



Student Online Workshop on
Application of Bioinformatics for Pedagogy of Plant Sciences



Organized by Kirori Mal College
(Under DBT Star Scheme)

23rd July 2021

Plant Cell Biology 101 meets
Bioinformatics

By

**Prof. Sudeshna Mazumdar-Leighton & Ph.D. Research Scholars of Plant-Biotic Interactions Lab,
Lab 15/18, Department of Botany, University of Delhi**

www.sml-botanydu.com

NOTE: This workshop presentation is not designed to infringe on any proprietary rights and is solely for teaching purposes without commercial benefits.

Online workshop contents

Sr. no.	Topics covered	Presenter	Duration
1.	Welcome address	Organizers	3.00 - 3.15 pm
2.	Bioinformatics and Actins in <i>Arabidopsis</i>	Prof. Sudeshna Mazumdar-Leighton	3.15 - 3.45 pm (35 min.)
3.	A brief overview of protein sorting; What is iPSORT and how does it work?	Aashima Mehra	3.45 - 4.00 pm (15 min)
4.	Retrieval of sequences from NCBI; Preparation of Fasta files; Prediction of localization sites of protein sequences	K. Tingneivah Mate	4.00 - 4.15 pm (15 min)
5.	Run amino acid index and Data analysis	Mansi Bansal	4.15 - 4.30 pm (15 min)
6.	Bayesian-based prediction software	Parul Bhardwaj	4.30 - 4.45 pm (15 min)
7.	Homework exercise and concluding session		4.45 – 5.00 pm

Overview

Cell (Animal, Plant, Bacteria, Virus + & -)

Genomes (Nuclear, Mitochondrial, Chloroplast)

recoDNA technologies (Cloning, PCR, DNA sequencing*)

Genome Sequencing Projects (HUGO, TAIR*)

Sequence Databases (NCBI, PDB)

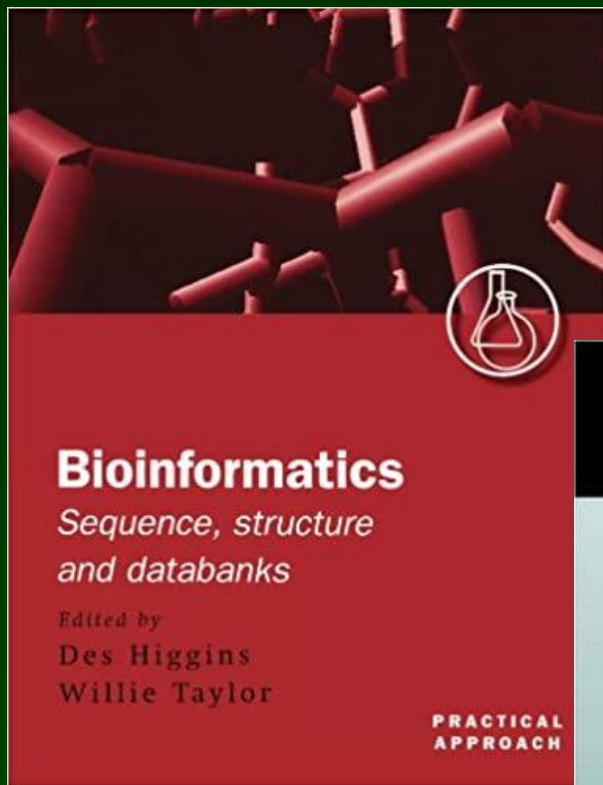
**Computational Biology/
Bioinformatics**

Sequence Databases

- ***History:*** Margaret Dayhoff (1970s), Walter Goad (Los Alamos National Lab, NM,USA) (1980s)
- **DDBJ, EMBL, NCBI (NA)**
- **PIR, Swiss-Prot, MIPS, JIPID (Proteins)**

WE RECOMMEND USING FOR CLASS:

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

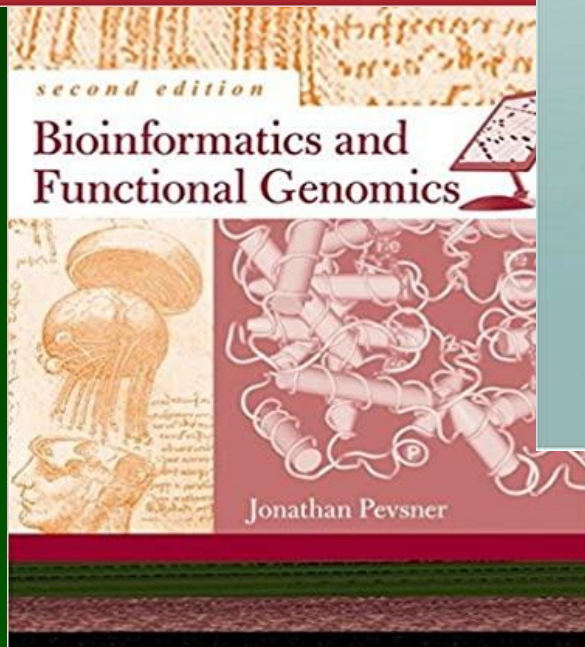


Bioinformatics

*Sequence, structure
and databanks*

Edited by
Des Higgins
Willie Taylor

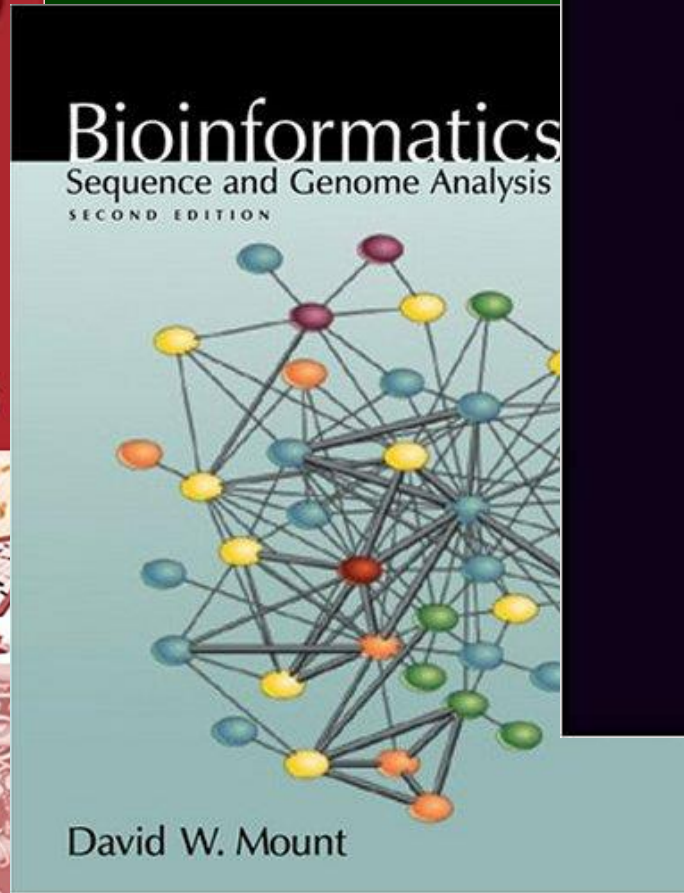
PRACTICAL
APPROACH



second edition

Bioinformatics and Functional Genomics

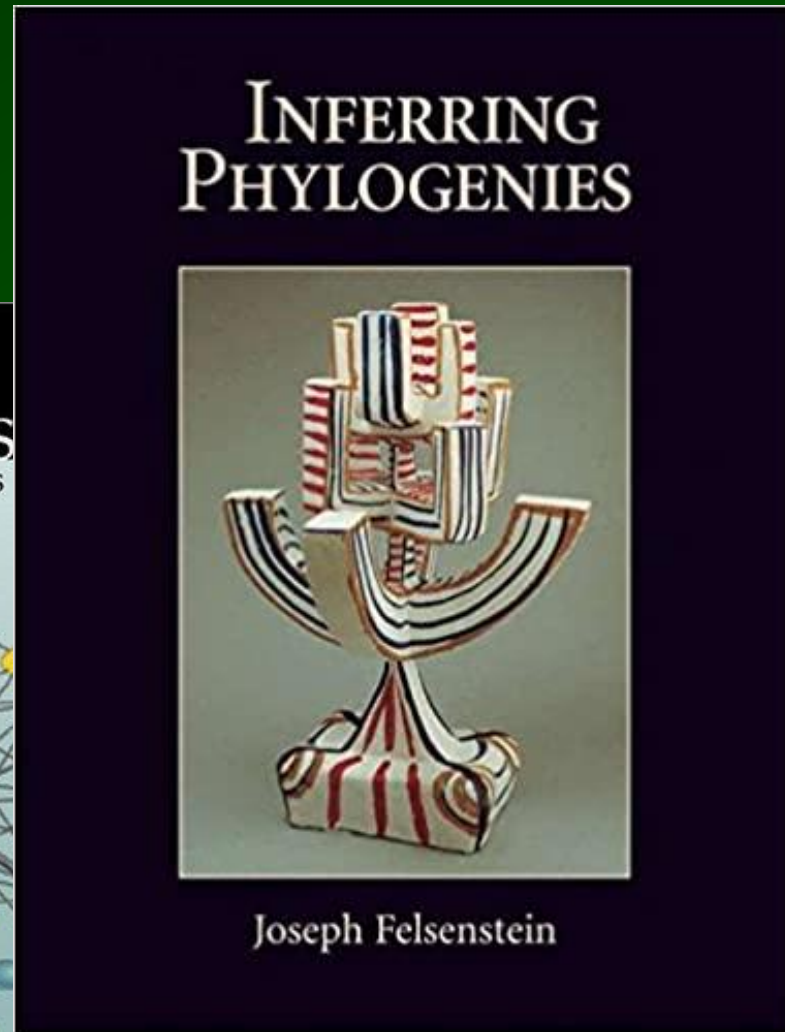
Jonathan Pevsner



Bioinformatics

Sequence and Genome Analysis
SECOND EDITION

David W. Mount

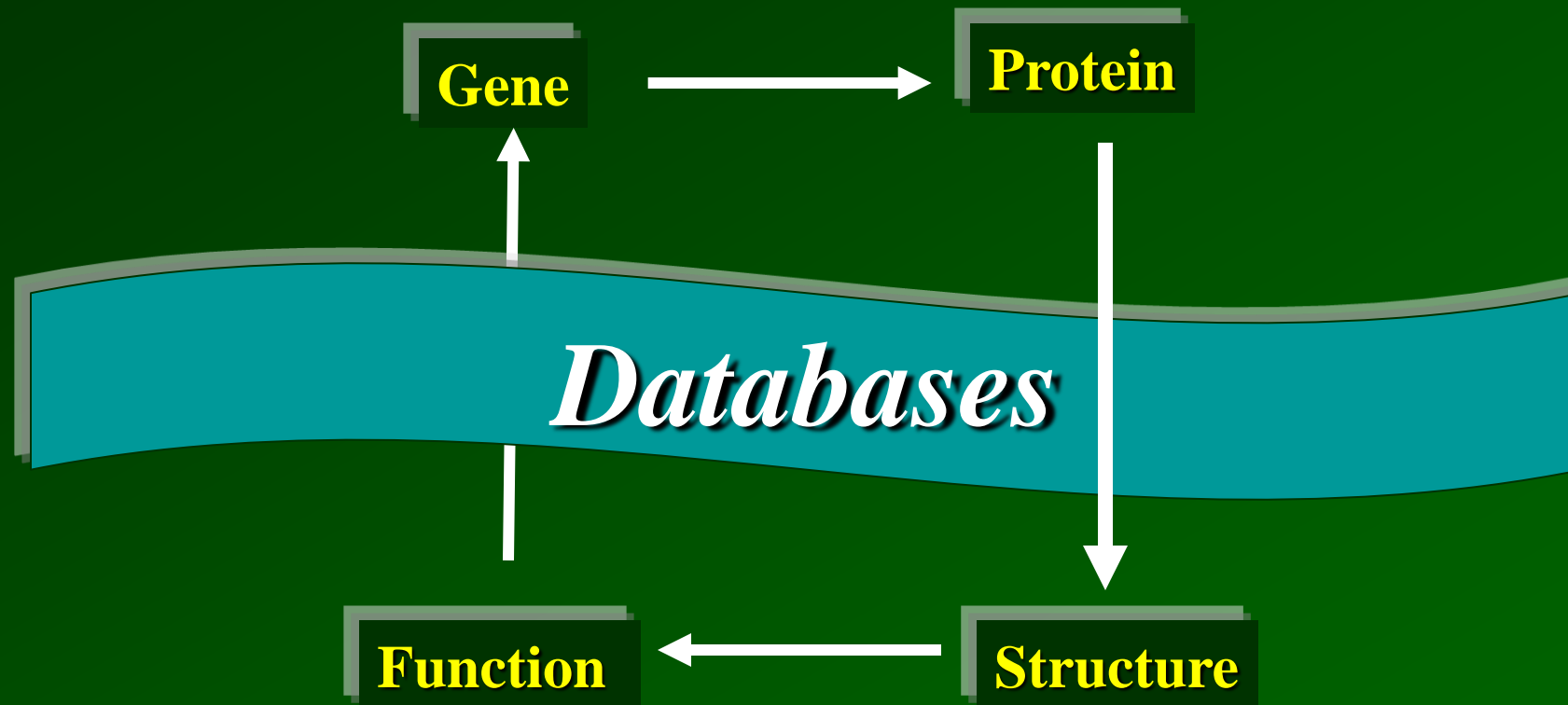


INFERRING PHYLOGENIES

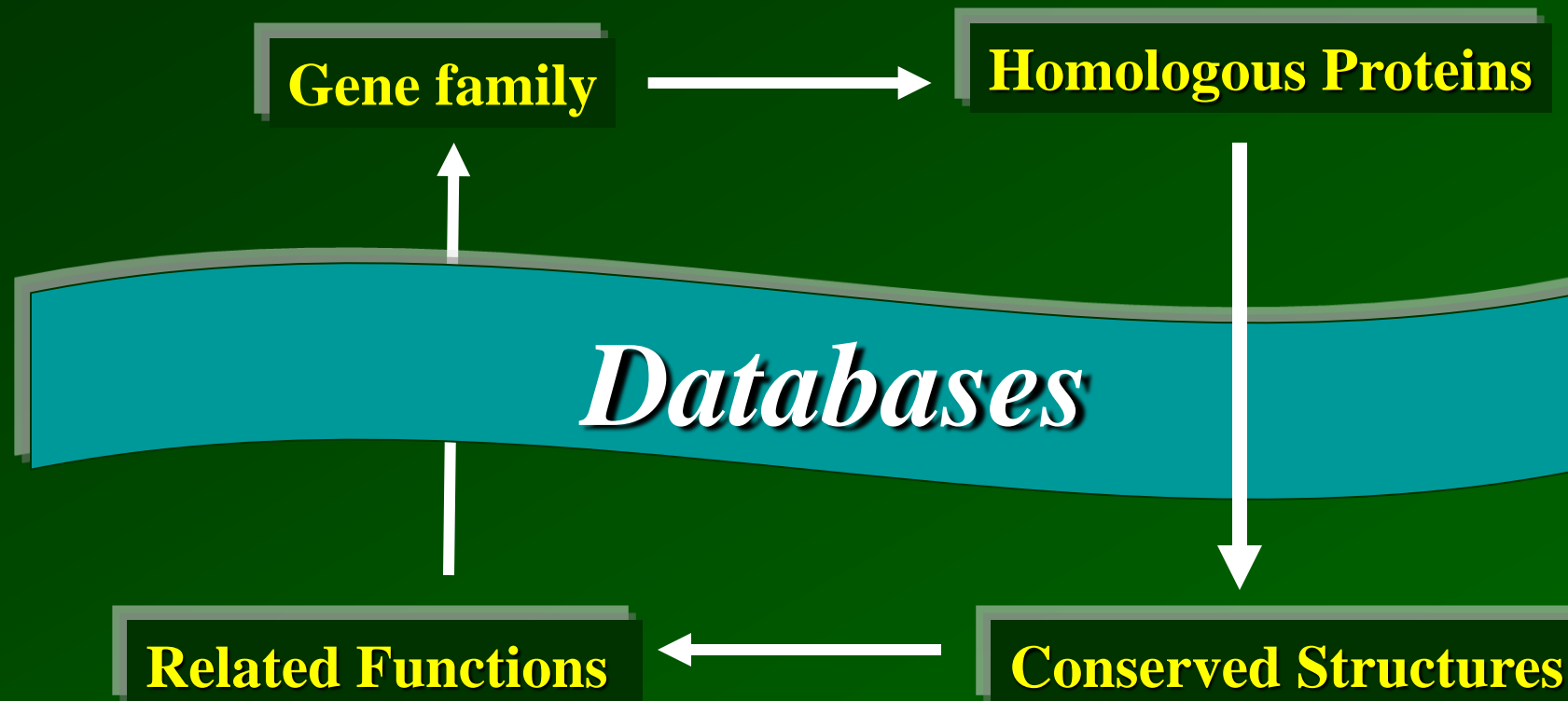


Joseph Felsenstein

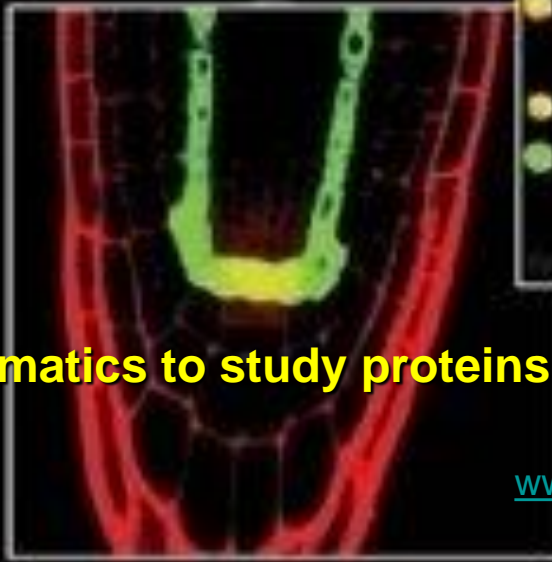
Bioinformatics/Computational Biology



Bioinformatics/Computational Biology



Molecular Evolution of Actin genes in *Arabidopsis thaliana*



***Bioinformatics to study proteins participating in functional organization of cells**



UNITE

A new NIH initiative to end structural racism and achieve racial equity in the biomedical research enterprise.

LEARN MORE

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
- Training & Tutorials
- Variation

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 libraries to build applications



Analyze

Identify an NCBI tool for your data analysis task



Research

Explore NCBI research and collaborative projects



Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Participating labs contribute over 70 tests for COVID-19 to the NIH Genetic Testing Registry

22 Jul 2021

During the COVID-19 pandemic, an

New RefSeq annotations for human, zebra finch, great white shark and more!

21 Jul 2021

In May and June, the NCBI Eukaryotic Genome Annotation Pipeline released

RefSeq release 207 is available!

16 Jul 2021

RefSeq release 207 is now available online, from the FTP site and through NCBI's Entrez programming utilities. F-

- Resources**
- NCBI Home
 - All Resources (A-Z)
 - Literature
 - DNA & RNA
 - Proteins
 - Sequence Analysis
 - Genes & Expression
 - Genomes
 - Maps & Markers
 - Domains & Structures
 - Genetics & Medicine
 - Taxonomy
 - Data & Software
 - Training & Tutorials
 - Homology
 - Small Molecules
 - Variation

Genome Reference Consortium

Formed to improve human and mouse reference assemblies, GRC will fix loci misrepresented in reference assembly, fill remaining gaps, and make alternate representations of complex loci.

1 2 3

- How To...**
- Obtain the full text of an article
 - Retrieve all sequences for an organism or taxon
 - Find a homolog for a gene in another organism
 - Find genes associated with a phenotype or disease
 - Design PCR primers and check them for specificity
 - Find the function of a gene or gene product
 - Find syntenic regions between the genomes of two organisms
- [See all ...](#)

About the NCBI

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

[More about the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS](#)

Search **All Databases** for **Search**

- All Databases
- NCBI Web Site
- Databases**
- PubMed
- Protein
- Nucleotide
- EST
- GSS
- Structure
- Genome
- Books
- CancerChromosomes
- Conserved Domains
- 3D Domains
- Gene
- Genome Project
- dbGaP
- GENSAT
- GEO Profiles
- GEO Datasets

- Newest H1N1 influenza sequences
- Submit flu sequences to GenBank
- Latest H1N1 citations in PubMed
- MedlinePlus (consumer health information)
- Enviro-Health links

Popular Resources

- PubMed

Search Protein for Arabidopsis, actin [Save Search](#)

Display Summary Show 20 Sort by Relevance Send to

All: 3811 Bacteria: 19 RefSeq: 296 Related Structures: 3137

Items 1 - 20 of 3811 Page 1 of 191 Next

This search in Gene shows [165 results](#), including:

- [ACT12](#) (Arabidopsis thaliana): ACT12 (ACTIN-12); structural constituent of cytoskeleton
- [ATFIM1](#) (Arabidopsis thaliana): ATFIM1 (Arabidopsis thaliana fimbrin 1); actin binding
- [VLN4](#) (Arabidopsis thaliana): VLN4 (ARABIDOPSIS THALIANA VILLIN 4); actin binding

Top Organisms [\[Tree\]](#)

- Oryza sativa Japonica Group (1620)
- Oryza sativa Indica Group (1401)
- Arabidopsis thaliana (652)
- Arabidopsis lyrata subsp. petraea (22)
- Physcomitrella patens (21)
- All other taxa (115)

- 1: [AAA98562](#) Reports
actin
gi|1145695|gb|AAA98562.1|1
- 2: [AAA98561](#) Reports
actin
gi|1145693|gb|AAA98561.1|1
- 3: [AA150956](#) Reports

Search Protein for Arabidopsis, actin

Display GenPept Show 20 Send to

Range: from begin to end Features: CDD

1: [AAA98562](#). Reports actin..[gi:1145695]

[BLink, Co](#)
[Domain](#)

[Features](#) [Sequence](#)

LOCUS AAA98562 377 aa linear PLN 29-APR-1996
 DEFINITION actin.
 ACCESSION AAA98562
 VERSION AAA98562.1 GI:1145695
 OBSOURCE locus ATU39480 accession [U39480.1](#)
 KEYWORDS .
 SOURCE Arabidopsis thaliana (thale cress)
 ORGANISM [Arabidopsis thaliana](#)
 Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
 Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
 rosids; eurosids II; Brassicales; Brassicaceae; Arabidopsis.
 REFERENCE 1 (residues 1 to 377)
 AUTHORS An,Y.Q., Huang,S., McDowell,J.M., McKinney,E.C. and Meagher,R.B.
 TITLE Conserved expression of the Arabidopsis ACT1 and ACT 3 actin
 subclass in organ primordia and mature pollen
 JOURNAL Plant Cell 8 (1), 15-30 (1996)
 PUBMED [8597657](#)

