

生物信息分析入门基础知识



目录

- 01 [生物信息学常见数据格式](#)
Fastq,fasta,gbk,gff,sam/bam,vcf
- 02 [生物信息学常用数据库](#)
Genbank , NCBI Refseq, Uniprot, SWISS-PROT, PDB, KEGG
- 03 [生信的重要基础--序列联配](#)
Blast系列运用
- 04 [微生物基因组学, 比较基因组学](#)



➤ 生物信息分析入门基础知识

1 生物信息学常见数据格式

[返回目录](#)



➤ 生物信息学常见数据格式

- 存储序列
fasta, fastq
- 展示注释信息
gbk, gff3
- 比对结果显示的文件
sam、bam
- 表示突变信息
vcf



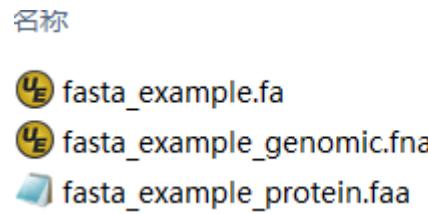
➤ 生物信息学常见数据格式

• 存储序列

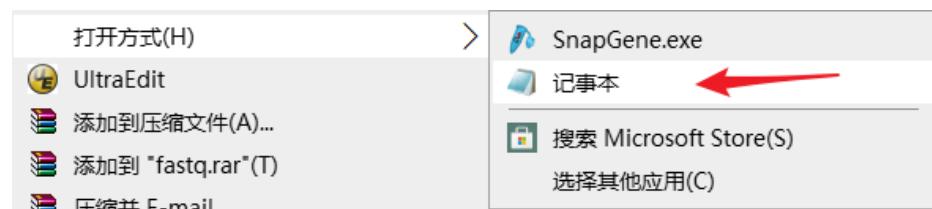
fasta

格式是最基本的表示序列信息（核苷酸或者蛋白质）的格式。

这里简单介绍下，fasta格式的文件通常后缀名为.fasta 或者.fa， 其实这都无所谓，因为都是**文本文件**。fasta格式文件（可以包含多条序列）中的一条序列的通常表示方法如下：



fasta格式的文件通常后缀名为.fasta, .fna 或者.fa,
后缀只是提示文件内容的格式只要满足两行序列格式的其实都是
fasta格式。



写字板，记事本查看文本文件

```

fasta_example.fa - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
>seq_id_1 sequencing_raw_data_read_1
CAACTAATAATAGACCTTGTGTATGTATATGTTGCAATGGTGGAAA
>seq_id_2 Illumina_sequencing_raw_data_read_2
TGATTACGCTGTTCCAATCTTGTGTCACTCATCATGAAGCACAATTG
~seq_id_3

```

fasta文件格式

两行格式, 第一行 “>” 符号起始到**第一个空格之前**是序列名称,一般
在同一个fasta文件中要求是**唯一的**. 其后是对于序列的简单描述.
第二行就是序列的详细碱基组成.
同一个fasta文件可以包含一个或多个序列记录



➤ 生物信息学常见数据格式

• 存储序列

fasta

细菌基因组fasta格式

基因组的登录号 (accession number) 基因组的描述信息

```
>NC_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome
AGCTTTCAATTCTGACTGCAACGGGAAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAAGTG
GTTACCTGCCGTGAGTAAATTAAAATTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGAC
AGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCAAGGT
AACGGTGGGGCTGACGCGTACAGGAAACACAGAAAAAGCCCGACCTGACAGTGGGGCTTTTTTCGACCAAAGG
TAACGAGGTAACAACCATGCGAGTGTGAAGTTCGGCGTACATCAGTGGCAAATGCAGAACGTTCTGCGTGTGCG
ATATTCTGGAAAGCAATGCCAGGCAGGGCAGGTGGCCACCGTCCTCTGCCCGCCAAAATACCAACCACCTGGTG
GCGATGATTGAAAAAACCATAGCGGCCAGGATGCTTACCCAATATCAGCGATGCCAACGTATTTGCCAACCTTT
GACGGGACTCGCCGCCAGCCGGGTTCCCGCTGGCGCAATTGAAAATTCGTCGATCAGGAATTGCCCAAATAA
AACATGTCCTGCATGGCATTAGTTGTTGGGGCAGTGGCGGATAGCATCACGCTGCGCTGATTGCCGTGGCGAGAAA
```

fasta_example_genomic.fna



➤ 生物信息学常见数据格式

• 存储序列

fasta

蛋白质的fasta格式

蛋白质的ID

蛋白质的功能描述

蛋白质的所属物种

```
>NP_414542.1 thr operon leader peptide [Escherichia coli str. K-12 substr. MG1655]
```

MKRISTTITTTITTTGAG

```
>NP_414543.1 fused aspartate kinase/homoserine dehydrogenase 1 [Escherichia coli str. K-12 substr. MG1655]
```

MRVLKFGGTSVANAERFLRVADILESNDARQGVATVLSAPAKITNHLVAMIEKTISGQDALPNISDAERIFAELLTGAA
AQPGFPLAQLKTFVDQEFAQIKHVLHGTLGQCPDSINAALICRGEKMSIAIMAGVLEARGHNVTVIDPVEKLAVGHY
LESTDVIAESTRRIAASRIPADHMVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTVDGVYTCDPRQV
PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRDEDELVKGISNLNNMAMFSV
SGPGMKGMVGMAARVFAAMSRARISVVLITQSSEYSISFCVPQSDCVRAERAMQEEFYLELKEGLLEPLAVTERLAIIS
VVGDGMRTL RGISAKFFAALARANINIVIAQGSSERSISVVNNDDATTGVRVTHQMLFNTDQVIEVFVIGVGGVGGAL
LEQLKRQQSWLKNKHIDLRCVGVANSKALLTNVHGLNLENWQEELAQAKEPFNLGRLIRLVKEYHLLNPVIVDCTSSQAV
ADQYADFLREGFHVVTPNKKANTSSMDYYHQLRYAAEKSRRKFLYDTNVGAGLPVIENLQNLLNAGDELMKFSGILSGSL
SYIFGKLDEGMSFSEATTLAREMGYTEPDPRDDLGMVDARKLLILARETGRELADIEIEPVLPAAEFNAEGDVAAFMA
NLSQLDDLFAARVAKARDEGKVLRVGNIDEDGVCRVKIAEVDGNDPLFKVKNGENALAFYSHYYQPLPLVLRGYGAGND
VTAAGVFADLLRTLSWKLGV

fasta_example_protein.faa



➤ 生物信息学常见数据格式

• 存储序列

fastq

文本形式来存储序列信息的格式，后缀名通常为.**fastq**或者**.fq**，区别于fasta，它除了存储序列本身外还存储了序列中每个碱基所对应的**质量分数**，所以fastq格式通常用于高通量测试数据的存储，是**高通量测序的标准格式**。

fastq格式文件中一个完整的单元分为四行，含义如下：

第一行：以@开头，内容同fasta的描述行类似

第二行：具体的碱基序列

第三行：以+开头，后面的内容可以和第一行类似，也什么都没有只留+

第四行：以ASCII字符集（分数）编码来表示对应碱基的测序质量(**后续软件读取判断碱基测序质量**)

```
*fastq.fq - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
@FCH2NFHBCX3:1:1214:13051:9728#CCAGCGCT_TCTTTCCC/1
ATATCTTTCCAACAATAATAGACCTTGTGTATGTATGTTGCAATGGTGGAAATTCAAC
+
ccccchhhhhhhhhhhheffgghhghhhhhhhggghhhhhhhhhghghhhhhhhhh
```

fastq格式



➤ 生物信息学常见数据格式

• 展示注释信息

Genbank格式

GenBank是美国国立卫生研究院维护的基因序列数据库，汇集并注释了所有公开的核酸以及蛋白质序列。

<https://www.ncbi.nlm.nih.gov/genbank/>

每个记录代表了一个单独的、连续的、带有注释的DNA或RNA片段。

便于人类阅读

LOCUS	NC_000913	4641652 bp	DNA	circular	CON 11-OCT-2018
DEFINITION	Escherichia coli str. K-12 substr. MG1655, complete genome.				
ACCESSION	NC_000913				
VERSION	NC_000913.3				
DBLINK	BioProject: PRJNA57779 BioSample: SAMN02604091 Assembly: GCF_000005845.2 RefSeq.				
KEYWORDS	Escherichia coli str. K-12 substr. MG1655				
SOURCE	Escherichia coli str. K-12 substr. MG1655				
ORGANISM	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.				
REFERENCE	1 (bases 1 to 4641652)				
AUTHORS	Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T., Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H., Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.				
TITLE	Escherichia coli K-12: a cooperatively developed annotation snapshot--2005				
JOURNAL	Nucleic Acids Res. 34 (1), 1-9 (2006)				
PUBMED	16397293				
REMARK	Publication Status: Online-Only				
COMMENT	PROVISIONAL REFSEQ: This record has not yet been subject to final NCBI review. The reference sequence is identical to U00096. On Nov 3, 2013 this sequence version replaced NC_000913.2. Changes to proteins and annotation made on September 24, 2018. Current U00096 annotation updates are derived from EcoCyc https://ecocyc.org/. Suggestions for updates can be sent to biocyc-support@ai.sri.com. These updates are being generated from a collaboration that includes EcoCyc, the University of Wisconsin, UniProtKB/Swiss-Prot, and the National Center for Biotechnology Information (NCBI). COMPLETENESS: full length.				
FEATURES	Location/Qualifiers				
source	1..4641652 /organism="Escherichia coli str. K-12 substr. MG1655" /mol_type="genomic DNA" /strain="K-12" /sub_strain="MG1655" /db_xref="taxon:511145"				
gene	190..255 /gene="thrL" /locus_tag="b0001" /gene_synonym="EC0001" /db_xref="ASAP:ABE-0000006" /db_xref="ECOCYC:EG11277"				
	(省略若干行...)				
ORIGIN	1 agctttcat tctgactgca acggcaata tgtctgtt tggattaaaa aaagagtgc 61 tgatagcagc ttctgaactg gttacctgcc gtgagtaat taaaattta ttgacttagg				

Genbank格式



➤ 生物信息学常见数据格式

- 展示注释信息

Genbank格式

LOCUS NC_000913 4641652 bp DNA circular CON 11-OCT-2018
DEFINITION Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION NC_000913
VERSION NC_000913.3
DBLINK BioProject: PRJNA57779
 BioSample: SAMN02604091
 Assembly: GCF_000005845.2
KEYWORDS RefSeq.
SOURCE Escherichia coli str. K-12 substr. MG1655
ORGANISM Escherichia coli str. K-12 substr. MG1655
 Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
 Enterobacteriaceae; Escherichia.
REFERENCE 1 (bases 1 to 4641652)
AUTHORS Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
 Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,
 Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H.,
 Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
TITLE Escherichia coli K-12: a cooperatively developed annotation
 snapshot--2005
JOURNAL Nucleic Acids Res. 34 (1), 1-9 (2006)
PUBMED 16397293
REMARK Publication Status: Online-Only
COMMENT PROVISIONAL REFSEQ: This record has not yet been subject to final
 NCBI review. The reference sequence is identical to U00096.
 On Nov 3, 2013 this sequence version replaced NC_000913.2.
 Changes to proteins and annotation made on September 24, 2018.
 Current U00096 annotation updates are derived from EcoCyc
<https://ecocyc.org/>. Suggestions for updates can be sent to
 biocyc-support@ai.sri.com. These updates are being generated from a
 collaboration that includes EcoCyc, the University of Wisconsin,
 UniProtKB/Swiss-Prot, and the National Center for Biotechnology
 Information (NCBI).
FEATURES COMPLETENESS: full length.
source Location/Qualifiers
 1..4641652
 /organism="Escherichia coli str. K-12 substr. MG1655"
 /mol_type="genomic DNA"
 /strain="K-12"
 /sub_strain="MG1655"
 /db_xref="taxon:511145"
gene 190..255
 /gene="thrL"
 /locus_tag="b0001"
 /gene_synonym="ECK0001"
 /db_xref="ASAP:ABE-0000006"
 /db_xref="ECOCYC:EG11277"
 |(省略若干行...)
ORIGIN
 1 agctttcat tctgactgca acggccaata tgctctgtg tggattaaaa aaagagtgtc
 61 ttagatcgacg ttctgaactg gttacccgttcc gtggatataat taaaattta ttgacttagg

