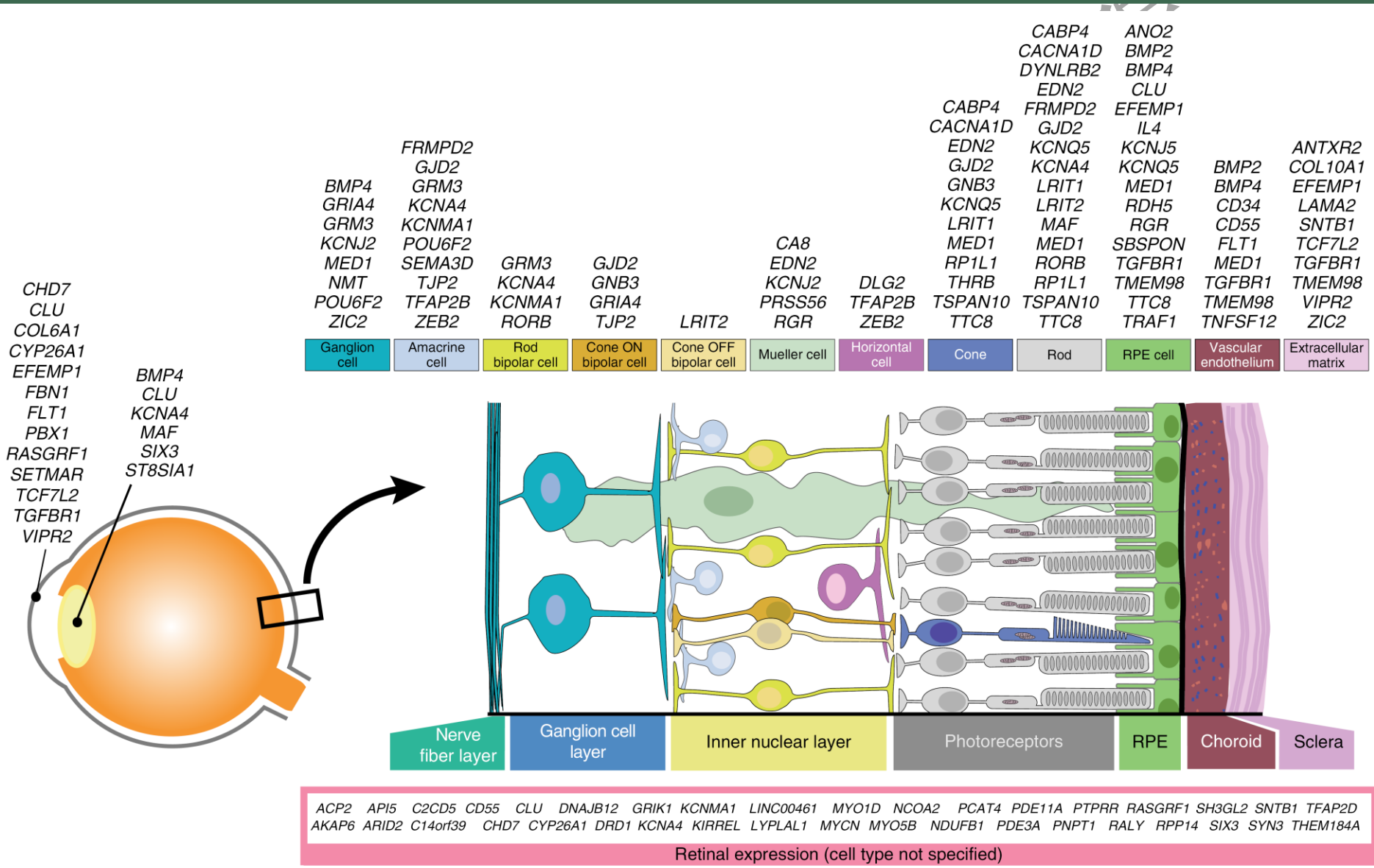
Chapter-07. 基因功能注释与富集分析



本章内容提要

- 7.1 引言：基因与基因组
- 7.2 什么是基因功能注释？
- 7.3 基因功能注释的方法
- 7.4 基因功能富集分析
- 7.5 基因富集分析常用方法
- 7.6 GSEA在R中的实现
- 7.7 总结与展望

第1节：概要 (Introduction)

随着后基因组(post-genomics)时代的来临，基因组学的研究重心开始从阐明所有遗传信息转移到在整体分子水平对功能进行研究。这种转变的一个重要标志是产生了功能基因组学。

功能基因组学的最主要任务之一就是进行基因组的功能注释，以了解基因的功能，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。

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2节：基因功能注释数据库(Database)

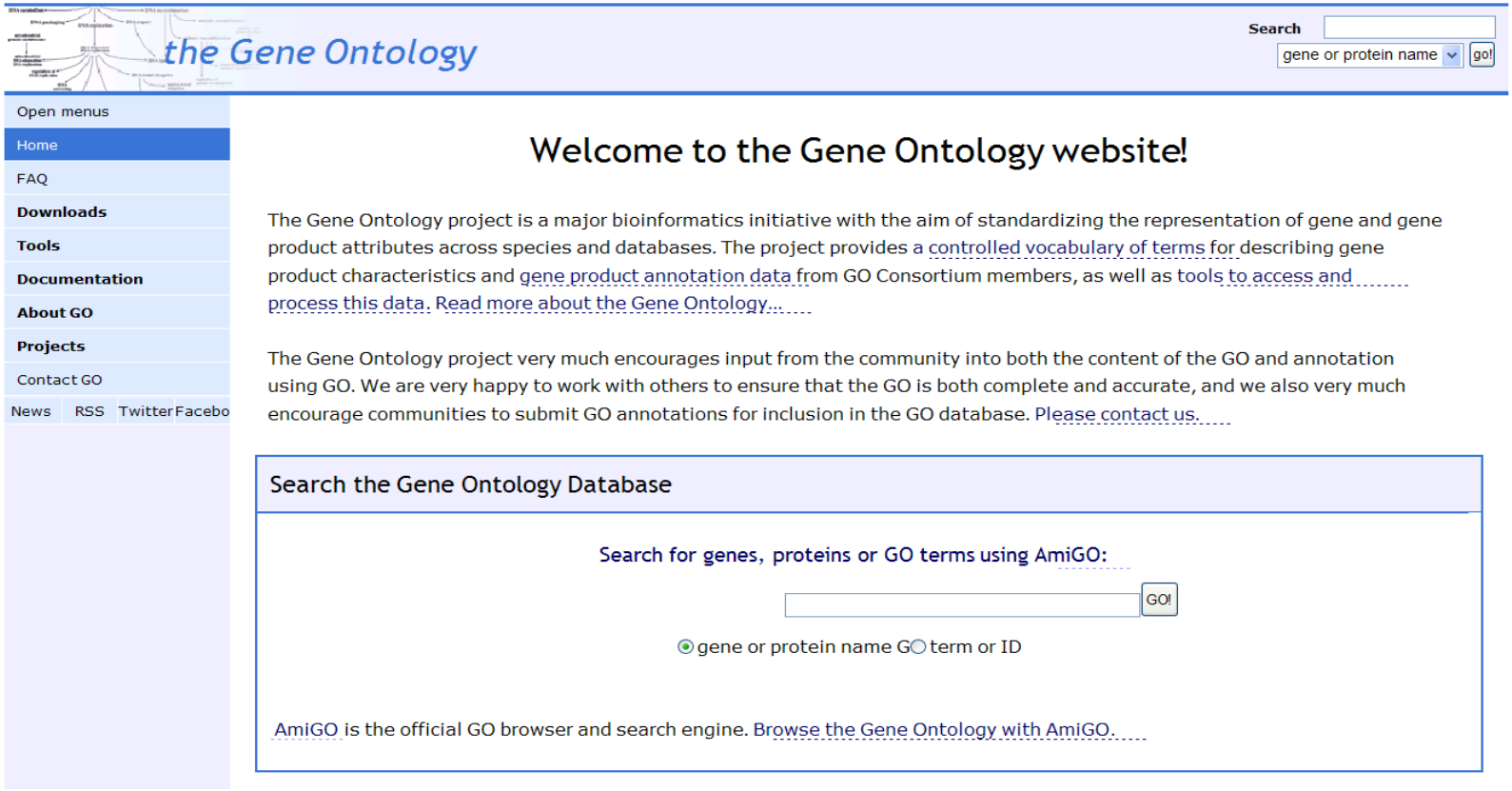
基因注释数据库产生的原因

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。

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。

❖ 1.1 基因本体学数据库

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cellular component、molecular function和biological process。



GO数据库收录的基因组数据列表

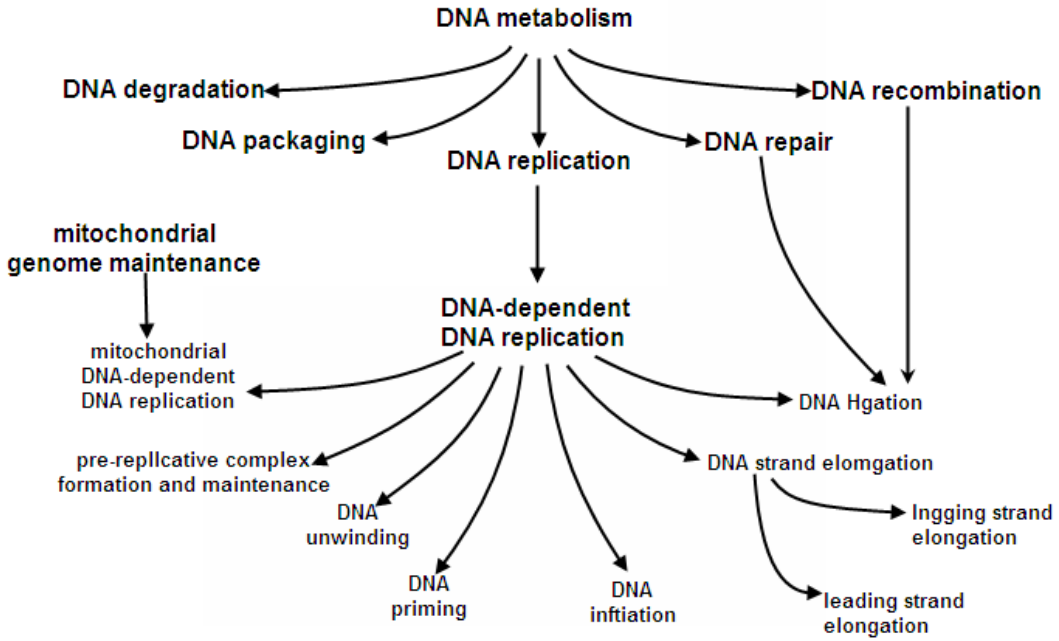
- GO数据库最初收录的基因信息来源于3个模式生物数据库：果蝇、酵母和小鼠，随后相继收录了更多数据，其中包括国际上主要的植物，动物和微生物基因组数据库。

- 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。

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	http://www.genedb.org
	恶性疟原虫	
	硕大利什曼原虫	
	布氏锥虫	
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	http://www.yeastgenome.org
	酿酒酵母	
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