LOCUS AAA98562 377 aa linear PLN 29-APR-1996
 DEFINITION actin.
 ACCESSION AAA98562
 VERSION AAA98562.1 GI:1145695
 DBSOURCE locus ATU39480 accession U39480.1
 KEYWORDS .
 SOURCE Arabidopsis thaliana (thale cress)
 ORGANISM Arabidopsis thaliana
 Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
 Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
 rosids; eurosids II; Brassicales; Brassicaceae; Arabidopsis.
 REFERENCE 1 (residues 1 to 377)
 AUTHORS An, Y.Q., Huang, S., McDowell, J.M., McKinney, E.C. and Meagher, R.B.
 TITLE Conserved expression of the Arabidopsis ACT1 and ACT 3 actin
 subclass in organ primordia and mature pollen
 JOURNAL Plant Cell 8 (1), 15-30 (1996)
 PUBMED 8597657
 COMMENT Method: conceptual translation.
 FEATURES Location/Qualifiers
 source 1..377
 /organism="Arabidopsis thaliana"
 /db_xref="taxon:3702"
 Protein 1..377
 /product="actin"
 CDS 1..377
 /gene="ACT3"
 /coded_by="join(U39480.1:1611..1670,U39480.1:1809..2202,
 U39480.1:2281..2894,U39480.1:2988..3053)"
 ORIGIN
 1 madgediqpl vcdngtgmvk agfagddapr avfpsiivgrp rhtgvmvngmg qkdayvgdea
 61 qskrgiltlk ypiehgivnn wddmekiwhh tfynelrvap eehpilltea plnpanrek
 121 mtqimfetfn apamyvaiqa visiyasgrt tgrivdsdgd vshtvpiyeg yalphailrl
 181 dlagrdltde lmkiltergv sftttaerei yrdikeklcy ialdvegele taktssvek

A FASTA FILE

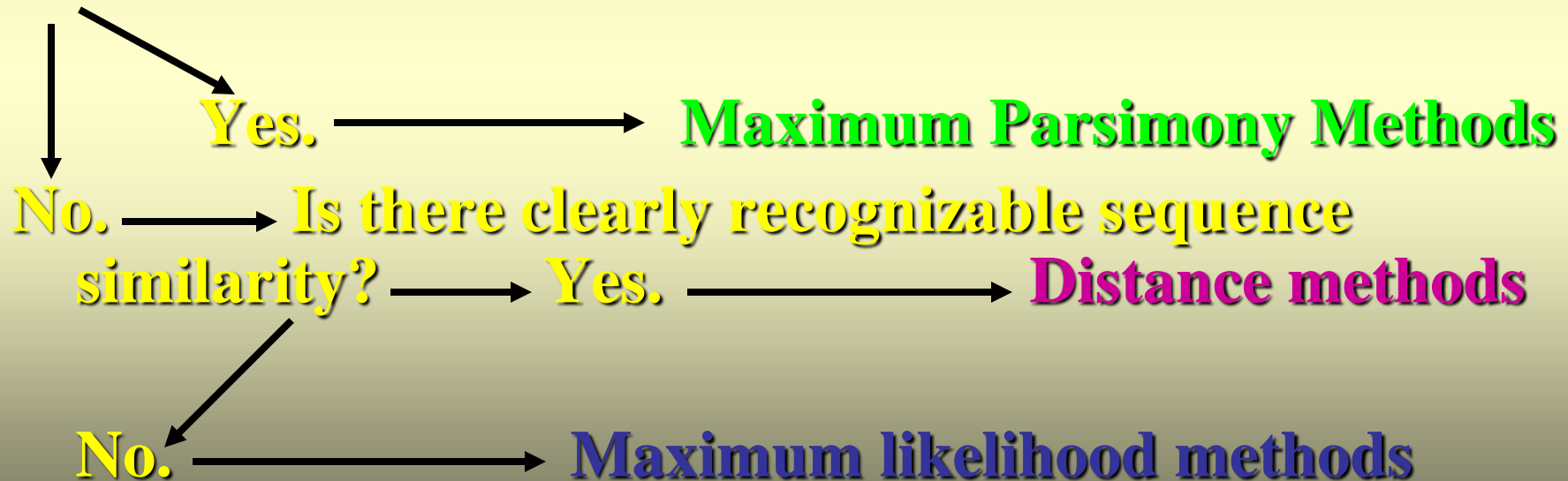
- >ACT1
MADGEDIQPLVCDNGTGMVKAGFAGDDAPRAVFPSIVGRPRHTGVMVGMGQKDAYVGDEA
QSKRGILTLKYPIEHGIVNNWDDMEKIWHHTFYNELRVAPEEHPILLTEAPLNPKANREKMTQI
MFETFNAPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILRLDLAGRDLT
DALMKILTERGYSFTTTAEREIVRDIKEKLCYIALDYEQELETAKTSSSVEKKNYELPDGQVITIGS
ERFRCPEVLYQPSMIGMENAGIHETTYNSIMKCDVDIRKDLYGNIVLSGGTTMFPGIADRMSKE
ITALAPSSMKIKVVAPPERKYSVWIGGSILASLSTFQQMWIAKAEYDESGPSIVHRKCF
- >ACT2
MAEADDIQPIVCDNGTGMVKAGFAGDDAPRAVFPSVVGRPRHHGVMVGMNPKDAYVGDEA
QSKRGILTLKYPIEHGVVSNWDDMEKIWHHTFYNELRIAPEEHPVLLTEAPLNPKANREKMTQI
MFETFNAPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGFSLPHAILRLDLAGRDLT
DYLKILTERGYMFTTTAEREIVRDIKEKLSFVAVDYEQEMETSKTSSSIEKKNYELPDGQVITIG
AERFRCPEVLFQPSFVGMEAAGIHETTYNSIMKCDVDIRKDLYGNIVLSGGTTMFSGIADRMSK
EITALAPSSMKIKVVAPPERKYSVWIGGSILASLSTFQQVKIDQILFRILLHAN
- >ACT3
MADGEDIQPLVCDNGTGMVKAGFAGDDAPRAVFPSIVGRPRHTGVMVGMGQKDAYVGDEA
QSKRGILTLKYPIEHGIVNNWDDMEKIWHHTFYNELRVAPEEHPILLTEAPLNPKANREKMTQI
MFETFNAPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILRLDLAGRDLT
DALMKILTERGYSFTTTAEREIVRDIKEKLCYIALDYEQELETAKTSSSVEKKNYELPDGQVITIGS
ERFRCPEVLYQPSMIGMENAGIHETTYNSIMKCDVDIRKDLYGNIVLSGGTTMFPGIADRMSKE
ITALAPSSMKIKVVAPPERKYSVWIGGSILASLSTFQQMWIAKAEYDESGPSIVHRKCF
- >ACT4
www.sml-botanydu.com

Methods

Choose set of **related** sequences

Obtain multiple sequence alignment

Is there strong sequence similarity?



CLUSTAL W (<http://www.ebi.ac.uk>)

The screenshot shows the ClustalW web interface in a Mozilla Firefox browser window. The browser's address bar shows the URL <http://www.ebi.ac.uk/Tools/clustalw/>. The page content includes a navigation menu with options like Databases, Tools, EBI Groups, Training, Industry, About Us, and Help. A left sidebar lists various help topics such as Help Index, General Help, Formats, Gaps, Matrix, References, ClustalW Help, ClustalW FAQ, Jalview Help, Scores Table, Alignment, Guide Tree, and Colours. The main content area features a heading for ClustalW, a descriptive paragraph, a link to the FAQ for new users, and a 'Download Software' section with icons for Windows, Mac, and Linux. Below this is a configuration form with several sections: 'YOUR EMAIL' (text input), 'ALIGNMENT TITLE' (text input with 'Sequence'), 'RESULTS' (dropdown menu with 'interactive'), 'ALIGNMENT' (dropdown menu with 'full'), 'KTUP (WORD SIZE)' (dropdown menu with 'def'), 'WINDOW LENGTH' (dropdown menu with 'def'), 'SCORE TYPE' (dropdown menu with 'percent'), 'TOPDIAG' (dropdown menu with 'def'), 'PAIRGAP' (dropdown menu with 'def'), 'MATRIX' (text input), 'GAP OPEN' (dropdown menu with 'def'), 'END GAPS' (text input), 'GAP EXTENSION' (text input), and 'GAP DISTANCES' (text input). The browser's taskbar at the bottom shows the Start button, several open applications, and the system clock displaying 7:01 PM.

CLUSTAL W (<http://www.ebi.ac.uk>)

The screenshot shows the ClustalW web interface in Mozilla Firefox. The browser window title is "EBI: Tools: ClustalW - Mozilla Firefox". The address bar shows the URL "http://www.ebi.ac.uk/Tools/clustalw/". The page features a sidebar on the left with links for "ClustalW Programmatic Access" and "www.clustal.org". The main content area is titled "OUTPUT" and "PHYLOGENETIC TREE". It contains several dropdown menus for "OUTPUT FORMAT" (set to "aln w/numbers"), "OUTPUT ORDER" (set to "aligned"), "TREE TYPE" (set to "none"), "CORRECT DIST." (set to "off"), and "IGNORE GAPS" (set to "off"). Below these settings is a large text input field with the placeholder text "Enter or paste a set of sequences in any supported format:" and a "Help" button. At the bottom of the input field, there is an "Upload a file:" label, a text box, and a "Browse..." button. To the right of the input field are two red buttons: "Run" and "Reset". The browser's status bar at the bottom shows "Done" and the Windows taskbar with various icons and the time "7:03 PM".

Sequence Alignments

- Pair-wise comparison....

LGPKSKQTGKGS--SRIWDN
LN--ITKSAGKKGAIMRLGDA

GLOBAL ALIGNMENT

LOCAL ALIGNMENT

Sequence Alignment

- 2 sequences (protein and nucleotide)
Needleman & Wunsch, 1970 alignments

GATCTA
GATCΔA

	G	A	T	C	T	A
G	1					
A		2				1
T			3		1	
C				4		
Δ → A		1				

5 minus gap penalty

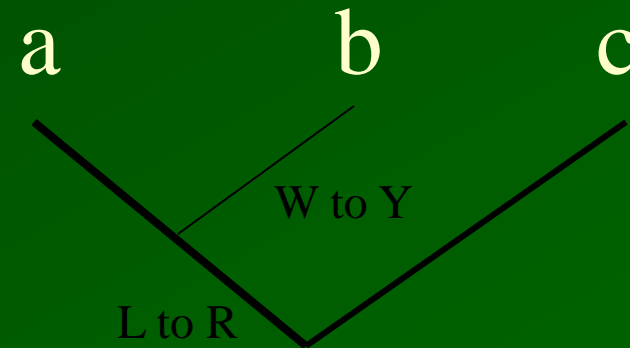
Aligning gene & protein sequences

- Margaret Dayhoff, 1972, 1978 (PIR) Matrices based PAM tables (percent amino acid substitutions)

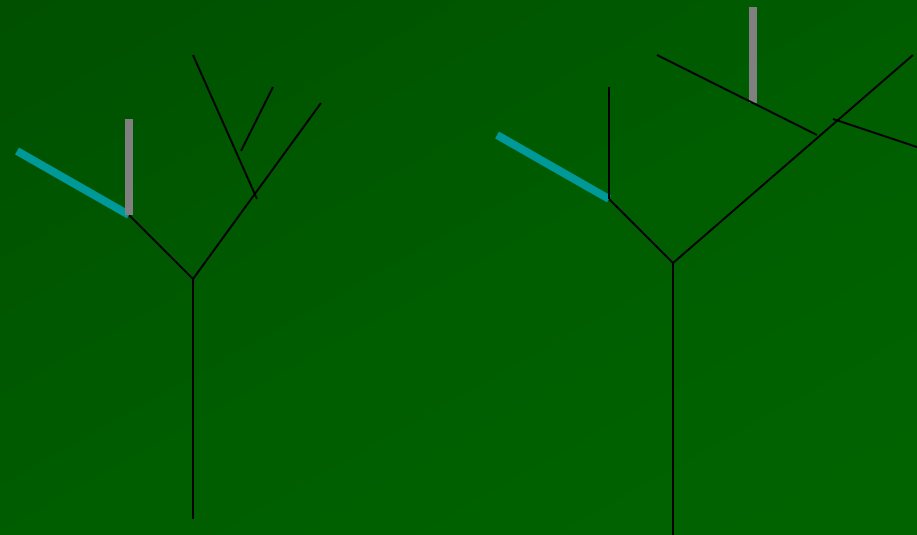
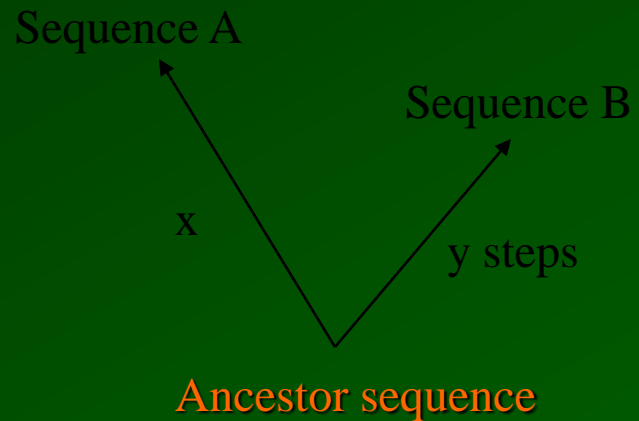
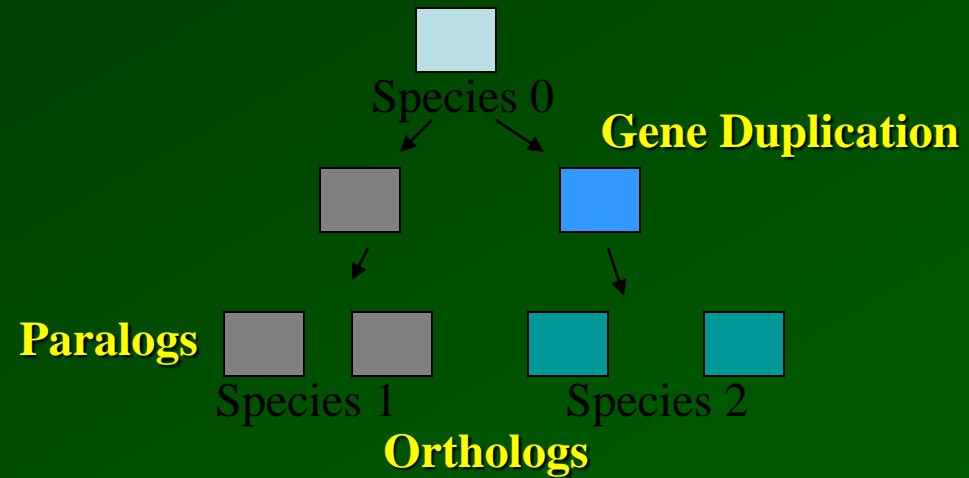
Organism a **A**WTVASAV**R**LSI

Organism b **A**YTVAAA**V**RLSI

Organism c **A**WTVAAA**V**LTSI

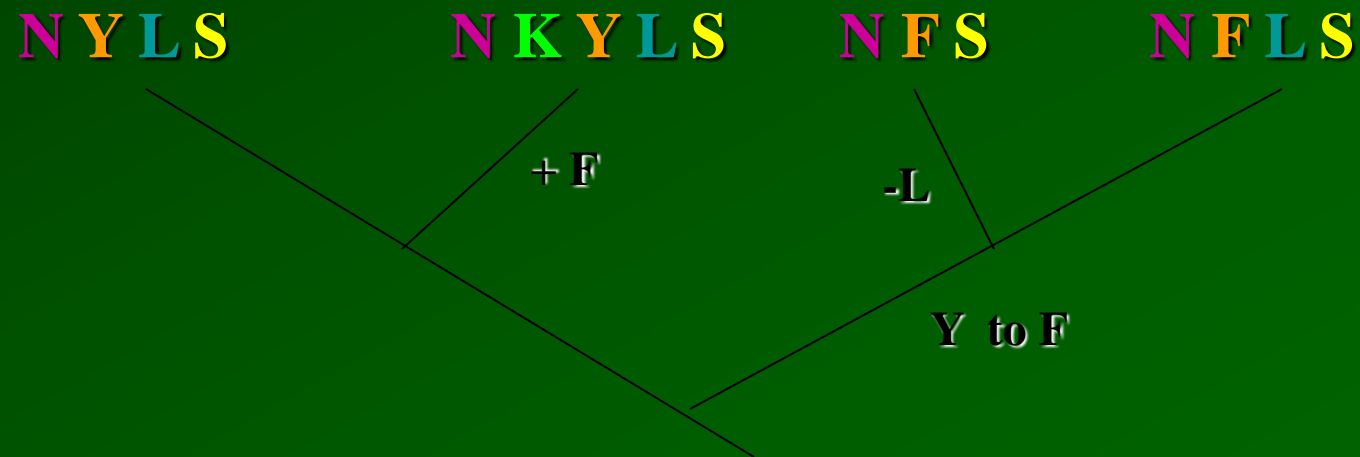


Significance of Sequence Alignments



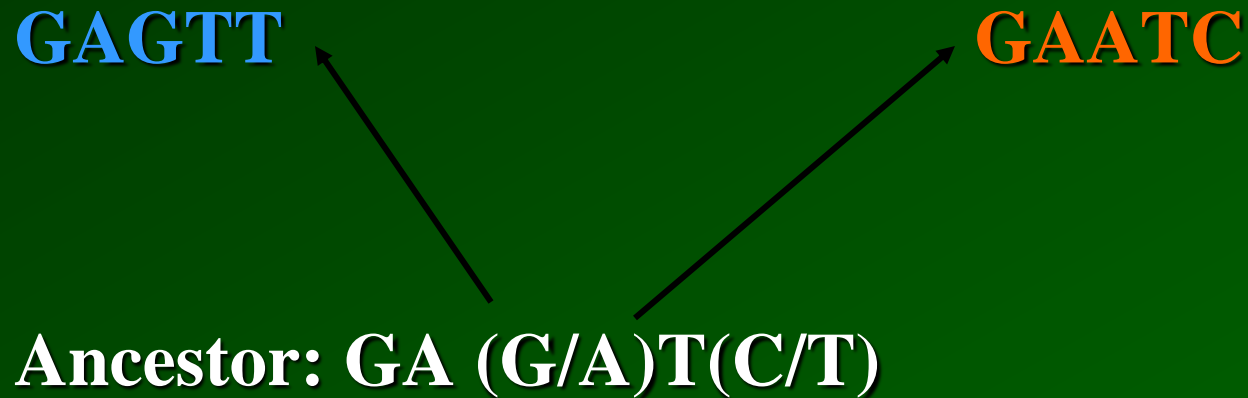
Multiple Sequence Alignment, MSA

seqA N* FLS
seqB N* F- S
seqC NKYLS
seqD N* YLS



Multiple Sequence Alignment & Trees

- MSA is a foundation step in the determination of evolutionary relationships among multiple sequences. They can be of proteins or nucleotides.



***Softwares convert MSA data into a tree using algorithms. Trees can be Rooted/Unrooted.**

MSA: Arabidopsis Actins

File Edit View History Bookmarks Tools Help

http://www.ebi.ac.uk/Tools/es/cgi-bin/clustalw/result.cgi?tool=clu

Customize Links Free Hotmail Windows Marketplace Windows Media Windows

Search Web Mail Shopping My Yahoo! News

Norton Phishing Protection on Identity Safe Log-ins

Additional plugins are required to display all the media on this page. Install Missing Plugins...

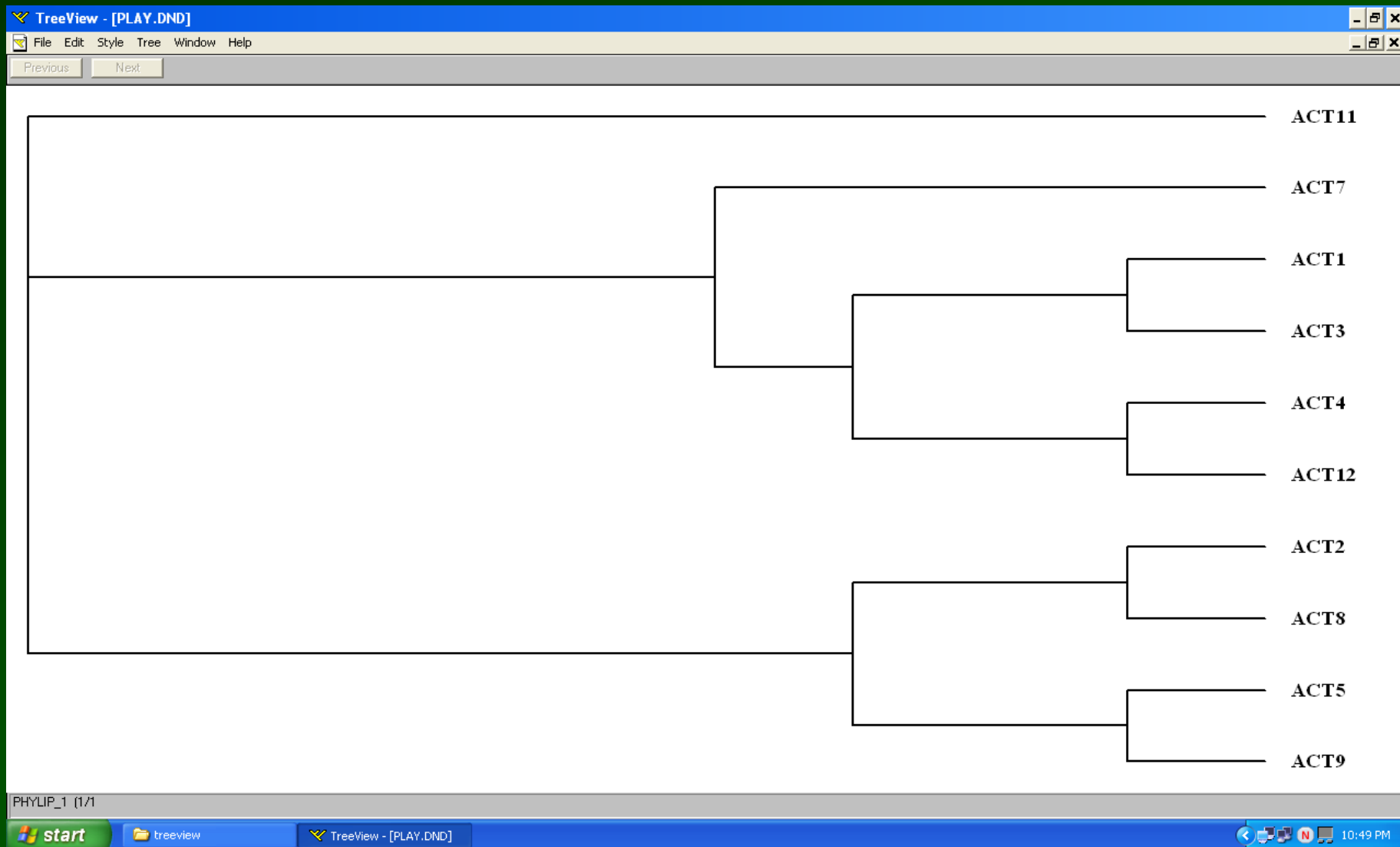
```
ACT2      KMTQIMFETFNSPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGFSLPHAILR 109
ACT8      KMTQIMFETFNSPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGFSLPHAILR 109
ACT1      KMTQIMFETFNAPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILR 179
ACT3      KMTQIMFETFNAPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILR 109
ACT4      KMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILR 179
ACT12     KMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILR 109
ACT7      KMTQIMFETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILR 179
ACT11     KMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILR 109
ACT5      KMTQIMFESFAPPSMYIGIQAVLSLYSSGRTTGIVLDSGDGVSHTVPIYEGYALPHAILR 180
ACT9      KMTQIMFESFDVPAMYVSMQSVLYLYSSGRTTGVVLDLGERVSHIVPVYEGYALPHGILR 98
*****:*  *:*:..*: **:*:*****:*** *: *****:****:***.***

ACT2      LDLAGRDLTDYLMKILTERGYMFTTAEIREIVRDIKEKLSFVAVDYEQEMETSKTSSSIE 169
ACT8      LDLAGRDLTDYLMKILTERGYMFTTAEIREIVRDIKEKLSFVAVDYEQEMETSKTSSSIE 169
ACT1      LDLAGRDLTDALMKILTERGYSFTTAEIREIVRDIKEKLCYIALDYEQELETAKTSSSVE 239
ACT3      LDLAGRDLTDALMKILTERGYSFTTAEIREIVRDIKEKLCYIALDYEQELETAKTSSSVE 169
ACT4      LDLAGRDLTDHLMKILTERGYSFTTAEIREIVRDMKEKLSYIALDYEQELETAKTSSSVE 239
ACT12     LDLAGRDLTDHLMKILTERGYSFTTAEIREIVRDMKEKLSYIALDYEQELETAKTSSSVE 169
ACT7      LDLAGRDLTDSLMLKILTERGYMFTTAEIREIVRDIKEKLAYVALDYEQELETAKSSSVE 239
ACT11     LDLAGRDLTDYLMKILTERGYSFTTAEIREIVRDIKEKLAYVALDYEQELETAKTSSSVE 169
ACT5      LDLAGRDLTEYLTKIMMERGYTYTTSAREIREIVRDIKEKLCYIAVDYEQEMEKATTSSATD 240
```