GenBank	含义
LOCUS	名称，长度，分子类型，数据分类，最后一次修订时间
DEFINITION	序列简单说明
ACCESSION	序列检索号
VERSION	序列版本号
KEYWORDS	与序列相关的关键词
SOURCE	序列物种来源
ORGANISM	序列来源的物种分类
REFERENCE	相关文献编号，或递交序列的注册信息
AUTHORS	相关文献作者，或递交序列的作者
TITLE	相关文献题目
JOURNAL	相关文献刊物杂志名，或递交序列的作者单位
PUBMED	相关文献 PUBMED引文代码
REMARK	相关文献注释
	相关文献其它注释
COMMENT	关于序列的注释信息
	相关数据库交叉引用号
FEATURES	序列特征表起始
...	序列特征表子项
ORIGIN	序列



➤ 生物信息学常见数据格式

• 展示注释信息

GFF格式

- GFF(General Feature Format)是一种用于描述基因或者其它序列元素的文件格式.
- GFF有几个版本，早期的第Version2和现在的Version3.
- GFF有统一的格式来表示基因等元素，被广泛的使用与mapping与基因组数据可视化方面。(**便于计算机处理**)

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build ASM584v2
#!genome-build-accession NCBI_Assembly:GCF_000005845.2
##sequence-region NC_000913.3 1 4641652
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=511145
NC_000913.3 RefSeq region 1 4641652 . + . ID=NC_000913.3:1..4641652;Dbxref=taxon:511145;Is_circular=true;Name=ANONYMOUS;gbkey=
NC_000913.3 RefSeq gene 190 255 . + . ID=gene-b0001;Dbxref=ASAP:ABE-0000006,ECOCYC:EG11277,EcoGene:EG11277,GeneID:944742;Name=
NC_000913.3 RefSeq CDS 190 255 . + 0 ID=cds-NP_414542.1;Parent=gene-b0001;Dbxref=UniProtKB/Swiss-Prot:P0AD86,Genbank:NP_414542
NC_000913.3 RefSeq gene 337 2799 . + . ID=gene-b0002;Dbxref=ASAP:ABE-0000008,ECOCYC:EG10998,EcoGene:EG10998,GeneID:945803;Na
NC_000913.3 RefSeq CDS 337 2799 . + 0 ID=cds-NP_414543.1;Parent=gene-b0002;Dbxref=UniProtKB/Swiss-Prot:P00561,Genbank:NP_4145
NC_000913.3 RefSeq gene 2801 3733 . + . ID=gene-b0003;Dbxref=ASAP:ABE-0000010,ECOCYC:EG10999,EcoGene:EG10999,GeneID:947498;
NC_000913.3 RefSeq CDS 2801 3733 . + 0 ID=cds-NP_414544.1;Parent=gene-b0003;Dbxref=UniProtKB/Swiss-Prot:P00547,Genbank:NP_41
NC_000913.3 RefSeq gene 3734 5020 . + . ID=gene-b0004;Dbxref=ASAP:ABE-0000012,ECOCYC:EG11000,EcoGene:EG11000,GeneID:945198;
NC_000913.3 RefSeq CDS 3734 5020 . + 0 ID=cds-NP_414545.1;Parent=gene-b0004;Dbxref=UniProtKB/Swiss-Prot:P00934,Genbank:NP_41
NC_000913.3 RefSeq gene 5234 5530 . + . ID=gene-b0005;Dbxref=ASAP:ABE-0000015,ECOCYC:G6081,EcoGene:EG14384,GeneID:944747;Na
NC_000913.3 RefSeq CDS 5234 5530 . + 0 ID=cds-NP_414546.1;Parent=gene-b0005;Dbxref=UniProtKB/Swiss-Prot:P75616,Genbank:NP_41
NC_000913.3 RefSeq repeat_region 5566 5601 . + . ID=id-NC_000913.3:5566..5601;Note=REP1a;gbkey=repeat_region;rpt_type=other
NC_000913.3 RefSeq repeat_region 5637 5670 . - . ID=id-NC_000913.3:5637..5670;Note=REP1b;gbkey=repeat_region;rpt_type=other
NC_000913.3 RefSeq gene 5683 6459 . - . ID=gene-b0006;Dbxref=ASAP:ABE-0000018,ECOCYC:EG10011,EcoGene:EG10011,GeneID:944749;
NC_000913.3 RefSeq CDS 5683 6459 . - 0 ID=cds-NP_414547.1;Parent=gene-b0006;Dbxref=UniProtKB/Swiss-Prot:P0A8I3,Genbank:NP_41
NC_000913.3 RefSeq gene 6529 7959 . - . ID=gene-b0007;Dbxref=ASAP:ABE-0000020,ECOCYC:EG11555,EcoGene:EG11555,GeneID:944745;
```

GFF格式



➤ 生物信息学常见数据格式

• 展示注释信息

GFF格式

第一行的`##gff-version 3`通常是要的，而且必须是在文件的第一行。

非必须注释信息行

```
##gff-version 3          注释生成器          基因组登陆号  
#!gff-spec-version 1.21  
#!processor NCBI annotwriter  
#!genome-build ASM584v2  
#!genome-build-accession NCBI_Assembly:GCF_000005845.2    基因组组装登陆号  
##sequence-region NC_000913.3 1 4641652                    gff注释区域  
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=511145
```

NCBI genbank数据库 GFF头部
`gff_example.gff`



- 生物信息学常见数据格式
- 展示注释信息

GFF格式

1	2	3	4	5	6	7	8	9
NC_000913.3	RefSeq	region	1	4641652	.	+	.	ID=NC_000913.3:1..4641652;Dbxref=taxon:511145;Is
NC_000913.3	RefSeq	gene	190	255	.	+	.	ID=gene-b0001;Dbxref=ASAP:ABE-0000006,ECOCY
NC_000913.3	RefSeq	CDS	190	255	.	+	0	ID=cds-NP_414542.1;Parent=gene-b0001;Dbxref=Un
NC_000913.3	RefSeq	gene	337	2799	.	+	.	ID=gene-b0002;Dbxref=ASAP:ABE-0000008,ECOCY
NC_000913.3	RefSeq	CDS	337	2799	.	+	0	ID=cds-NP_414543.1;Parent=gene-b0002;Dbxref=Un
NC_000913.3	RefSeq	gene	2801	3733	1	+	.	ID=gene-b0003;Dbxref=ASAP:ABE-0000010,ECOCY
NC_000913.3	RefSeq	CDS	2801	3733	.	+	0	ID=cds-NP_414544.1;Parent=gene-b0003;Dbxref=Un

NCBI genbank 数据库 GFF 主体部分

GFF文件格式是由tab(制表符)隔开的九列值，每一行的九个字段的含义如下



➤ 生物信息学常见数据格式

• 展示注释信息

GFF格式

NC_000913.3 RefSeq gene 190 255 . + 属性

GFF文件格式是由tab(制表符)隔开的九列值，每一行的九个字段的含义如下：

第一列: reference sequence, 该列表示的是特征元素所在的染色体（或者scaffold, 或者contig），也就是在基因组中的坐标系统，后续一切的注释信息都是基于此列。

第二列: source, 该列表示该行注释信息的来源，比如上述的一行表示该行的CDS注释信息来自名为“RefSeq”的注释。

第三列: feature, 或者说是method, type, 表示的是该注释的类型，比如上述表示改行注释为CDS信息，可以将source和feature结合起来描述的更加详细。

第四列: start position, 在reference sequence上的开始位置（坐标），通常是从1为起点而不是0。

第五列: end position, 在reference sequence上的结束位置（坐标），一般是大于start position的。

第六列: score, 表示该行feature的分数，比如序列相似性等，如果没有对应的分数可以用.代替。

第七列: strand, feature所在链，+表示正链，-表示负链，.表示不确定或者与链无关。

第八列: phase, 与蛋白质编码相关，一般用于CDS，值的范围为0-2，表示编码时阅读框的移动相位。

0表示指定区域在框架中，即其第一碱基对应于密码子的第一碱基。

1表示存在一个额外的碱基，即该区域的第二个碱基对应于一个密码子的第一个碱基，

2表示该区域的第三个碱基是一个密码子的第一个碱基。如果链为“-”，则该区域的第一个基数为<end>的值，因为相应的编码区域将从反向链的<end>到<start>。



➤ 生物信息学常见数据格式

• 展示注释信息

GFF格式

```
ID=gene-b0001;ECOCYC:EG11277,EcoGene:EG11277,GenelD:944742;Name=thrL;gbkey=Gene;gene=thrL;gene_biotype=protein_coding;gene_synonym=ECK0001;locus_tag=b0001
```

第九列：属性（ attributes ），是用于对该行注释更多的描述，以键值对的形式该列中可以存在多个属性，属性之间是用;隔开的。

有几个特定的键，具体如下：

- **ID**, feature在整个GFF3文件中唯一的标识符；
- **Name**, feature的名字，不同于ID，Name不要求唯一，只是方便用户浏览；
- **Alias**, 相当于feature的别名；
- **Parent**, 表明该feature所属的上一级feature 的ID，这种关系可用于exons-transcripts, transcripts-genes，可以看出一个feature可以拥有多个子feature；
- **Target**, 主要是用于序列比对结果的展示，value的格式为target_id start end [strand], 其中如果target_id中含有空格则需转换为%20；

后面还有些其它属性比如Note等，这里不再做详细描述。



➤ 生物信息学常见数据格式

• 比对结果显示的文件 sam、bam

在生物信息学中尤其是高通量测序数据分析中，大部分的操作都是在实现短片段序列与参考序列的比对（mapping），比如bowtie等，这就涉及到如何使用一个统一的格式来表示这种mapping结果呢，sam（Sequence Alignment/Map）格式就是来解决这个问题的。

```
131      141      151      161      171      181      191
SCATCTTCAACACCCGCCTCTCCCGCACCTCGGATATACTATCAAGCGAACACAGTCAAAACGCCCTCTGGG
SCATCTTCAACACCCGCCTCTCCCGCACCTCGGATATACTATCAAGCGAACACAGTCAAAACGCCCTCTGGG
SCATCTTCAACACCCGCCTCTCCCGCACCTCGGATATACTATCAAGCGAACACAGTCAAAACGCCCTCTGGG
SCATCTTCAACACCCGCCTCTCCCGCACCTCGGATATACTATCAAGCGAACACAGTCAAAACGCCCTCTGGG
SCATCTTCAACACCCGCCTCTCCCGCACCTCGGATATACTATCAAGCGAACACAGTCAAAACGCCCTCTGGG
SCATCTTCAACACCCGCCTCTCCCGCACCTCGGATATACTA          CGAACACAGTCAAAACGCCCTCTGGG
SCATCTTCAACACCCGCCTCTCCCGCACCTCGGATATACTATCAA      ACCACAGTCAAAACGCCCTCTGGG
CATCTTCAACACCCGCCTCTCCCGCACCTCGGATATACTATCAAGCGAA ACAGTCAAAACGCCCTCTGGG
SCATCTTCAA                                     GTCAAAACGCCCTCTGGG
SCATCTTCAACACC          CCCTCCTGGG
SCATCTTCAACACCCGCCTCTCCCGCACCTCGGATATACTATCAAGCGAA
CATCTTCAACACCCGCCTCTCCCGCACCTCGGATATACTATCAAGCGAA
CTTCAACACCCGCCTCTCCCGCACCTCGGATATACTATCAAGCGAACCA
```

Reads mapping 到基因组上

