生物信息学

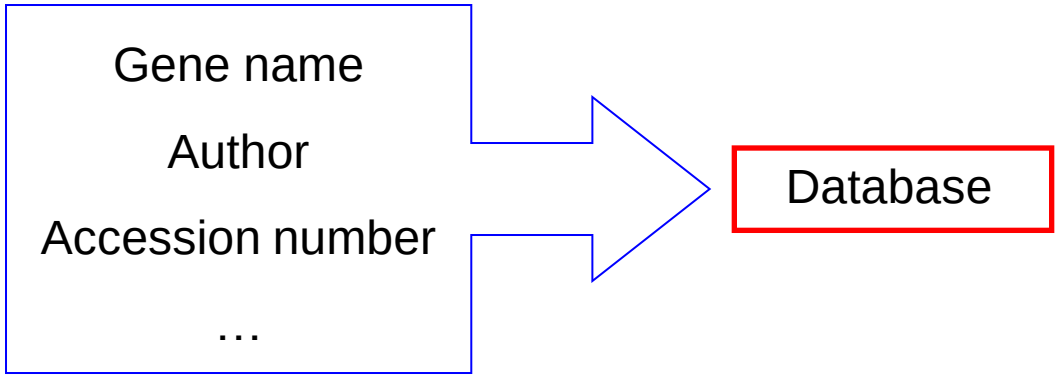
第三章

数据库检索与序列比对

检索数据库的方法

- 用关键词或词组进行数据库检索

  (Text-based database searching)



- 用核苷酸或蛋白质序列进行数据库检索

  (Sequence-based database searching)

```
G-IISKILRKE-KGGYEITIVDASNERQVID
GKIVAITALSEKKGGFEVSIEKA-NGEVVVD
```

关键词或词组为基础的数据库检索

关键词
- 名词、描述性词、词组
- 序列注册号 (Accession number)

检索体系
- Entrez
- Sequence Retrieval System (SRS)
- Integrated database retrieval system (DBGET)

检索须知（1）

- 连接词 **AND, OR, NOT（Boolean operators）**

  A AND B (同时包含检索词A和检索词B的信息)

  A NOT B（含有A但不含有B的信息）

  A OR B（含有A 或含有B 的信息）

注意事项：

1、AND, OR, NOT must be entered in UPPERCASE

2、Boolean operators are processed in a left-to-right sequence

3、The order can be changed by enclosing individual concepts in parentheses (processed first)

PubMed

rice AND microarray OR expression profile >284,000 records

rice AND (microarray OR expression profile) 4244 records

检索须知（2）

- 用引号将两个单词组成一个词组
  - **16S rRNA = 16S AND rRNA**
  - **“16S rRNA”**

exact match

Nucleotide

16S rRNA	~41749765 sequences
“16S rRNA”	~1433522 sequences

- *放在单词后使检索范围扩大，但专一性降低

- **pseudopod*** =pseudopod OR pseudopodia OR pseudopodium

1. Entrez

Entrez, The Life Sciences Search Engine

http://www.ncbi.nlm.nih.gov/gquery/

NCBI 的检索体系

优点：三种检索体系中最容易操作的体系
缺点：检索范围有限

Entrez可对8大类40个数据库进行检索

Sequence, Structure, Expression…

- **Nucleotide:** Core subset of nucleotide sequence records
- **EST:** Expressed Sequence Tag records
- **GSS:** Genome Survey Sequence records
- **Protein:** sequence database
- **Genome:** whole genome sequences
- **Structure:** three-dimensional macromolecular structures
- **Taxonomy:** organisms in GenBank
- **SNP:** single nucleotide polymorphism
- **dbVar:** Genomic structural variation
- **Gene:** gene-centered information
- **SRA:** Sequence Read Archive
- **BioSystems:** Pathways and systems of interacting molecules
- **HomoloGene:** eukaryotic homology groups
- **GENSAT:** gene expression atlas of mouse central nervous system
- **Probe:** sequence-specific reagents
- **Genome Project:** genome project information
- **dbGaP:** genotype and phenotype
- **UniGene:** gene-oriented clusters of transcript sequences
- **CDD:** conserved protein domain database
- **UniSTS:** markers and mapping data
- **PopSet:** population study data sets
- **GEO Profiles:** expression and molecular abundance profiles
- **GEO DataSets:** experimental sets of GEO data
- **Epigenomics:** Epigenetic maps and data sets
- **Cancer Chromosomes:** cytogenetic databases
- **PubChem BioAssay:** bioactivity screens of chemical substances
- **PubChem Compound:** unique small molecule chemical structures
- **PubChem Substance:** deposited chemical substance records
- **Protein Clusters:** a collection of related protein sequences
- **Peptidome:** MS/MS proteomic experiments
- **OMIA:** online Mendelian Inheritance in Animals
- **BioSample:** biological material descriptions

检索方法（1）：跨库检索 (cross-database search)

