

Bioinformatic Web-Tools: The Basics

sol genomics network

presented by Suzy Strickler Mueller Lab Rm 217

Slides can be found here: <u>ftp://ftp.solgenomics.net/</u> <u>bioinfo_class/interns/2015</u>





Part I: Biological Databases.



Source: Contributing Organizations at GMOD



Biological Databases:

I - Types.

2-Public Repositories.

3-Community specific databases.3.1- For species.3.2- For specific datatypes.

4- Genomic Browsers.



There is 3 types of biological databases (Rhee SY. et al. 2006):

• Public repositories with massive data storage.

• Community-specific databases.

• Project-specific databases.



- * Public repositories.
 - Maintained by public agencies or public international consortiums.
 - Massive data amounts (quantity).
 - No curated or poorly curated data.
 - Long term data storage.
 - Examples: GenBank, Uniprot, Gene Expression Omnibus (GEO), ArrayExpress.



- * Community-specific databases.
 - Maintained by scientific groups, frequently associated with an specific project or a research line.
 - Considerable data amount related with the community needs.
 - Curated or highly curated data (quality).
 - Long term data storage



* Community-specific databases.

- Different types:
 - Model Organism Databases (MODs): TAIR, MaizeGDB
 - Clade Organism Databases (CODs): Sol Genomics Network (SGN), Gramene.
 - Metabolic Pathways Databases: MetaCyc, KEGG pathways.
 - Specific Vocabulary (Ontologies) Databases: Gene Ontology, Plant Ontology.



- * Project specific databases.
 - Maintained by a group or a small consortium
 - Low data amount.
 - Variability for data curation (from poorly to highly).
 - Limited lifespan generally associated with a project.
 - Examples: Plant Genome Network (PGN)



Biological Databases:

- I-Types.
- 2-Public Repositories.

3-Community specific databases.3.1- For species.3.2- For specific datatypes.

4- Genomic Browsers.

2. Public Repositories.



NCBI (National Center for Biotechnology Information) http://www.ncbi.nlm.nih.gov/

S NCBI Resources 🖸 How To		Sign in to NCBI
All Data	abase:	Search
NCBI Home Resource List (A-Z)	Welcome to NCBI The National Center for Biotechnology Information advances science and health by providing access to biomedical and eccemic information	Popular Resources PubMed
All Resources Chemicals & Bioassays	About the NCBI Mission Organization Research NCBI News	PubMed Central PubMed Health
Data & Software DNA & RNA Domoins & Structures	Get Started	BLAST Nucleotide
Genes & Expression Genetics & Medicine	<u>Tools</u> : Analyze data using NCBI software <u>Downloads</u> : Get NCBI data or software How-To's: Learn how to accomplish specific tasks at NCBI	Genome SNP
Genomes & Maps Homology	Submissions: Submit data to GenBank or other NCBI databases	Gene Protein
Literature Proteins	NCBI Twitter feed	PubChem
Sequence Analysis Taxonomy	Keep up-to-date on data updates, resource announcements, and other information about what is going on at the NCBI.	NCBI Announcements New RefSeq Bacterial Protein Products and Emerging RefSeg Data Model
Training & Tutorials Variation	II 1 2 3 4 5 6 7 8	Jun 11, 2013 The NCBI Reference Sequence Project



NCBI (National Center for Biotechnology Information) http://www.ncbi.nlm.nih.gov/

Highlights:

- GenBank.
- PubMed.
- Gene Expression Omnibus (GEO)
- Taxonomy



GenBank, NIH database for sequences, an annotated collection of ALL publicly available DNA sequences (Benson DA. *et al.* 2011).

http://www.ncbi.nlm.nih.gov/genbank/

http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide

S NCBI Resources 🖂 How	ro 🖂		My NCBI Sign In
Nucleotide	Search: Nucleotide	Limits Advanced search Help	
Alphabet of Life		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	Nucleotide Tools	Other Resources
Quick Start Guide	Submit to GenBank	GenBank Home
FAQ	LinkOut	RefSeq Home
Help	E-Utilities	Gene Home
GenBank FTP	BLAST	SRA Home
RefSeq FTP	Batch Entrez	INSDC



GenBank:	Search Sect	ion
S NCBI Resources V How To V		My NCBI Sign In
Alphabet of Life Search: Nucleotide Sa	ave search Limits Advanced search Help Search Clear	
Display Settings: 🕞 Summary, 20 per page, Sorted by Default order	Send to: 🖂	Filter your results:
Found 770004 nucleotide sequences. Nucleotide (15269) EST (754652) GSS (8)	83)	All (15269)
 Results: 1 to 20 of 15269 Arabidopsis thaliana chromosome 1, complete sequence 30,427,671 bp linear DNA Accession: CP002684.1 GI: 332189094 GenBank FASTA Graphics Gossypium hirsutum mitogen-activated protein kinase (MAPK) gene, protein kinase (MA	<pre><< First < Prev Page 1 of 764 Next > Last >> </pre>	INSDE (GenBank) (14860) mRNA (9150) RefSeq (401) Tense Fitters • Cop Organisms [ree] Populus tremula x Populus alba (7835) Orga sativa (2155) Orga sativa Indica Group (1366) Pinus taeda (577) Orga sativa Japonica Group (524) Al other taxa (4292) More
Sections	Sequence Type Filter	laxonomic Filter



GenBank:	1 Fil	lter application box
Alphabet of Life ((Drought) AND "seed pla	Save search Limits Advanced search Help	
Display Settings: Summary, 20 per page, Sorted by Defau	ltorder	Send to: Send to: Filter your results:
Found 22653 nucleotide sequences. Nucleotide (559)	EST (22094)	All (559)
Results: 1 to 20 of 49	<< First < Prov Pag	ge 1 of 3 Next > Last >> INSDC (GenBank) (559)
Capsicum annuum chitinase class II (CAChi2) m	RNA, complete cds	mRNA (49)
1. 1,004 bp linear mRNA		RefSeq (0)
GenBank FASTA Graphics Related Sequences		Manage Filters
Capsicum annuum stellacyanin-like protein CAS	LP1 precursor, mRNA, complete cds	▼Taxonomic Groups [List]
 937 bp linear mRNA Accession: AE291179 1 GI: 9885805 		Solanaceae (49)
GenBank FASTA Graphics		Capsicum (12)
Nicotiana attenuata lipid transfer protein 1-like (I	LTP1) mRNA, partial sequence	Solanum (10)
3. 672 bp linear mRNA		
GenBank FASTA Graphics Related Sequences		Find related data
Nicotiana attenuata osmotin 1-like (OSM1) mRN	IA. complete sequence	Database: Select
4. 958 bp linear mRNA		Find items



Gei	nBank:	Tools Links
Nucleot Alphabet of L	ide Search: Nucleotide Limits Advanced search Help	
Display Settings	s offenBank	Send: Change region shown
GenBank: HMC FASTA Grap	a attenuata osmotin 1-like (OSM1) mknA, complete sequence 068893.1 hics	Customize view
Goto: LOCUS DEFINITION ACCESSION VERSION	HM068893 958 bp mRNA linear PLN 28-DEC-2010 Nicotiana attenuata osmotin 1-like (OSM1) mRNA, complete sequence. HM068893 HM068893.1 GI:298155393	Run BLAST Pick Primers Find in this Sequence
KEYWORDS SOURCE ORGANISM	Nicotiana attenuata <u>Nicotiana attenuata</u> Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; asterids; lamiids; Solanales; Solanaceae; Nicotianoideae;	LinkOut to external resources Gramene [Gramene]
REFERENCE AUTHORS TITLE	<pre>Nicotianeae; Nicotiana. 1 (bases 1 to 958) Re,D.A., Dezar,C.A., Chan,R.L., Baldwin,I.T. and Bonaventure,G. Nicotiana attenuata NaHD20 plays a role in leaf ABA accumulation during water stress, benzylacetone emission from flowers, and the timing of bolting and flower transitions J. Exp. Bot. 62 (1), 155-166 (2011)</pre>	All links from this record Full text in PMC PubMed
PUBMED REFERENCE AUTHORS TITLE JOURNAL	20713465 2 (bases 1 to 958) Bonaventure,G., Re,D. and Baldwin,I. Analysis of drought and ABA responsive genes in Nicotiana attenuata Unpublished	Recent activity Turn Off Clear Nicotiana attenuata osmotin 1-like (OSM1) Nucleotide

sol genomics network

GenBank:

	Format	File Storage	
Nucleot Alphabet of L	ide ife Search: Nucleotide Clear		
Display Setting	s: ⊙ GenBank	Send: Change region shown	•
GenBank: HM	a attenuata osmotin 1-like (OSM1) mRNA, complete sequence	Customize view	
Go to: LOCUS DEFINITION ACCESSION VERSION	HM068893 958 bp mRNA linear PLN 28-DEC-2010 Nicotiana attenuata osmotin 1-like (OSM1) mRNA, complete sequence. HM068893 HM068893.1 GI:298155393	Analyze this sequence Run BLAST Pick Primers Find in this Sequence	
KEYWORDS SOURCE ORGANISM	Nicotiana attenuata <u>Nicotiana attenuata</u> Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; asterids; lamiids; Solanales; Solanaceae; Nicotianoideae; Nicotianeae; Nicotiana.	LinkOut to external resources Gramene	ne]
REFERENCE AUTHORS TITLE JOURNAL	<pre>1 (bases 1 to 958) Re,D.A., Dezar,C.A., Chan,R.L., Baldwin,I.T. and Bonaventure,G. Nicotiana attenuata NaHD20 plays a role in leaf ABA accumulation during water stress, benzylacetone emission from flowers, and the timing of bolting and flower transitions J. Exp. Bot. 62 (1), 155-166 (2011)</pre>	All links from this record Full text in PMC PubMed	
PUBMED REFERENCE AUTHORS TITLE JOURNAL	20713465 2 (bases 1 to 958) Bonaventure,G., Re,D. and Baldwin,I. Analysis of drought and ABA responsive genes in Nicotiana attenuata Unpublished	Recent activity <u>Turn Off</u> Cle Nicotiana attenuata osmotin 1-like (OSM1) mRNA, complete sequence)



GenBank:





PubMed, NIH database for scientific literature and publications. <u>http://www.ncbi.nlm.nih.gov/pubmed/</u>

Publed.gov U.S. National Library of Medicine National Institutes of Health	Search: PubMed	RSS Save search Limits Advanced search Help Search Clear	
Display Settings: 🖂 Summary, 20	per page, Sorted by Recently Added		Send to: 🖂

Results: 1 to 20 of 117

<< First < Prev Page 1 of 6 Next > Last >>

An insertional mutagenesis programme with an enhancer trap for the identification and tagging of genes involved in abiotic stress

 tolerance in the tomato wild-related species Solanum pennellii. Atarés A, Moyano E, Morales B, Schleicher P, García-Abellán JO, Antón T, García-Sogo B, Perez-Martin F, Lozano R, Flores FB, Moreno V, Del Carmen Bolarin M, Pineda B. Plant Cell Rep. 2011 Jun 7. [Epub ahead of print] PMID: 21647638 [PubMed - as supplied by publisher] Related citations

- Identification and expression pattern of one stress-responsive NAC gene from Solanum lycopersicum.
- Han Q, Zhang J, Li H, Luo Z, Ziaf K, Ouyang B, Wang T, Ye Z. Mol Biol Rep. 2011 Jun 3. [Epub ahead of print] PMID: 21637957 [PubMed - as supplied by publisher] <u>Related citations</u>
- Atypical epigenetic mark in an atypical location: cytosine methylation at asymmetric (CNN) sites within the body of a non-repetitive
- 3. tomato gene.