```
@HD VN:1.0 SO:unsorted
@SQ SN:gi|189396501 LN:158
@SQ SN:gi|557878761 LN:184
@SQ SN:gi|330723203 LN:192
@SQ SN:gi|367460291 LN:889
@SQ SN:gi|385858114 LN:1330
@SQ SN:gi|526641975 LN:2587
@SQ SN:gi|330723203 LN:1170
@SQ SN:gi|367460291 LN:207
@SQ SN:gi|330723203 LN:286
@PG ID:bowtie2 PN:bowtie2 VN:2.2.9 CL:"/home/bio/bowtie2-2.2.9/bowtie2-align-s --wrapper basic-0 -x /home/script/data/reference -r query.fastq -S query.sam -q -p 40 --local"
FCH2NTJBCX3:2:2202:16757:32446_GTTGTGTA_TCCTTCCC/1 16 gi|189396501 1195 1 240M * 0 0 TGTTGGTACAAGAGAACGAACTATGGTGACATGGAGAACATTCGAAAAACCGATCTCGAACACTCGACTC
FCH2NTJBCX3:2:2116:4675:70521_GTTGTGTA_TCCTTCCC/1 0 gi|557878761 1203 1 240M * 0 0 ACGCTAAAGTGCCTGGATGACTTGTGGATAGCGGTGAATTCGAATCGAACCTGGAGATAGCTGTTCTCCGAAATAGCTTGGGCT
FCH2NTJBCX3:2:1207:8353:86305_GTTGTGTA_TCCTTCCC/1 16 gi|330723203 386946 1 240M * 0 0 TCAAGAAAAATACCAATTGGAAAGACTTGTGATACTCCGATTATTCCCATTTGTGATTTAAAGAAATTAATGTGCTAGATCTGAAGGAAG
FCH2NTJBCX3:2:2112:12518:70453_GTTGTGTA_TCCTTCCC/1 16 gi|367460291 21118 1 240M * 0 0 GGCTAGCGTATAGTTGTTAATGGGGTAGAGCCTGAATGTGGAAATGGCCGATCTAGCTGACTGACTATAATCAAACCTCGAA
FCH2NTJBCX3:2:2205:16628:88021_GTTGTGTA_TCCTTCCC/1 16 gi|385858114 1359 1 240M * 0 0 ATGATGACTTTAACTGCAATTACCACTGGAAATCCGAATCATCGGAGCTTCTAATGGTAGATCGTTGATTAATGGTGAATAC
FCH2NTJBCX3:2:1111:3800:29227_GTTGTGTA_TCCTTCCC/1 16 gi|526641975 635 1 240M * 0 0 TAAGCAGAAAAAGCGGAGCGTGGGAACCGAGCTGTAATAGGGGACTTTAGTATTTGGCATATAACCGGGACTTGTGATATTCGCAATAC
FCH2NTJBCX3:2:1205:11370:12921_GTTGTGTA_TCCTTCCC/1 0 gi|330723203 565196 1 240M * 0 0 AAAAATAGAAAAGGGGATTTTTGTGAAATTTCAATAGAATTAAATAGTGTGAAAAAAATACAAAAAAAGGAAATTGGAATTTAA
FCH2NTJBCX3:2:1213:9563:87046_GTTGTGTA_TCCTTCCC/1 16 gi|367460291 15856 1 240M * 0 0 TAAATACCCCTTTAAACCTGGCATTTCTGTGCTCATCAATTATTTGGACTTAATCTACCAAGATAGAAATTTGTG
FCH2NTJBCX3:2:1202:7873:29458_GTTGTGTA_TCCTTCCC/1 0 gi|330723203 480900 1 240M * 0 0 TACAGAAACCCATATCTAATAGAAAATATCAAATAGTAAATATTACTCTCAAACACCCAAATCTAGCAAAGGTCACCCATTGCA
FCH2NTJBCX3:2:1101:9345:26442_GTTGTGTA_TCCTTCCC/1 16 gi|557878761 100189 1 240M * 0 0 GGTAGTATCTGATCACAAATTTAGGTGAGCTTTAGTGGAAATGTTAGTGGTTAGGGTTATTGCTTAACGATTAAATATTGGG
```

Sam格式
[sam_example.sam](#)



➤ 生物信息学常见数据格式

• 比对结果显示的文件

sam、bam

➤ SAM分为两部分，注释信息（header section）

@HD VN:1.0 SO:unsorted (排序类型)

头部区第一行：VN是格式版本；SO表示比对排序的类型，有unknown (default) , unsorted, queryname和coordinate几种。
samtools软件在进行行排序后不能自动更新bam文件的SO值，而picard却可以。

@SQ SN:gi|189396501 LN:158 (序列ID及长度)

参考序列名，这些参考序列决定了比对结果sort的顺序，SN是参考序列名；LN是参考序列长度；每个参考序列为一行。

@RG ID:sample01 (样品基本信息)

Read Group。1个sample的测序结果为1个Read Group；该sample可以有多个library的测序结果，可以利用bwa mem -R 加上去这些信息。

例如：@RG ID:ZX1_ID SM:ZX1 LB:PE400 PU:Illumina PL:Miseq

ID：样品的ID号 SM：样品名 LB：文库名 PU：测序仪 PL：测序平台

这些信息可以在形成sam文件时加入，ID是必须要有的，后面是否添加看分析要求

@PG ID:bowtie2 PN:bowtie2 VN:2.2.9 ... (比对所使用的软件及版本)

例如：@PG ID:bowtie2 PN:bowtie2 VN:2.2.9 CL: "/home/bio/bowtie2-2.2.9/bowtie2-align-s --wrapper basic-0 -x /home/script/data/reference -r query.fastq -S query.sam -q -p 40 --local"

这里的ID是运行程序的id，PN是运行软件名，VN是软件的版本号。CL是生成sam文件运行的命令



➤ 生物信息学常见数据格式

• 比对结果显示的文件

sam、bam

➤ SAM比对结果部分 (alignment section)

```
FCH2NTJBCX3:2:2202:16757:32446 GTGTGTTA_TCTTCCC/1 16 gil189396501 1195 1 240M * 0 0 TGGTCGGTACAAAGAGA
```

1.第一列：read name, read的名字通常包括测序平台等信息；

2.第二列：sum of flags, 比对flag数字之和，比对flag用数字表示，分别为：

- 1 (1) 该read是成对的paired reads中的一个
- 2 (10) paired reads中每个都正确比对到参考序列上
- 4 (100) 该read没比对到参考序列上
- 8 (1000) 与该read成对的matepair read没有比对到参考序列上
- 16 (10000) 该read其反向互补序列能够比对到参考序列
- 32 (100000) 与该read成对的matepair read其反向互补序列能够比对到参考序列
- 64 (1000000) 在paired reads中，该read是与参考序列比对的第一条
- 128 (10000000) 在paired reads中，该read是与参考序列比对的第二条
- 256 (100000000) 该read是次优的比对结果
- 512 (1000000000) 该read没有通过质量控制
- 1024 (10000000000) 由于PCR或测序错误产生的重复reads
- 2048 (100000000000) 补充匹配的read

通过这个和可以直接推断出匹配的情况。假如说标记不是以上列举出的数字，比如说 $83 = (64 + 16 + 2 + 1)$ ，就是这几种情况值和。



➤ 生物信息学常见数据格式

• 比对结果显示的文件

sam、bam

➤ SAM比对结果部分 (alignment section)

```
FCH2NTJBCX3:2:2202:16757:32446 GTGTGTTA TCTTCCC/1    16    gil189396501  1195   1    240M *    0    0    TGGTCGGTACAAAGAGA
```

3.第三列：RNAM, reference sequence name, 实际上就是比对到参考序列上的染色体号。若是无法比对，则是*；

4.第四列：position, read比对到参考序列上，第一个碱基所在的位置。若是无法比对，则是0；

5.第五列：Mapping quality, 比对的质量分数，越高说明该read比对到参考基因组上的位置越唯一；

6.第六列：CIGAR值, read比对的具体情况，

- “M” 表示 match或 mismatch；但是无论reads与序列的正确匹配或是错误匹配该位置都显示为M
- “I” 表示 insert；表示read的碱基序列相对于第三列的RNAME序列，有碱基的插入
- “D” 表示 deletion；表示read的碱基序列相对于第三列的RNAME序列，有碱基的删除
- “N” 表示 skipped (跳过这段区域)；表示可变剪接位置
- “S” 表示 soft clipping (被剪切的序列存在于序列中)；
- “H” 表示 hard clipping (被剪切的序列不存在于序列中)；

clipped均表示一条read的序列被分开，之所以被分开，是因为read的一部分序列能匹配到第三列的RNAME序列上，而被分开的那部分不能匹配到RNAME序列上。

- “P” 表示 padding；
- “=” 表示 match；表示正确匹配到序列上
- “X” 表示 mismatch；表示错误匹配到序列上



➤ 生物信息学常见数据格式

• 比对结果显示的文件

sam、bam

➤ SAM比对结果部分 (alignment section)

