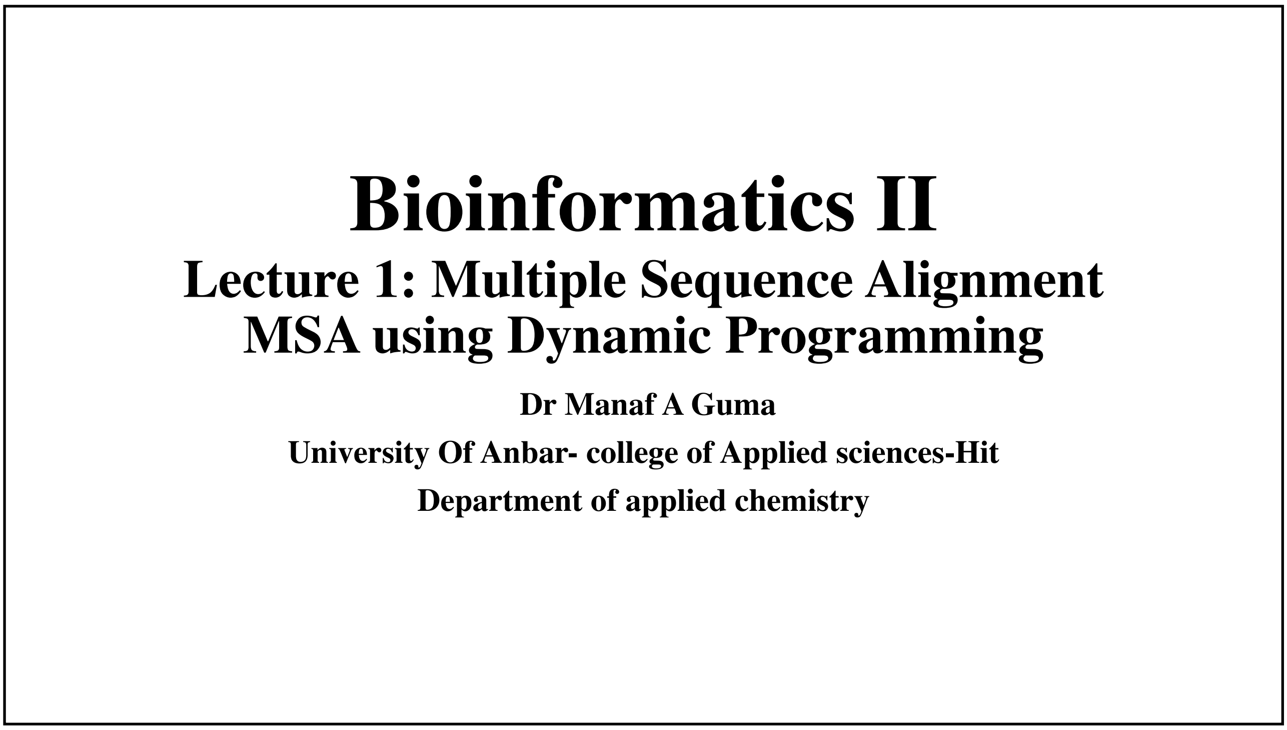


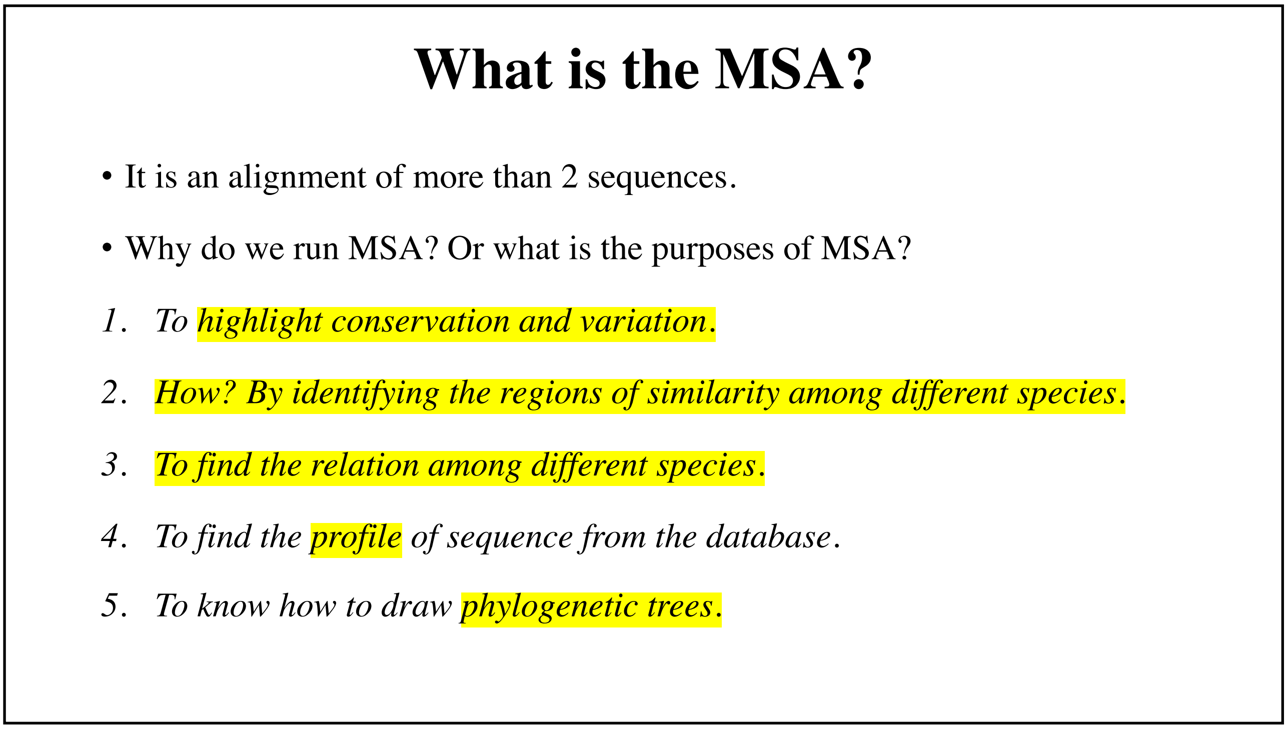
Bioinformatics II

Lecture 1: Multiple Sequence Alignment MSA using Dynamic Programming

Dr Manaf A Guma

University Of Anbar- college of Applied sciences-Hit

Department of applied chemistry

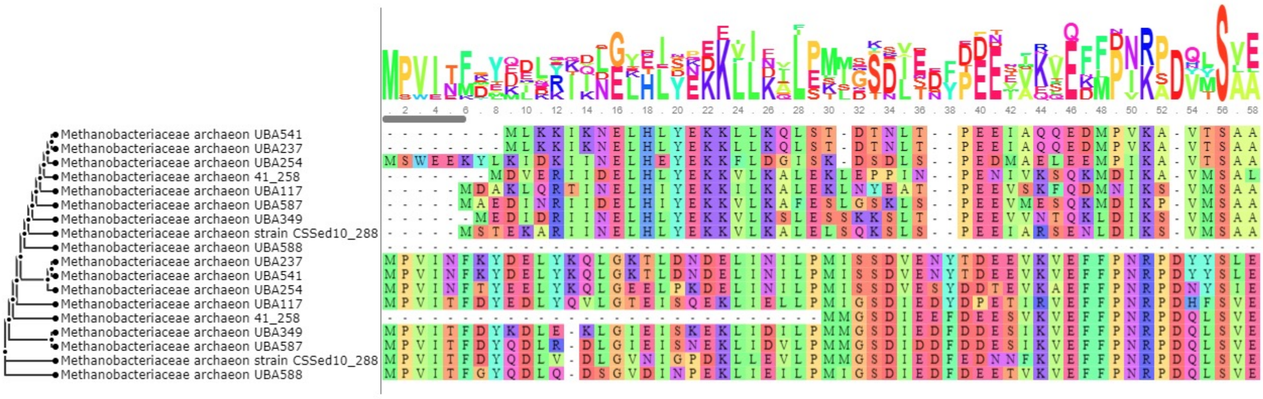


What is the MSA?

- It is an alignment of more than 2 sequences.
- Why do we run MSA? Or what is the purposes of MSA?

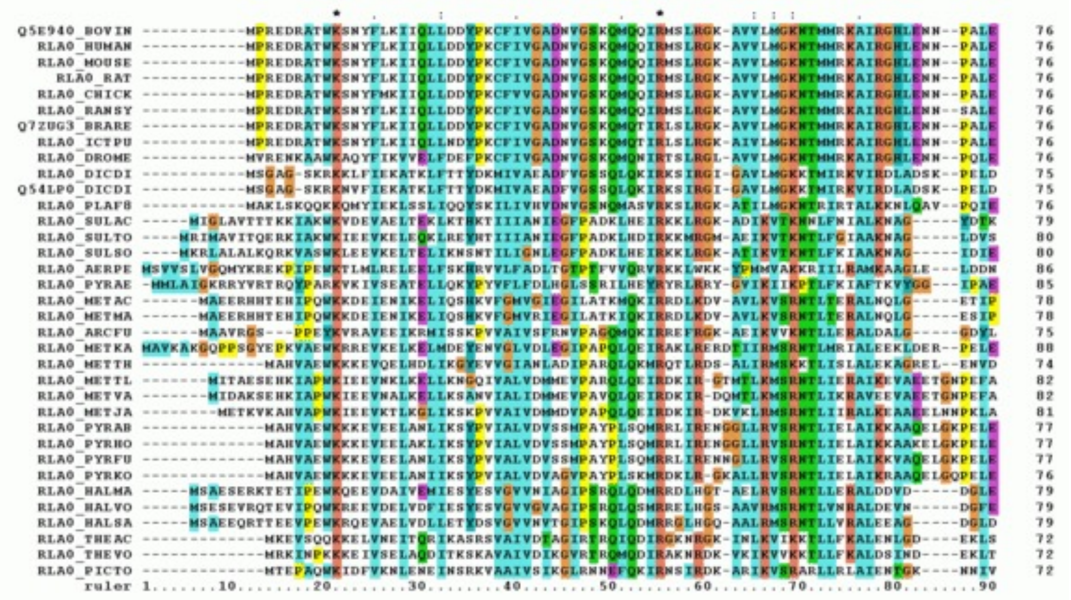
1. *To highlight conservation and variation.*
2. *How? By identifying the regions of similarity among different species.*
3. *To find the relation among different species.*
4. *To find the profile of sequence from the database.*
5. *To know how to draw phylogenetic trees.*

The purpose of MSA? *To highlight conservation and variation.*


Methanobacteriaceae archaeon UBA541
Methanobacteriaceae archaeon UBA237
Methanobacteriaceae archaeon UBA254
Methanobacteriaceae archaeon 41_258
Methanobacteriaceae archaeon UBA117
Methanobacteriaceae archaeon UBA587
Methanobacteriaceae archaeon UBA349
Methanobacteriaceae archaeon strain CSSed10_288
Methanobacteriaceae archaeon UBA588
Methanobacteriaceae archaeon UBA237
Methanobacteriaceae archaeon UBA541
Methanobacteriaceae archaeon UBA254
Methanobacteriaceae archaeon UBA117
Methanobacteriaceae archaeon 41_258
Methanobacteriaceae archaeon UBA349
Methanobacteriaceae archaeon UBA587
Methanobacteriaceae archaeon strain CSSed10_288
Methanobacteriaceae archaeon UBA588

3

When you have a huge data!

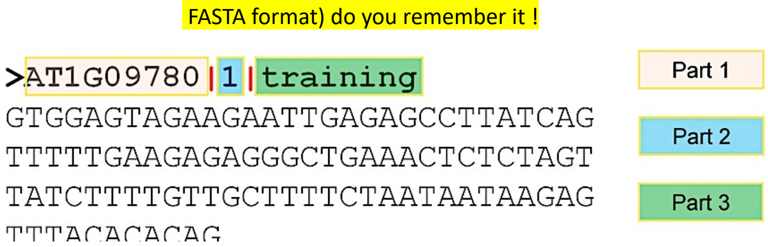


4

Why do we use dynamic programming in MSA?

- Because there is a huge database which makes the comparison very difficult if we run MSA by hand.
- Which software and websites are commonly used to do MSA?

1. BLAST.
2. FASTA.
3. ClustalW.



FASTA format) do you remember it !

```
>AT1G09780|1|training
GTGGAGTAGAAGAATTGAGAGCCTTATCAG
TTTTTGAAGAGAGGGCTGAAACTCTCTAGT
TATCTTTTGTTGCTTTTCTAATAATAAGAG
TTTACACACAG
```

5

How do you use BLAST to run MSA? (Tutorial)

1. We need to have a specific sequence for (protein or DNA for a specific species) that we need to find the similarity with it.
2. If we do not have it, then we go to https://www.uniprot.org and then find the Protein seq.
3. Copy the seq (in a FASTA format) do you remember it !
4. Open https://blast.ncbi.nlm.nih.gov/Blast.cgi and find blast protein-protein.
5. Paste the seq in the box labeled with Enter Query Sequence:
6. Click on BLAST to find the similarities.
7. The result will show the comparison (the identity and the scoring of the similarity) of the protein to various proteins in the database.
8. It also shows you the matrices used to generate the comparison.

