

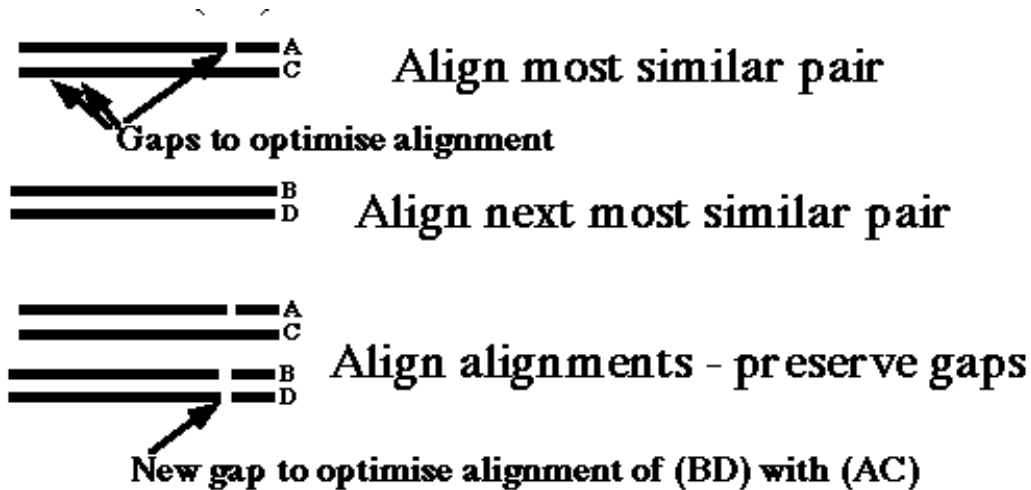


Figure 0.18 sequence alignments

CONCLUSION

MSA can help align multiple sequences. Progressive alignment can help perform MSA. Need to remove sequences with >80% similarity.

This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our [Terms of Use](#). OK

EMBL-EBI Services Research Training About us

Clustal Omega

[Input form](#) [Web services](#) [Help & Documentation](#) Share Feedback

[Tools](#) > [Multiple Sequence Alignment](#) > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

STEP 1 - Enter your input sequences

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

Or, upload a file: Choose File No file chosen

STEP 2 - Set your parameters

[OUTPUT FORMAT](#) Clustal w/o numbers

The default settings 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.)

Figure 0.19 CLUSTAL – Online tool

<http://www.ebi.ac.uk/Tools/msa/clustalo/>

Module 045: PROGRESSIVE ALIGNMENT FOR MSA

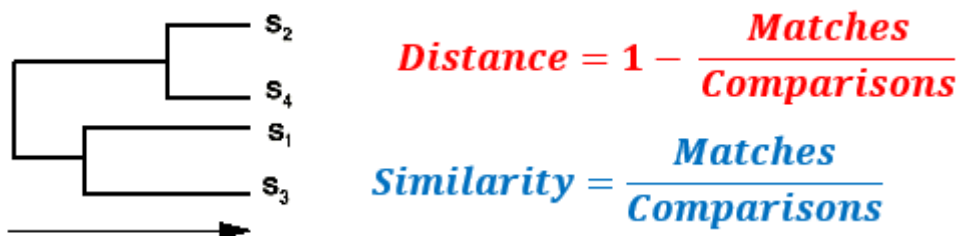
MSA involves progressive alignment of sequences. Doing so many progressive alignments can be slow.

STEPS:

- Step 1 : Pairwise Alignment of all sequences

Example: S_1, S_2, S_3, S_4 , so that is 6 pairwise comparisons.

- Step 2: Construct a Guide Tree (Dendrogram) using a *Distance Matrix*.



- Step 3: Progressive alignment following branching order in tree.

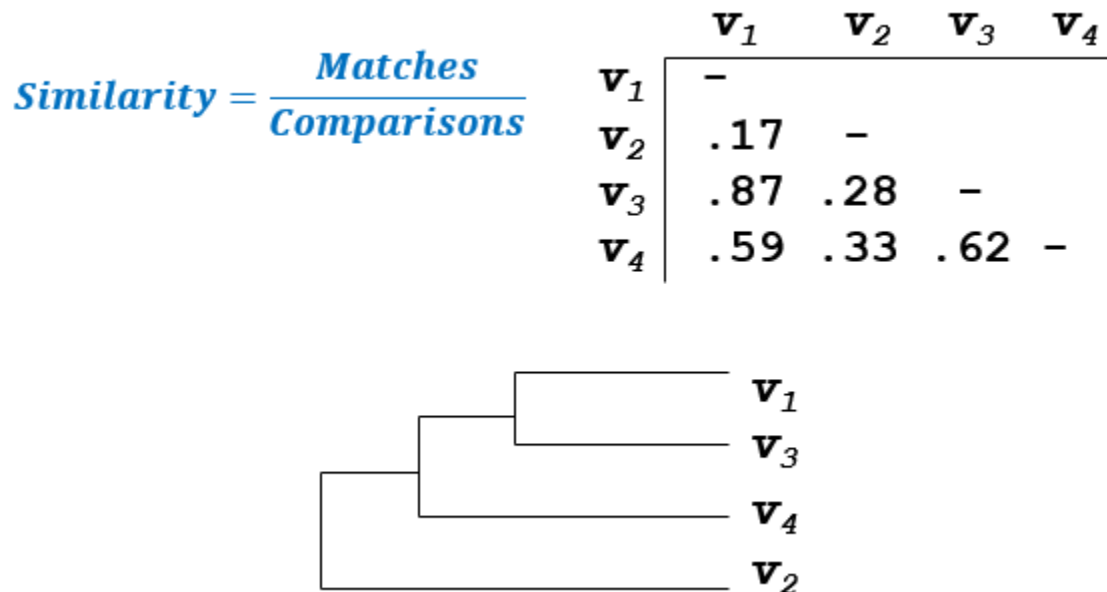


Figure 0.20 Similarity Matrix

- **SHORTCOMING OF THIS APPROACH**

- Dependence upon initial alignments
- If sequences are dissimilar, errors in alignment are propagated
- Solution: Begin by using an initial alignment, and refine it repeatedly

Progressive alignments are used in aligning multiple sequences. Iterative approaches can help refine results from progressive alignments.

Module 046: MSA-EXAMPLE

MSA involves progressive alignment of sequences. Doing so many progressive alignments can be slow.

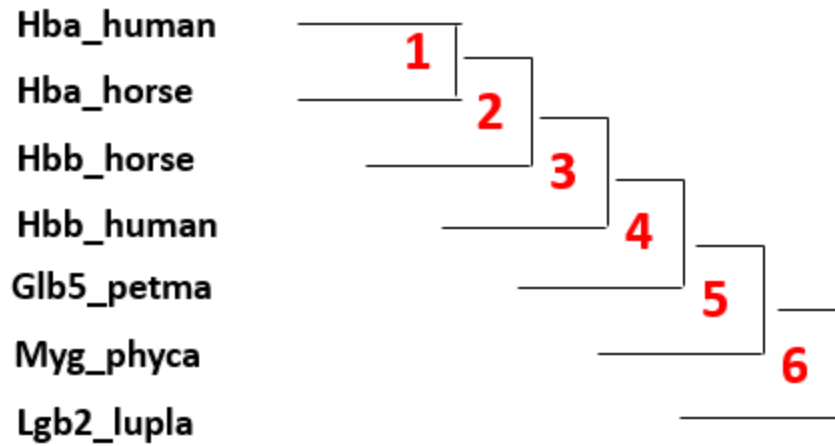
For example:

distance between 2 sequences = $1 - \frac{\text{No. identical residues}}{\text{No. aligned residues}}$

		Hbb_human	Hbb_horse	Hba_human	Hba_horse	Myg_phyca	Glb5_petma	Lgb2_lupla
Hbb_human	1	-						
Hbb_horse	2	.17	-					
Hba_human	3	.59	.60	-				
Hba_horse	4	.59	.59	.13	-			
Myg_phyca	5	.77	.77	.75	.75	-		
Glb5_petma	6	.81	.82	.73	.74	.80	-	
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90	-

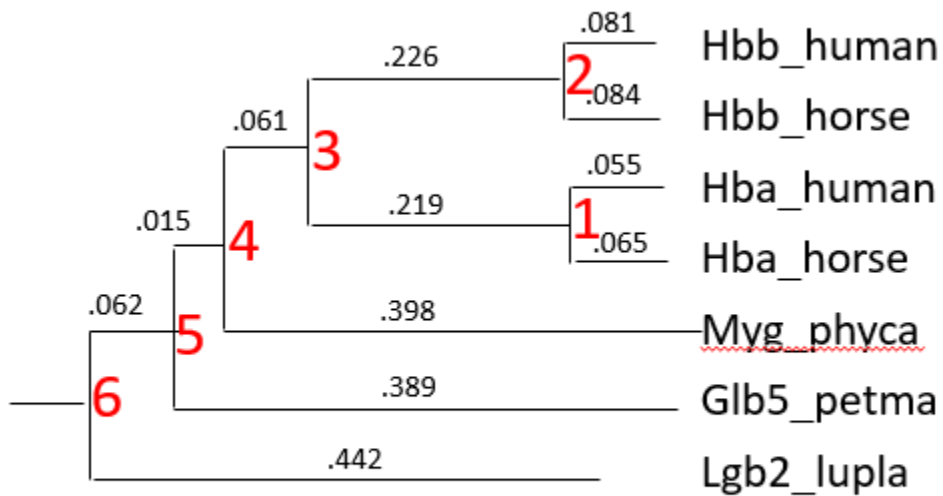
Thompson et al, IGBMC

Figure 0.21 MSA on globin sequences



Thompson et al, IGBMC

Figure 0.22 Progressive alignment using sequential branching



Thompson et al, IGBMC

Figure 0.23 Progressive alignment following a guide tree

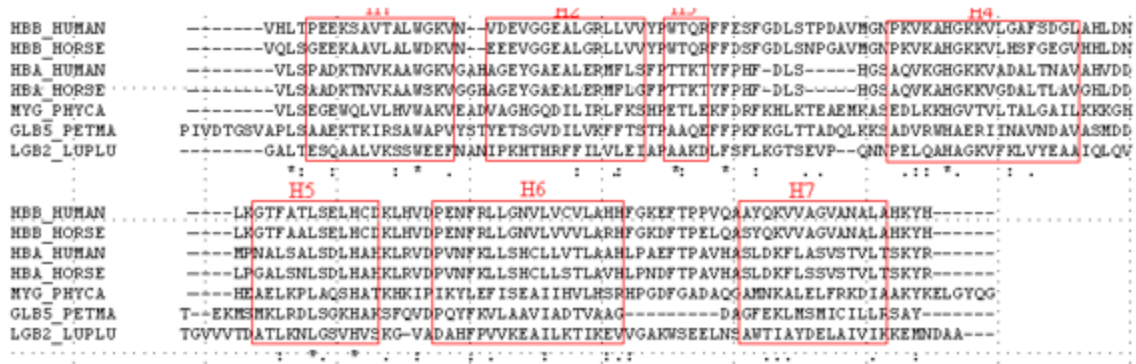


Figure 0.24 Alignment results

MSA can be better performed using clustering strategies followed by alignment of the alignments later. CLUSTAL is a free online tool that does all of this for us!

Module 047: CLUSTALW

MSA involves progressive alignment of sequences. Doing so many progressive alignments can be slow. CLUSTALW is an online tool to perform MSA.

Developed by European Molecular Biology Laboratory & European Bioinformatics Institute.

Performs alignment in:

- slow/accurate
- fast/approximate

SCOPE

- create multiple alignments,
- optimize existing alignments,
- profile analysis &
- create phylogenetic trees



Multiple Sequence Alignment by CLUSTALW

CLUSTALW	MAFFT	PRRN
Help		
General Setting Parameters: Output Format: CLUSTAL ▾ Pairwise Alignment: ☒ FAST/APPROXIMATE ☐ SLOW/ACCURATE Enter your sequences (with labels) below (copy & paste): ☒ PROTEIN ☐ DNA Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF 		
Or give the file name containing your query Choose File No file chosen Execute Multiple Alignment Reset		
More Detail Parameters... http://www.genome.jp/tools/clustalw		

More Detail Parameters...

Pairwise Alignment Parameters:

For FAST/APPROXIMATE:

K-tuple(word) size: , Window size: , Gap Penalty:

Number of Top Diagonals: , Scoring Method:

For SLOW/ACCURATE:

Gap Open Penalty: , Gap Extension Penalty:

Select Weight Matrix:

(Note that only parameters for the algorithm specified by the above "Pairwise Alignment" are valid.)

Multiple Alignment Parameters:

Gap Open Penalty: , Gap Extension Penalty:

Weight Transition: ☐ YES (Value:) , ☒ NO

Hydrophilic Residues for Proteins:

Hydrophilic Gaps: ☒ YES ☐ NO

Select Weight Matrix:

Type additional options (delimited by whitespaces) below:

(-options for help)

Table 1. Some recent and less recent available methods for MSAs.

Name	Algorithm	URL
MSA	Exact	http://www.ibc.wustl.edu/ibc/msa.html
DCA	Exact (requires MSA)	http://bibiserv.techfak.uni-bielefeld.de/dca
OMA	Iterative DCA	http://bibiserv.techfak.uni-bielefeld.de/oma
ClustalW, ClustalX	Progressive	ftp://ftp-igbmc.u-strasbg.fr/pub/clustalW or clustalX
MultAlin	Progressive	http://www.toulouse.inra.fr/multalin.html
DiAlign	Consistency-based	http://www.gsf.de/biodv/dialign.html
ComAlign	Consistency-based	http://www.dairi.au.df/~ocaprani
T-Coffee	Consistency-based/progressive	http://igs-server.cnrs-mrs.fr/~cnotred
Praline	Iterative/progressive	jhering@nimr.mrc.ac.uk
IterAlign	Iterative	http://giotto.Stanford.edu/~luciano/iteralign.html
Prp	Iterative/Stochastic	ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/
SAM	Iterative/Stochastic/HMM	rph@cse.ucsc.edu
HMMER	Iterative/Stochastic/HMM	http://hmmer.wustl.edu/
SAGA	Iterative/Stochastic/GA	http://igs-server.cnrs-mrs.fr/~cnotred
GA	Iterative/Stochastic/GA	czhang@watnow.uwaterloo.ca

Module 048: INTRODUCTION TO BLAST-I

National Center for the Biotechnology Information (NCBI) – USA. BLAST developed in 1990. “Basic Local Alignment Search Tool”. Searches databases for query protein and nucleotide sequences. Also searches for translational products etc. Online availability

www.blast.ncbi.nlm.nih.gov/Blast.cgi

NIH

U.S. National Library of Medicine

NCBI National Center for Biotechnology Information

BLAST®

BLAST finds regions of similarity between biological sequences. [more...](#)

BLAST Assembled Genomes

Find Genomic BLAST pages:

☐ [Human](#)
☐ [Rabbit](#)
☐ [Zebrafish](#)

☐ [Mouse](#)
☐ [Chimp](#)
☐ [Clawed frog](#)

☐ [Rat](#)
☐ [Guinea pig](#)
☐ [Arabidopsis](#)

☐ [Cow](#)
☐ [Fruit fly](#)
☐ [Rice](#)

☐ [Pig](#)
☐ [Honey bee](#)
☐ [Yeast](#)

☐ [Dog](#)
☐ [Chicken](#)
☐ [Microbes](#)

Basic BLAST

Choose a BLAST program to run.

[nucleotide blast](#)

Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast

Basic BLAST

Choose a BLAST program to run.

<u>nucleotide blast</u>	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
<u>protein blast</u>	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast, delta-blast</i>
<u>blastx</u>	Search protein database using a translated nucleotide query
<u>tblastn</u>	Search translated nucleotide database using a protein query
<u>tblastx</u>	Search translated nucleotide database using a translated nucleotide query

Standard Nucleotide BLAST

[blastn](#)
[blastp](#)
[blastx](#)
[tblastn](#)
[tblastx](#)

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter accession number(s), gi(s), 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

☐ Human genomic + transcript
☐ Mouse genomic + transcript
☒ Others (nr etc.):

Nucleotide collection (nr/nt)

Organism
Optional

Enter organism name or id—completions will be suggested
☐ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude
Optional

☐ Models (XM/XP)
☐ Uncultured/environmental sample sequences

Limit to
Optional

☐ Sequences from type material

Entrez Query

YouTube

Create custom database

BLAST can be used to search for local alignment of protein and nucleotide sequences. It is available online. Can perform searches across species and organisms

Module 049: INTRODUCTION TO BLAST-II

National Center for the Biotechnology Information (NCBI) – USA. BLAST developed in 1990. “Basic Local Alignment Search Tool”. Searches databases for query protein and nucleotide sequences. Also searches for translational products etc. Online availability

www.blast.ncbi.nlm.nih.gov/Blast.cgi

Smith Waterman can align complete sequences. BLAST does it in an approximate way. Hence, BLAST is faster BUT does not ensure optimal alignment. BLAST provides for approximate sequence matching. Input to BLAST is a FASTA formatted sequence and a set of search parameters

OUTPUT OF BLAST

Results are shown in HTML, plain text, and XML formats. A table lists the sequence hits found along with scores. Users can read this table off and evaluate results

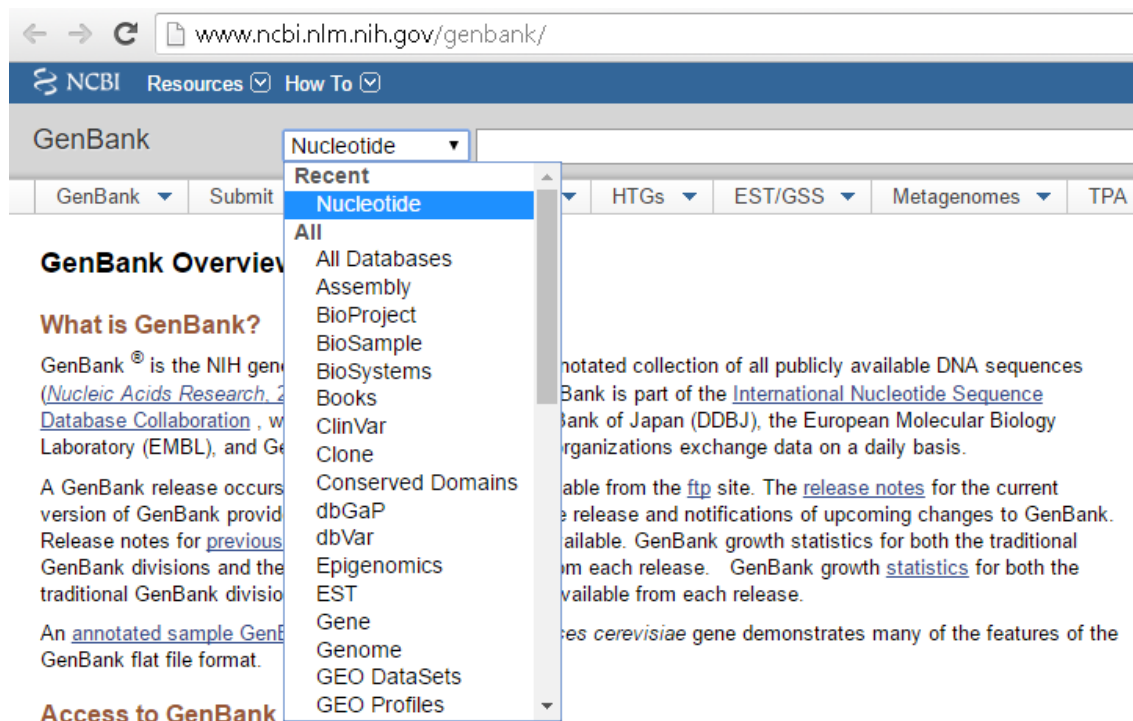


Figure 0.25 Input to BLAST: Gene IDs

← → ↻ www.uniprot.org/proteomes/ ☆ ⓘ

UniProt

Proteomes ▾ Advanced ▾ 🔍 Search

BLAST Align Retrieve/ID mapping Help Contact

Proteomes results

What are proteome sets?

