

FILE FORMATS



INTRODUCTION


- File formats is a part of bioinformatics that stores DNA and protein sequences which are of many types.
- The types of file format is used according to the different context.
- Following are some important types of file format:
 1. GenBank
 2. FASTA Format
 3. CluatalW2

GenBank

NCBI Resources How To Sign in to NCBI

GenBank

GenBank ▾ Submit ▾ Genomes ▾ WGS ▾ Metagenomes ▾ TPA ▾ TSA ▾ INSDC ▾ Other ▾

 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/>.

GenBank Overview

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). See [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programatically using [NCBI e-utils](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

GenBank Resources

[GenBank Home](#)

[Submission Types](#)

[Submission Tools](#)

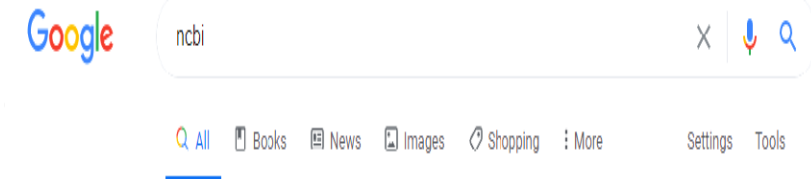
[Search GenBank](#)

[Update GenBank Records](#)

THE PROCEDURE IS SHOWN BELOW.

The image shows a Google search interface. The search bar contains the text 'ncbi'. Below the search bar, there are navigation links for 'All', 'Books', 'News', 'Images', 'Shopping', and 'More', along with 'Settings' and 'Tools'. The search results show 'About 10,40,00,000 results (0.56 seconds)'. The top result is for 'www.ncbi.nlm.nih.gov', titled 'National Center for Biotechnology Information'. A description follows: 'Welcome to NCBI. The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information. You've visited this page many times. Last visit: 14/10/20'. Below this are several related links: 'PubMed' (30 million citations), 'Nucleotide' (collection of sequences), 'Gene' (RefSeqGene, Workbench), 'All Resources' (NIH genetic sequence database), 'BLAST' (Basic Local Alignment Search Tool), and 'National Center for ...' (The National Center for Biotechnology Information). A 'People also ask' section is visible at the bottom left. On the right side, a knowledge panel for 'National Center for Biotechnology Information' is displayed, including the NCBI logo, the website URL 'ncbi.nlm.nih.gov', and a detailed description: 'The National Center for Biotechnology Information is part of the United States National Library of Medicine, a branch of the National Institutes of Health. The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper. Wikipedia'. Additional information includes 'Founder: Claude Pepper', 'Founded: 4 November 1988', 'Headquarters location: Bethesda, Maryland, United States', and 'Parent organization: National Library of Medicine'.

Search for NCBI in Google page.



www.ncbi.nlm.nih.gov
National Center for Biotechnology Information

Welcome to NCBI. The **National Center for Biotechnology Information** advances science and health by providing access to biomedical and genomic information.
You've visited this page many times. Last visit: 14/10/20

PubMed

PubMed® comprises more than 30 million citations for biomedical ...

Nucleotide

The Nucleotide database is a collection of sequences from ...

Gene

RefSeqGene - Genome
Workbench - Statistics - ...

All Resources

The NIH genetic sequence database, an annotated ...

