



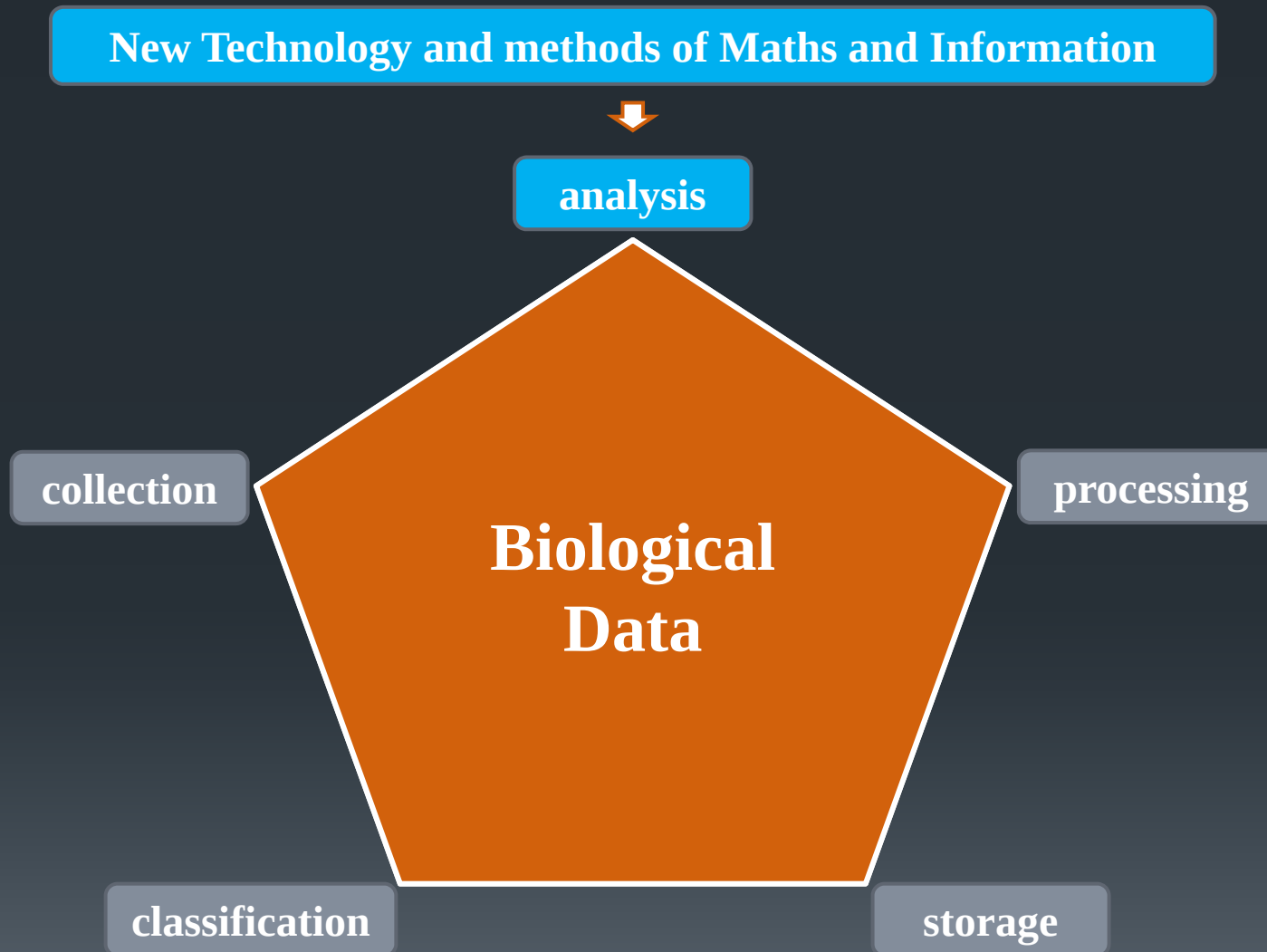
# Experiment Guide of Bioinformatics

**-农学院生物信息学教学实验课**

王煜 (Yu Wang)