Entrez系统中数据库之间的连接

NCBI主页选择“All Databases”或Entrez主页，输入关键词

↓

各个数据库中检索到的信息数量

↓

点击相应数据库查看信息目录，每一条信息与其它数据库的相关信息链接

检索方法（2）：选择数据库检索

NCBI主页选择数据库，输入关键词

↓

检索到的信息目录，每一条信息
与其它数据库的相关信息链接

↓

查看信息内容

应用举例

- 文本检索
  - 获取视黄醇结合蛋白（**Retinol binding protein, RBP4**）**Entrez**中的条目
    - 直接检索：**rbp4**

2. SRS (Sequence Reterieval System)

http://srs.ebi.ac.uk/
European Bioinformatics Institute (EBI) 的检索体系

优点：检索面宽
缺点：操作复杂

17大类194个数据库与 SRS 体系相连

- Literature, Bibliography and Reference databases
- Nucleotide sequence databases
- Uniprot Universal Protein Resource
- Other protein sequence databases
- Deprecated Protein Databases
- Nucleotide related databases
- Protein function databases
- Protein structure databases
- Enzymes, reactions and metabolic pathway databases
- Mutation and SNP databases
- Gene ontology resources
- Biological Resources Catalogues
- Mapping databases
- Other databases
- User owned databases
- Application result databases
- EMBOSS result databases

检索方法（1）：快速检索(Quick search)

- 操作简单，检索数据库有限
- 适用于目标明确的检索

在SRS主页选择检索类别，输入关键词

↓

检索到的信息目录，每一条信息
与其它数据库的相关信息链接

↓

查看信息内容

检索方法（2）：高级检索(advanced search)

- 操作稍微复杂，可以检索所有数据库
- 适用于范围广泛的检索

在SRS主页点击“Library Page”

↓

在“Library Page”网页选择数据库，然后点击“Query Form”

↓

在“Query Form”网页输入关键词检索

↓

检索到的信息目录，每一条信息与其它数据库的相关信息链接

3. DBGET (Integrated database retrieval system)

http://www.genome.jp/dbget/

日本GenomeNet的检索体系

优点：与 Kyoto Encyclopedia of Genes and Genomes (KEGG) database 相连

操作较SRS简单

缺点：检索面较 SRS 窄

DBGET与40多个数据库相连

DBGET检索体系中数据库之间的连接

检索方法（1）：单库检索（basic search）

在DBGET主页选择一个数据库

↓

输入关键词检索

↓

查看检索到的信息目录

↓

查看信息详细内容

检索方法（2）：跨库检索 （LinkDB）

在DBGET主页点击“LinkDB”

↓

在查询网页选择数据库

↓

输入关键词检索
（数据库:编号）

↓

结果

不是总能得到你所需要的信息

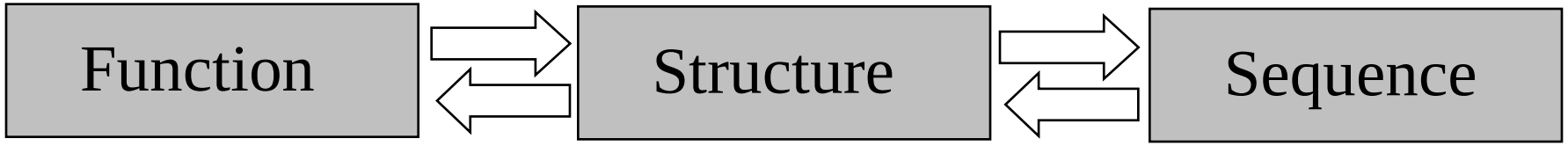
- 关键词的使用

  retrotransposon **retro-transposon**

- 数据库所包含数据的多少和范围
- 不同的数据库包含内容有限
- 关键词的拼写错误

核苷酸和蛋白质序列比对

核苷酸和蛋白质序列为基础的数据库检索



- 序列对位排列（**sequence alignment**）
- 将两条或多条序列对位排列，突出相似的结构区域

表示序列的字符

Most Common Letters Used for DNA Nucleotide Sequences		
1-Letter Code	*Nucleotide Name*	*Category*
A	Adenine	Purine
C	Cytosine	Pyrimidine
G	Guanine	Purine
T	Thymine	Pyrimidine
N	Any nucleotide (any base)	(n/a)
R	A or G	Purine
Y	C or T	Pyrimidine
--	-----	None (gap)

Nonpolar Amino Acids (hydrophobic)

amino acid	three letter code	single letter code
glycine	Gly	G
alanine	Ala	A
valine	Val	V
leucine	Leu	L
isoleucine	Ile	I
methionine	Met	M
phenylalanine	Phe	F
tryptophan	Trp	W
proline	Pro	P

Polar (hydrophilic)

serine	Ser	S
threonine	Thr	T
cysteine	Cys	C
tyrosine	Tyr	Y
asparagine	Asn	N
glutamine	Gln	Q