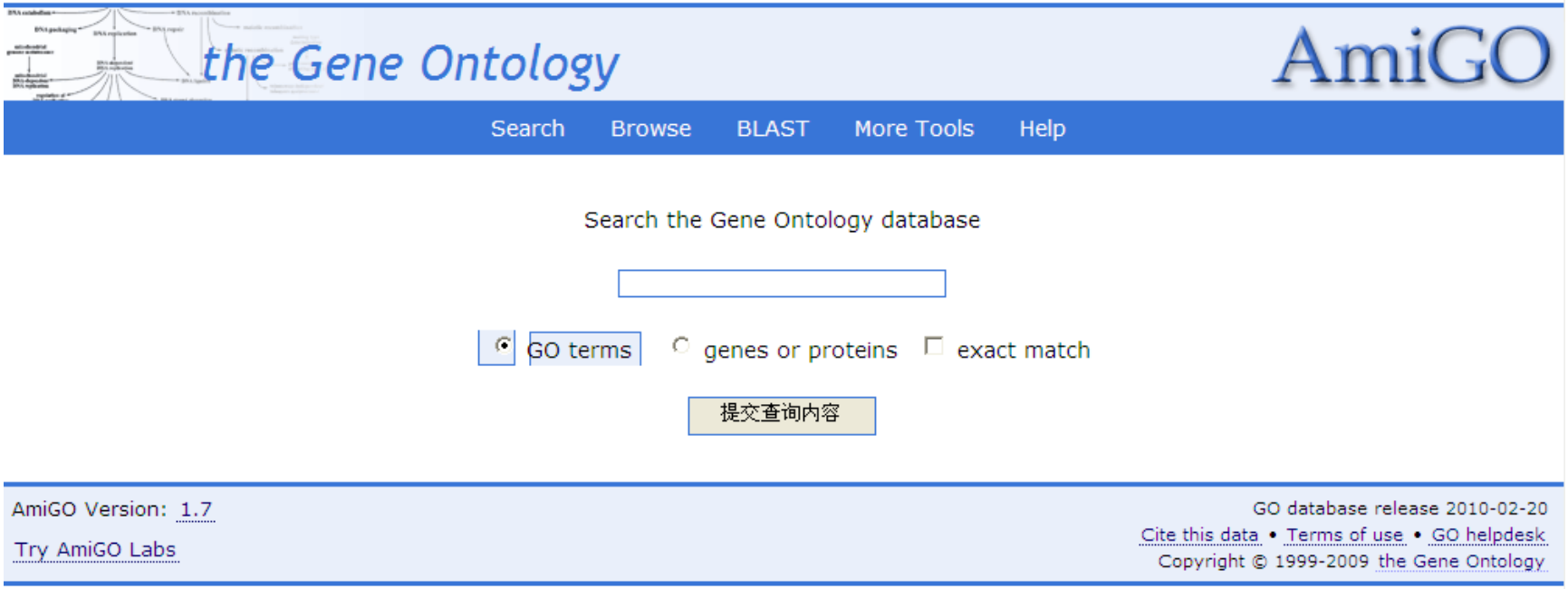
- GO通过控制注释词汇的层次结构使得研究人员能够从不同层面查询和使用基因注释信息。
- 从整体上来看GO注释系统是一个有向无环图(Directed Acyclic Graphs),包含三个分支,即：生物学过程(biological process)，分子功能(molecular function)和细胞组分(cellular component)。
- 注释系统中每一个结点(node)都是基因或蛋白的一种描述,结点之间保持严格的关系,即“is a”或“part of”。