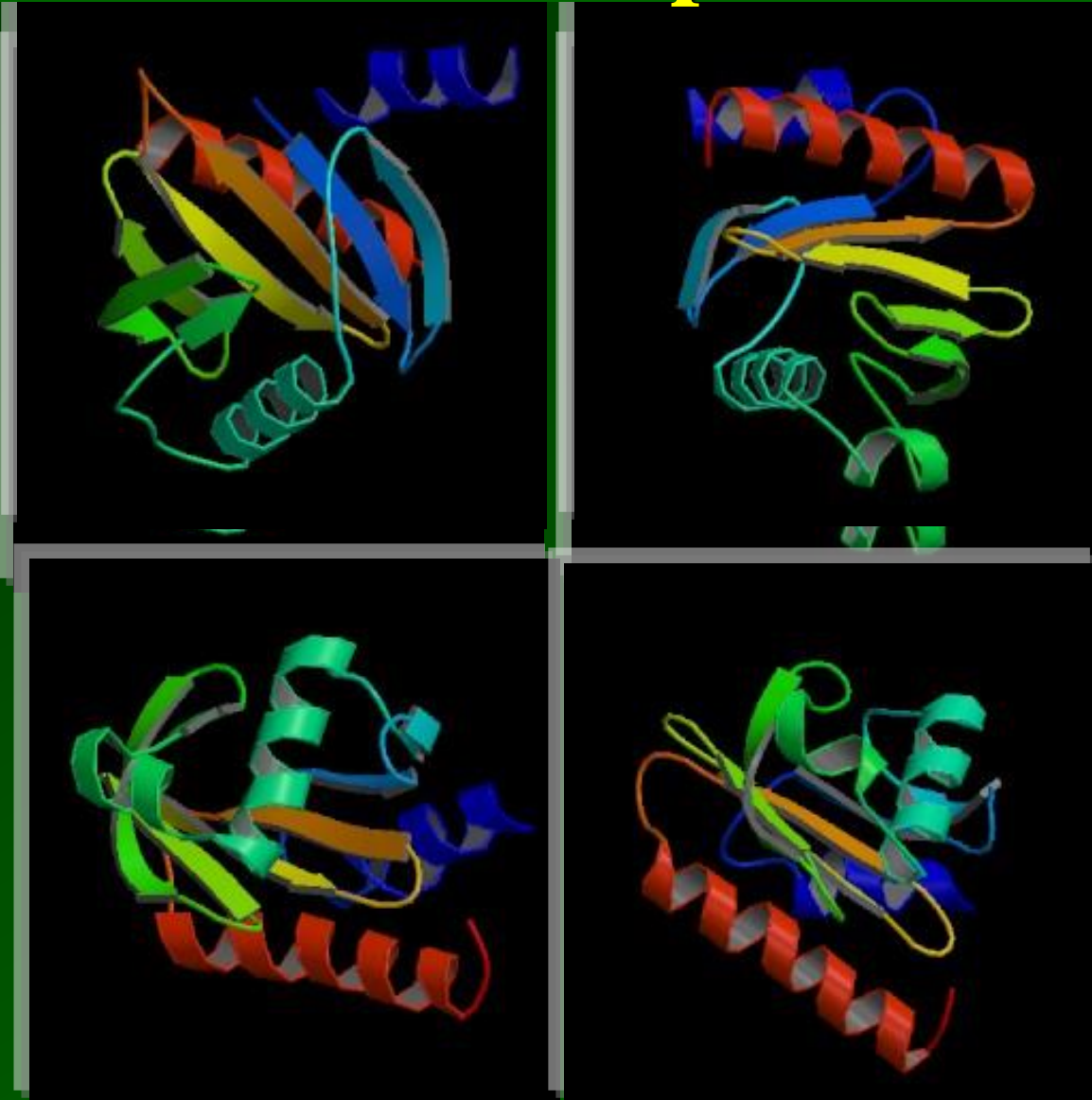
Done

start Yahoo... 3 Fi... gi.do... Micro... nero @SEARCH Norton™ 7:43 PM

Tree of Arabidopsis actins



MSAs also describe protein structures



LOCUS AAA98561 377 aa linear PLN 10-MAR-2011
DEFINITION actin 1 [Arabidopsis thaliana].
ACCESSION AAA98561
VERSION AAA98561.1
DBSOURCE locus ATU39449 accession [U39449.1](#) ****Arabidopsis, Actin, Meagher**
KEYWORDS .

SOURCE Arabidopsis thaliana (thale cress)
ORGANISM [Arabidopsis thaliana](#)
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
Pentapetalae; rosids; malvids; Brassicales; Brassicaceae;
Camelineae; Arabidopsis.

REFERENCE 1 (residues 1 to 377)
AUTHORS An,Y.Q., Huang,S., McDowell,J.M., McKinney,E.C. and Meagher,R.B.
TITLE Conserved expression of the Arabidopsis ACT1 and ACT 3 actin subclass in organ primordia and mature pollen
JOURNAL Plant Cell 8 (1), 15-30 (1996)
PUBMED [8597657](#)

REFERENCE 2 (residues 1 to 377)
AUTHORS An,Y.-Q., Huang,S., McDowell,J.M., McKinney,E.C. and Meagher,R.B.
TITLE Direct Submission
JOURNAL Submitted (25-OCT-1995) Yong-Qiang An, Department of Genetics, University of Georgia, Athens, GA 30602, USA

COMMENT Method: conceptual translation.

FEATURES
Location/Qualifiers
source 1..377
/organism="Arabidopsis thaliana"
/db_xref="taxon:[3702](#)"
[Protein](#) 1..377
/product="actin 1"
/name="ACT1"
[Region](#) 5..377
/region_name="NBD_sugar-kinase_HSP70_actin"
/note="Nucleotide-Binding Domain of the sugar kinase/HSP70/actin superfamily; cl17037"
/db_xref="CDD:[302596](#)"
[Region](#) 8..377
/region_name="ACTIN"
/note="Actin; smart00268"
/db_xref="CDD:[214592](#)"
[Site](#) order(13..16,18,20,139,156..159)
/site_type="other"
/note="nucleotide binding site [chemical binding]"
/db_xref="CDD:[212657](#)"
[CDS](#) 1..377
/coded_by="join(U39449.1:1528..1587,U39449.1:1726..2119,U39449.1:2220..2833,U39449.1:2945..3010)"

ORIGIN
1 madgediqpl vcdngtgmvk agfagddapr avfpsiivgrp rhtgvmvngm qkdayvgdea
61 qskrgiltlk ypiehgivnn wddmekiwhh tfynelrvap eehpilltea plnkanrek
121 mtqimfetfn apamyvaiqa vlslyasgrt tgivldsgdg vshtvpiyeg yalphailrl
181 dlagrdltda lmkiltergy sftttaerei vrdikeklcy ialdyeqele taktssvek
241 nyelpdgqvi tigserfrcp evlyqpsmig menagihett ynsimkcdvd irkdlygniv
301 lsggttmfpg iadrmskeit alapssmkik vvapperkys vwiggasilas lstlqqmwia
361 kaeydesgps ivhrkcf

//



[Identify Conserved Domains](#)

[Highlight Sequence Features](#)

[Find in this Sequence](#)

Related information

[Nucleotide](#)

[PubMed](#)

[Taxonomy](#)

[CDD Search Results](#)

[Conserved Domains \(Concise\)](#)

[Conserved Domains \(Full\)](#)

[Full text in PMC](#)

[Related Structures \(List\)](#)

[Related Structures \(Summary\)](#)

Recent activity

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

[actin 1 \[Arabidopsis thaliana\]](#) Protein

[Actin, Arabidopsis, meagher \(56\)](#) Protein

[actin 4 \[Arabidopsis thaliana\]](#) Protein

[actin-12 \[Arabidopsis thaliana\]](#) Protein

[actin \[Oryza sativa Japonica Group\]](#) Protein

[See more...](#)

Genome Data Viewer

GDV supports the exploration and analysis of *NCBI-annotated* and selected non-NCBI annotated eukaryotic genome assemblies. Currently, over 1250 assemblies are available.

Switch view



Search organisms

Homo sapiens (human)

To view more organisms in the tree, click on nodes that have '+' signs. Press and hold the '+' to expand and reveal all the subgroups. Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.



Homo sapiens (human)

Search in genome

 Examples: TXLNA, chr1:32178000-32200000, DNA repair

Assembly

[Browse genome](#) [BLAST genome](#)

[Download via NC](#) Explore this genome in Genome Data Viewer

Assembly details

Name GRCh38.p13
 RefSeq accession GCF_000001405.39
 GenBank accession GCA_000001405.28
 Submitter Genome Reference Consortium
 Level Chromosome
 Category Reference genome
 Replaced by GCF_000001405.25

Annotation details

Annotation Release 109
 Release date May 16, 2021

“Seeking more...”

Genome Data Viewer

GDV supports the exploration and analysis of *NCBI-annotated* and selected non-NCBI annotated eukaryotic genome assemblies. Currently, over 1250 assemblies are available.

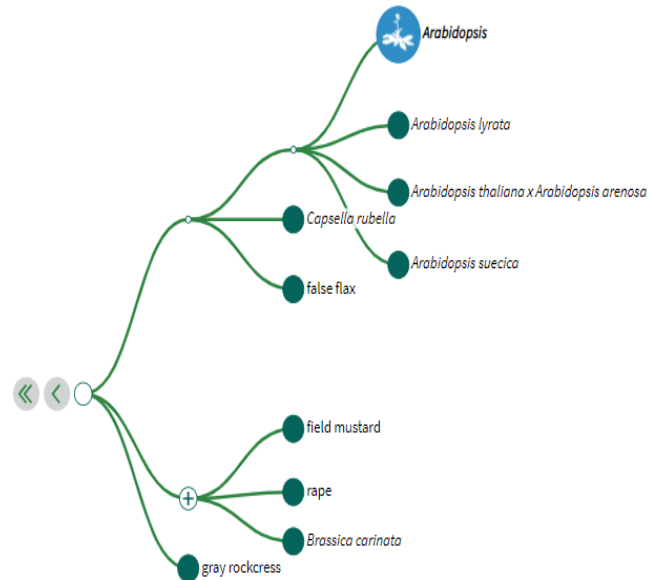
Switch view



Search organisms

Arabisopsis thaliana (thale cress)

To view more organisms in the tree, click on nodes that have '+' signs. Press and hold the '+' to expand and reveal all the subgroups. Or, search for an organism using the search box above.
New! Click on Switch view at the top to see another way of navigating genomes.



Arabisopsis thaliana (thale cress)

Search in genome

Actin

Genes Other

Name	Location
ACT12	Chr3: 17,128,211 - 17,131,466
ACT12	Chr3: 17,128,211 - 17,131,466
ARP2	Chr3: 9,952,479 - 9,955,982
ARP2	Chr3: 9,952,479 - 9,955,982
DIS1	Chr1: 4,494,924 - 4,498,469
DIS1	Chr1: 4,494,924 - 4,498,469
ADF1	Chr3: 16,909,217 - 16,910,845
ADF1	Chr3: 16,909,217 - 16,910,845

Examples: VHA-A1, chr2:12208000-12218000, DNA repair

Assembly

TAIR10.1

Browse genome BLAST genome

Download via NCBI Datasets

Assembly details

Name TAIR10.1
RefSeq accession GCF_000001735.4
GenBank accession GCA_000001735.2
Submitter The Arabidopsis Information Resource (TAIR)
Level Chromosome
Category Reference genome
Replaced by GCF_000001735.2

Annotation details

GDV, Actin genes of *A. thaliana*

Genome Data Viewer

Arabidopsis thaliana
(thale cress)

Assembly: TAIR10.1 (GCF_000001735.4)

Chr 3 (NC_003074.8)

Search assembly

Act12

Examples ▶

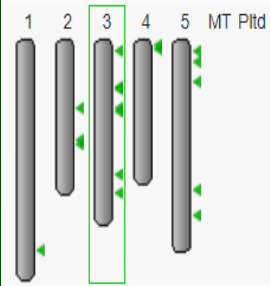
Genes

Other features

Name	Location
ACT12	Chr3: 17,128,211 - 17,131,466
VAM3	Chr5: 19,012,021 - 19,014,110
ACT2	Chr3: 6,474,710 - 6,477,204
ACT7	Chr5: 3,052,062 - 3,054,691
SYP21	Chr5: 5,532,846 - 5,535,295
MUR3	Chr2: 8,791,939 - 8,794,401
AFH1	Chr3: 9,251,093 - 9,255,083
SYP122	Chr3: 19,425,632 - 19,427,230

Pick Assembly

Ideogram View

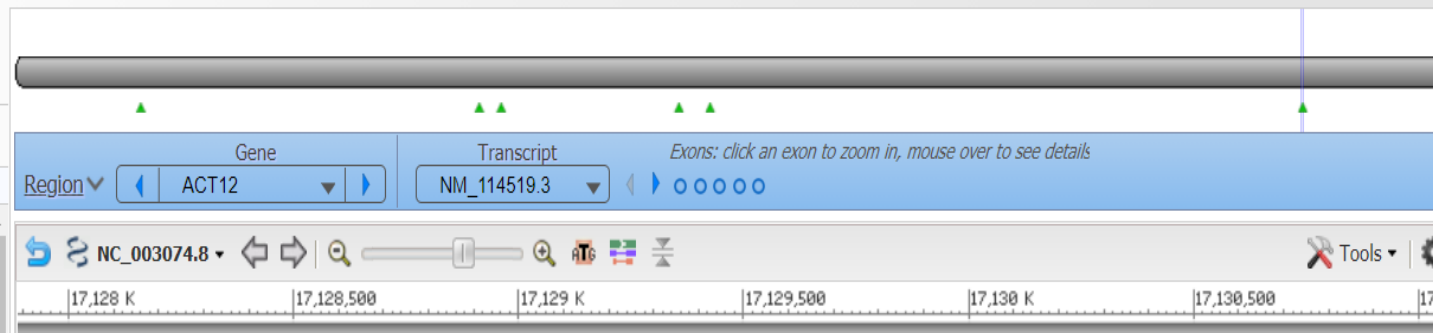


User Data and Track Hubs

BLAST

Add Tracks by Accession

NC_003074.8: 17,127,885 - 17,131,792



Genes, RefSeq propagation from TAIR and Araport, refreshed on 2019-02-...



ACT12

Gene: ACT12
Comment: Member of actin subclass composed of ACT12 and ACT4. RNA is expressed at very low levels in vegetative organs, low levels in flowers and very high levels in pollen. Expression of an ACT12/GUS fusion was found in vascular tissues, tapetum, developing and mature pollen, the root cap and in a ring of pericycle tissues during lateral root initiation and early development.
Location: 17,128,211..17,131,466
Length: 3,256 nt
[Positional Info]
NC_003074.8 position: 17,129,057
Gene position: 847
Merged features: 4

Links & Tools
Araport: AT3G46520
GeneID: 823805 (ACT12)
TAIR: AT3G46520

BLAST nr: NC_003074.8 (17,128,211..17,131,466)
BLAST to Genome: NC_003074.8 (17,128,211..17,131,466)
FASTA record: NC_003074.8 (17,128,211..17,131,466)
GenBank record: NC_003074.8 (17,128,211..17,131,466)

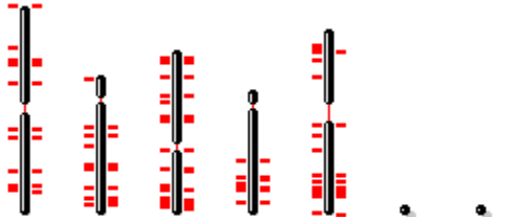
MAPVIEWER: *A. thaliana* ACTIN

NCBI   **NCBI Map Viewer**

PubMed Nucleotide Protein Genome Gene Structure PopSet Taxonomy Help


Search for on chromosome(s) Find

Arabidopsis thaliana (thale cress) genome view BLAST s
Build 8.1 statistics



Hits: **1** 2 3 4 5 MT Pltd
45 93 193 38 150

Search results for query "actin": 511 hits

Hits shown: 1 - 100  1 2 3 4 5 MT Pltd

Chr	Match	Map element	Type	Maps
1	all matches			
	Actin	Les.22414	Les_EST_CI	At UniG
	Transcribed locus, strongly... actin 7) [Arabidops...	Les.22407	Les_EST_CI	At UniG
	Cultivar PKM-1 actin	Les.4694	Les_EST_CI	At UniG

Quick Gene Transcrip all Ref

MAPVIEWER

Map Viewer - Mozilla Firefox

File Edit View History Bookmarks Tools

http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?TAXID=370

Customize Links Free Hotmail Windows Marketplace Windows Media Windows

Search Web

Norton Phishing Protection on Identity Safe Log-ins

PubMed Home Map Viewer

Map Viewer Home
Map Viewer Help
Arabidopsis Maps
Help
FTP
Data As Table View
Maps & Options
Compress Map
Region Shown:

 Go
out
zoom
in
You are here:

IRE1A	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm hm	exte
PFN1/PRF1	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm	exte
PRF5	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm hm	exte
AT2G19960	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm	exte
AT2G25409	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm	exte
VLN1	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm hm	exte
ADF6	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm hm	exte
AT2G31710	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm	exte
AT2G37470	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm hm	exte
ACT1	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm hm	exte
VLN2	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm hm	exte
ACT9	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm	exte
AT2G42100	+	TIGR MIPS TAIR SIGnAL	sv pr dl ev mm hm	exte

Done

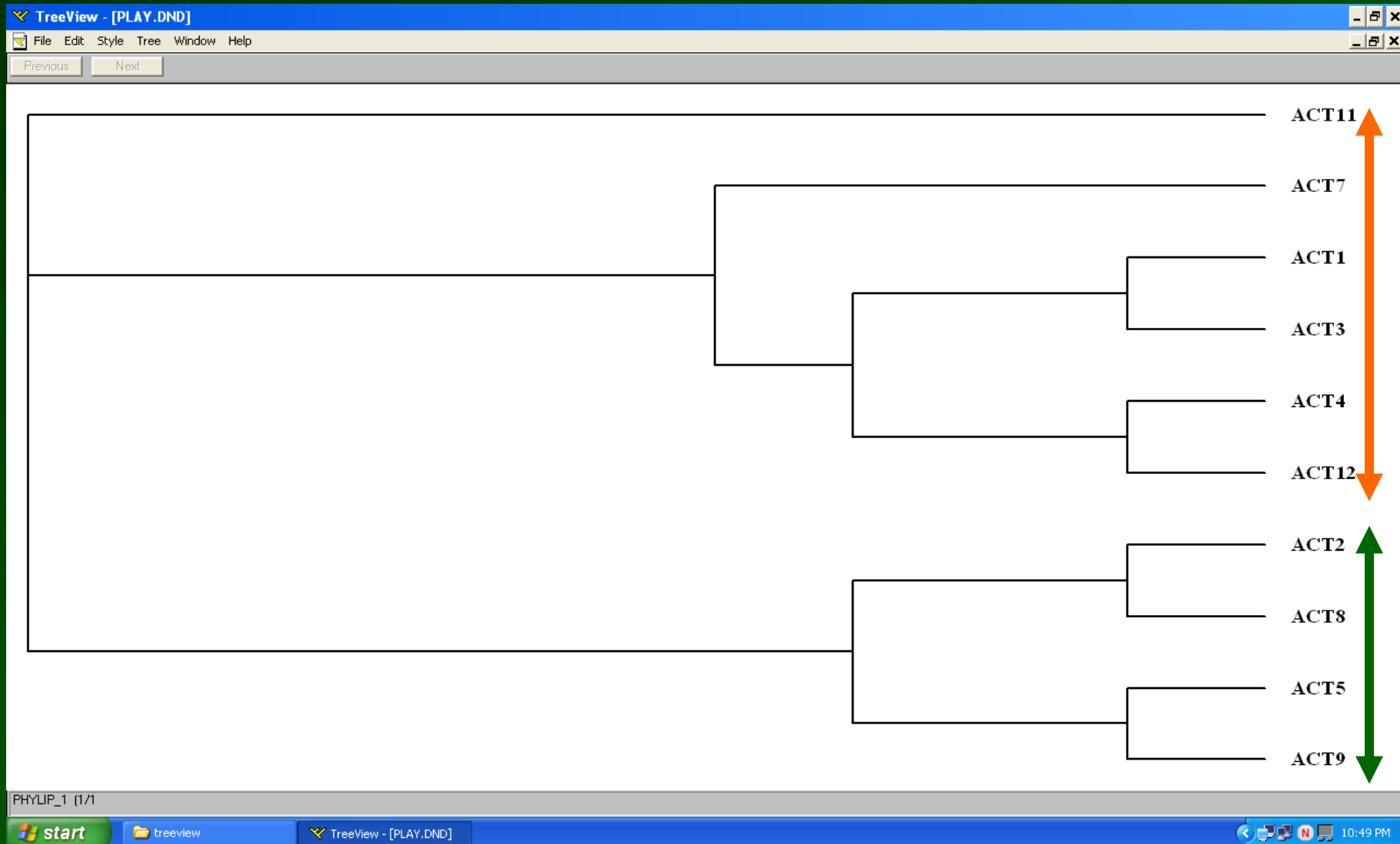
www.sml-botanydu.com

start Yahoo... Map Vi... CB tal... Micros... nero SEARCH Norton™ 8:17 PM

A FASTA FILE

- >ACT1 (**POLLEN**)
MADGEDIQPLVCDNGTGMVKAGFAGDDAPRAVFPSIVGRPRHTGVMVGMGQKDAYVGDEA
QSKRGILTLKYPIEHGIVNNWDDMEKIWHHTFYNELRVAPEEHPILLTEAPLNPKANREKMTQI
MFETFNAPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILRLDLAGRDLT
DALMKILTERGYSFTTTAEREIVRDIKEKLCYIALDYEQELETAKTSSSVEKKNYELPDGQVITIGS
ERFRCPEVLYQPSMIGMENAGIHETTYNSIMKCDVDIRKDL YGNIVLSGGTTMFPGIADRMSKE
ITALAPSSMKIKVVAPPERKYSVWIGGSILASLSTFQQMWIAKA EYDESGPSIVHRKCF
- >ACT2 (**ROOT, STEM, LEAVES of mature plants**)
MAEADDIQPIVCDNGTGMVKAGFAGDDAPRAVFPSVVGRPRHHGVMVGMNPKDAYVGDEA
QSKRGILTLKYPIEHGVVSNWDDMEKIWHHTFYNELRIAPEEHPVLLTEAPLNPKANREKMTQI
MFETFNAPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGFSLPHAILRLDLAGRDLT
DYLMKILTERGYMFTTTAEREIVRDIKEKLSFVAVDYEQEMETSKTSSSIEKKNYELPDGQVITIG
AERFRCPEVLFQPSFVGMEAAGIHETTYNSIMKCDVDIRKDL YGNIVLSGGTTMFSGIADRMSK
EITALAPSSMKIKVVAPPERKYSVWIGGSILASLSTFQQVKIDQILFRILLHAN
- >ACT3 (**POLLEN, OVULES**)
MADGEDIQPLVCDNGTGMVKAGFAGDDAPRAVFPSIVGRPRHTGVMVGMGQKDAYVGDEA
QSKRGILTLKYPIEHGIVNNWDDMEKIWHHTFYNELRVAPEEHPILLTEAPLNPKANREKMTQI
MFETFNAPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILRLDLAGRDLT
DALMKILTERGYSFTTTAEREIVRDIKEKLCYIALDYEQELETAKTSSSVEKKNYELPDGQVITIGS
ERFRCPEVLYQPSMIGMENAGIHETTYNSIMKCDVDIRKDL YGNIVLSGGTTMFPGIADRMSKE
ITALAPSSMKIKVVAPPERKYSVWIGGSILASLSTFQQMWIAKA EYDESGPSIVHRKCF
- >ACT4 (**SEEDS**)