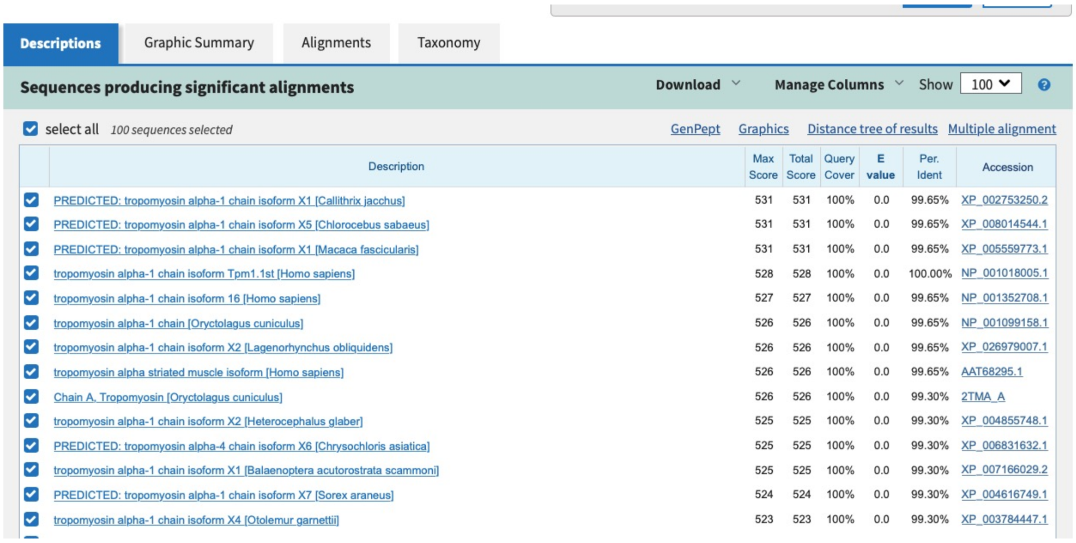
6

Can we get MSA form BLAST? What can we get?

- We can get only pairwise alignment using BLAST. (what is pairwise? do you remember?)
- But we can not get all of the sequences aligned together in the same screen using BLAST.
- We can get the profile of each sequence (the type of the species, the gene name and gene number etc.)

7

An example to see how BLAST works



Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	PREDICTED: tropomyosin alpha-1 chain isoform X1 [Callithrix jacchus]	531	531	100%	0.0	99.65%	XP_002753250.2
☑	PREDICTED: tropomyosin alpha-1 chain isoform X5 [Chlorocebus sabaeus]	531	531	100%	0.0	99.65%	XP_008014544.1
☑	PREDICTED: tropomyosin alpha-1 chain isoform X1 [Macaca fascicularis]	531	531	100%	0.0	99.65%	XP_005559773.1
☑	tropomyosin alpha-1 chain isoform Tpm1.1st [Homo sapiens]	528	528	100%	0.0	100.00%	NP_001018005.1
☑	tropomyosin alpha-1 chain isoform 16 [Homo sapiens]	527	527	100%	0.0	99.65%	NP_001352708.1
☑	tropomyosin alpha-1 chain [Oryctolagus cuniculus]	526	526	100%	0.0	99.65%	NP_001099158.1
☑	tropomyosin alpha-1 chain isoform X2 [Lagenorhynchus obliquidens]	526	526	100%	0.0	99.65%	XP_026979007.1
☑	tropomyosin alpha striated muscle isoform [Homo sapiens]	526	526	100%	0.0	99.65%	AAT68295.1
☑	Chain A, Tropomyosin [Oryctolagus cuniculus]	526	526	100%	0.0	99.30%	2TMA_A
☑	tropomyosin alpha-1 chain isoform X2 [Heterocephalus glaber]	525	525	100%	0.0	99.30%	XP_004855748.1
☑	PREDICTED: tropomyosin alpha-4 chain isoform X6 [Chrysochloris asiatica]	525	525	100%	0.0	99.30%	XP_006831632.1
☑	tropomyosin alpha-1 chain isoform X1 [Balaenoptera acutorostrata scammoni]	525	525	100%	0.0	99.30%	XP_007166029.2
☑	PREDICTED: tropomyosin alpha-1 chain isoform X7 [Sorex araneus]	524	524	100%	0.0	99.30%	XP_004616749.1
☑	tropomyosin alpha-1 chain isoform X4 [Otolemur garnettii]	523	523	100%	0.0	99.30%	XP_003784447.1

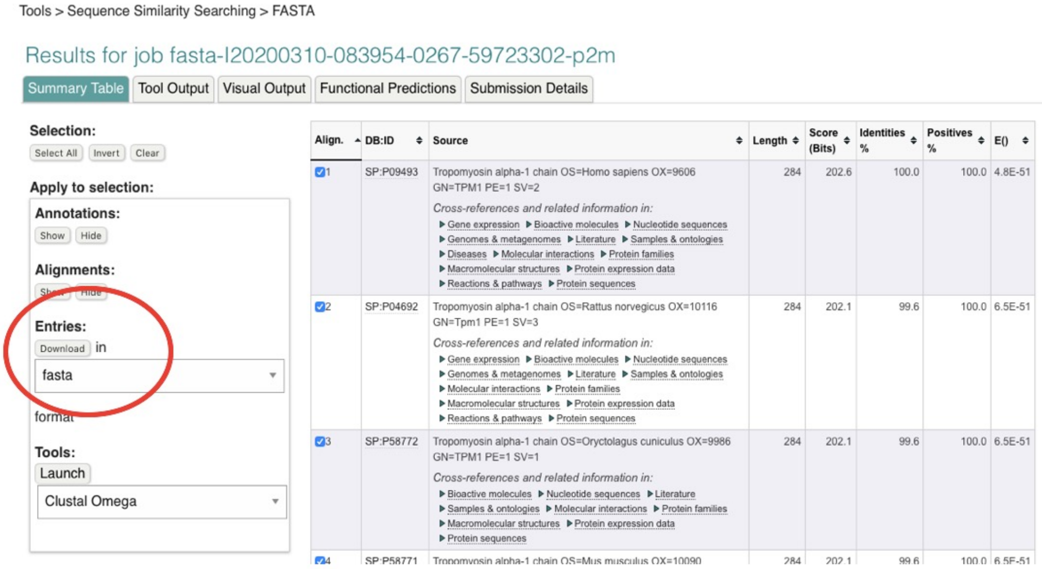
8

How do you use FASTA to run MSA?

1. Get the protein/DNA seq from https://www.uniprot.org.
2. copy the seq in FSATA format.
3. Open FASTA web page https://www.ebi.ac.uk/Tools/sss/fasta/.
4. Paste the seq.
5. The results will show different choses to get various bioinformatic analysis in a table.
6. You can show the MSA by clicking on visual output.
7. You can also download the seq by clicking on Download

9

The tables of FASTA results: an example



Tools > Sequence Similarity Searching > FASTA

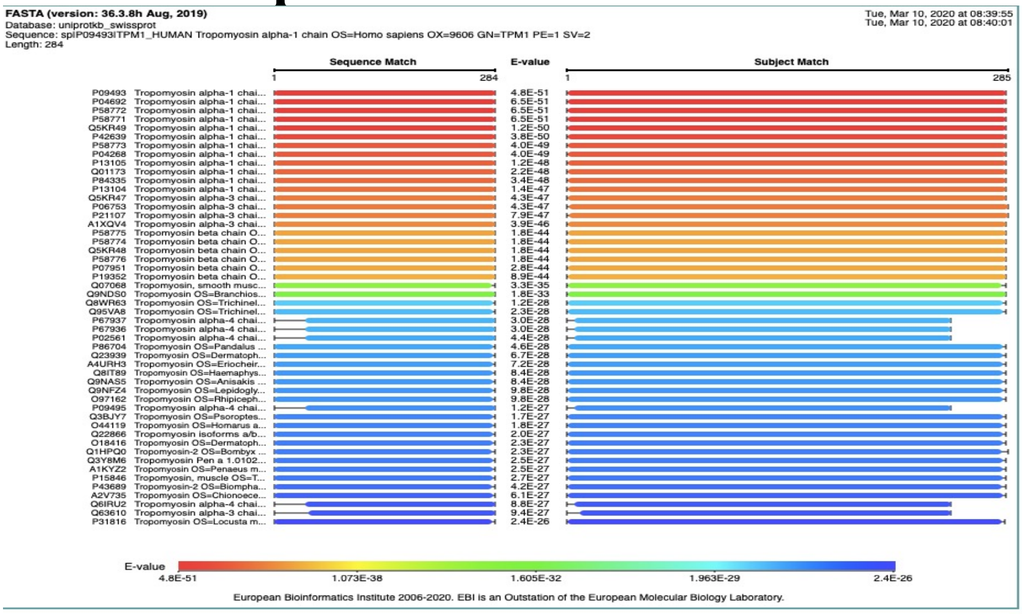
Results for job fasta-I20200310-083954-0267-59723302-p2m

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
☑1	SP:P09493	Tropomyosin alpha-1 chain OS=Homo sapiens OX=9606 GN=TPM1 PE=1 SV=2 *Cross-references and related information in:* ▶Gene expression ▶Bioactive molecules ▶Nucleotide sequences ▶Genomes & metagenomes ▶Literature ▶Samples & ontologies ▶Diseases ▶Molecular interactions ▶Protein families ▶Macromolecular structures ▶Protein expression data ▶Reactions & pathways ▶Protein sequences	284	202.6	100.0	100.0	4.8E-51
☑2	SP:P04692	Tropomyosin alpha-1 chain OS=Rattus norvegicus OX=10116 GN=Tpm1 PE=1 SV=3 *Cross-references and related information in:* ▶Gene expression ▶Bioactive molecules ▶Nucleotide sequences ▶Genomes & metagenomes ▶Literature ▶Samples & ontologies ▶Molecular interactions ▶Protein families ▶Macromolecular structures ▶Protein expression data ▶Reactions & pathways ▶Protein sequences	284	202.1	99.6	100.0	6.5E-51
☑3	SP:P58772	Tropomyosin alpha-1 chain OS=Oryctolagus cuniculus OX=9986 GN=TPM1 PE=1 SV=1 *Cross-references and related information in:* ▶Bioactive molecules ▶Nucleotide sequences ▶Literature ▶Samples & ontologies ▶Molecular interactions ▶Protein families ▶Macromolecular structures ▶Protein expression data ▶Protein sequences	284	202.1	99.6	100.0	6.5E-51
☑4	SP:P58771	Tropomyosin alpha-1 chain OS=Mus musculus OX=10090	284	202.1	99.6	100.0	6.5E-51

You can download all the seq form here

10

An example to see how FASTA works

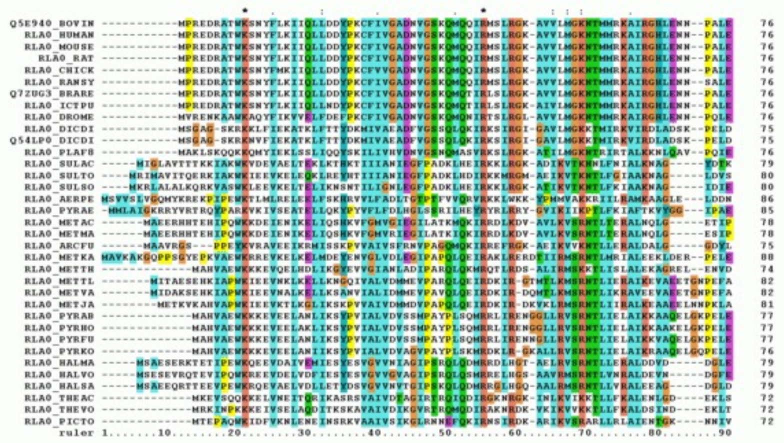


```
FASTA (version: 36.3.8h Aug, 2019)                                                    Tue, Mar 10, 2020 at 08:39:55
Database: uniprotkb_swissprot                                                         Tue, Mar 10, 2020 at 08:40:01
Sequence: spIP09493ITPM1_HUMAN Tropomyosin alpha-1 chain OS=Homo sapiens OX=9606 GN=TPM1 PE=1 SV=2
Length: 284

                                       Sequence Match        E-value           Subject Match
                                       1              284                      1                    285

  P09493  Tropomyosin alpha-1 chai...                        4.8E-51
  P04692  Tropomyosin alpha-1 chai...                        6.5E-51
  P58772  Tropomyosin alpha-1 chai...                        6.5E-51
  P58771  Tropomyosin alpha-1 chai...                        6.5E-51
  Q5KR49  Tropomyosin alpha-1 chai...                        1.2E-50
  P42639  Tropomyosin alpha-1 chai...                        3.8E-50
  P58773  Tropomyosin alpha-1 chai...                        4.0E-49
  P04268  Tropomyosin alpha-1 chai...                        4.0E-49
  P13105  Tropomyosin alpha-1 chai...                        1.2E-48
  Q01173  Tropomyosin alpha-1 chai...                        2.2E-48
  P84335  Tropomyosin alpha-1 chai...                        3.4E-48
  P13104  Tropomyosin alpha-1 chai...                        1.4E-47
  Q5KR47  Tropomyosin alpha-3 chai...                        4.3E-47
  P06753  Tropomyosin alpha-3 chai...                        4.3E-47
  P21107  Tropomyosin alpha-3 chai...                        7.9E-47
  A1XQV4  Tropomyosin alpha-3 chai...                        3.9E-46
  P58775  Tropomyosin beta chain O...                        1.8E-44
  P58774  Tropomyosin beta chain O...                        1.8E-44
  Q5KR48  Tropomyosin beta chain O...                        1.8E-44
  P58776  Tropomyosin beta chain O...                        1.8E-44
  P07951  Tropomyosin beta chain O...                        2.8E-44
  P19352  Tropomyosin beta chain O...                        8.9E-44
  Q07068  Tropomyosin, smooth musc...                        3.3E-35
  Q9NDS0  Tropomyosin OS=Branchios...                        1.8E-33
  Q8WR63  Tropomyosin OS=Trichinel...                        1.2E-28
  Q95VA8  Tropomyosin OS=Trichinel...                        2.3E-28
  P67937  Tropomyosin alpha-4 chai...                        3.0E-28
  P67936  Tropomyosin alpha-4 chai...                        3.0E-28
  P02561  Tropomyosin alpha-4 chai...                        4.4E-28
  P86704  Tropomyosin OS=Pandalus ...                        4.6E-28
  Q23939  Tropomyosin OS=Dermatoph...                        6.7E-28
  A4URH3  Tropomyosin OS=Eriocheir...                        7.2E-28
  Q8IT89  Tropomyosin OS=Haemaphys...                        8.4E-28
  Q9NAS5  Tropomyosin OS=Anisakis ...                        8.4E-28
  Q9NFZ4  Tropomyosin OS=Lepidogly...                        9.8E-28
  O97162  Tropomyosin OS=Rhipiceph...                        9.8E-28
  P09495  Tropomyosin alpha-4 chai...                        1.2E-27
  Q3BJY7  Tropomyosin OS=Psoroptes...                        1.7E-27
  O44119  Tropomyosin OS=Homarus a...                        1.8E-27
  Q22866  Tropomyosin isoforms a/b...                        2.0E-27
  O18416  Tropomyosin OS=Dermatoph...                        2.3E-27
  Q1HPQ0  Tropomyosin-2 OS=Bombyx ...                        2.3E-27
  Q3Y8M6  Tropomyosin Pen a 1.0102...                        2.5E-27
  A1KYZ2  Tropomyosin OS=Penaeus m...                        2.5E-27
  P15846  Tropomyosin, muscle OS=T...                        2.7E-27
  P43689  Tropomyosin-2 OS=Biompha...                        4.2E-27
  A2V735  Tropomyosin OS=Chionoece...                        6.1E-27
  Q6IRU2  Tropomyosin alpha-4 chai...                        8.8E-27
  Q63610  Tropomyosin alpha-3 chai...                        9.4E-27
  P31816  Tropomyosin OS=Locusta m...                        2.4E-26

  E-value
  4.8E-51        1.073E-38        1.605E-32        1.963E-29        2.4E-26

  European Bioinformatics Institute 2006-2020. EBI is an Outstation of the European Molecular Biology Laboratory.
```