❖ 使用GO数据库

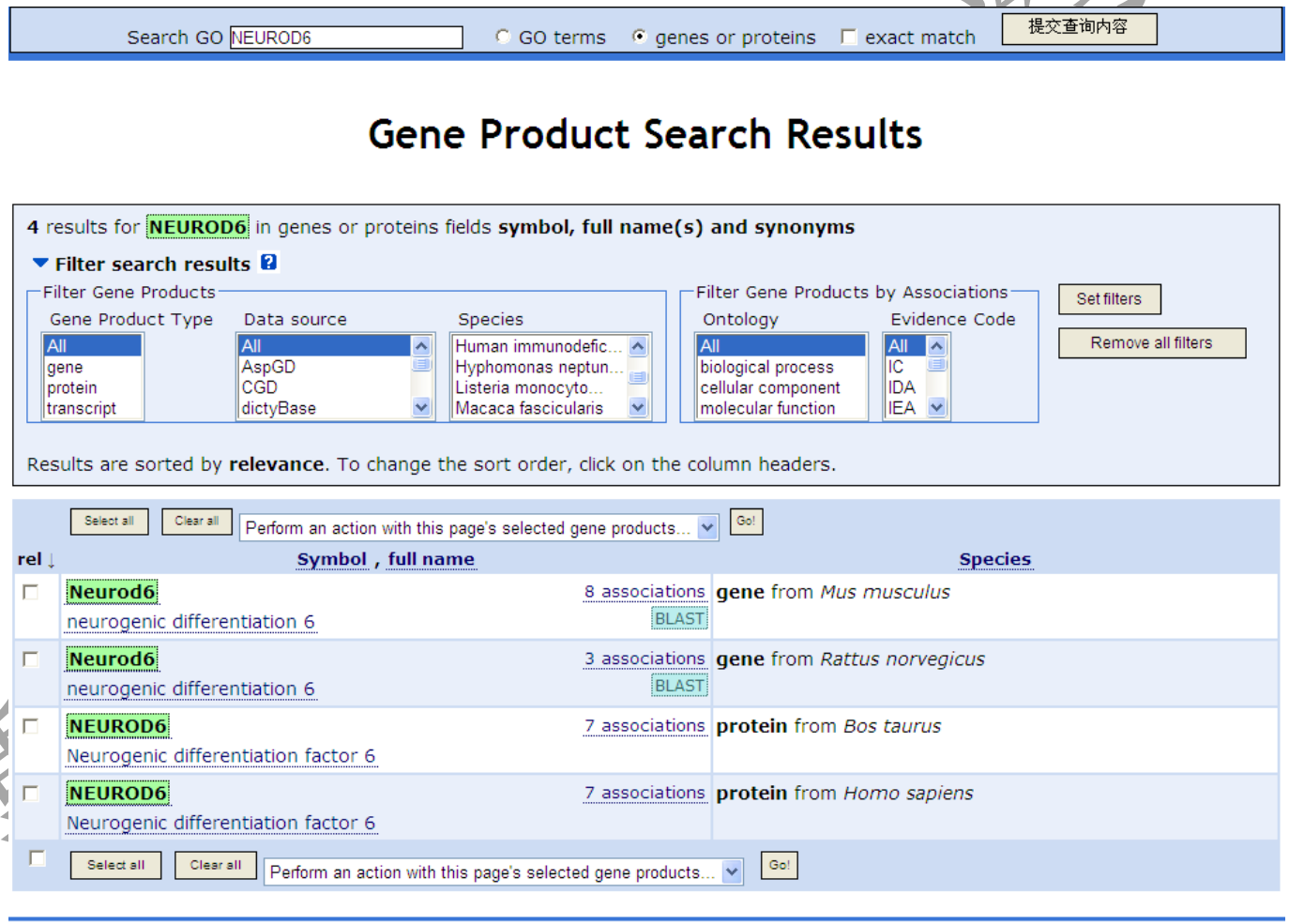
1. 用关键词检索GO数据库

- 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。

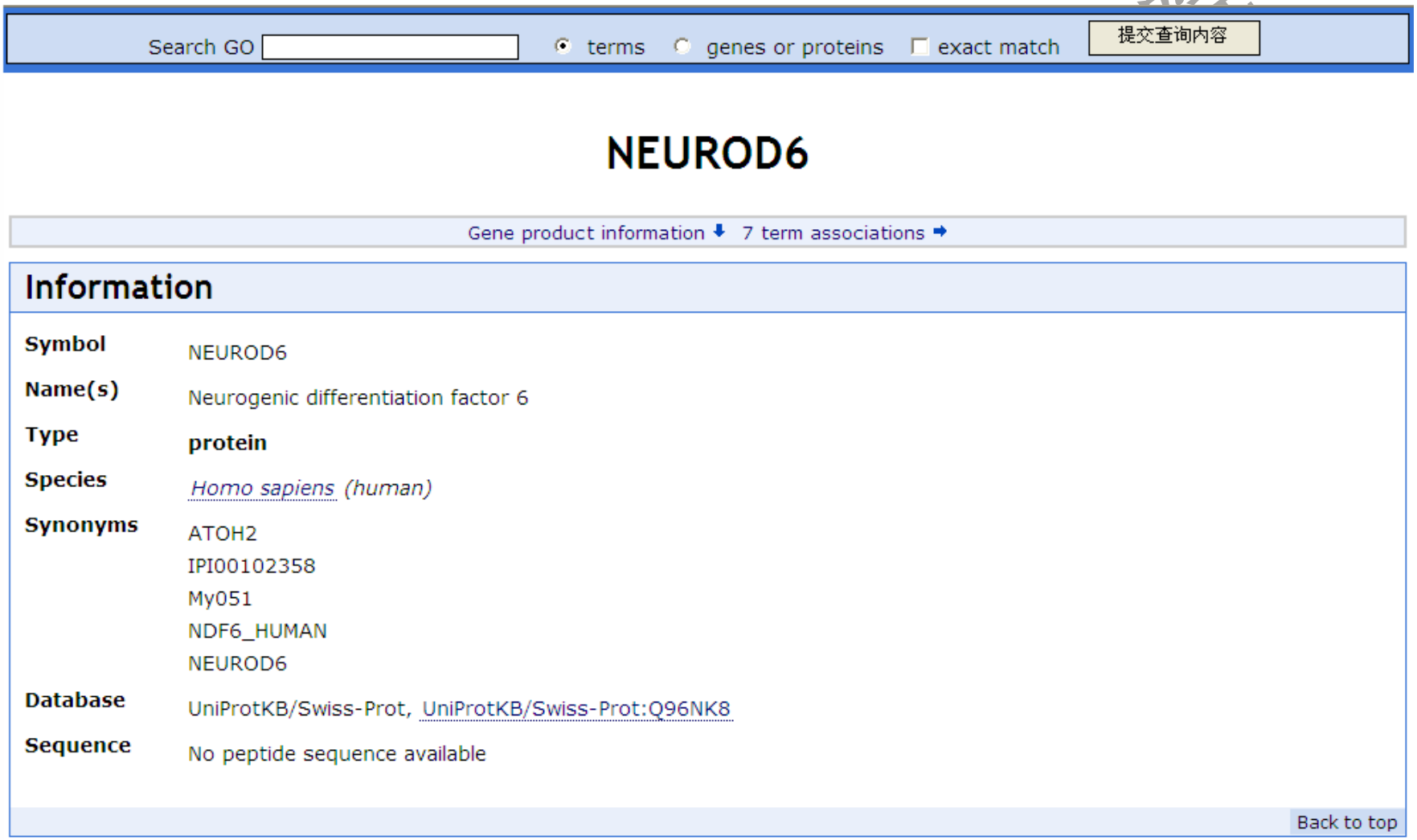


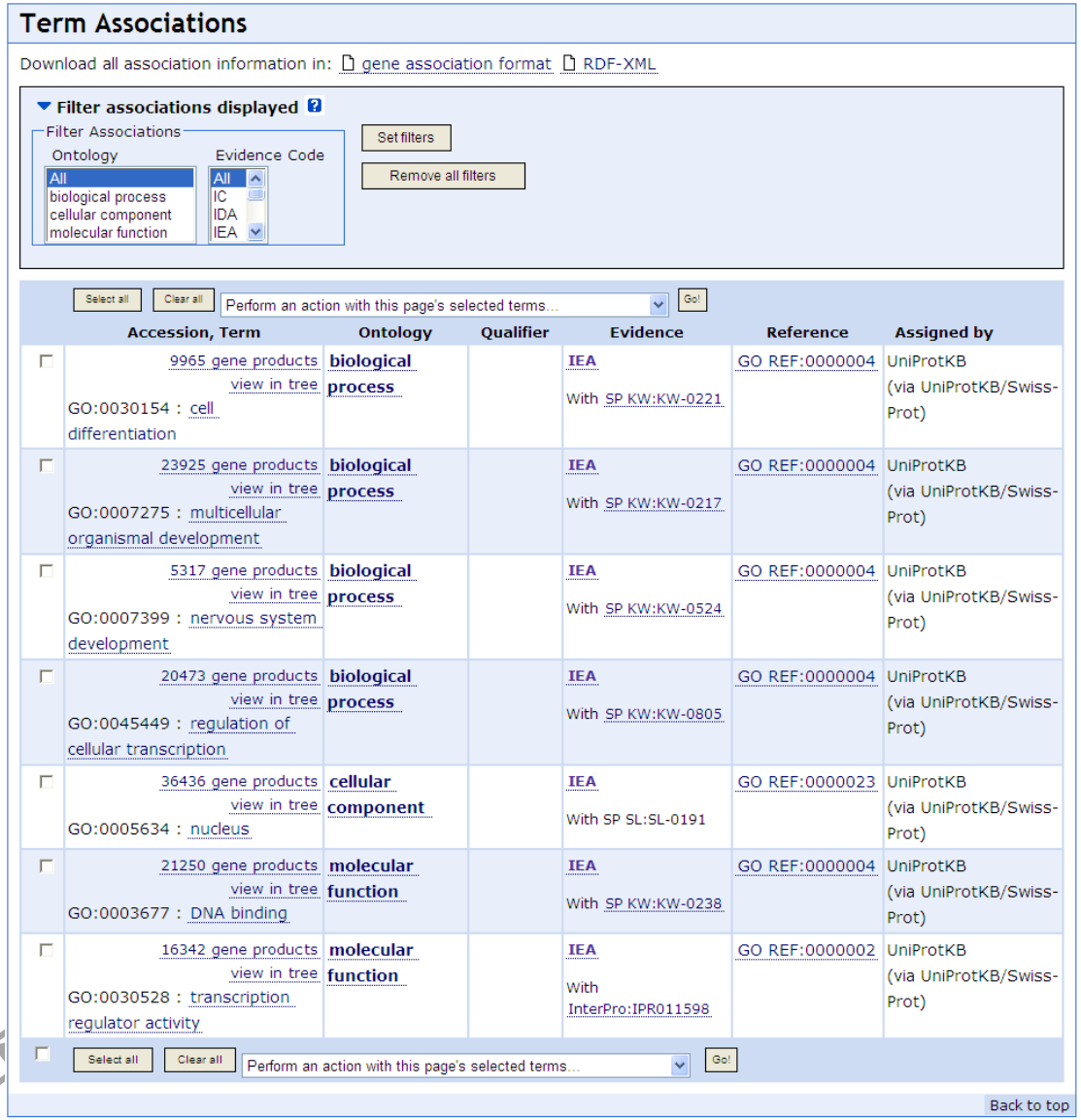
【举例】

- 这里以检索神经源性分化因子6（NEUROD6）为例。在检索框中输入“NEUROD6”并勾选“gene and proteins”和“exact match”,运行后所得基因产物检索结果如图所示。



此图显示了该基因产物的基本信息，包括类型、物种、别名来源和序列





此图显示了该基因产物的术语关联（term associations）图，图中记录名称“Term”是GO记录的名字，“Ontology”是该基因产物的特性，如要查看其分子功能，可点击其中的一条记录“nervous system development”。

nervous system development

Term information ⬇ Term lineage ⬇ External references ⬇ 5317 gene product associations ➡

Term Information

Accession	GO:0007399
Ontology	**biological process**
Synonyms	**related**: pan-neural process
Definition	The process whose specific outcome is the progression of nervous tissue over time, from its formation to its mature state. [source: GOC:dgh]
Comment	None
Subset	None
Community	There have been **0** comments for this term. If you would like to view or participate in the community annotation, please continue to the GONUTS page.

Back to top

Term Lineage

Switch to viewing term parents, siblings and children

Filter tree view

Filter Gene Product Counts — Data source: All, AspGD, CGD, dictyBase; Species: All, Anaplasma phagocy..., Arabidopsis thaliana, Bacillus anthraci...

View Options — Tree view: Full / Compact

Set filters | Remove all filters

- all : all [372469 gene products]
  - GO:0008150 : biological_process [274193 gene products]
    - GO:0032502 : developmental process [27802 gene products]
      - GO:0048856 : anatomical structure development [20054 gene products]
        - GO:0048731 : system development [15068 gene products]
          - **GO:0007399 : nervous system development [5317 gene products]**
      - GO:0007275 : multicellular organismal development [23925 gene products]
        - GO:0048731 : system development [15068 gene products]
          - **GO:0007399 : nervous system development [5317 gene products]**
    - GO:0032501 : multicellular organismal process [32735 gene products]
      - GO:0007275 : multicellular organismal development [23925 gene products]
        - GO:0048731 : system development [15068 gene products]
          - **GO:0007399 : nervous system development [5317 gene products]**

Actions...
Last action: Reset the tree
Graphical View
View in tree browser
Download...
OBO
RDF-XML
GraphViz dot

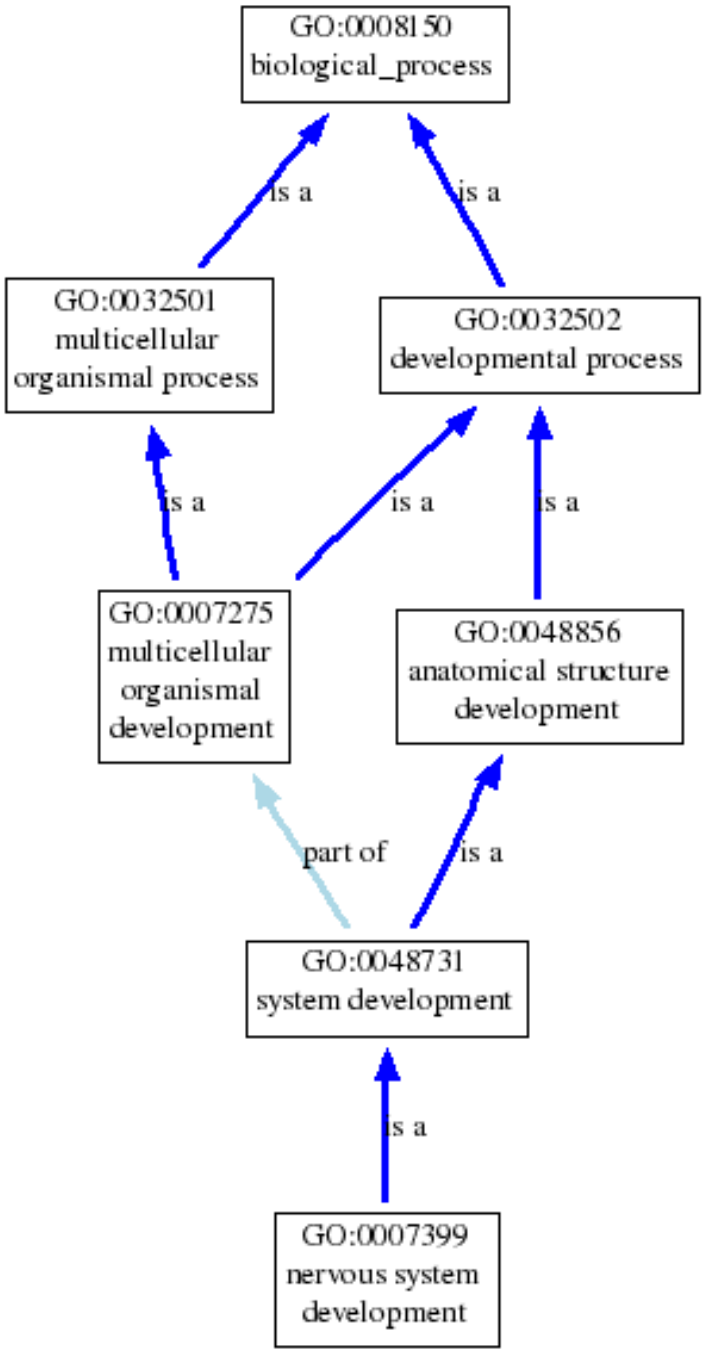
Back to top

External References

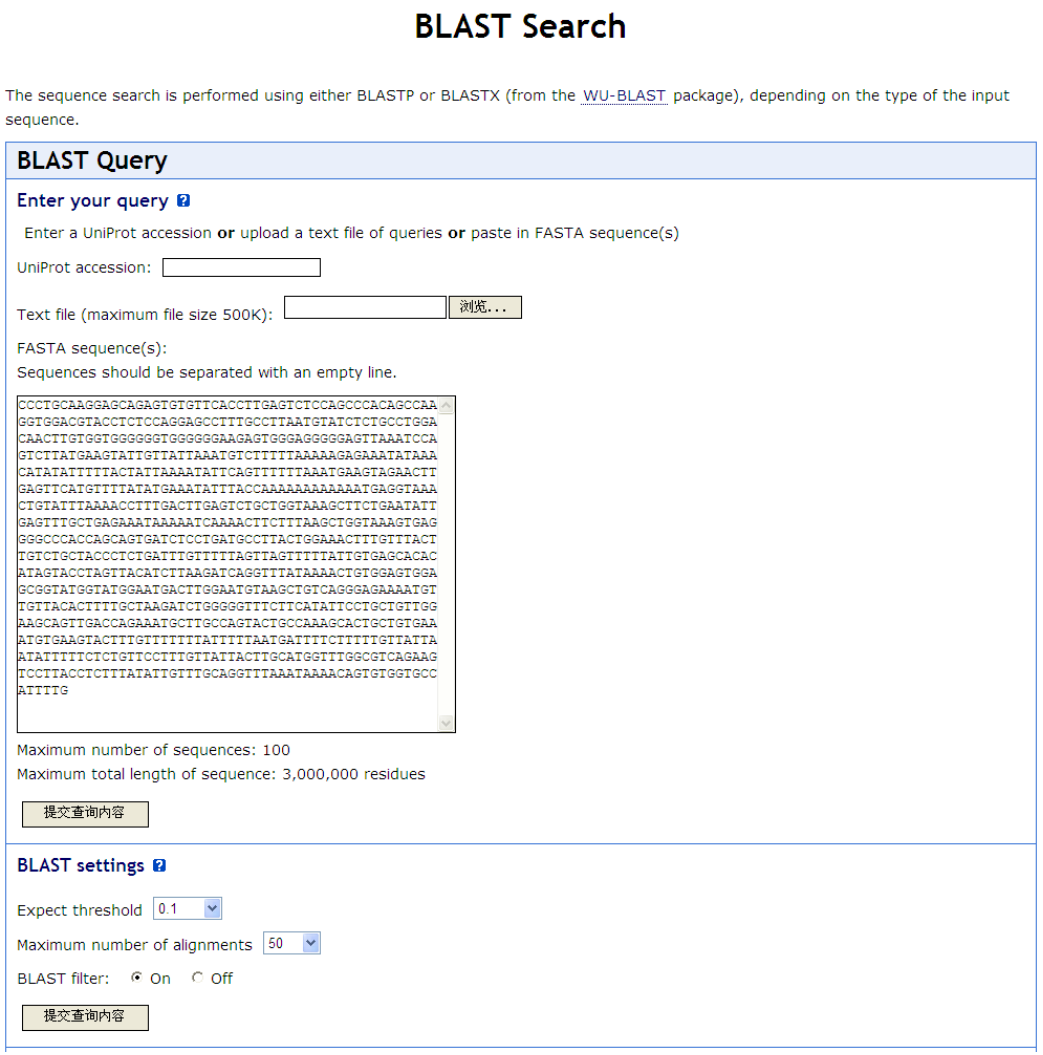
- MIPS_funcat (2)
  - 45.03.17
  - 47.03.01
- SP_KW (1)
  - KW-0524