Tree of Arabidopsis actins



Adobe Reader - [picrender.pdf]

File Edit View Document Tools Window Help

Save a Copy Search Select 155% Help

Adobe Reader 7.0

Copyright © 1996 by the Genetics Society of America

Structure and Evolution of the Actin Gene Family in *Arabidopsis thaliana*

John M. McDowell,* Shurong Huang,[†] Elizabeth C. McKinney,[‡] Yong-Qiang An[‡]
and Richard B. Meagher[‡]

**Department of Biology, University of North Carolina, Chapel Hill, North Carolina 27599, [†]Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720 and [‡]Department of Genetics, University of Georgia, Athens, Georgia 30602*

Manuscript received August 10, 1995
Accepted for publication October 27, 1995

ABSTRACT

Higher plants contain families of actin-encoding genes that are divergent and differentially expressed. Progress in understanding the functions and evolution of plant actins has been hindered by the large size of the actin gene families. In this study, we characterized the structure and evolution of the actin gene family in *Arabidopsis thaliana*. DNA blot analyses with gene-specific probes suggested that all 10 of the *Arabidopsis* actin gene family members have been isolated and established that *Arabidopsis* has a much simpler actin gene family than other plants that have been examined. Phylogenetic analyses suggested that the *Arabidopsis* gene family contains at least two ancient classes of genes that diverged early in land plant evolution and may have separated vegetative from reproductive actins. Subsequent divergence produced a total of six distinct subclasses of actin, and five showed a distinct pattern of tissue specific expression. The concordance of expression patterns with the phylogenetic structure is discussed. These subclasses appear to be evolving independently, as no evidence of gene conversion was found. The *Arabidopsis* actin proteins have an unusually large number of nonconservative amino acid substitutions, which mapped to the surface of the actin molecule, and should effect protein-protein interactions.

1 of 16

start CB Talk at KMC 2008, forcb workshop Adobe Reader - [picr... 10:17 PM

Adobe Reader - [picrender.pdf]

File Edit View Document Tools Window Help

Save a Copy Search Select 155% Help

Create an Adobe PDF online

```

Act2 1 .MAD DIQP VCDNGTGMVVKAGFAGDDAPRAVFFS VGRPRRHGVMVGMOKDAYVGD EAQSKRGILTLKYPTEHGVSNWDDMEKIWHHTFYN
Act8 1 .MAD DIQP VCDNGTGMVVKAGFAGDDAPRAVFFS VGRPRRHGVMVGMOKDAYVGD EAQSKRGILTLKYPTEHGVSNWDDMEKIWHHTFYN
Act12 1 .MADGEDIQPIVCDNGTGMVVKAGFAGDDAPRAVFFS VGRPRRHGVMVGMOKDAYVGD EAQSKRGILTLKYPTEHGVSNWDDMEKIWHHTFYN
Act4 1 .MADGEDIQPIVCDNGTGMVVKAGFAGDDAPRAVFFS VGRPRRHGVMVGMOKDAYVGD EAQSKRGILTLKYPTEHGVSNWDDMEKIWHHTFYN
Act1 1 .MADGEDIQPIVCDNGTGMVVKAGFAGDDAPRAVFFS VGRPRRHGVMVGMOKDAYVGD EAQSKRGILTLKYPTEHGVSNWDDMEKIWHHTFYN
Act3 1 .MADGEDIQPIVCDNGTGMVVKAGFAGDDAPRAVFFS VGRPRRHGVMVGMOKDAYVGD EAQSKRGILTLKYPTEHGVSNWDDMEKIWHHTFYN
Act11 1 .MADGEDIQPIVCDNGTGMVVKAGFAGDDAPRAVFFS VGRPRRHGVMVGMOKDAYVGD EAQSKRGILTLKYPTEHGVSNWDDMEKIWHHTFYN
Act7 1 .MADGEDIQPIVCDNGTGMVVKAGFAGDDAPRAVFFS VGRPRRHGVMVGMOKDAYVGD EAQSKRGILTLKYPTEHGVSNWDDMEKIWHHTFYN
Act5 1 MSDLG SVA VCDNGTGMVVKAGFAGDDAPRAVFFS VGRPRRHGVMVGMOKDAYVGD EAQSKRGILTLKYPTEHGVSNWDDMEKIWHHTFYN
Act9 1 ..ASED KP VCDNGTGMVVKAGFAGDDAPRAVFFS VGRPRRHGVMVGMOKDAYVGD EAQSKRGILTLKYPTEHGVSNWDDMEKIWHHTFYN

Scact1 93 ELRVAPEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCV HVVPIYAG LPHAILR DLGRDLTD
Rabsk 93 ELRVAPEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCV HNVPIYEGYALPHAILR DLGRDLTD
Act2 95 ELR APEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCVSHVPIYEGYALPHAILR DLGRDLTD
Act8 95 ELR APEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCVSHVPIYEGYALPHAILR DLGRDLTD
Act12 95 ELRVAPEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCVSHVPIYEGYALPHAILR DLGRDLTD
Act4 95 ELRVAPEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCVSHVPIYEGYALPHAILR DLGRDLTD
Act1 95 ELRVAPEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCVSHVPIYEGYALPHAILR DLGRDLTD
Act3 95 ELRVAPEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCVSHVPIYEGYALPHAILR DLGRDLTD
Act11 95 ELRVAPEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCVSHVPIYEGYALPHAILR DLGRDLTD
Act7 95 ELRVAPEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCVSHVPIYEGYALPHAILR DLGRDLTD
Act5 96 ELR APEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCVSHVPIYEGYALPHAILR DLGRDLTD
Act9 87 ELRVAPEEHPVLLTEAP NPK NREKMTQIMFEPNVPAMYVAIOAVLSLYASGRTTCIVLSDGCVSHVPIYEGYALPHAILR DLGRDLTD

Scact1 188 YLMKIL ERGYSF TTAEREIVRDIKEKLCYVALD EOBM TAAQSS ERSYELPDGQVITIGNERFRFCPEALPHPS G E AGID TTYNSI
Rabsk 188 YLMKIL ERGYSFV TTAEREIVRDIKEKLCYVALD E EMATAA SSS ERSYELPDGQVITIGNERFRFCPEALPHPS G E AGID TTYNSI
Act2 190 YLMKIL ERGYMFT TTAEREIVRDIKEKLS VA DYEQMET KTSS ENNYELPDGQVITIGNERFRFCPEALPHPS G E AGID TTYNSI
Act8 190 YLMKIL ERGYMFT TTAEREIVRDIKEKLS VA DYEQMET KTSS ENNYELPDGQVITIGNERFRFCPEALPHPS G E AGID TTYNSI
Act12 190 YLMKIL ERGYSF TTAEREIVRDIKEKLS VA DYEQMET KTSS ENNYELPDGQVITIGNERFRFCPEALPHPS G E AGID TTYNSI

```

Adobe Reader - [picrender.pdf]

File Edit View Document Tools Window Help

Save a Copy Search Select 155% Help

Create an Adobe PDF online

Arabidopsis Actin Gene Evolution 591

changes. Neighbor joining and maximum likelihood trees were reconstructed in the PHYLIP package (version 3.5) (FELSENSTEIN 1993). Third codon positions were assigned a weight of zero in every case. The inputs for the neighbor joining program were distance matrices constructed by KIMURA's two parameter method. Other methods of calculating genetic distance were also employed and yielded trees with very similar branching orders.

Construction of space-filling models for plant and human actins: Sequences for human Act-b, Act-g, Act-c, Act-s, Act-a, and Act-h were obtained from GenBank and aligned to identify variable residues. Structural coordinates for the model of rabbit skeletal actin-DNaseI complex (KABSCH *et al.* 1990) were retrieved from the Brookhaven Database. Rabbit skeletal actin is identical in sequence and presumably in structure to human Act-s and was used as the basis for building a model plant actin. The DNaseI molecule was deleted from this model using Nitro (Tripos Assoc. Inc.). Building plant ACT2 model structure required making 52 amino acid substitutions within the above skeletal muscle actin. Changing residue 7 from ala7 in Act-s to pro7 in ACT2 could not be done without making a minor change in conformation of the peptide backbone. Residues which varied within the plant or animal families were located within the Act-s and ACT2 structures, respectively, and displayed in a spacefilling model.

RESULTS

	ACT1	ACT2	ACT3	ACT4	ACT5	ACT7	ACT8	ACT9	ACT11	ACT12	SCER
ACT1	0	0.0666	0.0023	0.0251	0.0994	0.0420	0.0647	0.1787	0.0317	0.0257	0.1550
ACT2	1.297	0	0.0660	0.0635	0.1018	0.0426	0.0046	0.1814	0.0518	0.0648	0.1619
ACT3	6.055	1.133	0	0.0269	0.0994	0.0420	0.0641	0.1800	0.0347	0.0251	0.1561
ACT4	1.242	1.677	1.432	0	0.1026	0.0494	0.0604	0.1789	0.0293	0.0046	0.1570
ACT5	1.452	1.969	1.366	1.444	0	0.1068	0.0973	0.1608	0.0935	0.1059	0.1887
ACT7	0.9957	1.176	1.192	1.483	1.776	0	0.0408	0.1806	0.0403	0.0469	0.1616
ACT8	1.067	0.5571	1.228	1.351	1.603	1.045	0	0.1769	0.0493	0.0598	0.1634
ACT9	1.552	1.562	1.458	1.434	0.3960	1.676	1.292	0	0.1729	0.1759	0.2499
ACT11	1.100	1.097	1.289	1.438	1.342	1.250	1.218	1.170	0	0.0293	0.1571
ACT12	1.447	1.722	1.331	0.8351	1.535	1.569	1.764	1.667	1.426	0	0.1641
SCER	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	0

FIGURE 2.—Nonsynonymous and synonymous nucleotide substitutions among Arabidopsis actin genes. Nonsynonymous (RNS) and synonymous (SNS) substitutions (above and below the diagonal, respectively) were calculated for all pairwise comparisons of the Arabidopsis actin protein coding sequences (see MATERIALS AND METHODS). The numbers shown were corrected for multiple hit kinetics and are given as a fraction of the total number possible substitutions (*i.e.*, in the RNS comparison of ACT1 vs. ACT2 the value of 0.0666 means that 6.66% of the total possible changes occurred). Values which corrected to numbers >2.0 (200%) are listed as NAN.

Adobe Reader - [picrender.pdf]

File Edit View Document Tools Window Help

Save a Copy Search Select 155% Help

Create an Adobe PDF online

J. M. McDowell *et al.*

shared by ACT7 and subclasses T12. The s in actin genesis in t charged ERTMAN *et al.* a strong anges be-comitant significant ce of pro7 es relative lanine in ant actins o the ani-when this ctin back-rotein has nerization

Gene	Cy	Hy	Sc	Pr	Vy	Vo	Cy	Co	Le	Tr	Rc	Fl	P	Ov	Em	Si
ACT2	+	+/-	-	+	+	+	+	+	+	+	+	+	-	-	-	+/-
ACT8	-	-	+	+	+	+	+	+	+	+	+	+	+	-	-	+/-
ACT7	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ACT11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ACT1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ACT3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ACT4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ACT12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ACT5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ACT9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

FIGURE 8.—Summary of the expression patterns of the Arabidopsis gene actin genes. The phylogenetic relationships of the 10 Arabidopsis actin genes, based on protein coding sequence, is presented on the left. To the right is a table that outlines the expression of the genes in the following tissues and organs: Cy, tissues of cotyledons excluding young vascular tissue; Hy, hypocotyl; Sc, seed coat; Pr, organ primordia and meristems; Vy, young vascular tissue; Vo, old vascular tissue; Cy, young cortical tissue; Co, old cortical tissue, Le, leaf epidermis; Rc, root cap; P, pollen; Ov, developing ovules; Em, embryo; Fl, flower (*i.e.*, sepals, petals, stigma, style) but excluding pollen, pollen sack, developing ovules; and Si, siliques. The + and - score expression and no expression, respectively, and a +/- score implies very weak expression or that expression is limited to a small subset of the cells present in this organ or tissue. These data are summarized from several recent manuscripts from this laboratory (Y.-Q. AN, S. HUANG, J. M. MCDOWELL, E. C. MCKINNEY and R. B. MEAGHER; Y.-Q. AN, J. M. MCDOWELL, S. HUANG, E. C. MCKINNEY, S. CHAM

Hypothesis

“The macro-evolution of vegetative and reproductive plant structures is linked to the molecular evolution of two, differentially regulated classes of cytoskeletal genes that physically direct development.”

Link Out & Explore “A Brave New World”

COMMENT <http://www.genoscope.cns.fr>
Information on performance of analysis and a more detailed annotation of this entry and other sequences of chromosomes 3, 4 and 5 can be viewed at: <http://www.mips.biochem.mpg.de/proj/thal/>

FEATURES Location/Qualifiers

source 1..377
/organism="Arabidopsis thaliana"
/db_xref="taxon:3702"
/chromosome="3"
/ecotype="Columbia"

[Protein](#) 1..377
/product="actin 12"

[Region](#) 5..377
/region_name="NBD_sugar-kinase_HSP70_actin"
/note="Nucleotide-Binding Domain of the sugar kinase/HSP70/actin superfamily; cl17037"
/db_xref="CDD:418402"

[Site](#) order(13..16,18,20,139,156..159)
/site_type="other"
/note="nucleotide binding site [chemical binding]"
/db_xref="CDD:212657"

[CDS](#) 1..377
/gene="F12A12.40"
/coded_by="join(AL133314.1:27669..27728, AL133314.1:27811..28204, AL133314.1:28304..28917, AL133314.1:29018..29083)"
/note="Contains Actins signatures AA55-65;Actins signatures AA358-366;Actins signatures AA106-118"
/db_xref="GOA:P53497"
/db_xref="InterPro:IPR004000"
/db_xref="InterPro:IPR004001"
/db_xref="InterPro:IPR020902"
/db_xref="UniProtKB/Swiss-Prot:P53497"

ORIGIN

```
1 madgediqpl vcdngtgmvk agfagddapr avfpsivgrp rhtgvmvgmg qkdayvgdea
61 qskrgiltlk ypiehgivnn wddmekiwhh tfynelrvap eehpvlltea plnpanrek
121 mtqimfetfn tpamyvaiqa vlslyasgrt tgivldsgdg vshtvp iyeg yalphailrl
181 dlagrdltdh lmkiltergy sftttaerei vrdmkeklsy ialdyeqele tsktsssvk
```

More about the gene ACT12

Member of actin subclass composed of ACT12 and ACT4. RNA is expressed at very low levels in vegetative organs, low levels in flowers and ver... Also Known As: AT3G46520, ACTIN, actin...

Homologs of the ACT12 gene

The ACT12 gene is conserved in human, chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, fruit fly, rice, and frog.

Related information

BioProject

Nucleotide

Taxonomy

BioSystems

CDD Search Results

Conserved Domains (Concise)

Conserved Domains (Full)

Gene

Genome

PubMed (Weighted)

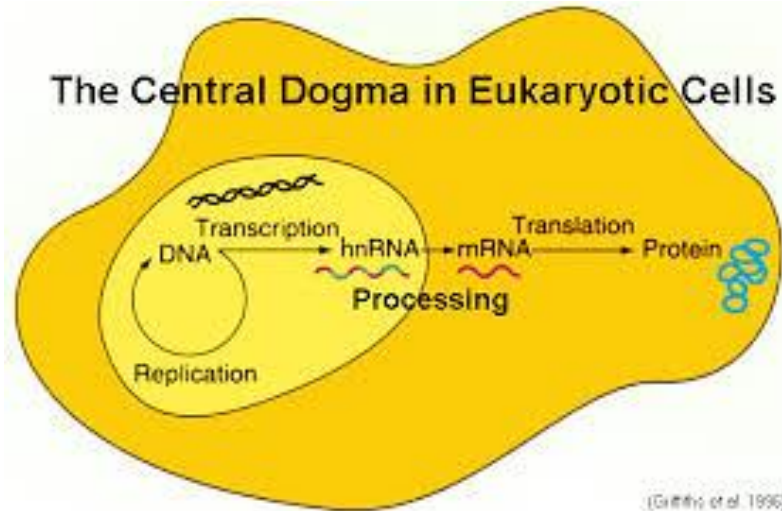
Related Structures (List)

Related Structures (Summary)

Suggested Readings

- RB Meagher, EC McKinney and AV Vitale, 1999 “The evolution of new structures: clues from plant cyto-skeletal genes” TIG 15:7, 278-284.
- Bioinformatics: A Practical Approach, Higgins & Taylor, OUP (Indian ed.) Rs. 295
- Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids, Durbin et al., CUP (Indian ed.) Rs. 395.
- *Surf the web!!*

Protein sorting in plant cells



1974 Nobel Prize in Physiology/Medicine

– George Palade

- “for discoveries concerning the structural and functional organization of the cell”



1999 Nobel Prize in Physiology/Medicine

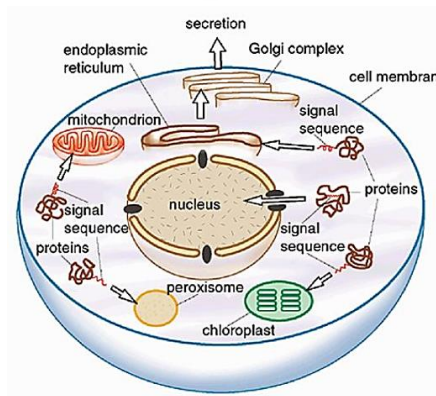
– Günter Blobel

- “for the discovery that proteins have intrinsic signals that govern their transport and localization in the cell”

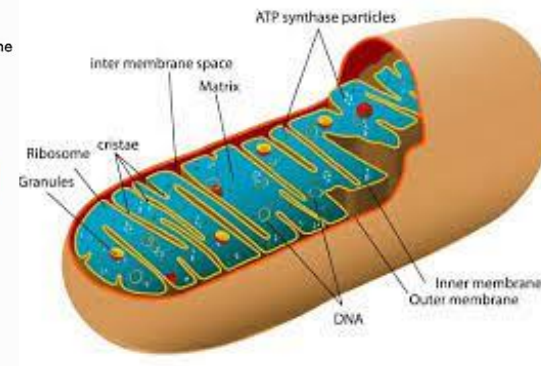


- **Protein sorting** is the mechanism by which a cell transports proteins to the appropriate positions in the cell or outside of it.
- Valuable annotations for novel proteins
- Proteins can move between compartments in different ways –
 - Gated transport (Nucleus)
 - Transmembrane transport (Mitochondria, Peroxisomes)
 - Vesicular transport (ER)

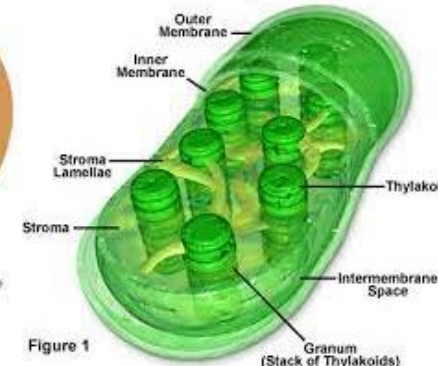
Compartmentalization in cellular organelles



Eukaryotic cell



Mitochondria



Chloroplast

Protein subcellular localization

Some signals are easily recognizable while others are difficult to understand

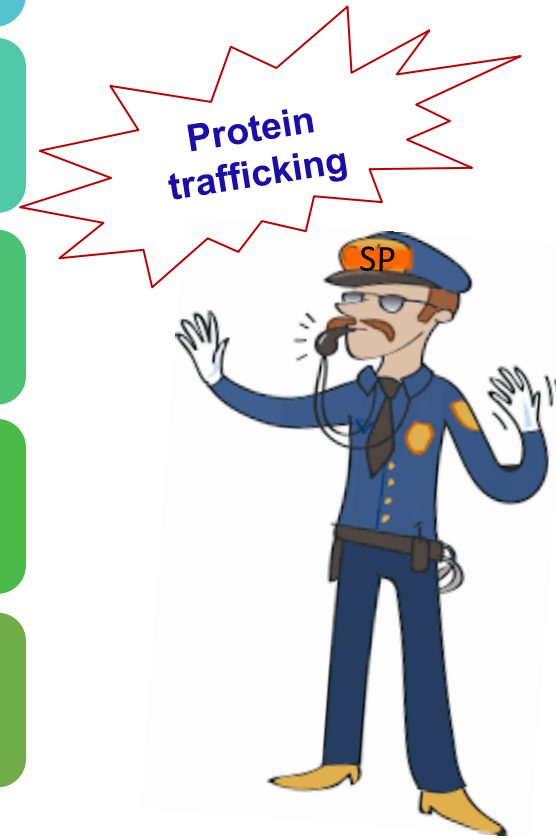
Proteins must have **intrinsic signals for their localization** – a cellular address (eg. **N-terminal signal sequences**)

➤ The pre-sequences of the targeting peptides are often found at the **N-terminal extension**.

➤ It is composed of between 6 to 136 basic and hydrophobic amino acids.

➤ Signal sequences are removed from the finished protein by specialized signal peptidases once the sorting process has been completed.

