

# 生物信息学

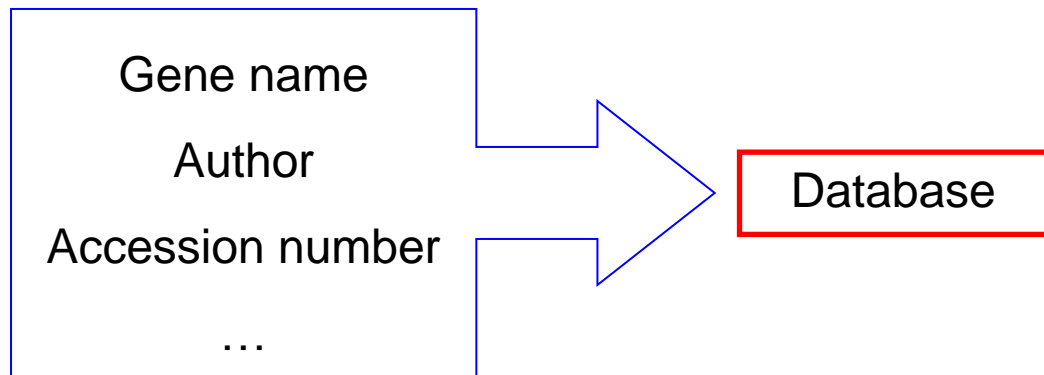
## 数据库检索与序列比对

张礼斌 华中科技大学生命科学与技术学院  
Email: libinzhang@hust.edu.cn

# 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  
GKI VAITALSEKKGGFEVSIEKA-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 (美国国家生物技术信息中心)**  
**Sequence Retrieval System (SRS) (欧洲生物信息研究所)**  
**Integrated database retrieval system (DBGET) (日本GenomeNet)**

# 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

PubMed

rice AND microarray OR expression profile

rice AND (microarray OR expression profile)

## 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  
“16S rRNA”

### ◆ \*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...

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

<https://www.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

- ❖ 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**

### 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

**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**

# 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**

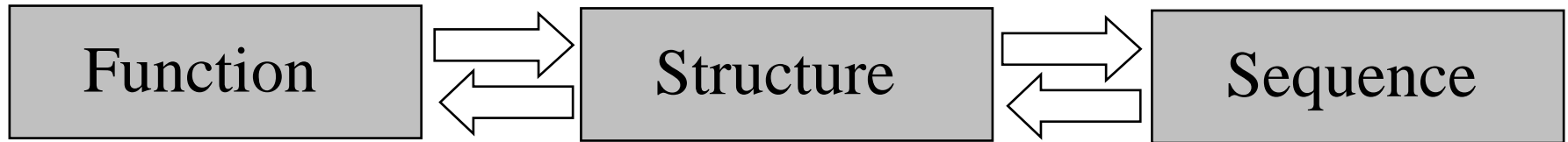
- ◆ **不同的数据库包含内容有限**

- ◆ **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**

<i>1-Letter Code</i>	<i>Nucleotide Name</i>	<i>Category</i>
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 NYLTGSI PDDLFNNTPLLTYLNVGNNSLSGLIPGCI GSLPI LQHLNFAQNNLTGAVPPAI 251

NYLTG IP+ LFNNTPL +L +GNNSLSG IP CIGSLP+L+ L Q NNLTG VPP+I

序列 2: 183 NYLTGLIPNGLFNNTPSLKHLIIGNNSLSGPIPSCIGSLP LLERLV LQGNLTGPVPPSI 242

序列 1: 252 FNMSKLSTISLISNGLTGPIPGNTSFSLPVLRHFAISKNNFFGQIPLGLAACPYLQVIAM 311

FNMS+L I+L SNGLTGPIPGN SF LP+L++F++ N F GQIPLGLAAC +L+V ++

序列 2: 243 FNMSRLHVI ALASNGLTGPIPGNKSFI LPI LQFFSLDYNYFTGQIPLGLAACRHLKVFSL 302

序列 1: 312 PYNLFEGVLPPWLGRLTNLDAISLGGNNFDAGPIPTELSNLTMLTVLDLTTCNLTGNI PA 371

NL EG LP WLG+LT L+ ISLG N GPI LSNLTML LDL CNLTG IPA

序列 2: 303 LDNLI EGPLPSWL GKLTCLNVISLGENLLVVGPI RDALSNLTMLNFDLAMDNLTGAI PA 362

# 序列比对的应用

- ❖ 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



# 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

	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

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**
- ◆ **两两序列对位排列分析**
- ◆ **从数据库找出所有同源序列**
- ◆ **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