Back to top

此图上部先对神经源性分化因子6的相关信息做简单描述，中间术语系谱（term lineage）成阶梯状分布，记录了GO数据库中全部分子功能所处的位置和关系。下方“External Reference”提供了与外部相关数据的链接。



点击上图右上方的可视化视图（graphical view）就更清晰地显示了分子功能记录之间构成的复杂网状结构，既有上下隶属关系，也存在平行关系。



2. 用序列检索GO数据库

- 对于未知基因名的序列，可以用序列直接检索GO 数据库。点击AmiGO首页上方的“BLAST”。
- 界面风格类似于其他数据库BLAST搜索的网页，在检索框中铁如氨基酸或核酸序列，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。

❖ 1.2 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。1

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

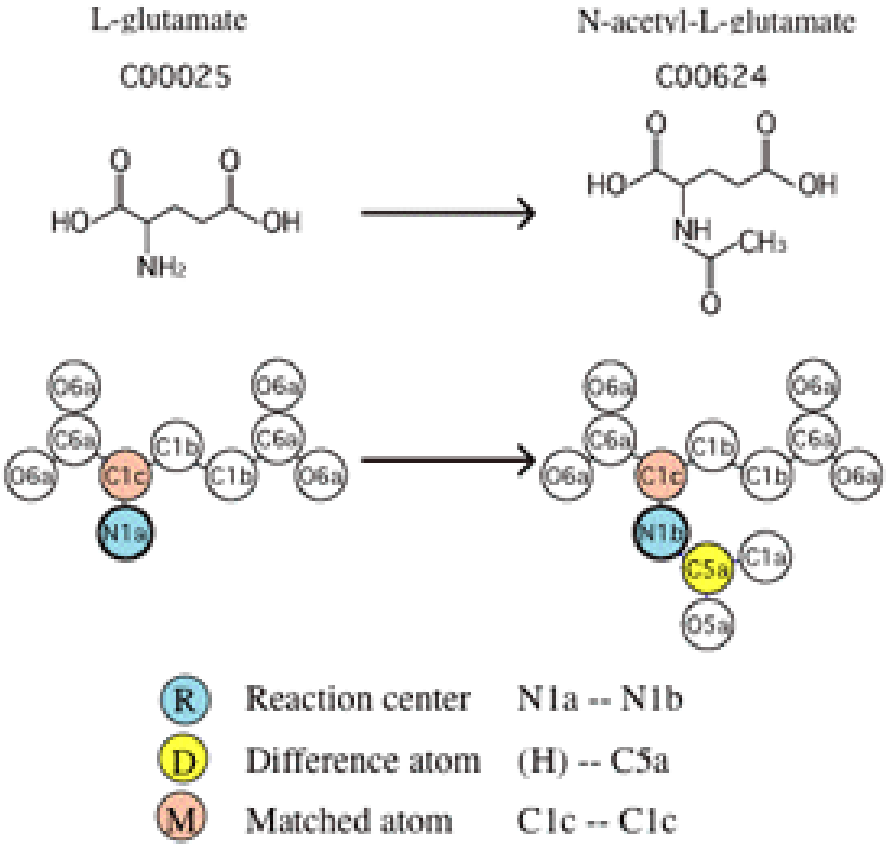
表 8-2　KEGG 的 13 个核心数据库的检索号

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

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。

RDM Pattern

- 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:diffevent atom M:matched atom)**模式。



(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类葡萄糖磷酸变位酶基因名称“PGM1”

PGM1 Go Clear

KEGG Home
Introduction
Overview
Release notes
Current statistics

KEGG Identifiers

KEGG XML

KEGG API

KEGG FTP

KegTools

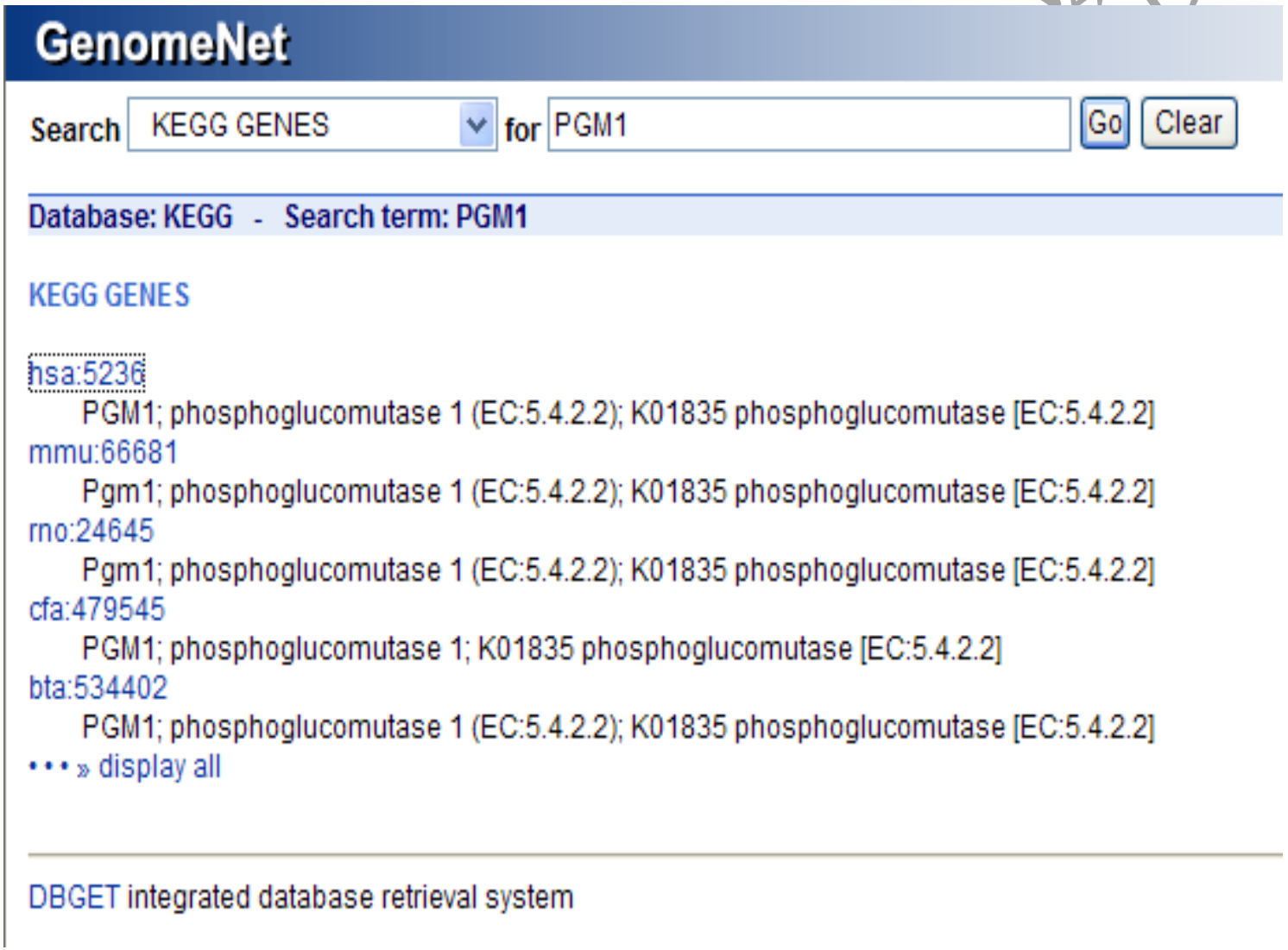
KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