Email: [bioinplant@zju.edu.cn](mailto:bioinplant@zju.edu.cn)

# Research Contents in Bioinformatics



# Contents



## ◆ Introduction of biological database

- ◆ Reference, DNA, protein, structure, et al.
- ◆ Public database, personal database, et al.
- ◆ Raw data, annotation information for raw data

## ◆ Search/Query in database (genBank)

- ◆ Reference (PubMed)
- ◆ DNA sequences
- ◆ Protein sequences

## ◆ Analysis and explanation of biological data

- ◆ Homologous sequence analysis and multiple sequence alignment
- ◆ Phylogenetic analysis
- ◆ Gene prediction and annotation
- ◆ Genome analysis

# ◆ Biological database

- ◆ genBank/EMBL/DDBJ <http://www.ncbi.nlm.nih.gov/>
  - ◆ dbEST/UniGene/dbSTS/dbGSS/HTG/Taxonomy/dbSNP
- ◆ SWISS-PROT/UniProt/PIR/Prosite <http://www.ebi.ac.uk/swissprot/>
- ◆ PDB <http://www.rcsb.org>
- ◆ Specialized species database: rice (RGAP), arabidopsis (TAIR), et al.
- ◆ Gene annotation database
  - ◆ GO <http://www.geneontology.org/>
  - ◆ KEGG <http://www.genome.ad.jp/kegg/>
  - ◆ PKR <http://www.nih.go.jp/mirror/Kinases/>
  - ◆ COG <http://www.ncbi.nlm.nih.gov/COG>
- ◆ PubMed/Web of science/Agricola
- ◆ Submission
  - ◆ BankIt
  - ◆ Sequin

# GenBank

## The divisions (分支) of GenBank

| Abbreviation | Full name                                        |
|--------------|--------------------------------------------------|
| PRI          | 灵长类序列 (primate sequences)                        |
| ROD          | 啮齿类序列 (rodent sequences)                         |
| MAM          | 其它哺乳类序列 (other mammalian sequences)              |
| VRT          | 其它脊椎动物序列 (other vertebrate sequences)            |
| INV          | 无脊椎动物序列 (invertebrate sequences)                 |
| PLN          | 植物、真菌和海藻类序列 (plant, fungal, and algal sequences) |
| BCT          | 细菌序列 (bacterial sequences)                       |
| VRL          | 病毒序列 (viral sequences)                           |
| PHG          | 噬菌体序列 (bacteriophage sequences)                  |



| Abbreviation | Full name                                    |
|--------------|----------------------------------------------|
| SYN          | 合成序列 (synthetic sequences)                   |
| UNA          | 未注释的序列 (unannotated sequences)               |
| EST          | 表达序列标签 (expressed sequence tags)             |
| PAT          | 已专利的序列 (patent sequences)                    |
| STS          | 序列标签位点 (sequence tagged sites)               |
| GSS          | 基因组序列 (genome survey sequences)              |
| HTG          | 高产出基因组序列 (high throughput genomic sequences) |
| HTC          | 高产出cDNA序列 (high throughput cDNA sequences)   |