11

We will carry on with MSA in the next lecture!

- How to get such like this !



```
                                       *   .      :      .          .    *            : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
   RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
   RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
     RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
   RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
   RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE   76
 Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
   RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE   76
   RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE   76
   RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD   75
 Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD   75
   RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE   76
   RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK   79
   RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS   80
   RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE   80
   RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEEKLFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN   86
   RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE   85
   RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP   78
   RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP   78
   RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL   75
   RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE   88
   RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD   74
   RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA   82
   RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA   82
   RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA   81
   RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE   77
   RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE   77
   RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE   77
   RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE   76
   RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE   79
   RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE   79
   RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD   79
   RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS   72
   RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT   72
   RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV   72
        ruler 1.......10........20........30........40........50........60........70........80........90
```

12

Bioinformatics I

Lecture 2: Multiple Sequence Alignment MSA using Dynamic Programming

Dr Manaf A Guma

University Of Anbar- college of Applied sciences-Hit

Department of applied chemistry

Obtaining MSA using ClustalW databases?

- **ClustalW is the "classic" MSA tool using C++ programming made by JD Thompson, DG Higgins, and TJ Gibson.**
- The original publication describing ClustalW is one of the 100 most cited publications in 'web of science'.
- Why do we run ClustalW? Or what is the purposes of ClustalW?

1. *To find the MSA.*
2. *Again, why do we find MSA? highlight conservation and variation.*
3. *How? By identifying the regions of similarity among different species.*
4. *To find the relation among different species.*
5. *To find the profile of sequence from the database.*
6. *To know how to draw phylogenetic trees.*

How CLUSTAL W deals with MSA??

CLUSTAL W: deals with multiple sequence alignment through:

1. **Sequence weighting.**
2. **position-specific gap penalties**
3. **weight matrix choice.**

- **What is the lasts version of** ClustalW?
- ClustalW It is an old version, the version is Clustal Omega which is much faster and better tools are available. http://www.ebi.ac.uk/Tools/msa/

3

How do you use ClustalW to run MSA? (very common)

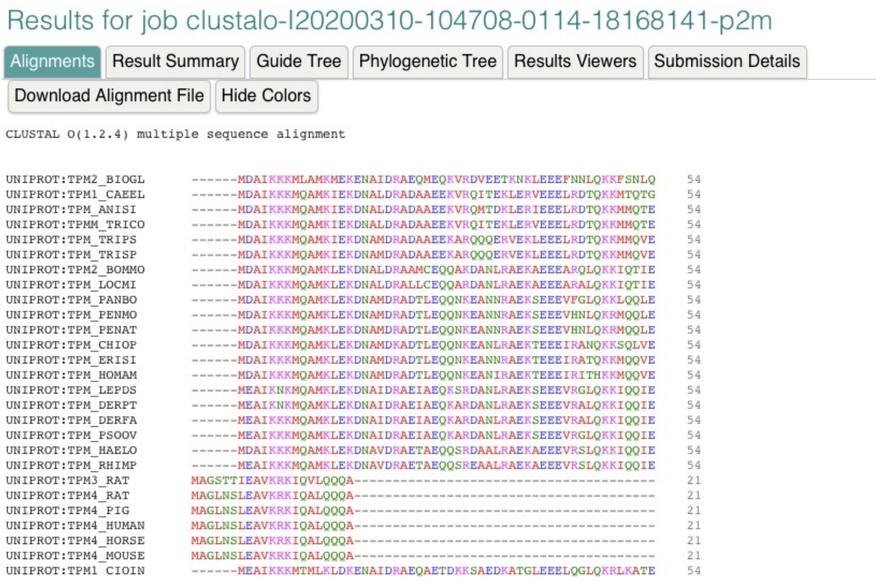
1. Get the protein/DNA seq from https://www.uniprot.org.
2. copy the seq in FSATA to download multiple seq.
3. Open FASTA web page

https://www.ebi.ac.uk/Tools/msa/clustalo/

1. Paste the multiple seq in the box.
2. Run the FASTA omega. You can color it.
3. You see also the phylogenetic tree as well. p

4

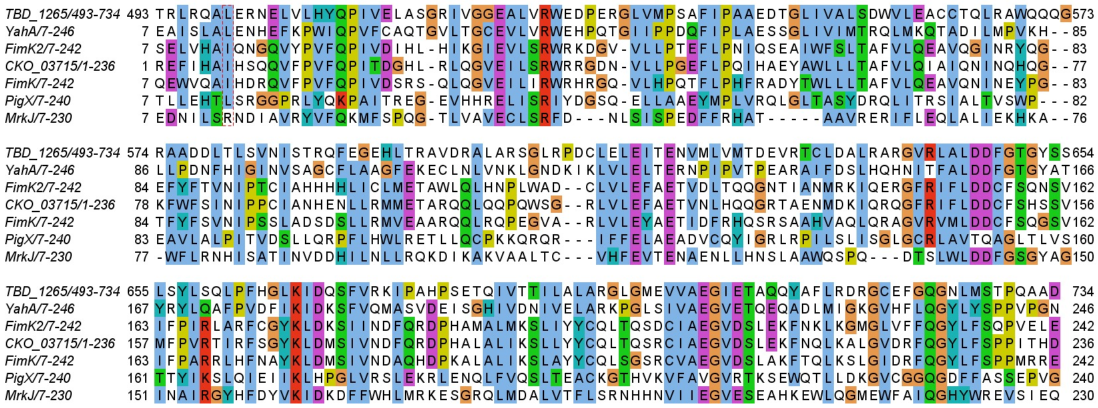
An example of ClustalW Omega



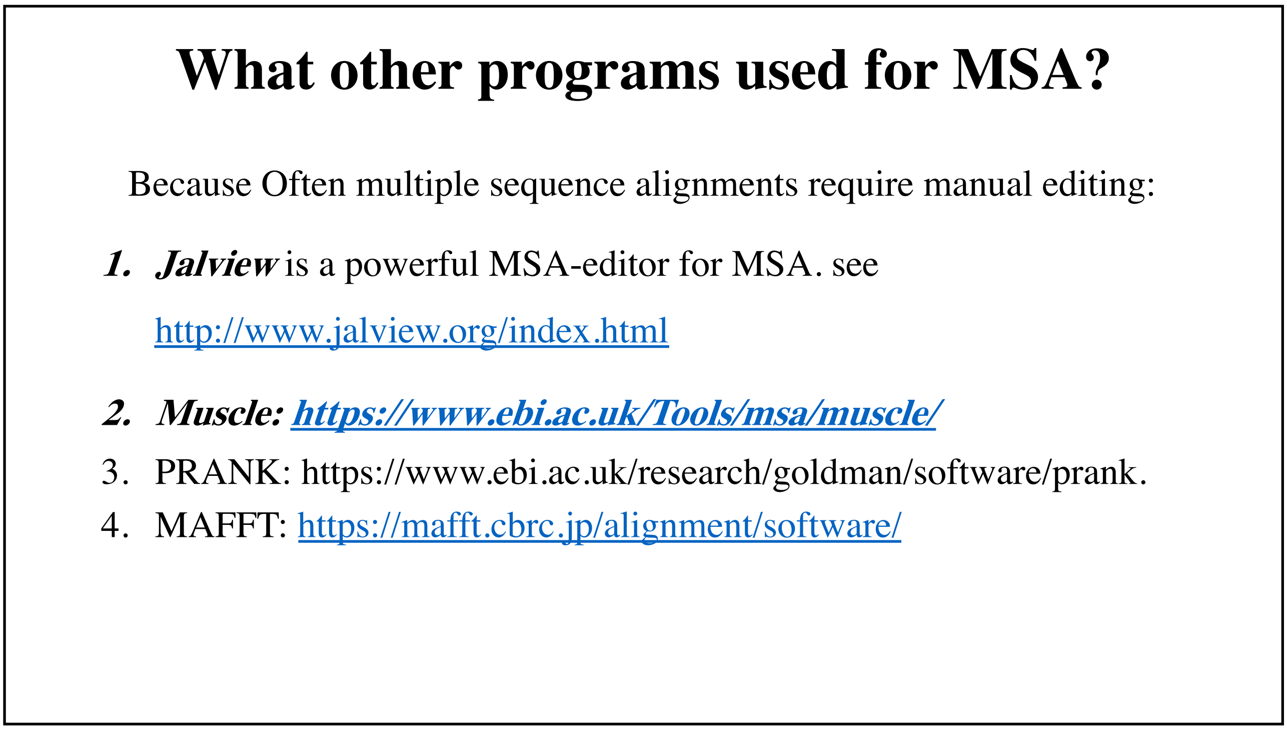
Results for job clustalo-I20200310-104708-0114-18168141-p2m

```
CLUSTAL O(1.2.4) multiple sequence alignment


UNIPROT:TPM2_BIOGL      ------MDAIKKKMLAMKMEKENAIDRAEQMEQKVRDVEETKNKLEEEFNNLQKKFSNLQ    54
UNIPROT:TPM1_CAEEL      ------MDAIKKKMQAMKIEKDNALDRADAAEEKVRQITEKLERVEEELRDTQKKMTQTG    54
UNIPROT:TPM_ANISI       ------MDAIKKKMQAMKIEKDNALDRADAAEEKVRQMTDKLERIEEELRDTQKKMMQTE    54
UNIPROT:TPMM_TRICO      ------MDAIKKKMQAMKIEKDNALDRADAAEEKVRQITEKLERVEEELRDTQKKMMQTE    54
UNIPROT:TPM_TRIPS       ------MDAIKKKMQAMKIEKDNAMDRADAAEEKARQQQERVEKLEEELRDTQKKMMQVE    54
UNIPROT:TPM_TRISP       ------MDAIKKKMQAMKIEKDNAMDRADAAEEKARQQQERVEKLEEELRDTQKKMMQVE    54
UNIPROT:TPM2_BOMMO      ------MDAIKKKMQAMKLEKDNALDRAAMCEQQAKDANLRAEKAEEEARQLQKKIQTIE    54
UNIPROT:TPM_LOCMI       ------MDAIKKKMQAMKLEKDNALDRALLCEQQARDANLRAEKAEEEARALQKKIQTIE    54
UNIPROT:TPM_PANBO       ------MDAIKKKMQAMKLEKDNAMDRADTLEQQNKEANNRAEKSEEEVFGLQKKLQQLE    54
UNIPROT:TPM_PENMO       ------MDAIKKKMQAMKLEKDNAMDRADTLEQQNKEANNRAEKSEEEVHNLQKRMQQLE    54
UNIPROT:TPM_PENAT       ------MDAIKKKMQAMKLEKDNAMDRADTLEQQNKEANNRAEKSEEEVHNLQKRMQQLE    54
UNIPROT:TPM_CHIOP       ------MDAIKKKMQAMKLEKDNAMDKADTLEQQNKEANLRAEKTEEEIRANQKKSQLVE    54
UNIPROT:TPM_ERISI       ------MDAIKKKMQAMKLEKDNAMDRADTLEQQNKEANNRAEKTEEEIRATQKKMQQVE    54
UNIPROT:TPM_HOMAM       ------MDAIKKKMQAMKLEKDNAMDRADTLEQQNKEANIRAEKTEEEIRITHKKMQQVE    54
UNIPROT:TPM_LEPDS       ------MEAIKNKMQAMKLEKDNAIDRAEIAEQKSRDANLRAEKSEEEVRGLQKKIQQIE    54
UNIPROT:TPM_DERPT       ------MEAIKNKMQAMKLEKDNAIDRAEIAEQKARDANLRAEKSEEEVRALQKKIQQIE    54
UNIPROT:TPM_DERFA       ------MEAIKKKMQAMKLEKDNAIDRAEIAEQKARDANLRAEKSEEEVRALQKKIQQIE    54
UNIPROT:TPM_PSOOV       ------MEAIKKKMQAMKLEKDNAIDRAEIAEQKARDANLRAEKSEEEVRGLQKKIQQIE    54
UNIPROT:TPM_HAELO       ------MDAIKKKMQAMKLEKDNAVDRAETAEQQSRDAALRAEKAEEEVRSLQKKIQQIE    54
UNIPROT:TPM_RHIMP       ------MEAIKKKMQAMKLEKDNAVDRAETAEQQSREAALRAEKAEEEVRSLQKKIQQIE    54
UNIPROT:TPM3_RAT        MAGSTTIEAVKRKIQVLQQQA---------------------------------------    21
UNIPROT:TPM4_RAT        MAGLNSLEAVKRKIQALQQQA---------------------------------------    21
UNIPROT:TPM4_PIG        MAGLNSLEAVKRKIQALQQQA---------------------------------------    21
UNIPROT:TPM4_HUMAN      MAGLNSLEAVKRKIQALQQQA---------------------------------------    21
UNIPROT:TPM4_HORSE      MAGLNSLEAVKRKIQALQQQA---------------------------------------    21
UNIPROT:TPM4_MOUSE      MAGLNSLEAVKRKIQALQQQA---------------------------------------    21
UNIPROT:TPM1 CIOIN      ------MEAIKKKMTMLKLDKENAIDRAEQAETDKKSAEDKATGLEEELQGLQKRLKATE    54
```

