

Introduction to Bioinformatics

The field of science in which **biology**, **computer science** and **information technology** merge into a single discipline

Biologists

collect molecular data:
DNA & Protein sequences,
gene expression, etc.

Bioinformaticians

Study biological questions by
analyzing molecular data

Computer scientists

(+Mathematicians, Statisticians, etc.)
Develop tools, softwares, algorithms
to store and analyze the data.

Bioinformatics

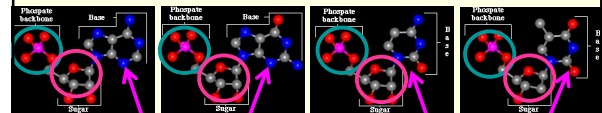
Bioinformatics is rapidly developing and growing field in current era. It includes the computational analysis of biological data, consisting of the information stored in the form of DNA, Protein and Genome sequences in various biological databases.

The branch of science concerned with information and information flow in biological systems, specially the use of computational methods in genetics and genomics.

The hereditary information of all living organisms, with the exception of some viruses, is carried by **deoxyribonucleic acid (DNA)** molecules.

2 purines:

2 pyrimidines:



adenine (A) guanine (G) cytosine (C) thymine (T)

two rings

one ring

Nucleic acids

Principle information molecule in the cell

DNA → RNA → Protein

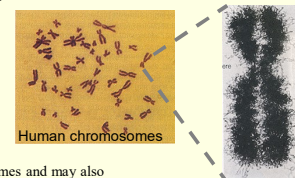
- All cells have a nucleus (Eukaryotic cell) or a nucleoid (Prokaryotic cell), in which the genome (the complete set of genes, composed of DNA) is stored and replicated.
- A segment of a DNA molecule that contains the information required for the synthesis of a functional biological product, whether protein or RNA, is referred to as a gene. A cell typically has many thousands of genes. The storage and transmission of biological information are the known functions of DNA.
- Structurally, nucleic acid is a linear polymer of nucleotides

Genome is an organism's complete set of DNA, including all of its genes.

The complete set of genes or genetic material present in an organism.

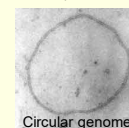
Eukaryotes may have up to 3 sub-cellular genomes:

- Nuclear
- Mitochondrial
- Plastid



Human chromosomes

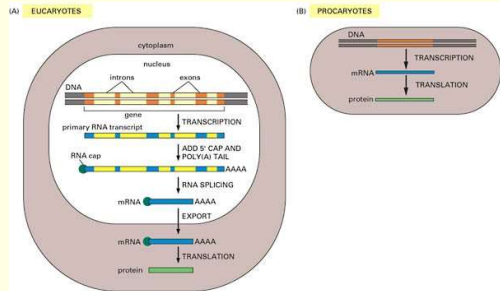
Bacteria have either circular or linear genomes and may also carry plasmids



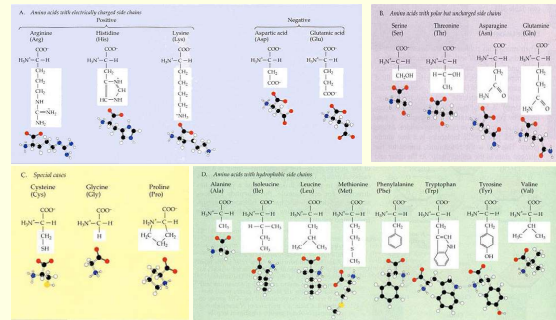
Circular genome

Central dogma: Sequential transfer of information from DNA to Protein via RNA (DNA makes RNA and RNA makes protein)

Modified central dogma: DNA makes RNA, RNA makes DNA via reverse transcription, RNA makes Protein



Amino acids - The protein building blocks



8

Genetic code: TRANSLATION

DNA-triplet → RNA-triplet = codon → amino acid

RNA codon table

There are **20 standard amino acids** used in proteins, here are some of the RNA-codons that code for each amino acid.