# GenBank Format

```

LOCUS       EF069996               536 bp    DNA        linear    PLN 08-JUN-2007
DEFINITION  Oryza sativa (japonica cultivar-group) strain IRGC seed accession
            no. Nipponbare granule-bound starch synthase (waxy) gene, partial
            cds.
ACCESSION   EF069996
VERSION     EF069996.1  GI:117297404
KEYWORDS    .
SOURCE      Oryza sativa Japonica Group
  ORGANISM  Oryza sativa Japonica Group
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BEP
            clade; Ehrhartoideae; Oryzoideae; Oryza.

REFERENCE   1  (bases 1 to 536)
  AUTHORS   Zhu,Q., Zheng,X., Luo,J., Gaut,B.S. and Ge,S.
  TITLE     Multilocus analysis of nucleotide variation of Oryza sativa and its
            wild relatives: severe bottleneck during domestication of rice
  JOURNAL   Mol. Biol. Evol. 24 (3), 875-888 (2007)
  PUBMED    17218640
REFERENCE   2  (bases 1 to 536)
  AUTHORS   Zhu,Q.-H., Zheng,X.-M., Luo,J.-C., Gaut,B.S. and Ge,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-OCT-2006) Laboratory of Systematic and Evolutionary
            Botany, Institute of Botany, The Chinese Academy of Sciences,
            Xiangshan, Nanxincun 20, Beijing 100093, China

FEATURES             Location/Qualifiers
     source            1..536
                     /organism="Oryza sativa Japonica Group"
                     /mol_type="genomic DNA"
                     /strain="IRGC seed accession no. Nipponbare"
                     /db_xref="taxon:39947"
                     /country="Japan"
     gene              <1..536
                     /gene="waxy"
     mRNA              join(<1..74,432..536)
                     /gene="waxy"
                     /product="granule-bound starch synthase"
     CDS               join(<1..74,432..536)
                     /gene="waxy"
                     /codon_start=3
                     /product="granule-bound starch synthase"
                     /protein_id="ABK33399.1"
                     /db_xref="GI:117297405"
                     /translation="IRVVGTPAYEEMVRNCLNQDLSWKGPKNWENVLLGLGVAGSAP
            GIEGDEIAPLAKENV"

ORIGIN
1  ccatacaaggt  cgtcgccacg  cggcggtacg  aggagatggt  caggaaactgc  atgaaccagg
61  acctctcctg  gaaggtataa  attacgaaac  aaatttaacc  caaacatata  ctatatactc
121  cctccgcttc  taaatatcca  acgcggttgt  ctttttttaa  tatgtttgac  cattcgcttc
181  attaaaaaaa  ttaataaatt  ataaattctt  ttcctatcat  ttgattcatt  gttaaaatata
241  cttatatgta  tacatatagt  tttacatatt  tcataaaaatt  ttttgaacaa  gacgaacggt
301  caaacatgtg  ctaaaaagtt  aacggtgtcg  aatattcaga  aacggaggga  gtataaacgt
361  cttgttcaga  agttcagaga  ttcacctgtc  tgatgctgat  gatgattaat  tgtttgcaac
421  atggatttca  ggggcctgcg  aagaactggg  agaattgtgt  cctgggcctg  ggcgtcgccg
481  gcagcgcgcc  ggggatcgaa  ggcgacgaga  tcgcgccgct  cgccaaggag  aacgtg

```

# FASTA Format

```

>gi|117297404|gb|EF069996.1| Oryza sativa (japonica cultivar-group) strain IRGC
CCATCAAGGTCGTCGGCAGCCCGGCGTACGAGGAGATGGTCAGGAACTGCATGAACCAGGACCTCTCCTG
GAAGGTATAAATTACGAAACAAATTTAACCCAAACATATACTATATACTCCCTCCGCTTCTAAATATTCA
ACGCCGTTGTCTTTTTTAAATAATGTTTGACCATTCTGCTTATTAATAAATAAATAAATATAAATCTTT
TTCCTATCATTTGATTTCATTGTTAAATATACTTATATGTATACATATAGTTTTACATATTTTATAAATTT
TTTTGAACAAGACGAACGGTCAAAACATGTGCTAAAAAGTTAACGGTGTGGAATATTGAGAAACGGAGGGA
GTATAAACGTCCTTGTTCAGAAGTTCAGAGATTACCTGTCTGATGCTGATGATGATTAAATTGTTTGAAC
ATGGATTTTCAAGGGCCTGCCAAGAACTGGGAGAATGTGCTCCTGGGCCTGGGCGTCCCGGCAGCGCGCC
GGGGATCGAAGGCGACGAGATCGCGCGCTCGCCAAGGAGAACGTG

```

# ◆ Search/Query in database

## ◆ Text-based database searching

- ◆ Noun, descriptive words, phrase
- ◆ Boolean logical operators: AND, OR, NOT
- ◆ **Entrez**, Sequence Retrieval System (SRS), DBGET
- ◆ Tips:
  - ◆ 1. Quote the words in a phrase: “disease resistance”
  - ◆ 2. disease resistance = disease AND resistance
  - ◆ 3. enzyme\* = enzyme + enzymes