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes. [? Help](#)

What are reference proteome sets?

Some proteomes have been (manually and algorithmically) selected as **reference proteomes**. They cover well-studied model organisms and other organisms of interest for biomedical research and phylogeny. [? Help](#)

[▶ Proteomes help video](#)
[▶ Other tutorials and videos](#)
[⬇ Downloads](#)

Filter byⁱ [Download](#) [Columns](#) ◀ 1 to 25 of 45,293 ▶ Show 25 ▾

Figure 0.26 Input to BLAST: Protein IDs

Graphic Summary

Distribution of 2 Blast Hits on the Query Sequence ⓘ

Mouse over to see the define, click to show alignments

Color key for alignment scores

Score Range	Color
<40	Black
40-50	Blue
50-80	Green
80-200	Pink
>=200	Red

Query

1 30 60 90 120 150

Descriptions

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

[Alignments](#)
[Download](#)
[GenPept](#)
[Graphics](#)
[Distance tree of results](#)
[Multiple alignment](#)

	Description	Max score	Total score	Query cover	E value	Ident	Acc
☐	hypothetical protein MIV003L [Invertebrate iridescent virus 3]	320	320	100%	2e-109	100%	YP_6
☐	unnamed protein product [Phytomonas sp. isolate EM1]	37.7	37.7	67%	4.3	28%	CCW

Figure 0.27 Results from BLAST

Module 050: BLAST ALGORITHM

BLAST can search sequence databases and identify unknown sequences by comparing them to the known sequences. This can help identify the parent organism, function and evolutionary history.

For example:

Query sequence: PQGELV

Make list of all possible words (length 3 for proteins)

PQG (score 15)

QGE (score 9)

GEL (score 12)

ELV (score 10)

Assign scores from Blosum62, use those with score > 11: PQG & GEL

Mutate words such that score still > 11

PQG (score 15) similar to PEG (score 13)

At the end, we get: PQG, GEL and PEG

Find all database sequences that have at least 2 matches among our 3 words: PQG, GEL & PEG.

Find database hits and extend alignment (High-scoring Segment Pair):

Query:	M	E	T	P	Q	G	I	A	V
Database:	-	-	-	P	Q	G	E	L	V
				8	5	5	2	0	8

High Scoring Pair: PQGI (score 8+5+5+2)

If 2 HSP in query sequence are < 40 positions away

Full dynamic alignment on query and hit sequences

BLAST performs quick alignments on sequences. The results are tabulated with alignment regions overlapping each other. Statistical evaluation is also provided alongside

Module 051: TYPES OF BLAST

BLAST can search sequence databases and identify unknown sequences by comparing them to the known sequences. This can help identify the parent organism, function and evolutionary history.

There are two main types of BLAST.

Nucleotides

- **Blastn:** Compares a nucleotide query sequence against a nucleotide database.

Proteins

- **Blastp:** Compares an amino acid query sequence against a protein database.

There are also many other types of BLAST:

- **Blastx:**

Compares a nucleotide query sequence against a protein sequence database.

Helps find potential translation products of unknown nucleotide sequences

- **tblastn:**

Compares a protein query sequence against a nucleotide sequence database

Nucleotide sequence dynamically translated into all reading frames

- **tblastx:**

Compares the six-frame translated proteins of a nucleotide query sequence against the six-frame translated proteins of a nucleotide sequence database.

- BLAST performs quick alignments on biological sequences
- Several types of BLAST exist which can assist in comparing nucleotide sequences with amino acids and vice versa

Module 052: SUMMARY OF BLAST

BLAST can search sequence databases and identify unknown sequences by comparing them to the known sequences. This can help identify the parent organism, function and evolutionary history.

Step1: obtain a query of sequence

NCBI Resources ☒ HowTo ☒ My NCBI Sign

Protein
Translations of Life

Search: Protein Limits Advanced search Help

Display Settings: ☒ FASTA ☒

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

[GenPept](#) [Graphics](#)

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDVAVMGNPVKVKAHGKKVLG
AFSDGLAHLDDLKGTFAATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAAYQKVAVGAVN
ALAHKYH
```

Analyze this sequence

[Run BLAST](#)

[Identify Conserved Domains](#)

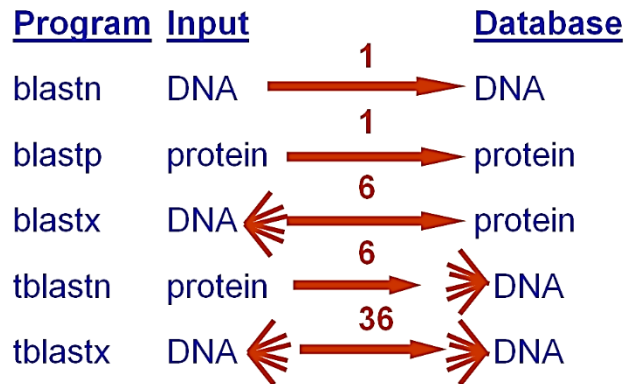
[Find in this Sequence](#)

Step2: choose a type of BLAST

Basic BLAST

Choose a BLAST program to run.

<u>nucleotide blast</u>	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontinuous megablast
<u>protein blast</u>	Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phi-blast, delta-blast
<u>blastx</u>	Search protein database using a translated nucleotide query
<u>tblastn</u>	Search translated nucleotide database using a protein query
<u>tblastx</u>	Search translated nucleotide database using a translated nucleotide query



Step3: search parameter

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

► NCBI/BLAST/blastp suite

blastn blastp **blastx** tblastn tblastx

Enter Query Sequence

BLASTP programs search protein databases using a protein query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)

NP_000509.1

From

To

Or, upload file [Browse...](#)

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database

Organism optional ☐ Exclude [+](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude optional ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query optional

Program Selection

Algorithm

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST Search database **Non-redundant protein sequences (nr)** using **Blastp (protein-protein BLAST)**

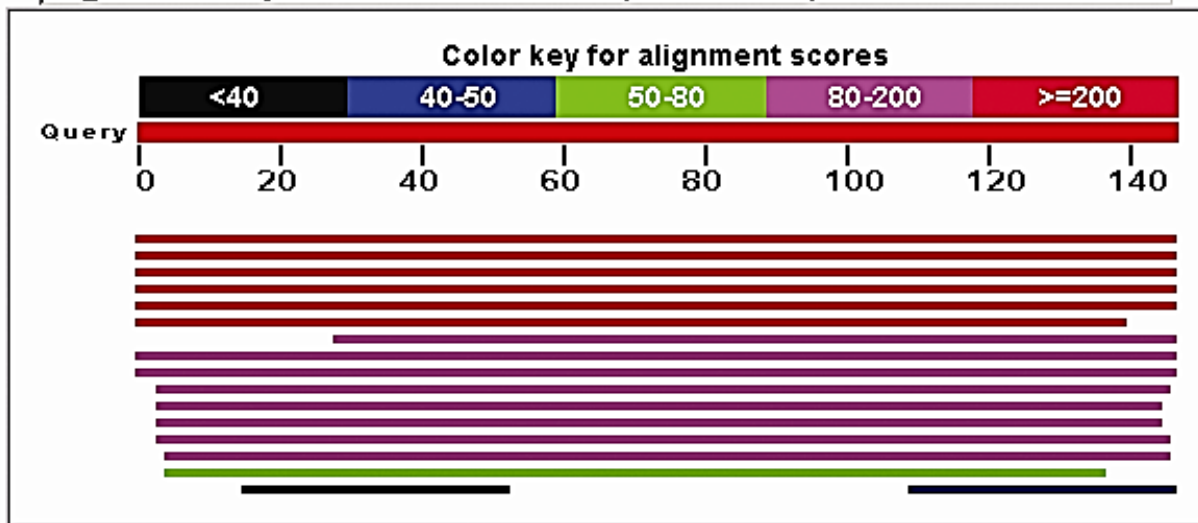
☐ Show results in a new window

► Algorithm parameters

Step4: tabulated search results

Distribution of 17 Blast Hits on the Query Sequence

NP_058652 hemoglobin, beta adult minor chain [Mus musculus] S=244 E=1.7e-65



BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/BLAST/blastp suite/ Formatting Results - GS1F74BK011

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

NP_000509:beta globin [Homo sapiens]

Query ID	gi 4504349 ref NP_000509.1	Database Name	nr
Description	beta globin [Homo sapiens] >gi 55635219 ref XP_508242.1 PREDICTED: hypothetical protein [Pan troglodytes] >gi 56749856 sp P68871.2 HBB_HUMAN RecName: Full=Hemoglobin subunit beta; AltName: Full=Hemoglobin beta chain; AltName: Full=Beta- hemoglobin, beta [synthetic construct] >gi 189053145 dbj BAG34767.1 unnamed protein product [Homo sapiens]	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects Program BLASTP 2.2.22+ Citation
Molecule type	amino acid		
Query Length	147		

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Related Structures](#) [Multiple alignment](#) **NEW**

Distance tree of results **NEW**

Sequences producing significant alignments:		Score (Bits)	E Value	
ref NP_058652.1 	hemoglobin, beta adult minor chain [Mus musculus]	244	2e-65	U G
ref NP_032246.2 	hemoglobin, beta adult major chain [Mus musculus]	228	2e-60	U G
ref XP_978992.1 	PREDICTED: similar to Hemoglobin epsilon-Y2 ...	226	3e-60	G
ref NP_032247.1 	hemoglobin Y, beta-like embryonic chain [Mus musculus]	223	4e-59	U G
ref NP_032245.1 	hemoglobin Z, beta-like embryonic chain [Mus musculus]	223	6e-59	U G
ref XP_998314.1 	PREDICTED: similar to Hemoglobin beta-H1 subunit ...	203	4e-53	G
ref XP_978924.1 	PREDICTED: similar to Hemoglobin epsilon-Y2 ...	187	2e-48	G
ref XP_912634.1 	PREDICTED: similar to Hemoglobin beta-2 subunit ...	161	2e-40	G
ref XP_488069.1 	PREDICTED: similar to Hemoglobin beta-2 subunit ...	154	3e-38	U G
ref NP_032244.1 	hemoglobin alpha 1 chain [Mus musculus]	105	1e-23	U G
ref XP_994669.1 	PREDICTED: similar to Hemoglobin alpha subunit ...	101	3e-22	G
ref XP_356935.3 	PREDICTED: similar to Hemoglobin alpha subunit ...	100	4e-22	U G
ref NP_034535.1 	hemoglobin X, alpha-like embryonic chain in ...	94.0	4e-20	U G
ref NP_001029153.1 	similar to hemoglobin, theta 1 [Mus musculus]	88.2	2e-18	U G
ref NP_778165.1 	hemoglobin, theta 1 [Mus musculus]	73.9	5e-14	U G
ref XP_978150.1 	PREDICTED: similar to hemoglobin, beta adult ...	41.6	2e-04	G
ref NP_795942.2 	5'-nucleotidase, cytosolic II-like 1 protein [Mus musculus]	28.9	1.5	U G

Figure 0.28 tabulated search results

Module 053: INTRODUCTION TO FASTA

For comparing two sequences we use pair wise sequencing and for the comparison of many sequences we use multiple sequence alignment. To handle the multiple alignments we perform alignment through smith-waterman algorithm for local one. And for global alignment we use Needleman-wunsch algorithm.

Both local and global alignments are the dynamic approaches. Many of the sequences are compared, which takes time and we use BLAST which is an approximate local alignment search tool BLAST compares a large number of sequences, quickly. FASTA took a similar approach.

Developed in 1988.it does Fast Alignment .Searches databases for query protein and nucleotide sequences. Was later improved upon in BLAST.

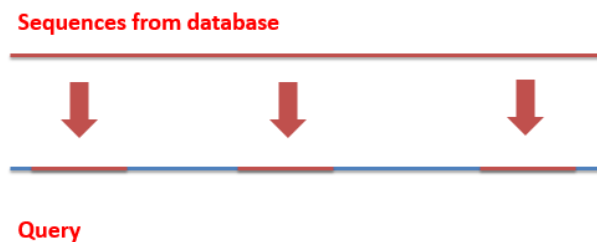


Figure 0.29 Regions of absolute identity

EMBL-EBI Services Research Training About us

FASTA

Protein Nucleotide Genomes Proteomes Whole Genome Shotgun Web services

Help & Documentation

Tools > Sequence Similarity Searching > FASTA

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query, FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected X Clear Selection

- ☒ UniProt Knowledgebase
- ☐ UniProtKB/Swiss-Prot
- ☐ UniProtKB/Swiss-Prot isoforms
- ☐ UniProtKB/TrEMBL
- ☐ UniProtKB Taxonomic Subsets
- ☐ UniProt Clusters
- ☐ Patents
- ☐ Structure
- ☐ Other Protein Databases

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

EMBL-EBI [Services](#) [Research](#) [Training](#) [About us](#)

FASTA

[Protein](#) **[Nucleotide](#)** [Genomes](#) [Proteomes](#) [Whole Genome Shotgun](#) [Web services](#) [Share](#) [Feedback](#)

[Help & Documentation](#)

[Tools](#) > [Sequence Similarity Searching](#) > FASTA

Nucleotide Similarity Search

This tool provides sequence similarity searching against nucleotide databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide query. TFASTX and TFASTY translate the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

NUCLEOTIDE DATABASES

111 Databanks Selected [X Clear Selection](#)

- ▼ ENA Sequence (formerly EMBL-Bank)
 - ☒ ENA Sequence Release
 - ☐ ENA Sequence Updates
 - ☐ ENA Coding Sequence Release
 - ☐ ENA Coding Sequence Updates
 - ☐ ENA Non-Coding Sequence Release
 - ☐ ENA Non-Coding Sequence Updates
- ▶ Others
- ▶ IMGT
- ▶ Patents
- ▶ Structure

Figure 0.30 Nucleotide FASTA

EMBL-EBI [Services](#) [Research](#) [Training](#) [About us](#)

FASTA

[Protein](#) [Nucleotide](#) [Genomes](#) [Proteomes](#) [Whole Genome Shotgun](#) [Web services](#) [Share](#) [Feedback](#)

[Help & Documentation](#)

[Tools](#) > [Sequence Similarity Searching](#) > FASTA

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected [X Clear Selection](#)

- ☒ UniProt Knowledgebase
- ☐ UniProtKB/Swiss-Prot
- ☐ UniProtKB/Swiss-Prot isoforms
- ☐ UniProtKB/TrEMBL
- ▶ UniProtKB Taxonomic Subsets
- ▶ UniProt Clusters
- ▶ Patents
- ▶ Structure
- ▶ Other Protein Databases

Figure 0.31 Protein FASTA

Module 054: INTRODUCTION TO FASTA-II

FASTA – Fast Alignment Algorithm. Classical global and local alignment algorithms are time consuming. FASTA achieves alignment by using short lengths of exact matches.

- **USES OF FASTA**

FASTA relies on aligning subsequences of absolute identity. Input to FASTA search can be in FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP or UniProt formats

- **OUTPUT OF BLAST**

Results are output in visual format along with functional prediction. Makes table lists the sequence hits found along with scores. Users can click on each reported match to look at the details.

Nucleotide Similarity Search

This tool provides sequence similarity searching against nucleotide databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide query. TFASTX and TFASTY translate the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

NUCLEOTIDE DATABASES

111 Databanks Selected X Clear Selection

- ▼ ENA Sequence (formerly EMBL-Bank)
 - ▶ ☒ ENA Sequence Release
 - ☐ ENA Sequence Updates
 - ☐ ENA Coding Sequence Release
 - ☐ ENA Coding Sequence Updates
 - ☐ ENA Non-Coding Sequence Release
 - ☐ ENA Non-Coding Sequence Updates
 - ▶ Others
- ▶ IMGT
- ▶ Patents
- ▶ Structure

STEP 2 - Enter your input sequence

Enter or paste a DNA sequence in any supported format:

```
CAGTCAGTCATAGTCGATAGTGTACGTAGCTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCT
AGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAG
TCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAG
TAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGT
```

or upload a file: Choose File No file chosen

Figure 0.32 Input to FASTA: Gene IDs

Tools > Sequence Similarity Searching > FASTA

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected X Clear Selection

- ☒ UniProt Knowledgebase
- ☐ UniProtKB/Swiss-Prot
- ☐ UniProtKB/Swiss-Prot isoforms
- ☐ UniProtKB/TrEMBL
- ▶ UniProtKB Taxonomic Subsets
- ▶ UniProt Clusters
- ▶ Patents
- ▶ Structure
- ▶ Other Protein Databases

STEP 2 - Enter your input sequence

Enter or paste a a [PROTEIN] sequence in any supported format:

```
MDASSSPSPSEESLKLDDLOKQLNKKLRFEASVCSIHNLRLDHYSSSSPSLRKQFYV
VSRVATVLKTRYTATGFVWAGLSLFEAEERLVSDASEKKHLKSCVAQAKEQLSEVDNQPT
ESSQGYLFEGHLTVDRPPQPQWLQQNLMSAFASIVGGESSNGPTENTIGETANLMQEL
INGLDMIIPIIDLDGGPPRAPPASKEVVEKLPVIFTEELLKKFGAEAECCICKENLVIG
DKMQELPCKHTFHPPCLKPWLDEHNSCPICRHELPTDDQYENWKEREKAEERKGAEN
AVRGGEYMYV
```

or Upload a file: Choose File No file chosen

Figure 0.33 Input to FASTA: Protein Sequence

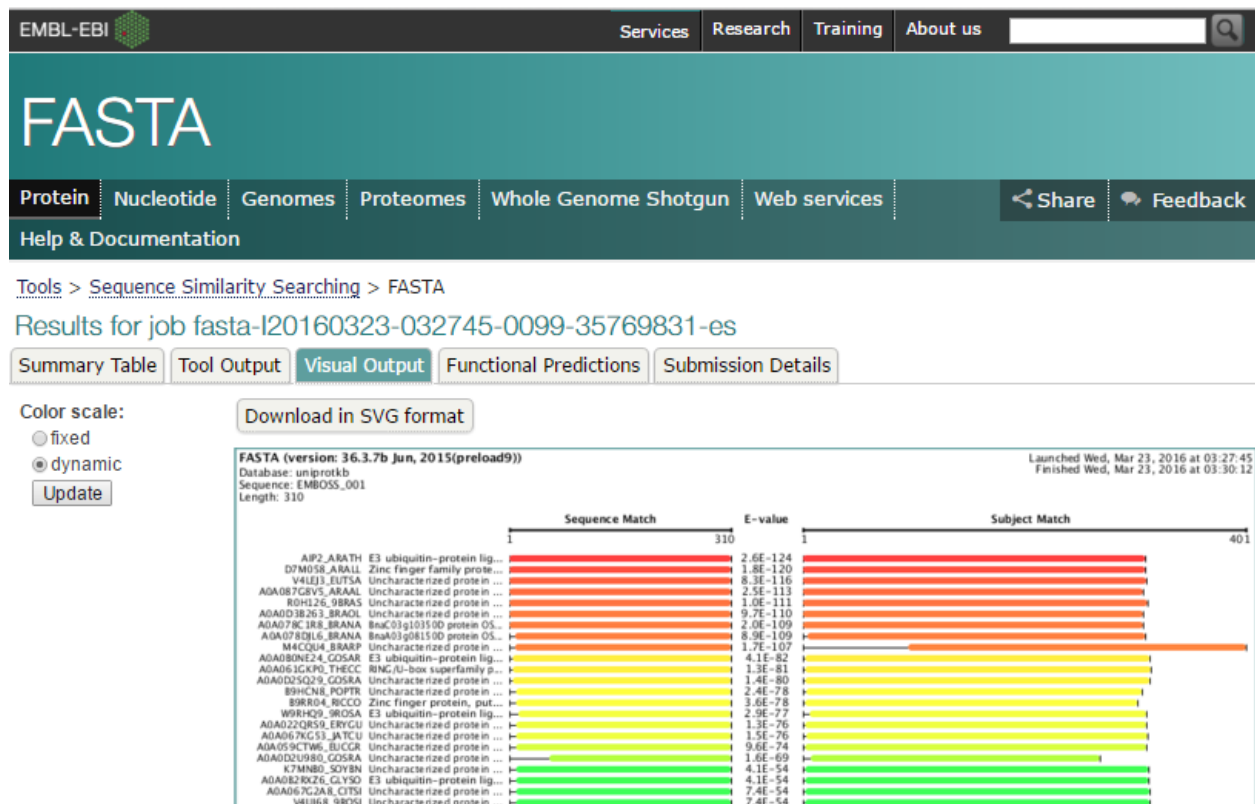
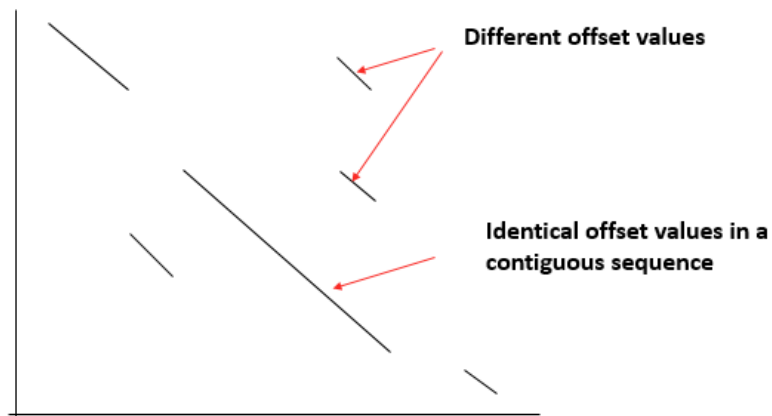


Figure 0.34 Results from FASTA

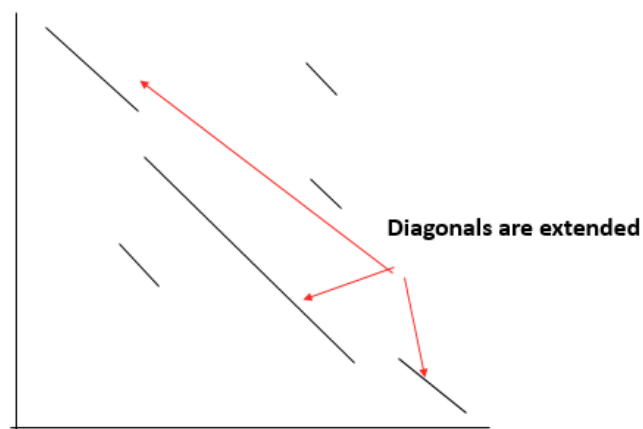
Module 055: FASTA ALGORITHM

FASTA can search sequence databases and identify unknown sequences by comparing them to the known sequence databases. This can help obtain information on the parent organism, function and evolutionary history.

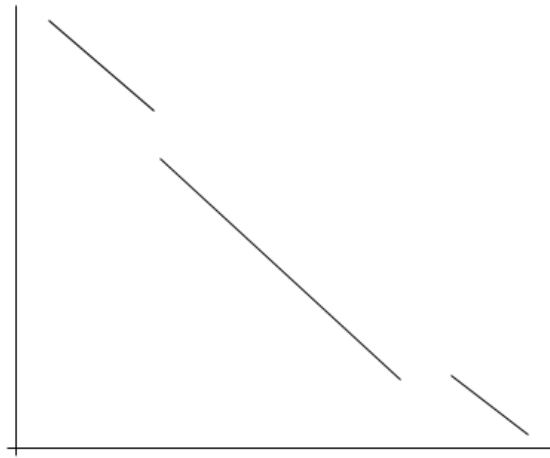
STEP1: Local regions of identity are found



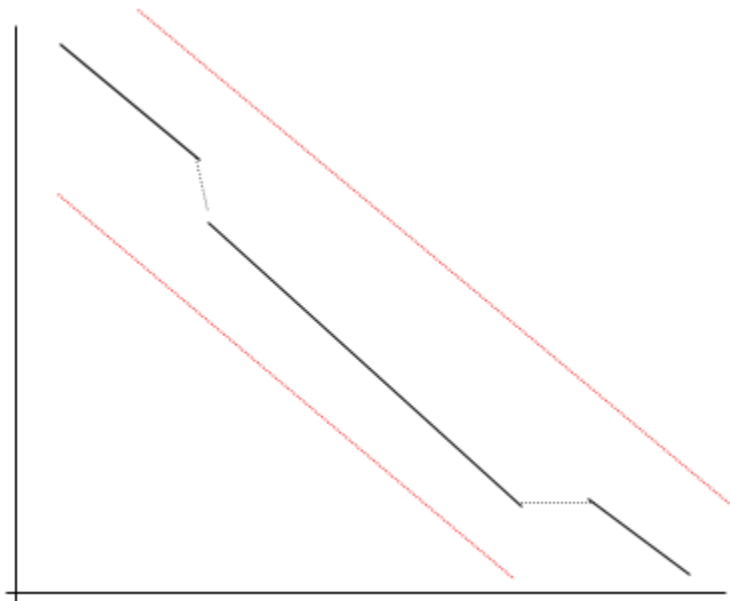
STEP2: Rescore the local regions using PAM or BLOSUM matrix



STEP3: Eliminate short diagonals below a cutoff score



STEP4: Create a gapped alignment in a narrow segment and then perform Smith Watermann alignment



Module 056: TYPES OF FASTA

There are six types of FASTS:

- **fasts35**

Compare unordered peptides to a protein sequence database

- **fastm35**

Compare ordered peptides (or short DNA sequences) to a protein (DNA) sequence database

- **Fasta35**

Scan a protein or DNA sequence library for similar sequences

- **Fastx35**

Compare a translated DNA sequence (6 ORFs) to a protein sequence database

- **tfastx35**

Compare a protein sequence to a DNA sequence database (6 ORFs)

- **fasty35**

Compare a DNA sequence (6ORFs) to a protein sequence database

FASTA performs quick alignments on biological sequences. Several types of FASTA exist which can assist in comparing DNA/RNA/Protein sequences with each other

Module 057: SUMMERY OF FASTA

FASTA can briskly perform sequence search databases if given a query sequence. Multiple types of FASTA exist which assist in aligning DNA/RNA/Protein sequences

NCBI Resources ☒ How To ☒ My NCBI Sign

Protein
Translations of Life

Search: Protein Limits Advanced search Help

Display Settings: ☒ FASTA ☒ **Change region shown**

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

[GenPept](#) [Graphics](#)

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDENVGGEALGRLLVVYPWTQRFFESFGDLSTPDVAVMGNPVKAHGKKVLG
AFSDGLAHLNDLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHHFGKEFTPPVQAAAYQKVVAGVAN
ALAHKYH
```

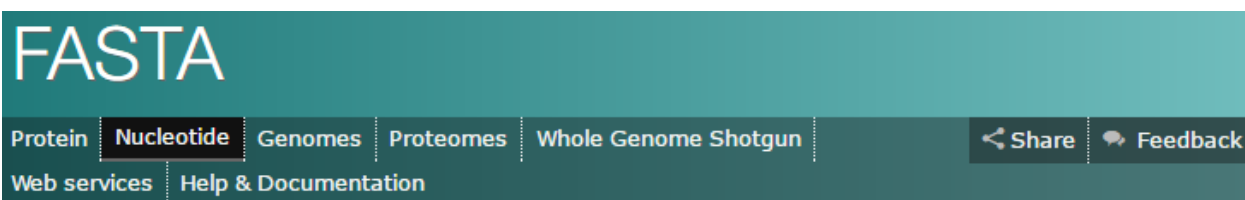
Analyze this sequence

[Run BLAST](#)

[Identify Conserved Domains](#)

[Find in this Sequence](#)

Figure 0.35 Step 1: Obtain a query sequence



Tools > Sequence Similarity Searching > FASTA

Nucleotide Similarity Search

This tool provides sequence similarity searching against nucleotide databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide query. TFASTX and TFASTY translate the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).



Figure 0.36 Step 2: Choose a type of FASTA

fasta35, fasta35_t*	scan a protein or DNA sequence library for similar sequences
fastx35, fastx35_t	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
tfastx35, tfastx35_t	compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.
fasty35, fasty35_t	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
tfasty35, tfasty35_t	compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.
fasts35, fasts35_t	compare unordered peptides to a protein sequence database
fastm35, fastm35_t	compare ordered peptides (or short DNA sequences) to a protein (DNA) sequence database
tfasts35, tfasts35_t	compare unordered peptides to a translated DNA sequence database
fastf35, fastf35_t	compare mixed peptides to a protein sequence database
tfastf35, tfastf35_t	compare mixed peptides to a translated DNA sequence database
ssearch35, ssearch35_t	compare a protein or DNA sequence to a sequence database using the Smith-Waterman algorithm.

Figure 0.37 Type of FASTA

http://fasta.bioch.virginia.edu/fasta_docs/fasta35.shtml

STEP 2 - Enter your input sequence

Enter or paste a DNA sequence in any supported format:

CAGTCAGTCATAGTCGTAGATGTACGTAGCTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTG
ATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGT
AGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGA
TGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGT

or upload a file: No file chosen

STEP 3 - Set your parameters

PROGRAM

FASTA 

MATCH/MISMATCH SCORES

GAP OPEN

GAP EXTEND

KTUP

EXPECTATION
UPPER VALUE

EXPECTATION
LOWER VALUE

+5/-4 ▼

-14 ▼

-4

6

10 ▼

0 (default)	
-------------	--

DNA STRAND

HISTOGRAM

FILTER

STATISTICAL ESTIMATES

both ▼

no

.....
none

Regress

SCORES

ALIGNMENTS

SEQUENCE RANGE

DATABASE RANGE

MULTIHSPs

50 ▼

50

START-END

START-END

no ▼

SCORE FORMAT

Default ▼

Figure 0.38 Step 3: Setup Search Parameters

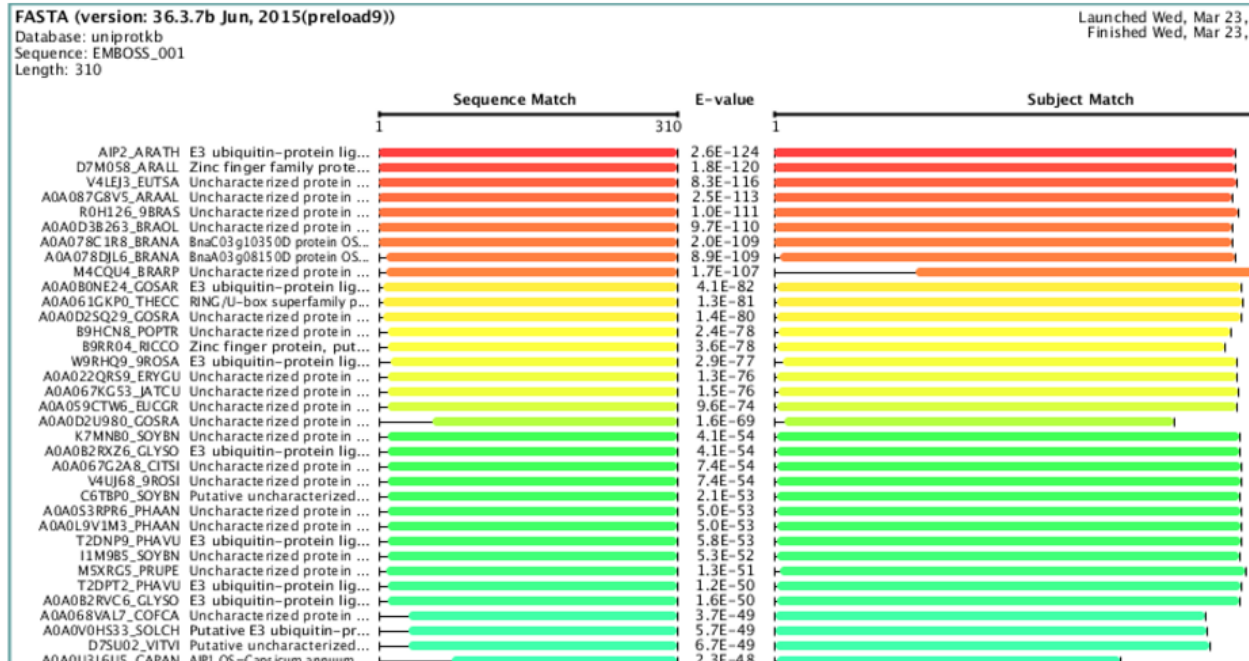


Figure 0.39 Step 4: Tabulated Search Results

Align. ▲	DB:ID ◆	Source ◆	Length ◆	Score (Bits) ◆	Identities % ◆	Positives % ◆	E() ◆
☒ 1	SP:AIP2_ARATH	E3 ubiquitin-protein ligase AIP2 OS=Arabidopsis thaliana GN=AIP2 PE=1 SV=1 <i>Cross-references and related information in:</i> ▶ Gene expression ▶ Small molecules ▶ Nucleotide sequences ▶ Genomes & metagenomes ▶ Enzymes ▶ Samples & ontologies ▶ Molecular interactions ▶ Protein families ▶ Literature ▶ Protein sequences	310	453.0	100.0	100.0	2.6E-124
☒ 2	TR:D7M058_ARALL	Zinc finger family protein OS=Arabidopsis lyrata subsp. lyrata GN=ARALYDRAFT_488997 PE=4 SV=1 <i>Cross-references and related information in:</i> ▶ Nucleotide sequences ▶ Genomes & metagenomes ▶ Samples & ontologies	310	440.3	96.5	99.7	1.8E-120

Figure 0.40 Tabulated data

Module 058: BIOLOGICAL DATABASE AND ONLINE TOOLS

All molecular information of RNA, DNA, Proteins have need to be stored and retrieved. Sequences are obtained from genome sequencing and mass spectrometry

Structures are obtained from X-Ray Crystallography, Atomic Force Microscopy & Nuclear Magnetic Resonance Spectroscopy.