BLAST

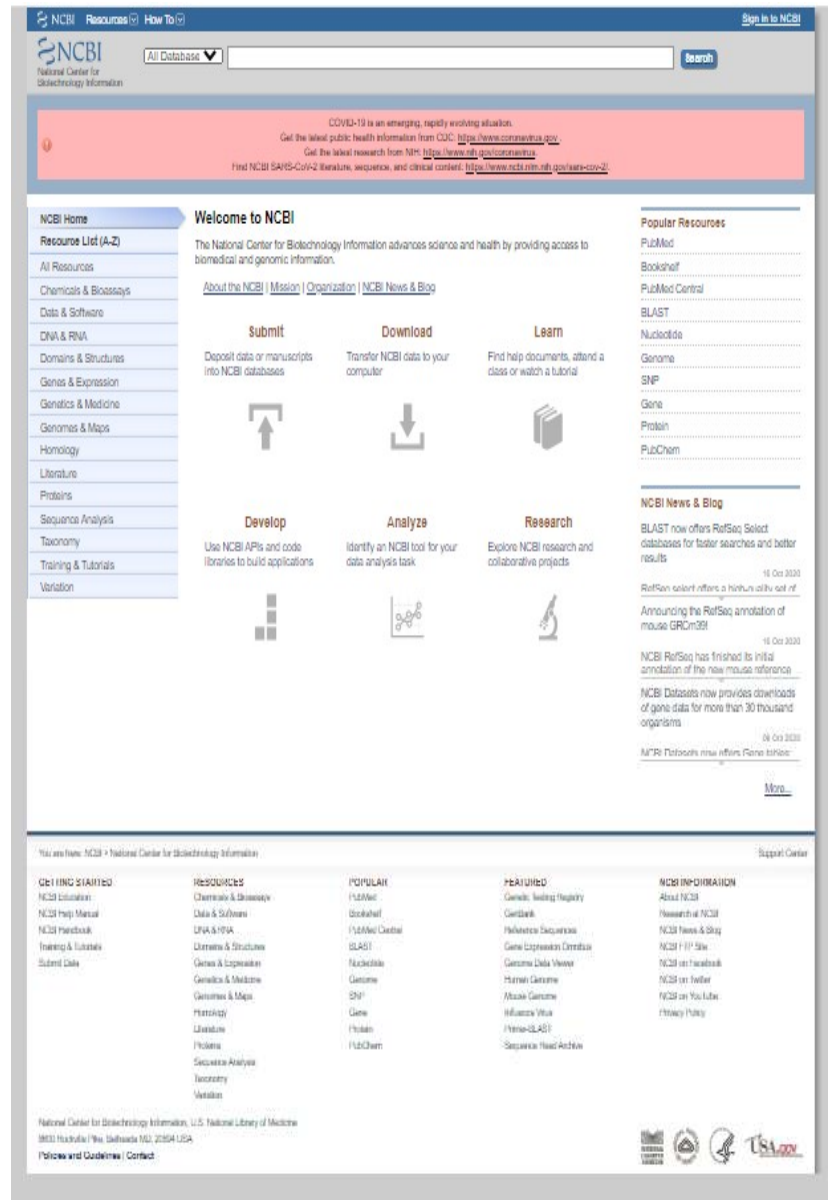
The Basic Local Alignment Search Tool (BLAST) finds regions of ...

National Center for ...

The National Center for Biotechnology Information ...

[More results from nih.gov »](#)

People also ask



Enter into the website.

NCBI home page.

Nucleotide

Search

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/>.

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or manuscripts into NCBI databases



Download

Transfer NCBI data to your computer



Learn

Find help documents, attend a class or watch a tutorial



Develop

Use NCBI APIs and code

Analyze

Identify an NCBI tool for your

Research

Explore NCBI research and

Popular Resources

[PubMed](#)

[Bookshelf](#)

[PubMed Central](#)

[BLAST](#)

[Nucleotide](#)

[Genome](#)

[SNP](#)

[Gene](#)

[Protein](#)

[PubChem](#)

NCBI News & Blog

BLAST now offers RefSeq Select databases for faster searches and better results

Enter nucleotide in the pop up box and search for an organism.
Here I've taken Rice.

EST (1,277,677)
GSS (822,690)

Genetic compartments
Chloroplast (3,011)
Mitochondrion (272)
Plasmid (152)
Plastid (3,045)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

[Clear all](#)
[Show additional filters](#)

Items: 1 to 20 of 2909523

<< First < Prev Page 1 of 145477 Next > Last >>

[Triticum aestivum chromosome 3B, genomic scaffold, cultivar Chinese Spring](#)

1. 774,434,471 bp linear DNA
Accession: HG670306.1 GI: 669026884
[BioProject](#) [Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Digitaria exilis annotation](#)

2. 48,169,323 bp linear DNA
Accession: LR792837.1 GI: 1833616914
[BioProject](#) [Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Oryza sativa Japonica Group genomic DNA, chromosome 1, complete sequence](#)

3. 43,342,410 bp linear DNA
Accession: BA000010.8 GI: 55417890
[Assembly](#) [BioProject](#) [Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Digitaria exilis annotation](#)

4. 52,445,811 bp linear DNA
Accession: LR792836.1 GI: 1833511964
[BioProject](#) [Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Digitaria exilis annotation](#)

5. 38,778,756 bp linear DNA
Accession: LR792829.1 GI: 1833595496
[BioProject](#) [Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

"Oryza sativa"[Organism] OR Oryza sativa[All Fields]

Search

See more...