Electrically Charged (negative and hydrophilic)

aspartic acid	Asp	D
glutamic acid	Glu	E

Electrically Charged (positive and hydrophilic)

lysine	Lys	K
arginine	Arg	R
histidine	His	H

两条蛋白质序列对位排列分析

```
序列 1: 192  NYLTGSIPDDLFNNTPLLTYLNVGNNSLSGLIPGCIGSLPILQHLNFQANNLTGAVPPAI 251
             NYLTG IP+ LFNNTP L +L +GNNSLSG IP CIGSLP+L+ L  Q NNLTG VPP+I
序列 2: 183  NYLTGLIPNGLFNNTPSLKHLIIGNNSLSGPIPSCIGSLPLLERLVLQCNNLTGPVPPSI 242

序列 1: 252  FNMSKLSTISLISNGLTGPIPGNTSFSLPVLRWFAISKNNFFGQIPLGLAACPYLQVIAM 311
             FNMS+L  I+L SNGLTGPIPGN SF LP+L++F++  N F GQIPLGLAAC +L+V ++
序列 2: 243  FNMSRLHVIALASNGLTGPIPGNKSFILPILQFFSLDYNYFTGQIPLGLAACRHLKVFSL 302

序列 1: 312  PYNLFEGVLPPWLGRLTNLDAISLGGNNFDAGPIPTELSNLTMLTVLDLTTCNLTGNIPA 371
               NL EG LP WLG+LT L+ ISLG N    GPI   LSNLTML  LDL  CNLTG IPA
序列 2: 303  LDNLIEGPLPSWLGKLTKLNVISLGENLLVVGPIRDALSNLTMLNFLDLAMCNLTGAIPA 362
```

序列比对的用途

- 分析功能
- 分析物种进化
- 检测突变、插入或缺失
- 序列延长
- 序列定位
- 基因表达谱分析

序列对位排列分析的种类

- 两序列对位排列分析
- 序列对库对位排列分析
  - ✓ 从数据库中寻找同源序列
  - ✓ 主要涉及核苷酸数据库和蛋白质数据库
- 多序列对位排列分析

（一）序列对位排列分析的基本原理

1、记分矩阵（scoring matrix）

- 记分矩阵中含有两条序列对位排列时具体使用的分值
- 长度一定时，分数越高，两条序列匹配越好

DNA序列对位记分

序列1	A	C	G	T	T	A	
序列2	A	C	T	T	T	G	
记分	2	2	-3	2	2	-3	=2

1、记分矩阵（scoring matrix）

- 蛋白质序列对位排列分析记分复杂
- 一致氨基酸的记分不同
  - 稀有氨基酸（C），分值高
  - 普通氨基酸（S），分值低
- 相似氨基酸也记分，如R-K

蛋白质序列对位记分

序列1	V	D	S	C	Y	
序列2	V	N	W	C	Y	
记分	4	1	-3	9	7	=18

1、记分矩阵（scoring matrix）

- 蛋白质有多种记分矩阵
  - **PAM矩阵（如PAM30、PAM70）**
  - **BLOSUM矩阵（如BLOSUM62、BLOSUM80）**

BLOSUM62 amino acid scoring matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

BLAST默认scoring matrix

2、空位（间隔）罚分（gap penalty）

- 基因进化过程中产生突变
  - 插入
  - 缺失

Indel

		A	T	G	T	G	A	
序列1	A	T	G	C	T	G	A
序列2	A	T	G	G	A		

- 序列对位排列分析时允许插入空位
- 空位罚分涉及两个参数
  - 空位开放（gap opening）
  - 空位延伸（gap extension）

序列1	A	T	G	C	T	G	A	
序列2	A	T	G	–	–	G	A	
	2	2	2	-5	-2	2	2	= 3

（二）序列对库对位排列分析

- 用待分析序列对数据库进行相似性分析
- 重复许多次的两两序列对位排列分析
- 从数据库中找出所有同源序列
- 主要检索体系
  - **BLAST**
  - **FASTA**
  - **Other methods**

1、基本概念

（1）Sequence identity和sequence similarity

Identity: 两条序列在同一位点上的核苷酸或氨基酸残基完全相同

The extent to which two (nucleotide or amino acid) sequences are invariant.

Similarity (positive): 两条序列在同一位点上的氨基酸残基的化学性质相似

The extent to which nucleotide or protein sequences are related. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation. In BLAST similarity refers to a positive matrix score

Homology

A is 80% identical to B

Identity

相同

Similarity

相似

A is 80% similar to B

Homology

同源

A is 80% homologous to B

If your sequences are more than 100 amino acids long (or 100 nucleotides long), you can label proteins as "homologous" if 25 percent of the amino acids are identical, for DNA you will require at least 70 percent identity