Vast amounts of such data exists. Moreover, this data is rapidly accumulating. Online Databases are formed to store and share this data.

- **OBJECTIVE**

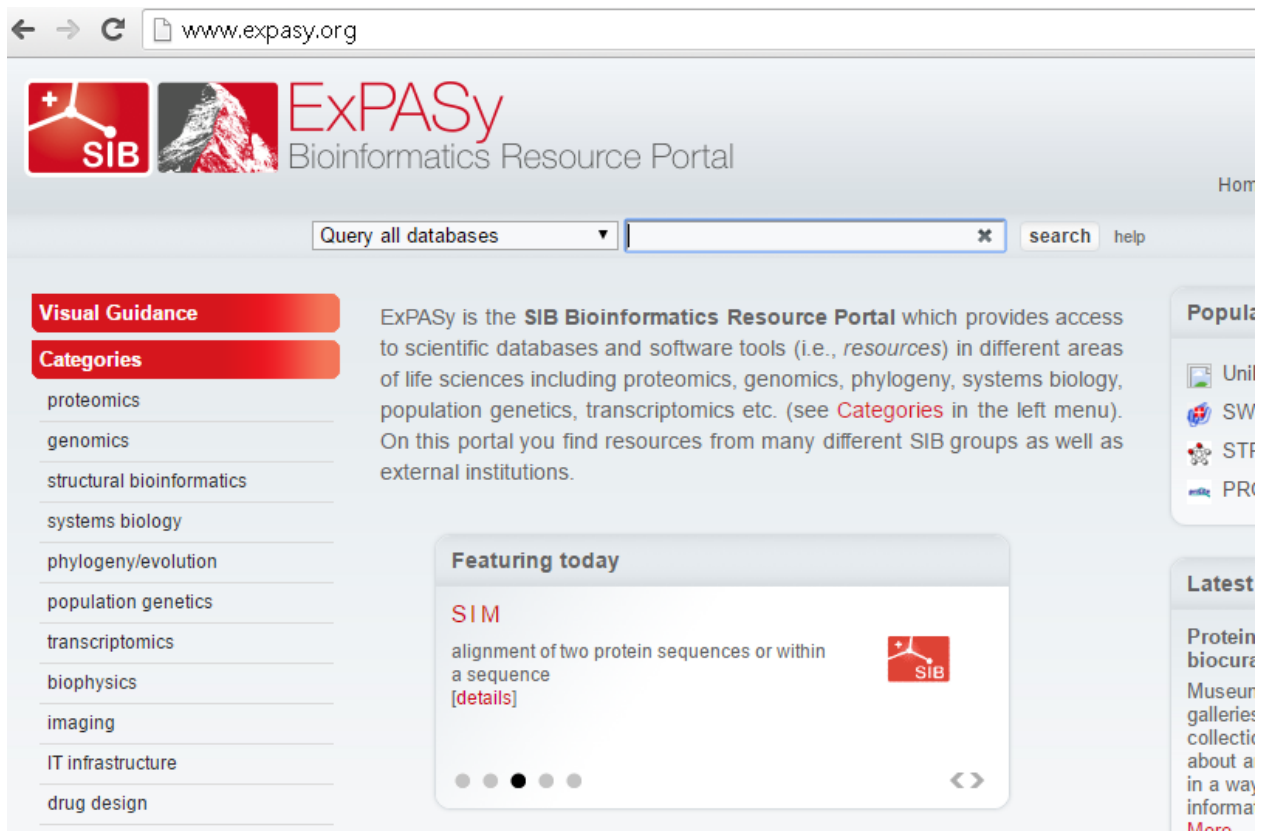
- Make biological data available to scientists in computer-readable form
- For handling, sharing and analysis of the data
- The best way to share is to keep this data on the web

Several sequence, structure and molecular interaction databases exist. These are available online on the web. Users can freely access and download such data

Module 059: EXPASY

It is developed by Swiss Bioinformatics Institute (SIB). Website provides access to databases and tools Proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. can be searched.

<http://www.expasy.org/>



The screenshot shows the ExPASy Bioinformatics Resource Portal homepage. At the top, there is a navigation bar with the SIB logo, the ExPASy logo, and the text "Bioinformatics Resource Portal". Below the navigation bar is a search bar with a dropdown menu set to "Query all databases", a search input field, and buttons for "search" and "help". On the left side, there is a "Visual Guidance" section with a red button and a "Categories" section with a red button. Below the "Categories" button is a list of categories: proteomics, genomics, structural bioinformatics, systems biology, phylogeny/evolution, population genetics, transcriptomics, biophysics, imaging, IT infrastructure, and drug design. In the center, there is a text block describing ExPASy as the SIB Bioinformatics Resource Portal, which provides access to scientific databases and software tools in various life science areas. Below this text is a "Featuring today" section with a slider showing a featured resource titled "SIM" (alignment of two protein sequences or within a sequence) with a [details] link and an SIB logo. On the right side, there are two sections: "Popular" with a list of popular resources (Unifac, SW, STF, PRO) and "Latest" with a section titled "Protein biocuration" and a link to "More".

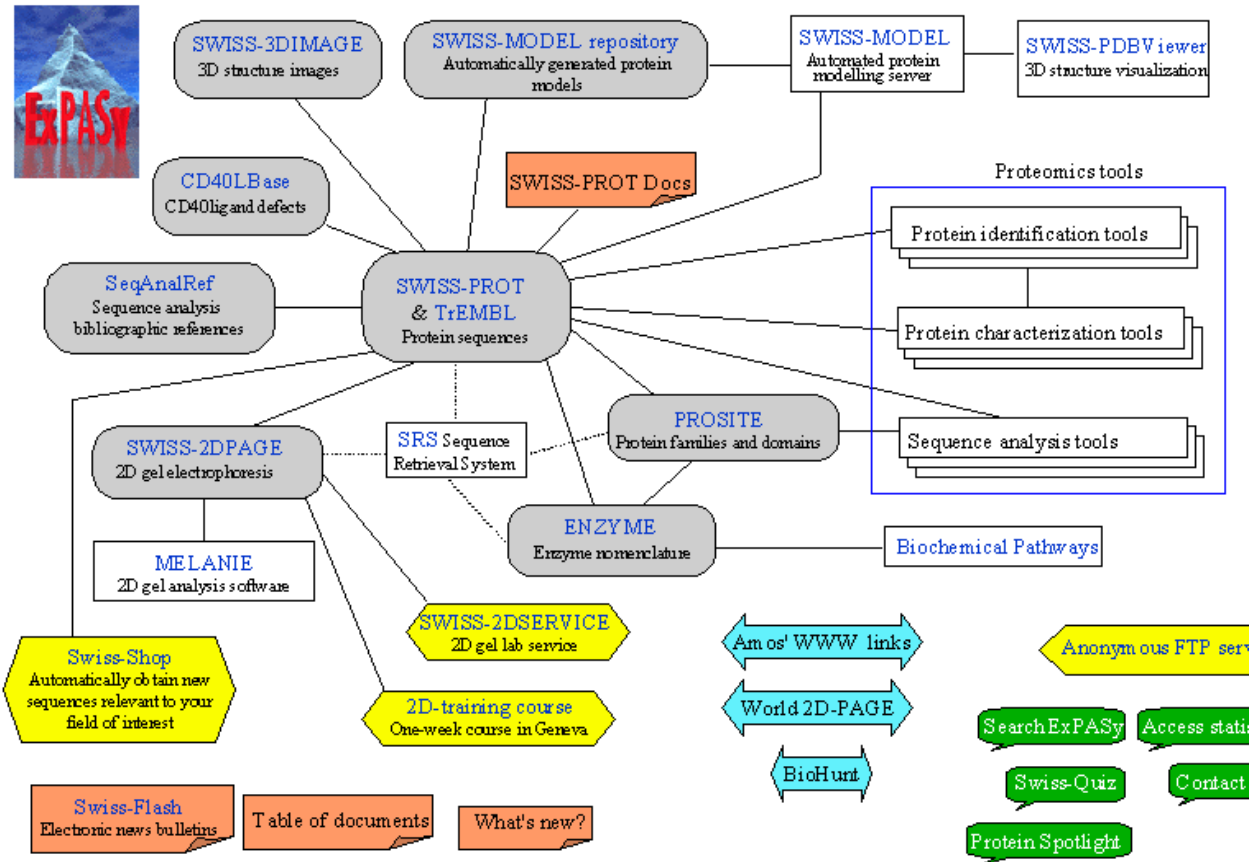


Figure 0.41 flowchart



The ScanProsite tool [[Help](#)] allows to scan protein sequence(s) (either from [Swiss-Prot](#) or [TrEMBL](#) or provided by the user) for the occurrence of patterns, profiles and rules (motifs) stored in the [PROSITE](#) database, or to search protein database(s) for hits by specific motif(s) [[Reference](#) / [Download ps_scan, the standalone version](#)]. The program [PRATT](#) can be used to generate your own patterns. You may either:

- Enter one or more PROSITE accession numbers and/or patterns [1 by line] to search the Swiss-Prot/TrEMBL and/or PDB databases, **OR**
- Enter one or more sequences [raw, Swiss_Prot or fasta format] and/or Swiss-Prot/TrEMBL accession numbers [1 by line] to be scanned with all patterns, profiles, rules in PROSITE, **OR**
- Fill in both fields to find all occurrences of a motif in a sequence.

Protein(s) to be scanned:	PROSITE pattern(s)/profile(s) to scan for:
Enter one or more Swiss-Prot/TrEMBL accession number(s) [AC] (e.g. P00747) and/or sequence identifier(s) [ID] (e.g. ENTK_HUMAN), and/or PDB identifier, and/or paste your own protein sequence(s) in the box below. (leave this box blank to scan PROSITE entries against selected protein databases) Clear	Enter one or more PROSITE accession number(s) (e.g. PS50240), and/or identifier(s) (e.g. CHEB), and/or type your pattern(s) in PROSITE format in the box below: (leave this box blank to scan sequence(s) against the entire PROSITE database) and specify your search limits (only used if no protein data specified) : • Protein database(s): ☒ Swiss-Prot ☐ TrEMBL ☐ PDB databases ☒ including splice variants randomize databases no (to test a pattern, see help) • Taxonomic lineage (OC) / species (OS) filter: (see NEWT Taxonomy; separate multiple taxa/species with a semicolon, e.g. <i>Eukaryota; Escherichia coli</i>; . Does not work on PDB sequences) • Description (DE) filter: e.g. protease
General options ☒ Exclude motifs with a high probability of occurrence ☐ Show low level score ☐ Do not scan profiles [User Manual] Show only sequences with at least 1 hit(s) Maximum of matched sequences 1000 ☒ Graphical rich view ☐ Simple HTML output ☐ Plain text output ☐ Plain text fasta output	pattern options:

Figure 0.42 prosite scanning section

[ExPASy Home page](#) [Site Map](#) [Search ExPASy](#) [Contact us](#) [Proteomics tools](#) [Swiss-Prot](#)

Search for



PeptideMass

PeptideMass [\[references\]](#) cleaves a protein sequence from the Swiss-Prot and/or TrEMBL databases **or** a user-entered protein sequence with a chosen enzyme, and computes the masses of the generated peptides. The tool also returns theoretical isoelectric point and mass values for the protein of interest. If desired, PeptideMass can return the mass of peptides known to carry post-translational modifications, and can highlight peptides whose masses may be affected by database conflicts, isoforms or splice variants.

[Instructions](#) are available.

Enter a Swiss-Prot protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406), **or** an amino acid sequence (e.g. 'SELVEGVIV'; you may specify post-translational modifications, but [PLEASE read this document first!](#)):

the fields. the cleavage of the protein.

The peptide masses are
with cysteines treated with:

Figure 0.43 peptide mass finding

PROWL for Intranet

"Global Solutions for Your Local Needs"

[Disclaimer](#)

[Contact ProteoMetrics](#)

PROWL for Intranet provides an interactive environment for protein analysis that can be easily accessed with any web browser over your intranet. PROWL consists of a set of interconnected databases and software tools that enable swift analysis of protein mass spectrometry experiments. PROWL is based on a powerful client-server model. Tasks that are highly interactive are performed on a local personal computer and computationally intensive tasks like database searches are sent to a local fast server. A demo of the system is available at <http://www.proteometrics.com/>. The following server-side tools for searching public or proprietary sequence databases are included in PROWL for Intranet:

- **ProFound** - a tool for searching a protein sequence database using information from mass spectra of peptide maps. A Bayesian algorithm is used to rank the protein sequences in the database according to their probability of producing the peptide map.
- **PepFrag** - a tool for searching protein or nucleotide sequences using information from fragmentation mass spectra of peptides.
- **ProteinInfo** - a tool for retrieval and analysis of information from protein sequence databases. The capabilities of the analysis tools include peptide mapping, mass spectrometric fragmentation analysis, disulfide mapping, etc.

PROWL for Intranet includes two client-side applications (available for Windows 95 and NT operating systems):

- **M/Z** - a tool for analysis of mass spectra obtained from protein chemistry experiments. M/Z employs an array of digital signal processing techniques and novel interface designs that make the analysis of protein mass spectra faster and easier.
- **PAWS** - a tool for analysis of protein sequences and post-translational modifications. PAWS greatly facilitates the interpretation of mass spectra of proteins and protein digest mixtures.



Figure 0.44 for local use of protein sequencing



FindMod tool

FindMod is a tool that can predict potential protein post-translational modifications (PTM) and find potential single amino acid sub

The experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot/TrE entered sequence, and mass differences are used to better characterise the protein of interest [[Documentation](#) / [Mass values and con](#)]

Swiss-Prot/TrEMBL ID or AC or user-entered sequence:

If your protein is **not** in Swiss-Prot or TrEMBL, please specify (if known) the source of your protein:

☐ Eukaryote or ☐ Bacteria or ☐ Archaea or ☐ Virus (☐ Phage).

(This information will be used to determine whether certain post-translational modifications are likely to occur in your sequence.)

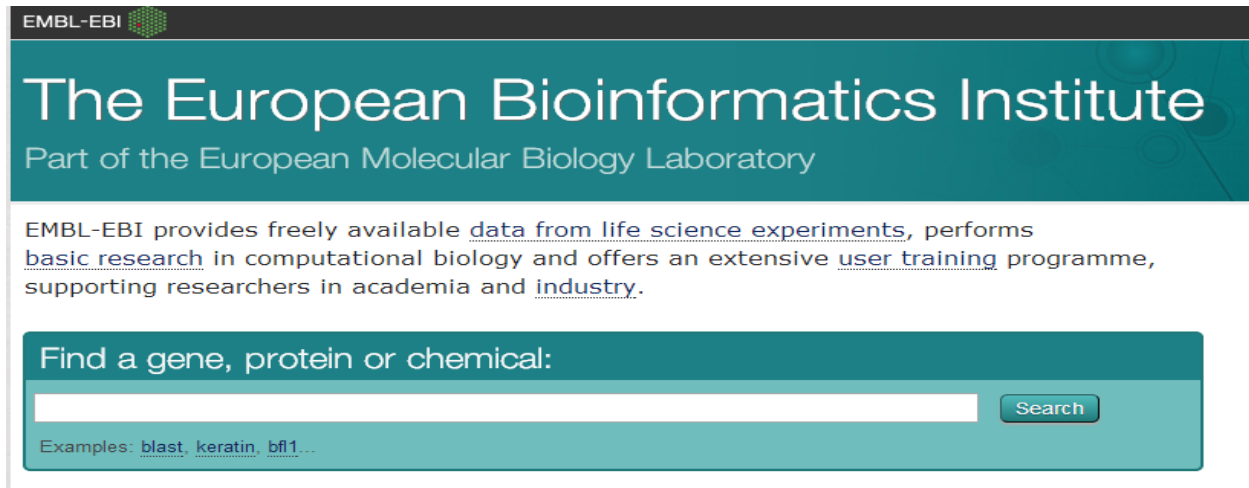
Enter a list of peptide masses and intensities (optional) that correspond to the specified protein. Enter one mass, space and its intensity per line:

Or upload a file from your computer, from
will be extracted ([supported formats](#)):

Figure 0.45 potential protein finding tool

Module 060: UNIPROT AND SWISSPROT

Both UniProt and SwissProt are the online database for proteins.