```
*      0      0      TGGTCGGTACAAAGAGAAGCAATATGGTA      gggcgedgghhhggggheh      AS:i:480      XS:i:480      XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:240      YT:Z:UU
```

7.第七列：MRNM(chr), mate的reference sequence name, 实际上就是mate比对到的染色体号，若是没有mate，则是*;

8.第八列：mate position, mate比对到参考序列上的第一个碱基位置，若无mate,则为0;

9.第九列：ISIZE, Inferred fragment size.详见Illumina中paired end sequencing 和 mate pair sequencing, 是负数，推测应该是两条read之间的间隔，若无mate则为0; (可以理解为文库插入片段长度) ;

10.第十列：Sequence, 就是read的碱基序列，如果是比对到互补链上则是reverse completed

11.第十一列：ASCII, read质量的ASCII编码。

12.第十二列之后：Optional fields, 可选的自定义区域；

详细sam文件信息请查阅附加数据中SAM格式详细说明.pdf



- 生物信息学常见数据格式
- **比对结果显示的文件**
 - sam、bam**
- **bam**

bam是**sam**的二进制格式，因此两者格式相同，只是BAM文件占用储存空间更小，运算更快



➤ 生物信息学常见数据格式

- 表示突变信息

➤ VCF

VCF(Variant Call Format) 是用于描述SNP, InDel和SV结果的文本文件。VCF格式在GATK软件中得到很好的支持。

以“#”开头的注释部分

没有 "#" 开头的主体部分



➤ 生物信息学常见数据格式

• 表示突变信息

➤ VCF

- 注释部分有很多对VCF的介绍信息
- 主体部分包含10列数据。主题部分每一行代表一个variant的信息。

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	p1
4	1470	.	T	G	607.3	.	AC=4;AF=0.500;AN=8;BaseQRankGT:AD:DP:GQ:PL	GT:AD:DP:GQ:PL	0/0:9,0:9:27:0,27,310

1. CHROM : 参考序列名称
2. POS : variant所在的left-most位置(1-base position) (发生变异的位置的第一个碱基所在的位置)
3. ID : variant的ID。同时对应着dbSNP数据库中的ID，若没有，则默认使用 ‘’
4. REF : 参考序列的Allele, (等位碱基，即参考序列该位置的碱基类型及碱基数量)
5. ALT : variant的Allele, **若有两个，则使用逗号分隔**, (变异所支持的碱基类型及碱基数量) **这里的碱基类型和碱基数量，对于SNP来说是单个碱基类型的编号，而对于Indel来说是指碱基个数的添加或缺失，以及碱基类型的变化**
6. QUAL : variants的质量。Phred格式的数值，代表着此位点是**纯合的概率**，**此值越大**，则概率越低，代表着次位点是**variants的可能性越大**。（表示变异碱基的可能性）
7. FILTER : 次位点是否要被过滤掉。**如果是PASS，则表示此位点可以考虑为variant.**
8. INFO : variant的相关信息
9. FORMAT : variants的格式，例如GT:AD:DP:GQ:PL
10. SAMPLES : 各个Sample的值，由BAM文件中的@RG下的SM标签所决定，这些值对应着第9列的各个格式，不同格式的值用冒号分开，每一个sample对应着1列；多个samples则对应着多列，这种情况下列的数多余10列。

重要
信息



➤ 生物信息学常见数据格式

• 表示突变信息

➤ VCF

● FORMAT和最后一列

两者和一起基因型信息，前者为格式，后者为对应的数据



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	p1
4	1470	.	T	G	607.3	.	AC=4;AF=0.500;AN=8;BaseQRankGT:AD:DP:GQ:PL	GT:AD:DP:GQ:PL	0/0:9,0:9:27:0,27,310

- **GT**: 表示样品的基因型，对于二倍体生物，GT值表示的是样本在这个位点所携带的两个等位基因。0表示跟REF一样，1表示跟ALT一样，2表示有第二个ALT；当只有一个ALT等位基因时：0/0表示纯合子并跟REF一致；0/1表示杂合子，有两个allele，一个是ALT，另一个是REF；1/1表示纯合子并都为ALT
- **AD**: 两个以逗号分隔的值，分别表示覆盖到REF和ALT碱基的reads数，也就是REF和ALT对应的测序深度
- **DP**: 表示覆盖在这个位点的总reads数，也就是这个位点的测序深度（并不是指具体有多少个reads数量，而是大概满足一定质量值要求的reads数）
- **GQ**: 表示基因型的质量值，Phred格式(Phred_scaled)的质量值， $\text{Phred值} = -10 * \log (1-p)$ p为基因型存在的概率，表示该位点基因型存在的可能性。
- **PL**: 三个逗号分隔的值，分别对应该位点的三个基因型0/0, 0/1, 1/1的没经过先验的标准化Phred-scaled似然值 (L)， $L=-10\lg P$ ，P为支持该基因型的概率，3个概率总和为1；因此，L这个值越小，支持概率就越大，也就是说是这个基因型的可能性越大。



➤ 生物信息学常见数据格式

• 表示突变信息

➤ VCF

- INFO列所包含的信息

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	p1
4	1470	.	T	G	607.3	.	AC=4;AF=0.500;AN=8;BaseQRankGT:AD:DP:GQ:PL	0/0:9,0:9:27:0,27,310	



- AC, AF和AN

AC (Allele Count) 表示基因型为与variant一致的Allele (等位碱基) 的数目；AF (Allele Frequency) 表示Allele的频率，AF值=AC值/AN值；AN (Allele Number) 表示Allele的总数目。比如：对2个sample的双倍体进行测序，则AN值为4。若REF上位点碱基为A，而2个sample在该位点分别为A/T和T/G，则AC值为2, 1；AF值为0.50, 0.25。**AC: variant数目, AF: 频率, AN: 总数目**

- DP (reads覆盖度)

表示reads被过滤后的覆盖度

- FS

FisherStrand的缩写，表示使用Fisher's精确检验来检测strand bias而得到的Fhred格式的p值，**该值越小越好**；如果该值较大，表示strand bias (正负链偏移) 越严重，即所检测到的variants位点上，reads比对到正负义链上的比例不均衡。一般进行filter的时候，**推荐保留FS<10~20的variants位点**。GATK可设定FS参数。

- MQ

表示覆盖序列质量的均方值RMS Mapping Quality



➤ 生物信息分析入门基础知识

2 生物信息学常用数据库

[返回目录](#)



➤ 生物信息学常用数据库

➤ 核酸基因组数据库

**Genbank ,
NCBI database (Nucleotide , Genome)**

➤ 蛋白质数据库

**Uniprot, SWISS-PROT,
NCBI protein database , NCBI CDD**

➤ 通路功能数据库

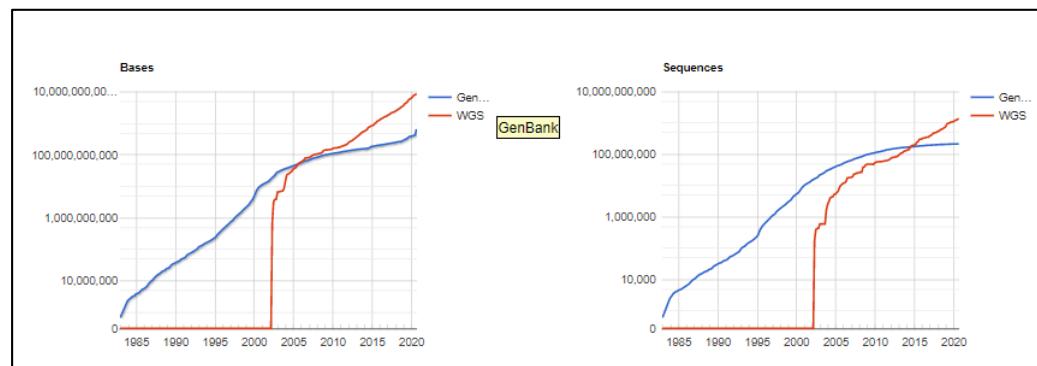
KEGG pathway database



- 生物信息学常用数据库
- 核酸基因组数据库

Genbank

- GenBank-核酸数据库是美国国家生物技术信息中心(National Center for Biotechnology Information , NCBI)建立的一个**全面的数据库**，包含近26万个物种的可公开的核苷酸序列。GenBank与洲的EMBL和日本的DDBJ之间每天进行数据交换确保了全球范围的数据覆盖。可通过FTP获取版本数据集和每日更新数据集。
- GenBank版本每两个月发布一次，可从 ftp站点（<https://ftp.ncbi.nih.gov/genbank/>）获得。



		GenBank		WGS	
Release	Date	Bases	Sequences	Bases	Sequences
3	Dec 1982	680338	606		
...
239	Aug 2020	654057069549	218642238	8841649410652	1408122887

GENBANK AND WGS STATISTICS



- 生物信息学常用数据库
- 核酸基因组数据库

访问GenBank

有几种方法可以从GenBank中搜索和检索数据。

- 用[Entrez核苷酸](#)在GenBank中搜索序列标识符和注释。
- 使用[BLAST](#)（基本局部比对搜索工具）搜索GenBank序列并将其与查询序列**比对**。BLAST独立搜索Core Nucleotide, dbEST和dbGSS；有关大量BLAST数据库的更多信息，请参见 [BLAST](#)。
- 使用[NCBI电子程序](#)以编程方式搜索，链接和下载序列。
- ASN.1和纯文本格式在NCBI的匿名FTP服务器上访问：<ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1>和 <ftp://ftp.ncbi.nlm.nih.gov/genbank>。



➤ 生物信息学常用数据库

- 核酸, 基因组数据库

NCBI database (Nucleotide , Genome)

- Nucleotide: 基因序列, 基因组序列, 片段序列
- Genome: 物种的基因组记录

Nucleotide Nucleotide MCR-1 Search Help

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.coronavirus.gov>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

Species Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

Animals (1)
Plants (1)
Fungi (3)
Bacteria (456,228)
Archaea (1)
Customize ...

Molecule types
genomic DNA/RNA (456,234)
mRNA (1)
Customize ...

Source databases
INSDC (GenBank) (230,749)
RefSeq (225,477)
Customize ...

Sequence Type
Nucleotide (456,243)

Genetic compartments
Plasmid (1,783)

Sequence length
Custom range...
Release date
...

ANTIMICROBIAL RESISTANCE GENE Was this helpful?

MCR-1 family, phosphoethanolamine–lipid A transferase gene family