（2）Global alignment 和 local alignment

Global alignment: 两条完整的序列相比较

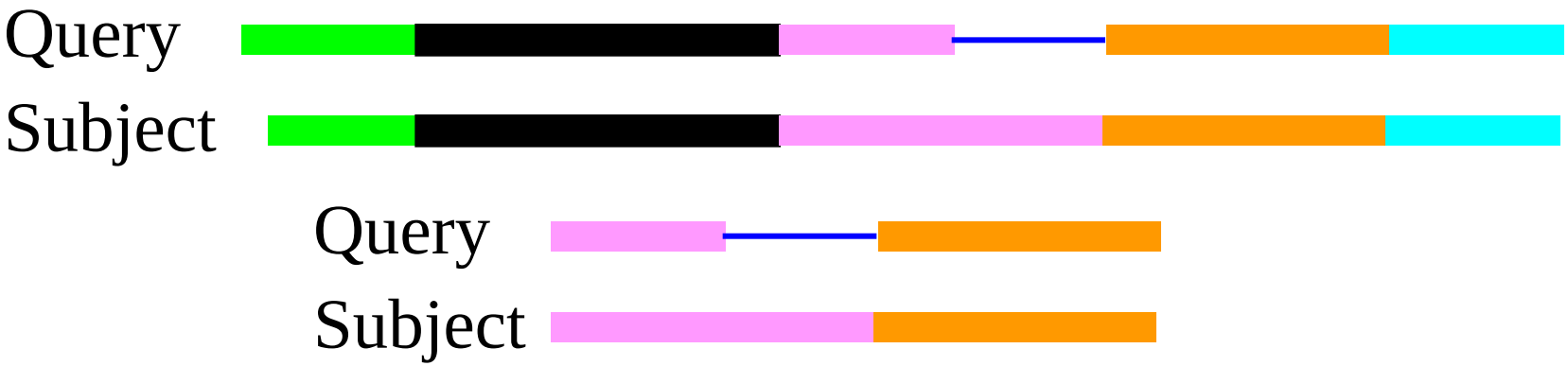
Local alignment: 两条序列中相似程度最高的部分相比较


```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

（3）Gapped alignment 和 ungapped alignment

Gapped alignment: 为达到最佳 **alignment**，序列中加入空位



Ungapped alignment： 相比较序列的核苷酸或氨基酸序列连续

（4）Alignment score 和 E (expect) value

衡量两条相比较序列相似程度的标准

(bits) Score: 分值越大，两个比较序列相似程度越高

E value: 期望得到的、完全由机会造成的、相当于或大于目前分值的alignment 次数

试验组存活率比对照组高20%（$p<0.05$）

- E值取决于 alignment 分值、相比较序列的长短和数据库中数据的数量
- Blast中E的阈值为10。$1e-66 = 1 \times 10^{-66}$
  E 值越小越好

（5）Low-complexity regions (LCRs)

核苷酸和蛋白质序列中短的重复序列或由少数几种核苷酸或氨基酸残基组成的序列（如 **Poly-A**）

- 数据库中半数以上的序列至少带有一个**LCR**
- **Sequence alignment** 时应避免 **LCR** 相互配对得分
- **BLAST** 用 **Filter** 功能避免比较 **LCR**

  - 在比对结果的**query**序列中用小写字母或**x**和**n**（分别代表氨基酸和核苷酸）代表 **LCR**

2. BLAST (Basic Local Alignment Search Tool) 检索

http://blast.ncbi.nlm.nih.gov/

Help

Basic BLAST

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search using SNP flanks
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA transcript and genomic libraries
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene
- Search WGS sequences grouped by organism

BLAST programs

blastn 用核苷酸序列检索核苷酸数据库

blastp 用蛋白质序列检索蛋白质数据库

blastx 将核苷酸序列通过 **6** 种阅读框翻译成不同的蛋白质序列检索蛋白质数据库

tblastn 用蛋白质序列检索核苷酸数据库（数据库中的序列被翻译出不同的蛋白质序列）

tblastx 将核苷酸序列通过 **6** 种阅读框翻译成不同的蛋白质序列检索核苷酸数据库（数据库中的序列也被翻译出不同的蛋白质序列）

BLAST databases

Human genomic plus transcript	人基因组和mRNA序列
Mouse genomic plus transcript	小鼠基因组和mRNA序列
nucleotide collection (nr/nt)	GenBank (无 EST, STS, GSS, HTGS)
non-redundant protein sequences (nr)	非冗余蛋白质数据库
refseq-rna	Reference mRNA sequences
refseq-genomic	Reference genomic sequences
refseq-protein	Reference protein sequences
est	EST 数据库

BLAST databases

est-others	非人和小鼠的EST数据库
gss	GSS 数据库
htgs	HTGS 数据库
pat	专利序列数据库
pdb	蛋白质三维结构数据库
alu_repeats	Alu 重复序列数据库
swissprot	swissprot蛋白质数据库
dbsts	STS 数据库
wgs	whole-genome shotgun reads
env_nt	Environmental samples (nt)
env_nr	Environmental samples (pro)

（1）BLASTN

- 将要查询的序列直接粘贴到序列框中或输入登陆号，GI 号
- 选择 database、organism
- 选择 Blast Algorithm
- 可进行其它项目的选择用于分析
  - 进一步选择检索范围：Limit by entrez query (如 protease NOT hivI [organism])
  - Filter (Human repeats)：遮盖重复序列可加快检索速度（特别是 > 100 kb 的片段）
  - 结果页面

BLAST结果解读

Sequence

Job Title: AB077992:Homo sapiens mRNA for AD24, complete...

Request ID	**HWH2UZHU01S**
Status	Searching
Submitted at	Tue Dec 8 07:19:38 2009
Current time	Tue Dec 8 07:19:44 2009
Time since submission	00:00:06

This page will be automatically updated in **37** seconds

Edit and Resubmit Save Search Strategies ▷Formatting options ▷Download

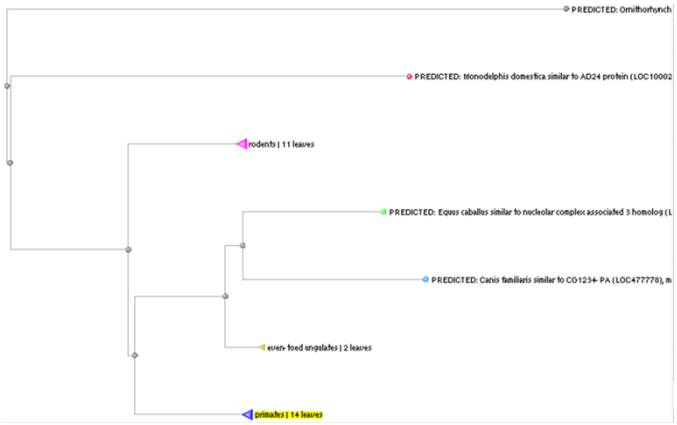
AB077992:Homo sapiens mRNA for AD24, complete...

Query ID gi|18389432|dbj|AB077992.1|
Description Homo sapiens mRNA for AD24, complete cds
Molecule type nucleic acid
Query Length 2450

Database Name nr
Description All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
Program BLASTN 2.2.25+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results]