The old version presentation of the ClustalW



```
TBD_1265/493-734 493 TRLRQALERNELVLHYQPIVELASGRIVGGEALVRWEDPERGLVMPSAFIPAAEDTGLIVALSDWVLEACCTQLRAWQQQG573
YahA/7-246         7 EAISLALENHEFKPWIQPVFCAQTGVLTGCEVLVRWEHPQTGIIPPDQFIPLAESSGLIVIMTRQLMKQTADILMPVKH--85
FimK2/7-242        7 SELVHAIQNGQVYPVFQPIVDIHL-HIKGIEVLSRWRKDGV-VLLPTEFLPNIQSEAIWFSLTAFVLQEAVQGINRYQG--83
CKO_03715/1-236    1 REFIHAIHSQQVFPVFQPITDGHL-RLQGVEILSRWRRGDN-VLLPGEFLPQIHAEYAWLLLTAFVLQIAIQNINQHQG--77
FimK/7-242         7 QEWVQAIHDRQVFPVFQPIVDSRS-QLQGVEILIRWRHRGQ-VLHPQTFLPHFRADYTWLLLTAFVLQEAVQNINEYPG--83
PigX/7-240         7 TLLEHTLSRGGPRLYQKPAITREG-EVHHRELISRIYDGSQ-ELLAAEYMPLVRQLGLTASYDRQLITRSIALTVSWP---82
MrkJ/7-230         7 EDNILSRNDIAVRYVFQKMFSPQG-TLVAVECLSRFD---NLSISPEDFFRHAT-----AAVRERIFLEQLALIEKHKA--76

TBD_1265/493-734 574 RAADDLTLSVNISTRQFEGEHLTRAVDRALARSGLRPDCLELEITENVMLVMTDEVRTCLDALRARGVRLALDDFGTGYSS654
YahA/7-246        86 LLPDNFHIGINVSAGCFLAAGFEKECLNLVNKLGNDKIKLVLELTERNPIPVTPEARAIFDSLHQHNITFALDDFGTGYAT166
FimK2/7-242       84 EFYFTVNIPTCIAHHHHLICLMETAWLQLHNPLWAD--CLVLEFAETVDLTQQGNTIANMRKIQERGFRIFLDDCFSQNSV162
CKO_03715/1-236   78 KFWFSINIPPCIANHENLLRMMETARQQLQQPQWSG--RLVLEFAETVNLHQQGRTAENMDKIQRQGFRIFLDDCFSHSSV156
FimK/7-242        84 TFYFSVNIPSSLADSDSLLRMVEAARQQLRQPEGVA--RLVLEYAETIDFRHQSRSAAHVAQLQRAGVRVMLDDCFSQGSV162
PigX/7-240        83 EAVLALPITVDSLLQRPFLHWLRETLLQCPKKQRQR---IFFELAEADVCQYIGRLRPILSLISGLGCRLAVTQAGLTLVS160
MrkJ/7-230        77 -WFLRNHISATINVDDHILNLLRQKDIKAKVAALTC---VHFEVTENAENLLHNSLAAWQSPQ---DTSLWLDDFGSGYAG150

TBD_1265/493-734 655 LSYLSQLPFHGLKIDQSFVRKIPAHPSETQIVTTILALARGLGMEVVAEGIETAQQYAFLRDRGCEFGQGNLMSTPQAAD 734
YahA/7-246       167 YRYLQAFPVDFIKIDKSFVQMASVDEISGHIVDNIVELARKPGLSIVAEGVETQEQADLMIGKGVHFLQGYLYSPPVPGN 246
FimK2/7-242      163 IFPIRLARFCGYKLDKSIINDFQRDPHAMALMKSLIYYCQLTQSDCIAEGVDSLEKFNKLKGMGLVFFQGYLFSQPVELE 242
CKO_03715/1-236  157 MFPVRTIRFSGYKLDMSIVNDFQRDPHALALIKSLLYYCQLTQSRCIAEGVDSLEKFNQLKALGVDRFQGYLFSPPITHD 236
FimK/7-242       163 IFPARRLHFNAYKLDMSIVNDAQHDPKALALIKSLAYYCQLSGSRCVAEGVDSLAKFTQLKSLGIDRFQGYLFSPPMRRE 242
PigX/7-240       161 TTYIKSLQIEIIKLHPGLVRSLEKRLENQLFVQSLTEACKGTHVKVFAVGVRTKSEWQTLLDKGVCGGQGDFFASSEPVG 240
MrkJ/7-230       151 INAIRGYHFDYVKIDKDFFWHLMRKESGRQLMDALVTFLSRNHHNVIIEGVESEAHKEWLQGMEWFAIQGHYWREVSIEQ 230
```



What other programs used for MSA?

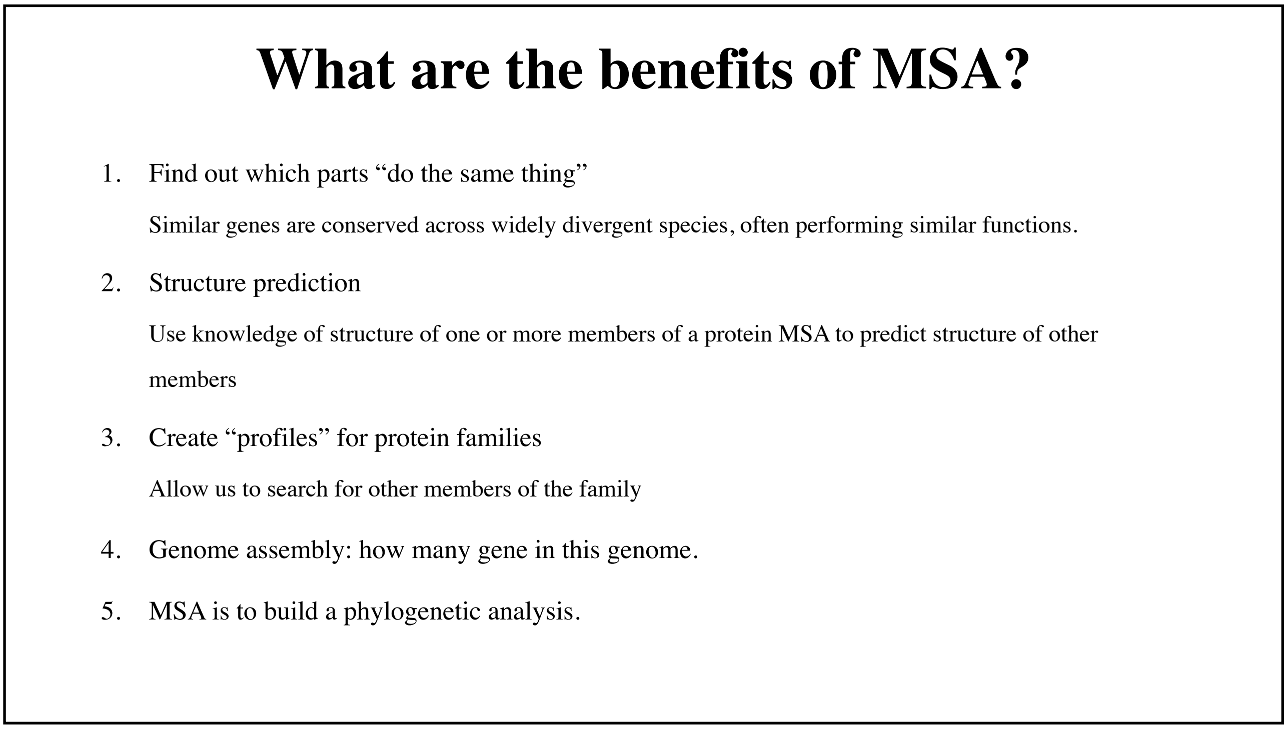
Because Often multiple sequence alignments require manual editing:

1. ***Jalview*** is a powerful MSA-editor for MSA. see

   http://www.jalview.org/index.html

2. ***Muscle: https://www.ebi.ac.uk/Tools/msa/muscle/***
3. PRANK: https://www.ebi.ac.uk/research/goldman/software/prank.
4. MAFFT: https://mafft.cbrc.jp/alignment/software/

7



What are the benefits of MSA?

1. Find out which parts "do the same thing"

   Similar genes are conserved across widely divergent species, often performing similar functions.

2. Structure prediction

   Use knowledge of structure of one or more members of a protein MSA to predict structure of other members

3. Create "profiles" for protein families

   Allow us to search for other members of the family

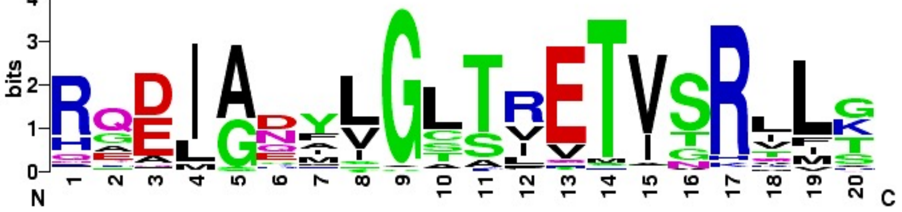
4. Genome assembly: how many gene in this genome.
5. MSA is to build a phylogenetic analysis.

8

How to find the most conservative amino acid in a seq among multiple species?

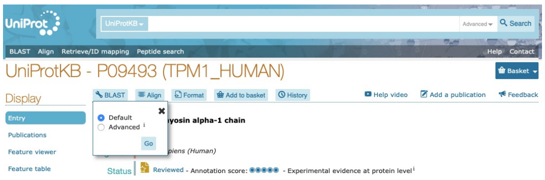
- **Sequence Logos and conservativity can be found using**
- http://weblogo.berkeley.edu/
- Sequence logos are based on **M**ultiple **S**equence **A**lignments
- Very useful to visualise Sequence profiles and motifs.

The DNA-binding helix-turn-helix motif of the CAP family



Tutorial

- Find a TPM1 (tropomyosin) gene for human by typing it in www.uniport.com. Type the gene name
- Go inside the page do click alignment.
- The job will take time.
- Download seq, paste it in https://www.uniprot.org/blast/uniprot/B202003208BC4D7ADE02784B0C2481C7F3DE0963A0E5076S
- Download the whole seq, paste it in http://weblogo.berkeley.edu/logo.cgi



UniProtKB - P09493 (TPM1_HUMAN)

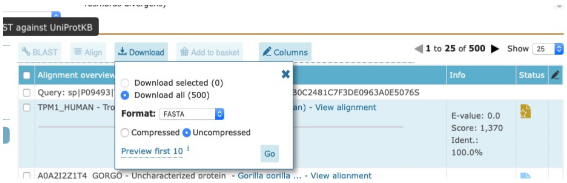
Display

● Default
○ Advanced

yosin alpha-1 chain

piens (Human)

Status Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level

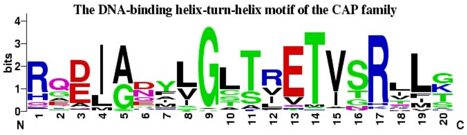


1 to 25 of 500

□ Alignment overview	Info	Status
□ Query: sp|P09493| … B0C2481C7F3DE0963A0E5076S		
□ TPM1_HUMAN - Tro … an) - View alignment	E-value: 0.0 Score: 1,370 Ident.: 100.0%	
□ A0A2I2Z1T4_GORGO - Uncharacterized protein - Gorilla gorilla … - View alignment		

○ Download selected (0)
● Download all (500)
Format: FASTA
○ Compressed ● Uncompressed
Preview first 10

The DNA-binding helix-turn-helix motif of the CAP family



Bioinformatics II
Lecture 3: Phylogenetic trees

Dr Manaf A Guma

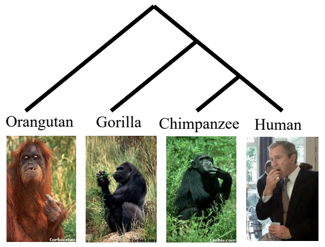
University Of Anbar- college of Applied sciences-Hit

Department of applied chemistry

1

What is a phylogenetics tree!