Main entry point to the KEGG web service

KEGG2 KEGG Table of Contents Update notes Help

- 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。



- 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。

KEGG

Homo sapiens (human): 5236

Help

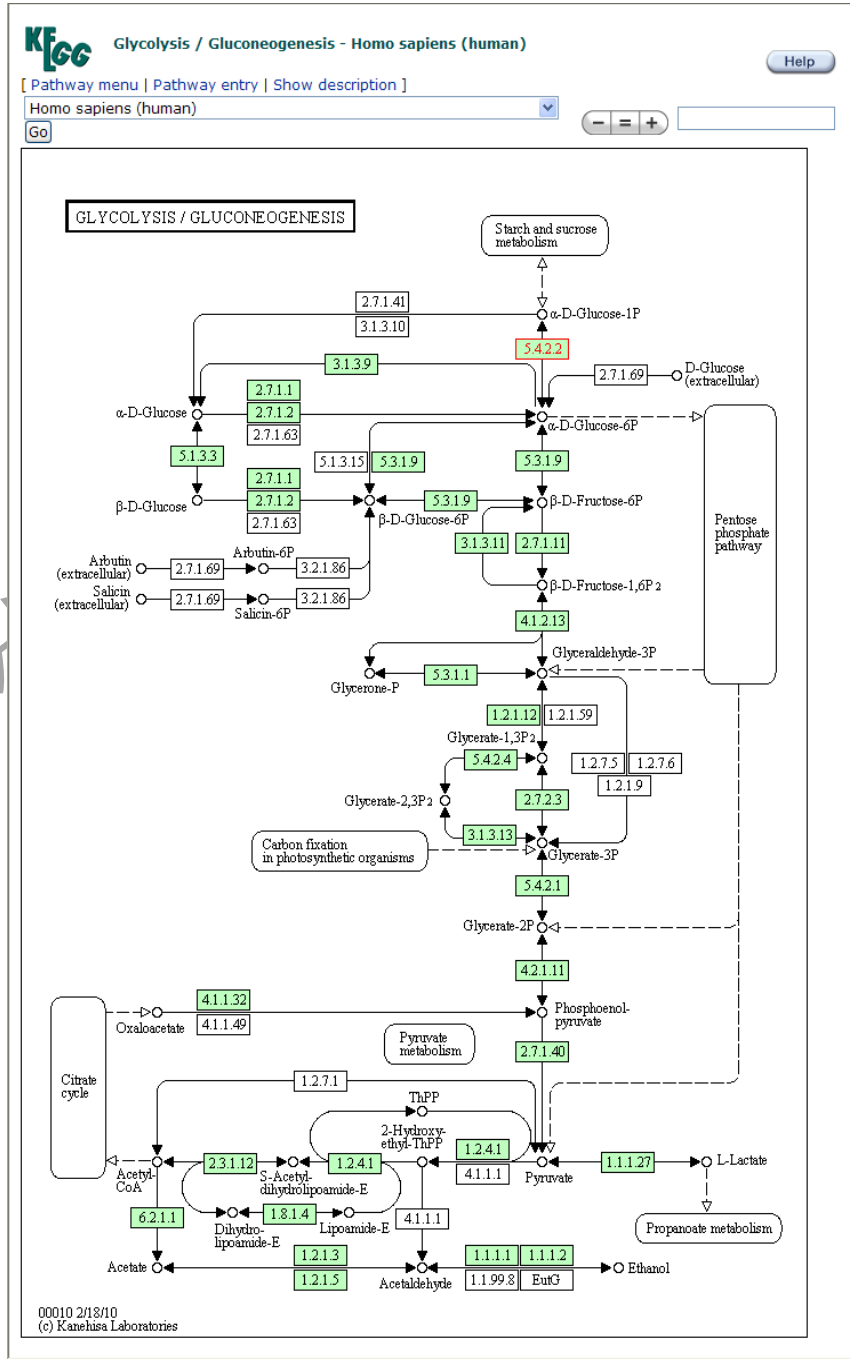
Entry	5236 CDS H.sapiens
Gene name	PGM1
Definition	phosphoglucomutase 1 (EC:5.4.2.2)
Orthology	K01835 phosphoglucomutase [EC:5.4.2.2]
Pathway	hsa00010 Glycolysis / Gluconeogenesis hsa00030 Pentose phosphate pathway hsa00052 Galactose metabolism hsa00500 Starch and sucrose metabolism hsa00520 Amino sugar and nucleotide sugar metabolism hsa01100 Metabolic pathways
Class	Metabolism; Carbohydrate Metabolism; Glycolysis / Gluconeogenesis [PATH:hsa00010] Metabolism; Carbohydrate Metabolism; Pentose phosphate pathway [PATH:hsa00030] Metabolism; Carbohydrate Metabolism; Galactose metabolism [PATH:hsa00052] Metabolism; Carbohydrate Metabolism; Starch and sucrose metabolism [PATH:hsa00500] Metabolism; Carbohydrate Metabolism; Amino sugar and nucleotide sugar metabolism [PATH:hsa00520] BRITE hierarchy
SSDB	Ortholog Paralog Gene cluster GFIT
Motif	Pfam: PGM_PMM_I PGM_PMM_III PGM_PMM_II PGM_PMM_IV PROSITE: PGM_PMM Motif
Other DBs	NCBI-GI: 21361621 NCBI-GeneID: 5236 OMIM: 171900 HGNC: 8905 HPRD: 01389 Ensembl: ENSG00000079739 UniProt: P36871
Position	1p31
AA seq	562 aa AA seq DB search MVKIVTVKTQAYQDQKPGTSGLRKRVKVFQSSANYAENFIQSIISTVEPAQRQEATLVVG GDGRFYMKEAIQLIARIAAANGIGRLVIGQNGILSTPAVSCIIRKIKAIGGIILTASHNP GGPNGDFGIKFNISNGGPAPEAITDKIFQISKTIEEYAVCPDLKVDLGVLGKQQFDLENK FKPFTVEIVDSVEAYATMLRSIFDFSALKELLSGPNRLKIRIDAMHGVVGPYVKKILCEE LGAPANSAVNCVPLEDFGGHHPDPNLTYAADLVETMKSGEHDFGAAFDGDGDRNMILGKH GFFVNPSDSVAVIAANIFSIPYFQQTGVRGFARSMPTSGALDRVASATKIALYETPTGWK FFGNLMDASKLSLCGEESFGTGSDHIREKDGLWAVLAWLSILATRKQSVEDILKDHWQKY GRNFFTRYDYEEVEAEGANKMMKDLEALMFDRSFVGKQFSANDKVYTVEKADNFEYSDPV DGSISRNQGLRLIFTDGSRIVFRLSGTGSAGATIRLYIDSYEKDVAKINQDPQVMLAPLI SIALKVSQLQERTGRTAPTVIT
NT seq	1689 nt NT seq atggtgaagatcgtgacagttaagacccaggcgtaccaggaccagaagccgggcacgagc gggctgcggaagcgggtgaaggtgttccagagcagcgccaactacgcggagaacttcatc cagagtatcatctccaccgtggagccggcgcagcgggcaggaggccacgctggtggtgggc ggggacggccggttctacatgaaggaggccatccagctcatcgctcgcatcgctgccgcc aacgggatcggtcgcttggttatcggacagaatggaatcctctccacccctgctgtatcc tgcatcattagaaaaatcaaagccattggtgggatcattctgacagccagtcacaaccca gggggccccaatggagattttggaatcaaattcaatatttctaatggaggtcctgctcca gaagcaataactgataaaattttccaaatcagcaagacaattgaagaatatgcagtttgc cctgacctgaaagtagaccttggtgttctgggaaagcagcagtttgacttggaaaataag ttcaaacccttcacagtggaaattgtggattcggtagaagcttatgctacaatgctgaga agcatctttgatttcagtgcactgaaagaactacttttctgggccaaaccgactgaagatc cgtattgatgctatgcatggagttgtgggaccgtatgtaaagaagatcctctgtgaagaa ctcggtgcccctgcgaactcggcagttaactgcgttcctctggaggactttggaggccac cacccctgacccccaacctcacctatgcagctgacctggtggagaccatgaagtcaggagag catgattttggggctgccttttgatggagatggggatcgaaacatgattctgggcaagcat gggttctttgtgaaccctcagactctgtggctgtcattgctgccaacatcttcagcatt ccgtatttccagcagactggggtccgcggctttgcacggagcatgcccacgagtggtgct ctggaccgggtggctagtgctacaaagattgctttgtatgagaccccaactggctggaag ttttttgggaatttgatggagcgcgagcaaactgtccctttgtgggggaggagcttcggg accggttctgaccacatccgtgagaaagatggactgtgggctgtccttgcctggctctcc atcctagccaccccgcaagcagagtgtggaggacattctcaaagatcattggcaaaagtat ggccggaatttcttcaccaggtatgattacgaggaggtggaagctgagggcgcaaacaaa atgatgaaggacttggaggccctgatgtttgatcgctcctttgtggggaagcagttctca gcaaatgacaagttacactgtggagaaggccgatcacttgaatacaggcgacccagtg gatggaagcatttcaagaaatcagggctgcgcctcatttttcacagatggttctcgaatc gtcttccgactgagcggcactgggagtgccgggggccaccattcggctgtacatcgatagc tatgagaaggacgttgccaagattaaccaggaccccaggtcatgttggccccccttatt tccattgctctgaaagtgtcccagctgcaggagaggacgggacgcactgcacccactgtc atcacctaa

All links

Pathway (6)
KEGG PATHWAY (6)
Disease (1)
OMIM (1)
Chemical reaction (1)
KEGG ENZYME (1)
Genome (1)
KEGG GENOME (1)
Gene (14)
KEGG ORTHOLOGY (1)
NCBI-Gene (1)
NCBI-GI (8)
UniGene (1)
HGNC (1)
HPRD (1)
ENSEMBL-HSA (1)
Protein sequence (4)
UniProt (1)
RefSeq(pep) (1)
IPI (2)
DNA sequence (11)
RefSeq(nuc) (1)
GenBank (5)
EMBL (5)
Protein domain (5)
Pfam (4)
PROSITE (1)
All databases (43)

DBGET integrated database retrieval system

- 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。

KEGG数据库的改进与更新

- 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多个新增加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3节：基因功能富集分析（GSEA）

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。

1).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过出现(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。