Recent activity

[Turn Off](#) [Clear](#)

[Oryza sativa \(2909523\)](#) Nucleotide

[rice \(5477782\)](#) Nucleotide

[Oryza sativa cultivar 9311 chromosome 1, whole genome shotgun sequence](#) Nucleotide

[ASM1463503v1 - Genome - Assembly - NCBI](#) Assembly

[Oryza sativa cultivar 9311 chromosome 12, whole genome shotgun sequence](#) Nucleotide

See more...

Enter on the GenBank/FASTA sequence of Oryza sativa japonica group.

GenBank Format

```
GenBank: M24845.1
FASTA  Gen2Seq

LOCUS       M24845.1                2883 bp            ssNA            linear            PLN 27-APR-1993
DEFINITION  Malva embry globulin 5 allele (75-like) ssNA, complete cds.
ACCESSION   M24845
VERSION     M24845.1
KEYWORDS    globulin.
SOURCE      Zea mays
  ORGANISM  Zea mays
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Rosales; Fabales; Fagaceae; Fagaceae;
            clade; Panicoideae; Andropogoneae; Andropogonoideae; Tripsacinae;
            Zea.
REFERENCE   1 (bases 1 to 2883)
AUTHORS    Belanger, F.C., and Kriz, A.L.
TITLE      Molecular Characterization of the Major Malva Embryo Globulin
  Encoded by the Glb1 Gene
JOURNAL    Plant Physiol 91 (2), 636-643 (1989)
PUBMED     1667888
COMMENT    Original source text: Malva (inbred line Va26) 27 DAP embryos. cDNA
  to ssNA, clone pcGlb15.
  Draft entry and computer-readable sequence for [1] kindly provided
```

```
FEATURES             Location/Qualifiers
     source           1..2883
                     /organism="Zea mays"
                     /mol_type="ssNA"
                     /db_xref="taxon:4577"
     ssNA             1..2883
     CDS              37..1758
                     /note="globulin precursor"
                     /codon_start=1
                     /protein_id="AAA33467.1"
                     /translation="MVSARIIVVLLAVLLCAAAAVASSWEDQKHHHGGHKSGRGVRRRC
EDRFWQWPNFLDQCRSEERERKRSRSHNDORSGGSSSDSRSRDEQKKEKQKDR
PYVFDKRSFRVYVRS EGGSLRVLRHVFDEVSRLLRGIRDYRVAVLLEANRYSFVAVSHTD
ANCIQYVAEGEGVVTIEMGERKSYTIKQGMFVAPAGAVTYLANIDGRNKLVTIKIL
HTISVPGEFQFFGPGGRNPFESFLSSFSKSDQRAAYKTSDDRLERLFRGHGQDQKGIIV
RATESQTRELRRHAS EGGHGFHVP LPPYFGESRGPYSLLDQRFSIANQHGQLYEADARS
FHDLAIEHDSVSVFANITAGSRSAPLVNTRSFKIAVYVWQKGYAEIVCFHNSQGGSE
RERKQDRHSEEESESSEQSEVGDQWHTLRALSPGTAIFYVPAQHPFVAWASRDSNLQ
IVCFEYHADRNRYL LAGADNV LKLDYRVALSFAKAAEEVDEVVLGSRNKGFLPGY
KESQGHHEERQEEEEERERHGGKGERENHGREEREKEEEEEERGRHGRGRREIWAETLL
RNVTAAP"
     sig_peptide     37..98
                     /note="globulin signal peptide"
     prot_initiator  295..1755
```

```
ORIGIN
1  Cggcacacacc  cggagcatatc  acagtgacac  tacacgatgg  tgagcgcag  aatcgttgtc
61  ctctctatg  cctctctatg  cctctctatg  cctctctatg  cctctctatg  cctctctatg
121  caccaccacc  accaccacc  accaccacc  accaccacc  accaccacc  accaccacc
181  taggacacag  gcccccggtg  cctggagcag  tacagagag  agagacagaa  ggaagcagaa
241  gtagccagac  gggccagagc  cggccagagc  agagccagag  gctcctcag  ggttagcagc
301  gtagccttc  gtcgagtggt  cggagcagag  caggagctcc  tgagagtgct  ccgagcgttc
361  gacagagtag  ccaggctcct  ccgagcctcc  caggagctcc  gcgttagcct  cctgagagcc
421  aacccgagct  cgttcgtggt  gccagcagc  accgagcagc  accgagcagc  accgagcagc
481  gtaggtagag  gtaggtagag  gtaggtagag  gtaggtagag  gtaggtagag  gtaggtagag
541  caagagcagc  tctctctatg  cctctctatg  cctctctatg  cctctctatg  cctctctatg
601  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
661  ctctctctc  gcccccggtg  gggccagagc  gggccagagc  gtagccttc  gtagccttc
721  ctctctctc  ctctctctc  ctctctctc  ctctctctc  ctctctctc  ctctctctc
781  atccagagag  ctctctctc  ctctctctc  ctctctctc  ctctctctc  ctctctctc
841  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
901  cagccttc  cagccttc  cagccttc  cagccttc  cagccttc  cagccttc
961  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1021  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1081  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1141  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1201  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1261  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1321  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1381  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1441  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1501  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1561  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1621  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1681  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1741  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1801  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1861  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1921  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
1981  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc  gtagccttc
```

