物的标准名称，以及药物的药效、靶点信息、类别信息等。



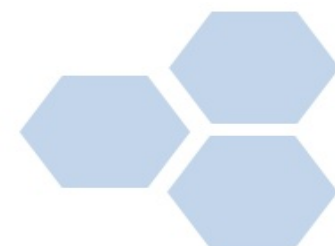


- 药物的靶点通过KEGG PATHWAY查询，药物的分类信息是KEGG BRITE数据库的一部分，通过药物的标准名称可以找到该药物的商品名，还可以找到药物销售的标签信息。此外，**DRUG**还包括一些天然的药物和中药的信息，有些药物被日本药典所收录。



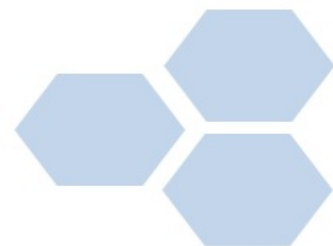


- 为了满足日益增长的科学研究需求，KEGG数据库在最近几年里不断扩充，新增加的50多个通路使KEGG PATHWAY数据库更加完善。这50多个新增的通路包括信号传导通路、细胞生物过程通路和人类疾病通路等。





- **KEGG**对通路数据新增了两个补充内容：第一个补充是一张全局通路图，这张全局通路图是通过手工拼接**KEGG**的**120**多个现存通路图生成的，存储为**SVG**文件。另一个补充内容是**KEGG MODULE**数据库，这是一个收集了通路模块以及其他一些功能单元的新型数据库，功能模块是在**KEGG**子通路中被定义为一些小的片段，通常包括几个连续的反应步骤、操纵子、调控单元，以及通过基因组比对得到的系统发生单元和分子的复合物等。





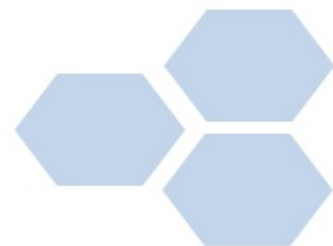
第三节

基因集功能富集分析

Gene Set Enrichment Analysis



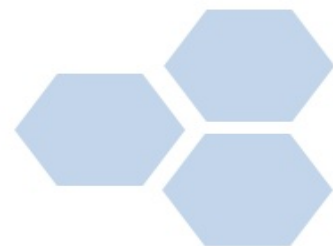
人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE





进行基因集功能富集分析的原因

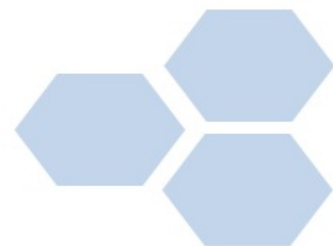
- 一组基因直接注释的结果是得到大量的功能结点。这些功能具有概念上的交叠现象，导致分析结果冗余，不利于进一步的精细分析，所以研究人员希望对得到的功能结点加以过滤和筛选，以便获得更有意义的功能信息。





一、富集分析算法

- 富集分析方法通常是分析一组基因在某个功能结点上是否出现过（**over-presentation**）。这个原理可以由单个基因的注释分析发展到大基因集合的成组分析。
- 由于分析的结论是基于一组相关的基因，而不是根据单个基因，所以富集分析方法增加了研究的可靠性，同时也能够识别出与生物现象最相关的生物过程。

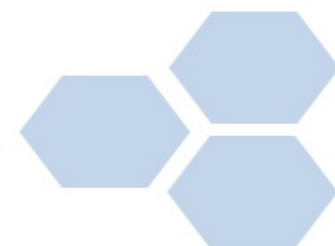




➤ 富集分析中常用的统计方法有累计超几何分布、Fisher精确检验等。

- 累计超几何分布：

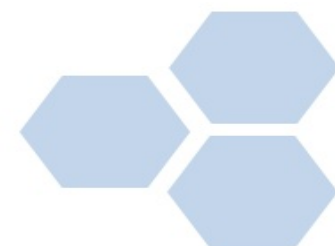
$$P(X > q) = 1 - \sum_{x=1}^q \frac{\binom{n}{x} \binom{N-n}{M-x}}{\binom{N}{M}}$$