## ◆ Sequence-based database searching

- ◆ Essence: sequence alignment
- ◆ Purpose: function analysis, evolution, detect SNPs and INDELs, elongation and location the sequence
- ◆ **BLAST** (blastn, blastp, tblastx, psi-blast)/FASTA/BLAT
- ◆ Tips:
  - ◆ Identity and similarity
  - ◆ E value
  - ◆ Bl2seq, Primer-BLAST

# A snapshot for Entrez

NCBI

Entrez, The Life Sciences Search Engine

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases    [Help](#)

- Result counts displayed in gray indicate one or more terms not found

|                                                                                       |                                                                   |
|---------------------------------------------------------------------------------------|-------------------------------------------------------------------|
| 38420 PubMed: biomedical literature citations and abstracts                           | 497 Books: online books                                           |
| 37249 PubMed Central: free, full text journal articles                                | 11229 Images: images from full text resources at NCBI             |
| 24 Site Search: NCBI web and FTP sites                                                | 153 OMIM: online Mendelian Inheritance in Man                     |
| 335235 Nucleotide: Core subset of nucleotide sequence records                         | 2218 dbGaP: genotype and phenotype                                |
| 1540421 EST: Expressed Sequence Tag records                                           | 41319 UniGene: gene-oriented clusters of transcript sequences     |
| 1027982 GSS: Genome Survey Sequence records                                           | 49 CDD: conserved protein domain database                         |
| 442805 Protein: sequence database                                                     | 996 UniSTS: markers and mapping data                              |
| 245 Genome: whole genome sequences                                                    | 1366 PopSet: population study data sets                           |
| 399 Structure: three-dimensional macromolecular structures                            | 571020 GEO Profiles: expression and molecular abundance profiles  |
| 1 Taxonomy: organisms in GenBank                                                      | 401 GEO DataSets: experimental sets of GEO data                   |
| 5418373 SNP: single nucleotide polymorphism                                           | 1 Epigenomics: Epigenetic maps and data sets                      |
| none dbVar: Genomic structural variation                                              | 4 Cancer Chromosomes: cytogenetic databases                       |
| 98787 Genes: gene-centered information                                                | 338 PubChem BioAssay: bioactivity screens of chemical substances  |
| 636 SRA: Sequence Read Archive                                                        | 5 PubChem Compound: unique small molecule chemical structures     |
| 603 BioSystems: Pathways and systems of interacting molecules                         | 195 PubChem Substance: deposited chemical substance records       |
| 10677 HomoloGene: eukaryotic homology groups                                          | 12990 Protein Clusters: a collection of related protein sequences |
| none GENSAT: gene expression atlas of mouse central nervous system                    | none Peptidome: MS/MS proteomic experiments                       |
| 110138 Probe: sequence-specific reagents                                              | none OMIA: online Mendelian Inheritance in Animals                |
| 74 Genome Project: genome project information                                         |                                                                   |
| 1089 NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections | 51 MeSH: detailed information about NLM's controlled vocabulary   |

## Blastn alignment

```

Score = 169 bits (91), Expect = 2e-36
Identities = 117/130 (90%), Gaps = 0/130 (0%)
Strand=Plus/Plus

Query 17450  TGCATGGTTTCCTTGTGCCAGTGTTTCGAGCTGGGCTTGCTTTGTTTCAGCCGACTCCCCAG 17509
          ||||||| ||||||| ||||||| ||||||| ||||||| ||||||| ||||||| ||
Sbjct 19594  TGCATGGCTTCCTTGTGCCAGTGTTTCGAGCTGGGCTTGCTCTGTTTCGGCTGACTCCCCCG 19653

Query 17510  AGCAAAGGATGGTGATGAGTGACGTGGTTGTGACACTGAAGAAGATTAGGAAGGACTATG 17569
          ||||||| || ||||||| || ||||||| || ||||||| || ||||||| || |||||||
Sbjct 19654  AGCAAAGGACGGCGATGAGCGATGTGGTCGTGACACTGAAGAAGATTACAAAGGACTATG 19713

Query 17570  TCAAATCGAT 17579
          ||||| |||
Sbjct 19714  TCAAATTGAT 19723
  
