

Sequence databases and online genomics tools

Bioinformatics

- Multiple levels from users to developers
- Custom or general use tools for data analysis
- Similar tools available from different sources
- Tools and databases are interconnected
- There are often multiple ways to reach an objective
- A lot of tools and information are freely available in the Web.

Sequence databases

- DDBJ: DNA Data bank of Japan
- EMBL: European Molecular Biology Laboratory
- NCBI: National Center for Biotechnology Information
- UniProt: (Universal Protein Resource)
 - Swiss Prot (Manually curated)
 - TrEMBL (automated annotation)

The screenshot displays the Entrez search engine interface. At the top, the Entrez logo and the text "Entrez, The Life Sciences Search Engine" are visible. Below the logo, there are navigation tabs for "PubMed", "All Databases", "Human Genome", "GenBank", and "Map Viewer". A search bar with a "GO" button and "Clear" and "Help" links is present. The main content area is titled "Welcome to the Entrez cross-database search page" and lists numerous databases in two columns, each with a small icon and a brief description:

- PubMed: Biomedical literature citations and abstracts
- PubMed Central: free, full text journal articles
- Site Search: NCBI web and FTP sites
- Books: online books
- OMIM: online Mendelian Inheritance in Man
- OMIA: online Mendelian Inheritance in Animals
- Nucleotide: Core subset of nucleotide sequence records
- EST: Expressed Sequence Tag records
- GSS: Genome Survey Sequence records
- Protein: sequence database
- Genome: whole genome sequences
- Structure: three-dimensional macromolecular structures
- Taxonomy: organisms in GenBank
- SNP: single nucleotide polymorphism
- dbVar: Genomic structural variation
- Gene: gene-centered information
- SRA: Sequence Read Archive
- BioSystems: Pathways and systems of interacting molecules
- HomoloGene: eukaryotic homology groups
- GENSAT: gene expression atlas of mouse central nervous system
- Probe: sequence-specific reagents
- Genome Project: genome project information
- dbGAP: genotype and phenotype
- UniGene: gene-oriented clusters of transcript sequences
- CDD: conserved protein domain database
- 3D Domains: domains from Entrez Structure
- UniSTS: markers and mapping data
- PopSet: population study data sets
- GEO Profiles: expression and molecular abundance profiles
- GEO DataSets: experimental sets of GEO data
- Epigenomics: Epigenetic maps and data sets
- Cancer Chromosomes: cytogenetic databases
- PubChem BioAssay: bioactivity screens of chemical substances
- PubChem Compound: unique small molecule chemical structures
- PubChem Substance: deposited chemical substance records
- Protein Clusters: a collection of related protein sequences
- Peptidome: MS/MS proteomic experiments
- Journals: detailed information about the journals indexed in PubMed and other Entrez databases
- NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections
- MeSH: detailed information about NLM's controlled vocabulary

Entrez – Nucleotide database

The screenshot shows the NCBI Nucleotide database search results for the query 'aquaporin'. The search returned 8228 nucleotide sequences, including 5446 Nucleotide entries, 2779 ESTs, and 3 GSSs. The results are displayed in a list format, showing the first 20 items. The top results include:

- 1. Flavobacteriales bacterium HTCC2170, complete genome**: 3,868,304 bp circular genomic, CP002157.1 | GI:304420054
- 2. Shewanella baltica OS183.ctg590, whole genome shotgun sequence**: 700,789 bp linear genomic, NZ_AECY01000002.1 | GI:304409268
- 3. Gordonia bronchialis DSM 43247, complete genome**: 5,208,602 bp circular genomic, CP001802.1 | GI:262083393
- 4. Salmo salar clone ssal-rgb2-537-119 Aquaporin FA-CHIP putative mRNA, complete cds**: 1,566 bp linear mRNA, BT059602.1 | GI:223649113
- 5. Salmo salar clone ssal-rgb2-574-205 Aquaporin FA-CHIP putative mRNA, complete cds**: 1,477 bp linear mRNA, BT047826.1 | GI:209733515

The interface includes a search bar at the top with the query 'aquaporin', a 'Search' button, and a 'Clear' button. On the right side, there are filters for 'Filter your results' (All (5446), Bacteria (1379), INSDC (GenBank) (3349), mRNA (1661), RefSeq (2062)) and 'Top Organisms' (Homo sapiens (692), Rattus norvegicus (203), Mus musculus (171), Oryza sativa (122), Escherichia coli (120), All other taxa (3880)).