- Ala** A GCU, GCC, GCA, GCG
- Leu** L UUA, UUG, CUU, CUC, CUA, CUG
- Arg** R CGU, CGC, CGA, CGG, AGA, AGG
- Lys** K AAA, AAG
- Asn** N AAU, AAC
- Met** M AUG
- Asp** D GAU, GAC
- Phe** F UUU, UUC
- Cys** C UGU, UGC
- Pro** P CCU, CCC, CCA, CCG

- ... AUG, GUG
- Start** AUG, GUG
- Stop** UAG, UGA, UAA

Table of Amino Acids and Their Abbreviations

Full Name	Abbreviation (3 Letter)	Abbreviation (1 Letter)
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartate	Asp	D
Aspartate or Asparagine	Asx	B
Cysteine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glutamate or Glutamine	Glx	Z
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

A cDNA sequence

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA
ACTCTTCGGTCCCACAGACTCAGAGAGACCACCACTGGTGTCTCTCCGCGACAAGCAACCTCAAGCCG
CCTGGGTAAGTTCGGCCGACCGCTGGCGAGTATGGTGGGAGGCCCTGGAGAGGATCTTCCTGCTCCACC
ACCAAGACTACTTCCCGACTTCGACTGAGCCACGGCTCTGCCAGGTAAAGGCCACGGCAAGAGGTGGCGA
CGCCGTGACCAACCGGTGGCGACGTGAGCACATGCCCAACGGCTGTCGCGCTGAGCGAAGCTGACGGGACA
AGCTTCGGTGGACCGGTCAACTTCAAGCTCCTAAGCCACTGCTGCTGAGCCTGGCGGCCCACTCCCGCC
GAGTTCACCCCTGGGTGACGCTCCCTGGACAAGTTCCTGCTGCTGAGCAGCCTGCTGACTCCAAATACGG
TTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGGCCCTTGGGCTCCGCCAGCCCTCCCTCCCTCTCTGACACCGT
ACCCCTCGGTCTTTGAATAAAGTCTGAGTGGCGGC
```

A cDNA sequence (reading frame)

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA
ACTCTTCGGTCCCACAGACTCAGAGAGACCACCACTGGTGTCTCTCCGCGACAAGCAACCTCAAGCCG
GCCTGGGTAAGTTCGGCCGACCGCTGGCGAGTATGGTGGGAGGCCCTGGAGAGGATGTTCTGCTCCACC
CACCAAGACTACTTCCCGACTTCGACTGAGCCACGGCTCTGCCAGGTAAAGGCCACGGCAAGAGGTGGCG
ACCCGTGACCAACCGGTGGCGACGTGAGCACATGCCCAACCGCTGTCGCGCTGAGCGAAGCTGACGGGACA
AAGCTTCGGTGGACCGGTCAACTTCAAGCTCCTAAGCCACTGCTGCTGAGCCTGGCGGCCCACTCCCGCC
CGAGTTCACCCCTGGGTGACGCTCCCTGGACAAGTTCCTGCTGCTGAGCAGCCTGCTGACTCCAAATACCG
GTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGGCCCTTGGGCTCCGCCAGCCCTCCCTCCCTCTCTGACACCGT
GTACCCCTCGGTCTTTGAATAAAGTCTGAGTGGCGGC
```

A protein sequence

```
>gi|4504347|ref|NP_000549.1| alpha 1 globin [Homo sapiens]
MVLSPADKTNVKAAGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDSLHGSQVKGHKVADALNTVAHV
VDDMPNLALSIDLHAHLKRVDFVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVTLVTSKYLK
```