```
Mammalia          [mammals]
. Theria            [mammals]
. . Eutheria          [placentals]
. . . Euarchontoglires [placentals]
. . . . Catarrhini       [primates]
. . . . . Hominidae        [primates]
. . . . . . Homininae        [primates]
. . . . . . . Homo sapiens (man) ------------------ 4479 31 hits [primates]
. . . . . . . Pan troglodytes ..................... 4451  2 hits [primates]
. . . . . . . Pongo abelii (Orang-utan) ------------- 4318  2 hits [primates]
. . . . . Macaca mulatta (rhesus macaque) --------- 2368  2 hits [primates]
. . . . Cricetulus griseus (Chinese hamsters) ----- 2791  1 hit  [rodents]
. . . . Mus musculus (mouse) ...................... 2580 14 hits [rodents]
. . . . Rattus norvegicus (brown rat) ............. 2562  3 hits [rodents]
. . . Equus caballus (equine) --------------------- 3469  1 hit  [odd-toed ungulates]
. . . Sus scrofa (wild boar) ...................... 3273  2 hits [even-toed ungulates]
. . . Canis lupus familiaris (dogs) ............... 3273  1 hit  [carnivores]
. . . Bos taurus (cow) ............................ 3208  2 hits [even-toed ungulates]
. . Monodelphis domestica ------------------------ 2097  1 hit  [marsupials]
. Ornithorhynchus anatinus (duck-billed platypus) -  876  1 hit  [monotremes]
```



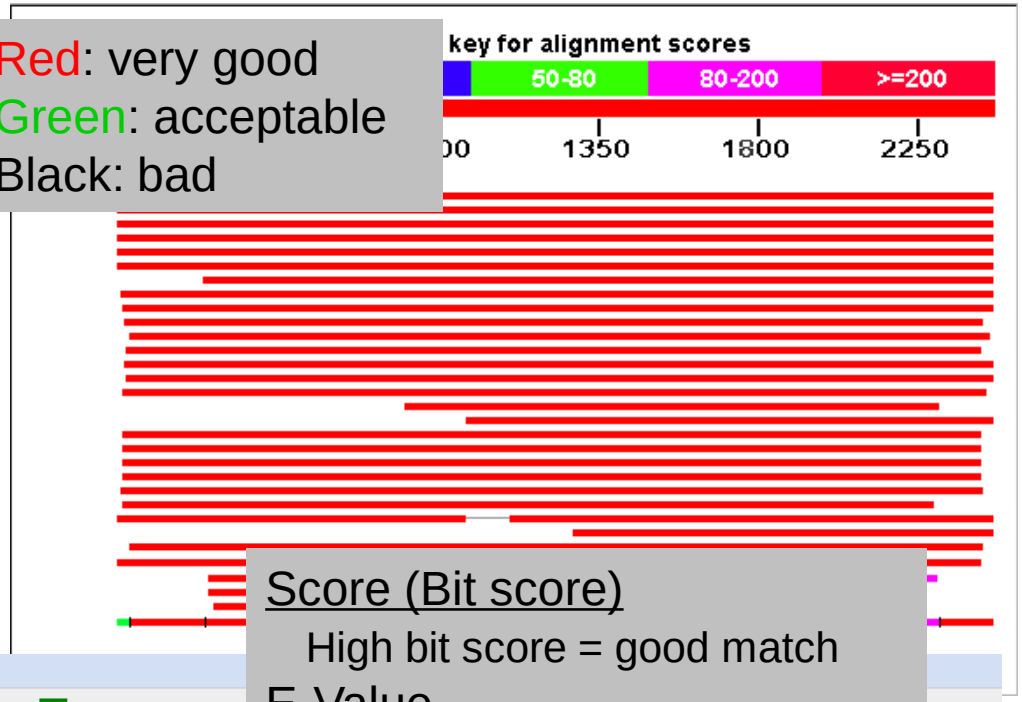
BLAST结果解读

1e-03 = borderline E-value
1e-04 = good E-value
1e-10 = very good E-value

E-values lower than 1e-4 indicate possible homology

E-values higher than 1e-4 require extra evidence to support homology

Red: very good
Green: acceptable
Black: bad



Score (Bit score)
High bit score = good match

E-Value