Gonzalez RM, Ricardi MM, Iusem ND. BMC Plant Biol. 2011 May 20;11(1):94. [Epub ahead of print] PMID: 21599976 [PubMed - as supplied by publisher] Free Article Related citations



PubMed:

- Relatively updated (Gap between publication and loading in PubMed database around 1-2 days).
- It doesn't have all plant science related journals (for example: Theoretical Applied and Genetics or Crop Science).

(More information: <u>http://wwwcf.nlm.nih.gov/serials/journals/index.cfm</u>)

• There are no links between articles and genes, sequences, expression or other information contained in the publication.



Sequence Read Archive (SRA), Database to store sequences produced by NGS such as Illumina, 454, Solid, Helicos... <u>http://www.ncbi.nlm.nih.gov/sra</u>

S NCBI Resources 🖸 How To 🖸		Sign in to NCBI			
SRA (SRA) Advanced		Search Help			
G ATATT AATAC	SRA				
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 Pacific Biosciences SMRT®.					
Using SRA	Tools	Other Resources			
Handbook	BLAST	SRA Home			
Download	SRA Run browser	Trace Archive			
E-Utilities	Submit to SRA	Trace Assembly			
	SRA software	GenBank Home			



Taxonomy, database with taxonomic names and classifications for NCBI organisms.

http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/

S NCBI	PORSOL T	axonomy Browser	
PubMed	Entrez B	LAST OMIM	Taxonomy Structure
Search for	As complete name 🔹 🗹 lock	Go Clear	
Taxonomy browser Archaea Bacteria Eukaryota Viroids	The NCBI Taxonomy Homepage	ms commonly used in molecular	research projects:
Viruses	These are uncer links to some of the organis		
Taxonomy	Arabidopsis thaliana	Escherichia coli	Pneumocystis carinii
common tree	Bos taurus	Hepatitis C virus	Rattus norvegicus
Taxonomy	Caenorhabditis elegans	Homo sapiens	Saccharomyces cerevisiae
information	Chlamydomonas reinhardtii	Mus musculus	Schizosaccharomyces pombe
Taxonomy	Danio rerio (zebrafish)	Mycoplasma pneumoniae	Takifugu rubripes
resources	Dictyostelium discoideum	Oryza sativa	Xenopus laevis
Taxonomic advisors	Drosophila melanogaster	Plasmodium falciparum	Zea mays



EBI (European Bioinformatics Institute) http://www.ebi.ac.uk/

EMBL-EBI European Bioinformatics Institute								
Databases	Tools	Research	Training	Industry	About Us	Help	Site Index	5
	Explore t	the EBI:						
	Examples:	ROA1_HUMAN, 1	tpi1, Sulston			Help	FIND	
		,						

Data Resoures and Tools

ENA	Genomes	Gene Expression		Literature
UniProt	Nucleotide Sequences	Molecular	1	Taxonomy
ArrayExpress	Protein Sequences	Interactions		Ontologies
Ensembl	Macromolecular	Reactions&	1	Patent
InterPro	Structures	Pathways		Resources
PDBe	Small Molecules	Protein Families		
		Enzymes		

- Sequence Similarity &
- Analysis
- Pattern & Motif Searches
- Structure Analysis
 - Text Mining
 - Downloads
 - Web Services



EBI (European Bioinformatics Institute) http://www.ebi.ac.uk/

Highlights:

- ENA (European Nucleotide Archive).
- UniProt
- ArrayExpress
- Ensembl
- InterPro



InterPro, protein domain database organized by superfamilies, families and subfamilies. It is frequently used for genome functional annotation, specially to link genes with gene ontologies associated with protein domains. (http://

www.ebi.ac.uk/interpro/).

BI > Databases > InterPro

InterPro protein sequence analysis & classification

InterPro is an integrated database of predictive protein "signatures" used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures.

Current release: 32.0 18th April 2011 (see Release Notes for further details)

Search 😡 InterPro:

Do a sequence search of InterPro, via InterProScan

Extract large datasets by querying our BioMart #

You can access our data programmatically, via Web Services

Use the updated InterProScan Web Service

If you have any questions or feedback please contact us.





Biological Databases:

- I-Types.
- 2-Public Repositories.

3-Community specific databases.3.1- For species.3.2- For specific datatypes.

4- Genomic Browsers.

3. Community specific databases



Name	Species	Data Link			
The Arabidopsis Information Resource (TAIR)	Arabidopsis	Single Species Genomes, Genetic Markers, SNPs, Genes, Expression, Proteins, Ontologies, Metabolic Pathways, Publications	<u>http://</u> www.arabidopsis.org/		
Gramene	Monocots (Grape and Arabidopsis)*	Multiple Species Genomes, Genetic Markers, SNPs, Genes, Proteins, Ontologies, Metabolic Pathways, QTLs	<u>http://</u> www.gramene.org/		
Sol Genomics Network (SGN)	Solanaceae, Rubiaceae	Multiple Species Genomes, Genetic Markers, SNPs, Genes, Expression*, Proteins, Ontologies, Metabolic Pathways, Publications, QTLs and Maps, Phenotypes	http://solgenomics.net/		
Genome Database for Rosaceae (GDR)	Rosaceae	Multiple Species Genomes, Genetic Markers, SNPs, Genes, Proteins, Ontologies, Phenotypes, Unigenes	<u>http://</u> www.rosaceae.org/		
Dendrome	Trees	Multiple Species Genomes, Genes, SNPs, Genetic Markers, Genetic Maps, Phenotypes, Literature, Expression	<u>http://</u> <u>dendrome.ucdavis.edu</u> /		
Curcurbit Genomics Database (ICuGI)	Cucurbitaceae	Multiple Sequence Genomes, Genes, Unigenes,Genetic Markers, Genetic Maps, Metabolic Pathways	http://www.icugi.org/		

3. Community specific databases



Name	Species	Data	Link			
The Compositae Genome Project	Compositaceae	Multiple Species Genomes, Genetic Markers, Genetic Maps, Genes	<u>http://</u> compgenomics.ucdavis. <u>edu</u> /			
MaizeGDB	Zea mays (Maize)	Single Species Genomes, Genetic Markers, SNPs, Genes, QTLs, Cytogenetics, Phenotypes	<u>http://</u> www.maizegdb.org/			
CottonDB	Gossypium spp.	Genetic Markers, Maps, Genes, Taxonomic. Data	http://cottondb.org/			
SoyBase	Glycine Max (Soybean)	Multiple Species Genomes, Genetic Markers, SNPs, Genes, Proteins, Ontologies, Metabolic Pathways, QTLs	http://soybase.org/			
Phytozome	Plants	Multiple Species Genomes	<u>http://</u> www.phytozome.net			
Plant Genome Database (PlantGDB)	Plants	Multiple Species Genomes, Genes, Unigenes	http://www.plantgdb.org/			



Biological Databases:

- I-Types.
- 2-Public Repositories.
- 3-Community specific databases.3.1- For species.3.2- For specific datatypes.

4- Genomic Browsers.

3. Community specific databases



There are other community driven databases focused in a knowledge area:

Metabolic databases: MetaCyc: http://metacyc.org/ KEGG: http://www.genome.jp/kegg/

Ontology databases: Gene Ontology: http://www.geneontology.org/ Plant Ontology: http://www.plantontology.org/

Transcription Factors database: TranscriptionFactorDB (DBD): <u>www.transcriptionfactor.org</u>



Biological Databases:

- I-Types.
- 2-Public Repositories.
- 3-Community specific databases.3.1- For species.3.2- For specific datatypes.
- 4- Genomic Browsers.



A Genome Browser is a graphical interface that shows aligned genomic data.

Each data type is in a track.

The tracks are hierarchically organized by track size. For example, the first track could be a *chromosome*, the second one a *region* and the third one, a *detailed region* with gene structures. 4. Genomic Browsers



Genome Browser most used:

- JBrowse (GMOD).
- GBrowse (GMOD).
- UCSC Genome Browser.
- Emsembl Genome Browser.
- -Vista Genome Browser.



http://solgenomics.net/

	ethent	to also	inder Frield Frie
search map	ps genomes	tools	sol search
111		Sequence Analysis	n new use
		BLAST	
	/	VIGS Tool	eF.
		Alignment Analyzer	
Mane & Markers	772	Tree Browser	nes
		Intron Finder	pes
СТ233		Mapping	
		Genome Browser (Jbrowse)	
C015		Comparative Map Viewer	
C2_At4g15790		CAPS Designer	
		solQTL: QTL Mapping	
		Molecular Biology	100
		In Silico PCR	
		Tomato Expression Database (TE	D)
		Systems Biology	
	ALL STATES	SolCyc Biochemical Pathways	
		Coffee Interactomic Data	
		Breeder Tools	
	40	Breeder Home	
Breeders Toolbox	н	Bulk Query	s
	¥	Unigene and BAC information	
	Genomes & Sequence	FTP Site	
		ID Converter (SGN <=> TIGR)	