This screenshot is identical to the one above, but with a dropdown menu open over the search bar. The dropdown menu lists various database categories and search options:

- All Databases
- PubMed
- Protein
- Nucleotide
- GSS
- EST
- Structure
- Genome
- BioSystems
- Books
- CancerChromosomes
- Conserved Domains
- dbCAP
- dbVar
- 3D Domains
- Epigenomics
- Care
- Genome Project
- GENSAT
- GEO Profiles
- GEO DataSets
- HomoloGene
- Journals
- MESH
- NCBI Web Site
- NLM Catalog
- OMA
- OMIM
- Peptidome
- PMC
- PopSet
- Probe
- Protein Clusters
- PubChem BioAssay
- PubChem Compound
- PubChem Substance
- SNP
- SRA
- Taxonomy
- Trace
- TrasK/Al
- UniCene
- UNSTS

GenBank record:

Display Settings: GenBank (full)

Arabidopsis thaliana ecotype Columbia aquaporin (Aqua) mRNA, complete cds
GenBank: HM217349.1

Features Sequence

LOCUS HM217349 988 bp mRNA linear PLN 29 JUN 2010

DEFINITION Arabidopsis thaliana ecotype Columbia aquaporin (Aqua) mRNA, complete cds.

ACCESSION HM217349

VERSION HM217349.1 GI:299507803

KEYWORDS .

SOURCE Arabidopsis thaliana (thale cress)

ORGANISM Arabidopsis thaliana
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids; malvids; Brassicales; Brassicaceae; Arabidopsids.

REFERENCE 1 (bases 1 to 988)
AUTHORS Venkatesh,J., Gururani,M.A., Nookaraju,A., Pandey,S.K., Park,S.W., Kim,D.H., Chul,C.S. and Upadhyaya,C.P.
TITLE Investigating the role of aquaporins in potato tuberization
JOURNAL Unpublished

REFERENCE 2 (bases 1 to 988)
AUTHORS Venkatesh,J., Gururani,M.A., Nookaraju,A., Pandey,S.K., Park,S.W., Kim,D.H., Chul,C.S. and Upadhyaya,C.P.
TITLE Direct Submission
JOURNAL Submitted (08-MAY-2010) Molecular Biotechnology, Konkuk University, 1-Nwyang-dong, Owangjin-gu, Seoul 143701, South Korea

FEATURES
Location/Qualifiers
source 1..988
/organism="Arabidopsis thaliana"
/mol_type="mRNA"
/db_xref="taxon:1702"
/ecotype="Columbia"
gene 1..988
/gene="Aqua"
cds 38..898
/gene="Aqua"
/note="involved in water channel activity, salt stress, tolerance, water transport, and drought stress tolerance"
/codon_start=1
/product="aquaporin"
/protein_id="AB321813.1"
/db_xref="GI:299507804"
/translation="MDREEDVGVGNKFFEROPIGTSAQSDSDYKPPFAPLFFPGE
LASHFPRACIAEFATFLYITVFLVWVFRSFWKAGVYQCIANRPGMIFALV
YCTACISGCHINPAVTFGLFLARKLSLFRVYIVMQLCAICACVVKCFQPKQVGA
LOGGANIAGVYKSGLSAEIIGTFVLYVYVPSATDAKRNARDSHVPLALPIGFA
VELVHLATIPITGTCINPARSLGAAIFPKNDANDHWHVFWGPFICAAALALYHIV
IKAIFFKRS"

ORIGIN
1 accctagaaa gctctagaga gaagagaga gagagatg gaagtaag aagaatgt
61 tagagtggg gctaacagt ttccggagag gcaaccgat ggaacttgg etoagatga
121 caagpactac aaagagacc cacttggcc gttgttggg ccggcagag tagctaatg
181 gctctctgg agagatggg ttgtgagtt tatagatag tttttttcc tganatcac
241 tgttttgact gtatagggt tgaagagtc accgaactg tgtgcttcc tggatcca
301 aactatgat tggatgctg gttatgatt atttattct gttcttacc gttatgatt

GenBank record (fasta format):