Enough experimental data exists to build highly accurate computational predictors of localization.



Common structure of N-terminal signal peptide

H. Owji et al.

European Journal of Cell Biology 97 (2018) 422-441

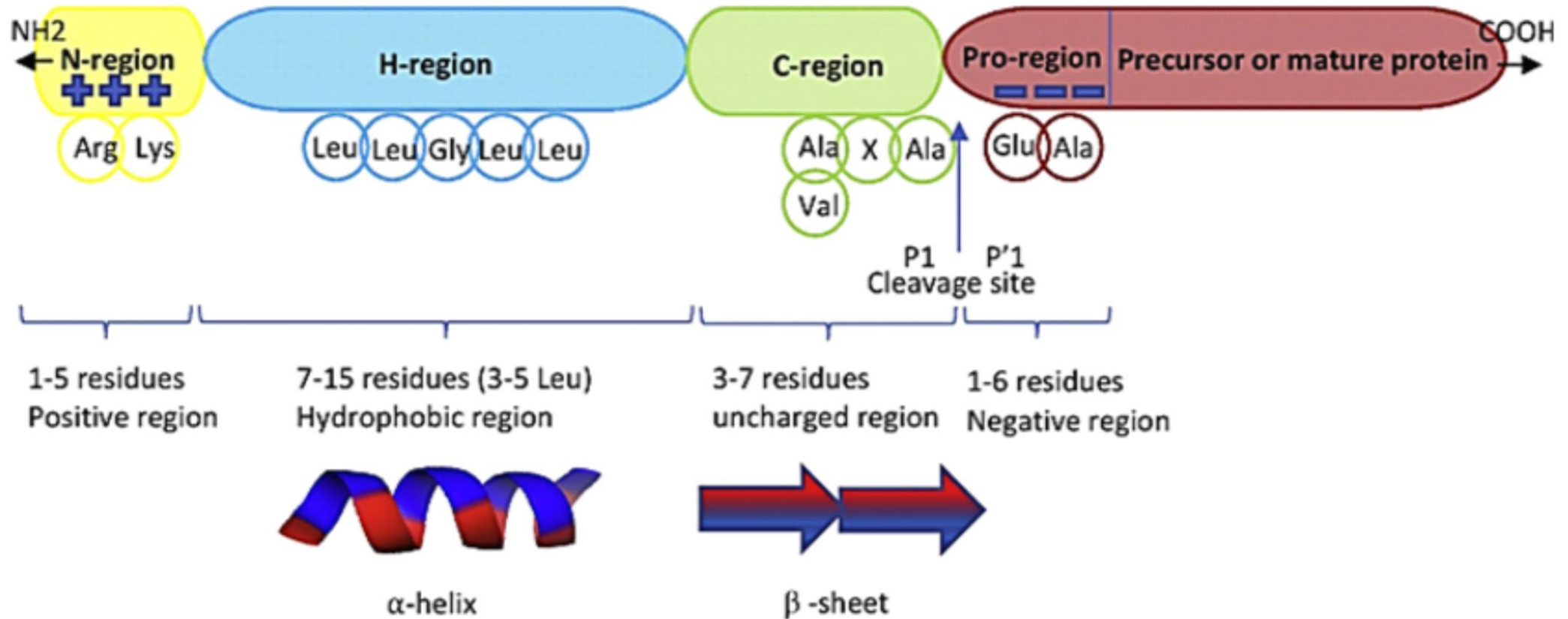


Fig. 2. The general structure of an SP. It is composed of three main parts: 1) N-region- the positive-charged domain 2) H-region- the hydrophobic core, forming α -helix 3) C-region- the cleavage site, forming β -sheet. The initial part of protein, important to protein secretion, is called Pro-region. Residues before the cleavage site are called as P1 and after the cleavage site are called as P'1. Cleavage occurs at AXA or VXA motif.

Predictions from known data

Different information can be used for predictions from known data

Sequence motifs

Amino acid composition

Homology

N-terminal

C-terminal

Mid-sequence

AA frequency

Dipeptide composition

Secretory signal peptides

Mitochondrial targeting peptides

Chloroplast transit peptides

Peroxisome import signal

ER retention signal

NLS

The PSORT family

The PSORT family server (<http://psort.nibb.ac.jp/>; <http://psort.org/>) contains several variant tools for the prediction of protein localization sites in cells.

Then, it analyzes the input sequence by applying the stored rules for various sequence features of known protein sorting signals.

It receives the information of an amino acid sequence and its source origin, e.g., Gram-negative bacteria, as inputs.

Finally, it reports the possibility for the input protein to be localized at each candidate site with additional information.

Variant tools of the PSORT family

- **PSORT**: an old version for plants and bacteria
- **PSORT II**: recommended for yeast and animal
- **iPSORT**: for N-terminal sorting signals for plants or non-plants
- **PSORT-B**: recommended for Gram-negative bacteria
- **WoLF PSORT**: recommended for animal, plant and fungi

What is iPSORT – Overview



iPSORT is a subcellular localization site predictor for **N-terminal** sorting signals.

Given a protein sequence, it will predict whether it contains a (Signal Peptide **SP**), Mitochondrial Targeting Peptide (**mTP**), or Chloroplast Transit Peptide (**cTP**).

Getting started with iPSORT



ipsort in bioinformatics



[All](#)

[Videos](#)

[Images](#)

[News](#)

[Maps](#)

[More](#)

[Tools](#)

About 15,100 results (0.43 seconds)

<https://ipsort.hgc.jp>

iPSORT Home Page

A command line version of iPSORT is available here. ... feature detection of N-terminal protein sorting signal", *Bioinformatics*, 18(2) 298-305, 2002.

You visited this page on 21/7/21.

<https://academic.oup.com> › [bioinformatics](#) › [article](#)

Extensive feature detection of N-terminal protein sorting ...

by H Bannai · 2002 · Cited by 806 — *Bioinformatics*, Volume 18, Issue 2, February 2002, Pages 298–305, ... usingrules obtained by our method is provided at...

<https://www.sciencedirect.com> › [topics](#) › [psort](#)

PSORT - an overview | ScienceDirect Topics

... (3) iPSORT: for N-terminal sorting signals for plants or non-plants; (4) PSORT-B: recommended for ... *Bioinformatics Tools for Reverse Vaccinology*.

<https://mybiosoftware.com> › [tag](#) › [ipsort](#)

iPSORT – My Biosoftware – Bioinformatics Softwares Blog

13-Nov-2013 — iPSORT / caml-iPSORT 20100316. :: DESCRIPTION. iPSORT is a subcellular localization site predictor for N-terminal sorting signals.

iPSORT homepage opens

← → ↻ 🏠 🔒 ipsort.hgc.jp



Welcome to the iPSORT WWW Service

This page is currently maintained by [Yoshinori Tamada at Hirosaki University](#) and [Division of Health Medical Intelligence, Human Genome Center, Institute of Medical Science, University of Tokyo](#).
Questions, comments, bug reports, etc. should be directed to: Hideo Bannai <hdbn.dsc@tmd.ac.jp> and [Yoshinori Tamada](#).
Best viewed with a style sheet capable browser.

A command line version of iPSORT is available [here](#). The program has been rewritten in the OCaml language, and is different from the one used for this web predictor (but should give the same predictions).

Contents

- [What is iPSORT and how does it work?](#)
- [Predict localization sites of protein sequences](#)
- [Run amino acid index analysis](#)
- [Reference](#)

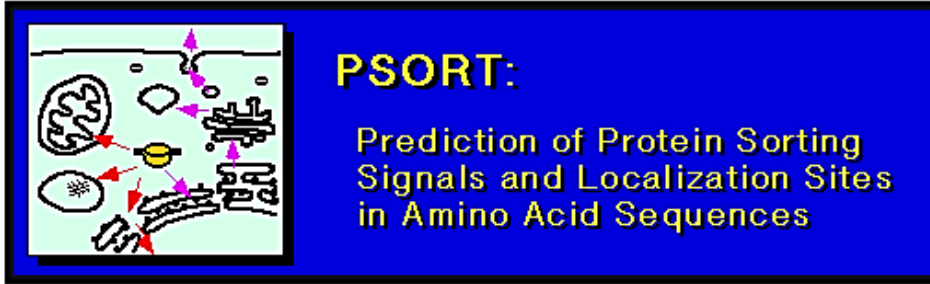
iPSORT Prediction

Paste your amino acid sequence below:

*** Characters except the standard 20 codes will be removed off

*** Only the first 30 residues are used for prediction

Link to PSORT www service (PSORT, PSORT II): The original PSORT and PSORT II



PSORT WWW Server

PSORT is a computer program for the prediction of protein localization sites in cells. It receives the information of an amino acid sequence and its source origin, *e.g.*, Gram-negative bacteria, as inputs. Then, it analyzes the input sequence by applying the stored rules for various sequence features of known protein sorting signals. Finally, it reports the possibility for the input protein to be localized at each candidate site with additional information.

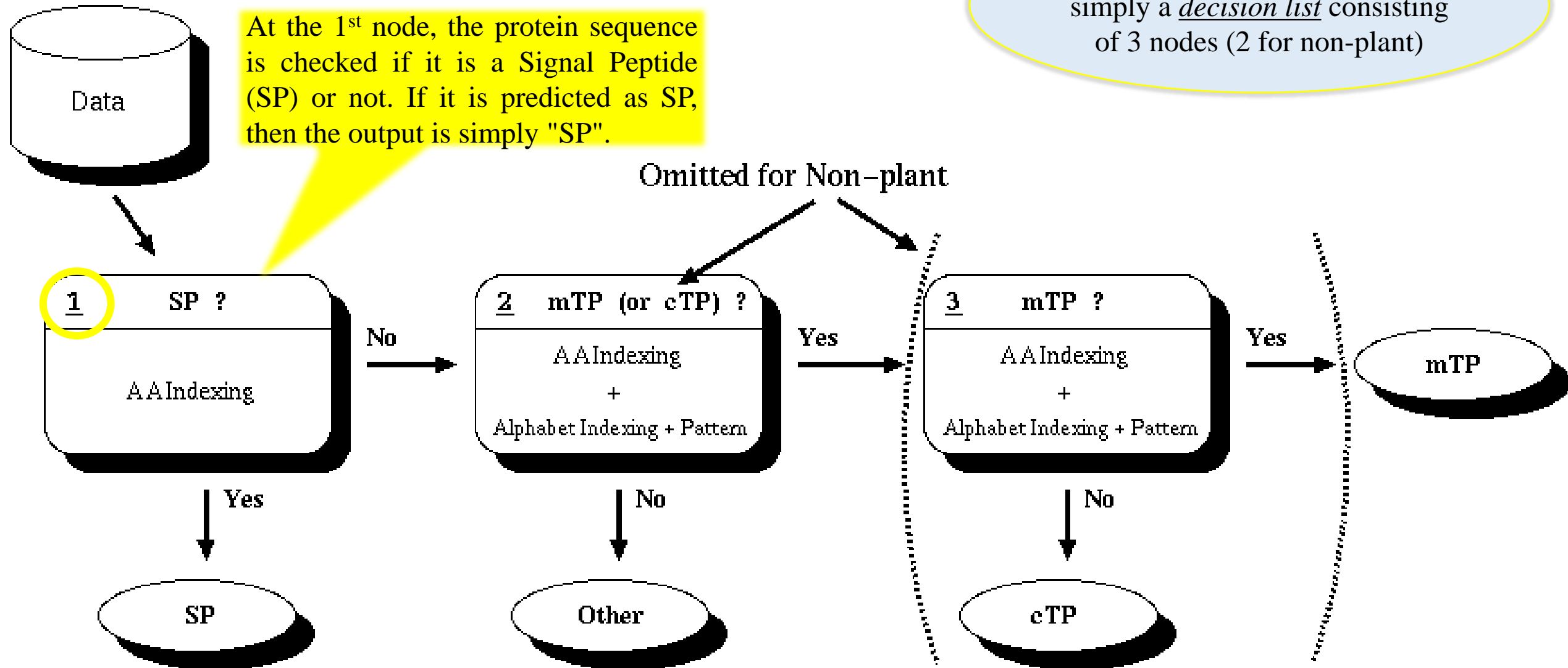
PSORT is mirrored at [Tokyo](#), [Okazaki](#), and [Peking](#)

- December 1, 1998, Official release of the PSORT II package
 - June 1, 1999, K. Nakai moved to Univ. Tokyo
 - October 13, 1999, The Web server has been moved from Osaka to Tokyo
 - March 11, 2001, Introduction of iPSORT
 - September 23, 2001, New mirror site at Peking University
 - December 22, 2001, Distribution of caml-iPSORT
 - January 18, 2003, Replacing the training data for PSORT II at Peking
 - February 22, 2003, Rebuilding the PSORT II server at Tokyo
 - April 16, 2003, Minor update of the top page
 - November 9, 2003, Minor updates of several pages
 - May 27, 2005, Link to WoLF PSORT; update some links
 - January 5, 2007, Modification of the link to WolfPSORT
-

iPSORT – Structure

The structure of iPSORT is simply a *decision list* consisting of 3 nodes (2 for non-plant)

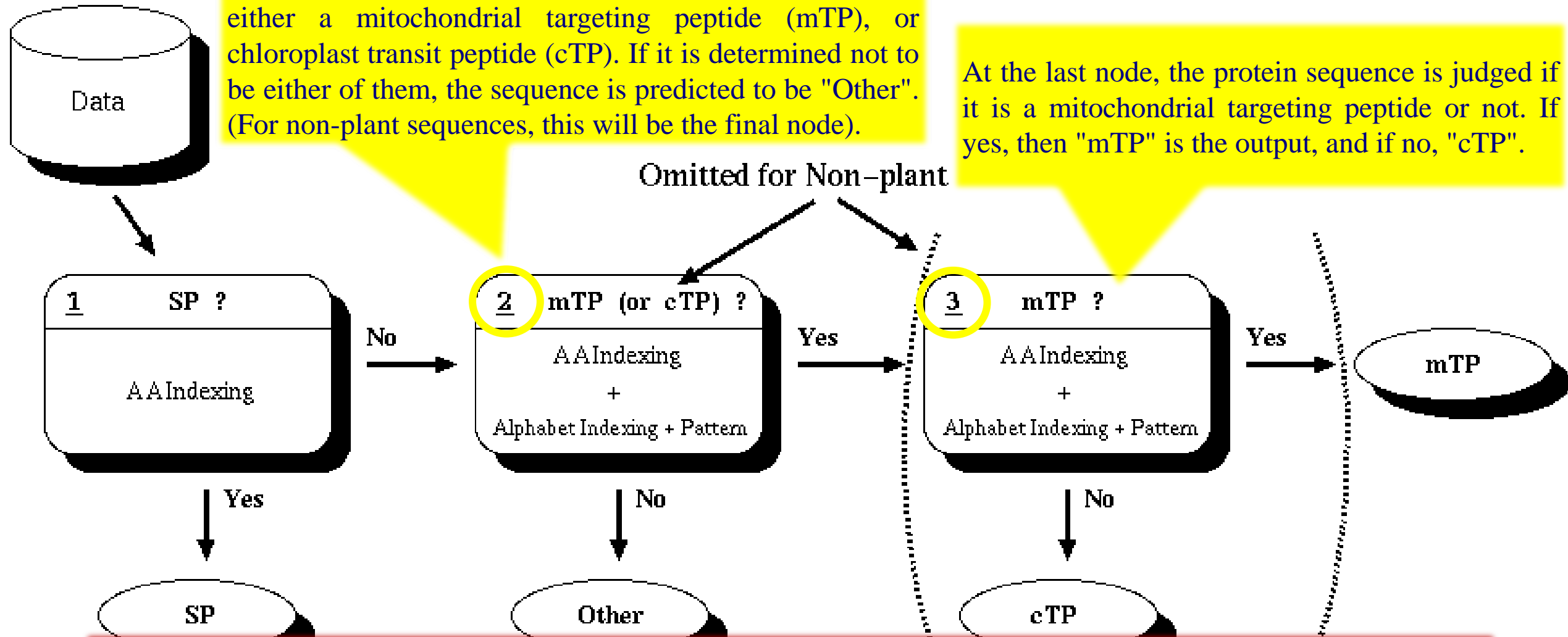
At the 1st node, the protein sequence is checked if it is a Signal Peptide (SP) or not. If it is predicted as SP, then the output is simply "SP".



iPSORT – Structure

At the 2nd node, the protein sequence is judged if it is either a mitochondrial targeting peptide (mTP), or chloroplast transit peptide (cTP). If it is determined not to be either of them, the sequence is predicted to be "Other". (For non-plant sequences, this will be the final node).

At the last node, the protein sequence is judged if it is a mitochondrial targeting peptide or not. If yes, then "mTP" is the output, and if no, "cTP".



The rules deciding whether or not the given signals contain a certain signal consists of two elements: an *amino acid index rule*, and an *alphabet indexing+pattern rule*. (except for SP with only an amino acid index rule). To be judged "yes" at each node, *the input amino acid sequence must satisfy both of the two rules* (except SP).

Suggested readings for iPSORT

- Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S., "*Extensive feature detection of N-terminal protein sorting signal*", *Bioinformatics*, 18(2) 298-305, 2002.
- Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S., *Views: Fundamental Building Blocks in the Process of Knowledge Discovery*, In Proceedings of the 14th International FLAIRS Conference, 233-238, AAAI Press, 2001.
- Gardy, J. L., Spencer, C., Wang, K., Ester, M., Tusnady, G. E., Simon, I., ... & Brinkman, F. S. (2003). *PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria*. *Nucleic acids research*, 31(13), 3613-3617.
- Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. (2007). *WoLF PSORT: protein localization predictor*. *Nucleic acids research*, 35(suppl_2), W585-W587.
- Hiller, K., Grote, A., Scheer, M., Münch, R., & Jahn, D. (2004). *PrediSi: prediction of signal peptides and their cleavage positions*. *Nucleic acids research*, 32(suppl_2), W375-W379.
- Zheng, Z., Chen, Y., Chen, L., Guo, G., Fan, Y., & Kong, X. (2012). *Signal-BNF: a Bayesian network fusing approach to predict signal peptides*. *Journal of Biomedicine and Biotechnology*, 2012.

Sequence retrieval from database

- ❑ Search the [Gene](#) database with the desired gene name, species, or accession number.
- ❑ Nucleotide/ Protein sequence: Basic Local Alignment Search Tool (BLAST) tool.

1

NCBI database website:

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

2

NCBI Resources How To Sign in to NCBI

All Databases XP_004135094 Search

COVID-19 Information
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

UNITE
A new NIH initiative to end structural racism and achieve racial equity in the biomedical research enterprise.
LEARN MORE

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine

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

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene

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

3

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

Web BLAST

Nucleotide BLAST
nucleotide → nucleotide

blastx
translated nucleotide → protein

Protein BLAST
protein → protein

blastn
protein → translated nucleotide

BLAST® » blastp suite Home Recent Results Saved Strategies Help

blastn **blastp** blastx tblastn tblastx

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence
Enter accession number(s), gi(e), or FASTA sequence(s) Clear

Query subrange
From To

Or, upload file Choose File No file chosen

Job Title
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism Optional Enter organism name or id—completions will be suggested exclude Add organism

Exclude Optional Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Program Selection

Algorithm
 Quick BLASTP (Accelerated protein-protein BLAST)
 blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm

BLAST Search database nr using Blastp (protein-protein BLAST)
 Show results in a new window

Preparation of FASTA files

GenPept Send to: ▾

ribulose biphosphate carboxylase small chain, chloroplastic [Cucumis sativus]

NCBI Reference Sequence: XP_004135094.1
[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to: ☑

LOCUS XP_004135094 185 aa linear PLN 17-DEC-2019
DEFINITION ribulose biphosphate carboxylase small chain, chloroplastic [Cucumis sativus].
ACCESSION XP_004135094
VERSION XP_004135094.1
DBLINK BioProject: [PRJNA182750](#)
DBSOURCE REFSEQ: accession [XM_004135046.3](#)
KEYWORDS RefSeq.
SOURCE Cucumis sativus (cucumber)
ORGANISM [Cucumis sativus](#)
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Cucurbitales; Cucurbitaceae; Benincaseae; Cucumis.
COMMENT MODEL REFSEQ: This record is predicted by automated computational analysis. This record is derived from a genomic sequence ([NC_026659.2](#)) annotated using gene prediction method: Gnomon, supported by mRNA and EST evidence.
Also see:
[Documentation](#) of NCBI's Annotation Process

##Genome-Annotation-Data-START##
Annotation Provider :: NCBI
Annotation Status :: Full annotation

1

Protein Protein ▾ Advanced

COVID-19 Information

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#)

FASTA Send to: ▾

ribulose biphosphate carboxylase small chain, chloroplastic [Cucumis sativus]

NCBI Reference Sequence: XP_004135094.1
[GenPept](#) [Identical Proteins](#) [Graphics](#)

```
>XP_004135094.1 ribulose biphosphate carboxylase small chain, chloroplastic [Cucumis sativus]
MASSILSSAAVASVNSASPAQASMVAPFTGLKSSAGFPITRKNNDITTLASNGGKVCMKVWPPPLGLRK
FETLSYLPDMSNEQLSKECDYLLRNGWPCVEFDIGSGFVYRENHRSPGYDGRYWTMVKLPMFGCTDSS
QVIQEIEEAKKEYPDAFIRVIGFDNVRQVCISFIAYKPPRFYSS
```

2

Preparation of FASTA files

>XP_004135094

MASSILSSAAVASVNSASPAQASMVAPFTGLKSSAGFPITRKNNVDITTLASN
GGKVQCMKVWPPLGLRKFETLSYLPDMSNEQLSKECDYLLRNGWVPCVEF
DIGSGFVYRENHRSPGYDGRYWTMWKLP MFGCTDSSQVIQEIEEAKKEYP
DAFIRVIGFDNVRQVQCISFIAYKPPRFYSS

>WP_140970218

MLKKAIAEFITFVLVLFGTGVAVLGGGIEGIGTLGIAMAFGLSIVAMAYSIG
TISGCHINPAVSVAMFINKRMNAMELCYYVLAQILGGLLGTATLV TILKSAK
APLDNLGQNGFGTLGLSGAFLVEFILTFVFLVIVAVTGKKGSSSLAGLVIGFT
LVLIHLLGIPLTGTSVNPARSIPALFAGGEALSQWVVFIVAPILGGIVAAIVGK
FILNTEK

Localization sites of protein sequences

❑ Protein Sorting Signals:

- ❑ Signal Peptides (SP)
- ❑ Chloroplast transit peptides (cTP)
- ❑ Mitochondrial targeting peptides (mTP)

❑ SP has conserved three-region structure: a positively charged n-region, a hydrophobic h-region, and a polar c-region.

❑ cTP are known to be rare in acidic residues.