Available Tracks		Tomato SL2.50 IT	AG2.4 -	File	View	Help						
X filter by text		0 5,000,000	10,000,000	15,0	00,000	20,000,000	25,000,000	30,000,000	35,000,000	40,000,000	45,000,000	50,0
Gene models	1			\rightarrow	Q	Q Q (Electric SL2.500	h11 👻 SL2.5	0ch11:3277059	132780130 (9.	54 Kb) Go	2
ITAG2.4 gene models			32,772,5	500			32,775,000)		32,777,50	0	
r Genetic loci	3	Reference sequence	e		Zoom in t	o see sequenc	e	Zoom in	n to see sequenc	e	Zoom in	to see s
SGN locus sequences SGN markers SolCAP_SNPs	5	n-like protein (Fragment) (AF	ets =_sol.tu); a	contains Int 4L1_SOLN	erpro domai I); contains I	 inte	-	•		Receptor-like k Solyc11g04	nase (AHRD V1 ***- 1370.1	A7VM20
ESTs and cDNAs - Other Solanaceae ESTs and cDNAs - Tomato MicroTom full-length cDNAs SGN unigenes SL2.50_assembly		€	Unknov Solyc	wn Protein (11g0443	AHRD V1); 80.1	contains Interpro	domain(s) IPR004	158 Protein			+	
 Prediction features (de novo) AUGUSTUS (de novo, Tomato trained) GlimmerHMM (de novo, Arabidopsis trained) GlimmerHMM (de novo, tomato trained) Infernal geneID (de novo, Tomato trained) tRNAscanSE 	6											
▼ Quantitative ▼ RNAseq Density	4											









gene Solyc06g069410.2

Primary	Data

	Name	Solyc06g069410.2						
	Туре	gene						
	Position	SL2.50ch06:43166656.						
	Length	3,295 bp						
At	tributes							
	Alias	Solyc06g069410						
	From_bogas	1						
	ld	gene:Solyc06g069410.2						
	Length	3295						
	Seq_id	SL2.50ch06						
	Source	ITAG_eugene						
	Region sequence							
	>SL2.50ch06 SL2.50ch06:431 class=gene length=3295 ATTAAGGAGGGGGAACTTGGGGCCTA TTTTCTGATGGGAGGAACAGCAGGCA AGTAAAGCTTTTGTTGCAGAATCAAG							

ATATGTGGGAATTGGTGATTGCTTTC GTGGAGGGGAAACCAGGCCAATGTTA TCATCTCCTTGTGATGTTTTAGACCT CTTAGCACACTGAACAGTTAACCTTC TAGATTGATGAAGTCCAACTTATTGA TGTCTCCCCTGGTTTGTGAGACTAGT

> RepeatMasker (aggressive) RepeatMasker (normal)

gene Solyc06g	g069410.2		×	Market Contraction			
Subfeatu	ires			%2CSo C Reader O			
Primary	y Data			1,000 90,000,000			
Nam	ne			› 🖉			
Mitod	chondrial AD	P/ATP carrier proteins (AHRD V1 **** Q2UU95_ASPOR); contains Interpro	93,770,000				
doma	domain(s) IPR002113 Adenine nucleotide translocator 1						
Туре	е	mRNA	M240_ARALY)				
Des	cription			Poptovi cis-trans isomerase (A)			
Mitod	chondrial AD	P/ATP carrier proteins (AHRD V1 **** Q2UU95_ASPOR); contains Interpro	Selyc01g105710.2				
doma	domain(s) IPR002113 Adenine nucleotide translocator 1						
Posi	ition	SL2.50ch06:4316665643169950 (+ strand)					
Leng	gth	3,295 bp	ck				
Attribut	Attributes			oom			
From	n_bogas	1					
Id		mRNA:Solyc06g069410.2.1		A			
Inter	pro2go_te	m GO:0016020 GO:0005743					
Leng	yth	3295					
Nb_e	exon	3					
Onto	ology_term	GO:0005471					
Seq	_id	SL2.50ch06					
Sifte	Sifter_term GO:0005471						
Sour	rce	ITAG_eugene					

×

load tracks: Fasta, GFF3, BAM, BigWig

Display a menu


JBrowse



)	20,000,000	40,00	0,000	60,000,000	80,000,000	
		Q @ 🕀	SL2.50ch01 -	Solyc06g0691		Go 📣
	35,000,000		35,500	Solyc06g069100),000
-	Genetic loci	3 Reference sequen	CO E E K G	Solyc06g069100.1		
8	SGN locus sequences SGN markers SolCAP_SNPs	MENL GGAGAACTIGG ACCTCTTGAACC	G K K R K G G C A A G A A G A A A G G G A C G C C C C C C C C C C C C C C C C C C	Solyc06g069110		L T V V CACAGOAGOCA ACOGOCAGOAGO
-	Genome data and reagents	H L V Q 5 P S S P	A L L S L S C S S F P 1	Solyc06g069110.2		Q C Y D S L L
5	ESTs and cDNAs - Other Solanaceae	1 5 F K P		Solyc06g069120		
ł	SGN unigenes	ITAG2.4_gene_mo	dels	Solyc06g069120.2		
Ľ	SL2.50_assembly	Coverage of RNA-S	eq reads on plus strand	Solyc06g069130		
Ľ	AUGUSTUS (de novo)	0		Solyc06g069130.2		
ł	GlimmerHMM (de novo, Arabidopsis trained) GlimmerHMM (de novo, tomato trained)	mean		Solyc06g069140		mean
l	geneID (de novo, Tomato trained) tRNAscanSE			Solyc06g069140.1		
- (Quantitative	4		Solyc06g069150		
ĿP	RNAseq Density			Solyc06g069150.1		
L	Density of RNAseq reads on minus strand Density of RNAseq reads on plus strand	Navios	ation zoom	Solyc06g069160	find featur	e hv name
ĿP	RNAseq XYPlot	i vaviga		Solyc06g069160.1	into roator	e by hame
L	 Coverage of RNA-Seq reads on minus strand Coverage of RNA-Seq reads on plus strand 	selecti	ng a region	Solyc06g069170		
- 1	Reference sequence	1		Solyc06g069170.2		
6	Reference sequence			Solyc06g069180		
-	Repetitive elements	2		Solvc06q069180.2		
8	RepeatMasker (aggressive) RepeatMasker (normal)			Solvc06c069190		
D	isplay a menu			Solycoogoos 150		
-				50lycubgub9190.2		

Sol Genomics Network

JBrowse



Tomato 360 variants SL2.50	•	File	View	Help		
Tomato SL2.40 ITAG 2.3		0,000	6	6,000,000	8,000),000
Tomato SL2.50 ITAG2.4						
Tomato variants SL2.40				\leftarrow		$\Theta \oplus \Theta$
Tomato 360 variants SL2.50		12,00	0,000			12,250,
Tomato 150 variants SL2.50						
Solanum pennellii				Zoom in to	see sequence	ce
N.benthamiana v1.0.1						
N.benthamiana v0.4.4						
Pepper 1.55						
N.tabacum TN90						
S. tuberosum Divi 1-3 v4.03]				



JBrowse: Nicotiana benthamiana



Sol Genomics Network

JBrowse: Pepper genome



Sol Genomics Network

JBrowse



Tomato 0	SL2.50 ITAG2.4 T File View Help Open files	×	000.0
35,000	Add any combination of data files and URLs, and JBrowse will automatically suggest tracks to display their contents.		10534
	Local files Remote URLs - one per line		
👌 ITAG2	Select Files http://paste.uris.here/example.bam		
	Select or drag files here.		
	Files and URLs		
	Add files and URLs using the controls above.		
	New Tracks		
	None		
	Open immediately		
	🗙 Cancel 🗁 Open		





GFF3 (<u>http://www.sequenceontology.org/gff3.shtml</u>)

- Column 1: seqid example: SL2.40ch08
- Column 2: source example: ITAG_eugene
- Column 3: type example: mRNA
- Columns 4 & 5: start & end example: 56809898 and 56812517
- Column 6: score example: .
- Column 7: strand example: -
- Column 8: phase example: .
- **Column 9:** attributes example: Name=Solyc08g075490.2.1;length=2620;

Exercise



- I. You are a coffee researcher and want to understand more about caffeine synthesis. Using the tools we discussed, do the following analyses with caffeine synthase.
 - I. Find some papers on caffeine synthase published since 2010.
 - 2. How many caffeine synthase protein sequences are in GenBank? How many are from *Coffea arabica*?
 - 3. How many species have a caffeine synthase homolog?
 - 4. Is caffeine synthase specific to the Gentianales clade or is it found elsewhere?
 - 5. Which of the homologs seem realistic? Download all *Coffea* homolog sequences in fasta format and select full-length proteins. How many appear full-length?
 - 6. What reaction(s) does caffeine synthase catalyze?

Please save your results for the next exercise.



Exercise I Solutions

Results: 6

Filters activated: Publication date from 2010/01/01 to 2014/01/01. Clear all to show 23 items.

- Identification and isolation of full-length cDNA sequences by sequencing and analysis of
- expressed sequence tags from guarana (Paullinia cupana), Figueirêdo LC, Faria-Campos AC, Astolfi-Filho S, Azevedo JL, Genet Mol Res. 2011 Jun 21;10(2):1188-99. doi: 10.4238/vol10-2gmr1124. PMID: 21732283 [PubMed - indexed for MEDLINE] Free Article Related citations
- Producing low-caffeine tea through post-transcriptional silencing of caffeine synthase mRNA.
- Mohanpuria P, Kumar V, Ahuja PS, Yadav SK. Plant Mol Biol. 2011 Aug;78(6):523-34. doi: 10.1007/s11103-011-9785-x. Epub 2011 May 12. PMID: 21562910 [PubMed - indexed for MEDLINE] Related citations
- Agrobacterium-mediated silencing of caffeine synthesis through root transformation in Camellia

 sinensis L. Mohanpuria P, Kumar V, Ahuja PS, Yadav SK. Mol Biotechnol. 2011 Jul;48(3):235-43. doi: 10.1007/s12033-010-9364-4. PMID: 21181507 [PubMed - indexed for MEDLINE] <u>Related citations</u>

A transcriptomic approach highlights induction of secondary metabolism in citrus fruit in response

 to Penicillium digitatum infection. González-Candelas L, Alamar S, Sánchez-Torres P, Zacarías L, Marcos JF. BMC Plant Biol. 2010 Aug 31;10:194. dok 10.1186/1471-2229-10-194. PMID: 20807411 [PubMed - indexed for MEDLINE] Free PMC Article Related citations

- Essential region for 3-N methylation in N-methyltransferases involved in caffeine biosynthesis.
- Mizuno K, Kurosawa S, Yoshizawa Y, Kato M. Z Naturforsch C. 2010 Mar-Apr;65(3-4):257-65. PMID: 20469646 [PubMed - indexed for MEDLINE] Related citations
- Expression for caffeine biosynthesis and related enzymes in Camellia sinensis.
- Kato M, Kitao N, Ishida M, Morimoto H, Irino F, Mizuno K. Z Naturforsch C. 2010 Mar-Apr;65(3-4):245-56.
 PMID: 20469645 [PubMed - indexed for MEDLINE] <u>Related citations</u>

I. use pubmed (<u>http://</u> <u>www.ncbi.nlm.nih.gov/</u> <u>pubmed</u>



 \odot

Search

Exercise I Solutions

2. II proteins total, I from C. arabica (<u>http://</u><u>www.ncbi.nlm.nih.gov/protein</u>)

Protein	
---------	--

٢

caffeine synthase[Protein Name]