- What is a phylogenetics tree?
- It is a diagram draws evolutionary relationships between a set of organism phylogenetic from Darwin notebook.
- It reflects the process of descent (from parents to children) with modification process ancestors to descendant.
- MSA multiped sequence alignment led us to reveal the evolutionary relationships between the species.
- Note:
- A little of sequences and species can be solved by hand.
- But we need a computational statistics to figure out the big groups.



From the Tree of the Life Website, University of Arizona

2

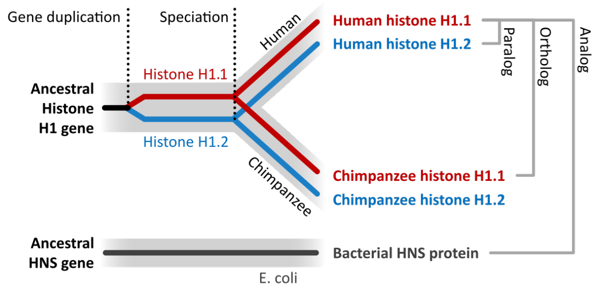
Some identifications and notes related to the topic.

- **Before we study this topic, we need to know the following information:**
- **Phylogeny** is the study of relationships among different groups of organisms and their evolutionary development. **It** attempts to trace the evolutionary history of all life on the planet.
- **Taxonomy** is the science of naming, describing and classifying organisms and includes all plants, animals and microorganisms of the world.
- Phylogenetics: the study of evolutionary relatedness among organisms.
- The term living is called (extant) and dead is called (extinct).
- What are the following terms: Homologous, Paralogous, Analog ad Orthologous genes?

3

What is a homologous gene (or homolog)?

- It is a gene inherited in two species by a common ancestor.
- Note: the homologous genes can be similar in sequence, similar sequences are not necessarily homologous.
- What is an Analog?
- One of two organs or parts in different species of animals or plants which differ in structure or development but are similar in function.



4

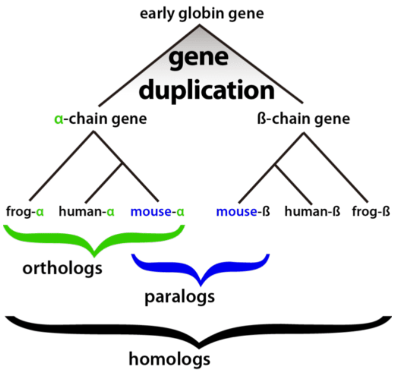
What are orthologous genes (or orthologs)?

- **Orthologous genes** (or **orthologs**) are a particular class of homologous **genes**.
- They are found **in** different species and have diverged following the speciation of the species hosting them.
- Therefore, **orthologous genes in** different species derive from a common ancestral **gene** found **in** the ancestor of those species.

5

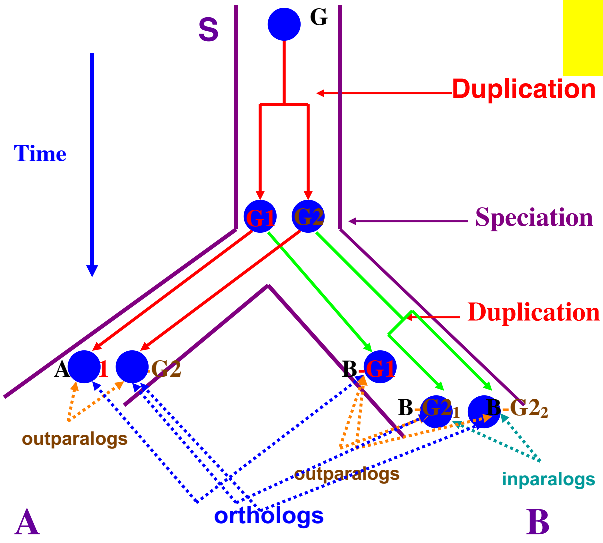
What are Paralogous genes (or paralogs)?

- Paralogous genes (or **paralogs**) are a particular class of homologous genes.
- They are the result of gene duplication.
- The gene copies resulting from the duplication are called paralogous of each other. ...
- Paralogous genes can be remined in the genome after their duplication, but some copies can also be lost.



6

Flow chart to describe Orthologs - Paralogs

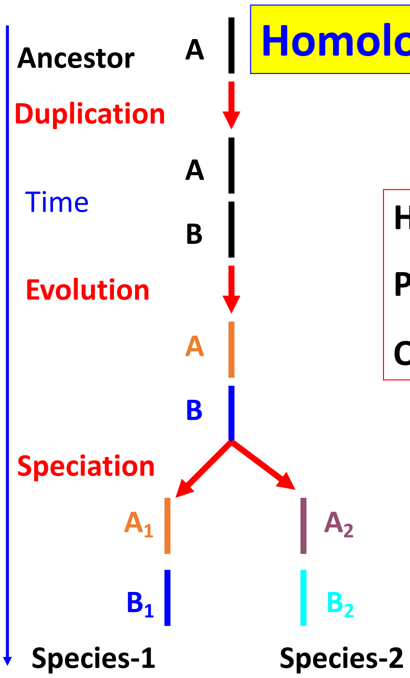


Koonin EV. 2005. Annu Rev Genet. 39:309-38.

Tekaia F. Inferring Orthologs: Open Question and Perspectives (Review). *Genomics Insights* 2016:9 17-28.

7

Homologs - Paralogs - Orthologs

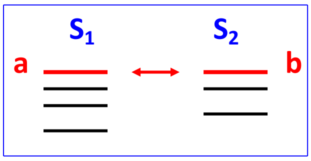


Homologs: A_1, B_1, A_2, B_2

Paralogs : A_1 vs B_1 and A_2 vs B_2

Orthologs: A_1 vs A_2 and B_1 vs B_2

Sequence analysis



8

Phylogeny analyses

1. Starting point to draw the tree needs a set of **homologous**, aligned DNA or protein sequences.
2. Result of the process: a tree describing evolutionary relationships between the considered sequences i.e.

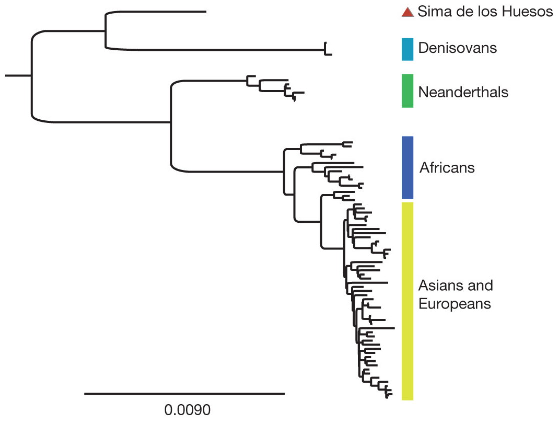
How many ways to describe the process of reconstructing a phylogeny?

- There have been two different ways and two different philosophies to the process of reconstructing a phylogeny:
- One approach is the phenetic approach. In this approach, a tree is constructed by considering the phenotypic similarities of the species without trying to understand the evolutionary pathways of the species.
- The trees constructed via this method are called phenograms.
- The second approach is called the cladistic approach. via these methods, a tree is reconstructed by considering the various possible pathways of evolution and choosing from amongst these the best possible tree.
- Trees reconstructed via these methods are called cladograms.

How does the phylogenetic tree describe the evolutionary relationship?

- By using large DNA sets to build tree underlying principle and logic sequences.
- The sequences are separated by shorter evolutionary distance which are expected to be similar to one another than sequences separated over longer distances.
- How about trees with long evolutionary distances ?

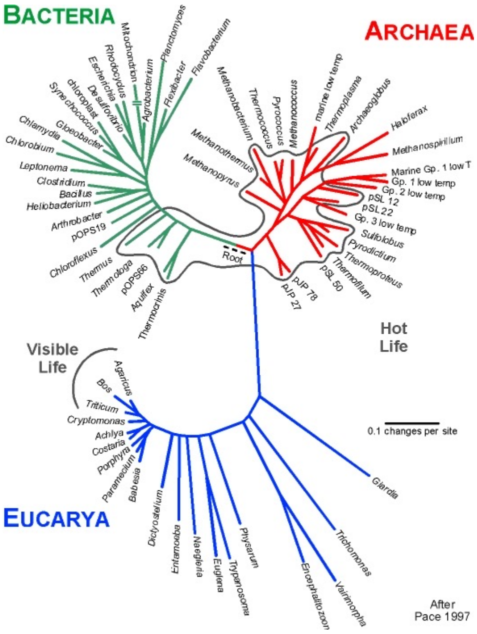
Bayesian phylogenetic tree of hominin mitochondrial relationships based on the Sima de los Huesos mtDNA sequence determined using the inclusive filtering criteria.



M Meyer *et al. Nature* **000**, 1-4 (2013) doi:10.1038/nature12788

Examples of phylogenetic trees

13



After
Pace 1997

Pace (2001) described a tree of life based on small subunit rRNA sequences.

Pace, N. R. (1997) *Science* **276**, 734-740

This tree shows the main three branches described by Woese and colleagues.

This tree is referred to as the tree of life or the universal tree.

14

Chlamydiae

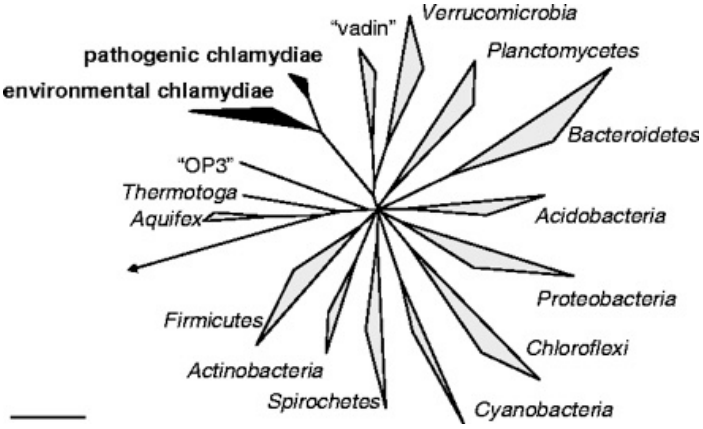


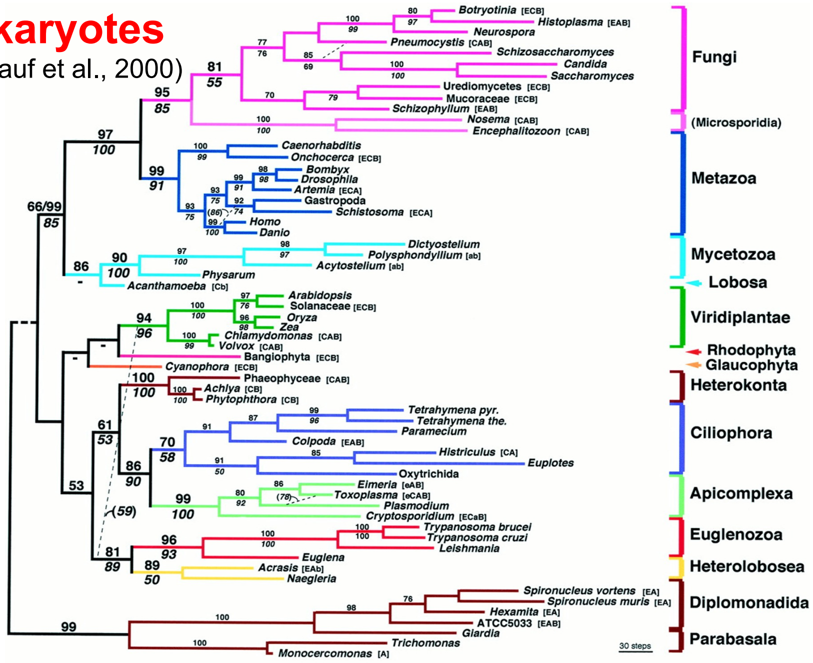
Fig. 1. Phylogeny of chlamydiae. 16*S* rRNA-based neighbor-joining tree showing the affiliation of environmental and pathogenic chlamydiae with major bacterial phyla. Arrow, to outgroup. Scale bar, 10% estimated evolutionary distance.

Science 304:728-30.2004.