EMBL-EBI

The European Bioinformatics Institute

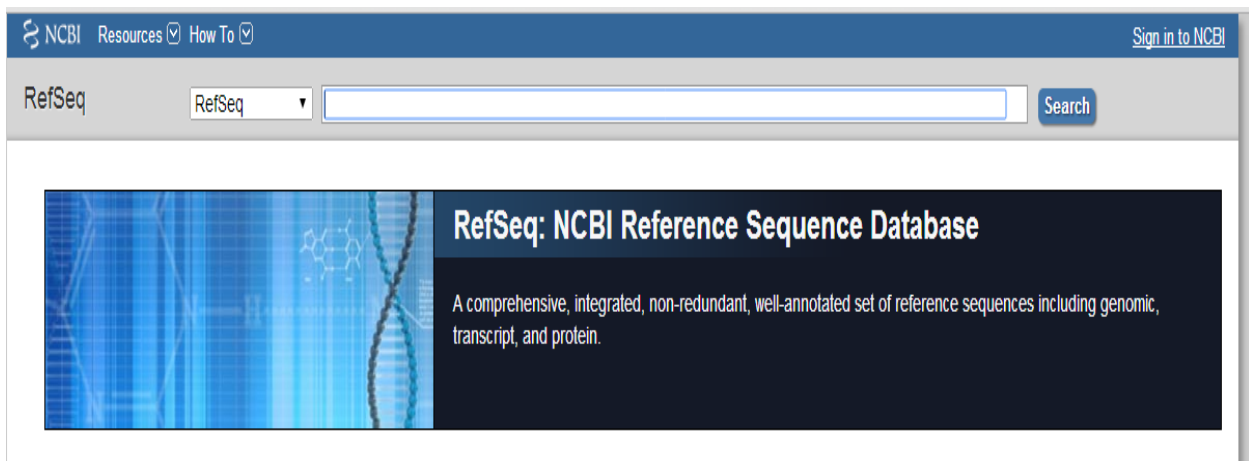
Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

Figure 0.46 gene, protein or chemical can be find



NCBI Resources ▾ How To ▾ Sign in to NCBI

RefSeq RefSeq

RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

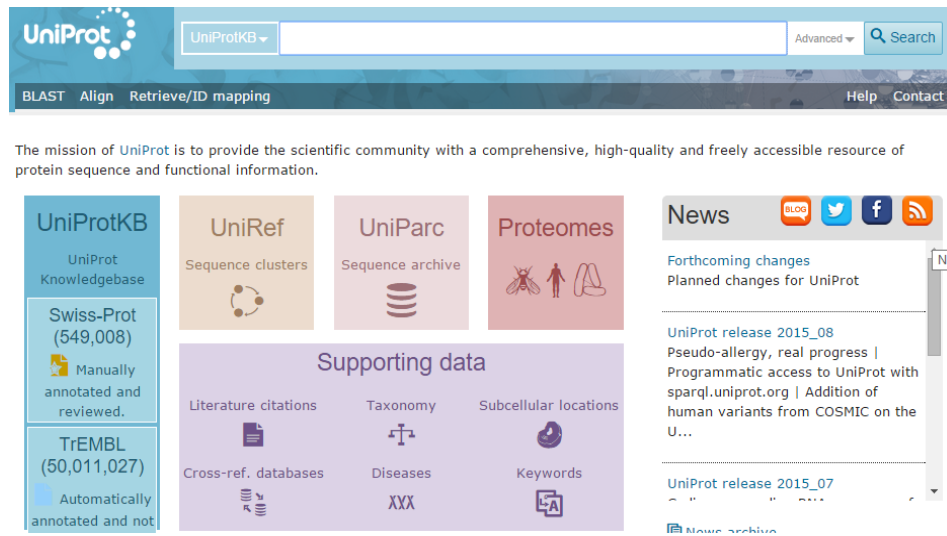
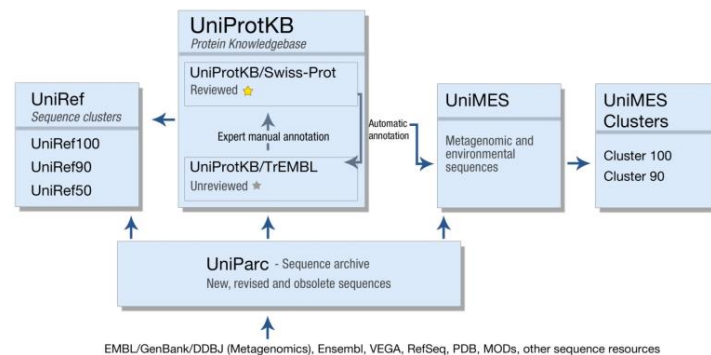


Figure 0.47 online database for proteins

Swiss-Prot contains human curated protein information

- Accession number, unique identifier
- The sequence
- Molecular mass
- Observed and predicted modifications



Protein sequences from various species and organisms can be found in uniprot. SwissProt is the manually annotated version of the UniProt Database.

Module 061: PROTEIN DATA BANK

Protein Data Bank is the premier resource of protein structures. These structures have been determined using experimental techniques. It's Open & Free

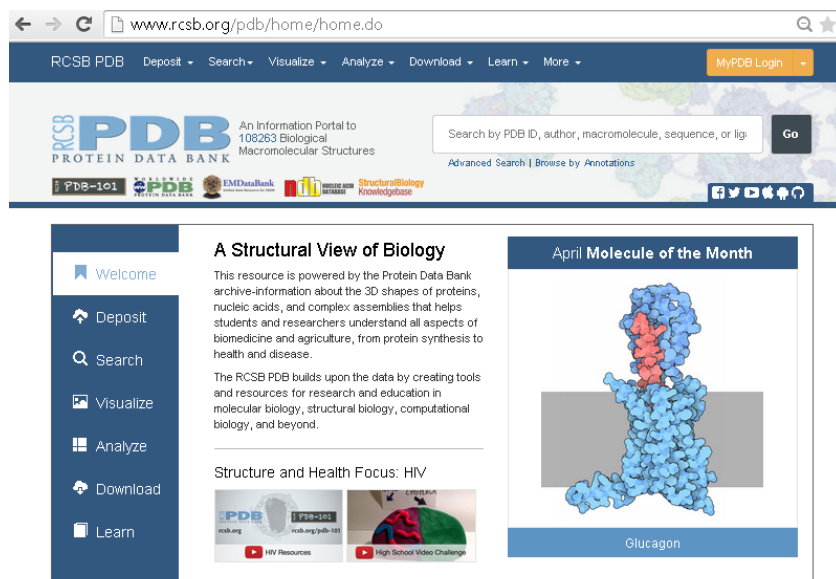


Figure 0.48 protein data bank

Summary 3D View Sequence Annotations Seq. Similarity 3D Similarity Literature DOI & Cites Methods Links

Solid-state NMR structure of ubiquitin **2MSG** Display Files Download Files Download Citation

DOI: 10.2210/pdb2msg/pdb

Primary Citation

Structural heterogeneity in microcrystalline ubiquitin studied by solid-state NMR.

Fasshuber, H.K., Lakomek, N.A., Habenstein, B., Loquet, A., Shi, C., Giller, K., Wolff, S., Becker, S., Lange, A.

Journal: (2015) Protein Sci.

PubMed: 25844685

DOI: 10.1002/pro.2854

Search Related Articles in PubMed

PubMed Abstract:

By applying [1-(13)C]- and [2-(13)C]-glucose labeling schemes to the folded globular protein ubiquitin, a strong reduction of spectral crowding and increase in resolution in solid-state NMR (ssNMR) spectra could be achieved. This allowed spectral resonance assignment in a... [Read More & Search PubMed Abstracts]

Molecular Description

Classification: Signaling Protein

Structure Weight: 8192.46

Molecule: Ubiquitin

Polymer: 1 **Type:** protein **Length:** 72

Chains: A

Fragment: UNP residue 1-72

Organism: Homo sapiens

Gene Name: UBB

UniProtKB: P0CG47

Structure Image

3D View More Images...

Downloadable viewers: Simple Viewer Protein Workshop Kiosk Viewer

MyPDB Personal Annotations Hide

To save personal annotations, please login to your MyPDB account.

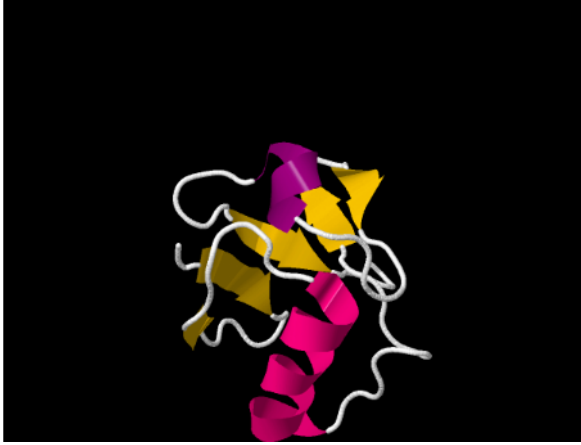
Deposition Summary Hide

Figure 0.49 P0CG47 - UBB_HUMAN

Summary 3D View Sequence Annotations Seq. Similarity 2D Similarity Literature Biol. & Chem. Methods Links

Solid-state NMR structure of ubiquitin

NOTE: Use your mouse to drag, rotate, and zoom in and out of the structure. ?



Structure Details ?

Structure PDB entry based on SOL

Symmetry Type Global Symmetry

Symmetry C1

Stoichiometry A

Select Orientation

Front

Select Display Mode ?

Secondary Structure

Subunit

Symmetry

Custom View

Figure 0.50 POCG47 - UBB_HUMAN

ATOM	1	N	MET	A	1	102.329	111.862	92.452	1.00	78.64	N
ATOM	2	CA	MET	A	1	103.332	112.165	93.516	1.00	77.39	C
ATOM	3	C	MET	A	1	103.877	113.584	93.255	1.00	76.87	C
ATOM	4	O	MET	A	1	103.802	114.075	92.129	1.00	78.54	O
ATOM	5	CB	MET	A	1	104.437	111.099	93.495	1.00	78.51	C
ATOM	6	CG	MET	A	1	105.176	110.881	94.812	1.00	76.25	C
ATOM	7	SD	MET	A	1	106.505	112.076	95.116	1.00	76.95	S
ATOM	8	CE	MET	A	1	107.077	111.551	96.763	1.00	73.31	C
ATOM	9	N	GLU	A	2	104.404	114.235	94.292	1.00	74.31	N
ATOM	10	CA	GLU	A	2	104.917	115.609	94.211	1.00	70.70	C
ATOM	11	C	GLU	A	2	105.963	115.909	93.143	1.00	68.10	C
ATOM	12	O	GLU	A	2	106.048	117.044	92.683	1.00	67.95	O
ATOM	13	CB	GLU	A	2	105.464	116.053	95.574	1.00	75.97	C
ATOM	14	CG	GLU	A	2	106.692	115.246	96.029	1.00	82.57	C
ATOM	15	CD	GLU	A	2	107.263	115.682	97.378	1.00	83.76	C
ATOM	16	OE1	GLU	A	2	106.611	115.396	98.412	1.00	86.90	O
ATOM	17	OE2	GLU	A	2	108.373	116.276	97.401	1.00	81.23	O
ATOM	18	N	ASN	A	3	106.789	114.924	92.784	1.00	64.50	N
ATOM	19	CA	ASN	A	3	107.834	115.134	91.773	1.00	59.93	C
ATOM	20	C	ASN	A	3	107.381	114.767	90.360	1.00	55.41	C
ATOM	21	O	ASN	A	3	108.159	114.856	89.416	1.00	53.33	O
ATOM	22	CB	ASN	A	3	109.086	114.308	92.095	1.00	62.56	C
ATOM	23	CG	ASN	A	3	109.531	114.441	93.535	1.00	62.42	C
ATOM	24	OD1	ASN	A	3	109.045	113.724	94.408	1.00	63.06	O
ATOM	25	ND2	ASN	A	3	110.484	115.326	93.787	1.00	62.31	N
ATOM	26	N	PHE	A	4	106.131	114.348	90.224	1.00	48.13	N
ATOM	27	CA	PHE	A	4	105.613	113.946	88.941	1.00	44.38	C
ATOM	28	C	PHE	A	4	104.326	114.601	88.521	1.00	48.72	C

Figure 0.51 searched results

Protein Data Bank provides Cartesian coordinates of each atom in the protein structure. Over 50,000 protein structures are reported and present in this database

Module 062: REVIEW OF SEQUENCE ALIGNMENT

We use next generation sequencing and whole genome sequencing to obtain the genetic information. For protein sequencing we use Mass Spectrometry and Edman Degradation

STORAGE:

- Sequence information is stored digitally
- Databases are designed to store sequence data
- Several databases exist depending on the type of sequence data

SHARING AND ACCESS:

- Sequence databases are shared via online websites
- Access to several such websites is free
- Data can be downloaded or searched on these website

USAGE OF DATA:

Sequence data can be used to obtain:

- Similarity of sequences
- Evolutionary History
- Predict the function of molecules

Module 063: GENBANK

- Developed by Swiss Bioinformatics Institute (SIB)
- Website provides access to databases and tools
- Proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc.

The screenshot shows the NCBI GenBank homepage. At the top, there's a navigation bar with 'NCBI', 'Resources', and 'How To'. Below this is a search bar with a dropdown menu set to 'Nucleotide' and a 'Search' button. A secondary navigation bar contains various database categories: GenBank, Submit, Genomes, WGS, HTGs, EST/GSS, Metagenomes, TPA, TSA, and INSDC. The main content area is divided into two columns. The left column, titled 'GenBank Overview', includes a section 'What is GenBank?' which describes the database as a collection of publicly available DNA sequences, part of the International Nucleotide Sequence Database Collaboration (INSDC). It also mentions release cycles and provides links to release notes and growth statistics. Below this is 'Access to GenBank', which lists several ways to search and retrieve data, including using Entrez Nucleotide, BLAST, and NCBI e-utils. The right column, titled 'GenBank Resources', contains links to GenBank Home, Submission Types, Submission Tools, Search GenBank, and Update GenBank Records. At the bottom of the left column is a section titled 'GenBank Data Usage'.

Figure 0.52 <http://www.ncbi.nlm.nih.gov/genbank/>

Several sequence, structure and molecular interaction databases exist. These are available online on the web. Users can freely access and download such data

Module 064: ENSEMBLE

As human brain is limited to remember and store the information for long time that's why we use online database for the storage of Molecular information.

ENSEMBLE is genome search engine which is used to search the genome of every recorded species.

You have been redirected to your nearest mirror. [Click here to go back to www.ensembl.org](http://www.ensembl.org) Close

Ensembl ASIA [BLAST/BLAT](#) [BioMart](#) [Tools](#) [Downloads](#) [Help & Documentation](#) [Blog](#) [Mirrors](#) Login/Register

Search all species...

Search: All species for Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Human GRCh38.p5 **Human** GRCh37

Mouse GRCm38.p4 **Zebrafish** GRCz10

★ [Log in to customize this list](#)

All genomes

-- Select a species --

[View full list of all Ensembl species](#)

asia.ensembl.org/Multi/psychic?r=rat:5:62797383-63627669

Still using Human GRCh37?

[Go to GRCh37](#)

Variant Effect Predictor

Gene expression in different tissues

Find SNPs and other variants for my gene

Retrieve gene sequence

Compare genes across species

[Use my own data in Ensembl](#) [ENCODE data in Ensembl](#)

What's New in Ensembl Release 84 (March 2016)

- 20 haematopoietic primary cell epigenomes from the BLUEPRINT project
- Mouse: update to Ensembl-Havana GENCODE gene set
- Track hub registry interface
- dbSNP 146 for Human, Cow and Dog
- Pairwise LD calculation on LD variant page

[Full details](#) | [All web updates, by release](#) | [More news on our blog](#)

- 29 Mar 2016: [Ensembl Genomes 31 has been released!](#)
- 10 Mar 2016: [Ensembl 84 has been released!](#)
- 16 Feb 2016: [Learn about Ensembl – online, live and free!](#)

[Go to Ensembl blog](#)

Tweets by @ensembl

Ensembl @ensembl
It's Spring! PERK14 is our #Geneoftheweek as it's involved in root formation #PlantOntology
buff.ly/1pJZXiC

<http://asia.ensembl.org/index.html>

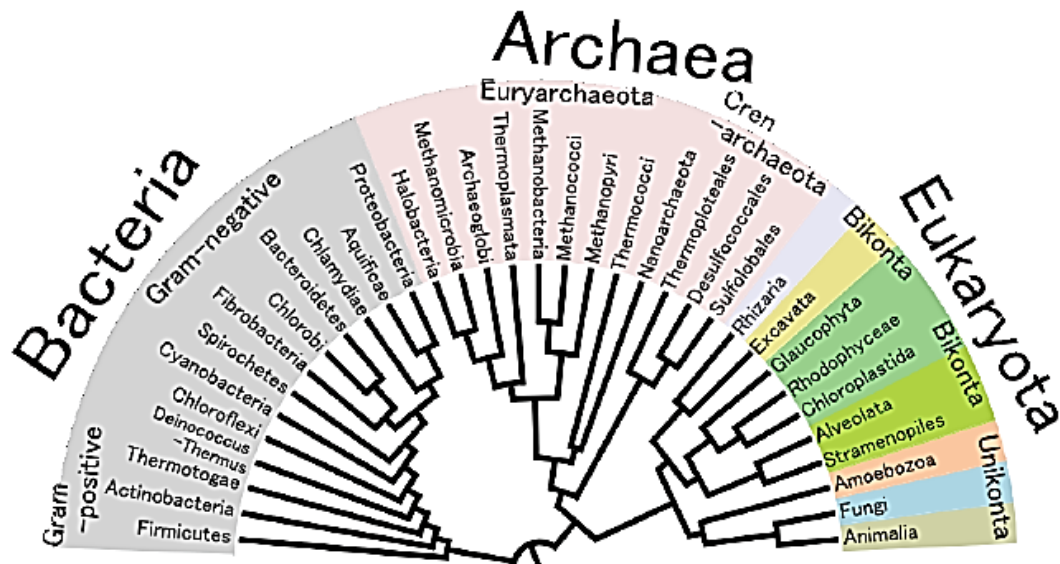
Chapter 3 - Molecular Evolution

Module 001: Molecular Evolution & Phylogeny

Molecular evolution is the process of change in the sequence composition of cellular molecules such as DNA, RNA, and proteins across generations. The field of molecular evolution uses principles of evolutionary biology and population genetics to explain patterns in these changes. Genes and Proteins are modified in this process.

All molecules have an evolutionary history. Phylogenetics is the science of studying evolutionary relationships. Phylogenetics has led to the creation of relationship trees between various species of Bacteria, Archaea, and Eukaryota.