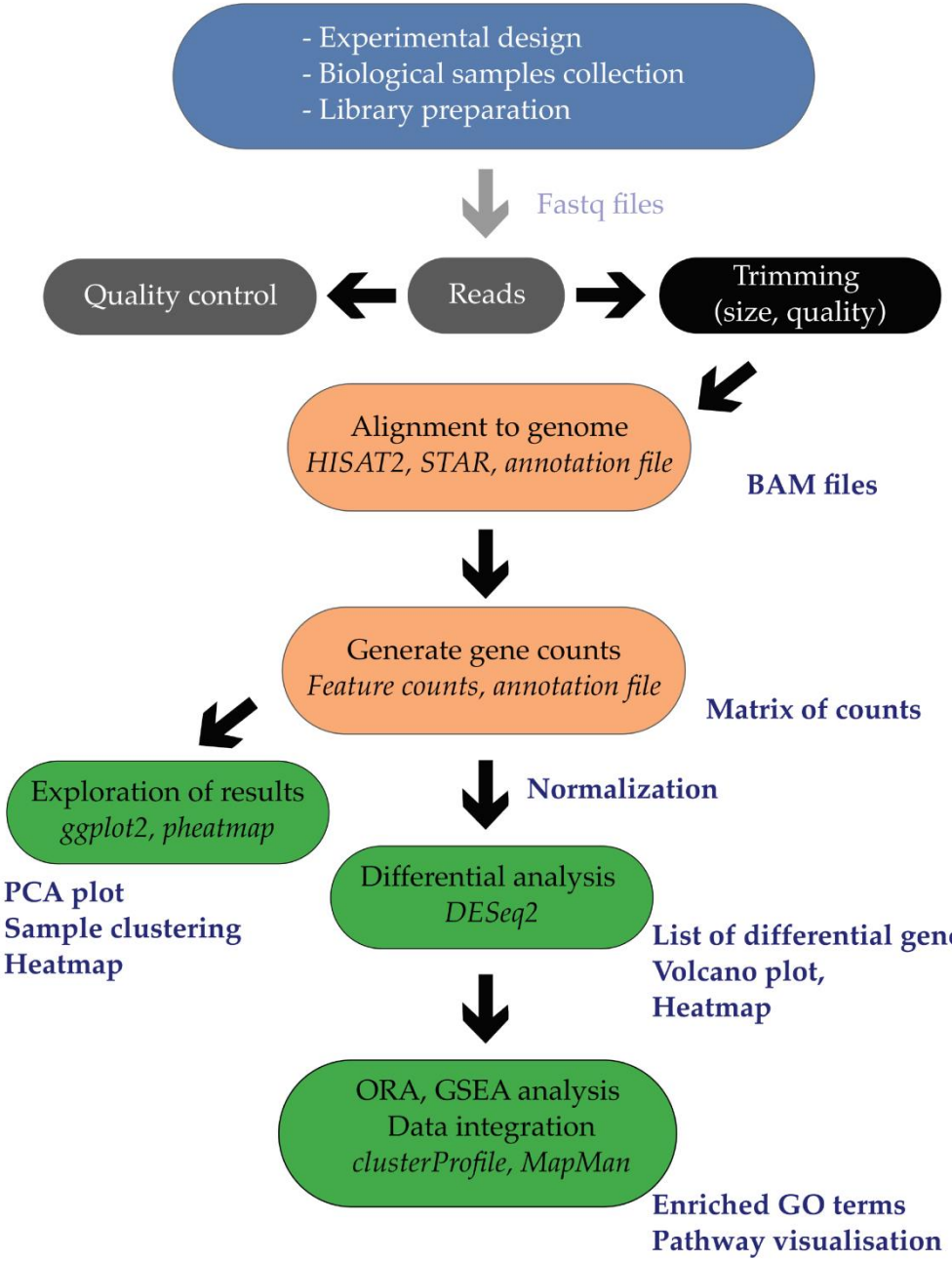
- 富集分析中常用的统计方法有累计超几何分布、Fisher精确检验等。

  - 累计超几何分布：

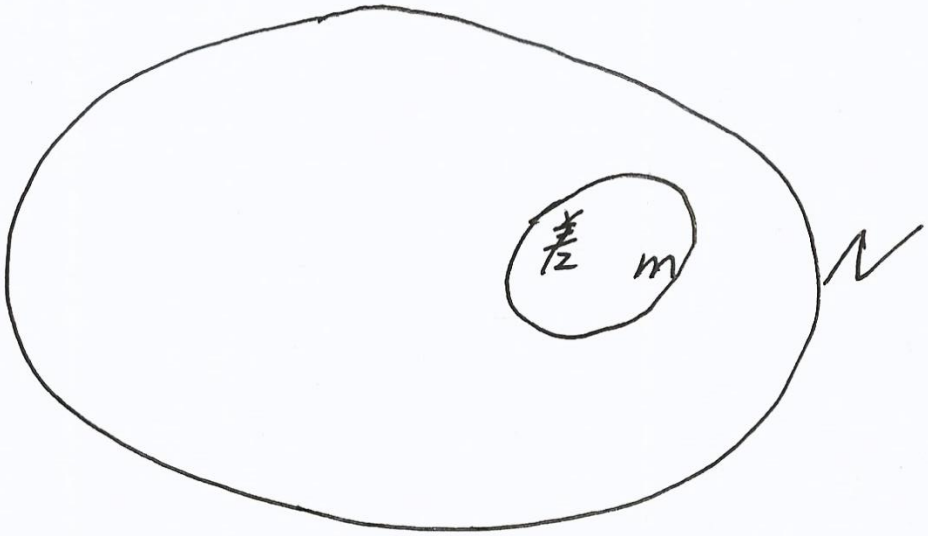
$$P(X > q) = 1 - \sum_{x=1}^{q} \frac{\binom{n}{x}\binom{N-n}{M-x}}{\binom{N}{M}}$$

  - Fisher精确检验：

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


- Experimental design
- Biological samples collection
- Library preparation
Fastq files
Quality control
Reads
Trimming
(size, quality)
Alignment to genome
HISAT2, STAR, annotation file
BAM files
Generate gene counts
Feature counts, annotation file
Matrix of counts
Exploration of results
ggplot2, pheatmap
Normalization
PCA plot
Sample clustering
Heatmap
Differential analysis
DESeq2
List of differential genes
Volcano plot,
Heatmap
ORA, GSEA analysis
Data integration
clusterProfile, MapMan
Enriched GO terms
Pathway visualisation

什么是超几何检验



从该N个中抽出n个，其中k个是不合格（差）的几率：

$$f(k;n;m;N)=\frac{C_m^k \cdot C_{N-m}^{n-k}}{C_N^n}$$

假如现在提取了***n***个基因（如差异表达基因），里面有***k***个属于某个特定的**pathway**。那么要判断这些基因在该**pathway**是否显著富集，则需要计算出一次性抽到不少于***k***个该**pathway**中基因的概率***p***：

$$p = 1 - \sum_{k=0}^{k-1} \frac{C_m^k \cdot C_{N-m}^{n-k}}{C_N^n}$$

背景知识：概率极小（如**p < 0.05**）的事件在一次试验中不可能出现！

那么，这n个差异基因在整个基因组中是随机分布的吗？

So，若该计算出的概率低于某个给定的显著性水平（如p <0.05），则可被判定为该pathway显著富集。

MMEASE

Meta-analysis of Metabolomics data by Enhanced metabolite Annotation and marker Selection and Enrichment analysis

Home | Analysis | Tutorial | Contact Us

MMEASE is a web-based platform for meta-analysis of multiple metabolomics datasets, and is designed for biologists with little background in statistics to perform sophisticated analysis on metabolomics data. Six analysis steps are included: Data Upload & Integration, Batch Effect Removal, Sample Separation, Marker Identification, Metabolite Annotation, and Enrichment Analysis. Four key features characterized MMEASE as a useful online tool for metabolomics are:

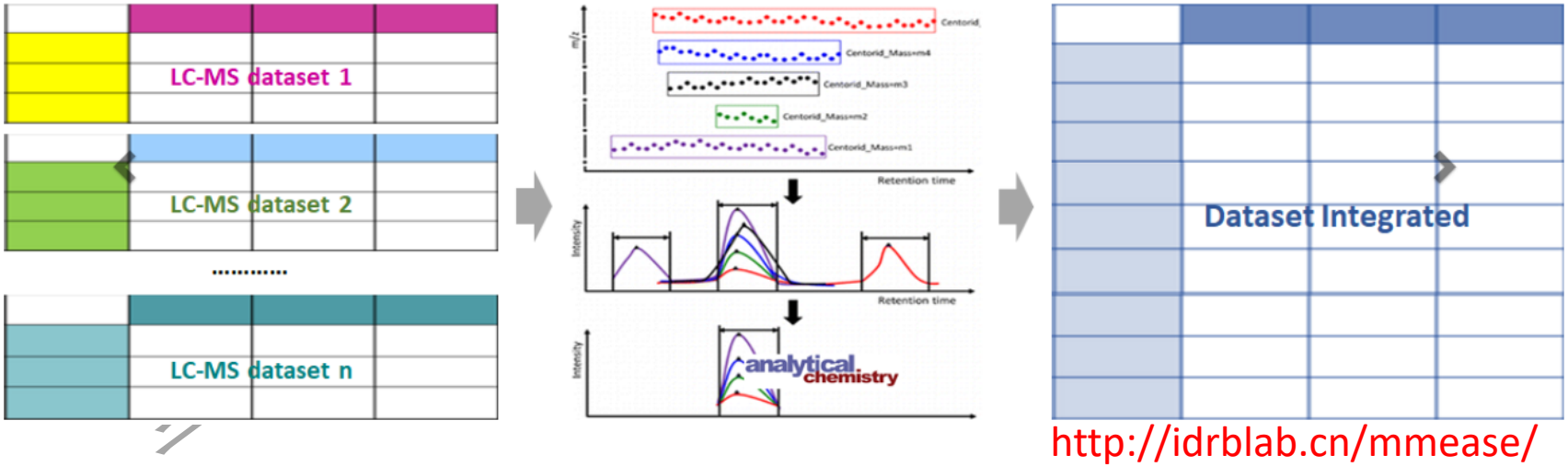
(1) Meta-analysis: integration of metabolomics datasets from multiple experiments or laboratories is conducted and tested based on Zhang's work (*Analytical Chemistry* , 2014. 86(13):6245-53).

(2) Enhanced metabolite annotation: 262,483 metabolites can be annotated including 169,352 peptides, 29,290 endogenous and 42,330 exogenous metabolites.

(3) Diverse statistical analyzing methods: MMEASE provides 15 statistic methods for identifying metabolic markers, 7 of which are applied for the first time among popular online servers for metabolomics data analysis.

(4) More optional databases for enrichment analysis: metabolites enrichment analysis is conduced based on KEGG and SMPD pathways, HMDB bio-functions, CFam structures and species/genus origin of traditionalmedicine.

Integration of metabolomics datasets from multiple experiments or laboratories based on Zhang's work
Anal Chem. 86(13):6245-53 (2014)



http://idrblab.cn/mmease/

Yang Q, **Li B**, Chen S, Tang J, Li Y, Li Y, Zhang S, Shi C, Zhang Y, Mou M. 2020. MMEASE: Online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis. *Journal of Proteomics*. 104023.

2).常用富集分析软件

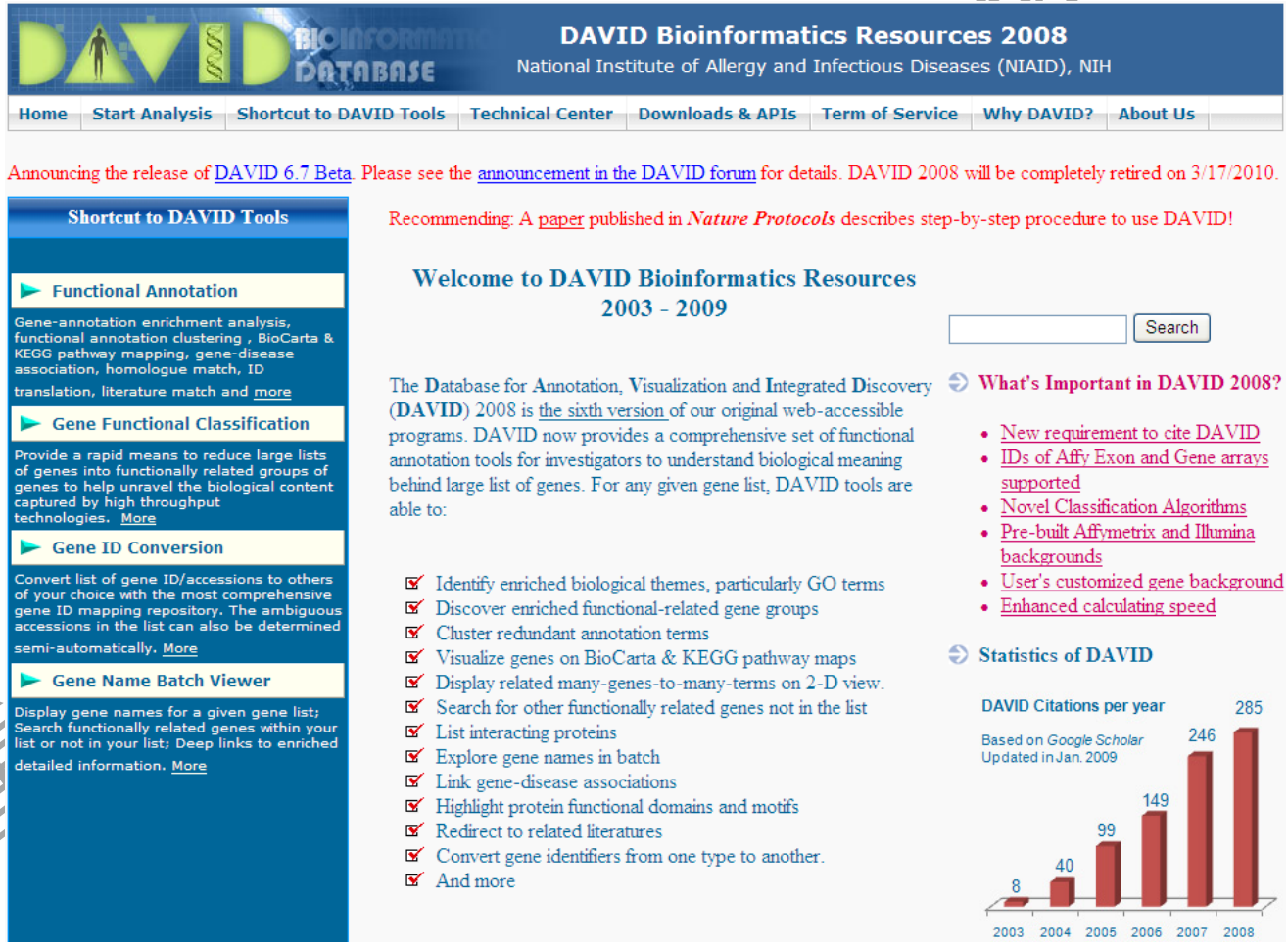
- 基于不同的算法原理，可以将目前的常用富集分析工具分为三类：单一富集分析（**singular enrichment analysis**），基因集富集分析（**gene set enrichment analysis**），模块富集分析（**modular enrichment analysis**）。