Save search Advanced



Exercise I Solutions

3.4 species:

Results by taxon

Top Organisms [Tree] Camellia sinensis (6) Paullinia cupana var. sorbilis (3) Coffea arabica (1) Theobroma cacao (1)

*click on "Tree" for next question



Exercise I Solutions (cont'd)

4. Found in Gentianales, Ericales, Sapindales, Malvales

Taxonomic Groups [List]

Ericales (6) Sapindales (3) Gentianales (1) Malvales (1)

*click on "List" for next answer



Exercise I Solutions (cont'd)

5. Coffea arabica, Coffea canephora, Camellia sinensis, Theobroma cacao, Paullinia. 7 sequences appear to be full-length.

sults: 11		
TPA_exp	caffeine synthas	e [Paullinia cupana var. sorbilis
360 aa pr Accession:	otein DAA64605.1 GI: 645	065978
GenPept	FASTA Graphics	Related Sequences
caffeine	synthase, partial [Paullinia cupana var. sorbilis]
351 aa p	rotein	
Accession:	AHA44434.1 GI: 557	942098
	EACTA Craphice	Related Sequences

Send to: V Filters: Manage Filters Choose Destination File Clipboard on Collections Analysis Tool pups [Li Download 11 items Format (1)0 FASTA Sort by Default order sequen Create File Align sequences with COI

Identify Conserved Doma



7.) <u>http://metacyc.org/</u>



MetaCyc Enzyme: (

Gene: CaDXMT1 Accession Number: G-9084 (MetaCyc)

Species: Coffea arabica

Summary:

The recombinant CaDXMT1 catalyzes the conversion of 7-methylxansine to theobromine and theobromine to caffeine. In addition, Ca the most preferred substrate. XMP is not an effective substrate. CaDXMT1 is predominantly expressed in immature fruits of coffee.

Ca-CaMXMT1

Ca-CaMXMT2

Ca-CaDXMT1

Ca-CCS1

Citations: [Uefuji03]

Molecular Weight of Polypeptide: 43.3 kD (from nucleotide sequence)

Relationship Links: Entrez-Nucleotide: PART-OF: AB084125

Gene-Reaction Schematic:



Coffea arabica



Part II: Web Tools





Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).
 - I.3 By Sequence/Chromosome coordinates (GBrowse).
- 2 Manipulation and Sequence Analysis Tools:
 - 2.1 Translators and Gene Predictors.
 - 2.2 Multiple Sequence Alignment(Clustalw).
 - 2.3 Protein Domain Analysis (InterProScan).
 - 2.4 Signal Peptide Analysis (SignalP).
- 3 Other Tools:
 - 3.1 Linkage Map Viewers (CViewer).
 - 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.



Text Searches:

One or more words are introduced in a box. The system use them to search coincidences with database fields or file sections such as genomic annotations.



NCBI: http://www.ncbi.nlm.nih.gov/



Text Searches:

EBI: http://www.ebi.ac.uk/

EMBL	-EBI	Europ	ean Bioin	formatics	Institute				
Databases	Tools	Research	Training	Industry	About Us	Help	Si	ite Index	5
	Explore the	he EBI:					FIND		
	Examples:	ROA1_HUMAN, t	oi1, <u>Sulston</u>			Help	Feedback		

Data Resoures and Tools

ENA

.

- UniProt
 - ArrayExpress = Protein Sequences

Genomes

Nucleotide Sequences

- Ensembl
 Macromolecular
- InterPro Structures
 - PDBe Small Molecules
- Gene Expression

Literature

Taxonomy

Ontologies

Resources

Patent

- Molecular
- Interactions
- Reactions& Pathways
- Protein Families
- Enzymes

- Sequence Similarity &
 - <u>Analysis</u>
- Pattern & Motif Searches
- Structure Analysis
- Text Mining
- Downloads
- Web Services



Text Searches:

TAIR: http://www.arabidopsis.org/

tair	Home He	lp Contact A	bout Us Login/Re	gister		Gene	\$ Search
Search	Browse	Tools	Portals	Download	Submit	News	ABRC Stocks

The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant Arabidopsis thaliana. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every two weeks from the latest published research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, preserves and distributes seed and DNA resources of Arabidopsis thaliana and related species. Stock information and ordering for the ABRC are fully integrated into TAIR.

TAIR is located at the Carnegie Institution for Science Department of Plant CARNEGIE Biology and funded by the National Science Foundation. SCIENCE



added to TAIR: Arabidopsis lyrata, Brachypodium distachyon, Oryza sativa japonica, Oryza sativa indica, Populus trichocarpa, Physcomitrella patens, Sorghum bicolor, Vitis vinifera, Zea mays.

Breaking News

[May 19, 2011]

Subscribe to news feed

Follow our Twitter feed

Join our Facebook group

GBrowse now available for

eight plant species at TAIR

GBrowse instances for the

following plants have been

Updates on TAIR funding are available here.



Text Searches:

GRAMENE: <u>http://www.gramene.org</u>/







Text Searches:

SGN: <u>http://solgenomics.net/</u>





Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).
 - I.3 By Sequence/Chromosome coordinates (GBrowse).
- 2 Manipulation and Sequence Analysis Tools:
 - 2.1 Translators and Gene Predictors.
 - 2.2 Multiple Sequence Alignment(Clustalw).
 - 2.3 Protein Domain Analysis (InterProScan).
 - 2.4 Signal Peptide Analysis (SignalP).
- 3 Other Tools:
 - 3.1 Linkage Map Viewers (CViewer).
 - 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.

I.I.Searches by Ontology



Ontologies in bioinformatics:

Controlled vocabulary of terms used to describe gene functions, localizations, plant organs... that allow the communication between different groups.

Example: Size increase, Mass increase...

Growth, The increase in size or mass of an entire organism, a part of an organism or a cell. (GO:0040007)

Ontology Types:

More used is *Gene Ontology* (<u>http://www.geneontology.org</u>/), to define biological process, cellular component and molecular function terms.

For plants, also is important *Plant Ontology* (<u>http://www.plantontology.org</u>/), to define plant parts such as organs and process associated to plants such as flowering or ripening.

A complete ontology list can be found at:

http://www.obofoundry.org/

I.I.Searches by Ontology



Searches by Ontology: Ontologies have parents-children relations. To optimize the search, it is common the use of tools such as the ontology browsers.

search	maps	genomes	tools	_ sol search
/W				log in new u
		Browse Onto	ologies	
ntology browser				
Find exact ID	Find	clear highlight	reset view	
Coursels for bout				
Search for text		GO (gene on	tology) 🛟	Search
GO:0003674 molecu	lar_function			
GO:0005575 cellular	_component			
GO:0008150 biologic	al_process			
Tis_a GO:0008283	cell proliferation			
Tis_a GO:0007587	sugar utilization			
Tis_a GO:0019740	nitrogen utilizati	on		
Tis_a GO:0009758	carbohydrate uti	lization		
"is_a GO:0006791	sulfur utilization			
Tis_a G0:0015976	carbon utilization	1 ration		
TIS_8 G0:0006794	cellular compone	ant organization or	hiogonosis	
Bis a G0:0023052	signaling	and organization of	biogenesis	
® is a G0:0000003	reproduction			
* is a GO:0009987	cellular process			
Tis_a GO:0016032	viral reproductio	n		
🖹 īs, a GO:0040007	growth			
🕆 is_a GO:00071	17 budding cell b	bud growth		
is_a GO:00550	17 cardiac musc	le tissue growth		

http://solgenomics.net/tools/onto/index.pl



Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).
 - I.3 By Sequence/Chromosome coordinates (GBrowse).
- 2 Manipulation and Sequence Analysis Tools:
 - 2.1 Translators and Gene Predictors.
 - 2.2 Multiple Sequence Alignment(Clustalw).
 - 2.3 Protein Domain Analysis (InterProScan).
 - 2.4 Signal Peptide Analysis (SignalP).
- 3 Other Tools:
 - 3.1 Linkage Map Viewers (CViewer).
 - 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.



Sequence homology/similarity searches:

It is based in the sequence comparison through a pair sequence alignment using different algorithms (blast, uses an approach to the Smith-Waterman algorithm). Matched sequences (hits) with some statistical values are selected and returned as result.

Most used programs are:

- Blast: (faster) http://blast.ncbi.nlm.nih.gov/Blast.cgi
- Fasta (sensitive): <u>http://www.ebi.ac.uk/Tools/sss/fasta/</u>

More information at: http://en.wikipedia.org/wiki/Sequence_alignment_software





Sequence homology/similarity searches:

NCBI: http://blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST®		Basic Local Alignment Search To	ol
Home Recent	Results Saved Strategies	Help	
NCBI/ BLAST Home			
BLAST finds regio	ons of similarity between biol	ogical sequences. more	
	New Aligning Multi	ple Protein Sequences? Try the COBALT Multiple	Alignment Tool. Go
BLAST Assem	bled RefSeq Genomes		
Choose a species g	enome to search, or list all gen	nomic BLAST databases.	
 <u>Human</u> <u>Mouse</u> <u>Rat</u> <u>Arabidopsis th</u> 	c naliana c	<u>Oryza sativa</u> <u>Bos taurus</u> <u>Danio rerio</u> Drosophila melanogaster	 <u>Gallus gallus</u> <u>Pan troglodytes</u> <u>Microbes</u> <u>Apis mellifera</u>
Basic BLAST			
Choose a BLAST p	rogram to run.		
nucleotide blast	Search a nucleotide database Algorithms: blastn, megat	e using a nucleotide query plast, discontiguous megablast	
protein blast	Search protein database usin Algorithms: blastp, psi-bla	g a protein query ast, phi-blast	
blastx	Search protein database usin	g a translated nucleotide query	
tblastn	Search translated nucleotide	database using a protein query	
tblastx	Search translated nucleotide	database using a translated nucleotide query	



Sequence homology/similarity searches:

EBI: <u>http://www.ebi.ac.uk/Tools/sss/</u>

EMBL-EBI			Enter Text Her	0		Find	Help Feedback		
Databases Tools	Research	Training	Industry	About Us	Help		Site Index 🔝 🚄		
= Tools Home	EBI > Tools > Seque	ance Similarity Sea	arching						
= Tools A-Z	Sequence Sim	ilarity Searc	hing						
= Web Services = Download	BLAST								
	NCBI BLAST 🛈	NCBI BLAST S tool is available	equence Simila for the followin	rity Search using databases:	ing the NCBI	BLAST (blasta	ll) program. This		
	WU-BLAST ()	rsity (WU) BLAS ing databases:	ST2 program						
	PSI-BLAST 🛈	Position Specific Iterative BLAST (PSI-BLAST) refers to a feature of BLAST 2.0 profile is automatically constructed from the first set of BLAST alignments. Q Launch PSI-BLAST							
	FASTA								
	FASTA 🕢	Sequence Simi following datab	ilarity Search us ases:	ing the FAST/	A program. Ti	his tool is availa	able for the		
		Q, Protein Q, Q, ASD Protein	Nucleotide Q	Proteomes Q	Genomes	Q Whole Genom	te Shotgun		
	SSEARCH ()	Sequence Similarity Search using the SSEARCH program. This tool is available for the following databases:							
		Q, Protein Q,	Nucleotide Q	Proteomes Q	Genomes	Q Whole Genom	ne Shotgun		
		Q, ASD Protein	Q, ASD Nucle	tide Q, LGIC	Protein Q, L	GIC Nucleotide			
	PSI-Search 🕡	PSI-Search cor with the PSI-BL protein sequen	nbines the sens AST (blastpgp) ces.	itivity of the Sr iterative profil	nith-Waterma e constructio	an search algor n strategy to fin	rithm (SSEARCH) d distantly related		
		Q, Launch PSI-	Search						



Sequence homology/similarity searches:

TAIR: <u>http://www.arabidopsis.org/Blast/index.jsp</u> <u>http://www.arabidopsis.org/cgi-bin/fasta/nph-TAIRfasta.pl</u>

						Gene	Search								Gene	\$ Search
tair	Home Help	Contact Ale	out Us Login/Regis	er.				tair	Home He	lp Conta	ct About	Us Login/Regis	lor			
Search	Browse	Tools	Portals	Download	Submit	News	ABRC Stocks	Search	Browse	Tools		Portals	Download	Submit	News	ABRC Stocks
Home > Tool	Is > BLAST ST 2.2.8	NCBI BLAST	228 and NOT W	LBI AST2 0					FASTA Name of query Enter a query	y: (option	al) æ: (forma	t: raw or fasta)]			
Blast	program		BLACTN: NT current	NT db					OR Upload a file	containir		samuence: Horris	nat: row or fact	a).		
Datasets	:	(TAIR10 Transcript	s (-introns, +UTF	s) (DNA)	•			Choose File	No file ch	osen	sequence: (ion	mail: raiw or fast	a)		
Input: e query locus (At1g0	sequence name)1030)								Datasets: [Der Submit () The type of query	Reset	TAIR10	Transcripts (-intr	e program used:	(A)		
Upload a Raw, FAS Filter o	file TA, GCG and RS tuery	F formats accep	Choose File No f	ile chosen					Query Dat DNA DN Protein Pro Protein DN	aset F A f tein f A t	Program Iasta3, sear Iasta3 fastx3, sea	rches both strands	5			
Advar	nced BLAST"	Parameter (Options				+		DNA Pro Options	tein t	astx3, forw	ard 3 frames, see	options for reverse	0		





Sequence homology/similarity searches:

GRAMENE: <u>http://www.gramene.org/multi/blastview</u>

GRAMEN .	BLAST I BioMart I Documentation I Help I Feedback	d.
	new SETUP +> CONFIG +> RESULTS +> OISPLAY	(refresh) Online Help) Summary
	We now use Blat as our default DNA search. This will make your query faster.	Setup Ø Not yet initialised
	Enter the Query Sequence	Configure Ø Not yet initialised
	Either Paste sequences (max 30 sequences) in FASTA or plain text:	results Ø Not yet initialised
	Or Upload a file containing one or more FASTA sequences	display Ø Not yet initialised
	Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq)	
	Or Enter an existing ticket ID: Retrieve o dna queries peptide queries	
	Select the databases to search against	1
	Select species: Use 'ctrl' key to select multiple species Oryza_glaberrima Oryza_indica • Oryza_sativa	





Sequence homology/similarity searches:

SGN: <u>http://solgenomics.net/tools/blast/index.pl</u>

search	maps	genomes	tools	sol search
		NCBI BL	AST	log in new user
		Simple	dvanced	
Sequence Set	SGN Tomato Com	bined - WGS, BAC,	and unigene sequence	es 🔹 🗘 db details
Program	BLASTN (nucleotid	=		
		Query seq	uence	
	single	sequence only, use A	dvanced for multiple	
	single	sequence only, use A	dvanced for multiple	
Expect (e-va	single	sequence only, use A	dvanced for multiple	Clear Search
Expect (e-va 1e-10	single	sequence only, use A	idvanced for multiple	Clear Search
Expect (e-va le-10 Substitution	single nlue) Threshold Matrix	sequence only, use A	Show Graphics	Clear Search



Blast:

It is a tool designed to find regions with local similarity for a sequence pair. It compare nucleotides or protein sequences and calculate the statistical significance.

Blast Programs:

	INPUT				
DATABASE		Nucleotide	Translated Nucleotide	Protein	
	Nucleotide	BlastN	-	-	
	Translated Nucleotide	-	TBlastX	TBlastN	
	Protein	-	BlastX	BlastP	

I.2 -Search by Sequence Homology



Blast uses:

Homologous gene search:

BlastX (input=cDNA, database=proteins). BlastP (input=protein, database=proteins). TBlastN (input=proteins, database=cDNA)

Intron-Exon alignment:

BlastN (input=cDNA, database=genomic DNA). (better Blat or GeneWise)

SNP search:

BlastN (input=cDNA,gDNA, database=cDNA,gDNA).

I.2 -Search by Sequence Homology



Blast terminology:

Query: Input sequence.

Subject: Sequence from the database

Query Coverage: Percentage of the input sequence cover by the database sequence.

E-value (expect value): Expected hits at random. It depends from the database size and it decrease exponentially with the sequence pair score.

% *identity*: Identity percentage for a sequence pair.



Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).

I.3 - By Sequence/Chromosome coordinates (GBrowse).

- 2 Manipulation and Sequence Analysis Tools:
 - 2.1 Translators and Gene Predictors.
 - 2.2 Multiple Sequence Alignment(Clustalw).
 - 2.3 Protein Domain Analysis (InterProScan).
 - 2.4 Signal Peptide Analysis (SignalP).
- 3 Other Tools:
 - 3.1 Linkage Map Viewers (CViewer).
 - 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.



Sequence Location Searches (By coordinates):

Based in the location at a chromosome region or a sequence. It has a start and end coordinates.

These searches generally uses Genome Browser as search software

More information at: <u>http://en.wikipedia.org/wiki/Genome_browser</u>



Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).
 - I.3 By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

- 2.1 Translators and Gene Predictors.
- 2.2 Multiple Sequence Alignment(Clustalw).
- 2.3 Protein Domain Analysis (InterProScan).
- 2.4 Signal Peptide Analysis (SignalP).

3 - Other Tools:

- 3.1 Linkage Map Viewers (CViewer).
- 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.


There are dozens of sequence manipulation tools with different licenses or for different operating systems.

- + Commercial package: LaserGene (DNAStar) (<u>http://www.dnastar.com/t-products-lasergene.aspx</u>)
- + Free packages: BioEdit (Windows) (<u>http://www.mbio.ncsu.edu/bioedit/bioedit.html</u>) eBioTools (MacOS) (<u>http://www.ebioinformatics.org/</u>)

Some databases have programs with similar functions integrated with the database interface.



Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).
 - I.3 By Sequence/Chromosome coordinates (GBrowse).
- 2 Manipulation and Sequence Analysis Tools:
 - 2.1 Translators and Gene Predictors.
 - 2.2 Multiple Sequence Alignment(Clustalw).
 - 2.3 Protein Domain Analysis (InterProScan).
 - 2.4 Signal Peptide Analysis (SignalP).

3 - Other Tools:

- 3.1 Linkage Map Viewers (CViewer).
- 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.



There are two tools types to find the right ORF for an expressed nucleotide sequence.

- Align with a mRNA or cDNA sequence, and select the longest ORF.
- Gene prediction based on the exon-intron structure

Tool types:

- Translators (DNA to proteins without exon-intron consideration, and analyzing all the possible ORFs). Use coding.
- Gene Predictors (DNA to CDS considering the intron-exon structure). They require software training with manually curated intron-exon structures.

2.1 - Translators and Gene Predictors.



Web-based translator programs:

- Translate Tool (ExPASy): http://expasy.org/tools/dna.html
- ORF Finder (NCBI): <u>http://www.ncbi.nlm.nih.gov/projects/gorf/</u>
- Transeq (EBI): <u>http://www.ebi.ac.uk/Tools/emboss/transeq/</u>
- RevTrans I.4 Server (CBS): <u>http://www.cbs.dtu.dk/services/RevTrans/</u>
- Transeq (UMass): <u>http://biotools.umassmed.edu/cgi-bin/biobin/transeq</u>
- Dnatoprotein (JHI): <u>http://www.dnatoprotein.com</u>/
- EstScan (embnet): <u>http://www.ch.embnet.org/software/ESTScan2.html</u>

2.1 - Translators and Gene Predictors.



• Transeq (EBI): <u>http://www.ebi.ac.uk/Tools/emboss/transeq/</u>

EMBL-EBI			Enter Text H	lere		Find	Help Feedback	
Databases Tools	Research	Training	Industry	About Us	Help		Site Index 🔊 🎒	
 Help Index General Help 	EBI > Tools > Sequence Analysis > EMBOSS							
 Formats Gaps Matrix 	EMBOSS Transeq <u>Transeq</u> translates nucleic acid sequences to the corresponding peptide sequence. It can translate in any of the 3 forward or three reverse sense frames, or in all three forward or reverse frames, or in all six frames.							
 References EMBOSS-Transeq Help Emboss Programmatic Access 	Fran 1 2 3 5T. F -1	ne	Sta Trim No 🛟	indard Code	Table Reverse	(¢ Colour No ¢	
	Enter or R 6 Upload a file: (Ucleic acid	Sequence in an	y format:		Run	Help	

2.1 - Translators and Gene Predictors.



Web-based gene predictor programs:

• FGENESH (ULondon):

http://mendel.cs.rhul.ac.uk/mendel.php?topic=fgen-file

• GENESCAN (MIT):

http://genes.mit.edu/GENSCAN.html

• GeneMark.hmm (GaTech):

http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi

• Augustus:

http://augustus.gobics.de/submission



Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).
 - 1.3 By Sequence/Chromosome coordinates (GBrowse).
- 2 Manipulation and Sequence Analysis Tools:
 - 2.1 Translators and Gene Predictors.
 - 2.2 Multiple Sequence Alignment (Clustalw).
 - 2.3 Protein Domain Analysis (InterProScan).
 - 2.4 Signal Peptide Analysis (SignalP).

3 - Other Tools:

- 3.1 Linkage Map Viewers (CViewer).
- 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.