This gene family can be found in the set of reference sequences used to annotate antimicrobial resistance genes from the [National Database of Antibiotic-Resistant Organisms \(NDARO\)](#).

RefSeq genomic (27) RefSeq protein (27) Pathogen Isolate Browser Reference Gene Catalog Download

Filter your results:
All (456243)
Bacteria (456228)
INSDC (GenBank) (230749)
RefSeq (225477)
WGS (454294)

Manage Filters

Results by taxon
Top Organisms [Tree]
Escherichia coli (424955)
Klebsiella pneumoniae (13757)
Salmonella enterica (8708)
Klebsiella quasipneumoniae (3110)
Klebsiella aerogenes (746)
All other taxa (4967)
More...

Find related data
Database: Select Find Items

Items: 1 to 20 of 456243

<< First < Prev Page 1 of 22813 Next > Last >

1. Escherichia coli strain C2-007R phosphoethanolamine–lipid A transferase (mcr-1).gene._partial cds
1. 1,508 bp linear DNA

NCBI Nucleotide database

<https://www.ncbi.nlm.nih.gov/nucleotide/>

Genome Genome escherichia coli[orgn] Search Help

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.coronavirus.gov>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

Tools
BLAST Genome

Related information
Assembly
BioProject

Gene
Components
Protein
PubMed
Taxonomy

Display Settings: ▾ Overview Send to: ▾

Organism Overview ; Genome Assembly and Annotation report [20922] ; Genome Tree report [14598] ; Plasmid Annotation Report [3442] ID: 167

Escherichia coli
A well-studied enteric bacterium

Lineage: Bacteria[27957]; Proteobacteria[9046]; Gammaproteobacteria[3385]; Enterobacteriales[623]; Enterobacteriaceae[345]; Escherichia[6]; Escherichia coli[1]

Escherichia coli. This organism is typically present in the lower intestine of humans, where it is the dominant facultative anaerobe present, but it is only one minor constituent of the complete intestinal microflora. E.coli is easily grown in a laboratory setting and is readily amenable to genetic manipulation making it one of the most [More...](#)

Search details
"Escherichia coli" [Organism]

NCBI Genome database

<https://www.ncbi.nlm.nih.gov/genome/>



➤ 生物信息学常用数据库

• 蛋白质数据库

Uniprot

- UniProt 是 Universal Protein 的英文缩写，是信息最丰富、资源最广的蛋白质数据库。它由整合Swiss-Prot、TrEMBL和PIR-PSD三大数据库的数据而成。他的数据主要来自于基因组测序项目完成，后续获得的蛋白质序列，它包含了大量来自文献的蛋白质的生物功能的信息，有质量保证的数据才被加入该数据库。（<https://www.uniprot.org/>）

The screenshot shows the UniProt homepage with several key sections:

- UniProtKB**:
 - UniProt Knowledgebase
 - Swiss-Prot (563,082)** (highlighted with a red box): Manually annotated and reviewed. Records with information extracted from literature and curator-evaluated computational analysis.
 - TrEMBL (188,961,949)** (highlighted with a red box): Automatically annotated and not reviewed. Records that await full manual annotation.
- UniRef**: The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.
- UniParc**: UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.
- Proteomes**: A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.
- Supporting data**: Includes sections for Literature citations, Cross-ref. databases, Taxonomy, Diseases, Subcellular locations, and Keywords.

On the left, there are two labels with red boxes:

- 人工注释的蛋白质** (highlighting the Swiss-Prot section)
- 计算机自动注释的蛋白质 (基于相似性等)** (highlighting the TrEMBL section)

At the bottom, there are links for "Getting started", "Text search", "UniProt data", and "Download latest release".

UniProt数据库主页



➤ 生物信息学常用数据库

• 蛋白质数据库

SWISS-PROT

- SWISS-PROT是经过人工注释的蛋白质序列数据库，由欧洲生物信息学研究所(EBI)维护。数据库由蛋白质序列条目构成，每个条目包含蛋白质序列、引用文献信息、分类学信息、注释等，注释中包括蛋质的功能、转录后修饰、特殊位点和区域、二级结构、四级结构、与其它序列的相似性、序列残缺与疾病的关系、序列变异体和冲突等信息。
[\(https://www.uniprot.org/uniprot/?query=reviewed:yes \)](https://www.uniprot.org/uniprot/?query=reviewed:yes)
- 在uniprot网站主页直接访问，或者使用blast进行序列查询。
- 利用blast等工具可以构建本地数据库，进行蛋白序列的批量功能注释（文件下载地址：
<https://www.uniprot.org/downloads>）



➤ 生物信息学常用数据库

• 蛋白质数据库

NCBI protein database

- 使用蛋白，基因名称，功能描述等信息检索对应蛋白

The screenshot shows the NCBI Protein search results for the query 'MCR-1'. The search bar at the top contains 'MCR-1'. Below the search bar, there is a red banner with information about COVID-19. The main content area displays the 'ANTIMICROBIAL RESISTANCE GENE' section for the 'MCR-1 family, phosphoethanolamine-lipid A transferase gene family'. It includes a brief description, links to RefSeq genomic and RefSeq protein databases, and buttons for Pathogen Isolate Browser, Reference Gene Catalog, and Download. On the left, there is a sidebar with various filters and links related to species, source databases, and genetic compartments. At the bottom, there are search details showing 'MCR-1 [All Fields]' and pagination information.

NCBI蛋白数据库
<https://www.ncbi.nlm.nih.gov/protein/>

The screenshot shows the NCBI Protein advanced search interface. The search bar at the top contains '(mcr-1[Title]) AND salmonella[Organism]'. Below the search bar, there is an 'Advanced' button with a red arrow pointing to it. The main search area contains a 'Builder' section with three dropdowns: 'Title' set to 'mcr-1', 'Organism' set to 'salmonella', and 'All Fields' set to an empty field. There are also 'Search' and 'Add to history' buttons. To the right of the search area, there are three 'Show index list' buttons.

检索词：(mcr-1[Title]) AND salmonella[Organism]

NCBI蛋白数据库高级检索页面