```

## Blastp alignment

```

Score = 1722 bits (4461), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 890/1103 (81%), Positives = 963/1103 (88%), Gaps = 6/1103 (0%)

Query 1      MALGLLVWIYIVLLI-ALSTVSAASPPGPSKSNGETNLAALLAFKAQLSDPLGILGGNW 59
          MA + V I +VLLI ALS V+ AS  PSKSNST+T+ AALLAFKAQL+DPLGIL  NW
Sbjct 1      MAFRMPVVRISVLLIIALSAVTCASAV-PSKSNSTDTDYAALLAFKAQLADPLGILASNW 59

Query 60     TVGTPFCRWVGVSCSHHRQRTALDLRDTPLLGLSPQLGNLSFLSILNLTNTGLTGSVP 119
          TV TPFCRWVG+ C   QRVT L L   PL GELS LGNLSFLS+LNLTN LTGSVP
Sbjct 60     TVNTPFCRWVGIRCGRRHQRTGLVLPGLPLQELSSHLGNLSFLSVLNLTNASLTGSVP 119

Query 120    NDIGRLHRLEILELGYNLTSGSIPATIGNLRLQVLDLQFNLSLGGPIPADLQNLQNLSSI 179
          DIGRLHRLEILELGYN+LSG IPATIGNLRL+VL L+FN LSG IPA+LQ L ++ +
Sbjct 120    EDIGRLHRLEILELGYNLSGGIPATIGNLRLRLVLYLEFNQLSGSIPAEIQGLGSIGLM 179

Query 180    NLRNRYLIGLIPNNLFNNTLLTYLNIGNNSLSGPIPGCIGSLPILQTLVLQVNNLTGPV 239
          +LRRNYL G IPNNLFNNT LL Y NIGNNSLSG IP IGSL +L+ L +QVN L GPV
Sbjct 180    SLRRNYLTGSIPNNLFNNTPLLAYFNIGNNSLSGIPASIGSLSMLEHLNMQVNNLAGPV 239
  
```

# ◆ Gene prediction and annotation

◆ FGENESH, GenScan, Glimmer, Genemark, Augustus

◆ Web-server (FGENESH) or standalone package (Augustus, Glimmer)

◆ Steps:

- ◆ Eliminate repetitive sequence
- ◆ Confirm ORF (open reading frame)
- ◆ Promoter

◆ Tips:

- ◆ Difference between prokaryotes and eukaryotes
- ◆ Difference between different softwares
- ◆ comprehensive consideration all results

◆ GO, KEGG, COG

# A snapshot for FGENESH



HOMEPRODUCTSNEW PRODUCTSSERVICESMANAGEMENT TEAMCORPORATE PROFILECONTACT

TEST ON LINE

- GENE FINDING in Eukaryota
- GENE FINDING WITH SIMILARITY
- OPERON AND GENE FINDING IN BACTERIA
- GENE FINDING IN VIRUSES
- ALIGNMENT /Sequences&genomes
- GENOME EXPLORER /Infogene
- SEARCH FOR MOTIFS /promoters&functional
- PROTEIN LOCATION /patterns/Epitops
- RNA STRUCTURE COMPUTING
- PROTEIN STRUCTURE
- PROTEIN / DNA 3D-Visual Works
- SEQMAN
- MULTIPLE ALIGNMENTS
- CLUSTERING ESTs
- ANALYSIS OF EXPRESSION DATA
- HUMAN-MOUSE-RAT SYNTENY
- PLANT PROMOTERS DATABASE
- REPEATS /find&map repeats
- SNP Extracting known SNPs

## FGENESH

HMM-based gene structure prediction (multiple genes, both chains)

Paste nucleotide sequence here:

Alternatively, load a local file with sequence in Fasta format:

Local file name:

Organism: ☐ Bos taurus ☐ Chicken ☐ Fish ☐ Frog (Xenopodinae) ☒ Human ☐ Mouse

☐ Anopheles gambiae ☐ Culex ☐ Drosophila ☐ Honey Bee ☐ Tribolium (red flour beetle)

☐ Brugia malayi (parasitic nematode) ☐ C.elegans ☐ Sea urchin

☐ Diatom ☐ Plasmodium falciparum ☐ Phytophthora

☐ Dicot plants (Arabidopsis) ☐ Medicago (legume plant) ☐ Monocot plants (Corn, Rice, Wheat, Barley) ☐ Nicotiana tabacum

☐ Tomato ☐ Vitis vinifera

☐ Chlamydomonas (single celled green algae)

☐ Alternaria ☐ Aspergillus ☐ Batrachochytrium ☐ Botrytis ☐ Coccidioides immitis ☐ Coprinopsis cinerea

☐ Cryptococcus neoformans ☐ Fusarium graminearum ☐ Histoplasma (fungus) ☐ Leptosphaeria ☐ Magnaporthe

☐ Neurospora crassa ☐ Paracocci ☐ Phanerochaete chrysosporium (white rot) ☐ Penicillium ☐ Puccinia ☐ Pyrenophora

☐ Rhizopus\_oryzae ☐ Schizosaccharomyces pombe ☐ Sclerotinia sclerotiorum ☐ Stagnospora nodorum ☐ Uncinocarpus reesii

☐ Ustilago

[\[Help\]](#) [\[Show advanced options\]](#)  
[\[Example: Homo sapiens genomic beta globin region \(HBB@\) on chromosome 11\]](#)

## Tools at the EBI

We provide a comprehensive range of bioinformatics tools.

This page shows a selection of those that are used most frequently. These include tools for the analysis and comparison of nucleotide and protein sequences, data from functional genomics experiments, text mining of the scientific literature and tools for determination and visualisation of macromolecular structures.

All these tools can be accessed over the web and most provide [Web Services](#) interfaces using SOAP or REST APIs.

### Nucleotide and Protein sequence searching

#### Nucleotide sequence searches

The sequence databases that can be searched with the tools outlined below include EMBL-Bank, Coding Sequences, immunoglobulins and High throughput cDNA:

- [ENA Search](#)
- [BLAST Nucleotide](#)
- [Fasta Genomes](#)
- [Ssearch Genomes](#)

#### Protein sequence searches

The protein sequence databases available to search below include UniProtKB, sequences derived from macromolecular structures, immunoglobulins and sequences from patents:

- [BLAST Protein](#)
- [PSI-Search](#)
- [Fasta Proteomes](#)
- [Ssearch Proteomes](#)

### Multiple Sequence Alignment

Alignment of three or more sequences to identify regions of conservation which may indicate functional constraints and infer evolutionary relationships.

- [Clustal Omega](#)
- [ClustalW2](#)
- [WebPRANK](#)
- [T-Coffee](#)
- [MUSCLE](#)

### Pairwise Sequence Alignments

Alignment of two sequences to identify regions where the sequence is conserved and conversely regions where the sequence is not conserved.

- [Needle](#)
- [LAlign](#)

### Protein Functional Analysis

**Identification** of potential protein families, **domains** and functional sites allowing the function of a protein sequence to be determined.

- [InterProScan](#)
- [Pratt](#)
- [Phobius](#)
- [RADAR](#)

### Functional Genomics Tools

Tools to examine and explore data from gene expression experiments.

- [Gene Expression Atlas](#)
- [EFO Tools](#)
- [EBI R-workbench](#)

### Structure Tools

Tools for visualisation and analysis of molecular & cellular structure.

- [PDBFold](#)
- [PDBsum](#)

### Scientific literature text mining

Analysis of text documents to identify important terms and relate these to ontologies.

- [EBIMed](#)
- [Whatizit](#)

### Sequence Translation

Tools to translate or back-translate between nucleotide and peptide sequences.