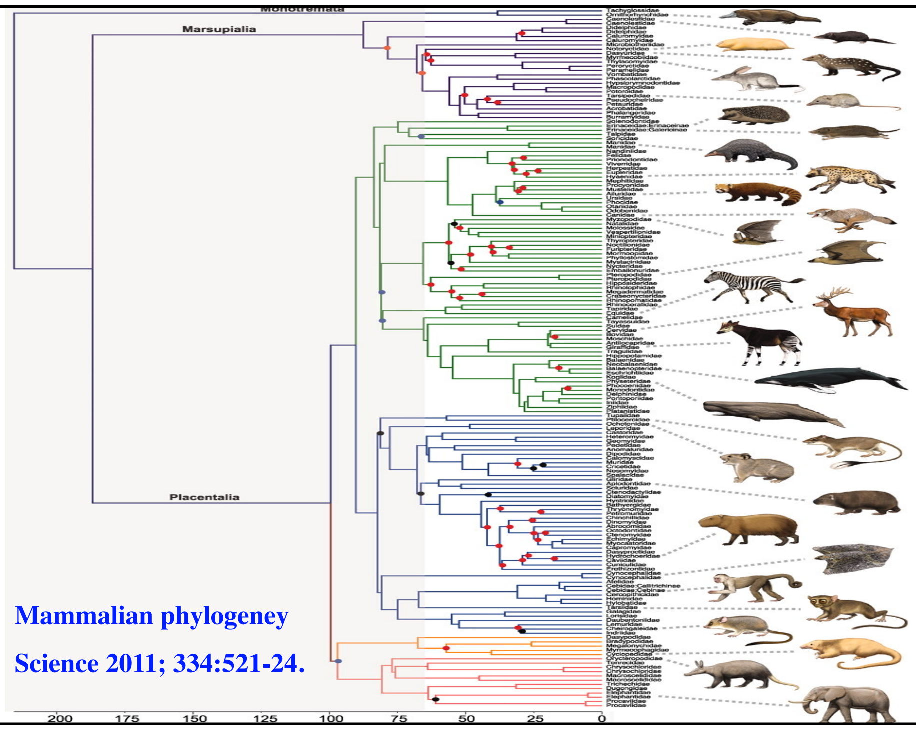
15

Eukaryotes

(Baldauf et al., 2000)



16



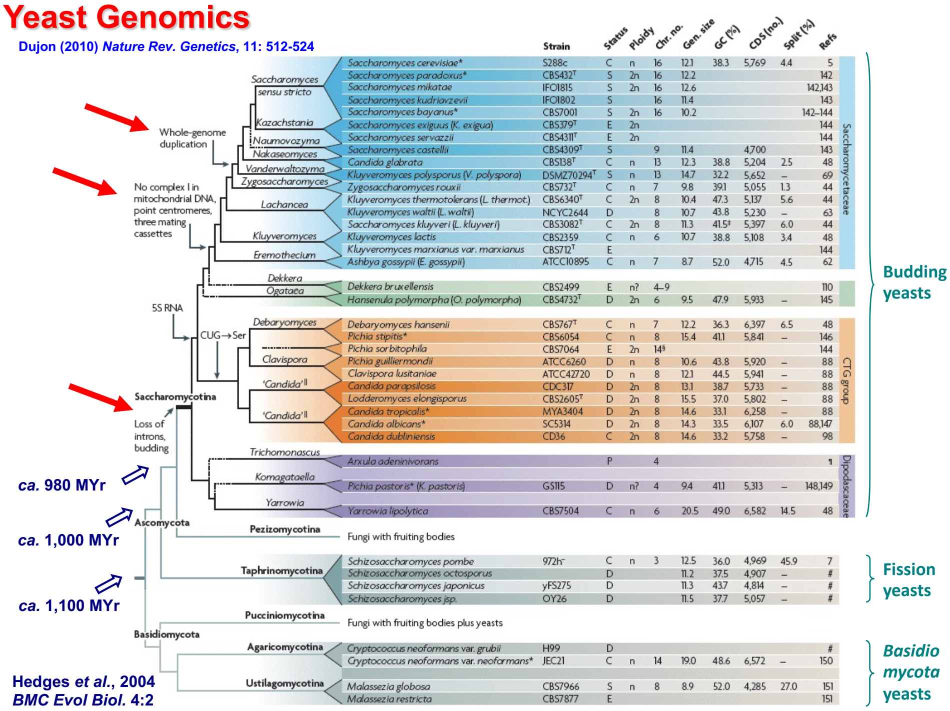
Mammalian phylogeny

Science 2011; 334:521-24.

17

Yeast Genomics

Dujon (2010) *Nature Rev. Genetics*, 11: 512-524



	Strain	Status	Ploidy	Chr. no.	Gen. size	GC (%)	CDS (no.)	Split (%)	Refs
*Saccharomyces cerevisiae**	S288c	C	n	16	12.1	38.3	5,769	4.4	5
*Saccharomyces paradoxus**	CBS432^T	S	2n	16	12.2				142
Saccharomyces mikatae	IFO1815	S	2n	16	12.6				142,143
Saccharomyces kudriavzevii	IFO1802	S		16	11.4				143
*Saccharomyces bayanus**	CBS7001	S	2n	16	10.2				142–144
Saccharomyces exiguus (K. exigua)	CBS379^T	E	2n						144
Saccharomyces servazzii	CBS4311^T	E	2n						144
Saccharomyces castellii	CBS4309^T	S		9	11.4		4,700		143
Candida glabrata	CBS138^T	C	n	13	12.3	38.8	5,204	2.5	48
Kluyveromyces polysporus (V. polyspora)	DSMZ70294^T	S	n	13	14.7	32.2	5,652	–	69
Zygosaccharomyces rouxii	CBS732^T	C	n	7	9.8	39.1	5,055	1.3	44
Kluyveromyces thermotolerans (L. thermot.)	CBS6340^T	C	2n	8	10.4	47.3	5,137	5.6	44
Kluyveromyces waltii (L. waltii)	NCYC2644	D		8	10.7	43.8	5,230	–	63
Saccharomyces kluyveri (L. kluyveri)	CBS3082^T	C	2n	8	11.3	41.5‡	5,397	6.0	44
Kluyveromyces lactis	CBS2359	C	n	6	10.7	38.8	5,108	3.4	48
Kluyveromyces marxianus var. marxianus	CBS712^T	E							144
Ashbya gossypii (E. gossypii)	ATCC10895	C	n	7	8.7	52.0	4,715	4.5	62
Dekkera bruxellensis	CBS2499	E	n?	4–9					110
Hansenula polymorpha (O. polymorpha)	CBS4732^T	D	2n	6	9.5	47.9	5,933	–	145
Debaryomyces hansenii	CBS767^T	C	n	7	12.2	36.3	6,397	6.5	48
*Pichia stipitis**	CBS6054	C	n	8	15.4	41.1	5,841	–	146
Pichia sorbitophila	CBS7064	E	2n	14§					144
Pichia guilliermondii	ATCC6260	D	n	8	10.6	43.8	5,920	–	88
Clavispora lusitaniae	ATCC42720	D	n	8	12.1	44.5	5,941	–	88
Candida parapsilosis	CDC317	D	2n	8	13.1	38.7	5,733	–	88
Lodderomyces elongisporus	CBS2605^T	D	2n	8	15.5	37.0	5,802	–	88
*Candida tropicalis**	MYA3404	D	2n	8	14.6	33.1	6,258	–	88
*Candida albicans**	SC5314	D	2n	8	14.3	33.5	6,107	6.0	88,147
Candida dubliniensis	CD36	C	2n	8	14.6	33.2	5,758	–	98
Arxula adeninivorans		P		4					¶
Pichia pastoris (K. pastoris)*	GS115	D	n?	4	9.4	41.1	5,313	–	148,149
Yarrowia lipolytica	CBS7504	C	n	6	20.5	49.0	6,582	14.5	48
Fungi with fruiting bodies									
Schizosaccharomyces pombe	972h⁻	C	n	3	12.5	36.0	4,969	45.9	7
Schizosaccharomyces octosporus		D			11.2	37.5	4,907	–	#
Schizosaccharomyces japonicus	yFS275	D			11.3	43.7	4,814	–	#
Schizosaccharomyces jsp.	OY26	D			11.5	37.7	5,057	–	#
Fungi with fruiting bodies plus yeasts									
Cryptococcus neoformans var. grubii	H99	D							#
*Cryptococcus neoformans var. neoformans**	JEC21	C	n	14	19.0	48.6	6,572	–	150
Malassezia globosa	CBS7966	S	n	8	8.9	52.0	4,285	27.0	151
Malassezia restricta	CBS7877	E							151

Budding yeasts

Fission yeasts

Basidio mycota yeasts

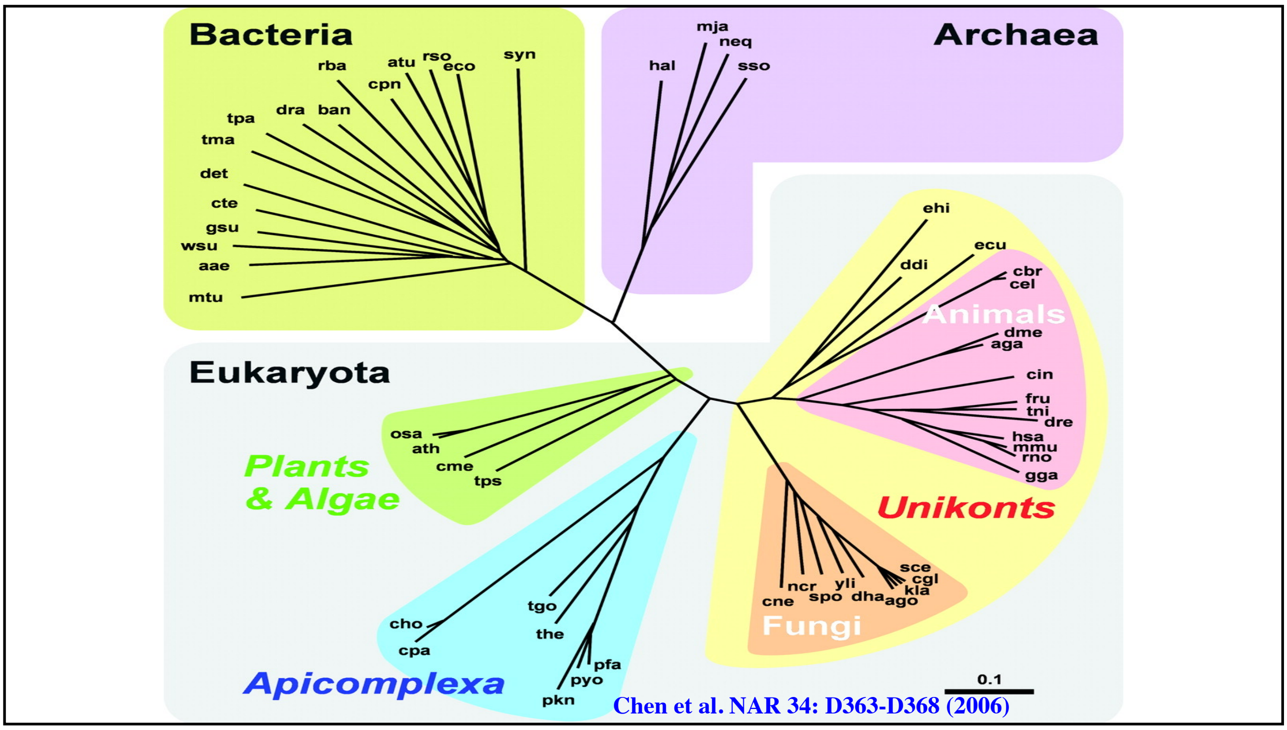
ca. 980 MYr

ca. 1,000 MYr

ca. 1,100 MYr

Hedges *et al.*, 2004
BMC Evol Biol. 4:2

18

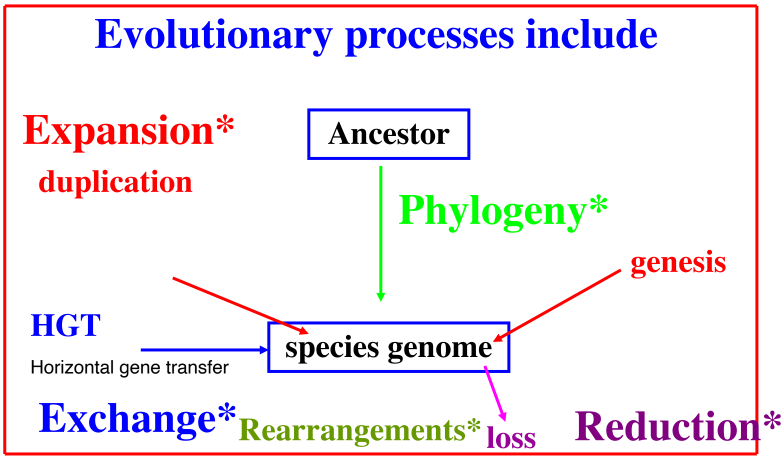


Chen et al. NAR 34: D363-D368 (2006)

19

How the analyses of predicted proteomes revealed significant evolutionary processes?

Expansion, Exchange and Reduction.



Evolutionary processes include

Expansion*

duplication

Ancestor

Phylogeny*

genesis

HGT

Horizontal gene transfer

species genome

Exchange*

Rearrangements*

loss

Reduction*

20

Bioinformatics II:

Lecture 4: Building Phylogenies using different methods

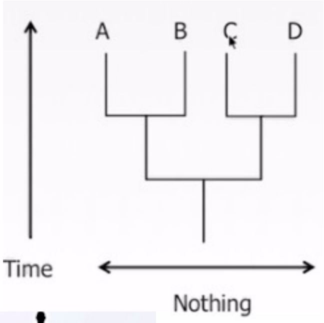
Dr Manaf A Guma

University of Anbar- college of applied sciences-Heet.