<https://www.ncbi.nlm.nih.gov/protein/advanced>



➤ 生物信息学常用数据库

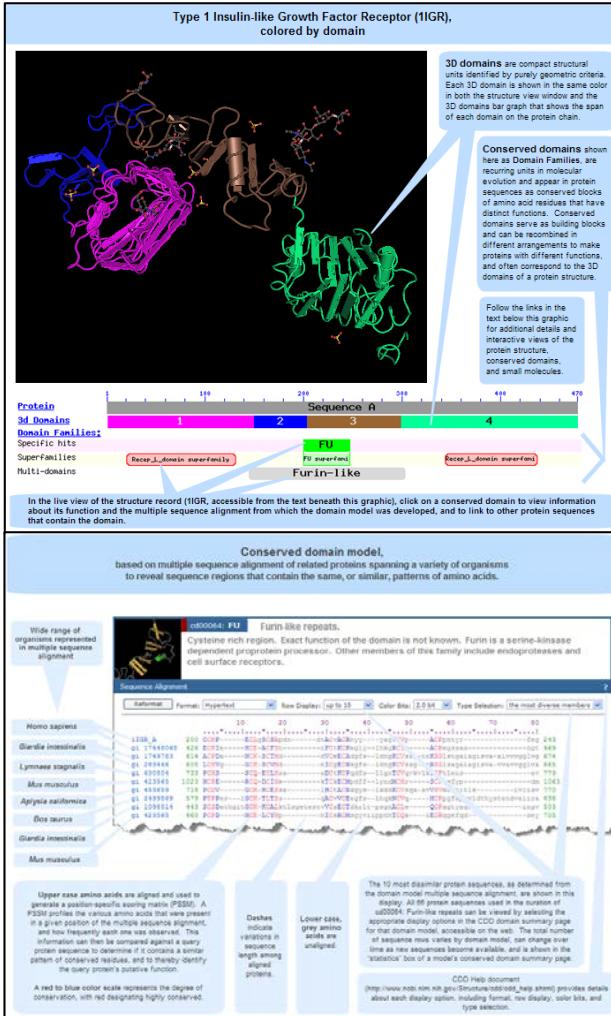
- 蛋白质数据库

NCBI Conserved Domain Database (CDD)

- 域可以被认为是蛋白质的不同功能和/或结构单元。保守结构域包含保守的序列模式或基序，其允许在多肽序列中进行检测。
- 保守域模型基于跨越各种生物体的相关蛋白的多重序列比对，揭示了包含相同或相似氨基酸模式的序列区域。
- 域的存在大概率意味着功能的存在

CDD: <https://www.ncbi.nlm.nih.gov/cdd/>

实用工具: 1、[CD-Search](#)
2、[Batch CD-Search](#)



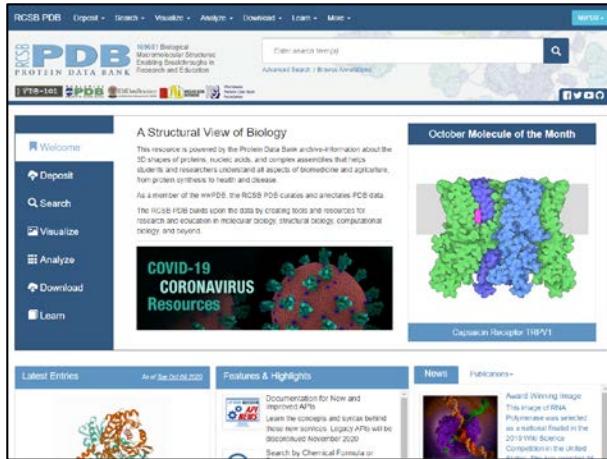
多序列比对为基础的蛋白的保守域



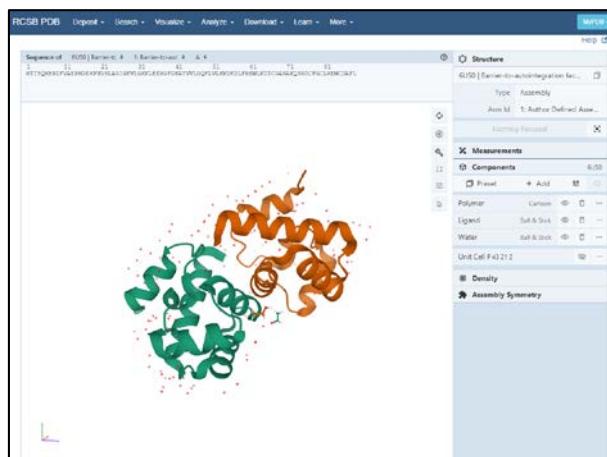
- 生物信息学常用数据库
- 蛋白质数据库

Protein Data Bank (PDB)

- Protein Data Bank (PDB)可以访问大型生物分子（蛋白质，DNA和RNA）的3D结构数据。
<https://www.rcsb.org/>
- 拥有大于1 TB的蛋白质，DNA和RNA结构数据
- 可以使用名称，功能，氨基酸序列检索已有的蛋白质三维结构。



PDB主页



查看蛋白质的三维结构



➤ 生物信息学常用数据库

- 通路功能数据库

KEGG pathway database

- KEGG全称为 Kyoto Encyclopedia of Genes and Genomes (京都基因与基因组百科全书)
<https://www.genome.jp/kegg/>
- 子数据库中最重要也是最常用的就是**KEGG PATHWAY**, 它包括大量由科研人员根据已有研究文献, 通过手动绘制的KEGG通路图, 代表着代谢过程, 环境信息过程, 细胞过程, 生物系统, 人类疾病和药物开发。

快速检索栏

内容导航

KEGG各个子库列表

按物种检索kegg信息

kegg分析工具

Copyright 1995-2020 Kanehisa Laboratories

KEGG主页
<https://www.genome.jp/kegg/>

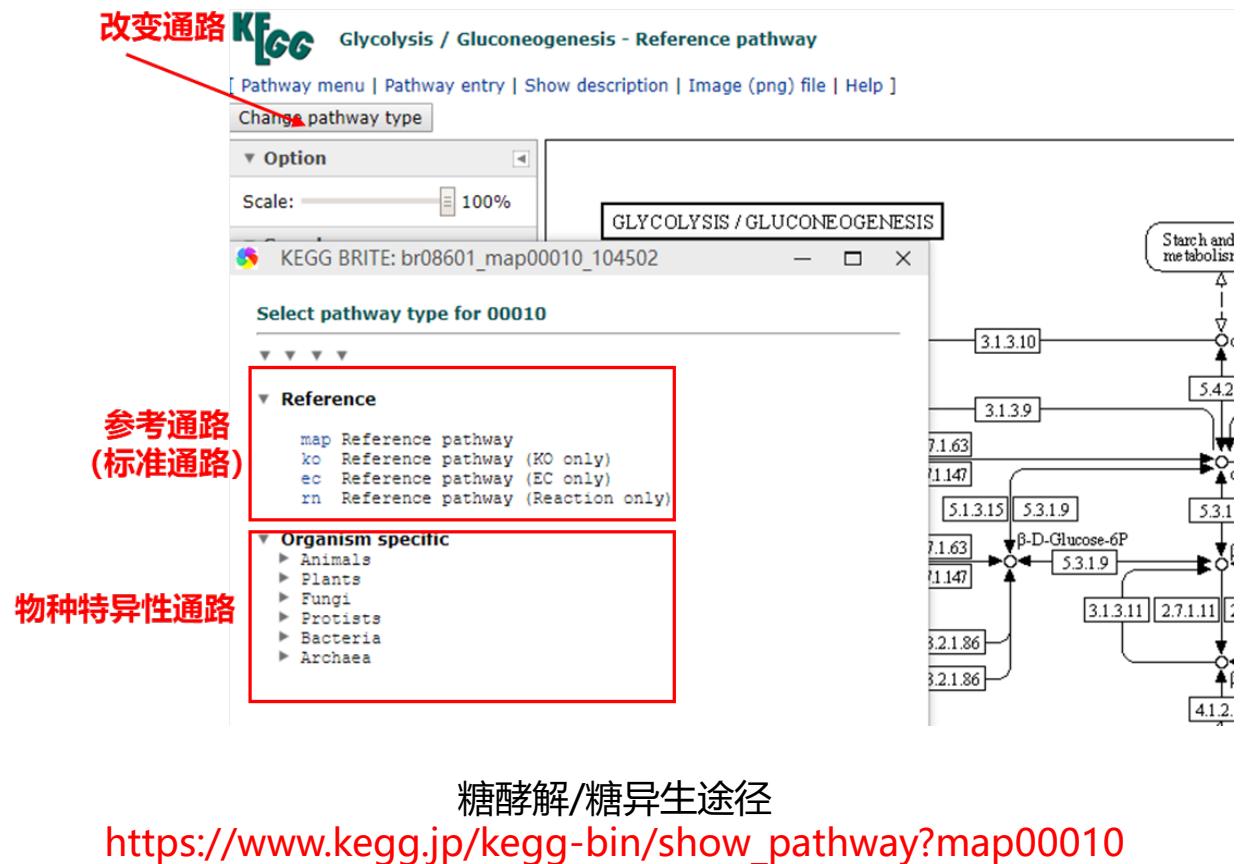


➤ 生物信息学常用数据库

• 通路功能数据库

KEGG pathway database

- **map编号**: 代表reference pathway, 根据已有的知识绘制的、概括的、详尽的具有一般参考意义的代谢图。一个点同时表示一个基因, 这个基因编码的酶或这个酶参加的反应
- **ko编号**: KO通路中的点表示直系同源基因
- **ec编号**: EC通路中的点表示相关的酶
- **rn编号**: 化学反应通路中的点只表示该点参与的某个反应、反应物及反应类型
- **org编号**: 物种特异性通路, 这里就是将K编号基因(直系同源基因, 后面会介绍)换为每个物种中对应的基因





➤ 生物信息学常用数据库

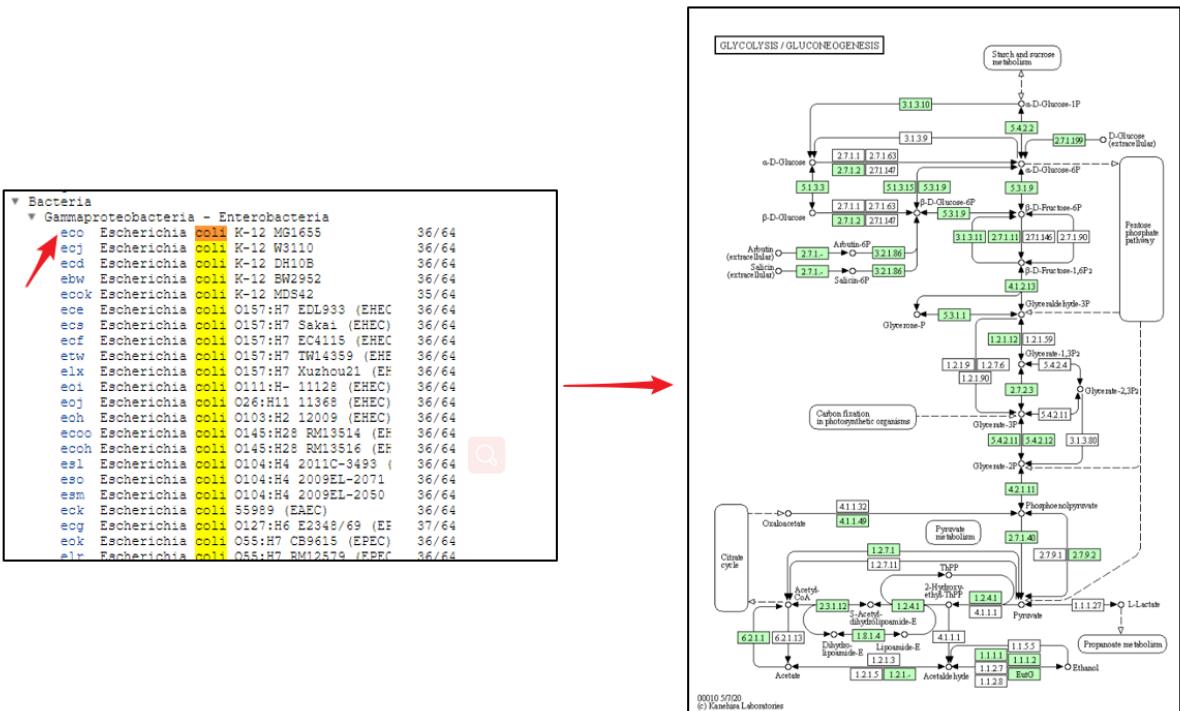
- 通路功能数据库

KEGG pathway database

➤ 物种特异性通路

- ★ 1. eco : Escherichia coli K-12 MG1655
- ★ 2. 无色表示不存在该基因
- 3. 点击绿色基因，会进入Gene详细信息

▼ Bacteria		
▼ Gammaproteobacteria - Enterobacteri		
eco	Escherichia coli K-12 MG1655	36/64
ecoJ	Escherichia coli K-12 W3110	36/64
ecoD	Escherichia coli K-12 DH10B	36/64
ehw	Escherichia coli K-12 BW2592	36/64
ecoli	Escherichia coli K-12 MDS42	35/64
ece	Escherichia coli O157:H7 EDL933 (EHEC)	36/64
ecs	Escherichia coli O157:H7 Sakai (EHEC)	36/64
ecf	Escherichia coli O157:H7 EC4115 (EHEC)	36/64
etw	Escherichia coli O157:H7 TW14359 (EHE)	36/64
elx	Escherichia coli O157:H7 Kuzhou21 (EF)	36/64
eoi	Escherichia coli O111:H7 11128 (EHEC)	36/64
eoJ	Escherichia coli O26:H11 11368 (EHEC)	36/64
echo	Escherichia coli O103:H2 12009 (EHEC)	36/64
ecoo	Escherichia coli O145:H28 RM13514 (EE)	36/64
ecol	Escherichia coli O145:H28 RM13516 (EE)	36/64
esl	Escherichia coli O104:H4 2011C-3493 (36/64
eso	Escherichia coli O104:H4 2009EL-2071	36/64
esm	Escherichia coli O104:H4 2009EL-2050	36/64
eck	Escherichia coli S5989 (EAEC)	36/64
ecg	Escherichia coli O127:H6 E2348/69 (EE)	37/64
eko	Escherichia coli O55:H7 C89615 (EPEC)	36/64
elr	Escherichia coli O55:H7 RM12579 (EPEC)	36/64



Escherichia coli K-12 MG1655 糖酵解/糖异生途径
https://www.kegg.jp/kegg-bin/show_pathway?eco00010



➤ 生物信息学常用数据库

• 通路功能数据库

KEGG pathway database

➤ 物种特异性通路

1. eco : Escherichia coli K-12 MG1655
 2. 无色表示不存在该基因
 3. 点击绿色基因，会进入Gene详细信息

Escherichia coli K-12 MG1655: b1780基因详细信息
https://www.kegg.jp/dbget-bin/www_bget?eco:b1780



➤ 生物信息分析入门基础知识

3 生信的重要基础--序列联配

返回目录



- 生物信息学常用数据库
- **生信的重要基础--序列联配**

- 序列联配基本概念
- 计分矩阵
- 联配的应用场景



- 生物信息学常用数据库
- **生信的重要基础--序列联配**

序列联配基本概念

- **序列分析**是生物信息学最主要的研究内容之一, 它可以分为两个主要部分:
 - 一、是序列组成分析 (包括基因和基因组层次)
 - 二、是序列之间的比较分析
- **联配**目的是对序列**相似性**进行评估, 找出这些序列中结构或功能相似性区域等。通过联配未知序列与已知序列 (其功能或结构等已知) 的相似程度, 我们可以判断或推测未知序列的结构与功能。
- **序列联配** (sequence alignment) 也叫**序列对比**, 指在允许错配和插入空格的情况下, 对两条或多条序列的位点进行匹配。**序列联配**是生物信息学中的重要内容之一, 许多生物信息学分析均涉及序列联配方法。

seq1: TGC^GGAGC
seq2: TCGGAGC

最简单的联配规则

TGC^GGAGC
| |
TCGGAGC

引入空格

TGC^GGAGC
| | | | |
T CGGAGC

简单的序列联配示意



- 生物信息学常用数据库
- **生信的重要基础--序列联配**

序列联配的表示方法

- '-': 表示gap，可以认为是候选的InDel
- '|': 表示匹配，一致的碱基或氨基酸残基
- '.': 表示错配，可能有碱基发生了突变

ATGCAAATGACAAATAC	. .
ATGC --- TGATAACT --	

包含 “gap”、 “match”、 “mismatch”
的序列联配表示方法



➤ 生物信息学常用数据库

- 生信的重要基础--序列联配

计分矩阵

- **计分矩阵 (scoring matrix)** 是序列联配过程中使用的计分规则, 是序列比对的重要组成部分, 它给出序列联配中碱基或氨基酸匹配或错配值, 故又称替换矩阵 (substitution matrix)。
- DNA 序列相对比较简单, 只有 4 种碱基, 而蛋白质序列有 20 种氨基酸, 如何给出这些氨基酸匹配和错配一个科学准确的评价值, 即准确反映它们的生物学特征, 是生物信息学发展之初就面临的问题, 也是最早被解决的序列联配关键问题。

	A	C	G	T
A	0.9	-0.1	-0.1	-0.1
C	-0.1	0.9	-0.1	-0.1
G	-0.1	-0.1	0.9	-0.1
T	-0.1	-0.1	-0.1	0.9

常用DNA 序列联配的替换矩阵

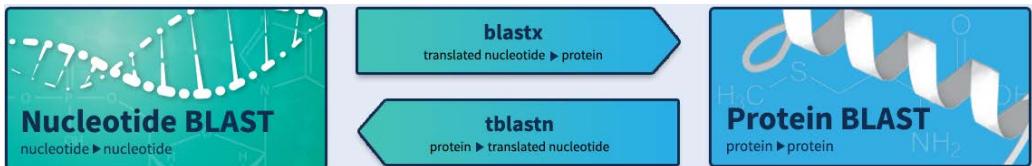
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9	
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	2	3	2	6	2	1	1	1	1	1	2	
F	2	1	2	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3	
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	1	0	1	0	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	3	2	2	1	2	15	1	2	2	3	31	2	
V	7	4	4	4	4	4	5	4	15	10	4	10	5	5	5	5	7	2	4	17

250PAM 突变概率矩阵(Dayhoff 等,1979)



- 生物信息学常用数据库
- 生信的重要基础--序列联配

联配的应用场景--Blast工具



BLAST (Basic Local Alignment Search Tool) Web版 Blast

The screenshot shows the Web版 Blastn interface. At the top, there are tabs for blastn, blastp, blastx, tblastn, and tblastx. The blastn tab is selected. Below the tabs, there are two main sections: 'Enter Query Sequence' and 'Enter Subject Sequence'. Each section has fields for entering accession numbers, FASTA sequences, or uploading files, along with 'From' and 'To' date range inputs. A checkbox 'Align two or more sequences' is checked. In the 'Program Selection' section, the 'More dissimilar sequences (discontiguous megablast)' option is selected. At the bottom, there is a large 'BLAST' button and a note about using Discontiguous megablast for more dissimilar sequences.