Arabidopsis thaliana ecotype Columbia aquaporin (A... - Nucleotide result

http://www.ncbi.nlm.nih.gov/nuccore/HM217349.1?report=fasta

```
>gi|299507803|gb|HM217349.1| Arabidopsis thaliana ecotype Columbia aquaporin
(Aqua) mRNA, complete cds
AACCTAGAAAAGCTCTAGAGAGAAAGAGAGAGAGAGATGGAAGGTAAGAAGAAGATCTTAGAGTCGGA
GCTAACAAAGTTCCGGAGAGGCAACCGATCGGAACCTCGGCTCAGAGTGACAAGGACTACAAAGAGCCAC
CACCTCGCCCGTTGTCGAGCCCGCGGAGCTAGCTTCATGGTCTCTCGGAGAGCTGGGATTGCTGAGTT
TATAGCTACGTTTTTCTCTGTACATCACATGTTTGGCTGTTATGGGTGTGAAGAGGTCACCGAACATG
TGTGCTTCGTCGGAATCCAAGGTATCGCTTGGCTTTCGGTGGTATGATCTTCGCTCGTCTACTGCA
CCGCTGGTATCCTCGGTGGACACATCAACCAGCGGTTACGTCGGTTTGTCTTAGCTAGGAAGCTTTC
GCTCACAGAGCTGTACTACTATAGTATGCAGTGTCTAGGAGCTATCTGGAGCTGCTGCTGGTCAAG
GGTTCACCCAAAGCAATACCAGGCTTTGGAGGCTGGACCAACCCATAGCTCATGGCTACACCAAAG
GAAGTGGTCTTGGAGCTGAGATTATGGAACCTTTGCTCTTGTACACCGCTTCTCTGCCACTGATGC
CAAGAGAAAACGCTCGTGACTCTCATGTTCCATTCTAGCACCGCTCCCTATCGGATTCGCTGTCTCTG
GTTCACTTAGCAACCATCCCCATTACTGGAACCTGGAATCAACCCAGCAAGAAGCTCTGGAGTGCAATCA
TCTTCAACAAGGCAACGCTTGGGATGACCACCTGGTCTTTGGGTTGGACCATTCATTGGTCTGCACT
TGCTGCTCTACACAGTTATAGTCAATCAGAGCCATCCCATTTCAAGTCCAGAGCTAAAGCTGATTGACT
TCTATTTAAAATCGGCTTTTGTCTTAGTTTGTCTTTGTGAACTACTACCTGTGTAAACGTGT
GTATCTCG
```

GB Protein record:

aquaporin [Arabidopsis thaliana]