And, a whole genome...

```

ACTTCTTGGTCCCAAGACTCAGAGAAACCCACATGGTGTCTCTTCCGACAGACCAAGCTCAAGCCCGCTGGGGT
AAGTGGCGCGCAGCTGGGGAGTATGGTGGAGGCCCTGGAGAGATGTTCTGTCTTCCACCAACCAAGACTACTTCC
CGACTTGGACTGAGCCACGGCTCTCCAGGTTAAGGGCCACGGCAAGAAAGTGGCCGACGGCTGACCAAGCCCTGGGCA
CGTGGACGATGCCAACGGCTGTCCGCTGAGGACCTGACAGCGCAAGCTTGGGTGGACCGGCTCAACTCAAGCTTC
CTAAGCCACTGGTGGTGGACTGGGCGCCGACCTCCCGCGAGTTACCTGGGGTGGAGGCTCTCTGGACAAGTTC
TGGCTCTGGAGCACCGTGGACTCAAAATACCTTAACTGGAGCCTCGTGGCCATGCTTCTTGGCCCTGGGGCTCCCC
CGACCCCTCTCCCTTCTCGAACCCGTACCCCGTGTCTTGAATAAGTCTGAGTGGGGGGACTCTTCTGGTCCCAAG
ACTGAGAGAAACCACTGGTGTCTCTTCCGACAAAGCAAGCTCAAGGCCGCTGGGGTAAAGTGGGGGGGAGACTG
GGAGTATGGTGGAGGCGCTGGAGAGAGTGTCTCTGCTCTCCCAACCAAGACTACTTCCGGACTTCCGACTGGCCA
CGCTCTGCCAGGTTAAGGGCCACGGCAAGAGGTGGCCGACGGCTGACCAAGCCCTGGGCGACTGGAGGACATGCCAAC
GGCTGTCCGCTGAGGACCTGACGGCACAAGCTTGGGTGGACCGGCTCAACTCAAGCTCTAAGCCACTGCTGCTGG
TGACTTGGGCGCCACTCCCGCGAGTTCACTCCCTGGTGCAGCCCTCTGGACAAGTCTGGCTCTTGTGGACAGCTG
GGTGAACCTCAATAACCTTAACTGGAGCCTGGTGGCCATGCTTCTGGCCCTGGGGCTCCCGCCAGCCCTCTCCCTTC
CTGACCCGTACCCCGTGTCTTGAATAAGTCTGAGTGGGGGCACTCTTCTGGTCCCAAGACTCAGAGAGAACCA
TGTGTCTCTCTCGCGCAAGCAACCAAGTCAAGGCCCTGGGGTAAAGTGGGCGCACGCTGGCGAGTATGGTGGAGGC
CTTGGAGAGATGTTCTCTCTCCCAACCAAGACTACTTCCCGACTTGGACTGGAGCGGCTCTGGCGAGTAAAG
GGCCAGGCAAGAGAGTGGCGACGGCTGACCAAGCCGCTGGCGACTGGAGAGACTGCCCAAGCCGCTGTCCGCTGAGCG
ACTGACGGCACAAGCTTGGGTGGACCGGCTCAACTCAAGCTCTAAGCCACTGCTGCTGGTGGACCTGGCGCCACTCCCGCGG
AGTTCACCCCTGGGTGCAGCCTCCCTGGACAAGTCTGGCTTCTGTGGACCGCTGCTGACTCAAAATACCTGTAAGTGG
AGCTGGGTGGCAAGTCTTGGCCCTGGGCTCCCGCGCCCTGCTCCCTTCTGGACCGCTAGCCCGTGGCTTGA
ATAAGTCTGAGTGGGGGCACTTCTGGTCCCAAGACTCAGAGAGAACCAACTGGTGTCTCTCCGACAGACCA
ACGTCAAGGCCCTGGGGTAAAGTGGGCGCACGCTGGCAGTATGGTGGAGGCCCTGGAGAGATGTTCTGTCTTCC
CACCAAGAGACTACTTCCCGCACTTCCACTGAGCCAGCTTGGCCAGGTTAAGGGCCAGGCAAGAGGTGGCG...
    
```

How big are whole genomes?



E. coli 4.6 x 10⁶ nucleotides
 - Approx. 4,000 genes



Yeast 15 x 10⁶ nucleotides
 - Approx. 6,000 genes



Human 3 x 10⁹ nucleotides
 - Approx. 30,000 genes

Smallest human chromosome 50 x 10⁶ nucleotides

What do we actually do with bioinformatics?

Origin of bioinformatics and biological databases:

The first protein sequence reported was that of bovine insulin in **1956**, consisting of 51 residues.

Nearly a decade later, the first nucleic acid sequence was reported, that of yeast tRNA^{alanine} with 77 bases.

Genomic era

- 1958: **Frederick Sanger** (Cambridge, UK): Nobel prize for developing protein sequencing techniques
- 1978: **Frederick Sanger** : First complete viral genome
- 1980: **Frederick Sanger** : First mitochondrial genome
- 1980: **Frederick Sanger** : Nobel prize for developing DNA sequencing techniques
- 1995: **Craig Venter** (TIGR): complete genome of *Haemophilus influenza*
- 2001: Human Genome Project – Sequencing of entire genome of *Homo sapiens*

Complete Genomes in Plants as of January 2020

Algae	43
Bryophytes	2
Pteridophytes	2
Gymnosperm	6
Monocots	29
Dicots	104

Animal genomes	More than 300
Bacteria -	More than 500
Cyanobacteria	17
Fungi	More than 300



databases

1.5 Finding data: GenBank,

- Online databases
- FASTA: a standard data format

FASTA is a DNA and protein sequence alignment software package first described (as FASTP) by David J. Lipman and William R. Pearson in 1985. Its legacy is the FASTA format which is now ubiquitous in bioinformatics.