HEADER

FEATURE TABLE

SEQUENCE

Bank Header section

4845.1

[hics](#)

MZEGLB1SA 2003 bp mRNA linear PLN 27-APR-1993

Maize embryo globulin S allele (7S-like) mRNA, complete cds.

M24845

M24845.1

globulin.

Zea mays

[Zea mays](#)

Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;

Spermatophyta; Magnoliopsida; Liliopsida; Poales; Poaceae; PACMAD

clade; Panicoideae; Andropogonodae; Andropogoneae; Tripsacinae;

Zea.

[1] (bases 1 to 2003)

Belanger, F.C. and Kriz, A.L.

Molecular Characterization of the Major Maize Embryo Globulin

Encoded by the Glb1 Gene

Plant Physiol 91 (2), 636-643 (1989)

[16667080](#)

Original source text: Maize (inbred line Va26) 27 DAP embryos, cDNA

to mRNA, clone pcGlb1S.

Draft entry and computer-readable sequence for [1] kindly provided

by A.L.Kriz, 16-MAY-1989.

GenBank sequence section

Sequence field identifier

ORIGIN

```
1 cgcacacacc cgagcatatc acagtgcac tacacgatgg tgagcgccag
61 ctctcgccg tcctcctatg cgctgccgcc gcagtcgcgt cgtcctggga
121 caccaccacc acgggggcca caagtccggg cgatgcgtgc ggccggtcga
181 tggcaccagc gccccgggtg cctggagcag tgcagggagg agggcgggga
241 gagcggagca ggcacgaggc cgacgaccgc agcggcgagg gctcgtcgga
301 gagcgcgagc aggagaagga ggagaagcag aaggaccggc ggccgtactgt
361 cgagccttc gtcgcgtggt ccggagcgag cagggggtccc tgagggtgct
421 gacgaggtgt ccaggctcct ccgcgccatc cgggactacc gcgtggcggt
481 aaccgcgct cgttcgtggt gccagccac accgacgcgc actgcatcgg
541 gaaggcgagg gagtgggtgac gacgatcgag aacggcgaga ggccggtcgt
601 caaggccacg tcttcgtggc gccggccggg gcggtcacct acctggccaa
661 cggaagaaac tggatcacac caagatcctc cataccatct ccgtgcctgg
721 ttcttcttcg gccccggcgg gaggaacccc gaatcgttcc tgtcagactt
781 atccagagag ctgcgtacaa gacctcgagc gaccggctgg agaggctgtt
841 gggcaggaca aggggatcat cgtgcgtgcc acggaggagc agaccgcgga
901 cacgcctcgg agggcggcca cggcccgcac tggcccctgc cgccgttcgg
961 ggcccctaca gctcctgga ccagcggccc agcatcgcca accagcacgg
1021 gaggccgacg cgcgcagctt ccacgacctc gccgagcacg acgtcagcgt
1081 aacatcaccg cgggttccat gagcgcgcca ttgtacaaca cccgttcgtt
1141 tacgtccga acggcaaggg ctacgccgag atcgtgtgcc cgcaccgcca
1201 ggcgagagcg agcgcgagc cggcaagggc agggagagc aagaagaaga
1261 gaggagcagg aggaagtcgg gcaggggtac cacaccatcc gggcgcggct