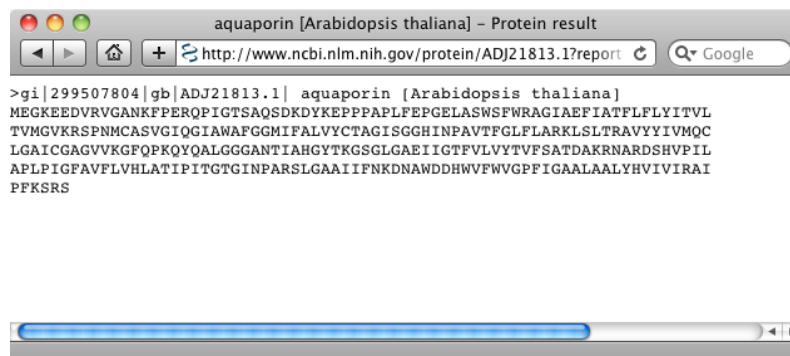
GenBank: ADJ21813.1

Features Sequence

```

LOCUS       ADJ21813                286 aa                linear   PLN 29-JUN-2010
DEFINITION aquaporin [Arabidopsis thaliana].
ACCESSION   ADJ21813
VERSION     ADJ21813.1 GI:299507804
DBSOURCE    accession HM21749.1
KEYWORDS    -
SOURCE      Arabidopsis thaliana (thale cress)
ORGANISM    Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; malvids; Brassicales; Brassicaceae; Arabidopsis.
REFERENCE   1 (residues 1 to 286)
AUTHORS     Venkatesh,J., Gururani,M.A., Mookaraju,A., Pandey,S.K., Park,S.M.,
            Kim,D.H., Chul,C.S. and Upadhyaya,C.F.
TITLE       Investigating the role of aquaporins in potato tuberization
JOURNAL     Unpublished
REFERENCE   2 (residues 1 to 286)
AUTHORS     Venkatesh,J., Gururani,M.A., Mookaraju,A., Pandey,S.K., Park,S.W.,
            Kim,D.H., Chul,C.S. and Upadhyaya,C.F.
TITLE       Direct Submission
JOURNAL     Submitted (08-MAY-2010) Molecular Biotechnology, Konkuk University,
            1-Weayangdong, Gwangjin-gu, Seoul 143701, South Korea
COMMENT     Method: conceptual translation.
FEATURES             Location/Qualifiers
     source           1..286
                     /organism="Arabidopsis thaliana"
                     /db_xref="taxon:2204"
                     /ecotype="Columbia"
     protein          1..286
                     /product="aquaporin"
                     /name="involved in water channel activity, salt stress,
                     tolerance, water transport, and drought stress tolerance"
     region           52..276
                     /region_name="MIP"
                     /note="Major intrinsic protein (MIP) superfamily. Members
                     of the MIP superfamily function as membrane channels that
                     selectively transport water, small neutral molecules, and
                     ions out of and between cells. The channel proteins share
                     a common fold: the N-...; cd00333"
     site             /db_xref="CDD:23422"
                     order(94..114,231..232,235,238)
                     /site_type="other"
                     /note="amphipathic channel"
                     /db_xref="CDD:23421"
     site             order(114..116,235..237)
                     /site_type="other"
                     /note="Asn-Pro-His signature motifs"
     CDD              1..286
                     /ggseq="Aqua"
                     /coded_by="HM21749.1:38..898"
ORIGIN       1 megkeedrvr gankfperqp lgtsaqskdk ykepppapif epgelaswf wraqlaeia
            41 tflilylvtl tvmpkprgn meavqlegli awqggmia lvytaspig qhnapavtfg
            121 lflarkslst ravvyivmge lgaicgavpv kgfqpkyga lggantiah pytkgqgla
            181 eiigtvlvy tvfatdakr nardehvpil apipqfavf lvhatipit gtingparal
            241 qaalfnkn awdhwvfv gfigaalaa lnyhviral ptkars
    //
  
```

Protein sequence in fasta format:



```

>gi|299507804|gb|ADJ21813.1| aquaporin [Arabidopsis thaliana]
MEGKEEDVRVGANKFPERQP IGTSAQSDKDYKEPPAPLFEPEGLASWSFWRAGIAEFIATPFLFLYITVL
TVMGVKRSPNMCASVGIQGIAWAFGGMIFALVYCTAGISGGHINPAVTFGLFLARKLSLTRAVVYIVMQC
LGAICGAVVKGFPQKQYQALGGGANTIAHGYTKGSGLGAEI IGTFLVLYTVFSATDAKRNARDSHVPII
APLPIGFVFLVHLATIPITGTGINPARSLGAAIFNKNDAWDHWFVWVGFPIGAALALYHVIVIRAI
PFKSRS
  
```

Searching GenBank

NCBI Resources How To My NCBI Sign In

Nucleotide
Alphabet of Life

Search: Nucleotide Limits Advanced search Help

Search Clear

Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

Using Nucleotide

- [FAQ](#)
- [Help](#)
- [GenBank FTP](#)
- [RefSeq FTP](#)

Nucleotide Tools

- [Submit to GenBank](#)
- [LinkOut](#)
- [EUtilities](#)
- [Sequence Revision History](#)
- [BLAST](#)

Other Resources

- [GenBank Home](#)
- [RefSeq Home](#)
- [Gene Home](#)
- [SRA Home](#)
- [INSDC](#)

You are here: NCBI > DNA & RNA > Nucleotide Database [Write to the Help Desk](#)

RefSeq

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

UniProt

Q41951 (TIP21_ARATH) Reviewed.

UniProtKB/Swiss-Prot

Last modified August 10, 2010. Version 96. [History...](#)

Clusters with 100%, 90%, 50% identity | Documents (2) | Third-party data

Customize display: Names - Attributes - General annotation - Ontologies - Sequence annotation - Sequences - References - Cross-refs - Entry info - Documents

Names and origin

Protein names	Recommended name: Aquaporin TIP2-1 Alternative name(s): Tonoplast intrinsic protein 2-1 Short name: TIP2-1 Delta-tonoplast intrinsic protein Short name-Delta-TIP
Gene names	Name: TIP2-1 Ordered Locus Names: AT3G16240 ORF Names: MYA6.10
Organism	Arabidopsis thaliana (Mouse-ear cress) [Complete proteome]
Taxonomic identifier	3702 [NCBI]
Taxonomic lineage	Eukaryota - Viridiplantae - Streptophyta - Embryophyta - Tracheophyta - Spermatophyta - Magnoliophyta - eudicotyledons - core eudicotyledons - rosids - malvids - Brassicales - Brassicaceae - Arabidopsis

Protein attributes

Sequence length	250 AA
Sequence status	Complete
Protein existence	Evidence at protein level.