- [Transeq](#)
- [Backtranseq](#)

### Data retrieval and ID mapping

- [dbfetch](#)
- [ENA / SRA](#)
- [ENA / EMBL-SVA](#)
- [UniProt / UniSave](#)
- [SRS](#)
- [PICR](#)

## ◆ Multiple sequence alignment

- ◆ ClustalW, MAFFT, MUSCLE, T-Coffee
  - ◆ Web-server or standalone package (ClustalX)
  - ◆ Visualization of alignment results
    - ◆ Gendoc/Bioedit

## ◆ Phylogenetic analysis

- ◆ MEGA5.0, phylip, PUAP\*, MrBayes
- ◆ NJ, maximum parsimony, maximum likelihood, Bayesian, UPGMA

# A snapshot for ClustalW2

EBI > Tools > Multiple Sequence Alignments > ClustalW2

## ClustalW2 - Multiple Sequences Alignment

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.

### Use this tool

#### STEP 1 - Enter your input sequences

Enter or paste a set of Protein sequences in any supported format:

Or, upload a file:

浏览...

#### STEP 2 - Set your Pairwise Alignment Options

Alignment Type: ☒ Slow ☐ Fast

##### Slow Pairwise Alignment Options

| Protein Weight Matrix | GAP OPEN | GAP EXTENSION |
|-----------------------|----------|---------------|
| Gonnet                | 10       | 0.1           |

#### STEP 3 - Set your Multiple Sequence Alignment Options

| Protein Weight Matrix | GAP OPEN | GAP EXTENSION | GAP DISTANCES | NO END GAPS |
|-----------------------|----------|---------------|---------------|-------------|
| Gonnet                | 10       | 0.20          | 5             | no          |

| ITERATION | NUMITER | CLUSTERING |
|-----------|---------|------------|
| none      | 1       | NJ         |

##### OUTPUT Options

| FORMAT        | ORDER   |
|---------------|---------|
| Aln w/numbers | aligned |