There are programs for multiple sequence alignment (nucleotide or protein) such as ClustalW or Muscle

Some of them, as ClustalW, can create simple phylogenetic trees based in simple algorithms such as *Neighbor-Joining*.

- ClustalW (EBI): <u>http://www.ebi.ac.uk/Tools/msa/clustalw2/</u>
- Kalign (EBI): <u>http://www.ebi.ac.uk/Tools/msa/kalign</u>
- MAFFT (EBI): <u>http://www.ebi.ac.uk/Tools/msa/mafft</u>
- MUSCLE (EBI): <u>http://www.ebi.ac.uk/Tools/msa/muscle</u>
- T-Coffee (EBI): <u>http://www.ebi.ac.uk/Tools/msa/tcoffee</u>



🙆 🔿 🔿 📄 se	quence.fasta			
>gil152250891refINP_180705.11 F-box protein DOR [Arad MKSRRQNVSVARQTILGRDENFEPIPIDLVIEIFSRSPVKSIARCRCVSKLW ARPRLLFACQKHRELFFFSTPQPHNPNESSSPLAASFHMKIPFDGRFNIISP PEFVSAICNPSTGQSLTLPKPKTRKRIWGTSHFGYDPIEKQFKVLSMNIGDG RIECSIPHVHGSKGICINGVLYYRAKADMFSGTLMIVCFDVRFEKFSYIKIL EGPSYICGKRFEMWVLGDPEKHEWLKHTYELRPRWQNVLGEDLLIFAGMTGT YNLERNTIRRVEIQGMGAFKVNEDYIFLDHVEDVKLI	Didopsis thaliana] ASILRLPYFTELYLTKSC IGGLVFVRYEQILKGRKT VYKEHYVLTLGTENLSWR KPTTTLISYNGKLASLVW NEIVLSPKYPSHPFYVFY	Input: Set	of 5 pr	oteins
>gil2978372291refIXP_002886496.11 hypothetical prote MKTERQNVSEDVVVVTERNKRAKTSNNGGEPIPFDLTVEICSRLPAKSISRF LFLTRSLARPQLLFACHKDNHVFVFSSPQPQNIDDNNASSLLAANYHMKIPF DERYSNGKEHKVSVICNPSTRQSLTLPKLKTRKRIGVRSYFGFEPIEKQYKV QVLTLGTRKPSWRMIECWIPHSLYHTYNNVCINGVLYYPAVNTSSKGFIIVS SSYYGPHLINYNGKLGSLGSGGFGGIGASCTSITLRVLEDAEKHEWSEHIVV VGVTRTNEIVLSLRFPSTPFYVFYYNTERNAIRRVEIQGQEAFKDHSVYTFL	in ARALYDRAFT_893290 RCVLKLWGSILRLPYFTE YASSFERCSSVRGLVFFG LSMTWGIYGTRDMDSEEH FDFRSEEFRFVEDTDTSI LPAWWKNIFGGECTVLSV DHVENVNMKLLEGF	[Arabidopsis lyrata subs	p. lyrata]	
>gil152391821refINP_201386.11 F-box protein [Arabido MRTLRRNVTENRLTISRRRTEKKTSPNKTEKSVQIPVDIIIEILLRLPAKSI FTELFLTRSLHRPQLLFCCKKDGNLFFFSSPQLQNPYENSSAISLKNFSLCY ETVTVICNPSTGHTLSLPKPMKTSIGPSRFFVYEPIQKQFKVLLSYKSDEHQ PHILGMSEICINGVLYYPAINLSSGDYIIVCFDVRSEKFRFITVMEEFIKAA RYCFVDGRSKSIELWVLQDAEKKEWSKHTYVLPAWWQHRIGTLNLRFVGVTR YFNIERKTMMSVAIQGMEAFQGHLVFTYLDHVENVKLLHNMF	DSIS THALIANA] ATCRCVSKLWISVICRQD KISRPVNGLICFKRKEMN VLTLGTGELSWRIIECSM HDGTLINYNGKLASLVSE TNEIMLSPCYQTVPFDVY			
>gil152295531refINP_189038.11 putative F-box protein MRSRQLHNVSEDRETLSRRNKRSKTSLNGHIPIDLLIEIFLKLPVKSIATCR FLTKSSSRPQLLFACANDNGYFFFSSNQPQNLDENSSPIAAYPLTHVPKSRD GRIRPVDVSIIYNPSTGESLTLPKTNMTRKKIYTVTSFLGYDPIEKQYKVLS LGTGKLSWRMIKCCLNYQHPLKNSEICINGVLYYLAMVNGSSWPTRAVVCFD TTLINYNNGKLGMLMGQEAHKTISGICRSFELWVLEDTVKHEWSKHVYLLPP TSEIVLFRPDEPLCVFYYNIDRNTIKRVGIRGLEAFKYFRIFLNHVENVKLF	[Arabidopsis thalian SVSKFWTYVLGRQDFTEL LGPPINGLVSLRGERILK MNMSYEKHPKCEGYQVLT IRSEMFNFMEVYRELSYT LWKDAVANTRLYFAGMIG	ia]		
>gil297819588 reflXP_002877677.1 hypothetical prote MSTMMKKRKRHVSKEDVALTISSSLGEYGENSGTLPMDLMVEILSRVPAKSA FTNLYLTRSPTRPRLLITFQAEGKWSFFSSPEYLISDQNSNLVVVDNHMDVPI DEWVLSRKKDARMMICNPSTRQFQSLPKVRSRRNKVITYIGYDPIEKEYKVL LTLGTGKLKWRMLKCFVEHFPHHKEICINGVLYYLAVKDETREDIIVCFHVK YNGKLGGIRHGFMEGGVAGYELWDLDIEKEDWTRHIHILPPMWKQVVGETRV SNPFYIFHLNIERNSITRVEIQGTGPLEGQQVYTFINHIENVKLIM	in ARALYDRAFT_906230 AKFHCVSKNWNSLLRSSY KDYSFGVCEPVCGLLCTR CMTICERPYMFKAEEHQV HEKFQFILNKAPLSTLIN YVVGMIGTSEIVFSPFVK	[Arabidopsis lyrata subs	p. lyrata]	



ClustalW2 - Mu	Itiple Sequence Alignment	
lustalW2 is a gene	ral purpose multiple sequence alignment program for DNA or proteins.	
se this tool		
STEP 1 - Enter yo	ur input sequences	
Enter or paste a se	t of Protein 🔷 sequences in any supported format:	
SUBSP. Jyratal MSTMMKKRKRHVS FTNLYLTRSPTRPRL DEWVLSRKKDARM LTLGTGKLKWRMLK YNGKLGGIRHGFME SNPFYIFHLNIERNSI	KEDVALTISSSLGEYGENSGTLPMDLMVEILSRVPAKSAAKFHCVSKNWNSLLRSSY LITFQAEGKWSFFSSPEYLISDQNSNLVVVDNHMDVPKDYSFGVCEPVCGLLCTR MICNPSTRQFQSLPKVRSRRNKVITYIGYDPIEKEYKVLCMTICERPYMFKAEEHQV CFVEHFPHHKEICINGVLYYLAVKDETREDIIVCFHVKHEKFQFILNKAPLSTLIN GGVAGYELWDLDIEKEDWTRHIHILPPMWKQVVGETRVYVVGMIGTSEIVFSPFVK TRVEIQGTGPLEGQQVYTFINHIENVKLIM	
Or, upload a file: (Choose File No file chosen	
STEP 2 - Set your Alignment Type:	●Slow ○Fast	
The default setting	s will fulfill the needs of most users and, for that reason, are not visible.	
More options	(Click here, if you want to view or change the default settings.)	
STEP 3 - Set your	Multiple Sequence Alignment Options	
The default setting	s will fulfill the needs of most users and, for that reason, are not visible.	
More options	(Click here, if you want to view or change the default settings.)	
STEP 4 - Submit v	our job	
orer 4 outlinty	mail (Tick this box if you want to be notified by email when the results are available)	
Be notified by e		



EBI > Tools > N	Multiple Sequence Alig	nment > Clusta	IW2						
ClustalW2	Results								
Alignments	Result Summary	Guide Tree	Submissi	on Details	Submit	Another Jo	b		
Alignment									
Download Al	ignment File Sho	w Colors							
CLUSTAL 2.1	multiple seque	nce alignme	nt						
gi 15225089	ref NP_180705.	1 -	MKSR-R	ONVSVARQT	ILGRDE-	WTONN-	-NFEPIPID	LVIEIFSR	36
gi 15239182	ref NP 201386.	1 -	MRTL-R	RNVTENRLT	TSRRTE	KTSPNKT	KSVOTPVD	DITELLR	46
gi 15229553	ref NP 189038.	1 -	MRSRQL	INVSEDRET	LSRRNKR	SKTSLN	GHIPID	LLIEIFLK	42
gi 29781958	8 ref XP_002877	677. M	STMMKKRK	RHVSKEDVA	LTISSSL	GEYGEN	SGTLPMD	LMVEILSR	46
			* •	*	:		:*.*		
gi 15225089	ref NP_180705.	1 S	PVKSIARC	RCVSKLWAS	ILRLPYF	PELYLTKS	CARPRLLFA	COKHRELF	86
gi 29783722	9 ref XP_002886	496. L	PAKSISRF	RCVLKLWGS	ILRLPYF	TELFLTRSI	LARPQLLFA	CHKDNHVF	93
gi 15239182	ref NP_201386.	1 I	PAKSIATC	RCVSKLWIS	VICRODF	PELFLTRSI	LHRPQLLFO	CKKDGNLF	96
gi 15229553	ref NP_189038.	1 L	PVKSIATC	RSVSKFWTY	VLGRQDF	PELFLTKS:	SSRPQLLFA	CANDNGYF	92
g1 29/81928	8 rer XP_0028//	6//. V	*.** :	:.* * *	:: *	*:*:**:*	**:**:		90
ai 15225089	ref NP 180705.	11 12	FSTPOPHN	PNESSSP	TAASFHM	TPEDG-RI	NTTSPTCO	UVEVEVEO	133
gi 29783722	9 ref XP 002886	496. V	FSSPOPON	IDDNNASSI	LAANYHM	KIPFYASSI	FERCSSVRO	LVFFGDER	143
gi 15239182	ref NP 201386.	1 F	FSSPQLQN	PYENSSAIS	LKN	FSLC	KISRPVNO	LICFK	135
gi 15229553	ref NP_189038.	1 F	FSSNOPON	LDENSSPIA	AYP	-LTHVPKSH	RDLGPPING	LVSLRGER	137
gi 29781958	8 ref XP_002877	677. F	FSSPEYLI	SDQNSNLVV	VDNHI	DVPKDYSI	GVCEPVCG	LLCTRDEW	143
			**: :	*.			.: *	*:	
gi 15225089	ref NP 180705.	1 І	LKGRKTPE	FVSAICNPS	TGOSLTL	PKPK-TRK	RIW-GTSHP	GYDPIEKO	181
gi 29783722	9 ref XP_002886	496. Y	SNGKEH	KVSVICNPS	TROSLTLI	PKLK-TRK	RIG-VRSYP	GFEPIEKQ	189
gi 15239182	ref NP_201386.	1 -	RKEMNE	IVTVICNPS	TGHTLSL	PKPMKTS	SIG-PSRFF	VYEPIQKQ	179
gi 15229553	ref NP_189038.	1 І	LKGRIRPV	DVSIIYNPS	TGESLTL	PKTNMTRKI	XIYTVTSFI	GYDPIEKQ	187
gi 29781958	8 ref XP_002877	677. V	LSRKKD	ARMMICNPS	TROFOSLI	PKVRSRRN	(VITYI	GYDPIEKE	188
			:	* ***	* . :**	** :.		**:*:	