- **Fisher精确检验:**

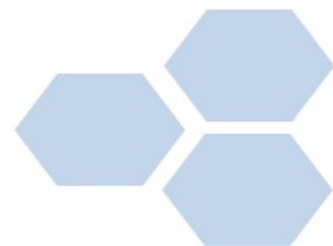
$$P = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}}$$





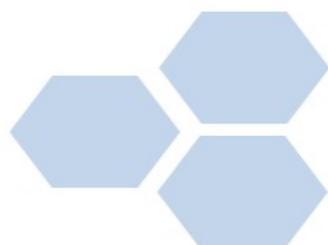
二、常用富集分析软件

- 基于不同的算法原理，可以将目前的常用富集分析工具分为三类：单一富集分析（**singular enrichment analysis**），基因集富集分析（**gene set enrichment analysis**），模块富集分析（**modular enrichment analysis**）。





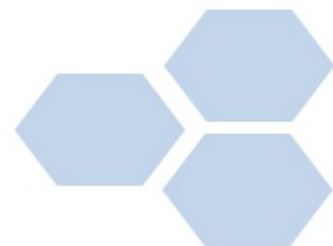
Enrichment tool name	Year of release	Key statistical method	Category
FunSpec	2002	Hypergeometric	Class I
Onto-express	2002	Fisher's exact; hypergeometric; binomial; chi-square	Class I
EASE	2003	Fisher's exact (modified as EASE score)	Class I
FatiGO/FatiWise/FatiGO+	2003	Fisher's exact	Class I
FuncAssociate	2003	Fisher's exact	Class I
GARBAN	2003	Hypergeometric	Class I
GeneMerge	2003	Hypergeometric	Class I
GoMiner	2003	Fisher's exact	Class I
MAPPFinder	2003	Z-score; hypergeometric	Class I
CLENCH	2004	Hypergeometric; chi-square; binomial	Class I
GO::TermFinder	2004	hypergeometric	Class I
GOAL	2004	Permutation	Class I
GOArray	2004	Hypergeometric; Z-score; permutation	Class I
GOSat	2004	Fisher's exact; chi-square	Class I
GoSurfer	2004	Chi-square	Class I






三、富集应用分析实例

- 这里以目前应用较为广泛的**DAVID**为例对基因集进行具体分析。**DAVID**是一个综合工具，不但提供基因富集分析，还提供基因间**ID**的转换、基因功能的分类等。





DAVID应用工具首页

**DAVID Bioinformatics Resources 6.7**
National Institute of Allergy and Infectious Diseases (NIAID), NIH

HomeStart AnalysisShortcut to DAVID ToolsTechnical CenterDownloads & APIsTerm of ServiceWhy DAVID?About Us

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering, BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and [more](#)

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. [More](#)

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. [More](#)

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. [More](#)

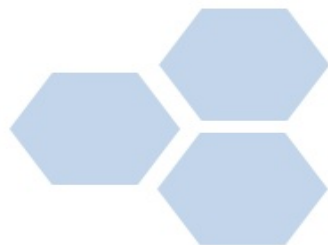
Recommending: A [paper](#) published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2014

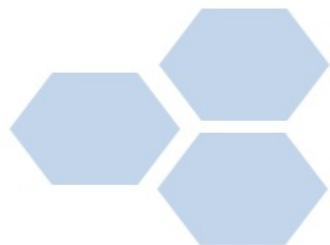
The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 is an [update to the sixth version](#) of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- ☑ Identify enriched biological themes, particularly GO terms
- ☑ Discover enriched functional-related gene groups
- ☑ Cluster redundant annotation terms
- ☑ Visualize genes on BioCarta & KEGG pathway maps





- 点击“**Functional Annotation**”后，第一步为提交基因集，选择基因标识名和基因集类型；第二步得到注释结果摘要，包括多种注释数据；然后选择感兴趣的注释内容得到富集分析结果。