(Page and Holmes 2009)



<http://bacterialphylogeny.info/overview.html>

Figure 0.1 Phylogenetic tree

Types of Phylogenetic Trees

Scaled Trees

- Branch lengths are equal to the magnitude of change in the nodes

Unscaled Trees

- Only representing the relationship between sequences

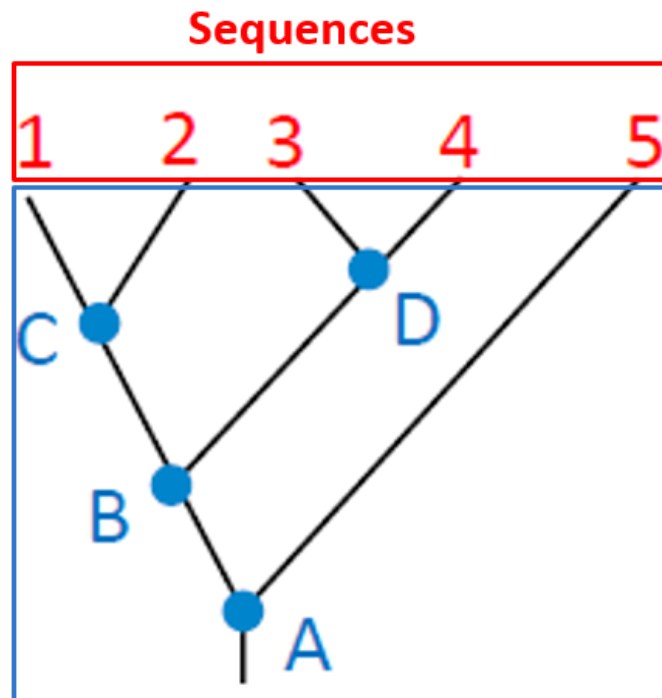


Figure 0.2 phylogenetic tree interference

Conclusion

Phylogenetics is the study of extracting evolutionary relationships between species. Sequence information from each species is used to measure the difference between the species.

Module 002: Evolution of Sequences

DNA acts as cellular memory unit and protein are the translated product of DNA coded information. And evaluation is very important to survive in different type of environments. There are some methods which brings change or evolution in any organism. (Kluger 2015)

Method of Change

DNA gets modified by:

- Mutation & Substitution
- Insertion
- Deletion

Discussion

Over time, species evolve to adapt to their circumstances. Since the environment and circumstances may be different for each species, they evolve uniquely. Unique evolutionary pressures may be encountered by each cell for struggle of life. However, in which sequence they are presented to the cells is also unique. Combinations of evolutionary factors are involve in evolution. The evolutionary events and their combination impart relationships between sequences. These relationships are explored in Phylogenetics .Several algorithms exist for finding such relationships

Kluger, M. J. (2015). Fever: its biology, evolution, and function, Princeton University Press.

Page, R. D. and E. C. Holmes (2009). Molecular evolution: a phylogenetic approach, John Wiley & Sons.

Module 003: Concepts and Terminologies - I

To understand the concept of evolution we follow some rules. Phylogenetics involves processing sequence information from different species to find evolutionary relationships. Output from such studies include Phylogenetic Trees

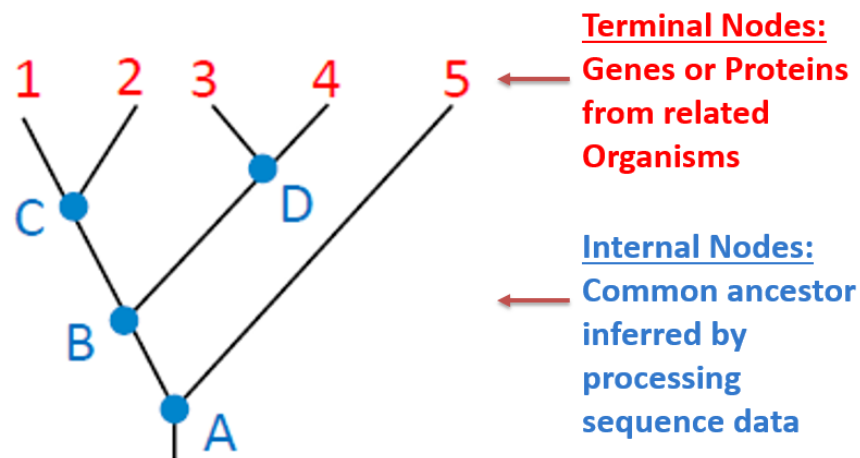
Phylogenetic Trees

Figure 3 phylogenetic tree from ancestor to evolution

In above figure the point A stands for ancestor and with the passage of time the evolution occurred with and the genome sequence of organisms changed.

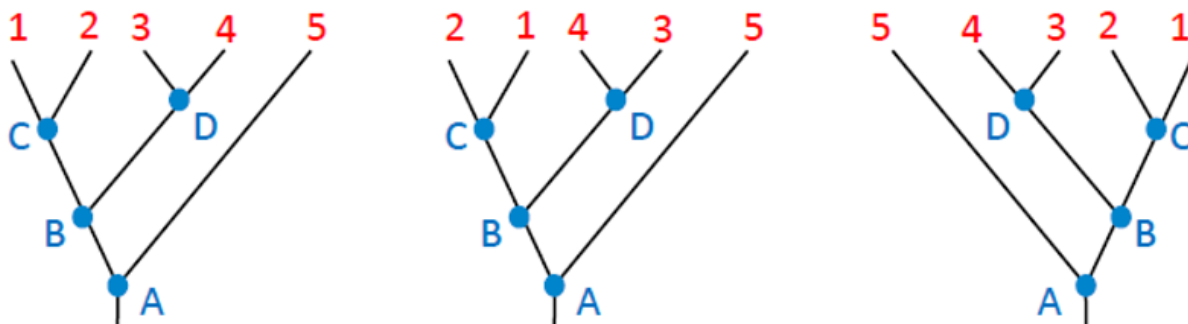


Figure 4 layout of trees

All trees have same meanings.

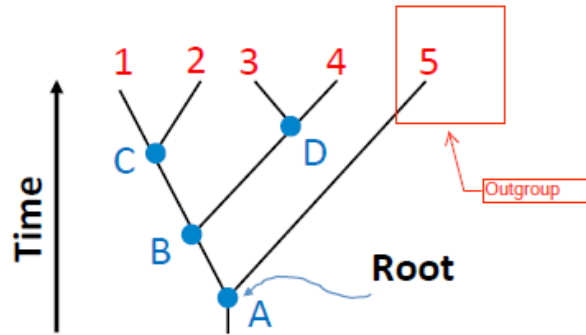


Figure 5 rooted tree

Root node is the ancestor of all other nodes. The direction of evolution is from ancestor to the terminal nodes.

Conclusion

Phylogenetics specifies evolutionary relationship with the help of trees. Trees can be rooted or unrooted. Rooted trees can show temporal evolutionary direction.

Module 004: Concepts and Terminologies - II

Rooted and Unrooted trees can be used to show phylogenetic relationships between sequences. Let's examine the properties of these trees further.

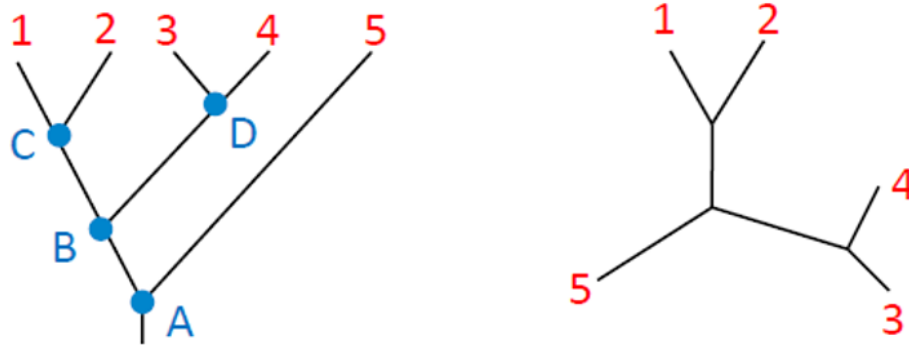


Figure 6 rooted tree vs unrooted

Rooted trees are computationally expensive.

http://everythingexplained.today/Computational_phylogenetics/

<https://github.com/joey711/phyloseq/issues/597>

Number of Sequences	Number of Rooted Trees	Number of Unrooted Trees
3	3	1
4	15	3
5	105	15
10	34,459,425	2,027,025
15	213,458,046,767,875	7,905,853,580,625
...	...	...

How many trees?

Rooted: $(2n-3)! / 2^{n-2}(n-2)!$

Unrooted: $(2n-5)! / 2^{n-3}(n-3)!$

Figure 7 computation comparison

Conclusion

Rooted and Unrooted trees have their own advantages and disadvantages. Depending on our requirement, we can choose between them.

Module 005: Algorithms and Techniques

Rooted and Unrooted trees can be used to show phylogenetic relationships between sequences. Several types of algorithms exist which are divided into two classes. There are many methods for constructing evolutionary trees.

Clustering Approach	Objective based Methods
UPGMA	Least Square Distances
WPGMA	Maximum Likelihood
Neighbor Joining	Maximum Parsimony
Single Linkage	
Complete Linkage	

Figure 9 construction methods

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is a simple agglomerative (bottom-up) [hierarchical clustering](#) method. The method is generally attributed to Sokal and Michener.

In this method two sequences with the shortest evolutionary distance between them are considered and these sequences will be the last to diverge, and represented by the most recent internal node.

Clustering Approach	Objective based Methods
UPGMA	Least Square Distances
WPGMA	Maximum Likelihood
Neighbor Joining	Maximum Parsimony
Single Linkage	
Complete Linkage	

Least Squares Distance Method. Branch lengths, represent the “observed” distances between sequences (i & j).

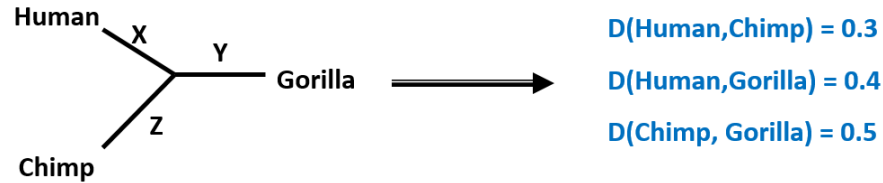


$$D(\text{Human}, \text{Chimp}) = 0.3$$

$$D(\text{Human}, \text{Gorilla}) = 0.4$$

$$D(\text{Chimp}, \text{Gorilla}) = 0.5$$

Find X, Y and Z such that $D(i, j)$ are conserved?



Conclusion

- Several methods exist for constructing phylogenetic trees.
- Broadly, they belong to objective methods or clustering methods.
- We will study UPGMA and Distance Methods.

Module 006: Introduction to UPGMA

Phylogenetic trees can be used to show phylogenetic relationships between sequences. To construct these trees, several types of algorithms exist which are divided into two classes.

UPGMA: Unweighted Pair – Group Method using arithmetic Averages

- Calculating distance between two clusters:

Cluster X + Cluster Y = Cluster Z

Calculate the distance of a cluster (e.g. W) to the new cluster Z

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

N_X is the number of sequences in cluster x

- **Calculating distance between two trees:**

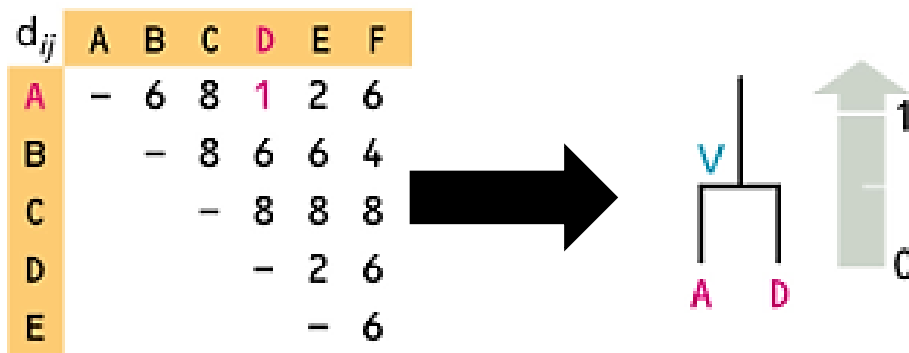
Assume we have N sequences

Cluster X has N_X sequences, cluster Y has N_Y sequences

d_{XY} : the evolutionary distance between X and Y

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$

- **Methods for constructing trees**



The distance matrix is obtained using pairwise sequence alignment.

- **Calculating distance between two clusters:**

Cluster X + Cluster Y = Cluster Z

Calculate the distance of a cluster (e.g. W) to the new cluster Z

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

N_X is the number of sequences in cluster x

- **Calculating distance between two trees:**

Assume we have N sequences

Cluster X has N_X sequences, cluster Y has N_Y sequences

d_{XY} : the evolutionary distance between X and Y

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$

- **Methods for constructing trees**

A – D becomes a new cluster let's say V. We have to modify the distance matrix. What are the distances between:

- V and B (Calculate),
- V and C,
- V and E,
- V and F.

d_{ij}	A	B	C	D	E	F
A	-	6	8	1	2	6
B		-	8	6	6	4
C			-	8	8	8
D				-	2	6
E					-	6

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

$$d_{VB} = \frac{N_A d_{AB} + N_D d_{DB}}{N_A + N_D} = \frac{1*6 + 1*6}{1+1} = 6$$

Conclusion

UPGMA is a clustering algorithm which can help us compute phylogenetic trees. We will see the detailed working of this approach in later modules.

Module 007: UPGMA-I

UPGMA has two components to it. These include distance calculations between two clusters and between two trees.

- **Building trees using UPGMA**

Combining Clusters: Cluster X + Cluster Y = Cluster Z

Calculate the distance of each cluster (e.g. W) to the new cluster Z

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

N_X is the number of sequences in cluster x

- **Calculating the distance between two trees**

Assume we have N sequences

Cluster X has N_X sequences, cluster Y has N_Y sequences

d_{XY} : the evolutionary distance between X and Y

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$

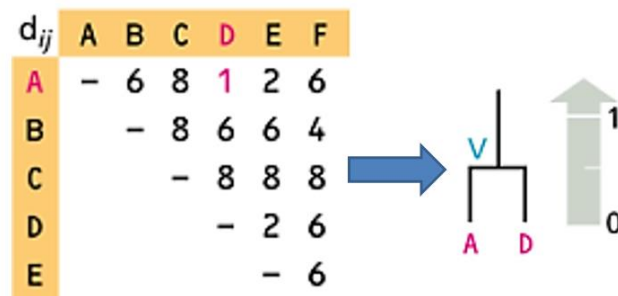


Figure 10 the distance matrix is obtained using pairwise sequence alignment

- **Methods for constructing trees**

A – D becomes a new cluster lets say V.

We have to modify the distance matrix!

What are the distances between:

d_{ij}	A	B	C	D	E	F
A	-	6	8	1	2	6
B		-	8	6	6	4
C			-	8	8	8
D				-	2	6
E					-	6

- V and B (Calculate),

- V and C,
- V and E,
- V and F.

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

$$d_{VB} = \frac{N_A d_{AB} + N_D d_{DB}}{N_A + N_D} = \frac{1*6 + 1*6}{1+1} = 6$$

Conclusion

UPGMA starts with creating clusters of sequences which are the closest. Next, distance is computed between the new cluster and the remaining sequences. The process is repeated for all sequences.

Module 008: UPGMA-II

UPGMA steps include distance calculations between two clusters and between two trees. We formed clusters from sequences which had the shortest distance.

Building trees using UPGMA

Combining Clusters: Cluster X + Cluster Y = Cluster Z

Calculate the distance of each cluster (e.g. W) to the new cluster Z

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

N_X is the number of sequences in cluster x

Calculating the distance between two trees

Assume we have N sequences

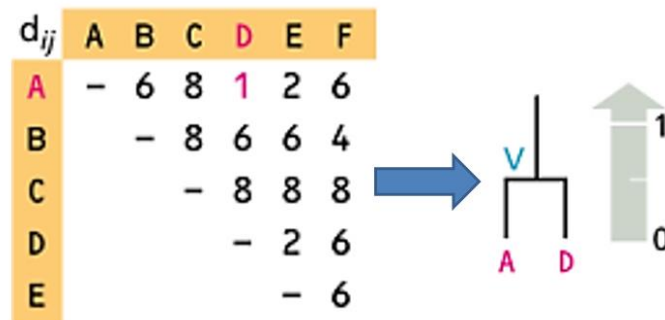
Cluster X has N_X sequences, cluster Y has N_Y sequences

d_{XY} : the evolutionary distance between X and Y

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$

Methods for constructing trees

The distance matrix is obtained using pairwise sequence alignment.



$$d_{vB} = \frac{N_A d_{AB} + N_D d_{DB}}{N_A + N_D} = \frac{1*6 + 1*6}{1+1} = 6$$

$$d_{vC} = \frac{N_A d_{AC} + N_D d_{DC}}{N_A + N_D} = \frac{1*8 + 1*8}{1+1} = 8$$

$$d_{vE} = \frac{N_A d_{AE} + N_D d_{DE}}{N_A + N_D} = \frac{1*2 + 1*2}{1+1} = 2$$

$$d_{vF} = \frac{N_A d_{AF} + N_D d_{DF}}{N_A + N_D} = \frac{1*6 + 1*6}{1+1} = 6$$

d_{ij}	A	B	C	D	E	F
A	-	6	8	1	2	6
B		-	8	6	6	4
C			-	8	8	8
D				-	2	6
E					-	6

→

d_{ij}	B	C	E	F	V
B	-	8	6	4	6
C		-	8	8	8
E			-	6	2
F				-	6

V – E becomes a new cluster lets say W

Now we have to modify the distance matrix again.

What are the distances between:

W and B,

W and C,

W and F.

Conclusion

Once a cluster is selected and its distance is computed with all other sequences, we update the distance matrix. Next, we select the shortest distance from the new matrix and repeat the process.

Module 009: UPGMA-III

UPGMA has two components to it. These include progressive distance calculations between two clusters or between two trees.

Building trees using UPGMA

Combining Clusters: Cluster X + Cluster Y = Cluster Z. Calculate the distance of each cluster (e.g. W) to the new cluster Z.