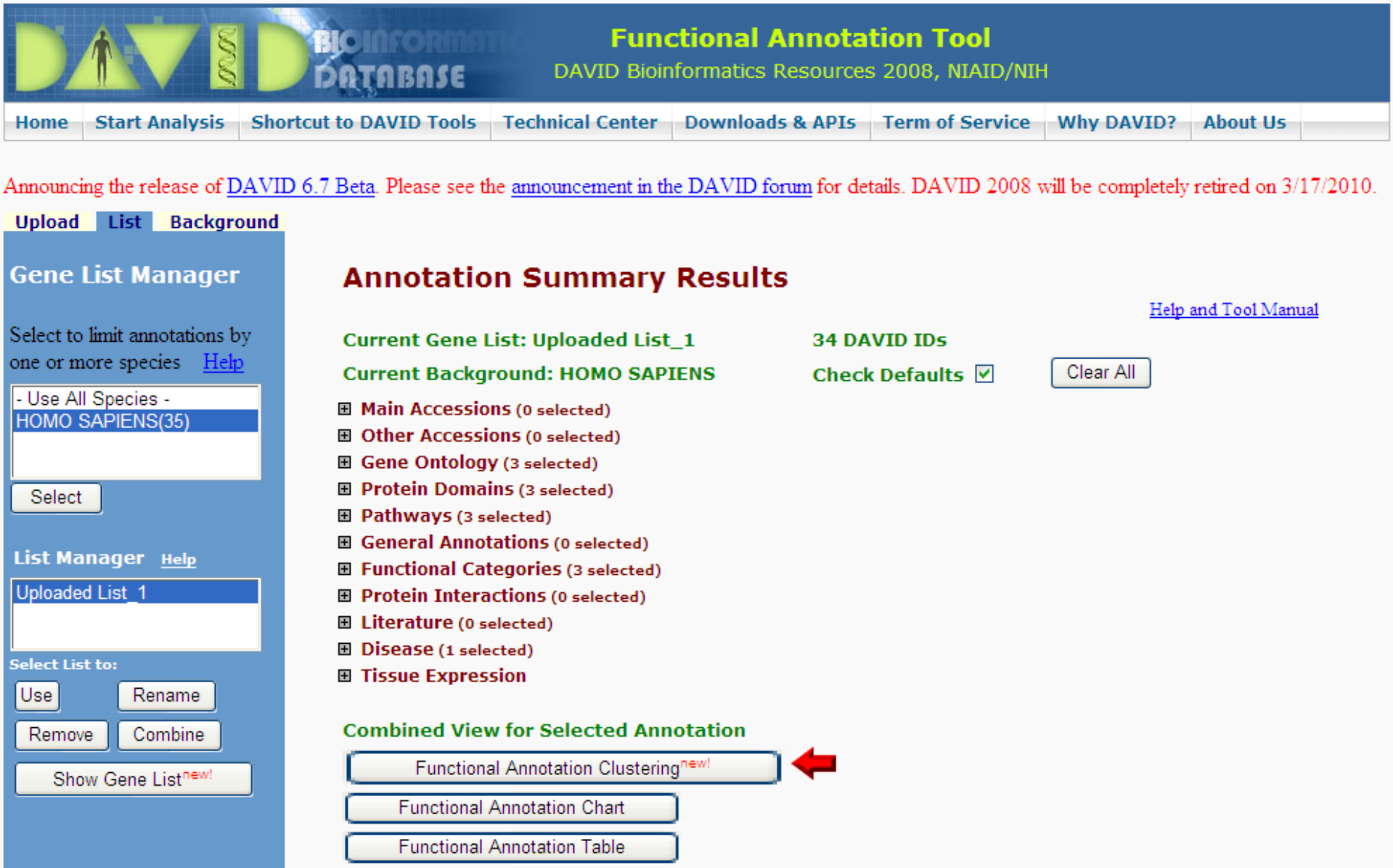
表 8-4 常用富集分析工具集

Enrichment tool name	Year of release	Key statistical method	Category
FunSpec	2002	Hypergeometric	Class I
Onto-express	2002	Fisher's exact;hypergeometic; 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
GOStat	2004	Fisher's exact; chi-square	Class I
GoSurfer	2004	Chi-square	Class I

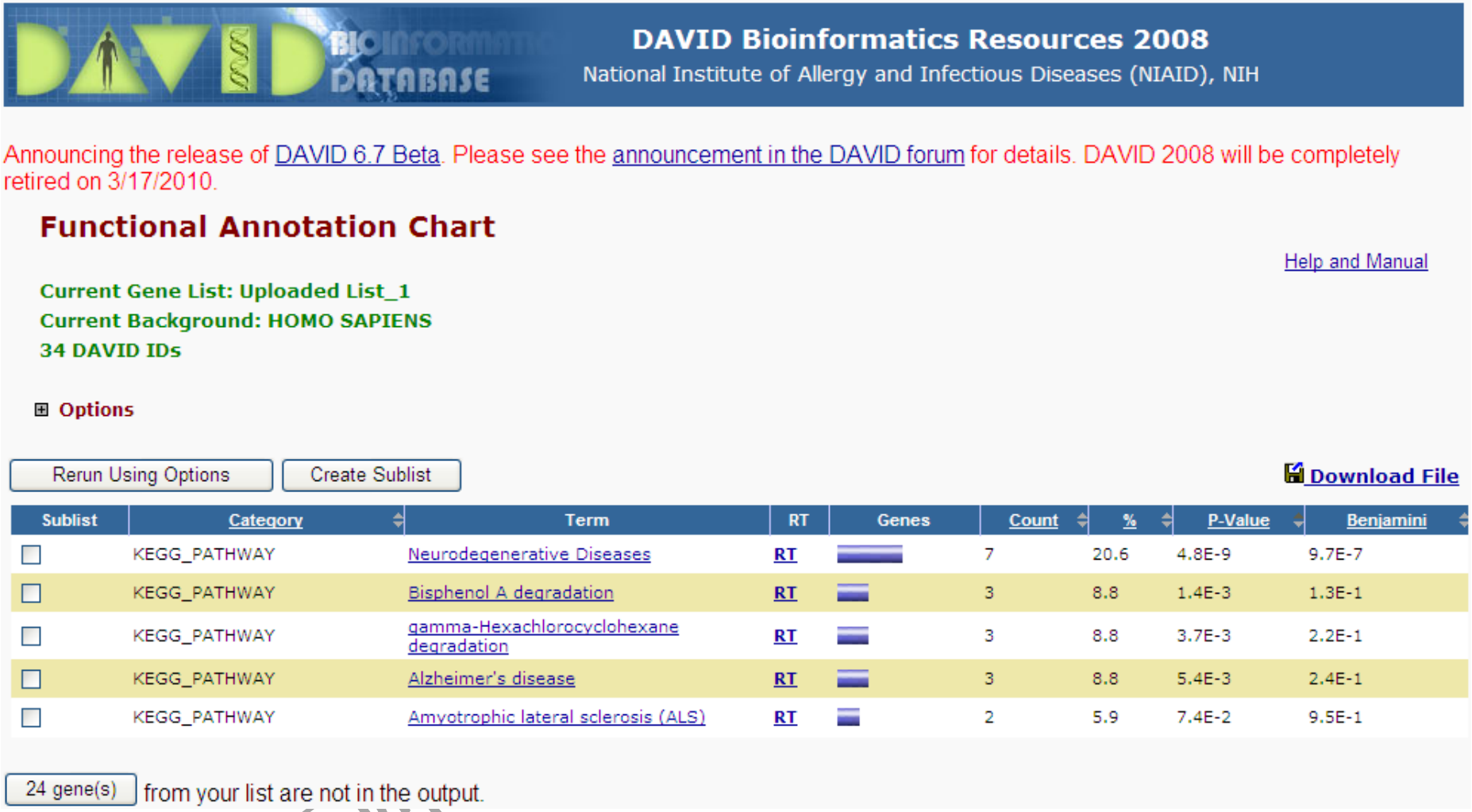
3).富集分析应用实例

- 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。





点击“Start Analysis”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。



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。

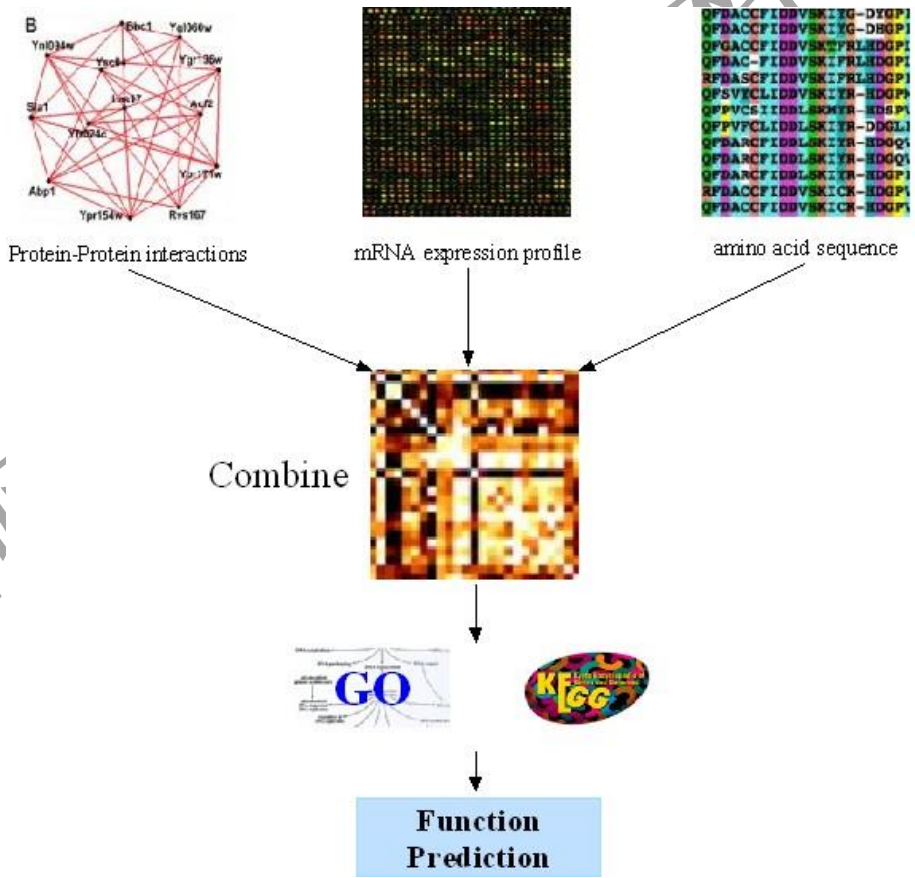
第4节：基因功能预测

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。



1).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

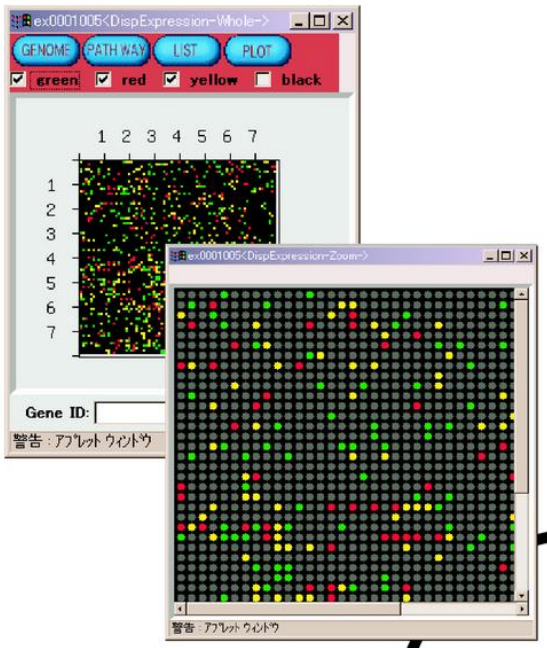
- 目前，利用相互作用网络进行功能注释主要有两种方法，即直接注释方法（direct annotation schemes）和基于模块的方法（module assisted schemes）。