❑ mTP are rich in arginine (R), alanine (A), and serine (S), while negatively charged amino acid residues (aspartic acid (D) and glutamic acid (E)) are rare.

iPSORT domain address:

<https://ipsort.hgc.jp/>

iPSORT Prediction

Paste your amino acid sequence below:

*** Characters except the standard 20 codes will be removed off

*** Only the first 30 residues are used for prediction

```
MASSILSSAAVASVNSASPAQASMVAPFTGLKSSAGFPITRKNVNDITTLASNGGKVQCMKV  
WPPLGLRKFETLSYLPDMSNEQLSKECDYLLRNGWVPCVEFDIGSGFVYRENHRSPGYDGR  
YWTMWKLPFMFGCTDSSQVIQIEIEEAKKEYPDAFIRVIGFDNVRQVQCISFIAYPKPPFYSS
```



Plant Protein

Non-plant Protein

Submit

Localization sites of protein sequences

1. XP_004135094 (*Cucumis sativus* - Rubisco)

Predicted as: *having a chloroplast transit peptide*

Sequence (Type: plant)

1 PMASS ILSSA AVASV NSASP AQASM VAPFT GLKSS AGFPI TRKNN VDITT
 51 LASNG GKVQC MKVWP PLGLR KFETL SYLPD MSNEQ LSKEC DYLLR NGWVP
 101 CVEFD IGS GF VYREN HRSPG YYDGR YWTMW KLPMF GCTDS SQVIQ EIEEA
 151 KKEYP DAFIR VIGFD NVRQV QCISF IAYKP PRFYS S

Values used for reasoning

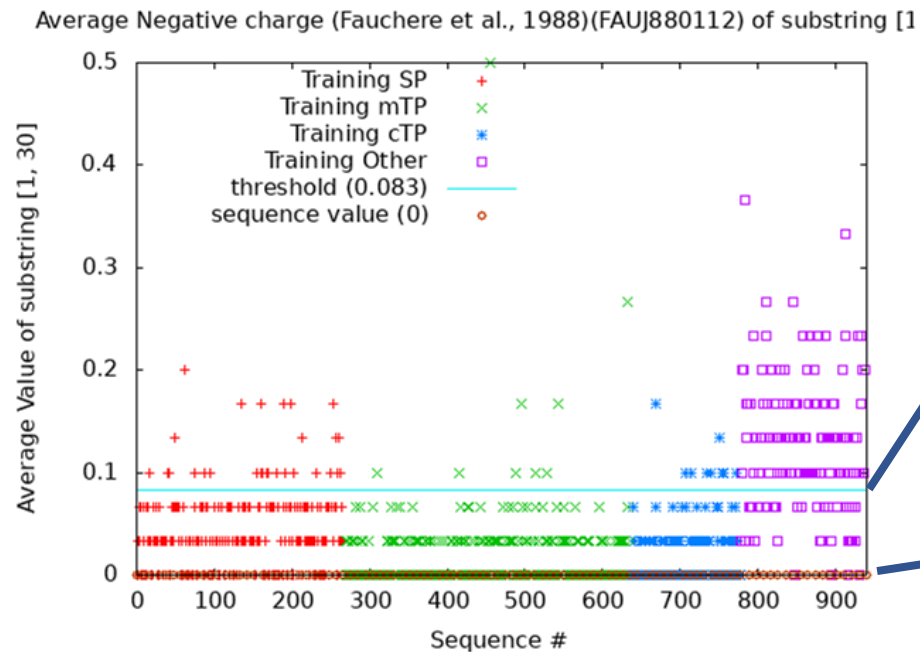
Node	Answer	View	Substring	Value(s)	Plot
1. Signal peptide?	No	Average Hydrophathy (KYTJ820101)	[6,25]	0.8 (>= 0.9225? No)	show
2. Mitochondrial or chloroplast ?	Yes	Average Negative Charge (FAUJ880112)	[1,30]	0 (< 0.083? Yes)	show
		Indexing: AI1 Pattern: 22121222 (ins/del <= 2)	[1,30]	PMAS-SILSSAAVASVNSASPAQASMVAPFT 2222-21222222222202222222222222 22121222	--
3. Mitochondrial?	No	Average Isoelectric Point (ZIMJ680104)	[1,15]	5.89067 (>= 6.21? No)	show
		Indexing: AI2 Pattern: 100100110 (ins/del <= 3)	[1,15]	PMASSILSSAAVASV 2000020000000000 NOMATCH	--

* This color means "not used".

Localization sites of protein sequences

1. XP_004135094 (*Cucumis sativus*- Rubisco)

Plot against Training Data



graph generated by [gnuplot](#)

Threshold = 0.083

Sequence Value = 0

Localization sites of protein sequences

2. WP_140970218 (*Bacillus*- Aquaporin)

Predicted as: *having a signal peptide*

Sequence (Type: nonplant)

1 MLKKA IAEFI GTFVL VLFGT GVAVL GGGIE GIGTL GIAMA FGLSI VAMAY
51 SIGTI SGCHI NPAVS VAMFI NKRMN AMELC YYVLA QILGG LLGTA TLVTI
101 LKSAK APLDN LGQNG FGTLG LSGAF LVEFI LTFVF VLVIV AVTGG KGSSS
151 LAGLV IGFTL VLIHL LGIPL TGTSV NPARS IAPAL FAGGE ALSQL WVFIV
201 APILG GIVAA IVGKF ILNTE K

Values used for reasoning

Node	Answer	View	Substring	Value(s)	Plot
1. Signal peptide?	Yes	Average Hydropathy (KYTJ820101)	[6,20]	1.96667 (>= 0.953? Yes)	show
2. Mitochondrial ?	No	Average Net Charge (KLEP840101)	[1,30]	0 (>= 0.083? No)	show
		Indexing: All Pattern: 221121122 (ins/del <= 3)	[1,30]	MLKKAIAEFIGTFVLVLFGTGVAVLGGGIE 22002120210222222020222200010 NOMATCH	--

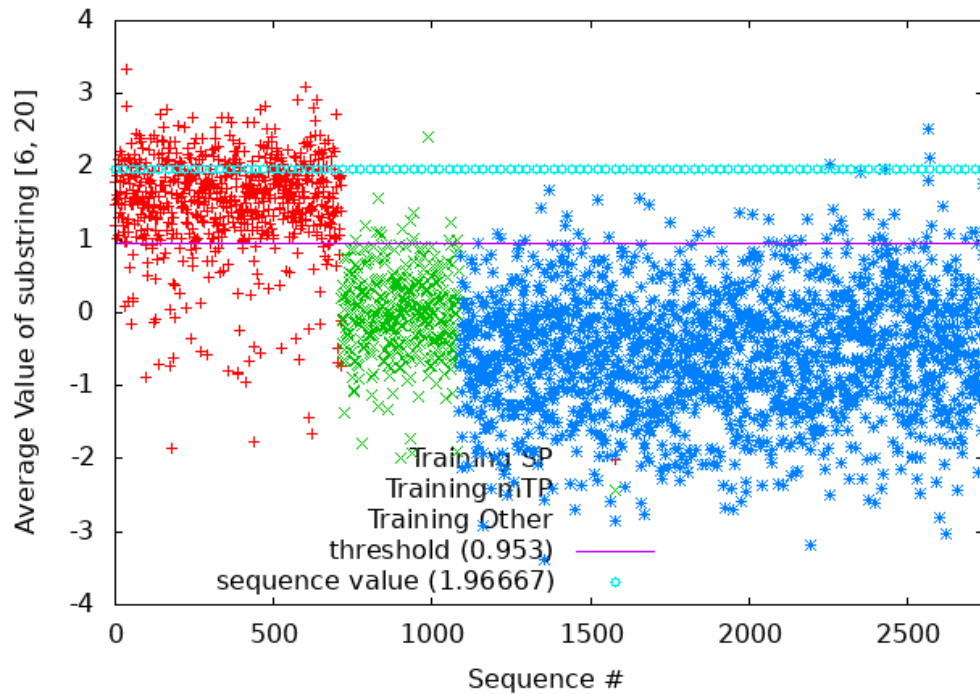
* **This color** means "not used".

Localization sites of protein sequences

2. WP_140970218 (*Bacillus*- Aquaporin)

Plot against Training Data

Average Hydropathy index (Kyte-Doolittle, 1982)(KYTJ820101) of substring [6,:



Sequence Value = 1.96667

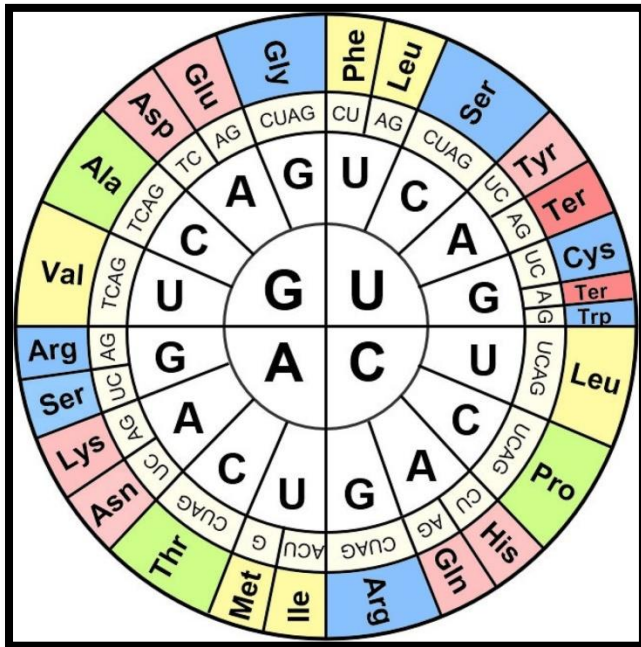
Threshold = 0.953

graph generated by [gnuplot](#)

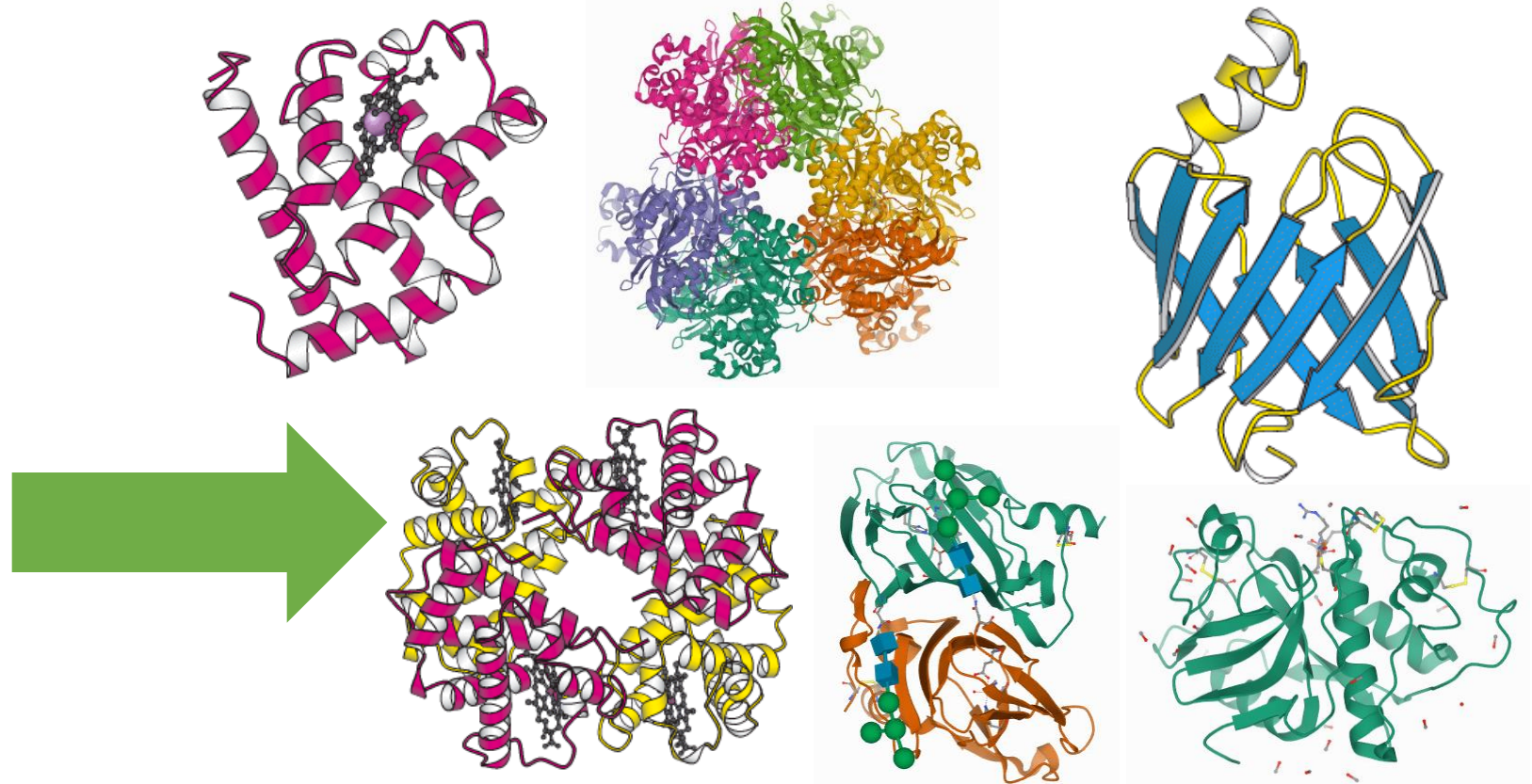
Amino acid index analysis

- **Amino Acid indices**
- **Amino Acid Index analysis tool at iPSORT WWW service**

Proteins are built from a repertoire of 20 Amino Acids



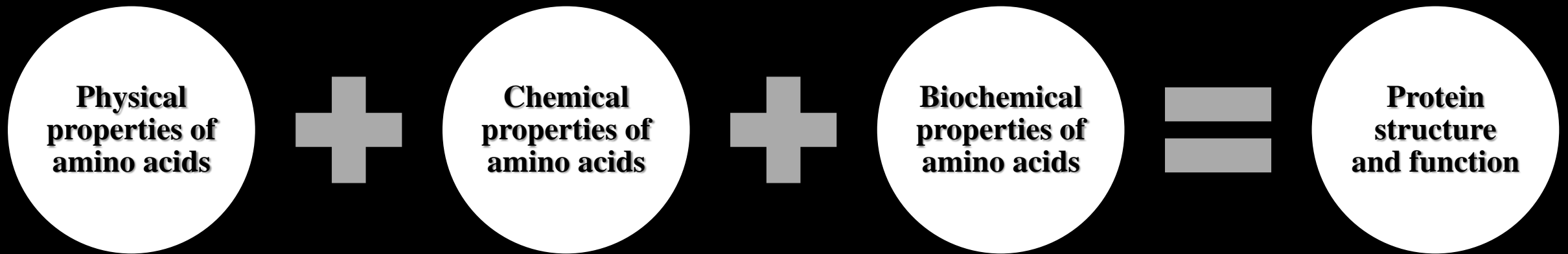
Genetic code encodes 20 basic amino acids which are the building blocks of all proteins



Three dimensional structure of Myoglobin, RuBisCo, Fatty acid binding protein , Haemoglobin tetramer, Auxin binding protein, Papain.

Source : Berg et al., 2002, www.rcsb.org

Properties of Amino Acids determine Protein structure



- **The amino acid composition determines the primary structure of the protein.**
- **The spatial arrangement of the amino acid sequence in three-dimensional space determine the secondary and tertiary structure of the protein.**
- **The constituent amino acid has multifaceted properties that are responsible for the specificity and diversity of protein structure and function.**

Source: Berg et al., 2002; Kawashima & Kanehisa. 2000

AminoAcid Index or AA Index

Examples: Isoelectric point, Hydrophobicity, Hydropathy index, Alpha helix propensity

AA INDEX

AA index is a database of amino acid indices and amino acid mutation matrices

AA INDEX I

An amino acid index is a set of 20 numerical values representing various physico-chemical and biochemical properties of amino acids.

AA INDEX II

An amino acid mutation matrix is generally 20×20 numerical values representing similarity of amino acids

Running Amino Acid Index analysis tool at iPSORT WWW Service

Welcome to the iPSORT WWW Service

This page is currently maintained by [Yoshinari Tanada at Hiroshi University](#) and [Division of Health Medical Intelligence, Human Genome Center, Institute of Medical Science, University of Tokyo](#).
Questions, comments, bug reports, etc. should be directed to: [Hideo Barnai <hdbn.dsc@tmd.ac.jp>](mailto:Hideo.Barnai@tmd.ac.jp) and [Yoshinari Tanada](#).
Best viewed with a style sheet capable browser.

A command line version of iPSORT is available [here](#). The program has been rewritten in the OCaml language, and is different from the one used for this web predictor (but should give the same predictions).

Contents

- [What is iPSORT and how does it work?](#)
- [Predict localization sites of protein sequences](#)
- [Run amino acid index analysis](#)
- [Reference](#)

iPSORT Prediction

Paste your amino acid sequence below:

*** Characters except the standard 20 codes will be removed off

*** Only the first 30 residues are used for prediction

Plant Protein

Non-plant Protein

AAindex Analysis

Choose amino acid index:

QIANBB0131 - Weights for coil at the window position of -2 (Qian-Sejnowski, 1988)
QIANBB0132 - Weights for coil at the window position of -1 (Qian-Sejnowski, 1988)
QIANBB0133 - Weights for coil at the window position of 0 (Qian-Sejnowski, 1988)
QIANBB0134 - Weights for coil at the window position of 1 (Qian-Sejnowski, 1988)
QIANBB0135 - Weights for coil at the window position of 2 (Qian-Sejnowski, 1988)
QIANBB0136 - Weights for coil at the window position of 3 (Qian-Sejnowski, 1988)
QIANBB0137 - Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)
QIANRR0138 - Weights for coil at the window position of 5 (Qian-Sejnowski, 1988)

Paste your sequences here:

Step 1 : Visit iPSORT WWW Service home page through your web browser

Welcome to the iPSORT WWW Service

iPSORT homepage URL :
<https://ipsort.hgc.jp>

This page is currently maintained by [Yoshinori Tamada](#) at [Hirosaki University](#) and [Division of Health Medical Intelligence, Human Genome Center, Institute of Medical Sciences, Hirosaki University](#).
Questions, comments, bug reports, etc. should be directed to: [Hideo Bannai <hdbn.dsc@tmd.ac.jp>](mailto:hdbn.dsc@tmd.ac.jp) and [Yoshinori Tamada](#).

Best viewed with a style sheet capable browser.

A command line version of iPSORT is available [here](#). The program has been rewritten in the OCaml language, and is different from the one used for this web predictor (but should give the same predictions).

Contents

- [What is iPSORT and how does it work?](#)
- [Predict localization sites of protein sequences](#)
- [Run amino acid index analysis](#)
- [Reference](#)

iPSORT Prediction

Paste your amino acid sequence below:

*** Characters except the standard 20 codes will be removed off

*** Only the first 30 residues are used for prediction

Step 2 : Click on the 'Run amino acid index analysis' option in 'Contents'

Welcome to the iPSORT WWW Service

This page is currently maintained by [Yoshinori Tamada at Hirosaki University](#) and [Division of Health Medical Intelligence, Human Genome Center, Institute of Medical Science, University of Tokyo](#).

Questions, comments, bug reports, etc. should be directed to: Hideo Bannai <hdbn.dsc@tmd.ac.jp> and [Yoshinori Tamada](#).

Best viewed with a style sheet capable browser.

Contents

- [What is iPSORT and how does it work?](#)
- [Predict localization sites of protein sequences](#)
- [Run amino acid index analysis](#)
- [Reference](#)

iPSORT Prediction

Paste your amino acid sequence below:

*** Characters except the standard 20 codes will be removed off

*** Only the first 30 residues are used for prediction

Step 2 : Click on the 'Run amino acid index analysis' option in 'Contents'

AAindex Analysis

Choose amino acid index:

ANDN920101 - alpha-CH chemical shifts (Andersen et al., 1992)
ARGP820101 - Hydrophobicity index (Argos et al., 1982)
ARGP820102 - Signal sequence helical potential (Argos et al., 1982)
ARGP820103 - Membrane-buried preference parameters (Argos et al., 1982)
BEGF750101 - Conformational parameter of inner helix (Beghin-Dirkx, 1975)
BEGF750102 - Conformational parameter of beta-structure (Beghin-Dirkx, 1975)
BEGF750103 - Conformational parameter of beta-turn (Beghin-Dirkx, 1975)
BHAR880101 - Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)

Paste your sequences here:

Specify Window Size: (Default: 10)

of residues to look (from N-terminal): (Default: 0 - meaning the whole sequence)

Submit

The amino acid index analysis or the AAindex Analysis tool of iPSORT

A long list of amino acid indices to select from

GUI input box to enter/paste the query sequence

GUI input box to customize the window size for analysis of query sequence

GUI input boxes to customize/specify a part of query sequence for analysis

Submit button

Step 3 : Choose amino acid index of your interest

AAindex Analysis

Choose amino acid index:



- NAGK730101 - Normalized frequency of alpha-helix (Nagano, 1973)
- NAGK730102 - Normalized frequency of beta-structure (Nagano, 1973)
- NAGK730103 - Normalized frequency of coil (Nagano, 1973)
- NAKH900101 - AA composition of total proteins (Nakashima et al., 1990)
- NAKH900102 - SD of AA composition of total proteins (Nakashima et al., 1990)
- NAKH900103 - AA composition of mt-proteins (Nakashima et al., 1990)
- NAKH900104 - Normalized composition of mt-proteins (Nakashima et al., 1990)
- NAKH900105 - AA composition of mt-proteins from animal (Nakashima et al., 1990)

The selected amino acid index highlights in blue

Step 4 : Enter/Paste query sequence

Paste your sequences here:



```
MASSILSSAAVASVNSASPAQASMVAFETGLKSSAGFPITRKNNDITTLASNGGKVOCKV  
WPPLGLRK  
FETLSYLPDMSNEQLSKECDYLLRNGWVPCVEFDIGSGFVYRENHRSPGYDGRYWTMVKLP  
MFGCTDSS  
QVIOEIEEAKKEYPDAFIRVIGFDNVROVOCISFIAYKPPRFYSS|
```

The query protein sequence is pasted in the GUI box provided

The search can be customized for window size or a part of protein

Specify Window Size: (Default: 10)

of residues to look (from N-terminal): (Default: 0 - meaning the whole sequence)

Submit

If not stated, the default window size of 10 is selected and entire sequence is analysed

Step 5 : Click on 'Submit' Button

Paste your sequences here:

```
MASSILSSAAVASVNSASPAQASMVAFETGLKSSAGFPITRKNNVDIITLASNGGKVOCKMV  
WPPLGLRK  
FETLSYLPDMSNEQLSKECDYLLRNGWVPCVEFDIGSGFVYRENHRSPGYDGRYWTMWKLP  
MFGCTDSS  
QVIOEIEEAKKEYPDAFIRVIGFDNVROVOCISFIAYKPPRFYSS|
```

Specify Window Size: (Default: 10)

of residues to look (from N-terminal): (Default: 0 - meaning the whole sequence)

Submit



Example 1

AAindex Analysis

Choose amino acid index:

- NAGK730101 - Normalized frequency of alpha-helix (Nagano, 1973)
- NAGK730102 - Normalized frequency of beta-structure (Nagano, 1973)
- NAGK730103 - Normalized frequency of coil (Nagano, 1973)
- NAKH900101 - AA composition of total proteins (Nakashima et al., 1990)
- NAKH900102 - SD of AA composition of total proteins (Nakashima et al., 1990)
- NAKH900103 - AA composition of mt-proteins (Nakashima et al., 1990)
- NAKH900104 - Normalized composition of mt-proteins (Nakashima et al., 1990)
- NAKH900105 - AA composition of mt-proteins from animal (Nakashima et al., 1990)