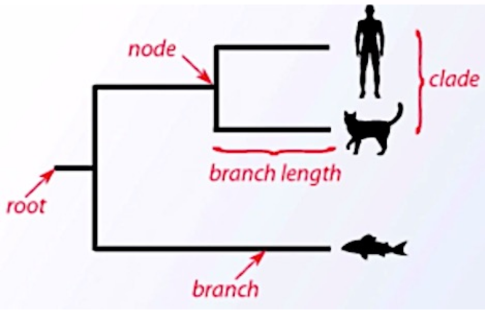
Department of chemistry

1

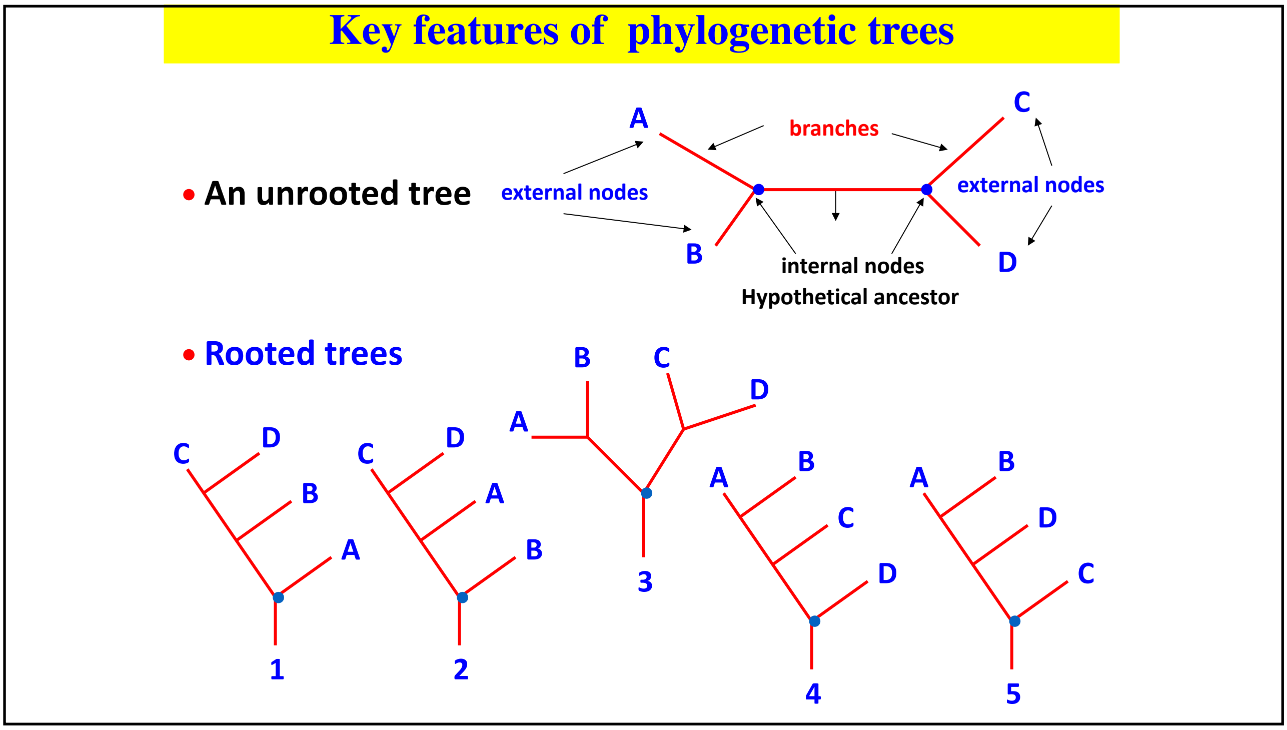
What does tree consist of:



1. Root: where tree begins.
2. branch represent different species.
3. Node to separate the branches.
4. Tips of the tree (called leaves as in a plant tree).
5. There are 2D of the tree:
6. The length of the branch to the tips tells the time over.
7. The other dimension does not tell any things about the relationships.



2

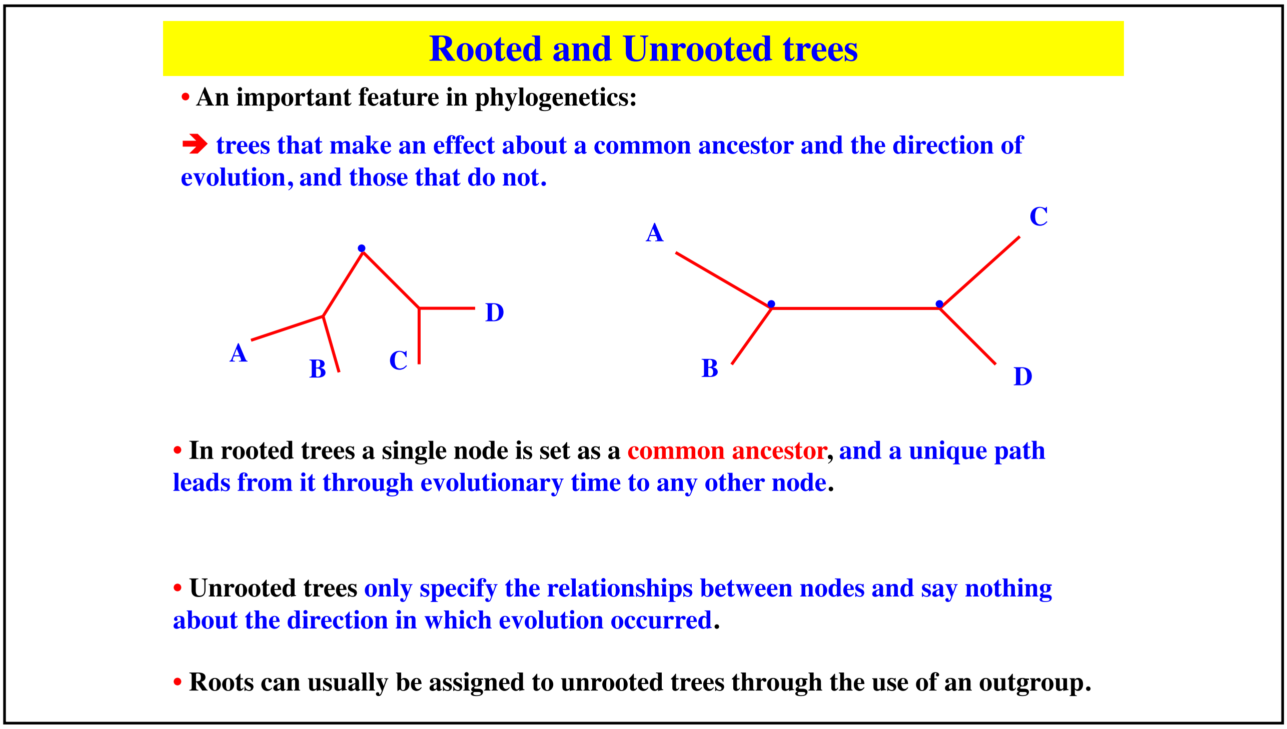


Key features of phylogenetic trees

- An unrooted tree

- Rooted trees

3



Rooted and Unrooted trees

- An important feature in phylogenetics:

➔ trees that make an effect about a common ancestor and the direction of evolution, and those that do not.

- In rooted trees a single node is set as a common ancestor, and a unique path leads from it through evolutionary time to any other node.

- Unrooted trees only specify the relationships between nodes and say nothing about the direction in which evolution occurred.

- Roots can usually be assigned to unrooted trees through the use of an outgroup.

4

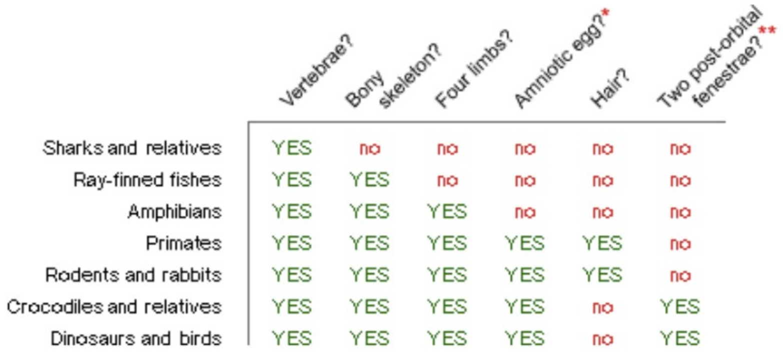
Reconstructing trees: A simple example: step 1

- **Choose the taxa.** You decide to study the major clades of vertebrates shown in the leftmost column of the table below.
- (Note that many vertebrate lineages are excluded from this example for the sake of simplicity.)

Taxa: a classification group.

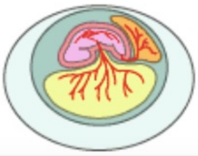
Or Taxon.

Moisture	ندي او رطب
Amphibians	البرمائيات
Retaining	الاحتفاظ
Amniotic	الذي يحيط بالجنين
Fenestrae	النوافذ
Rodents	القوارض
Primates	الرئيسيات such as human, monkeys ets
ray finned fishes	أسماك الراي ذات الزعانف



	Vertebrae?	Bony skeleton?	Four limbs?	Amniotic egg?*	Hair?	Two post-orbital fenestrae?**
Sharks and relatives	YES	no	no	no	no	no
Ray-finned fishes	YES	YES	no	no	no	no
Amphibians	YES	YES	YES	no	no	no
Primates	YES	YES	YES	YES	YES	no
Rodents and rabbits	YES	YES	YES	YES	YES	no
Crocodiles and relatives	YES	YES	YES	YES	no	YES
Dinosaurs and birds	YES	YES	YES	YES	no	YES

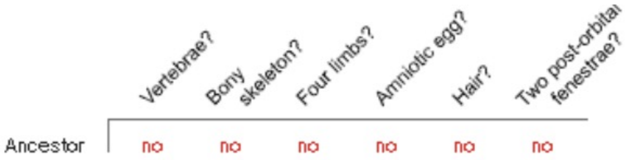
*amniotic egg: an egg in which the embryo is surrounded by the moisture-retaining amnion membrane



**post-orbital fenestrae: holes in the skull behind the eye

Step 2 & step 3

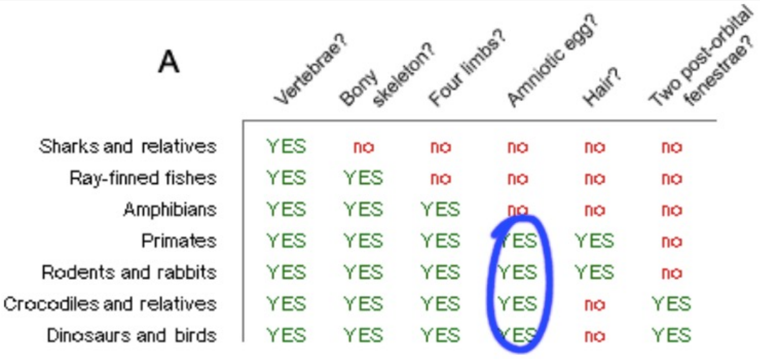
- **Determine the characters.** After studying the vertebrates, you select a set of traits, which seem to be homologies, and build the following data table to record your observations.
- (Note that many relevant vertebrate characters are excluded from this example for the sake of simplicity.
- **Determine the polarity of characters.** From studying fossils and outgroups closely related to the vertebrate clade, you hypothesize that the ancestor of vertebrates had none of these features.
- Clade: a group of organisms believed to have evolved from a common ancestor, according to the principles of cladistics.



	Vertebrae?	Bony skeleton?	Four limbs?	Amniotic egg?	Hair?	Two post-orbital fenestrae?
Ancestor	no	no	no	no	no	no

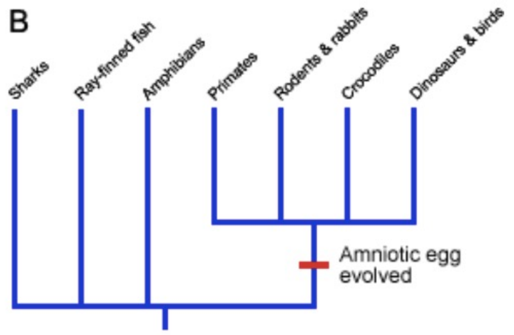
Step 4

- **Group taxa by traits .** We might start out by examining the egg character.
- We focus in on the group of lineages that share the **traits** form of this character, an amniotic egg (A,) and hypothesize that they form a clade (B):



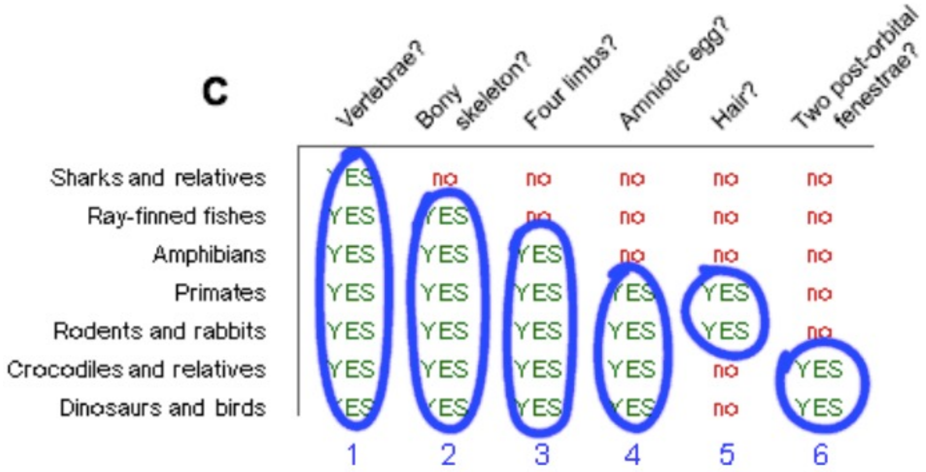
A

	Vertebrae?	Bony skeleton?	Four limbs?	Amniotic egg?	Hair?	Two post-orbital fenestrae?
Sharks and relatives	YES	no	no	no	no	no
Ray-finned fishes	YES	YES	no	no	no	no
Amphibians	YES	YES	YES	no	no	no
Primates	YES	YES	YES	YES	YES	no
Rodents and rabbits	YES	YES	YES	YES	YES	no
Crocodiles and relatives	YES	YES	YES	YES	no	YES
Dinosaurs and birds	YES	YES	YES	YES	no	YES