DAVID富集分析注释结果摘要

Upload

List

Background

Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -

Homo sapiens(201)

Unknown(4)

Select Species

List Manager [Help](#)

top5% unique gene

Select List to:

Use

Rename

Remove

Combine

Show Gene List

[View Unmapped Ids](#)

Annotation Summary Results

Current Gene List: top5% unique gene

Current Background: Homo sapiens

199 DAVID IDs

Check Defaults ☒

☒ Disease (1 selected)

☒ Functional_Categories (3 selected)

☒ Gene_Ontology (3 selected)

☒ General_Annotations (0 selected)

☒ Literature (0 selected)

☒ Main_Accessions (0 selected)

☒ Pathways (3 selected)

☒ Protein_Domains (3 selected)

☒ Protein_Interactions (0 selected)

☒ Tissue_Expression (0 selected)

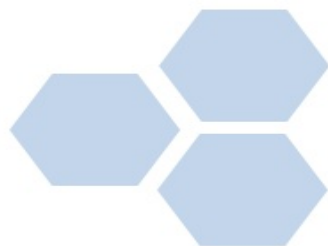
Red annotation categories denote DAVID defined defaults

Combined View for Selected Annotation

Functional Annotation Clustering

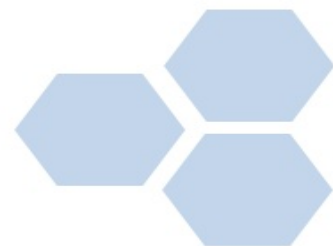
Functional Annotation Chart

Functional Annotation Table






- 这里以KEGG通路的富集分析为例。提交之后的结果如图，可以看到，对提交的基因集做富集分析，找到5个具有显著性的通路。这里的“**P-Value**”是通过**Fisher**精确检验得到的P值，“**Benjamini**”指的是本杰明假阳性率校正方法。





DAVID在KEGG上富集结果实例

**DAVID Bioinformatics Resources 6.7**
National Institute of Allergy and Infectious Diseases (NIAID), NIH



Functional Annotation Chart

[Help and Manual](#)

Current Gene List: top5% unique gene
Current Background: Homo sapiens
199 DAVID IDs

☐ Options

2 chart records [Download File](#)

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	KEGG_PATHWAY	Insulin signaling pathway	RT		5	2.5	4.0E-2	9.7E-1
<input type="checkbox"/>	KEGG_PATHWAY	Neuroactive ligand-receptor interaction	RT		6	3.0	9.8E-2	9.9E-1

from your list are not in the output.

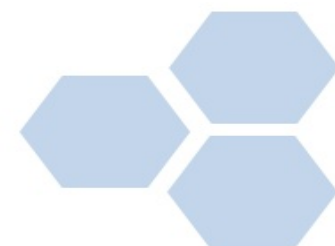




第四节

基因功能预测

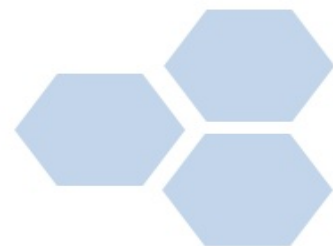
Gene Function Prediction





基因功能预测算法

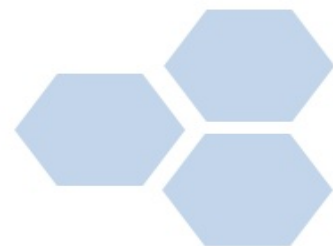
- 近来已经发展了很多基于GO数据库或KEGG数据库的方法，利用高通量的基因表达和蛋白质互作数据进行功能预测，其中一些新开发的方法试图整合多种数据类型，通过构建功能相关网络的方式预测基因功能。