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

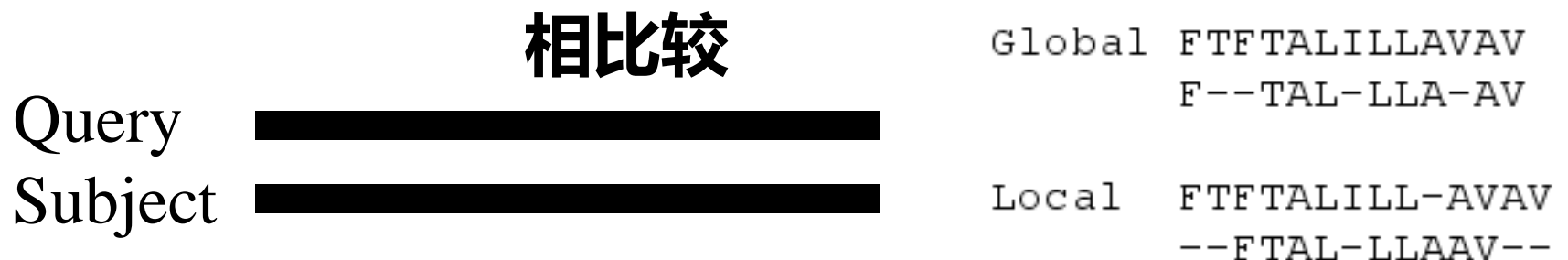
## (2) Global alignment 和 local alignment



**Global alignment: 两条完整的序列相比较**



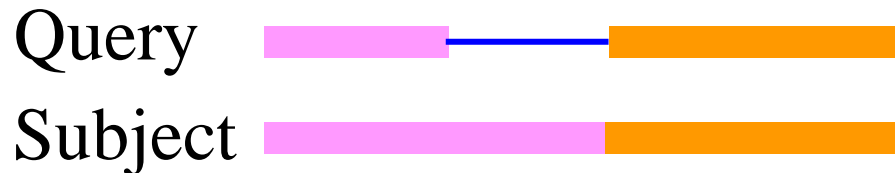
**Local alignment: 两条序列中相似程度最高的部分相比较**



### (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

## 2. BLAST (Basic Local Alignment Search Tool) 检索

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

### Basic BLAST

<a href="#">nucleotide blast</a>	Search a <b>nucleotide</b> database using a <b>nucleotide</b> query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
<a href="#">protein blast</a>	Search <b>protein</b> database using a <b>protein</b> query <i>Algorithms: blastp, psi-blast, phi-blast</i>
<a href="#">blastx</a>	Search <b>protein</b> database using a <b>translated nucleotide</b> query
<a href="#">tblastn</a>	Search <b>translated nucleotide</b> database using a <b>protein</b> query
<a href="#">tblastx</a>	Search <b>translated nucleotide</b> database using a <b>translated nucleotide</b> 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

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

## **(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**

## (2) BLASTP

### ◆ 基本操作同 blastn

[Edit and Resubmit](#) [Save Search Strategies](#) [▶Formatting options](#) [▶Download](#)

#### ACL12051:FAD24 protein [Sus scrofa]

**Query ID** [gi|218855168|gb|ACL12051.1|](#)  
**Description** nucleolar complex protein 3 homolog [Sus scrofa]  
>gi|218855168|gb|ACL12051.1| FAD24 protein [Sus scrofa]  
**Molecule type** amino acid  
**Query Length** 799

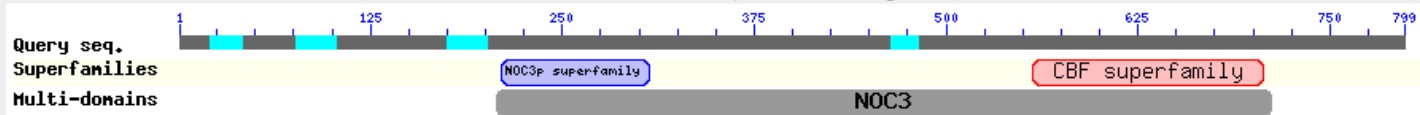
**Database Name** nr  
**Description** All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF  
excluding environmental samples from WGS projects  
**Program** BLASTP 2.2.25+ [▶Citation](#)

Other reports: [▶Search Summary](#) [\[Taxonomy reports\]](#) [\[Distance tree of results\]](#) [\[Multiple alignment\]](#)

#### ▼ Graphic Summary

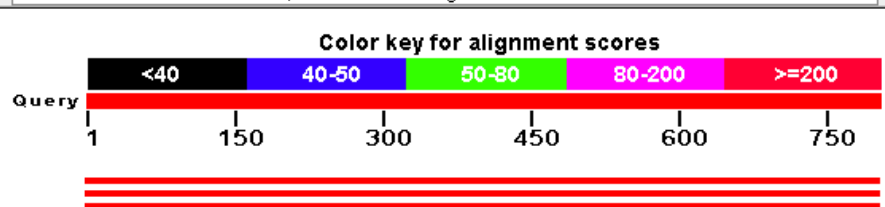
##### ▼ [Show Conserved Domains](#)

Putative conserved domains have been detected, click on the image below for detailed results.