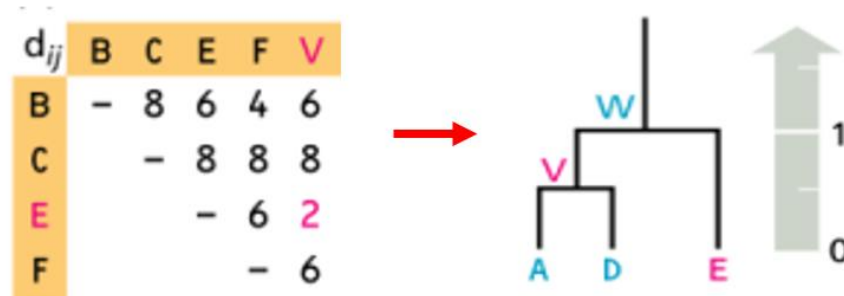
$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

N_X is the number of sequences in cluster x

Calculating the distance between two trees

- Assume we have N sequences
- Cluster X has N_X sequences, cluster Y has N_Y sequences
- d_{XY} : the evolutionary distance between X and Y

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$



V – E becomes a new cluster lets say W. Now we have to modify the distance matrix again.

What are the distances between:

W and B,

W and C,

W and F.

$$d_{WB} = \frac{N_V d_{VB} + N_E d_{EB}}{N_V + N_E} = \frac{2*6 + 1*6}{2+1} = 6$$

$$d_{WC} = \frac{N_V d_{VC} + N_E d_{EC}}{N_V + N_E} = \frac{2*8 + 1*8}{2+1} = 8$$

$$d_{WF} = \frac{N_V d_{VF} + N_E d_{EF}}{N_V + N_E} = \frac{2*6 + 1*6}{2+1} = 6$$

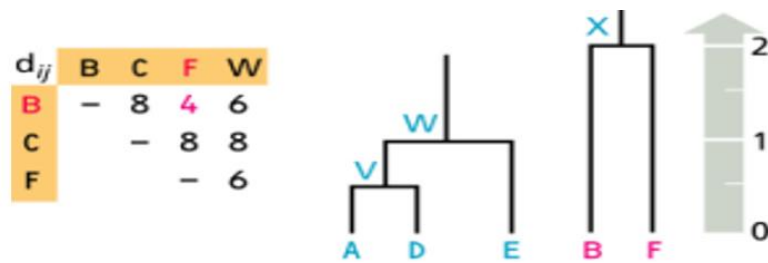
New matrix

d_{ij}	B	C	E	F	V
B	-	8	6	4	6
C		-	8	8	
E			-	6	2
F				-	6

→

d_{ij}	B	C	F	W
B	-	8	4	6
C		-	8	8
F			-	6

Cluster according to min distance

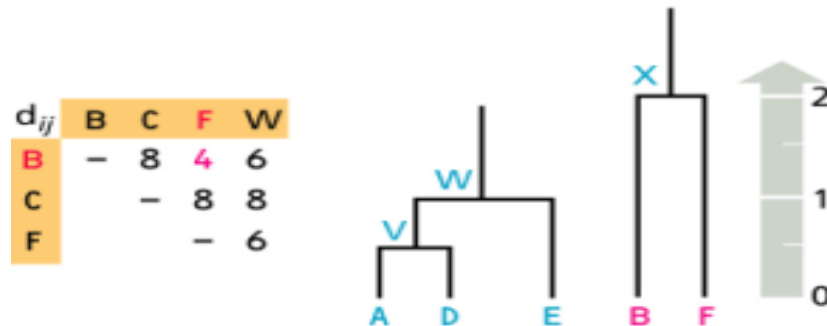


Conclusion

Now we have formed three clusters. Also, two separate trees have been formed. Next, we need to join these trees to create a complete tree.

Module 010: UPGMA-IV

Application of UPGMA resulted in formation of two sub-trees. The need now was to join them into a single tree. Let's see how that is done.



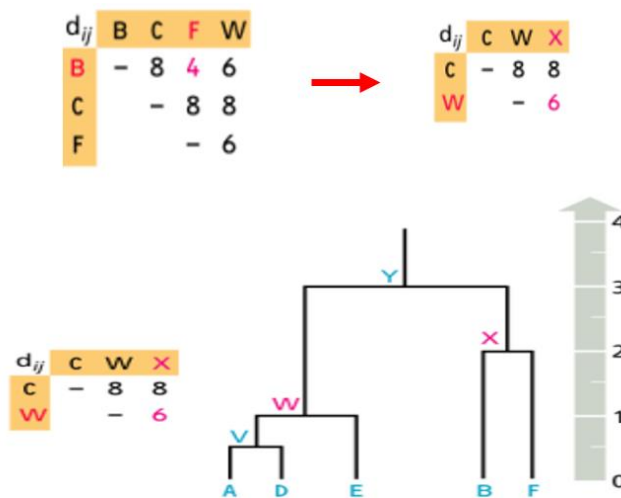
F – B becomes a new cluster let's say X. We have to modify the distance matrix yet again. What is the distance between trees: W and X.

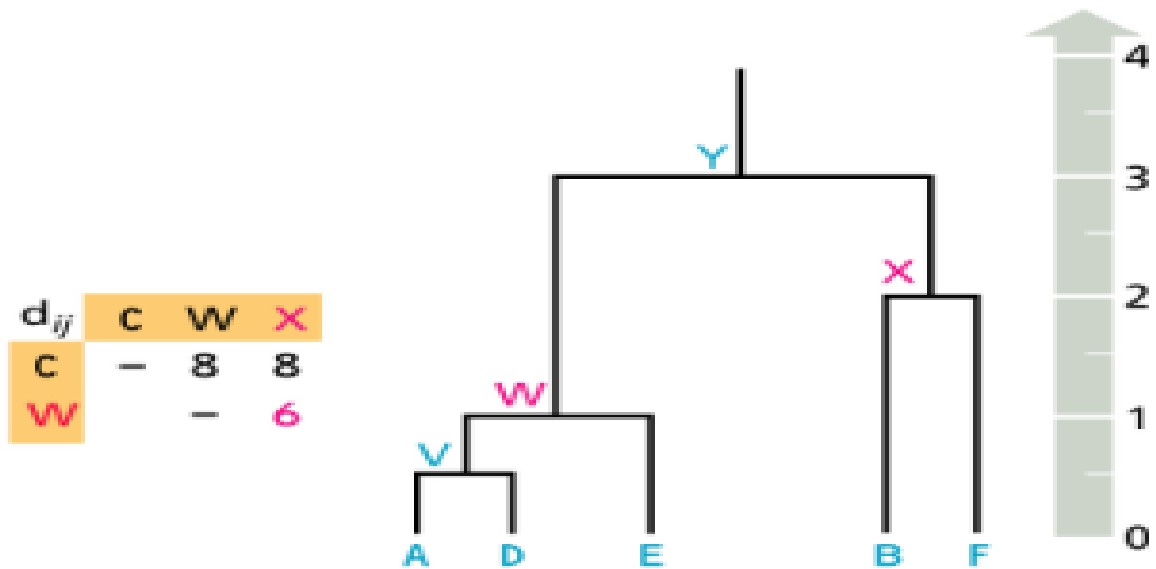
$$d_{WX} = \frac{1}{N_W N_X} \sum_{i \in W, j \in X} d_{ij} =$$

$$\frac{1}{N_W N_X} (d_{AB} + d_{AF} + d_{DB} + d_{DF} + d_{EB} + d_{EF}) =$$

$$\frac{1}{3 \times 2} * (6 + 6 + 6 + 6 + 6 + 6) = 6$$

d_{ij}	A	B	C	D	E	F
A	-	6	8	1	2	6
B		-	8	6	6	4
C			-	8	8	8
D				-	2	6
E					-	6





X – W becomes a new cluster lets say Y. We have to modify the distance matrix

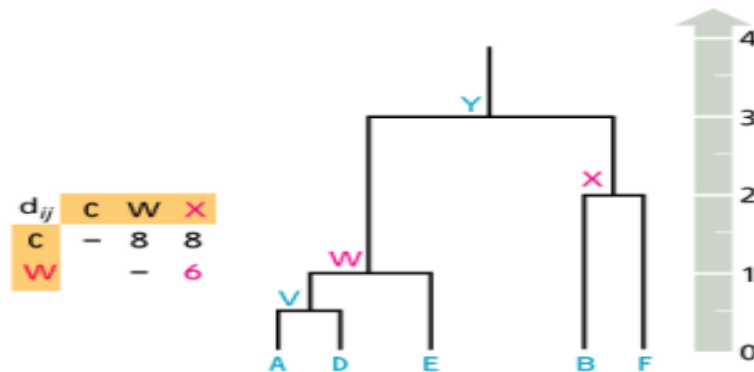
What is the distance between: Y and C.

Conclusion

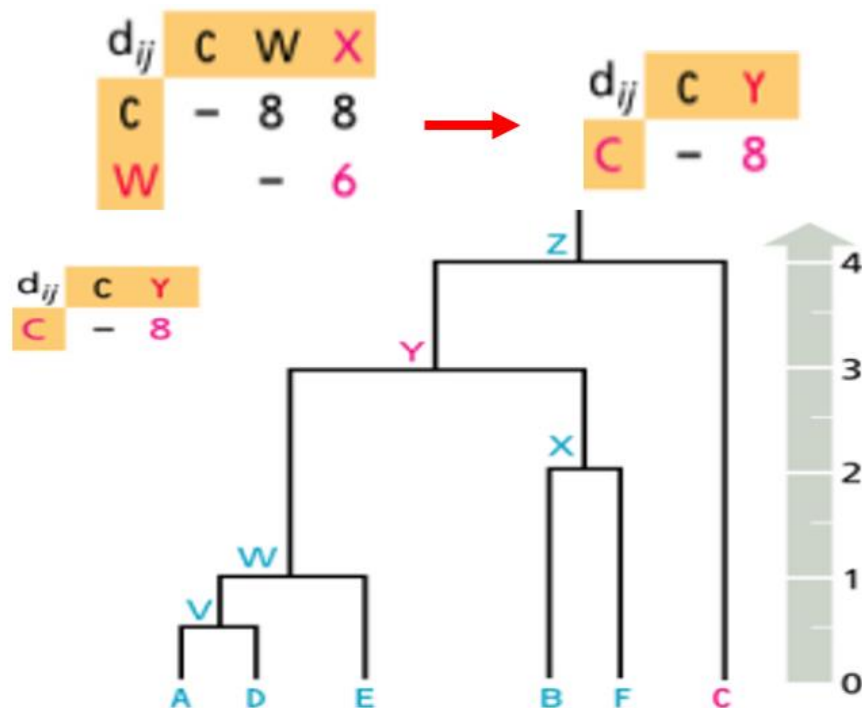
We have now seen how trees are generated and connected. Next, we need to finalize the tree by adding the last two clusters.

Module 011: UPGMA-V

Application of UPGMA resulted in formation of two sub-trees. The need now was to join them into a single tree. Let's see how that is done.

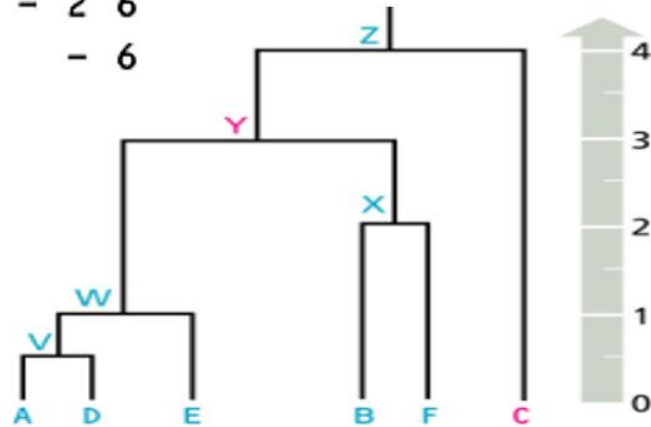


X – W becomes a new cluster let's say Y. We have to modify the distance matrix. What is the distance between: Y and C.



d_{ij}	A	B	C	D	E	F
A	-	6	8	1	2	6
B		-	8	6	6	4
C			-	8	8	8
D				-	2	6
E					-	6

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$



Conclusion

Un-weighted Pair Group Method using Arithmetic Averages is a clustering method to construct phylogenetic trees. Non-clustering methods such as Maximum Parsimony may be used for making trees as well.

Chapter 4 - RNA Secondary Structure Prediction

Module 001: DNA TO RNA SEQUENCES

• MOTIVATION

In early days RNA was considered as a structure which was involved between DNA and protein, means takes information from DNA and converts that information into protein synthesis. Now we know that it has multiple types like mRNA, tRNA, miRNA and siRNA. And they perform most of the work in gene expression and proteins. Not All RNA molecules are same, they differ in nucleotide sequences and functions also.

Many viruses assemble their genomes from RNAs. They are therefore called RNA viruses. Examples include Human Immunodeficiency Virus and Hepatitis C Virus.

There is little difference between RNA and DNA:

- Thymine is replaced by Uracil in RNA molecule.
- RNA molecule is single stand.
- RNA contain ribose sugar.

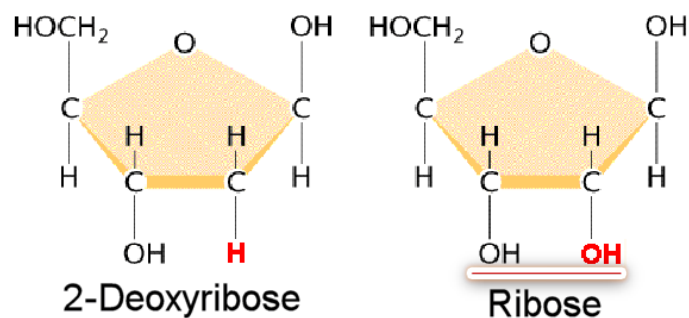


Figure 0.1 Ribose sugar has (OH) and Deoxyribose (H)

Because RNA has two (OH) groups that's why it has short life span because of both (OH) repulsion.

Module 002: TYPES OF RNA & THEIR FUNCTIONS

There are two categories of RNA:

- Coding RNA
- Non-Coding RNA

Coding RNA perform their coded function in protein synthesis. And Non-coding RNA helps in translation process.

• TYPES OF RNA

There are many types of RNA according to their functions like:

- Messenger RNA (mRNA)
- Transfer RNA (tRNA)
- Ribosomal RNA (rRNA)
- Micro RNAs (miRNA)
- Small Interfering RNA (siRNA)
-

• MESSENGER RNA

Only 5-10% of this RNA type is present in cell. Which has variable sequence, variable size and it carries the genetic information from DNA to Ribosomes where proteins to be assembled.

Messenger RNA 5' end is capped with (**7-Methyl Guanosine Triphosphate**) which helps the Ribosomes to identify the mRNA. And 3' end of the mRNA is poly A tail (around 30-200 adenylate residues) which help shield against 3' exonucleases)

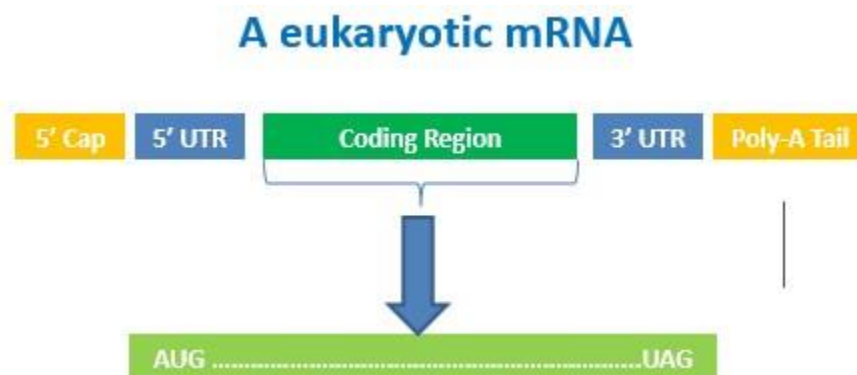


Figure 2 RNA sequence is complementary to the DNA sequence and is translated as codons of three nucleotides

As RNA has differ in nucleotides sequences therefore differ in functions.

Module 003: SIGNIFICANCES OF RNA STRUCTURE

RNA can form 3D structures {Sarver, 2008 #5}, such structural properties helps the RNA molecule to perform different functions.

As RNA is composed of sugars, phosphate and nucleotides and these nucleotides have ability to form hydrogen bonds.

- A' can make hydrogen bonds with 'U'
- 'G' makes hydrogen bonds with 'C'
- 'G' can also make hydrogen bonds with 'U' (Wobble Pair)

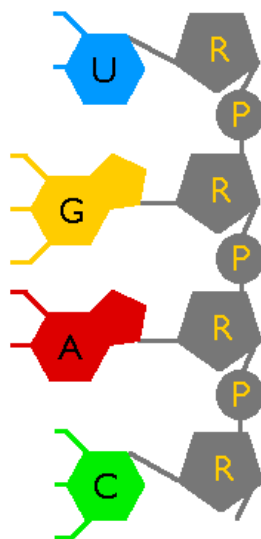


Figure 3 In RNA ribose is used in place of deoxyribose 3 In RNA uracil is used in place of thymine

Due to this ability of bonding RNA forms many structures and due to variety of structures RNA performs many functions in cell like:

- DNA information transfer

- Regulatory roles
- Catalytic roles
- Defense & immune response
- Structure-based special roles

Module 004: RNA FOLDING AND ENERGY FOLDING

RNA molecules form many structures for stability and different functions. "Gibbs Free Energy" (LANGRIDGE and KOLLMAN 1987) is the free energy available for RNA molecule for reactions and RNA structure formation takes place at this lower energy. In case if RNA has two structure we can select the one with lowest energy state.

http://chemwiki.ucdavis.edu/Core/Physical_Chemistry/Thermodynamics/State_Functions/Free_Energy/Gibbs_Free_Energy

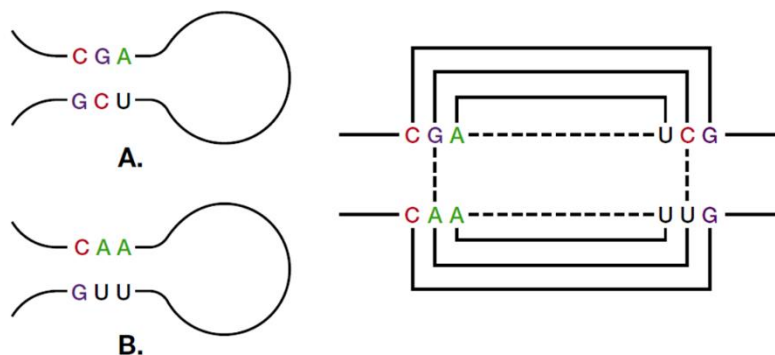


Figure 4 Energy is continuously given out as the RNA molecule folds by pairing complementary bases

		Y: A C G U			

X:	A	. . . -2.1			
	C	. . -3.3 .			
	G	. -2.4 . -1.4			
	U	-2.1 . -2.1 .			