Low E-value = good match

Hit list

▼ Descriptions

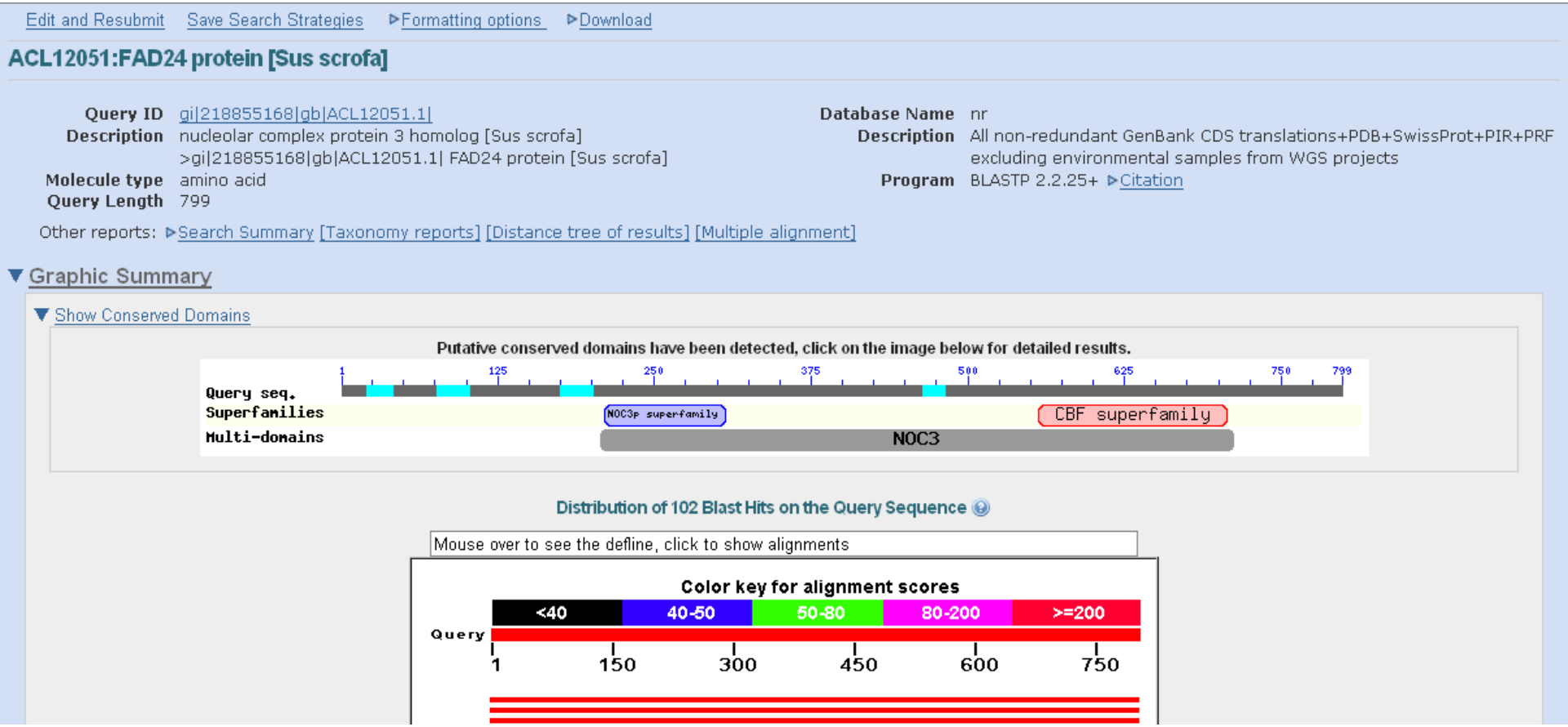
Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
AB077992.1	Homo sapiens mRNA for AD24, complete cds	4479	4479	99%	0.0	100%	U G
AK292520.1	Homo sapiens cDNA FLJ77615 complete cds, highly similar to Homo :	4468	4468	99%	0.0	99%	U G
NM_022451.9	Homo sapiens nucleolar complex associated 3 homolog (S. cerevisiae	4468	4468	99%	0.0	99%	U E G
XM_507934.2	PREDICTED: Pan troglodytes nucleolar complex associated 3 homolo	4451	4451	99%	0.0	99%	G
AK091246.1	Homo sapiens cDNA FLJ33927 fis, clone CTONG2017429	4436	4436	99%	0.0	99%	U G
NM_001133981.1	Pongo abelii nucleolar complex associated 3 homolog (S. cerevisiae) (	4318	4318	99%	0.0	98%	U G
XM_001149829.1	PREDICTED: Pan troglodytes nucleolar complex associated 3 homolo	4061	4061	90%	0.0	99%	G
XM_001917312.1	PREDICTED: Equus caballus similar to nucleolar complex associated 3	3469	3469	99%	0.0	92%	U G
NM_001144115.1	Sus scrofa nucleolar complex associated 3 homolog (NOC3L), mRNA	3273	3273	98%	0.0	91%	U G
AL355341.19	Human DNA sequence from clone RP11-146P21 on chromosome 10	394	4561	98%	1e-105	100%	
AF086377.1	Homo sapiens full length insert cDNA clone ZD69A07	243	243	5%	5e-60	99%	U E
AC158131.4	Mus musculus chromosome 19, clone RP24-398D6, complete sequer	235	663	19%	8e-58	96%	
AC101775.10	Mus musculus chromosome 19, clone RP24-116B5, complete sequer	182	329	8%	1e-41	96%	
AK089396.1	Mus musculus B6-derived CD11 +ve dendritic cells cDNA, RIKEN full-	182	182	5%	1e-41	92%	U E G

（2）BLASTP

◆ 基本操作同 blastn



（3） Translated BLAST

- **blastx, tblastn, tblastx**
- **基本操作同 blastn**