##### Distribution of 102 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments



### **(3) 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**

举例

>序列

ATGGCCATGGGTCCACACGCAGTGAGATGAATGCTAGATCTCACGAGAAAAAAGAA  
ATACATCTCAGGGGGTTGTGATGTACTGGATAATTTGCTCGTCATATTAACCATTAGCTT  
ACTCTAGTTGATGTGGGCATGGATGGAGCCGGCAGCCGGCGATCCTATTAA ...

>序列

MALVRLPVWIFVAALLIASSSTVPCASSLGPIASKSNSDDTLAALLAFKAQLSDPNN  
ILAGNWTTGTPF CRWVGVSCESSHRRRRQRVTALELPNVPLQGELSS...

序列

mngtldhpdqpdl daikmfvgqvprt wsekdlrel feqygavyeinvl rdrsqnppq skgccfvtfytrkaaleaqnalhnmkvl  
pgmhhpiqmkpadseknn avedrklfigmisk kctendirvmfssfgqieecrilrgpdglsrc

序列

mngtldhpdqpdl daikmfvgqvprt wsekdlrel feqygavyeinil rdrsqnppaskgccfvtfytrkalleaqnalhnmkvlp  
gmhipiqmapadseknnavedrklfigmisk kctendirvmfssfgqieecrilrgpdglsrc

序列

gttcgttgcaacaaattgatgagcaatgctttttataatgccaaactttgtacaaaaaag

### 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)

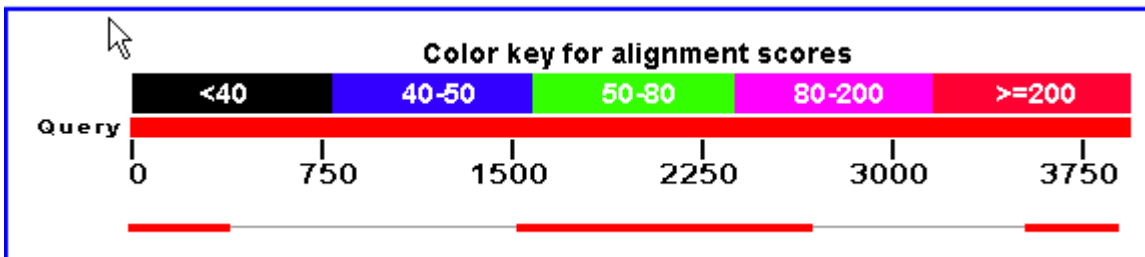
- ❖ tblastx: 两条核苷酸序列 (翻译成蛋白质序列) 比较

## 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	AATAACTTGGAGGGATCAATACCACAAGAAATAGGGCATCTCAAARATCTAGTAGAATT	1578
Sbjct	6205	AATAACTTGGAGGGATCAATACCAAGAAATAGGGAACCTAAAAATATTGTGCAATTC	6264
Query	1579	CATGCAGAATCGAATAGATTATCAGSTAAAATCCCTAACACGC-TTGGTGATTGCCAGC-	1636
Sbjct	6265	CATGCTGATTGAACAAATTATCGGGTGAGATCCCTAGCAC-CATTGGTGAATGCCAACT	6323
Query	1637	TCTTACGGTATCTTTAT-CTGCAAAATAATTTGTTATCTGGTAGCATCCC-ATCAGCCT-	1693
Sbjct	6324	TC-TGCAGCATCTTT-TCCTGCAAAACAATTTCTTAATGGTAGCATCCAAT-AG-CTC	6379
Query	1694	TGGGTCAGCTGAAAGGTCTCGAAACTCTTGATCTCTCAAGCAACAATTTGTGAGGCCAGA	1753
Sbjct	6380	TGACTCAGTTGAAAGGTCTGGACACACTTGATCTCTCAGGCAACAATTTGTGAGGTGAGA	6439

# 上机

在NCBI中使用Entrez检索系统用ACE2（angiotensin-converting enzyme 2）搜索，观察所有数据库的响应，分别有多少核酸序列（Nucleotide），多少基因序列（gene），多少蛋白质序列（protein），多少论文（pubmed）？最新一篇论文是什么？

点击Nucleotide，然后分别查看Homo sapiens和Mus musculus的ACE2前2条核酸序列（fasta格式），任选一条Homo sapiens和Mus musculus的ACE2核酸序列，点击进入，查看它编码蛋白质的信息，并在uniprot中查看它的功能信息和蛋白质三级结构信息。

分别用Homo sapiens的ACE2核酸序列与蛋白质序列进行blastn与blastp比对分析