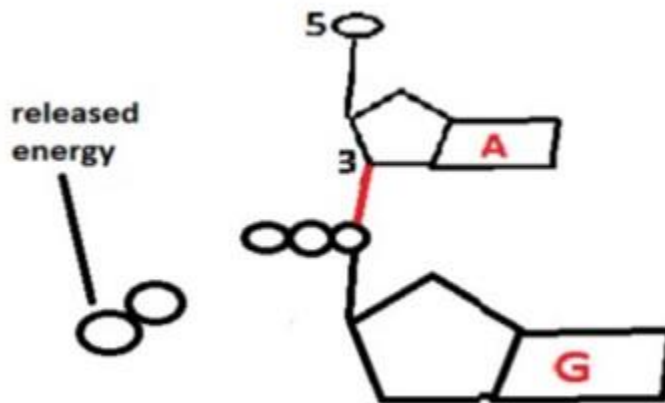
We can calculate the overall energy of RNA structures by summing up energies given out during the process of folding. For knowing the positive and negative values of calculations of stabilizing and destabilizing energies we may factor in ways in which RNA can be destabilized.

Figure 5 calculation of stabilizing and destabilizing values

Module 005: CALCULATING ENERGIES OF FOLDING-AN EXAMPLE

RNA is composed of four nucleotides (A, U, C and G) and these nucleotides are attached with ribose sugar in backbone. And these nucleotides have hydrogen bonding between them. G always bond with C and Always bonds with U through hydrogen bonding and energy is released.

That's why RNA molecule become more stable.



Y:	A	C	G	U
X:	A	C	G	U
A	.	.	-2.1	.
C	.	-3.3	.	.
G	-2.4	.	-1.4	.
U	-2.1	.	-2.1	.

Figure 0.2 nucleotides are held together by strong bonds which are created by the release of energy

"AUGUC.....GACAU"

A	—	U	= -2.1
U	—	A	= -2.1
G	—	C	= -2.4
U	—	A	= -2.1
C	—	G	= -3.3

5 nucleotides formed H-Bonds. This bond formation released energy (-12.0 kcal/mol) RNA molecule took up a 2' structure. Hence became more stable.

Module 006: TYPES OF RNA SECONDARY STRUCTURES-I

All the complimentary bases of RNA combine together to form RNA secondary structures. A simple nucleotide sequences of RNA is called as Primary structure and denoted by 1' while when these nucleotides fold together and form a complex structure that is called secondary structure and denoted by 2'.

The preferred structure of RNA is 2' which has many structural patterns like **Helices, Loops, Bulges and Junctions**

A. Single-stranded RNA

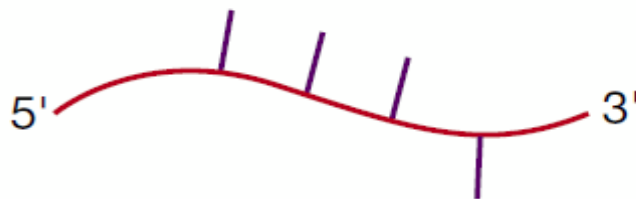


Figure 8 RNA sequence extends from its 5' end to 3' end. Upon folding, 3' end may fold on to the 5' end

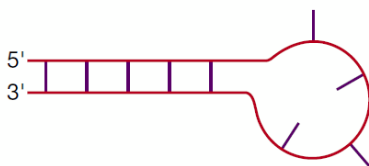
The first 2' RNA structure is called helix. Unlike the DNA helix, the RNA helix is formed when the RNA folds onto itself.

B. Double-stranded RNA helix of stacked base pairs



The second 2' structure is the hairpin loop

C. Stem and loop or hairpin loop.



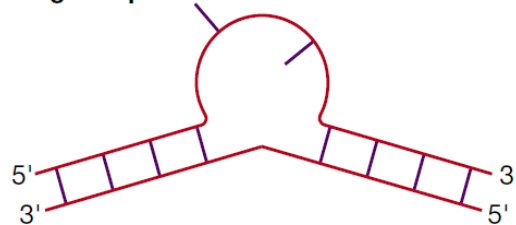
The loop of the hairpin must at least four bases long to avoid steric hindrance with base-pairing in the stem part of the structure.

Note that hairpins reverse the chemical direction of the RNA molecule.

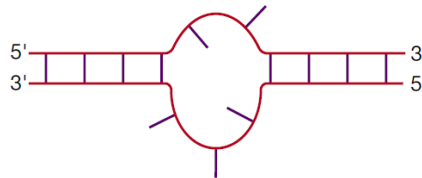
Module 007: TYPES OF RNA SECONDARY STRUCTURES-II

RNA 1' structure fold the (5'-3') ends and make RNA 2' structure just like helix and hairpin structure.

The third type of 2' structure is bulge loop.

D. Bulge loop

Bulges, are formed when a double-stranded region cannot form base pairs perfectly. Bulges can be asymmetric with varying number of base pairs on one side of the loop. Bulge loops are commonly found in helical segments of cellular RNAs and used to measure the helical twist of RNA in solution. (Tang and Draper 1990) The forth type of 2' RNA structure is interior loop.

E. Interior loop

Interior loops are formed by an asymmetric number of unpaired bases on each side of the loop. (Turner, Sugimoto et al. 1988)

Module 008: TYPES OF RNA SECONDARY STRUCTURES-III

Another 2' RNA structure is the Junction or Intersection.

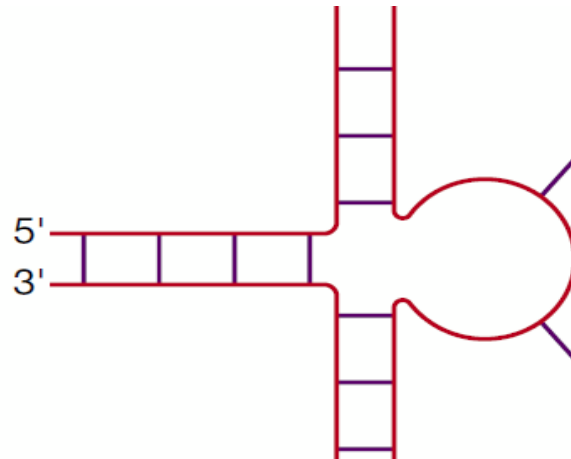


Figure 9 2' RNA structure called junction

Junctions include two or more double-stranded regions converging to form a closed structure. The unpaired bases appear as a bulge. (Zuker and Sankoff 1984)

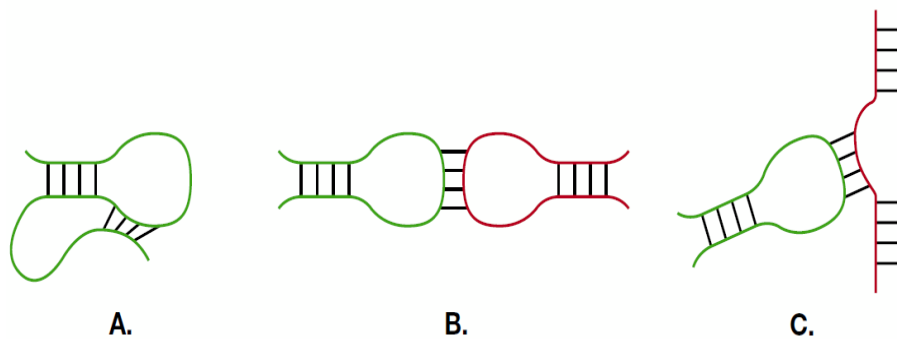


Figure 10 Unpaired bases in two 2' structures form hydrogen bonds with each other

RNA tertiary structures are formed when RNA unpaired base bond in 2' region bond.

Module 009: RNA TERTIARY STRUCTURES

2' RNA structures is formed due to folding of nucleotide with in RNA molecule but after folding some nucleotides remain open for interaction. And they form hydrogen bonds together.

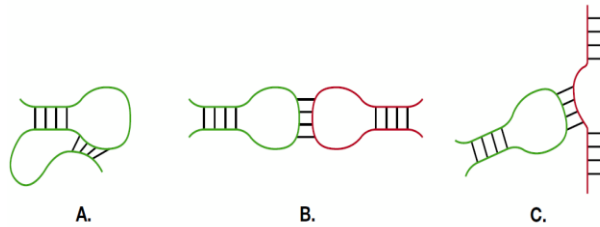
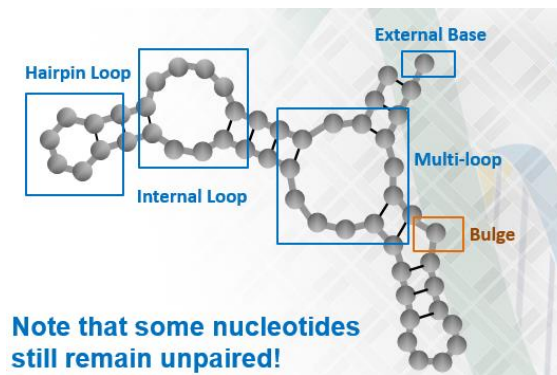


Figure 11 Hydrogen bonding formation in open nucleotides.

These unpaired nucleotides of 2' structure interact with other unpaired nucleotides and form a third structure called tertiary 3' structure. For example 4 nucleotides in hairpin loop structure does.



The above figure:

1. Indicate how these 2' structures come together
2. Indicate the difference between internal loop and multi loop
3. Indicate the yet unpaired bases

The unpaired bases in 3' structure remain paired by abnormal folding called (pseudoknots) but instead of pairing they remain available for pairing.

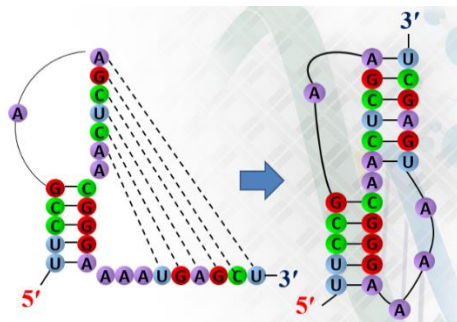
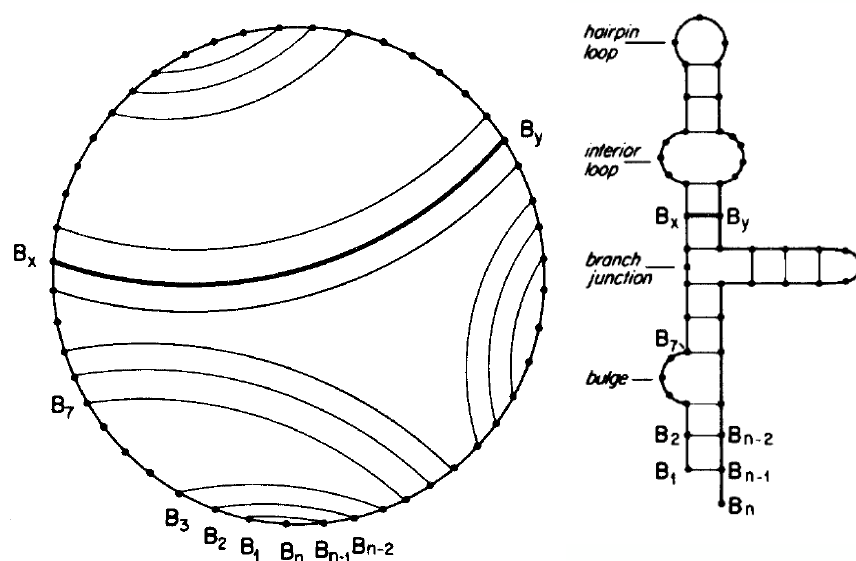


Figure 12 pseudoknots

Module 010: CIRCULAR REPRESENTATION OF STRUCTURES

Tertiary or 3' structure of RNA may form pseudoknots to detect the pseudoknots in RNA structure we need "circular plot" which is a graphical approach.



Intersecting arcs in circular plot are the pseudoknot.

Module 011: EXPERIMENTAL METHODS TO DETERMINE RNA STRUCTURES

RNA has 1', 2' and 3' structures. 1' has simple nucleotide sequence and 2' has nucleotides folding and 3' has knots.

For measuring the RNA structure we use **X-ray crystallography** (Smyth and Martin 2000), which works according to the principle of diffraction. Crystallized RNA diffracts X-rays which helps estimate atomic positions

All isotopes that contain an odd number of protons and/or of neutrons (see Isotope) have an intrinsic magnetic moment and angular momentum, in other words a nonzero spin, while all nuclides with even numbers of both have a total spin of zero. The most commonly studied nuclei are ^1H and ^{13}C , al

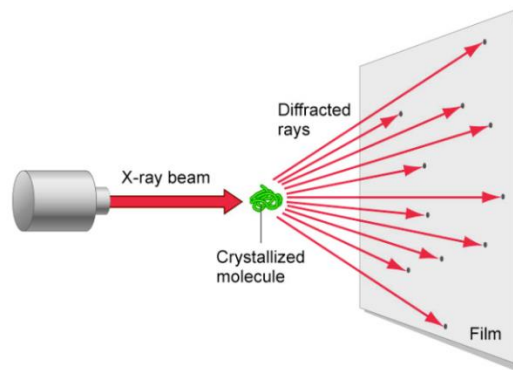


Figure 13 X-ray Crystallography

<https://260h.pbworks.com/w/page/30814223/X%20Ray%20Crystallography>

Another method to measure the RNA structure is called as **Atomic Force microscopy** in this technique a laser connected to a Si_3N_4 piezoelectric probe scans an RNA sample. It works well in air and liquid environment.

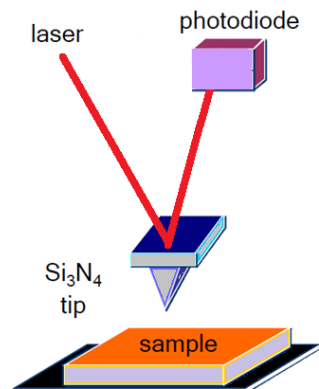
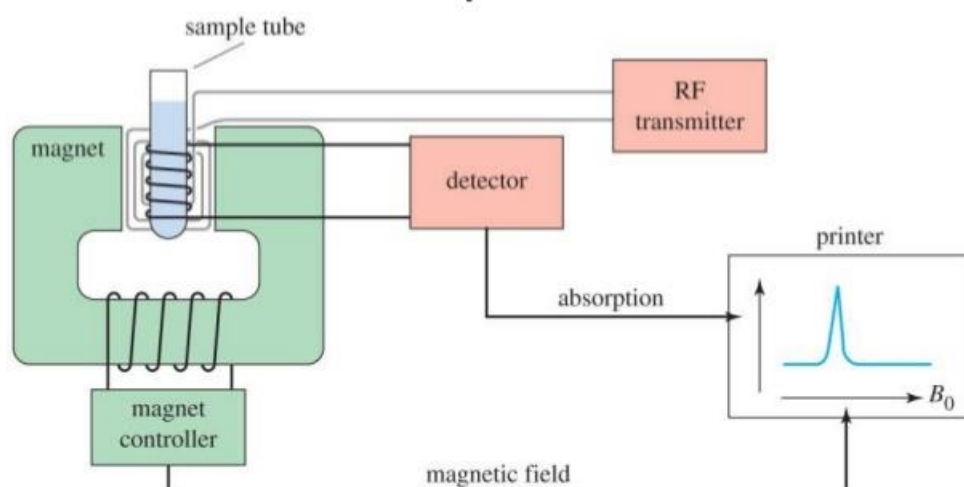


Figure 14 Atomic microscopy

The third method for measuring the RNA structure is **Nuclear Magnetic Resonance Imaging** in this method Hydrogen atoms in RNA resonate upon placement in a high magnetic field. It Works well without crystallizing RNA

The NMR Spectrometer



Copyright © 2010 Pearson Prentice Hall, Inc.

Chapter 13

13

<http://www.slideshare.net/Oatsmith/13-nuclear-magnetic-resonance-spectroscopy-wade-7th>

- **STORAGE OF STRUCTURES**

Reported structures are stored in online databases. Example includes RNA Bricks and RMDB etc. Bioinformaticians can refer to these databases for RNA structure studies

RNA Bricks is a database of RNA 3D structure motifs and their contacts, both with themselves and with proteins

Stanford University's RNA Mapping Database is an archive that contains results of diverse structural mapping experiments performed on ribonucleic acids.

Module 012: Strategies for RNA Structure Prediction

RNA structure 2' and 3' can be measured experimentally, but RNA molecule readily degrade due to their short shelf life.

Give 1' RNA structure creates the 2' structure because the simple nucleotides folds and form 2' structure. And on the base of folding we can predict the stability of the RNA molecule.

For example.

UAGUGUGUA (2 pairs)

UAGUGUGUA (1 pair)

UAGUGUGUA (1 pair)

Figure 15 pairs represent the stability of the RNA molecule

Maximizing the number of nucleotides can increase the structure and we have to select the structure according to the stability.

Module 013: Dot Plots for RNA 2' Structure Prediction

Structure measurement through experiments is slow and costly and there is maximum chances of more than one structure existence.