```
>ref|XP_392682.2| PREDICTED: similar to C-terminal-binding protein (CtBP protein)
(dCtBP) isoform 1 [Apis mellifera]
Length=472

 GENE ID: 409158 LOC409158 | similar to C-terminal-binding protein (CtBP
protein) (dCtBP) [Apis mellifera] (10 or fewer PubMed links)

 Score =  314 bits (804),  Expect = 2e-84
 Identities = 155/155 (100%), Positives = 155/155 (100%), Gaps = 0/155 (0%)
 Frame = +1

Query  1    MMPKRPRMDNLRGPIANGPIQTRPLVALLDGRDCSIEMPILKDVATVAFCDAQSTSEIHE  180
            MMPKRPRMDNLRGPIANGPIQTRPLVALLDGRDCSIEMPILKDVATVAFCDAQSTSEIHE
Sbjct  6    MMPKRPRMDNLRGPIANGPIQTRPLVALLDGRDCSIEMPILKDVATVAFCDAQSTSEIHE  65

Query  181  KVLNEAVGALMWHTIILTKEDLEKFKTLRIIVRIGSGVDNIDVKAAGELGIAVCNVPGYG  360
            KVLNEAVGALMWHTIILTKEDLEKFKTLRIIVRIGSGVDNIDVKAAGELGIAVCNVPGYG
Sbjct  66   KVLNEAVGALMWHTIILTKEDLEKFKTLRIIVRIGSGVDNIDVKAAGELGIAVCNVPGYG  125