Paste your sequences here:

```
>XP_004135094.1 ribulose bisphosphate carboxylase small chain,  
chloroplastic [Cucumis sativus]  
MASSILSSAAVASVNSASPAQASMVAPFTGLKSSAGFPITRKNVVDITLASNGGKVOCKMV  
WPPLGLRK  
FETLSYLPDMSNEQLSKECDYLLRNGWVPCVEFDIGSGFVYRENHRSPGYDGRYWTMWKLP  
MFGCTDSS  
QVIQEIIEEAKKEYPDAFIRVIGFDNVRQVQCISFIAYKPPRFYSS
```

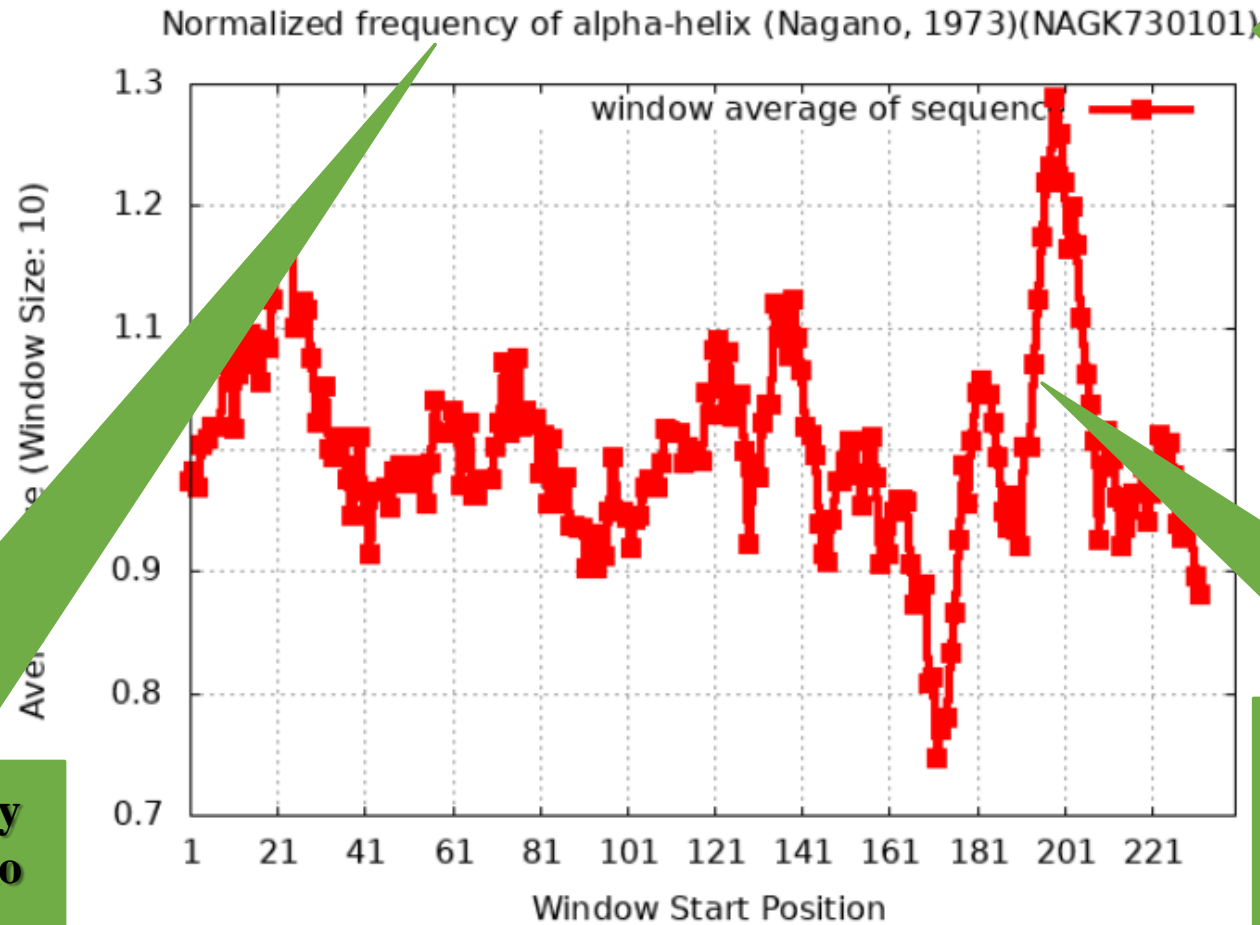
Specify Window Size: (Default: 10)

of residues to look (from N-terminal): (Default: 0 - meaning the whole sequence)



Example 1

AAIndex Analysis



It is the overall frequency calculated for each amino acid to occur in a alpha helix

The output graph shows the plot of average values of normalized alpha helix frequency for amino acids in selected window size

Example 2

AAindex Analysis

Choose amino acid index:

- ZASB820101 - Dependence of partition coefficient on ionic strength (Zaslavsky et al., 1982)
- ZIMJ680101 - Hydrophobicity (Zimmerman et al., 1968)
- ZIMJ680102 - Bulkiness (Zimmerman et al., 1968)
- ZIMJ680103 - Polarity (Zimmerman et al., 1968)
- ZIMJ680104 - Isoelectric point (Zimmerman et al., 1968)**
- ZIMJ680105 - RF rank (Zimmerman et al., 1968)
- AURR980101 - Normalized positional residue frequency at helix termini N4' (Aurora and Rose, 1998)
- AURR980102 - Normalized positional residue frequency at helix termini N''' (Aurora and Rose, 1998)

Paste your sequences here:

```
>XP_004135094.1 ribulose biphosphate carboxylase small chain,  
chloroplastic [Cucumis sativus]  
MASSILSSAAVASVNSASPAQASMVAPFTGLKSSAGFPITRKNNVDITILASNGGKVOCMKV  
WPPLGLRK  
FEILSYLPDMSNEQLSKECDYLLRNGWVPCVEFDIGSGFVYRENHRSPGYDGRYWTMWKLP  
MFGCTDSS  
OVIOEIEEAKKEYPDAFIRVIGFDNVROVOCISFIAYKPPRFYSS
```

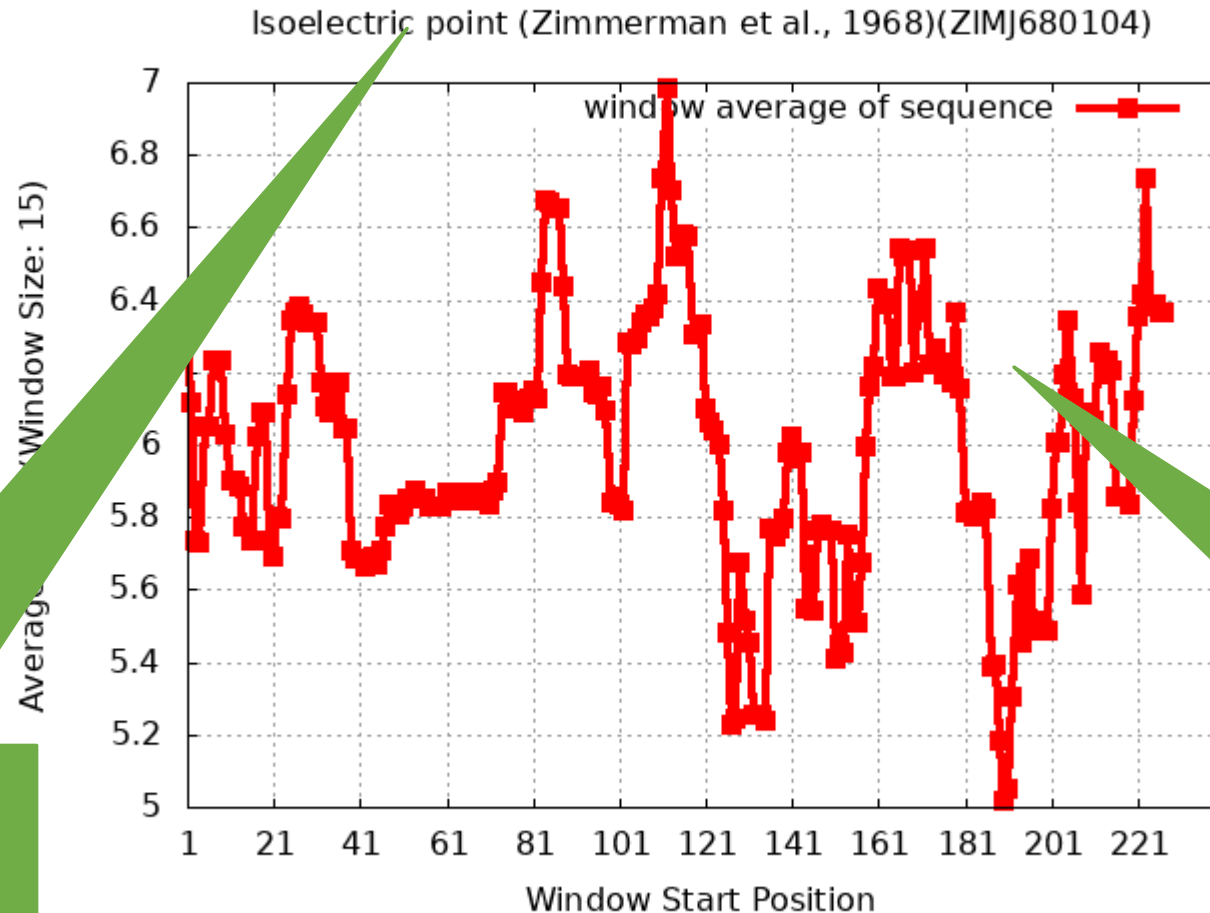
Specify Window Size: (Default: 10)

of residues to look (from N-terminal): (Default: 0 - meaning the whole sequence)



Example 2

AAIndex Analysis



The isoelectric point of an amino acid is the point at which the amino acid has no net electrical charge.

The output graph shows the plot of average values of isoelectric points for amino acids in selected window size

Exploring AAindex database further

Paste your sequences here:

```
>XP_004135094.1 ribulose biphosphate carboxylase small chain,  
chloroplastic [Cucumis sativus]  
MASSILSSAAVASVNSASPAQASMVAPFTGLKSSAGFPITRKNNVDITLASNGGKVCQMKV  
WFPLGLRK  
FETLSYLPDMSNEQLSKECDYLLRNGWVPCVEFDIGSGFVYRENHRSPGYDGRYWTMWKLP  
MFGCTDSS  
QVIQEIFEAKKEYPDAFIRVIGFDNVROVOCISFIAVKPRREYSS
```

Specify Window Size: (Default: 10)

of residues to look (from N-terminal): (Default: 0 - meaning the whole sequence)

Reference

- Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S., "Extensive feature detection of N-terminal protein sorting signal", *Bioinformatics*, 18(2) 298-305, 2002. (My version that I'm allowed to post, is [here](#). The copyright has been transferred to [Oxford University Press](#).) [[Pubmed](#)]
- Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S., *Views: Fundamental Building Blocks in the Process of Knowledge Discovery*, In Proceedings of the 14th International [FLAIRS](#) Conference, 233-238, AAAI Press, 2001. ([postscript](#), [gzipped postscript](#), [PDF](#)) (C) [AAAI](#)

Links

- [PSORT www service \(PSORT, PSORT II\)](#): The original PSORT and PSORT II.
- [AAindex Database](#): A database of amino acid indices.

Exploring AAindex database further



AAindex

Amino acid indices, substitution matrices and pair-wise contact potentials

AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids. AAindex consists of three sections now: AAindex1 for the amino acid index of 20 numerical values, AAindex2 for the amino acid mutation matrix and AAindex3 for the statistical protein contact potentials. All data are derived from published literature.

Search or Download

Search by DBGET bfind for

(e.g. Hydrophobicity, MIRL960101)

- [AAindex Help](#)
- [Search AAindex by DBGET/LinkDB](#) [\[DBGET Help\]](#)
- [Download by anonymous FTP](#)
- [Download by HTTPS](#)

Figures and Lists

- Lists of the current database entries

Exploring AAindex database further

AAindex: Amino Acid Index Database

Release 9.2, Feb 2017

Introduction

An amino acid index is a set of 20 numerical values representing any of the different physicochemical and biological properties of amino acids. The AAindex1 section of the Amino Acid Index Database is a collection of published indices together with the result of cluster analysis using the correlation coefficient as the distance between two indices. This section currently contains 566 indices.

Another important feature of amino acids that can be represented numerically is the similarity between amino acids. Thus, a similarity matrix, also called a mutation matrix, is a set of 210 numerical values, 20 diagonal and $20 \times 19/2$ off-diagonal elements, used for sequence alignments and similarity searches. The AAindex2 section of the Amino Acid Index Database is a collection of published amino acid mutation matrices together with the result of cluster analysis. This section currently contains 94 matrices.

In the release 9.0, we added a collection of published protein pairwise contact potentials to AAindex as AAindex3. This section currently contains 47 contact potential matrices.

How to search AAindex

AAindex entries can be retrieved by keyword search against the description lines of AAindex (D line in each entry) using DBGET. For example, the user can retrieve a list of entries to input keywords such as 'alpha-helix' or 'hydrophobicity' on the text-box of AAindex home page or the DBGET search page. In the DBGET system, the "bfind" mode is used to get the list of entries including the specified keywords and the "bget" mode is used to get an database entry by specifying an entry ID.

AAindex Home <https://www.genome.jp/aaindex/>

AAindex search by DBGET https://www.genome.jp/dbget-bin/www_bfind?aaindex

Exploring AAindex database further

List of 566 Amino Acid Indices in AAindex ver.9.2

The columns correspond to the AAindex accession number and the description of each index.

ANDN920101 alpha-CH chemical shifts (Andersen et al., 1992)

ARGP820101 Hydrophobicity index (Argos et al., 1982)

ARGP820102 Signal sequence helical potential (Argos et al., 1982)

BEGF750101 Conformational parameter of inner helix (Beghin-Dirkx, 1975)

BEGF750102 Conformational parameter of beta-structure (Beghin-Dirkx, 1975)

BEGF750103 Conformational parameter of beta-turn (Beghin-Dirkx, 1975)

BHAR880101 Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)

BIGC670101 Residue volume (Bigelow, 1967)

BIOV880101 Information value for accessibility; average fraction 35% (Biou et al., 1988)

BIOV880102 Information value for accessibility; average fraction 23% (Biou et al., 1988)

BROC820101 Retention coefficient in TFA (Browne et al., 1982)

BROC820102 Retention coefficient in HFBA (Browne et al., 1982)

BULH740101 Transfer free energy to surface (Bull-Breese, 1974)

BULH740102 Apparent partial specific volume (Bull-Breese, 1974)

BUNA790101 alpha-NH chemical shifts (Bundi-Wuthrich, 1979)

BUNA790102 alpha-CH chemical shifts (Bundi-Wuthrich, 1979)

BUNA790103 Spin-spin coupling constants $3J_{\text{H}\alpha\text{-NH}}$ (Bundi-Wuthrich, 1979)

BURA740101 Normalized frequency of alpha-helix (Burgess et al., 1974)

BURA740102 Normalized frequency of extended structure (Burgess et al., 1974)

CHAM810101 Steric parameter (Charton, 1981)

CHAM820101 Polarizability parameter (Charton-Charton, 1982)

CHAM820102 Free energy of solution in water, kcal/mole (Charton-Charton, 1982)

CHAM830101 The Chou-Fasman parameter of the coil conformation (Charton-Charton, 1983)

CHAM830102 A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton, 1983)

CHAM830103 The number of atoms in the side chain labelled 1+1 (Charton-Charton, 1983)

CHAM830104 The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)

CHAM830105 The number of atoms in the side chain labelled 3+1 (Charton-Charton, 1983)

CHAM830106 The number of bonds in the longest chain (Charton-Charton, 1983)

CHAM830107 A parameter of charge transfer capability (Charton-Charton, 1983)

CHAM830108 A parameter of charge transfer donor capability (Charton-Charton, 1983)

CHOC750101 Average volume of buried residue (Chothia, 1975)

CHOC760101 Residue accessible surface area in tripeptide (Chothia, 1976)

CHOC760102 Residue accessible surface area in folded protein (Chothia, 1976)

CHOC760103 Proportion of residues 95% buried (Chothia, 1976)



PrediSi

PREDiction of Signal peptides



PrediSi

PREDiction of Signal peptides

- **PrediSi** allows the evaluation of whole proteome datasets i.e. analysis of large datasets in real time with high accuracy.
- The method employed is based on a **position weight matrix** approach.
- A **freely available**, user-friendly web interface tool (<http://www.predisi.de>)
- Presentation of the **results** in **user-** as well as **computer-friendly** formats such as HTML, XML and CSV.
- Free availability as a **Java package** for integration into other software projects.

(a)

```
CCCATTGTTCTC
TTTCTGGTTCTC
TCAATTGTTTAG
CTCATTGTTGTC
TCCATTGTTCTC
CCTATTGTTCTC
TCCATTGTTTCGT
CCAATTGTTTTG
TCAATTGTTCTC
CCCATTGTTTAG
```



(b)

A	0.0	0.0	0.3	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
C	0.5	0.8	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.6
G	0.0	0.0	0.0	0.0	0.0	0.1	1.0	0.0	0.0	0.1	0.1	0.3
T	0.5	0.2	0.2	0.0	1.0	0.9	0.0	1.0	1.0	0.3	0.7	0.1

Steps to follow

<http://www.predisi.de/>



PrediSi

PREDIction of Signal peptides

Home

Submission

Download

Literature

Links

NEW Detailed graphical information about submitted sequences are now available. Just click [Details](#) after submitting your request.

1. Sequence submission

- Type/paste sequences below:
Amino acid sequences must be in FASTA-format

```
>WP_140970218.1 aquaporin [Bacillus  
sp. (in: Bacteria)]  
MLKKAIAEFIGTFVLVLFGTGVAVLGGGIEGIGTLGIA  
MAFGLSIVAMAYSIGTISGCHINPAVSVAMFI  
NKRMNAMELCYYVLAQILGGLLGATLVITILKSAKAPL  
DNLGQNGFGTLGLSGAFLVEFILTFFVFLVIV
```

- Submit local FASTA-file:

No file chosen

1

2. Truncation

Truncate sequences to residues for faster calculation

3. Organism group

- Eukaryotic
 Gram-negative bacteria

Other bioinformatic
tools from our team:

JVirGel

PRODORIC Release 2

Steps to follow

2. Truncation

Truncate sequences to residues for faster calculation

3. Organism group

- Eukaryotic
- Gram-negative bacteria
- Gram-positive bacteria

4. Output format

- HTML
- MS-EXCEL
- Text
- XML
- CSV
- Get results as compressed download

Delimiter:

Recently we observed some problems analyzing submitted FASTA files. The error is fixed now. However, if your request fails feel free to contact [us](#)

Submit

2

Settings:

Matrix:	null
Truncation:	70 residues
Output:	null

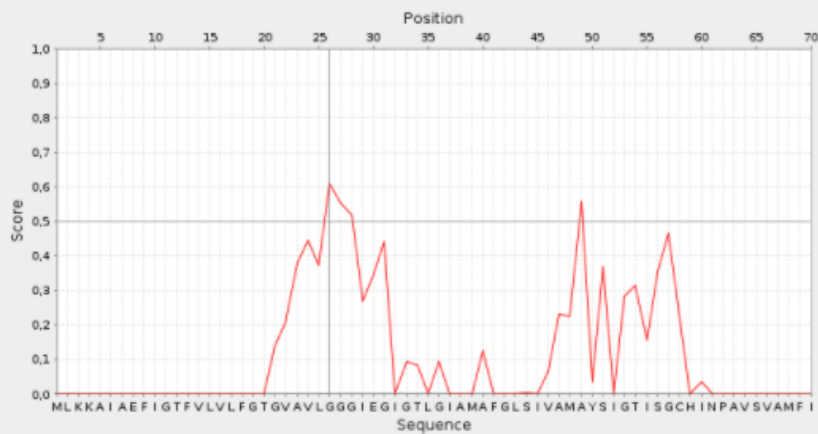
Result

FASTA-ID	Score	Cleavage Position	Signal Peptide ?	Chart
WP_140970218.1 aquaporin [Bacillus sp. (in: Bacteria)]	0.6092	26	Y	Details

Details:

Matrix:	Eukarya
Truncation:	70 residues
Cleavage position:	26
Score:	0.6092
Secreted protein:	predicted for secretion

PredISI - Signal Peptide Prediction



WP_140970218.1 aquaporin [Bacillus sp. (in: Bacteria)]:

MLKKKIAEFIGTFVLLVLFGTGVAVLGGGIEGIGTLGIAMAFGLSIVAMAYSIGTISGCHINPAVSVAMFI
1 10 20 30 40 50 60 70

Detailed Results

Bibliography



PrediSi PREDiction of Signal peptides

Home

Submission

Download

Literature

Links

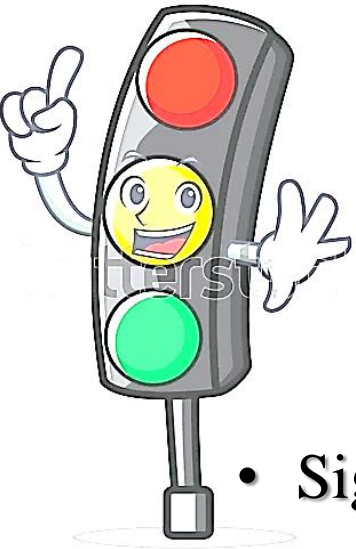
Literature:

- von Heijne, G. (1985) **Signal sequences. The limits of variation.** *Journal of molecular biology* **184** , 99-105 [\[PUBMED\]](#)
- Rapoport, TA., Jungnickel, B. & Kutay, U. (1996) **Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes.** *Annu. Rev. Biochem.* **61**, 271-303 [\[PUBMED\]](#)
- Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997) **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng.* **10**, 1-6 [\[PUBMED\]](#)
- Zheng, N. and Gierasch, L.M. (1996) **Signal sequences: the same yet different.** *Cell* **86**, 849-852 [\[PUBMED\]](#)
- Rapoport T.A., Jungnickel B. and Kutay U. (1996) **Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes.** *Annual review of biochemistry* **65**, 271-303 [\[PUBMED\]](#)
- Nielsen, H., Engelbrecht, J., von Heijne, G. and Brunak, S. (1996) **Defining a Similarity Threshold for a Functional Protein Sequence Pattern: The Signal Peptide Cleavage Site.** *Proteins* **24** , 165-177 [\[PUBMED\]](#)
- Menne, K. M. L., Hermjakob, H. and Apweiler, R. (2000) **A comparison of signal sequence prediction methods using a test set of signal peptides.** *Bioinformatics* **16**, 741-742 [\[PUBMED\]](#)
- Schreiber M. and Brown C. (2002) **Compensation for nucleotide bias in a genome by representation as a discrete channel with noise.** *Bioinformatics* **18** , 507-512 [\[PUBMED\]](#)
- Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S. and Schneider M. (2003) **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Research.* **31**, 365- 370 [\[PUBMED\]](#)
- Hiller, K., Schobert, M., Hundertmark, C., Jahn, D. & Münch, R. (2003) **JVirGel: calculation of virtual two-dimensional protein gels.** *Nucleic Acids Res.* **31**, 3862-3865 [\[PUBMED\]](#)

Other bioinformatic
tools from our team:

JVirGel

PRODORIC Release 2



Algorithms for Prediction Softwares

- Signal peptides serve as “address tags”. It is highly desirable to develop the fast and accurate algorithms to identify the signal sequences and predict their cleavage sites. Based on different kinds of characteristics, several algorithms such as neural networks, and **Hidden Markov Models** have been used.
- **Bayesian network** is a method of statistical inference in which some kind of evidence or observations are used to calculate the probability if a hypothesis may be true.