Web版 Blastn -双序列比对

➤ **Blast**是生信入门过程中使用频率最高的软件之一了，而且一些软件的原理也是基于序列比对的基础上的。



➤ 生物信息学常用数据库
• 生信的重要基础--序列联配

联配的应用场景--Blast工具

➤ Blast工具使用选择

1. 使用核酸序列检索核酸数据库使用blastn
2. 使用蛋白质序列检索蛋白质数据库使用blastp
3. 使用核酸序列检索蛋白质数据库使用blastx
4. 使用蛋白质序列检索核酸数据库使用tblastn
5. 使用核酸序列检索核酸数据库,但是都翻译成蛋白序列以后再联配使用tblastx

Query sequence type	Database sequence type	Alignment level type	What the program should be called	What the program is actually called
nucleotide	nucleotide	nucleotide	blastNN	blastn
peptide	peptide	peptide	blastPP	blastp
nucleotide	peptide	peptide	blastNP	blastx
peptide	nucleotide	peptide	blastPN	tblastn
nucleotide	nucleotide	peptide	blastNNP	tblastx

不同的需求对应不同的blast工具



➤ 生物信息学常用数据库

- 生信的重要基础--序列联配

Web版 Blast工具使用

**限制比对的物种
勾选exclude则排除该物种**

Query序列粘贴输入

Query序列文件输入 (fasta格式)

数据库选择 可选数据库如下

**比对结果相似性限制
向下依次相似性更低**

**在新页面显示比对结果
而非刷新当前页面 (推荐勾选)**

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear From To

Or, upload file 未选择任何文件

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

BLAST has New Default Parameters and Search Limits. [New](#)

Click [here](#) for more info.

Choose Search Set

Database: Standard databases (nr etc.) rRNA/ITS databases Genomic + transcript databases Betacoronavirus

Nucleotide collection (nr/nt)

Organism: Optional Enter organism name and completions will be suggested exclude

Exclude: Optional Models (XM/XP) Uncultured/environmental sample sequences

Limit to: Optional Sequences from type material

Entrez Query: Optional Enter an Entrez query to limit search [YouTube](#) [Create custom database](#)

Program Selection

Optimize for: Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

Algorithm parameters

Nucleotide collection (nr/nt)
Reference RNA sequences (refseq_rna)
RefSeq Representative genomes (refseq_representative_genomes)
RefSeq Genome Database (refseq_genomes)
Whole-genome shotgun contigs (wgs)
Expressed sequence tags (est)
Sequence Read Archive (SRA)
Transcriptome Shotgun Assembly (TSA)
High throughput genomic sequences (HTGS)
Patent sequences(pat)
PDB nucleotide database (pdb)
Human RefSeqGene sequences(RefSeq_Gene)
Genomic survey sequences (gss)
Sequence tagged sites (dbsts)



➤ 生物信息学常用数据库

- 生信的重要基础--序列联配

Web版 Blast工具使用

新版 Blastn比对结果页面

显示窗口

The screenshot shows the BLASTn search results page with the following details:

- Job Title:** NC_000913.3 Escherichia coli str. K-12 substr....
- RID:** RVJPK4KF014 (Search expires on 10-08 22:04 pm)
- Program:** BLASTN
- Database:** nt
- Description:** NC_000913.3 Escherichia coli str. K-12 substr. MG1655, comp...
- Molecule type:** dna
- Query Length:** 323
- Other reports:** Distance tree of results, MSA viewer

Filter Results

Percent Identity: [] to [] **E value:** [] to [] **Query Coverage:** [] to []

显示窗口 (highlighted in red box) contains the following tabs: Descriptions (selected), Graphic Summary, Alignments, Taxonomy.

Sequences producing significant alignments

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	Shigella flexneri strain FDAARGOS_713 chromosome	597	597	100%	1e-166	100.00%	CP054892.1
<input checked="" type="checkbox"/>	Escherichia coli strain NEB5-alpha_F'lg chromosome, complete genome	597	597	100%	1e-166	100.00%	CP053607.1
<input checked="" type="checkbox"/>	Escherichia coli strain T7Express_LysY chromosome, complete genome	597	597	100%	1e-166	100.00%	CP053597.1
<input checked="" type="checkbox"/>	Escherichia coli strain Dam/Dcm chromosome, complete genome	597	597	100%	1e-166	100.00%	CP053603.1
<input checked="" type="checkbox"/>	Escherichia coli strain T7Express_LysYlq chromosome, complete genome	597	597	100%	1e-166	100.00%	CP053595.1
<input checked="" type="checkbox"/>	Escherichia coli strain T7Express_Crystal chromosome, complete genome	597	597	100%	1e-166	100.00%	CP053594.1
<input checked="" type="checkbox"/>	Escherichia coli strain NEBExpress_lq chromosome, complete genome	597	597	100%	1e-166	100.00%	CP053592.1

过滤一致性和覆盖度 (highlighted in red box)

一致性 (highlighted in red box)

比对上的序列的登录号 (highlighted in red box)

期望值越小越好 (不会为负) (highlighted in red box)

覆盖度 (相对于query seq) (highlighted in red box)

新版 Blastn比对结果页面 (2020年5月中旬更新)



- 生物信息学常用数据库
- 生信的重要基础--序列联配

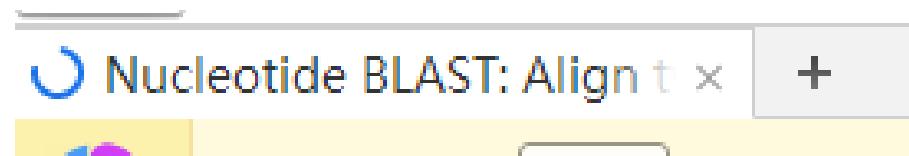
Web版 Blast工具使用

Blast双序列比对

The screenshot shows the 'blastn' tab selected in the top navigation bar. On the left, there is a form for 'Enter Query Sequence' with fields for 'Enter accession number(s), gi(s), or FASTA sequence(s)' and 'Query subrange'. Below these are fields for 'Or, upload file' and 'Job Title'. A red arrow points to the 'Align two or more sequences' checkbox at the bottom of the form. A large blue arrow points to the right, leading to a second screenshot of the same interface.

The second screenshot shows the same 'blastn' tab selected. The 'Align two or more sequences' checkbox is now checked (indicated by a blue checkmark). Other fields like 'Enter Query Sequence', 'Query subrange', 'Or, upload file', 'Job Title', and 'Enter Subject Sequence' are also visible.

Web版 Blast工具使用 -- 双序列比对 (勾选 Align two or more sequences)



页面加载未完成时所有比对操作都无效



➤ 生物信息学常用数据库

• 生信的重要基础--序列联配

➤ BLAST+工具本地化

BLAST+的一般用法如下：

格式化数据库