General annotation (Comments)

Function: Aquaporin required to facilitate the transport of water from the vacuolar compartment to the cytoplasm. Does not promote glycerol permeability. Its function is impaired by Hg²⁺. Transports urea in yeast cells and *Xenopus laevis* oocytes in a pH-independent manner. Transports methylammonium or ammonium in yeast cells and *Xenopus laevis* oocytes, preferentially at high medium pH. May participate in vacuolar compartmentation and detoxification of ammonium. [UniProtKB/Swiss-Prot](#)

Subunit structure: Interacts with cucumber mosaic virus (CMV) Protein 1a. [UniProt](#)

Q0EEM0 (Q0EEM0_CRYJA) Unreviewed.

UniProtKB/TrEMBL

Last modified August 10, 2010. Version 23. [History...](#)

Clusters with 100%, 90%, 50% identity | Third-party data

Customize display: Names - Attributes - General annotation - Ontologies - Sequence annotation - Sequences - References - Cross-refs - Entry info

Names and origin

Protein names	Submitted name: Putative aquaporin UniProt
Gene names	Name: AQU UniProt
Organism	Cryptomeria japonica (Japanese cedar) UniProt
Taxonomic Identifier	3369 [NCBI]
Taxonomic lineage	Eukaryota - Viridiplantae - Streptophyta - Embryophyta - Tracheophyta - Spermatophyta - Coniferopsida - Coniferales - Cupressaceae - Cryptomeria

Protein attributes

Sequence length	275 AA
Sequence status	Fragment
Protein existence	Inferred from homology.

General annotation (Comments)

Sequence similarities: Belongs to the MIP/aquaporin family. [UniProt](#)

Ontologies

Keywords

Biological process	Transport UniProt
Cellular component	Membrane UniProt
Domain	Transmembrane UniProt

Gene Ontology (GO)

Biological process	transmembrane transport Inferred from electronic annotation. Source: InterPro
Cellular component	integral to membrane Inferred from electronic annotation. Source: UniProtKB-KW
Molecular function	transporter activity Inferred from electronic annotation. Source: InterPro

[Complete GO annotation...](#)

Next-gen sequence databases

HOME | SEARCH | SITE MAP

NCBI » GEO

Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

Not logged in | Login

Gene Expression Omnibus: a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. [More information »](#)

GEO navigation

QUERY

- DataSets [GO](#)
- Gene profiles [GO](#)
- GEO accession [GO](#)
- GEO BLAST

BROWSE

- DataSets
- GEO accessions
 - Platforms
 - Samples
 - Series

Submitter login

User id:

Password:

[LOGIN](#)

[» New account](#)
[» Recover password](#)

Site contents

Public data

Platforms	7,766
Samples	470,041
Series	18,486

Documentation

- [Overview](#) | [FAQ](#) | [Find](#)
- [Submission guide](#)
- [Linking & citing](#)
- [Journal citations](#)
- [Programmatic access](#)
- [DataSet clusters](#)
- [GEO announce list](#)
- [Data disclaimer](#)
- [GEO staff](#)

Query & Browse

- [Repository browser](#)
- [Submitters](#)
- [SAGEmap](#)
- [FTP site](#)
- [GEO Profiles](#)
- [GEO DataSets](#)
- [Submit](#)
- [New account](#)

| NLM | NIH | Email GEO | Disclaimer | Section 508

Announcements - Main : Sequence Read Archive : NCBI/NCIM/NH

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi? Search

NCBI Site map: All databases: PubMed: Search

Sequence Read Archive

Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST

Announcements Provisional SRA Tracking History About

The Sequence Read Archive (SRA) stores raw sequencing data from the "next" generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, Complete Genomics®, and others.