FASTA – FAST ADAPTIVE SHRINKAGE THRESHOLDING ALGORITHM

Generalized DNA, protein and carbohydrate databases

Primary sequence databases

[GenBank](http://www.ncbi.nlm.nih.gov) (at National Center for Biotechnology information, [NCBI](http://www.ncbi.nlm.nih.gov), Bethesda, MD, USA)

[EMBL](http://www.ebi.ac.uk) (European Molecular Biology Laboratory nucleotide sequence database at [EBI](http://www.ebi.ac.uk), Hinxton, UK)

[DDBJ](http://www.ddbj.nig.ac.jp) (DNA Data Bank Japan at [CIB](http://www.ddbj.nig.ac.jp), Mishima, Japan)



NCBI: National Center for Biotechnology information

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

Database	Count	Database	Count
Literature	0	EST	0
Books	0	Gene	0
BLAST	0	Gene	0
Bookshelf	0	Gene	0
Gene	0	Gene	0
Genome	0	Gene	0
Nucleotide	0	Gene	0
OMIM	0	Gene	0
Protein	0	Gene	0
PubChem	0	Gene	0
PubMed	0	Gene	0
PubMed Central	0	Gene	0
SNP	0	Gene	0
NCBI News	0	Gene	0
Assembly	0	Gene	0
BioCatalysis	0	Gene	0
BioProject	0	Gene	0
BioSample	0	Gene	0
Clustal	0	Gene	0
dbVar	0	Gene	0
Genome	0	Gene	0
Genomes & Maps	0	Gene	0
Genomics & Medicine	0	Gene	0
Genetics & Biocassays	0	Gene	0
Chemicals & Bioassays	0	Gene	0
Data & Software	0	Gene	0
DNA & RNA	0	Gene	0
Domains & Structures	0	Gene	0
Genes & Expression	0	Gene	0
Genetics & Medicine	0	Gene	0
Genomes & Maps	0	Gene	0
Homology	0	Gene	0
Literature	0	Gene	0
Proteins	0	Gene	0
Sequence Analysis	0	Gene	0
Taxonomy	0	Gene	0
Training & Tutorials	0	Gene	0
Variation	0	Gene	0

Generalized DNA, protein and carbohydrate databases

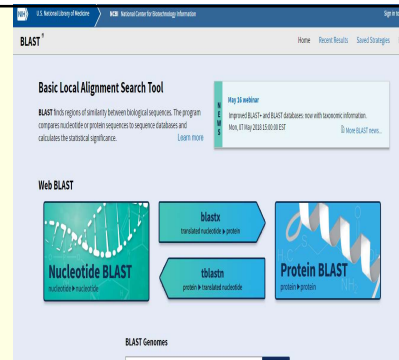
Protein sequence databases

[SWISS-PROT](#) (Swiss Institute of Bioinformatics, [SIB](#), Geneva, CH)
[TrEMBL](#) (=Translated EMBL: computer annotated protein sequence database at [EBI](#), UK)
[PIR-PSD](#) (PIR-International Protein Sequence Database, annotated protein database by PIR, MIPS and JIPIID at NBRF, Georgetown University, USA)
[UniProt](#) (Joined data from Swiss-Prot, TrEMBL and PIR)
[UniRef](#) (UniProt NREF (Non-redundant REFerence) database at [EBI](#), UK)
[IPI](#) (International Protein Index; human, rat and mouse proteome database at [EBI](#), UK)

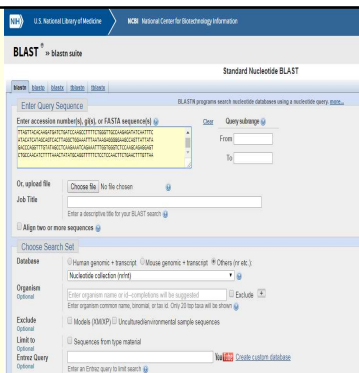
NCBI BLAST

BLAST for Basic Local Alignment Search Tool is a bioinformatics program which widely used in sequence searching and comparing primary biological sequence information, such as the nucleotides of DNA sequences and the amino-acid sequences of proteins. A BLAST search enables a researcher to compare a query sequence with sequences available in a public database like NCBI (National Center for Biotechnology Information).