```
makeblastdb -in db.fasta -dbtype prot -out dbname
```

参数说明:

-in: 待格式化的序列文件

-dbtype: 数据库类型, prot或nucl

-out: 数据库名

蛋白序列比对蛋白数据库 (blastp)

```
blastp -query seq.fasta -out seq.blast -db dbname -outfmt 6 -evalue 1e-5 -num_threads 4
```

参数说明:

-query: 输入文件路径及文件名

-out: 输出文件路径及文件名

-db: 格式化了的数据库路径及数据库名

-outfmt: 输出文件格式, 总共有12种格式, 6是tabular格式对应之前BLAST的m8格式

-evalue: 设置输出结果的e-value值

-num_threads: 线程数

FTP 目录 /blast/executables/blast+/LATEST/ 位于 ftp.ncbi.nlm.nih.gov		
转到高层目录		
06/09/2020 01:56下午	85 ChangeLog	
06/09/2020 01:56下午	53,635,918 ncbi-blast-2.10.1+-1.src.rpm	
06/09/2020 01:56下午	63 ncbi-blast-2.10.1+-1.src.rpm.md5	
06/09/2020 01:56下午	185,108,504 ncbi-blast-2.10.1+-1.x86_64.rpm	
06/09/2020 01:56下午	66 ncbi-blast-2.10.1+-1.x86_64.rpm.md5	
06/09/2020 01:56下午	58,444,148 ncbi-blast-2.10.1+-1.src.tar.gz	
06/09/2020 01:56下午	64 ncbi-blast-2.10.1+-1.src.tar.gz.md5	
06/09/2020 01:56下午	62,897,696 ncbi-blast-2.10.1+-1.src.zip	
06/09/2020 01:56下午	61 ncbi-blast-2.10.1+-1.src.zip.md5	
06/09/2020 01:57下午	91,278,541 ncbi-blast-2.10.1+-win64.exe	
06/09/2020 01:57下午	63 ncbi-blast-2.10.1+-win64.exe.md5	
06/09/2020 01:57下午	235,550,284 ncbi-blast-2.10.1+-x86_64-linux.tar.gz	
06/09/2020 01:57下午	70 ncbi-blast-2.10.1+-x86_64-linux.tar.gz.md5	
06/09/2020 01:57下午	147,981,543 ncbi-blast-2.10.1+-x86_64-macosx.tar.gz	
06/09/2020 01:57下午	71 ncbi-blast-2.10.1+-x86_64-macosx.tar.gz.md5	
06/09/2020 01:57下午	90,981,746 ncbi-blast-2.10.1+-x86_64-win64.tar.gz	
06/09/2020 01:57下午	70 ncbi-blast-2.10.1+-x86_64-win64.tar.gz.md5	
06/09/2020 01:57下午	149,292,474 ncbi-blast-2.10.1+-dmg	
06/09/2020 01:57下午	57 ncbi-blast-2.10.1+-dmg.md5	

win安装版

linux版

win直接运行版

Blast+ 最新版下载

<http://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>

Index of /blast/db

Name	Last modified	Size
Parent Directory		-
FASTA/	2020-10-04 16:43	-
cloud/	2020-02-11 16:27	-
r4/	2020-06-30 10:29	-
r5/	2020-10-07 14:04	-
16S ribosomal RNA.tar.gz	2020-10-06 05:36	35M
16S ribosomal RNA.tar.gz.md5	2020-10-06 05:36	59
16S fungal sequences.tar.gz	2020-10-06 05:37	30M
16S fungal sequences.tar.gz.md5	2020-10-06 05:37	62
nr.UU.tar.gz	2020-10-07 13:30	19G
nr.00.tar.gz.md5	2020-10-07 13:30	47
nr.01.tar.gz	2020-10-07 13:30	1.9G
nr.01.tar.gz.md5	2020-10-07 13:30	47
nr.02.tar.gz	2020-10-07 13:30	1.8G
nr.02.tar.gz.md5	2020-10-07 13:30	47
nr.03.tar.gz	2020-10-07 13:31	2.3G
nr.03.tar.gz.md5	2020-10-07 13:31	47

Blast+ 数据库下载 (文件巨大谨慎下载nrnt数据库在250G左右)

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



➤ 生物信息分析入门基础知识

4 微生物基因组学，比较基因组学

返回目录



➤ 微生物基因组学，比较基因组学

➤ **微生物基因组学**

微生物基因组学研究内容

➤ **比较基因组学**

种间比较基因组学研究

种内比较基因组学研究



➤ 微生物基因组学，比较基因组学

• **微生物基因组学**

微生物基因组学及研究内容

- **微生物基因组学(Microbial genomics)** 是研究基因组组成、进化和生物如何利用基因的一门学科包括两方面内容：
 - 第一方面是**结构基因组学**，研究生物基因组的组成、进化和选择压力；
 - 第二方面是**功能基因组学**，研究基因功能发现、基因表达分析和突变检测。
- **高通量测序技术**为微生物基因组学提供了新的研究方向和丰富的研究材料。基因组测序技术主要应用两种：
 - 第二代测序技术 (Next-Generation Sequencing, **NGS**)，高通量，高精度，短reads；
Illumina测序平台为代表
 - 第三代测序技术，高通量，长读长。**Nanopore的单纳米孔测序，和PacBio的单分子实时测序**
- **全基因组功能注释与分析**，分析整个基因组中**序列的特征和结构，基因的组成特征**



➤ 微生物基因组学，比较基因组学

• 微生物基因组学

微生物全基因组功能注释与分析

基因预测

1.相似性预测

- 已知的 mRNA 或蛋白质序列为线索在 DNA 序列中搜寻

2.从头预测

- 利用统计学模型训练出相应参数 (原核生物主要的预测方式)

NR <https://ftp.ncbi.nlm.nih.gov/blast/db/>

EggNOG <http://eggnog5.embl.de/#/app/home>

Cog <http://www.ncbi.nlm.nih.gov/COG/>

Swiss-Prot <http://www.ebi.ac.uk/uniprot/>

KEGG (KAAS) <https://www.genome.jp/kegg/kaas/>

GO(Gene Ontology) <http://www.geneontology.org/>

详细介绍：附加数据\功能注释数据库详细介绍.pdf

IslandViewer 4 <http://www.pathogenomics.sfu.ca/islandviewer/>

VFDB <http://www.mgc.ac.cn/VFs/main.htm>

CARD <https://card.mcmaster.ca/>

ARDB <http://ardb.cbcn.umd.edu/>

ResFinder 4.1 <https://cge.cbs.dtu.dk/services/ResFinder/>

dbCAN <http://bcb.unl.edu/dbCAN/>

antiSMASH5 [https://antismash.secondarymetabolites.org/#!/start](https://antismash.secondarymetabolites.org/#!/)

蛋白基本注释

各数据库比对

- 基于BLAST算法将样本物种的蛋白序列与公共数据gene进行比较，通过gene的相似性进行功能注释(蛋白水平相似度>30%，
*evalue<1e-5*的注释)
- NR、EggNOG、COG、Swissprot、GO、KEGG

基因组及基因高级注释

1.基因岛

- 基因岛 (Genomics Islands, GIs) 是一些细菌、噬菌体或质粒中有横向起源迹象的一部分基因组。预测工具 IslandViewer 4

2.毒力因子, 耐药基因

- 毒力因子包括细菌毒素, 附着相关的表面蛋白, 水解酶等, 工具 VFDB
- 耐药基因赋予细菌抗生素耐受能力工具 CARD, ARDB (未更新), ResFinder 4.1

3.碳水化合物酶

- CAZy碳水化合物酶相关的专业数据库, 内容包括能催化碳水化合物降解、修饰、以及生物合成的相关酶系家工具 dbCAN

4.次级代谢产物合成基因簇

- antiSMASH可以对细菌和真菌基因组中的次级代谢产物生物合成基因簇进行快速的全基因组鉴定, 注释和分析



➤ 微生物基因组学，比较基因组学

• 微生物基因组学

微生物全基因组功能注释与分析

微生物全基因组功能注释与分析工具一览

分析工具	网址
基因预测:	http://prodigal.ornl.gov/ http://ccb.jhu.edu/software/glimmer/index.shtml
RNamer	http://www.cbs.dtu.dk/services/RNAmer/
tRNAscan	http://lowelab.ucsc.edu/tRNAscan-SE/
Go功能富集DAVID	http://david.abcc.ncifcrf.gov/ http://rast.nmpdr.org/ https://www.patricbrc.org/
功能注释	http://tandem.bu.edu/trf/trf407b.linux.download.html
trf预测	http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html http://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/
CPG岛预测	http://bioinfo-mml.sjtu.edu.cn/TADB/
TADB	http://tubic.tju.edu.cn/deg/
必须基因注释	http://csbl.bmb.uga.edu/DOOR/index.php http://operondb.cbcn.umd.edu/cgi-bin/operondb/operons.cgi
操纵子预测:	http://www.islander.com/
基因岛预测	IslandViewer: http://www.pathogenomics.sfu.ca/islandviewer/query.php
MobilomeFINDER	http://db-mml.sjtu.edu.cn/MobilomeFINDER/
CRISPR预测	http://crispr.u-psud.fr/Server/Advanced.CRISPRfinder.php
预测启动子	http://www-bimas.cit.nih.gov/molbio/proscan/
预测致病岛	https://www.gem.re.kr/paidb/about_paidb.php
预测分析转录终止信号	http://linux1.softberry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter
预测噬菌体	http://phaster.ca
预测信号肽	http://www.cbs.dtu.dk/services/SignalP/



- 微生物基因组学，比较基因组学
- **比较基因组学**

- 比较基因组学（Comparative genomics）是在已知基因组数据的基础上，对已知基因或基因组数据进行比较，以此来预测基因的功能、相互之间的作用关系以及物种的进化等的学科。
- 种间比较基因组学研究
 - 横向基因转移（Horizontal gene transfer, HGT）外源功能DNA片段的横向转移可以导致目标菌株产生遗传变异，直接使目标菌株获得新的生物学性。细菌的接合、转化以及噬菌体携带外源片段插入等均能导致细菌间发生基因横向转移，其中，由噬菌体介导的是主要方式
 - 横向基因转移的预测方法，基本准则是某基因的遗传方向和菌株的系统发育方向不一致。
- 种内比较基因组学研究
 - DNA序列多态性：由于菌株自身的突变，产生的序列水平上的多样性，如SNPs, InDels等
 - 泛基因组：在物种水平来研究一种细菌的全部遗传特性，而非局限于某部分基因组，包含core genes, accessory genes, Unique genes
 - 上述分析内容结合表型数据就能够解释菌株生物学特性，阐释基因功能。



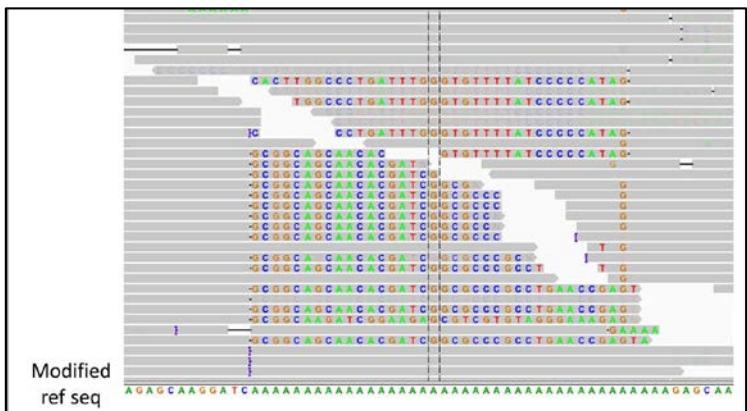
➤ 微生物基因组学，比较基因组学

• 比较基因组学

DNA 序列多态性

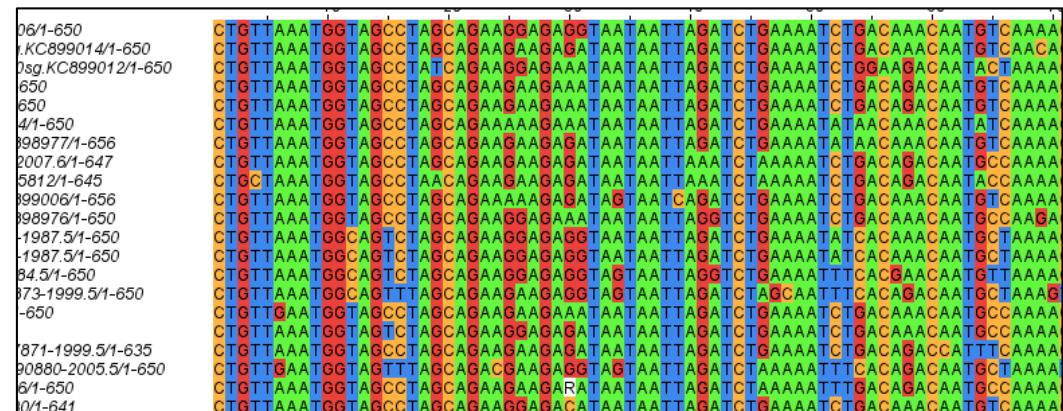
基于原始数据

- 将NGS测序后reads（即NGS测序结果）直接比对或者组装后比对到参考基因组
- GATK (<https://gatk.broadinstitute.org/hc/en-us>)
- Snippy (<https://github.com/tseemann/snippy>)



基于基因组

- 将组装好的基因组进行多序列比对，根据比对的结果提取snps, indels等信息。
- MUMmer3套件 (<http://mummer.sourceforge.net/>)





➤ 微生物基因组学，比较基因组学

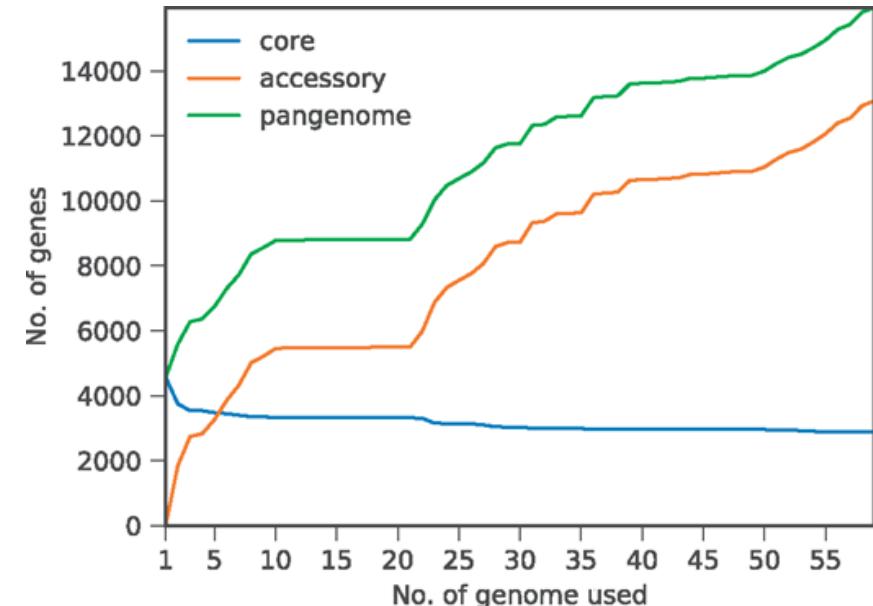
• 比较基因组学

泛基因组

- **核心基因** (Core genes) : 主要为信息基因和管家基因，其功能主要为生命活动必需的代谢和功能
- **非必须基因** (accessory genes) : 与菌株特定的生存环境相关，影响菌株的毒力、抗生素抗性、血清型、抗逆性等特征。
- **特异性基因** (Strain special genes):主要是由菌株横向基因转移或基因缺失引起。

增长趋势划分

- 开放型泛基因组 (Open pan-genome)
- 封闭型泛基因组 (Closed pan-genome)



随着大肠杆菌基因组数量的增加，泛基因组大小、核心基因簇数和附属基因簇数的增长速度

(Her et al. Bioinformatics 2018)

泛基因组优势：能最大程度的阐释某个物种的遗传特征，以及菌株间的差异和多样性，能够在多个角度进行分析，是比较基因组学分析的高级版本。



➤ 微生物基因组学, 比较基因组学

• 比较基因组学

泛基因组构建方法

- 同种菌株间基因或蛋白序列进行**同源性比较**, 构建同源基因簇, 进而构建泛基因组
- 近几年文献中报道使用的泛基因组构建软件主要有: pgap-X, **BPGA**, **Roary**, PanOCT, OrthoFinder and OrthoMCL, 从易用性, 流程完备性来讲**BPGA**, **Roary**比较好的, 引用分别是241次, 1345次

[HTML] BPGA-an ultra-fast pan-genome analysis pipeline

[NM Chaudhari, VK Gupta, C Dutta - Scientific reports, 2016 - nature.com](#)

Recent advances in ultra-high-throughput sequencing technology and metagenomics have led to a paradigm shift in microbial genomics from few genome comparisons to large-scale pan-genome studies at different scales of phylogenetic resolution. Pan-genome studies ...

☆ 99 被引用次数: 241 相关文章 所有 12 个版本

Roary: rapid large-scale prokaryote pan genome analysis

[AJ Page, CA Cummins, M Hunt, V](#)

A typical prokaryote population se of isolates. Interrogating these da structure of prokaryotic genomes.

☆ 99 被引用次数: 1345 相

[Roary: rapid large-scale prokaryote pan genome analysis](#)

在引用文章中搜索

Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread

[S David, S Reuter, SR Harris, C Glasner, T Feltwell... - Nature ..., 2019 - nature.com](#)

Public health interventions to control the current epidemic of carbapenem-resistant Klebsiella pneumoniae rely on a comprehensive understanding of its emergence and spread over a wide range of geographical scales. We analysed the genome sequences and ...

☆ 99 被引用次数: 92 相关文章 所有 8 个版本

谢谢！