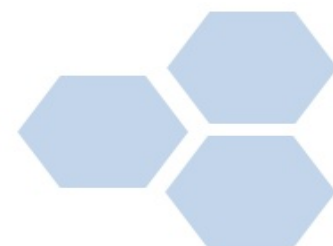
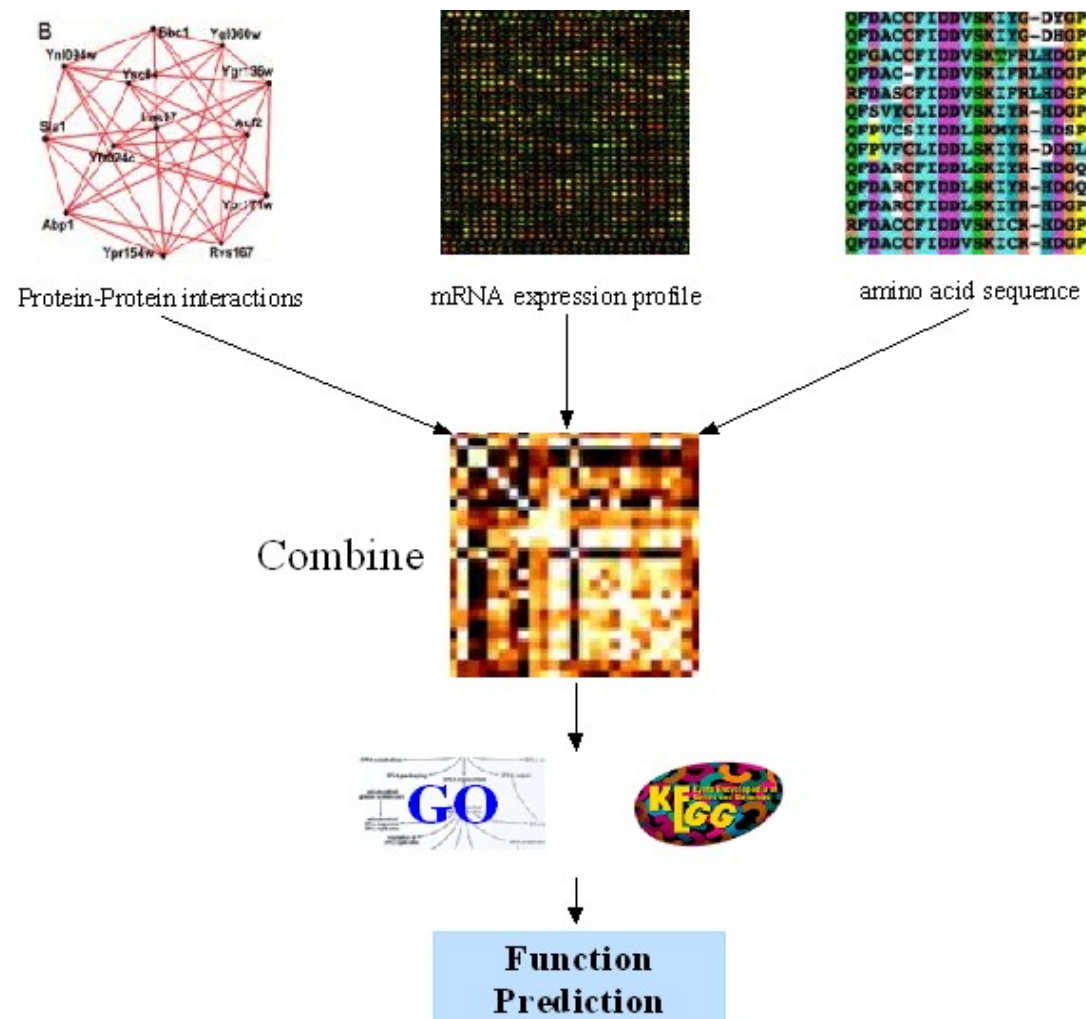
当前基于GO或KEGG的基因功能预测策略

- 首先，从总体上宏观地概括抽取信息，如不同样本间、不同时间点间全部差异基因；
- 其次，通过GO或KEGG分析，即从GO分类结果找到实验涉及的显著功能类别或将差异基因映射到通路中，根据基因在通路中的位置及表达水平的变化算出受影响显著的通路，从而预测未知的基因功能等。