3. 利用GO体系结构比较基因功能

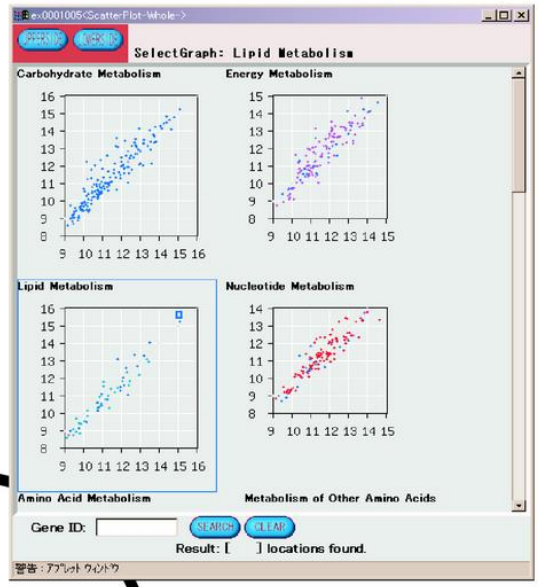
- 通常认为如果两个基因产物的功能相似，那么它们的表达也就相近，同时它们在GO中注解的结点就相似，所以只要能找出GO中结点对的相似度，就可以近似估计两基因表达的相似度，从而判断两基因产物的功能的相似度。

2).基于KEGG的基因功能预测

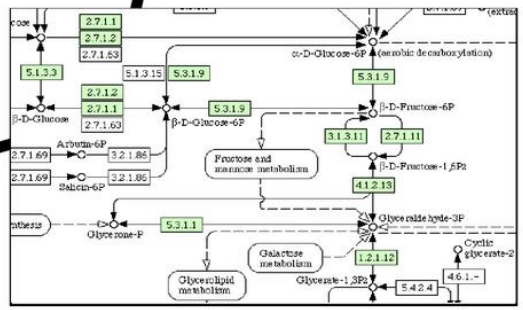
通路分析是现在经常被使用的芯片数据基因功能分析法。与GO分类法（应用单个基因的GO分类信息）不同，通路分析法利用的资源是许多已经研究清楚的基因之间的相互作用，即生物学通路。研究者可以把表达发生变化的基因集导入通路分析软件中，进而得到变化的基因都存在于哪些已知通路中，并通过统计学方法计算哪些通路与基因表达的变化最为相关。



array browser applet



scatterd plot browser applet



3).常用基因功能预测软件

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

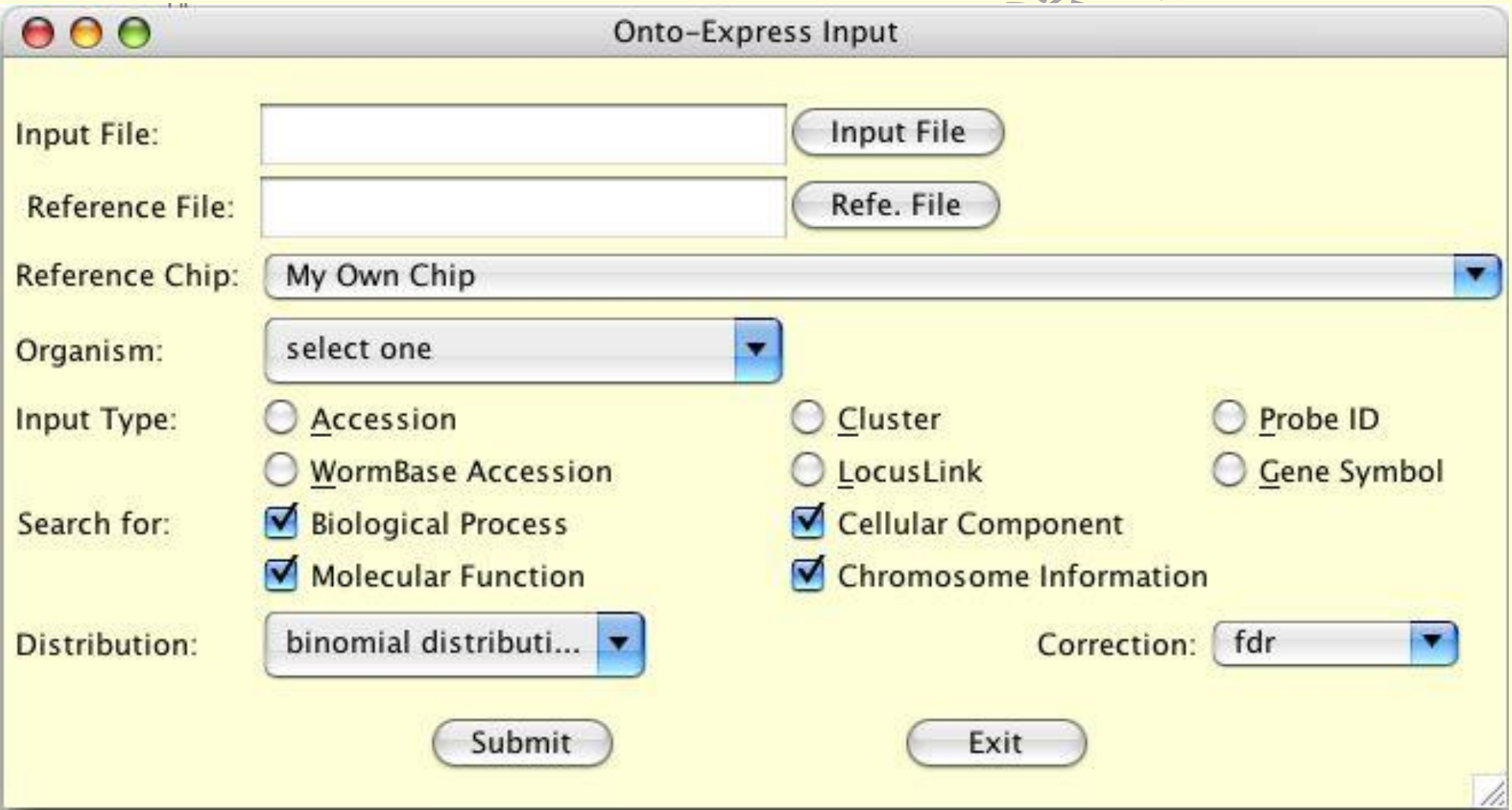
【举例】

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。

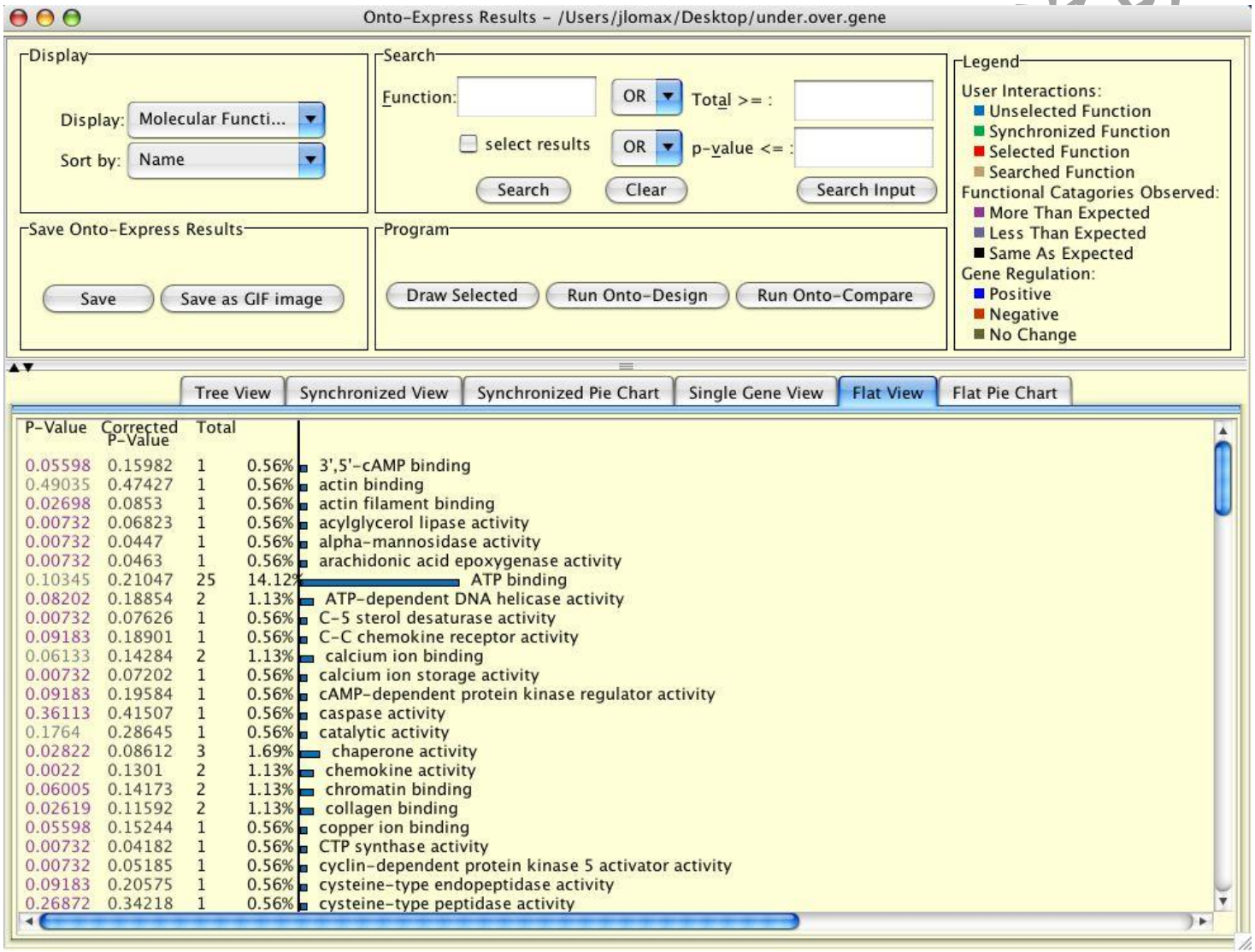
1. 数据输入

- **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：**



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。



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、PathwaryMiner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。

重庆师范大学
CHONG QING NORMAL UNIVERSIYT

Thanks for your attention!

Acknowledgement

College of Life Sciences, Chongqing Normal University

2022, Chongqing of P. R. C