ClustalW2	Results				
Alignments	Result Summary	Guide Tree	Submission Details	Submit Another Job	
Guide Tree					
Download G	uide Tree File				
(gi 15225089	ref NP 180705.	11.0 21002			
gi 29783722 :0.01981, gi 15239182 :0.01344, gi 15229553 gi 29781958 Phylogram	9 ref XP_002886 ref NP_201386. ref NP_189038. 8 ref XP_002877	1 :0.21903, 496.1 :0.220 1 :0.26252) 1 :0.25932, 677.1 :0.329	024) 906);		
gi 29783722 :0.01981, gi 15239182 :0.01344, gi 15229553 gi 29781958 Phylogram Show as Cla	9 ref XP_002886 ref NP_201386. ref NP_189038. 8 ref XP_002877	496.1 :0.22 1 :0.26252) 1 :0.25932, 677.1 :0.32	024) 906);		



• ClustalW (EBI): <u>http://www.ebi.ac.uk/Tools/msa/clustalw2/</u>

ClustalW2	Results			
Alignments	Result Summary	Guide Tree	Submission Details	Submit Another Job
lignment				
Download A	lignment File Sho	ow Colors		

The alignment can be downloaded to be used by phylogenetic programs like Protpars (from Phylip package).

	Protein parsimony algorithm, version 3.
phylip 3.67: protpars	
Protein Sequence Parcimony Method ?	One most parsimonious tree found:
Alignment File 2 [use example data] EDT CLEAR Enter your data below: EDCMCAFKY NEDYHELDHY EDVKU DOCCOFAFKY NEDYHELDHY EDVKU DOCCOFAFKY INSVITTEDHY ENVKULISME - DOCCOFAFKY INSVITTEDHY ENVKULISME - GRELLARY FRIFLNHY DNYKLF DOCCOFAFKY FRIFLNHY DNYKLF DOCCOFAFKY INSVITTEDHY ENVKULISME - GRELLARY FRIFLNHY DNYKLF GRELLARY FRIFLNHY DNYKLF DOCCOFAFKY INSVITTEDHY ENVKULISME - GRELLARY FRIFLNHY DNYKLF GRELLARY FRIFLNHY DNYKLF DOCCOFAFKY INSVITTEDHY ENVKULISME - GRELLARY FRIFLNHY DNYKLF GRELLARY FRIFLNHY DNYKLF DOCCOFAFKY INSVITTEDHY ENVKULISME - GRELLARY FRIFLNHY DNYKLF	+ArFbox2 +3 1 +AtFbox1 +2 1 1 +ARALY_9062 1 +4 1 +4 1 +ARALY_8932 1 +AtDOR remember: this is an unrooted tree!
Web-based Phylip package:	requires a total of 1060.000
http://mobyle.pasteur.fr/cgi-bin/portal.py?#welcome	



Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).
 - I.3 By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

- 2.1 Translators and Gene Predictors.
- 2.2 Multiple Sequence Alignment (Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

- 3.1 Linkage Map Viewers (CViewer).
- 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.



Some of the functional annotations are made by homology search with conserved protein fragments or **domains**.

InterPro (<u>http://www.ebi.ac.uk/interpro/</u>) is an EBI resource with several protein domain databases such as *ProSite*, *Pfam* or *Superfamily*.



The tools used for functional domain search is InterProScan (<u>http://www.ebi.ac.uk/Tools/pfa/iprscan/</u>).

2.3 - Protein Domain Analysis



InterProScan (<u>http://www.ebi.ac.uk/Tools/pfa/iprscan/</u>).

BI > Tools > Protein F	Functional Analysis > I	nterProScan		
InterProScan Se	equence Search			
This form allows you locumentation for the nanual or help page Jse this tool	to query your seque e perl stand-alone in s.	nce against InterPro. F terProScan package (For more detailed info Readme file or FAQ	ormation see the s), or the InterPro user
STEP 1 - Enter you	r innut seguence			
Enter or paste a PR	OTFIN sequence in	any supported format:		
PEFVSAICNPSTGQSL RIECSIPHVHGSKGIC EGPSYICGKRFEMWV YNLERNTIRRVEIQGM Or, upload a file:	TLPKPKTRKRIWGTSH INGVLYYRAKADMFSO LGDPEKHEWLKHTYEL MGAFKVNEDYIFLDHV Choose File No file	FGYDPIEKQFKVLSMNIG TLMIVCFDVRFEKFSYIK RPRWQNVLGEDLLIFAG EDVKLI	GDGVYKEHYVLTLGTEI ILKPTTTLISYNGKLASI MTGTNEIVLSPKYPSH	NLSWR LVW PFYVFY
STEP 2 - Select the	applications to run			
Select All Clear	All			
 BlastProDom HMMTigr SignalPHMM 	 ✓ FPrintScan ✓ ProfileScan ✓ TMHMM 	 ✓ HMMPIR ✓ HAMAP ✓ HMMPanther 	 ✓ HMMPfam ✓ PatternScan ✓ Gene3D 	✓ HMMSmart ✓ SuperFamily
CTED 2 Cubmiture	uur lah			
Be potified by on	our job	ou want to be notified	by omail when the m	eulte are available)
Be notilied by en	ian (nok uns box if y	ou want to be notified	by email when the fe	suns are available)
		Submit		

2.3 - Protein Domain Analysis



InterProScan (<u>http://www.ebi.ac.uk/Tools/pfa/iprscan/</u>).

	Results				
Summary Table	Tool Output Visual O	utput Submission Details	Submit Another Job		
terProScan Vi	isual Output				
Download in SV	G format				
nterProScan (ve equence: NP_1807 angth: 387 RC64: ABA57A26	rsion: 4.8) 705.1 79C00184			Launched Mon, Jun 13 Finished Mon, Jun 13	, 2011 at 18:04:2 , 2011 at 18:04:5
InterPro Match	.h	Query S	equence	Bescription 387	
IPR001810 PF00 SM00	F-box domain, cyclin	-like		F-box FBOX	
IPR013187 PF08	F-box associated dor	main, type 3		FBA_3	
IPR017451 TIGR01	F-box associated inte 640 •	eraction domain		F_box_assoc_1	
	E-box domain Sko2	-like			
IPR022364 SSF81	1383			F-box_dom_Skp2-like	
IPR022364 SSF81 noIPR G3DSA:1.20.1280 PTHR116 PTHR11603:S	unintegrated			G3DSA:1.20.1280.50 PTHR11603 PTHR11603:SF79	



Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).
 - I.3 By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

- 2.1 Translators and Gene Predictors.
- 2.2 Multiple Sequence Alignment (Clustalw).
- 2.3 Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

- 3.1 Linkage Map Viewers (CViewer).
- 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.

2.4 - Signal Peptide Analysis



A signal peptide is a short (3-60 amino acids long) peptide chain that directs the transport of a protein. Signal peptides may also be called targeting signals, signal sequences, transit peptides, or localization signals. (wikipedia).

Examples:	Transport to the nucleus (NLS)	-Pro-Pro-Lys-Lys-Lys-Arg-Lys-Val-
	Transport to the endoplasmic retice	ulum H ₂ N-Met-Met-Ser-Phe-Val-Ser-Leu- Leu-Leu-Val-Gly-Ile-Leu-Phe- Trp-Ala-Thr-Glu-Ala-Glu-Gln- Leu-Thr-Lys-Cys-Glu-Val-Phe- Gln-
	Retention to the endoplasmic retice	ulum -Lys-Asp-Glu-Leu-COOH
	Transport to the mitochondrial mat	rix H ₂ N-Met-Leu-Ser-Leu-Arg-Gln-Ser- Ile-Arg-Phe-Phe-Lys-Pro-Ala- Thr-Arg-Thr-Leu-Cys-Ser-Ser- Arg-Tyr-Leu-Leu-
	Transport to the peroxisome (PTS1)	-Ser-Lys-Leu-COOH
	Transport to the peroxisome (PTS2)	H ₂ NArg-Leu-X ₅ -His-Leu-

SignalP (<u>http://www.cbs.dtu.dk/services/SignalP/</u>) is a program to predict signal peptides.



Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).
 - I.3 By Sequence/Chromosome coordinates (GBrowse).
- 2 Manipulation and Sequence Analysis Tools:
 - 2.1 Translators and Gene Predictors.
 - 2.2 Multiple Sequence Alignment (Clustalw).
 - 2.3 Protein Domain Analysis (InterProScan).
 - 2.4 Signal Peptide Analysis (SignalP).
- 3 Other Tools:
 - 3.1 Linkage Map Viewers (CViewer).
 - 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.

3.2 - Primer Design.



There are some web-based tools to design primers or to check the possible amplify fragments with the primers designed.

• Primer-Blast (NCBI) (design):

http://www.ncbi.nlm.nih.gov/tools/primer-blast/

• Primer3 (design):

http://frodo.wi.mit.edu/primer3/

• In-Silico PCR (SGN) (fragment analysis):

http://solgenomics.net/tools/insilicopcr/index.pl

3.2 - Primer Design.