整合蛋白质互作数据、表达谱和序列数据的功能预测

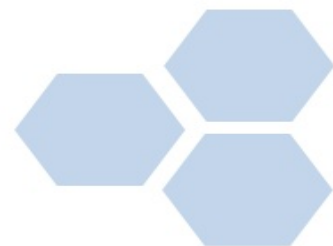




一、基于GO的基因功能预测

1. 对差异表达基因进行功能预测

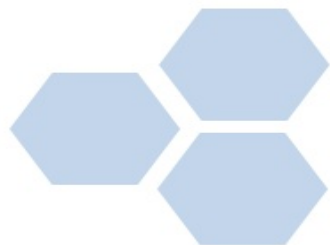
- 在基因芯片的数据分析中，研究者可以找出哪些差异表达基因属于一个共同的GO功能分支，并用统计学方法检验结果是否具有统计学意义，从而得出差异表达基因主要参与了哪些生物功能。





2. 蛋白质互作网络用于基因功能预测

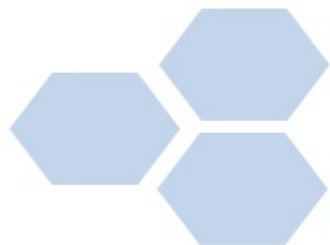
- 目前，利用相互作用网络进行功能注释主要有两种方法，即直接注释方法（**direct annotation schemes**）和基于模块的方法（**module assisted schemes**）。





3. 利用GO体系结构比较基因功能

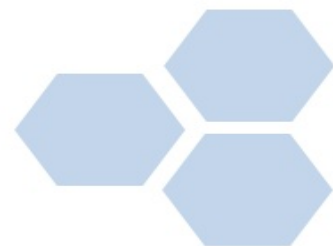
- 通常认为如果两个基因产物的功能相似，那么它们的表达也就相近，同时它们在GO中注解的结点就相似，所以只要能找出GO中结点对的相似度，就可以近似估计两基因表达的相似度，从而判断两基因产物的功能的相似度。