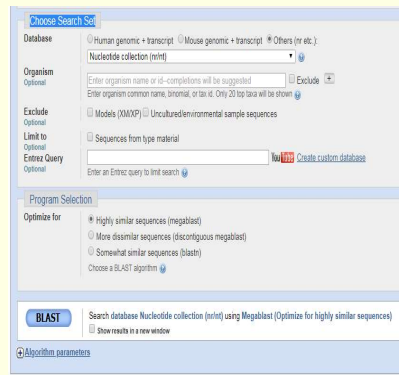
Open homepage of NCBI search database and click on BLAST. If the given sequence is nucleotide sequence, then we will go for Nucleotide BLAST. Enter the given nucleotide sequence (or accession no. or FASTA sequence) in 'Enter Query Sequence box'. Set the database for 'nucleotide collection' and program selected as 'highly similar sequences'. Finally, click on BLAST for sequence search.



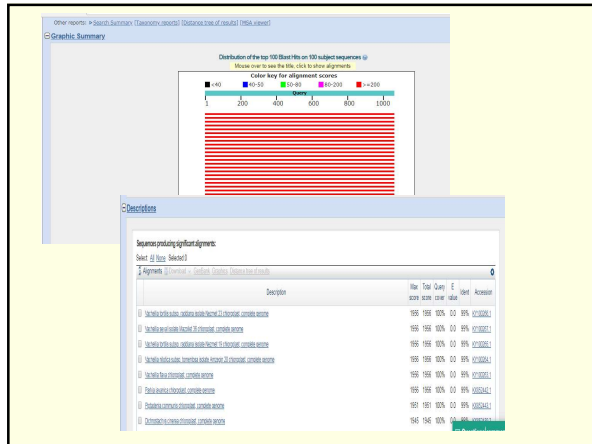
Open NCBI BLAST
Click Nucleotide blast



Paste nucleotide sequences in 'enter Query sequence' and choose search set 'nucleotide collection'.



Click blast



Multiple sequence alignment using CLUSTAL W tools

Multiple sequence alignment was performed by using Clustal W, which is online software that performs optimum alignment for sequence.

Alignment of multiple nucleotides sequences in FASTA format

>KX385911.1 *Acacia auriculiformis* isolate AA1 ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcl) gene, partial cds; chloroplast

```

TTGAC TATTATA CTCTGACTATGAAACCAAGAGATGATATCTGGACGATTCGGAGTAACCTCTCAACCTGGAGTTC
GCTGAAAGAGCAAGTGGCCGGTAGCTGCTGAATCTCTACTGTACATGCAACAACCTGTGTGGACCGATGGCTTACCA
GCTCTGATCGTTACAAAGGACGATGCTACCACATCGATCGCTGCTGGAGAGAAATCAATATATGCTTATGATGCTTAT
CCCTTAGACCTTTTGAAGAAGGTTCTGTACTACATGTTTACTTCGATTTGGGTAATGATTTGGGTCAGGCGCTTC
GGCTCTAGCTGTGGAAGATTGGGATCCCTCTTATTGTAAACCTTCCAAAGTGGCGCTCAAGGGATCAAGTTGA
GAGAGATAAATGAACAAGTACGGCCGCTCCCTTATGGGATGTACTATAACCAAAATGGGGTATCCCGGAAAGATAC
GGTAGAGCGGTTGATGATGCTCCGTTGGTGAATTTATTAACAAGATGATGAAATGTAATGCCAACCAATTATGTC
GTTGGAGAGCCGTTTCTGCTTTGTCGGCAAGCACTTTTAAAGCAGCGCCGAAACAGGTGAAATCAAGGGCATTATGC
TGATGCTACTGC

```

>JX856822.1 *Acacia catechu* isolate 294 ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcl) gene, partial cds;

```

chloroplastAATTGACTTATTATCTGACTATGAAACCAAGGATAGTATGATCTTGGGACGATTCGGAGTAACCTCTCAAC
TGAGTTCGCGCTGAAAGAGCGGTTGCCGGTAGCTGCTGAATCTCTACTGTACATGCAACAACCTGTGTGGACCGATG
GGCTTACCGTCTGATGCTTACAAAGGACGATGCTACCACATCGATCGCTGCTGGAGAGAAATCAATATATGCTTATG
TACCTTACCTTACCTTACCTTGAAGAAGGTTCTGTACTACATGTTTACTTCGATTTGGGTAATGATTTGGGTCAGG
GGCTCTAGCTGTGGAAGATTGGGATCCCTCTTATTGTAAACCTTCCAAAGTGGCGCTCAAGGGATCAAGTTGA
CCAAGTTGAAAGATAAATGAACAAGTACGGCCGCTCCCTTATGGGATGTACTATAACCAAAATGGGGTATCCCGG
AAGAATACGGTAGAGCGGTTTGAATGCTCCGTTGGTGAATTTATTAACAAGATGATGAAATGTAATGCCAACCA
CTTATGGGTTGGAGAGCGGTTTCTTATTTGTCGGCAAGCAATTTATAAGCACAGCGCGAA