Current capabilities include:

- [Run Browser](#)
- [Study/Sample/Experiment/Analysis browsers](#)
- [Download facility](#)
- [Search SRA \(using Entrez\)](#)
- [Interactive submissions facility](#)
- [Automated submissions](#)

See [Sequence Read Archive Overview](#) for more information.

Write to the Help Desk | Privacy Notice | Disclaimer | Accessibility

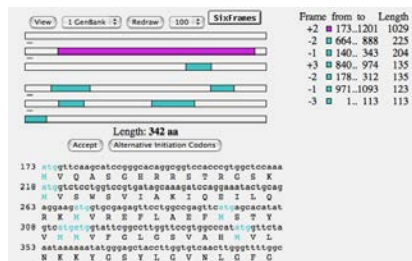
National Center for Biotechnology Information | U.S. National Library of Medicine

Last update: Mon, 02 Aug 2010 Rev. 199407

NCBI logo and FIRST GOV .GOV logo

ORF Finder (Open Reading Frame Finder)

- Graphical analysis tool which finds all ORFs in a sequence
- Looks for start and stop codons
- Different displays:
 - Graphical view of ORFs
 - Translated DNA sequence
 - Nucleotide sequence
 - Amino acid sequence
- Integral TBLASTN and BLASTP



Multiple sequence alignment

CLUSTAL W (1.83) multiple sequence alignment

```

swall|Q8CGP5|H2A1F_MOUSE      SGRGKQGGKARAKAKTRSSRAGLQFPVGRVHLLRKGNYSERVGAGAPVY 50
swall|P0C0S9|H2A1_BOVIN       SGRGKQGGKARAKAKTRSSRAGLQFPVGRVHLLRKGNYAERVGAGAPVY 50
swall|P0C169|H2A1C_RAT        SGRGKQGGKARAKAKSRSSRAGLQFPVGRVHLLRKGNYAERVGAGAPVY 50
swall|Q96QV6|H2A1A_HUMAN      SGRGKQGGKARAKSKSRSSRAGLQFPVGRVHLLRKGNYAERIGAGAPVY 50
*****_*:*****_*:*****_*:*****_*:*****_*

swall|Q8CGP5|H2A1F_MOUSE      LAAVLEYLTAEIILELAGNAARDNKKTRIIIPRHLQLAIRNDEELNKLGRV 100
swall|P0C0S9|H2A1_BOVIN       LAAVLEYLTAEIILELAGNAARDNKKTRIIIPRHLQLAIRNDEELNKLGRV 100
swall|P0C169|H2A1C_RAT        LAAVLEYLTAEIILELAGNAARDNKKTRIIIPRHLQLAIRNDEELNKLGRV 100
swall|Q96QV6|H2A1A_HUMAN      LAAVLEYLTAEIILELAGNASRDNKKTRIIIPRHLQLAIRNDEELNKLGGV 100
*****_*:*****_*:*****_*:*****_*:*****_*

```

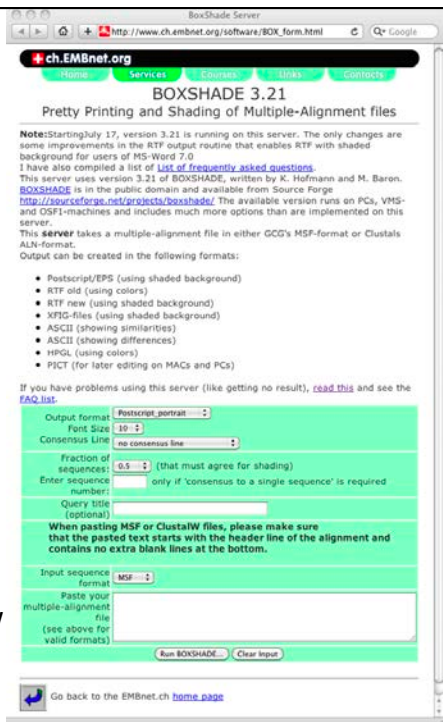
- Maximize the number of similar letters (aa or nt) per column
- Find functionally important sites
- Alignments can be viewed and edited in Jalview, GeneDoc, and other programs

Progressive Alignment

- Start by finding the pair-wise alignment of the two most related sequences
- Add new (less related) sequences, one at a time to the first pair-wise alignment
- Very sensitive to initial alignments (particularly for distantly related sequences)



BOXSHADE:



Input ALN file from CLUSTAL W

BOXSHADE output: postscript file