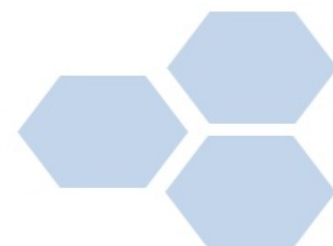
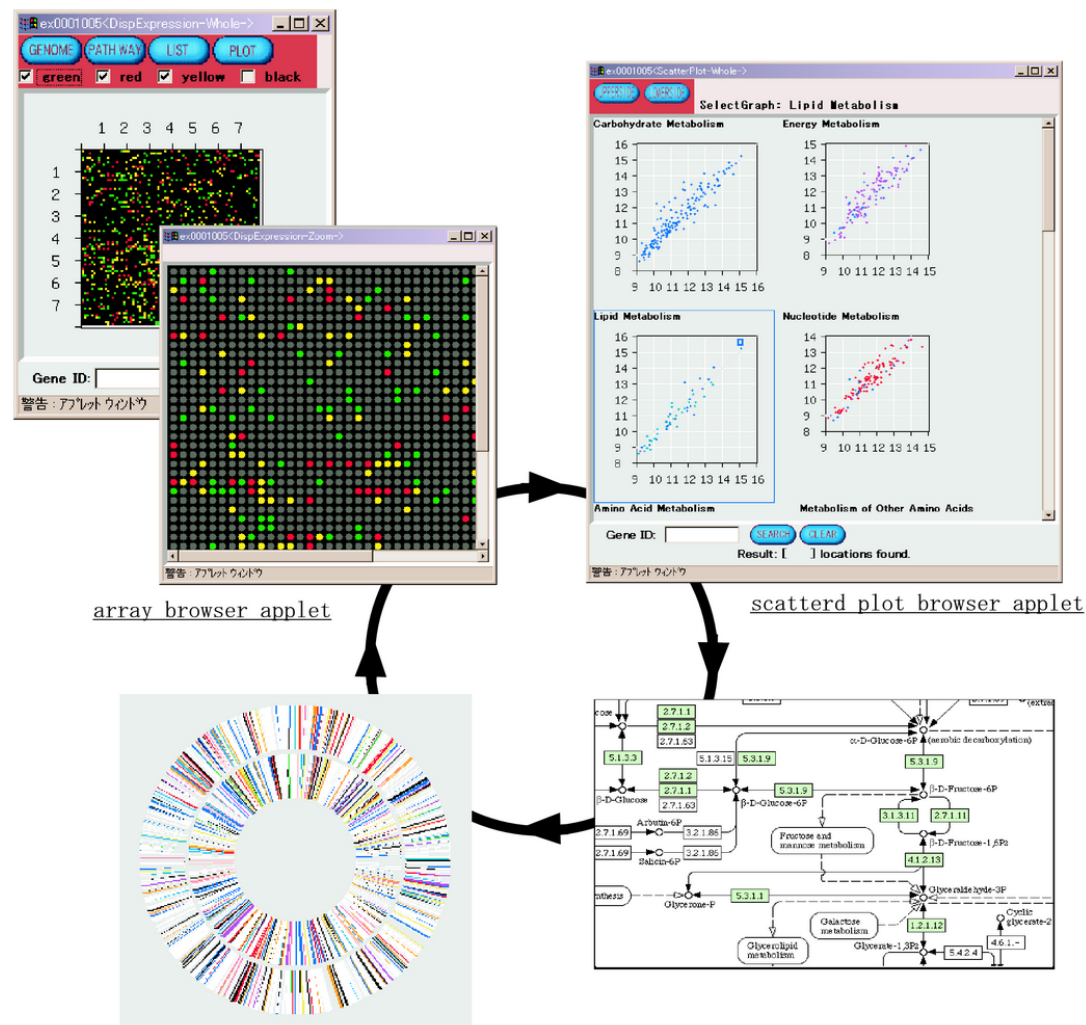
二、基于KEGG通路分析的基因功能预测

- 通路分析是现在经常被使用的芯片数据基因功能分析法。与GO分类法（应用单个基因的GO分类信息）不同，通路分析法利用的资源是许多已经研究清楚的基因之间的相互作用，即生物学通路。研究者可以把表达发生变化的基因集导入通路分析软件中，进而得到变化的基因都存在于哪些已知通路中，并通过统计学方法计算哪些通路与基因表达的变化最为相关。





通过表达谱数据进行通路定位

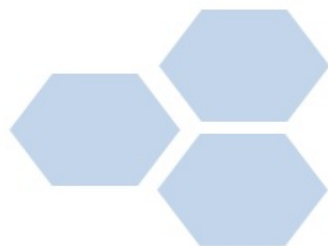




三、常用基因功能预测软件

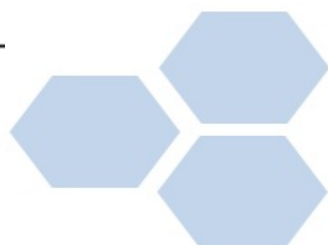
用GO分类法进行芯片功能分析的网络平台

Name	Internet Site
Onto-Tools	http://vortex.cs.wayne.edu/projects.htm
ROSETTA	http://rosetta.lcb.uu.se/general/
GOToolBox	http://burgundy.cmmt.ubc.ca/GOToolBox/
GOstat	http://gostat.wehi.edu.au/
GFINDER	http://www.medinfopoli.polimi.it/GFINDER/
FatiGO	http://www.fatigo.org/
EASE	http://david.abcc.ncifcrf.gov/ease/ease.jsp