Query  361  VEEVADTTLCLILNLYRRTYWLANMVREGKKFTGP  465
            VEEVADTTLCLILNLYRRTYWLANMVREGKKFTGP
Sbjct  126  VEEVADTTLCLILNLYRRTYWLANMVREGKKFTGP  160
```

(4) Conserved Domain Search

- 检索conserved domain database
- 只适用于蛋白质序列的检索分析
- 检测被检索的序列中是否含有保守结构域

点击“Search for similar domain architectures”查看相关结构域

点击结构域图标查看多序列对位排列

（5） Primer-BLAST

http://www.ncbi.nlm.nih.gov/tools/primer-blast/

- 设计PCR引物
- 分析引物特异性
- 在GenBank检索结果页面中提供了链接
- 结果

3、FASTA 检索

http://www.ebi.ac.uk/Tools/sss/

- **Programs**

BLAST 和 FASTA 检索体系有时不能检测出某些远缘序列的相关性

一些特殊设计的序列检索体系在发现基因和蛋白质家族成员方面可能更为可靠

FASTA

FASTA Sequence Similarity Search using the FASTA program. This tool is available for the following databases:

Protein | Nucleotide | Proteomes | Genomes | Whole Genome Shotgun | ASD Protein | ASD Nucleotide | LGIC Protein | LGIC Nucleotide

SSEARCH Sequence Similarity Search using the SSEARCH program. This tool is available for the following databases:

Protein | Nucleotide | Proteomes | Genomes | Whole Genome Shotgun | ASD Protein | ASD Nucleotide | LGIC Protein | LGIC Nucleotide

PSI-Search PSI-Search combines the sensitivity of the Smith-Waterman search algorithm (SSEARCH) with the PSI-BLAST (blastpgp) iterative profile construction strategy to find distantly related protein sequences.

Launch PSI-Search

GGSEARCH GGSEARCH performs a sequence search using alignments that are global in the query and global in the database (Needleman-Wunsch).

Protein | Nucleotide

GLSEARCH GLSEARCH performs a sequence search using alignments that are global in the query and local in the database.

Protein | Nucleotide

FASTM Peptide similarity searching using the FASTM/FASTS/FASTF programs. This tool is available for the following databases:

Protein | Nucleotide | Proteomes | ASD Protein | LGIC Protein

（三）两序列对位排列分析

Specialized BLAST

Align two (or more) sequences using BLAST (bl2seq)

Needleman-Wunsch Global Sequence Alignment Tool

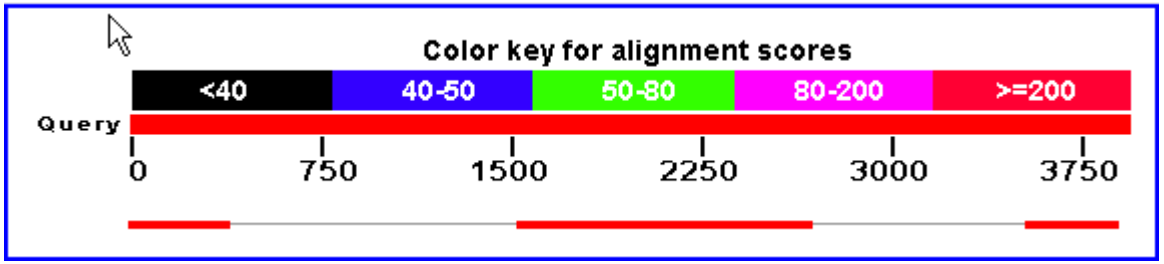
- **NCBI的分析工具**
- 对任意两条序列进行对位排列分析
- 允许空位

BLAST 2 sequences

- 序列来源
  - 输入 **Accession number**
  - 直接粘贴序列
- 适用于 **blastn, blastp, blastx, tblastn, tblastx**
  - **blastn:** 两条核苷酸序列相比较
  - **blastp:** 两条蛋白质序列相比较
  - **tblastn:** 比较蛋白质序列（**sequence 1**）和核苷酸序列（翻译成蛋白质序列）（**sequence 2**）
  - **blastx:** 比较核苷酸序列（翻译成蛋白质序列）（**sequence 1**）和蛋白质序列（**sequence 2**）
  - **tblastx:** 两条核苷酸序列（翻译成蛋白质序列）比较

BLAST 2 sequences

- 结果格式
  - 两种图形
  - 两序列对位排列




```
 Score =  955 bits (517),  Expect = 0.0
 Identities = 989/1207 (81%), Gaps = 71/1207 (5%)
 Strand=Plus/Plus

Query  1519  AATAACTTGGAGGGATCAATACCACAAGAAATAGGGCATCTCAAAAATCTAGTAGAATTT  1578
             ||||||||||||||||||||||| |||||||||||  | || |||||| | || |||||
Sbjct  6205  AATAACTTGGAGGGATCAATACCAAAAGAAATAGGGAAACTTAAAAATATTGTCGAATTC  6264

Query  1579  CATGCAGAATCGAATAGATTATCAGGTAAAATCCCTAACACGC-TTGGTGATTGCCAGC-  1636
             ||||| || ||||| | |||||| ||| | ||||||| ||| | ||||||| ||||| |
Sbjct  6265  CATGCTGATTCGAACAAATTATCGGGTGAGATCCCTAGCAC-CATTGGTGAATGCCAACT  6323

Query  1637  TCTTACGGTATCTTTAT-CTGCAAAATAATTTGTTATCTGGTAGCATCCC-ATCAGCCT-  1693
             || | | | |||||| | |||||||| ||||| |||  ||||||||||| || || ||
Sbjct  6324  TC-TGCAGCATCTTT-TCCTGCAAAACAATTTCTTAAATGGTAGCATCCCAAT-AG-CTC  6379

Query  1694  TGGGTCAGCTGAAAGGTCTCGAAACTCTTGATCTCTCAAGCAACAATTTGTCAGGCCAGA  1753
             ||  |||| |||||||||| || || ||||||||||| |||||||||||||||| ||||
Sbjct  6380  TGACTCAGTTGAAAGGTCTGGACACACTTGATCTCTCAGGCAACAATTTGTCAGGTCAGA  6439
```