1321 acggcgttcg tggtgcccgc gggccacccg ttcgtcgcgg tggcgtcccg
1381 ctccagatcg tgtgcttcga ggtccacgcc gacaggaacg agaagggttt
1441 gccgacaacg tgctgcagaa gctcgaccgg gtcgccaagg cgctgtcatt
1501 gcgaggaggg tggacgaggt gctcggctcg cggcgcgaga aggggttcct
1561 aaggagagcg gcggccacga ggagcgggag caggaggagg aggaacgcga
1621 ggcgggcgtg gggagaggga acgccacgga cgtgaggagc gggagaaaag
1681 cgcaaggac gccacggccg cgggcgccgc gaggaagtgg cggagacgct
1741 gtgaccgcca ggatgtgagg ccggccgtgc tcgcaaaac gacgaggaag
1801 gtggcgcgcg accgacgtgc gtacgtagca tgagcctgag tggagacgtt
1861 gtatatacct ctctgcgtgt taactatgta cgtaagcggc aggcagtgca
1921 ctctgtagta tgtacgtgcg ggtacgatgc tgtaagctac tgaggcaagt
1981 ataatgcac gtgcgtgttc tat
```

//

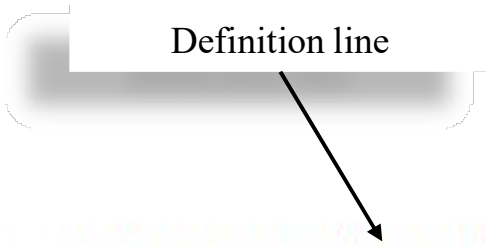
Termination line

FASTA Format

- FASTA format, the widely used format and can be easily read by wide range of programs.
- It is a standard format used for DNA and protein sequences.
- Basically, represented as '>'.
>
M...
...
>
M...
...

FASTA Format

Definition line



```
>M24845.1 Maize embryo globulin S allele (7S-like) mRNA, complete cds  
CGCACACACCCGAGCATATCACAGTGACACTACACGATGGTGAGCGCCAGAATCGTTGTCCTCCTCGCCG  
TCCTCCTATGCGCTGCCGCCGCAGTCGCGTCGTCTGGGAGGACGACAACCACCACCACCACGGGGGCCA  
CAAGTCCGGGCGATGCGTGCGGCGGTGCGAGGACCGGCCCTGGCACCAGCGCCCCGGTGCCTGGAGCAG  
TGCAGGGAGGAGGAGCGGGAGAAGCGGCAAGAGCGGAGCAGGCACGAGGCCGACGACCGCAGCGGCGAGG
```

ClustalW2

ClustalW2

[Input form](#)[Web services](#)[Help & Documentation](#)[Bioinformatics Tools FAQ](#)[Feedback](#)[Share](#)

[Tools](#) > [Multiple Sequence Alignment](#) > [ClustalW2](#)

ClustalW2 is a general purpose DNA or protein multiple sequence alignment program for **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

Please Note

The ClustalW2 services have been retired. To access similar services, please visit the [Multiple Sequence Alignment tools](#) page. For protein alignments we recommend [Clustal Omega](#). For DNA alignments we recommend trying [MUSCLE](#) or [MAFFT](#). If you have any questions/concerns please contact us via the [feedback](#) link above.

ClustalW2 is a general purpose DNA or protein multiple sequence alignment program for three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Thank You