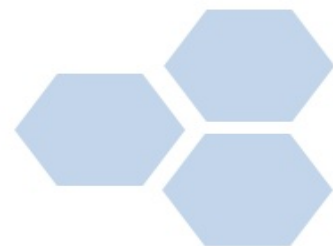
Name	Internet Site
GenMAPP	http://www.genmapp.org/
PathwayMiner	http://www.biorag.org/pathway.html
KOBAS	http://kobas.cbi.pku.edu.cn
GEPAT	http://gepat.bioapps.biozentrum.uni-wuerzburg.de/GEPAT/index.faces
VitaPad	http://bioinformatics.med.yale.edu/group
KEGGanim	http://biit.cs.ut.ee/kegganim/
WholePathwayScope	http://www.abcc.ncifcrf.gov/wps/wps_index.php
VisANT 3.0	http://visant.bu.edu/
Eu.Gene	http://www.ducciocavalieri.org/bio/Eugene.htm





举例

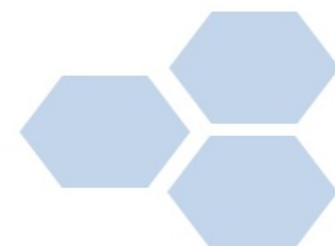
- 利用**Onto-Express**预测基因功能
- **Onto-Express**是Wayne State University开发的**Onto-Tools**软件包中的一个表达谱数据分析工具，利用**Gene Ontology**中的数据信息对基因的功能进行分析，
可以免费下载该软件。





1. 数据输入

- 下面通过提供的测试数据阐述Onto-Express的使用方法，该芯片的测试数据可在<http://www.ebi.ac.uk/~jane/TestData/>下载，输入数据为total和under.over，输入数据为文本格式，包含accession numbers, cluster identifiers 或 probe identifiers。进入Onto-Express的输入窗口，如图所示：





Onto-Express输入窗口

Onto-Express Input

Input File:

Reference File:

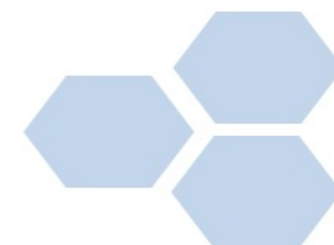
Reference Chip:

Organism:

Input Type: ☐ Accession ☐ Cluster ☐ Probe ID
☐ WormBase Accession ☐ LocusLink ☐ Gene Symbol

Search for: ☒ Biological Process ☒ Cellular Component
☒ Molecular Function ☒ Chromosome Information

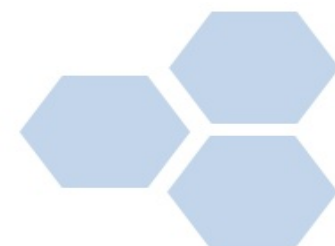
Distribution: Correction:





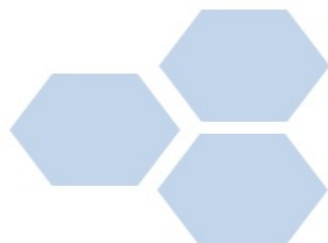
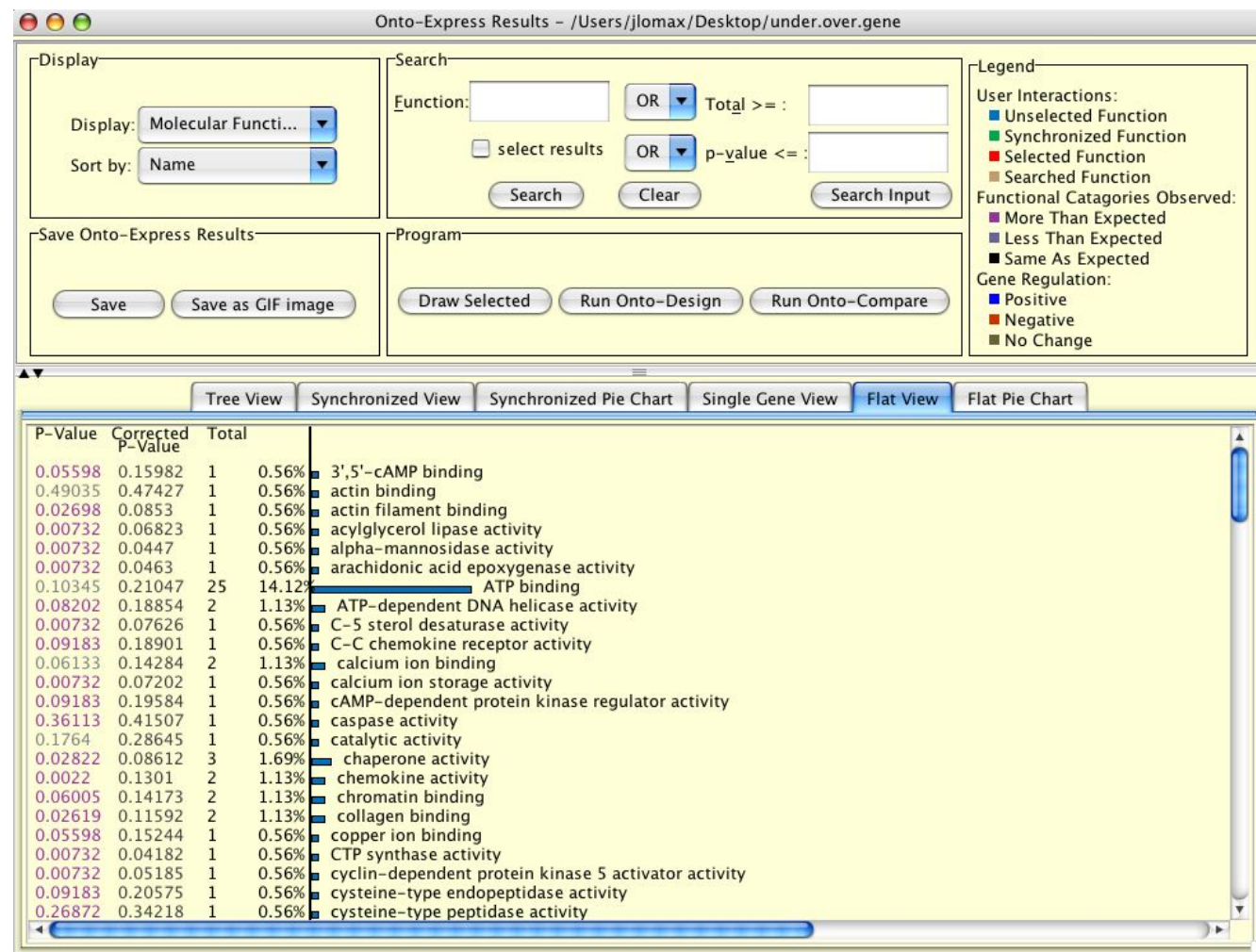
2. 结果页面

- 选择“Tree View”，将显示GO的树状图，可以单击收缩或展开显著term的信息。GO term上的黑体字是输入的上调或下调基因集合注释到该term上的数目。P值是该结点含有上调或下调基因的数目大于随机期望的概率。





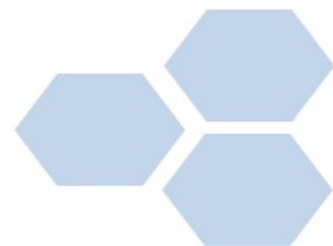
Onto-Express结果窗口





小 结

- 基因注释与功能分类是功能基因组学和计算系统生物学的重要基础。本章重点介绍了**Gene Ontology (GO)** 数据库 和 **Kyoto Encyclopedia of Genes and Genomes (KEGG)** 数据库。分别从基因功能注释和通路注释两个层面阐述功能注释与分类。





- 随着功能基因组学在人类复杂疾病研究中应用的逐步深入，基因功能注释的尺度也逐步从单基因注释发展到多基因注释和通路（或特定功能的基因集合）注释。基于GO和KEGG发展起来的David、GOEAST、GOSim、KEGGSpider、KEGGArray、PathwayMiner等软件从不同角度实现注释、富集分析和功能预测，方便临床医学工作人员对感兴趣的基因或基因组进行研究。