#### STEP 4 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

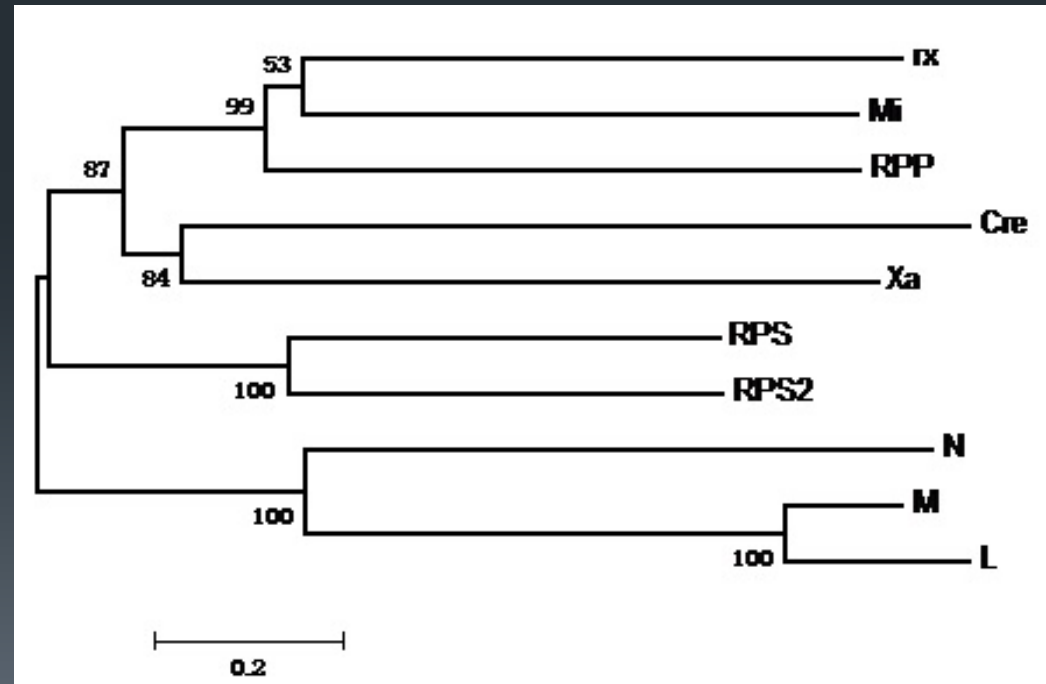
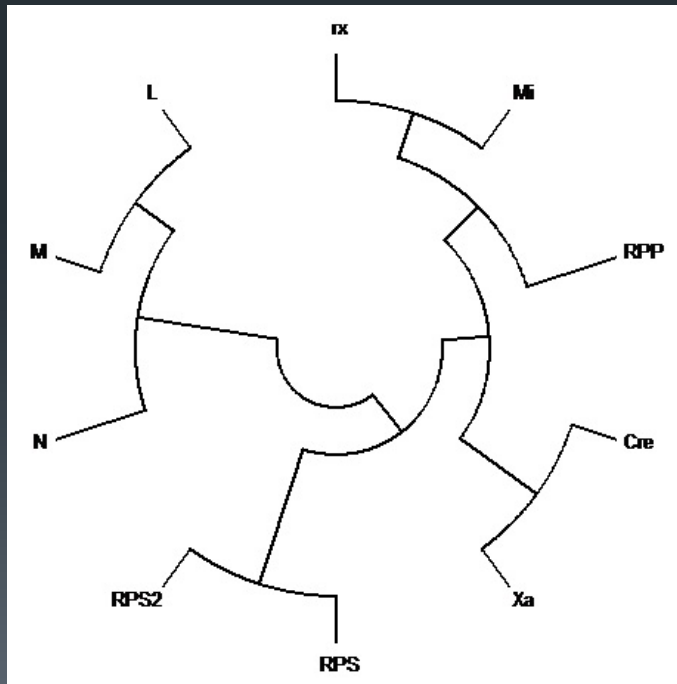
## Multi-alignment (MSF format in Gendoc)

```

520          *          540          *          560          *          580          *
----IVGQDSMLDKVWNCIMEDK---VWIVGLYGMGGVGKTTILTQINNKFSLKG-----GGFDVVIWVVVSKN-
----VVGNTTMMEQVLEFLSEEEE--RGIIGVYGGGGVGKTTIMQSINNELITKG-----HQYDVLIWVQMSRE-
ETDELVGIDDHVEVILEMISLDSKS-VTMVGLYGMGGIGKTTTAKAVYNKISSHF-----DRCCFVDNVRAMQEQ
ETDELVGIDDHITAVLEKISLDSN-VTMVGLYGMGGIGKTTTAKAVYNKISSCF-----DCCCFIDNIRETQE-
LQN-IVGIDTHLEKTESLLEIGING-VRIMGIVGMGGVGKTTIARAFDITLLGRM-----DSSYQFDGACFLKDI
SESDLVGVEQSVKEIVGHIVEND-V-HQVVSIAAGMGGIGKTTLARQVEHHDLVR-----RHFDGFAWVCVSQQ-
---IMVGRENEFEMMLDQIARGGRE-LEVVSIVGMGGIGKTTLATKLY-SDPCIM-----SRFDIRAKATVSQE-
TDKITVGFEETNLIIRKITSGSAD-IDVISITGMFGSGKTTLAYKVYNDKSVS-----SRFDLRAWCTVDQG-
----IIAMLHEKEGGDPSTSKGLC-FSVIGIHGVSGSGKSTLAQFVYAHEKNDKQDNKEDHFDLVMWVHVSQD-
PEPIVYGRAAEMETIKQLIMSNRSNGITVLPFIVGNGGIGKTTLAQLVCKDLVIKS-----QFNVKIWVYVSDK-
g                                66 6 G gG GK3T a 6

```

## Phylogenetic Tree (NJ methods in MEGA)





# Happy new year!

- Please submit your homework in **PDF** format to [bioinplant@zju.edu.cn](mailto:bioinplant@zju.edu.cn) at the **8<sup>th</sup>** week in winter term.