• Primer3 (design): <u>http://frodo.wi.mit.edu/primer3/</u>

Copy the downloaded sequence to Primer3. Change min. size to I 23 pb (intron size) Change target to 200 (intron start), I 23 (intron length)

Drimer?	Checks for mispriming in template,	disclaimer	Primer3 Home
FIIIIEIS (v. 0.4.0) Pick primers from a DNA sequence.	Primer3plus interface	cautions	FAQ/WIKI
Paste source sequence below (5'->3', string of ACGTNacgtn other letters treated as N number ALUs, LINEs, etc.) or use a Mispriming Library (repeat library): NONE	bers and blanks ignored). FASTA format ok. Please	N-out undesiral	ble sequence (vector,
AACGTCAATGAATAGATAGATGGCTGCCGCGGCAATCCAAAGTTCCCCGGCTGCTTCCCG CCACCACCACTTCCACCCTGACGGGGCTCATTACCAAAGTTCTTGAAATGATAATTA CTCCCCATTTCACTAAAACTCCTCAGTCCTCACACAAATCCGCCTTCAAACCTCAGCTCTG TTATTCAAGAATCACAAAACCTACATATCAGATCAACAAGTTAATTCCCTTCCCTTTGAA CCTTTTTCCTTATCATACTGTTCAACCCTTCACACAAGATCAACTCTATTTACAAACACA CAGTTAATTAAAAGCAAAATATACCTGGAAAGAGATCAAAAATCAATTTACAGCTAAAAC			
Pick left primer, or use left primer below:	e oligo below: Pick right primer, or use right primer	ner below (5' to	3' on opposite strand)
Pick Primers Reset Form Sequence Id: A string to identify your output. Targets: 200,123 E.g. 50,2 requires primers to surround the 2 bases at position	ons 50 and 51. Or mark the source sequence with [a	nd]: e.gATC	T[CCCC]TCAT
Excluded E.g. 401,7 68,3 forbids selection of primers in the 7 bases s ATCT <cccc>TCAT forbids primers in the central CCCC</cccc>	tarting at 401 and the 3 bases at 68. Or mark the sou	irce sequence wi	th < and >: e.g.
Product Size Ranges 123-223			
Number To Return 5 Max 3' Stability 9.0 Max Repeat Mispriming 12.00 Pair Max Repeat Mispriming 24.00 Max Template Mispriming 12.00 Pair Max Template Mispriming 24.00			

3.2 - Primer Design.



• Primer3 (design): <u>http://frodo.wi.mit.edu/primer3/</u>

<<<<< right primer

```
No mispriming library specified
Using 1-based sequence positions
OLIGO
              <u>start len</u>
                           tm gc% any 3' seg
LEFT PRIMER
              157 19 60.20 52.63 3.00 3.00 ATCCGCCTTCAAACCTCAG
                373 21 59.51 47.62 2.00 2.00 AAGGGGTTGGTGAGTTTTAGC
RIGHT PRIMER
SEQUENCE SIZE: 524
INCLUDED REGION SIZE: 524
PRODUCT SIZE: 217, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 3.00
TARGETS (start, len)*: 200,123
   1 AACGTCAATGAATAGATAGATGGCTGCCGCGGCAATCCAAAGTTCCCCGGCTGCTTCCCG
  61 CCACCACCACCTCCACCTCGCTGGCTCATTACCAAAGTTCTTGAAATGATAATTA
 121 CTCCCCATTTCACTAAAACTCCTCAGTCCTCACACAATCCGCCTTCAAACCTCAGCTCTG
                                    181 TTATTCAAGAATCACAAAACCTACATATCAGATCAACAAGTTAATTCCCTTCCCTTTGAA
                      241 CCTTTTTCCTTATCATACTGTTCAACCCTTCACATAAATGTACATCTATTTACAAACACA
     301 CAGTTAATTAAAAGCAAAATATACCTGGAAAGAGATCAAAAATCAATTTACAGCTAAAAC
     ********************
                                                   <<<<<<
 361 TCACCAACCCCTTATCAATAAAATCATCAAAAAACAAATCCTATTTGAAATTCACTTCATT
     <<<<<<<
 421 CAACTAAATTGACTGCATTTTCAGTTCACCCCAAGAACCCCCCAAAACCACCTTCCCCAC
 481 CCACCAATCCAATAAAGAACACACCTTTTGACCTTCAAATACAC
KEYS (in order of precedence):
****** target
>>>>> left primer
```



Bioinformatic Web Tools:

- I Search Tools:
 - I.I By Ontology.
 - I.2 By Sequence Homology/Similarity (Blast).
 - I.3 By Sequence/Chromosome coordinates (GBrowse).
- 2 Manipulation and Sequence Analysis Tools:
 - 2.1 Translators and Gene Predictors.
 - 2.2 Multiple Sequence Alignment (Clustalw).
 - 2.3 Protein Domain Analysis (InterProScan).
 - 2.4 Signal Peptide Analysis (SignalP).
- 3 Other Tools:
 - 3.1 Linkage Map Viewers (CViewer).
 - 3.2 Primer Design (Primer3).
- 4 Web Pages with Multiple Tools.

4 - Web Pages with Multiple Tools.



Useful bioinformatic web-portals with classical bioinformatic tools on-line:

• EBI (European Bioinformatic Institute): Analysis of sequences.

http://www.ebi.ac.uk/Tools/

• Mobyle (Instituto Pasteur): Phylogenetic analysis.

http://mobyle.pasteur.fr/cgi-bin/portal.py?#welcome

- ExPASy (SwissProt): Analysis of proteins and sequences.
 http://expasy.org/tools/
- CBS (Center For Biological Sequence Analysis).

http://www.cbs.dtu.dk/biotools/

• Phylemon2: Molecular evolution analysis

http://phylemon.bioinfo.cipf.es/evolutionary.html



Exercise 2

- I. Select a protein from exercise I part 5, what domains can be found?
- 2. Find the Arabidopsis thaliana best protein match to the protein.
- 3. Find the tomato best protein match to the protein
- 4. What sequences are upstream and downstream of the tomato match from part 2? How many introns does the match have?
- 5. Align all sequences from exercise 1.4 with the Arabidopsis and tomato protein matches.
- 6. Make a phylogenetic tree with the alignment from 5. Which sequences appear to be most closely related?



Exercise 2 Solutions (cont'd)

I. I used the gene from *C. arabica*: interpro scan (<u>http://www.ebi.ac.uk/interpro/</u>) : SAM

Protein family membership

SAM dependent carboxyl methyltransferase (IPR005299)

Domains and repeats

None predicted.

Detailed signature matches





Exercise 2 Solutions

2. At5g04380 (http://arabidopsis.org/Blast/index.jsp)

3. Solyc04g05560 (<u>http://solgenomics.net/tools/blast/index.pl</u>)



Exercise 2 Solutions



http://solgenomics.net/gbrowse/bin/gbrowse/ITAG2.3_genomic/

Exercise 2 Solutions

CLUSTAL O(1.2.1) multiple sequence alignment

MDMKDVLCMNTGEGESSYLLNSKFTNVTAIKSIPT
MEVKEMLFMNKGDGENSYVKTSGYTQKVAAVTQPV
MELATAGKVNEVLFMNRGEGESSYAQNSSFTQQVASMAQPA
MELATAGKVNEVLFMNRGEGESSYAQNSSFTQQVASMAQPA
MKEVKEALFMNKGEGESSYAQNSSFTQTVTSMTMPV
MKEVKEALFMNKGEGESSYAQNSSFTQTVTSMTMPV
MELQEVLHMNGGEGEASYAKNSSFNQLVLAKVKPV
MSLCLILCRCDCKSEYKVDEERSSKYPFVGALCMNGGDVDNSYTTKSLLQKRVLSITNPI
MEVTKVLHMNGGMGDASYAKNSLLQQKVILMTKSI

* ** * : ** .* : .

LKRAIESLFKEESPPFEHLLNVADLGCASGSTSNTIMPTVVQTVVNKCRE--LNHKIPEF VYRAAQSLFTGRNSCSYQVLNVADLGCSSGPNTFTVMSTVIESTRDKCSE--LNWQMPEI LENAVETLFSR-DFHL-QALNAADLGCAAGPNTFAVISTIKRMMEKKCRE--LNCQTLEL LENAVETLFSR-DFHL-QALNAADLGCAAGPNTFAVISTIKRMMEKKCRE--LNCQTLEL LENAVETLFSK-DFHLLQALNAVDLGCAAGPTTFTVISTIKRMVEKKCRE--LNCQTLEL LENAVETLFSK-DFHLLQALNAVDLGCAAGPTTFTVISTIKRMMEKKCRE--LNCQTLEL LEQCVRELLRANLPNINKCIKVADLGCASGPNTLLTVWDTVQSIDKVKQEMKNELERPTI LVKNTEEMLTN--LDFPKCIKVADLGCSSGQNTFLAMSEIVNTINVLCQK--WNQSRPEI TDEAISSLYNN--LSSRETICIADLGCSSGPNTFLSVSQFIQTIDKERKKK-GRHKAPEF

gi|645065978|tpg|DAA64605.1| gi|87887929|dbj|BAE79730.1| gi|145952324|gb|ABP98983.1| gi|9967143|dbj|BAB12278.1| gi|59611829|gb|AAW88351.1| gi|51968288|dbj|BAD42854.1| gi|13365694|dbj|BAB39213.1| At5g04380 Solyc04g055260.2.1

gi | 645065978 | tpg | DAA64605.1 |

gi|87887929|dbj|BAE79730.1|

gi | 145952324 | gb | ABP98983.1 |

gi|9967143|dbj|BAB12278.1|

gi|59611829|gb|AAW88351.1| gi|51968288|dbj|BAD42854.1|

gi|13365694|dbj|BAB39213.1|

At5g04380

Solyc04g055260.2.1

http://www.ebi.ac.uk/Tools/msa/clustalw2/



sol genomics network

6.

Phylogram

Branch length: O Cladogram O Real



gi|645065978|tpg|DAA64605.1| 0.26164Pagi|87887929|dbj|BAE79730.1| 0.22594Thegi|145952324|gb|ABP98983.1| 0.00247Cagi|9967143|dbj|BAB12278.1| 0.00566Cagi|59611829|gb|AAW88351.1| 0.00778Cagi|51968288|dbj|BAD42854.1| 0.00592Cagi|13365694|dbj|BAB39213.1| 0.28947CaAt5g04380 0.28929AraSolyc04g055260.2.1 0.27805Ta

Paullinia Theobroma Camellia Camellia Camellia Camellia Coffea Arabidopsis Tomato



When using web tools remember:

I.) Often not all program options are available

2.) Jobs are run on another server, large jobs may be better run locally

Additional Bioinformatics Classes

- I. slides at ftp://ftp.solgenomics.net/bioinfo_class/interns/2015/
- 2. If you are interested in hands on command line training, come to the advanced bioinformatics sessions:
 - I. Introduction to Unix Command-line Noe Fernandez (6/25)
 - 2. Mapping NGS Data Aimin Yan (7/2/15)
 - 3. SNP calling from NGS Data Naama Menda (7/9)
- 3. Sign-up: email <u>srs57@cornell.edu</u>
- 4. You will need to have a virtual machine installed prior to next class

<u>https://btiplantbioinfocourse.wordpress.com/how-to/set-a-virtual-machine-using-vm-virtualbox/</u>