```

>JX195517.1 *Acacia ferruginea* ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcl) gene, partial cds;

```

chloroplastGGTGTAAAGATTAAATGACTTATTACTCTGACTATGAAACCAAGGATAGTATGATCTTGGCAGGATCCG
AGTAACTCTCAACCTGGAGTTCGCGCTGAAAGAGCGGTTGCCGGTAGCTGCTGAATCTCTCAACCTGTGTGGAC
CTGCTGGAGCAAGTGGCTTACCAAGTCTGATCGTTAGCAAGAGATGCTACCACATCGATCGCTGCTGGAGAGAA
AGTCAATATTGCTTATGACTTATCCCTAGACCTTTTGAAGAAGTCTGTTACTAACGTTTACTGATGTTGGGTT
AATGTTATGGTTCAGGCGCTGCGCTGCTAGCTGCTGGAAGATTGGGAACTCCCTCTTATTCTAAACTTTCCAA
GTCCCGCTACCGCATCAAGTTGAGAGATAAATGAAACAAGTACGGCCGCTCCCTATGGGATGTACTATAACCAAA
ATTTGGGGTATCCCGGAAAGTACGGTAGAGCGGTTTATGAATGCTCCGTTGGGACTTATTACCACCAAGATGATG
AATGTAATCCCAACCTTATGCGTTGGAGAGCGGTTTC

```

And So on.....

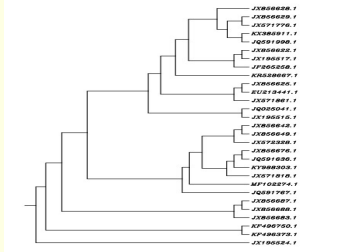
Open homepage of ClustalW tool (<http://www.genome.jp/tools-bin/clustalw>)

Since, multiple sequences in FASTA format are nucleotide sequence, then set 'DNA' for analysis and enter sequences in box 'Enter your sequences'. Select Weight Matrix as IUB (for DNA). Clicks execute multiple Alignment.

Multiple sequence alignment using CLUSTALW tools

Construction of phylogenetic tree using UPGMA

The sequence data was analyzed by Unweighted Pair Group Mean Average (UPGMA) methods using CLUSTAL W program. After multiple sequence alignment, we will go for rooted phylogenetic tree (UPGMA -unweighted pair-group with arithmetic mean) and finally click exec for construction of phylogenetic tree.



Construction of phylogenetic tree

Software	Version	Application	References
Trimmomatic	0.36	Adapter and low-quality end trimming	Bolger et al. (2014)
Trinity	2.1.1	RNA Seq denovo assembly	Grabherr et al. (2011)
Cd-hit	4.6.1	Transcript clustering to generate unigenes	Li and Godzik (2006)
Trans Decoder	2.0.1	CDS prediction from unigenes	http://transdecoder.github.io/
Blastx	2.2.30	Functional Annotation against NR and TF database	Altschul et al. (1997)
Blast2GO	Pro 3.0.9	GO analysis	Conesa et al. (2005)
WEGO	-	Plotting GO annotation	Ye et al. (2006)
KAAS	Web server	Pathway Analysis against KEGG database	Moriya et al. (2007)
MicroSatellite (MISA)	A perl script	SSR identification	Thiel et al. (2003)
BatchPrimer3	2.0	mining SSR markers and primer design	You et al. (2008)
CLC genomics workbench	V6	Mapping of reads to CDS for Expression profiling	http://www.clcbio.com

Fields of Bioinformatics

- › Molecular Medicine
- › Gene Therapy
- › Drug Development
- › Microbial genome applications
- › Crop Improvement
- › Forensic Analysis of Microbes
- › Biotechnology
- › Evolutionary Studies