Hidden Markov Model & Bayes Theorem

Life of a science student



**Experiment
worked!!**



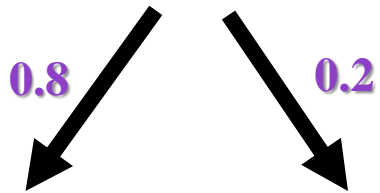
**Experiment
failed!!**



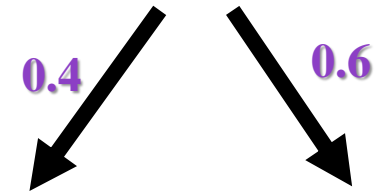
Exceptions



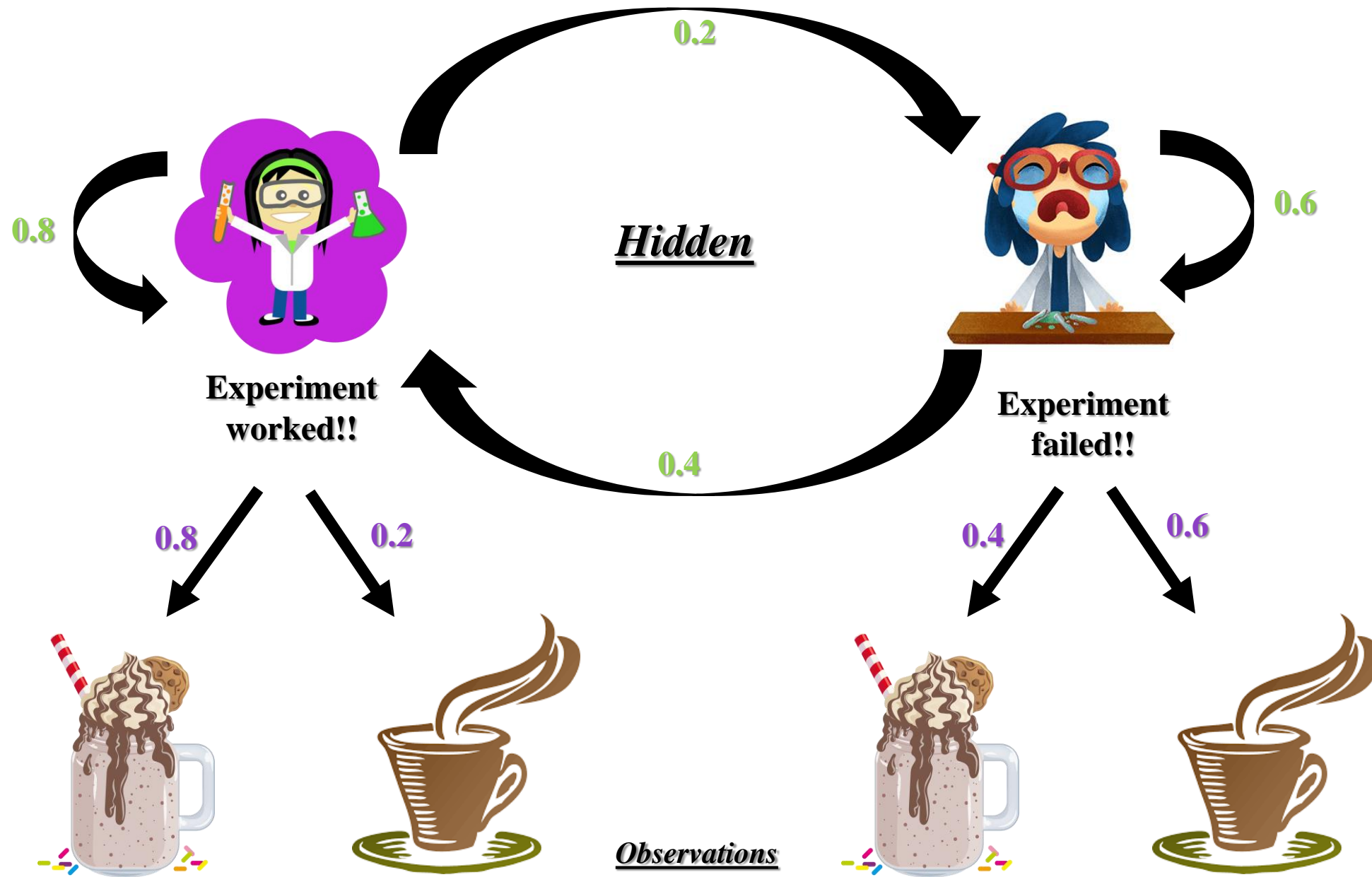
**Experiment
worked!!**



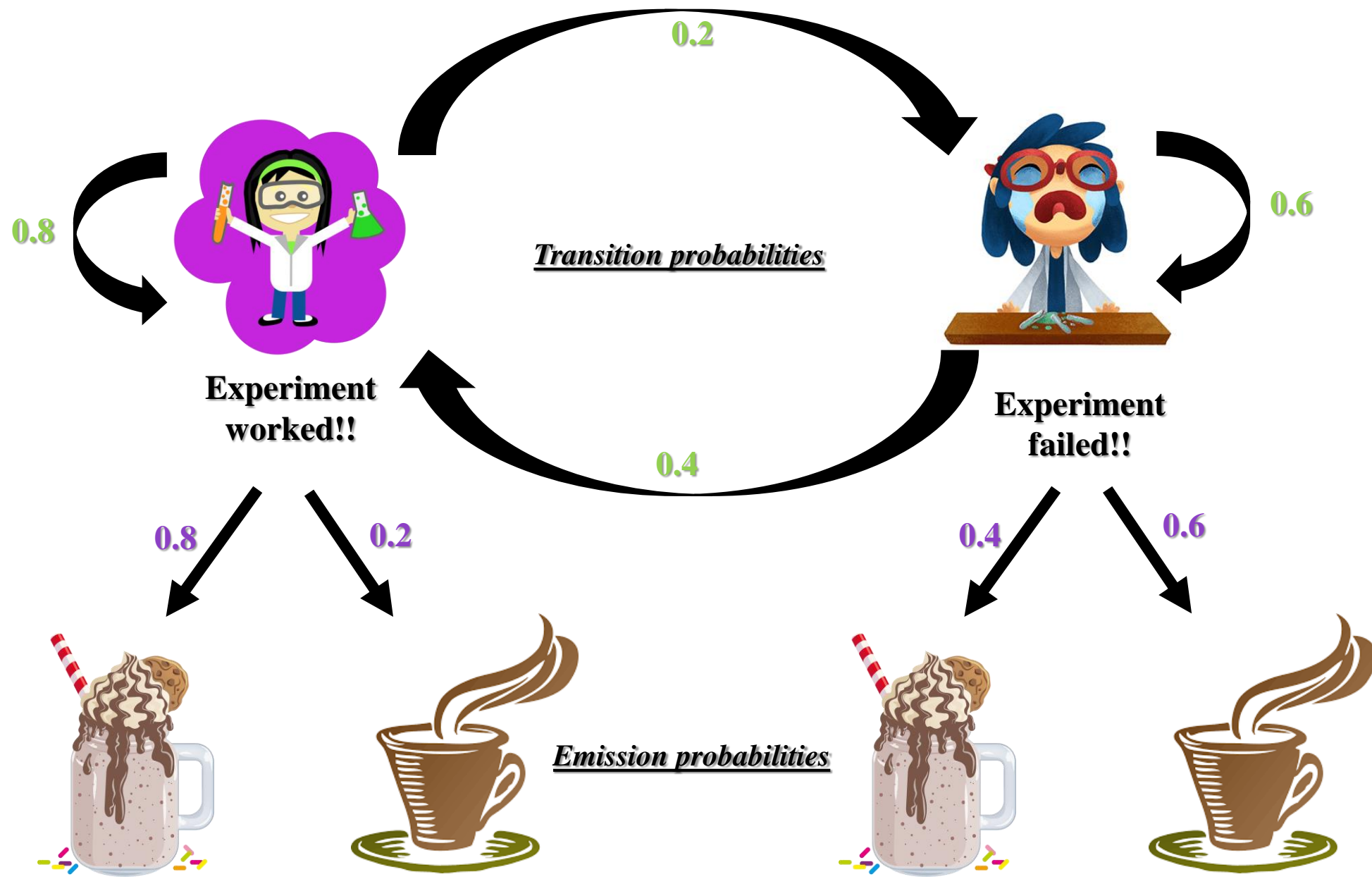
**Experiment
failed!!**



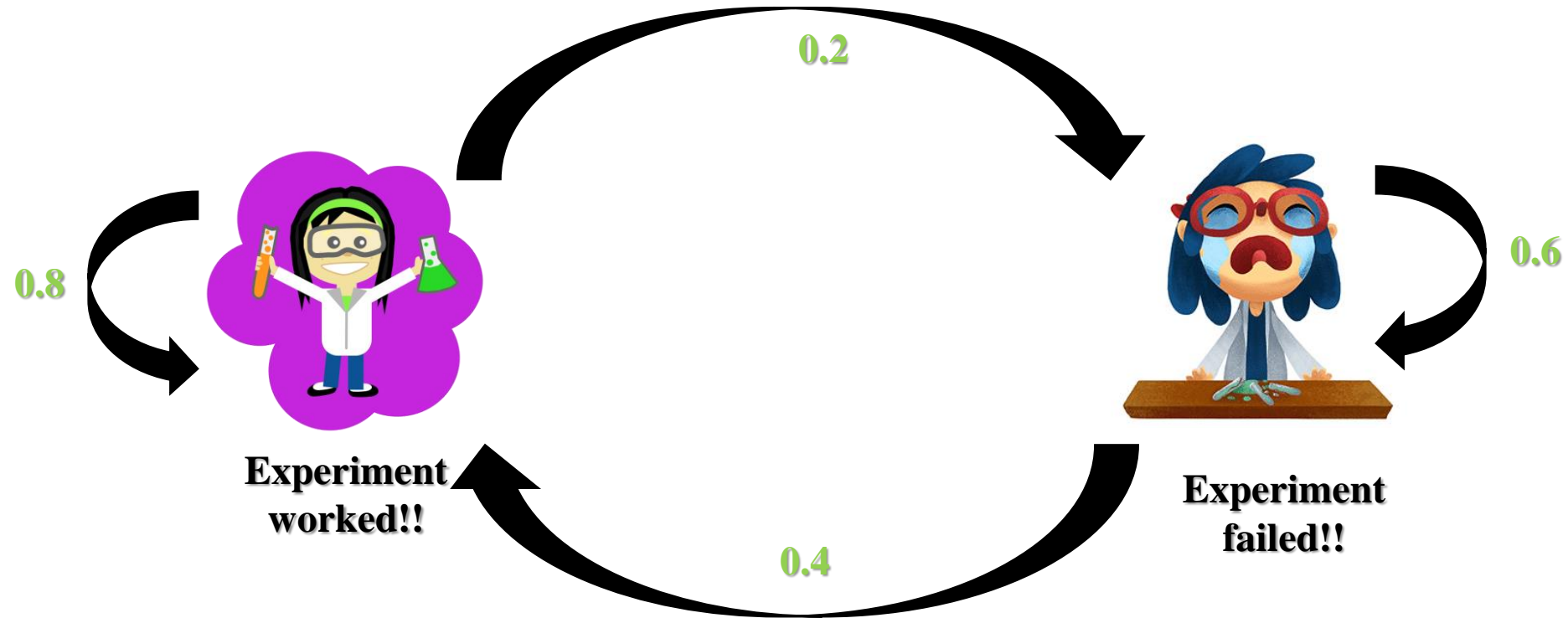
Hidden Markov Model



Hidden Markov Model



Probability of an experiment working or failing on a Random day?

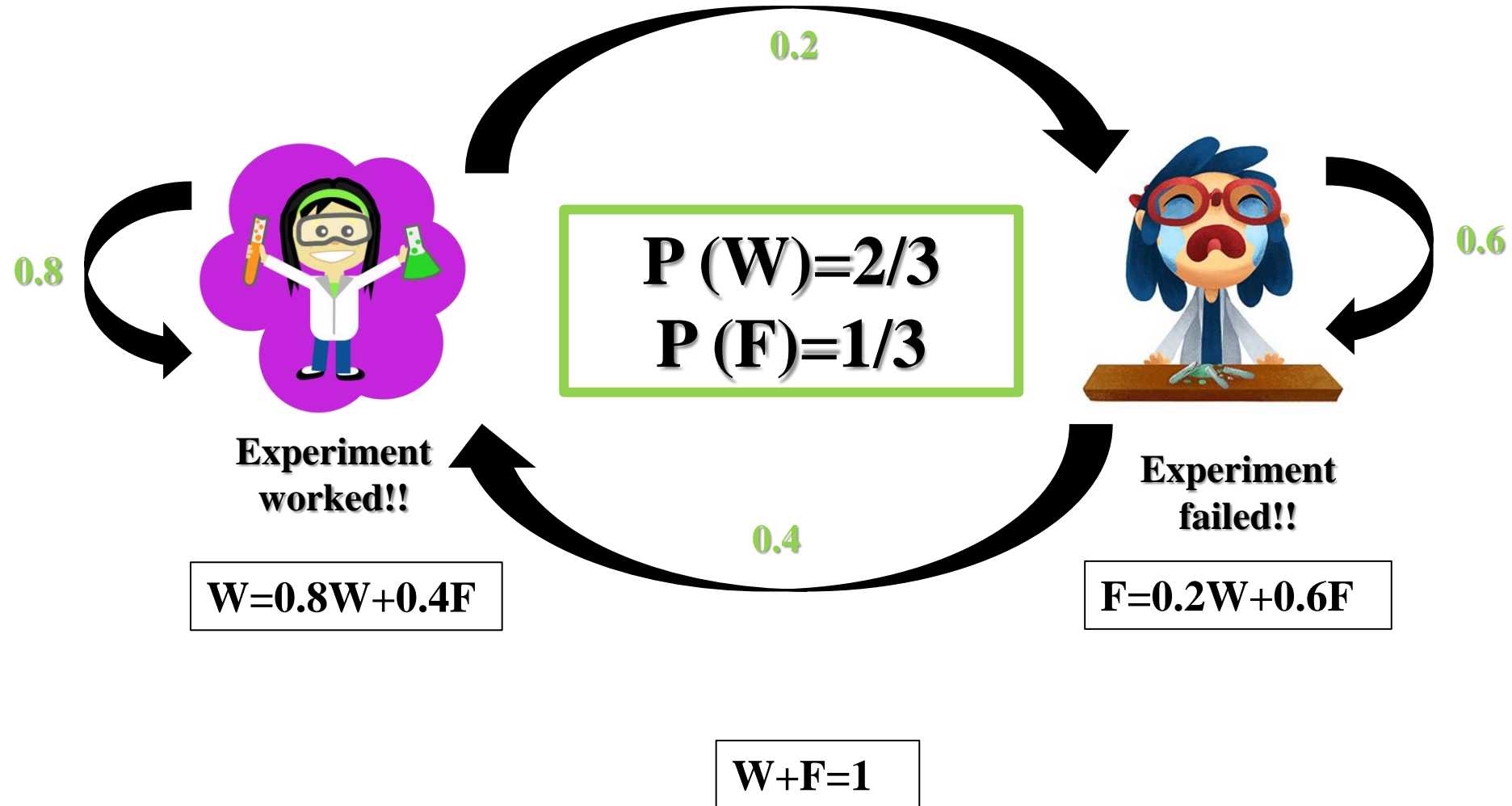


$$W = 0.8W + 0.4F$$

$$F = 0.2W + 0.6F$$

$$W + F = 1$$

Probability of an experiment failing or working on a Random day?



Probability of an experiment failing or working if I ordered Oreo shake?



**Experiment
worked!!**



**Experiment
failed!!**

$$P(W) = 2/3$$

$$P(F) = 1/3$$

**Prior
Probability**

Bayes Theorem

$$P(W)=2/3$$

Prior Probability

$$P(F)=1/3$$



0.8

0.2

Bayes Theorem

$P(W)=2/3$

Prior Probability

$P(F)=1/3$



0.8

0.4

0.2

0.6

Bayes Theorem

$$P(W)=2/3$$

Prior Probability

$$P(F)=1/3$$



0.8

0.4

0.2

0.6

If OS,

$$P(W|OS) = 8/10$$

$$P(F|OS) = 2/10$$



Posterior Probabilities

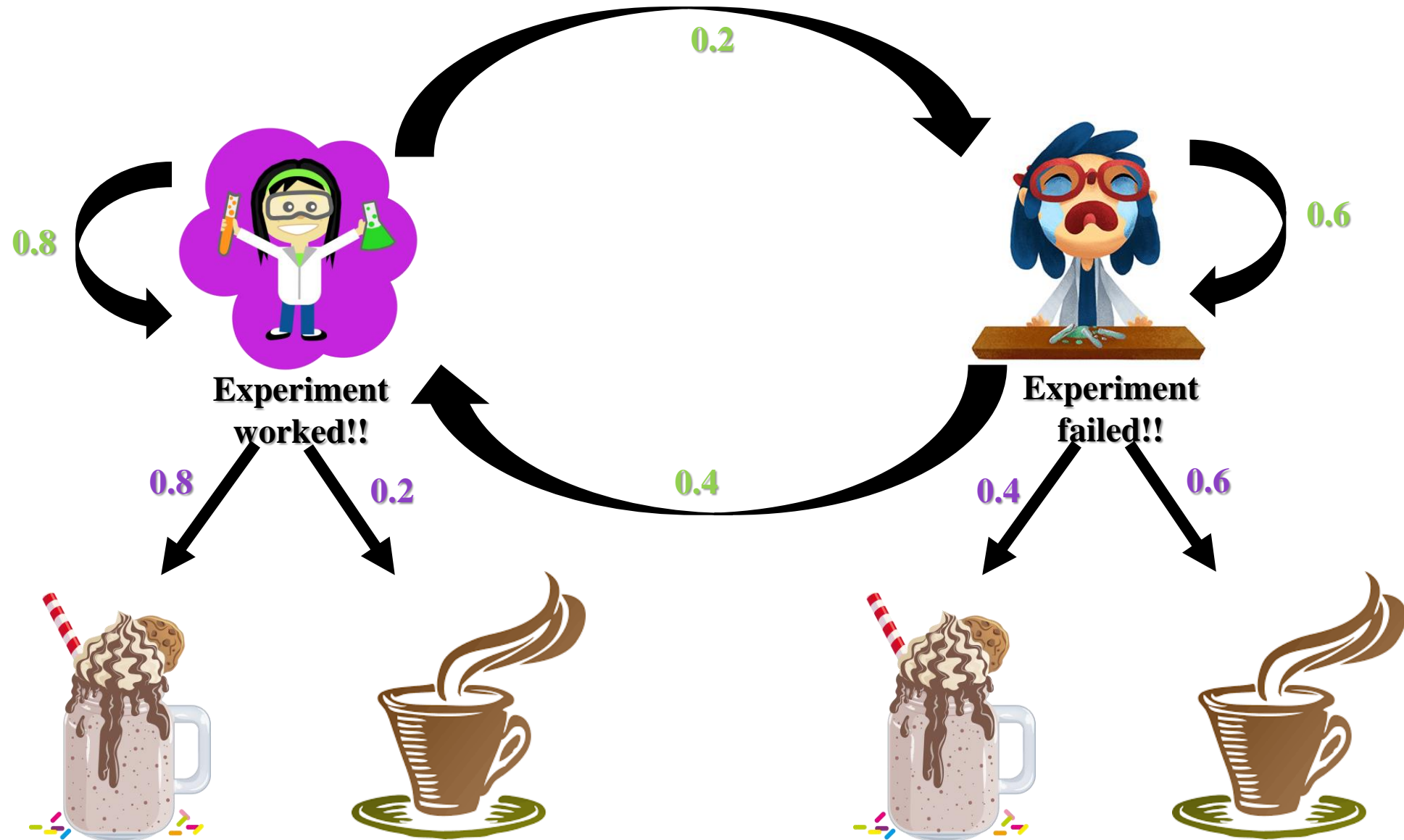
If C,

$$P(W|C) = 2/5$$

$$P(F|C) = 3/5$$



Maximum Likelihood- Home Exercise



If for 2 consecutive days I ordered Oreo shake, what was the status of my experiments?

Hint

No. of possibilities = 2^n where $n=2$ in this case

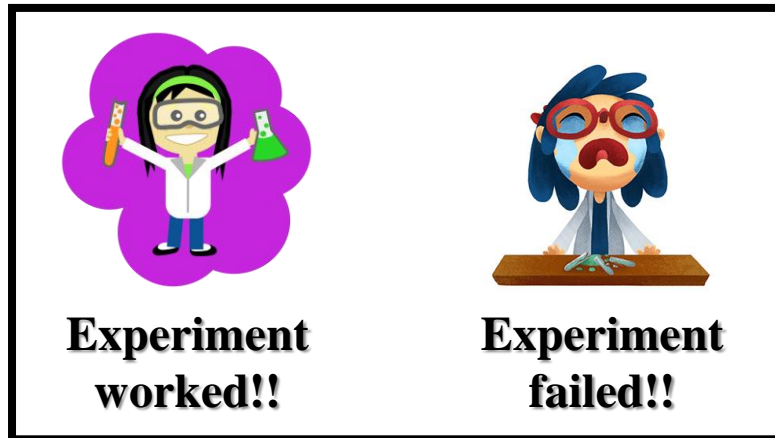
I



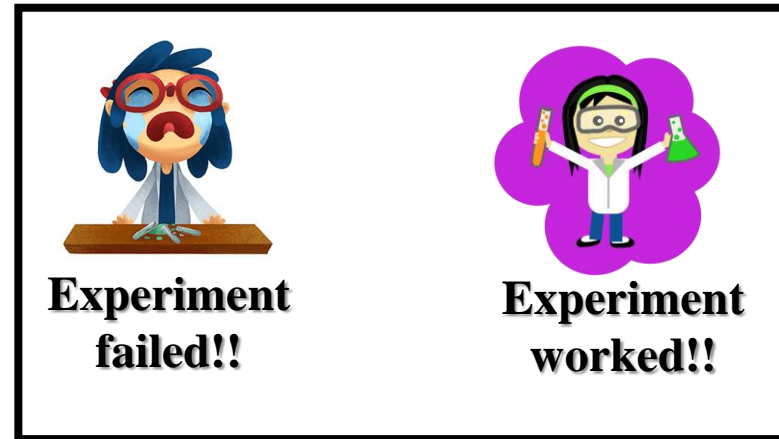
II



III



IV



Acknowledgements



Students of B. Sc (H) Botany, KMC, University of Delhi

Student Online Workshop on

Application of Bioinformatics for Pedagogy of Plant Sciences



Organized by Kirori Mal College

(Under DBT Star Scheme)

23rd July 2021

***Plant Cell Biology 101 meets
Bioinformatics***