Then,

- We go through the whole table like this, grouping clades according to traits

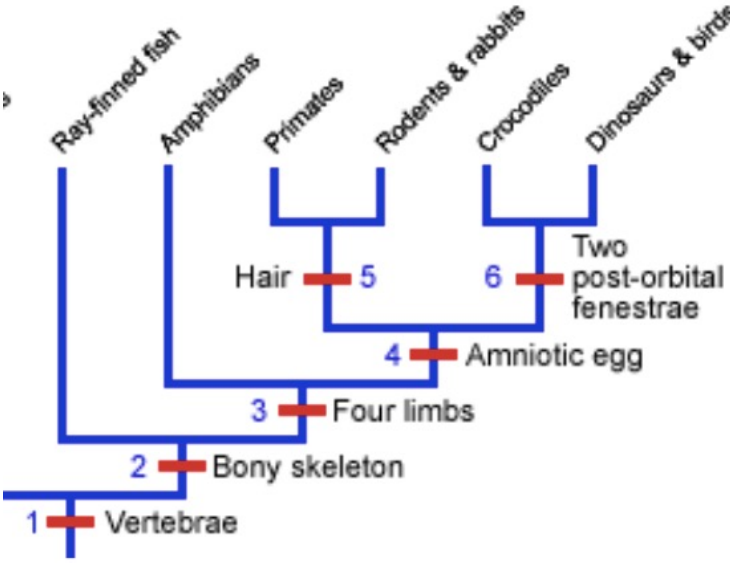


C

	Vertebrae?	Bony skeleton?	Four limbs?	Amniotic egg?	Hair?	Two post-orbital fenestrae?
Sharks and relatives	YES	no	no	no	no	no
Ray-finned fishes	YES	YES	no	no	no	no
Amphibians	YES	YES	YES	no	no	no
Primates	YES	YES	YES	YES	YES	no
Rodents and rabbits	YES	YES	YES	YES	YES	no
Crocodiles and relatives	YES	YES	YES	YES	no	YES
Dinosaurs and birds	YES	YES	YES	YES	no	YES
	1	2	3	4	5	6

Step 5

- **Work out conflicts that arise.** There are no conflicts here. Every group is a subset of another group (see C).
- **Build your tree.** Based on the groupings above, you produce this tree:

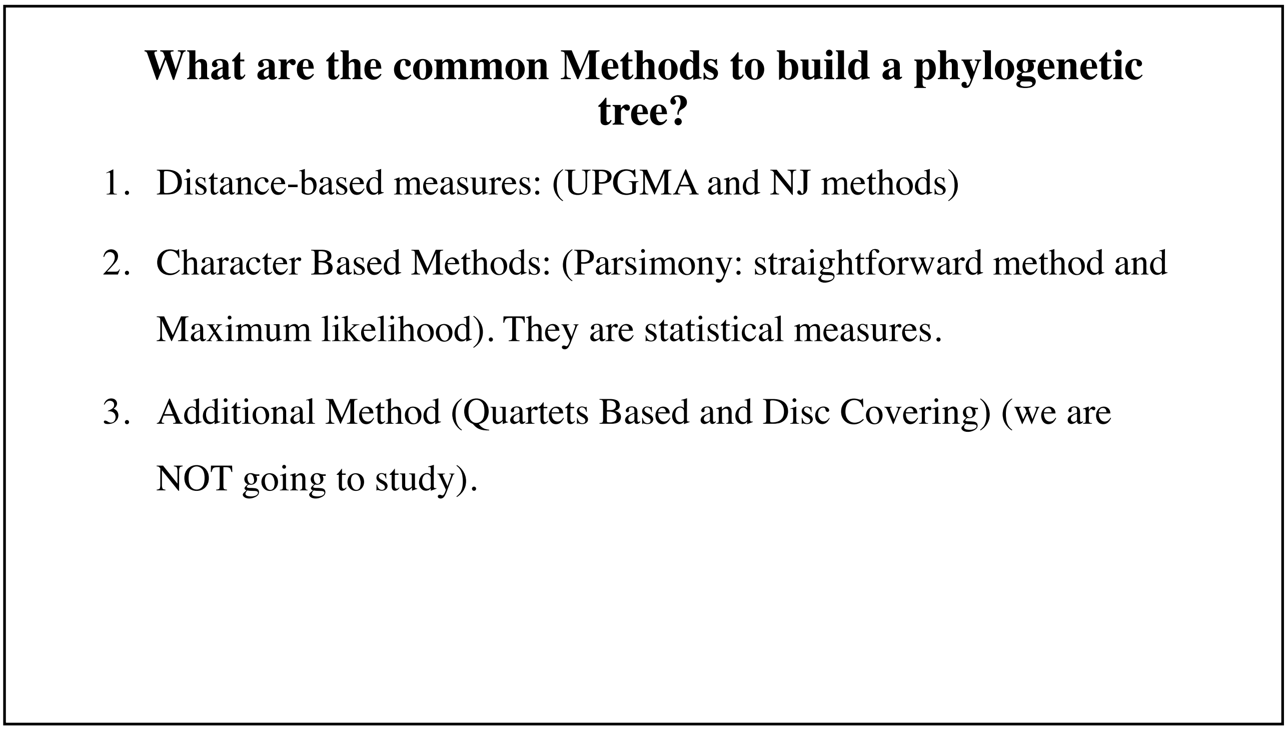


9

Step 7

- **You have made a phylogeny.**
- Of course, this was just an example of the tree-building process.
- Phylogenetic trees are generally based on many more characters and often involve more lineages.

10



What are the common Methods to build a phylogenetic tree?

1. Distance-based measures: (UPGMA and NJ methods)
2. Character Based Methods: (Parsimony: straightforward method and Maximum likelihood). They are statistical measures.
3. Additional Method (Quartets Based and Disc Covering) (we are NOT going to study).

Bioinformatics II:

Lecture 5: Building a phylogenetic tree using Parsimony methods

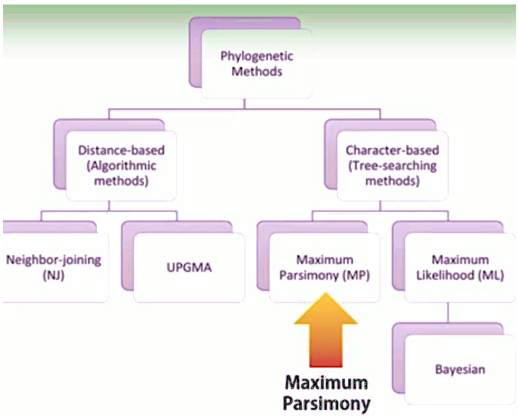
Dr Manaf A Guma

University of Anbar- college of applied science-Heet.

Department of chemistry

What is Parsimony methods

- **It is a method taxa (for classification) in ways that minimize the number of evolutionary changes.**
- **What is the Goal of the method?**
- Find the tree that allows evolution of the sequences with the fewest changes.
- This is called a ***most parsimonious*** (MP) tree
- **NOTE: Taxonomists : scientists who do classification through evaluations.**



Phylogenetic Methods

Distance-based (Algorithmic methods)

Character-based (Tree-searching methods)

Neighbor-joining (NJ)

UPGMA

Maximum Parsimony (MP)

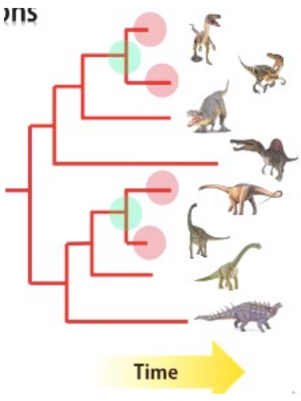
Maximum Likelihood (ML)

Bayesian

Maximum Parsimony

What is the idea of the Parsimony method?

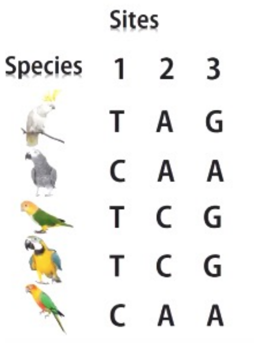
- It is based on, all other species being equal.
- For e.g: 4 nucleotide changes, is more likely to be true than a complex one. (sub-grouping).
- The method assumes that change in characteristic occurs in lineage over time.
- But, the rate of mutation is NOT constant over time.
- So, leaves (descendent) can have different distances from the root.



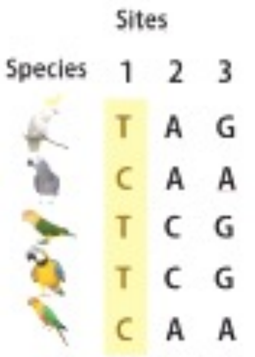
3

Example? Construct a tree that represent minimum evolutionary changes

- For site 1, the tree that represents minimum changes groups all the (parrots) -a type of birds-showing ''T'' in one group and all those showing ''C'' in another group.



	Sites		
Species	1	2	3
	T	A	G
	C	A	A
	T	C	G
	T	C	G
	C	A	A

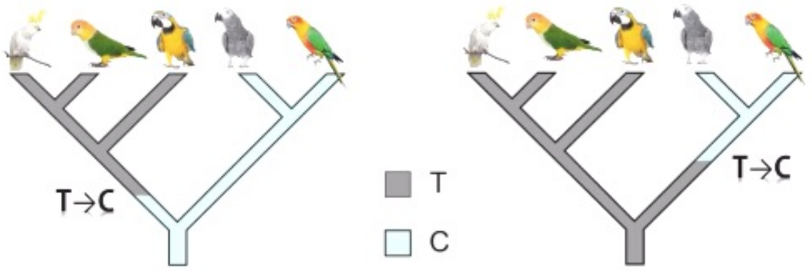


	Sites		
Species	1	2	3
	T	A	G
	C	A	A
	T	C	G
	T	C	G
	C	A	A

4

The question: How thee taxa divergence started from an ancestor?

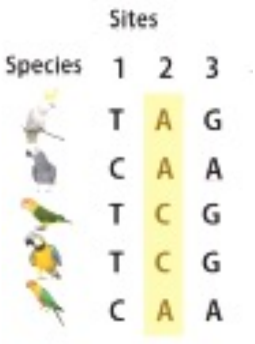
- The two trees are topologically identical, but they assume different. ancestral states



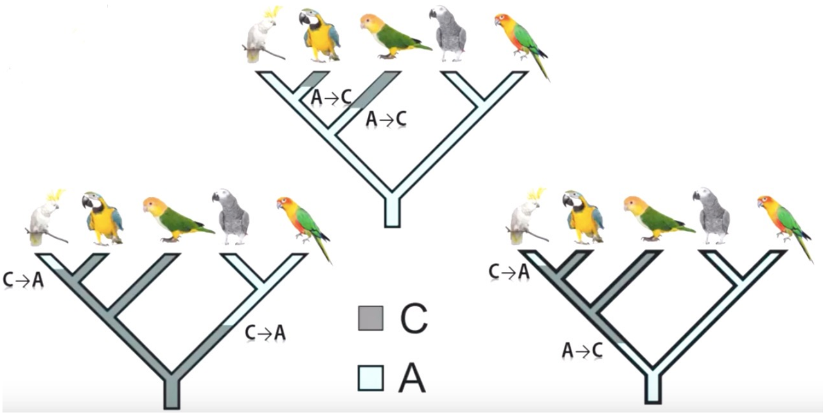
5

2

- For site 2, there are three possible parsimonious trees.



	Sites		
Species	1	2	3
	T	A	G
	C	A	A
	T	C	G
	T	C	G
	C	A	A



6

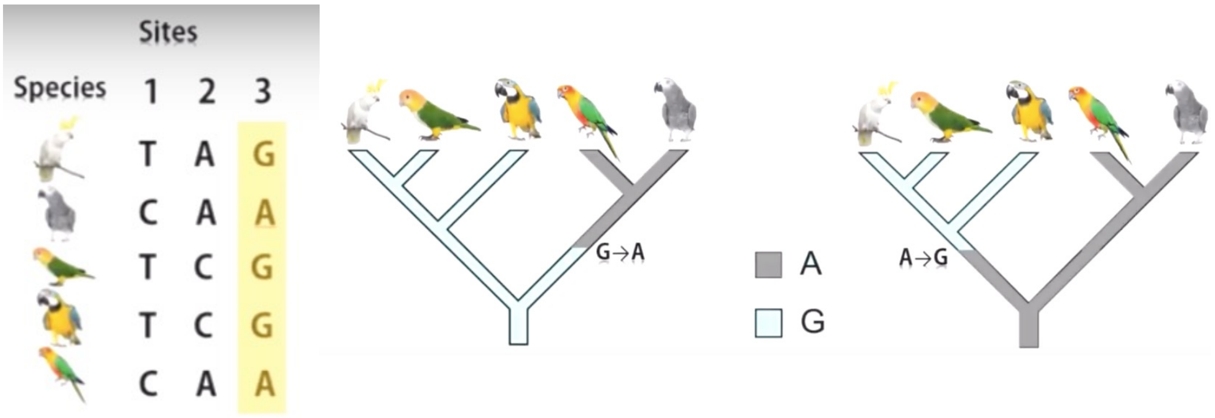
2

- All these trees agree on:
- How to cluster parrots in different groups.
- The minimum number of substitution.
- But they disagree on:
- How this evolutionary divergence has occurred in the course of evolution.
- Again, trees are topologically identical, but they assume different ancestral state

7

3

- For site 3, there are 2 possible parsimonious tree.

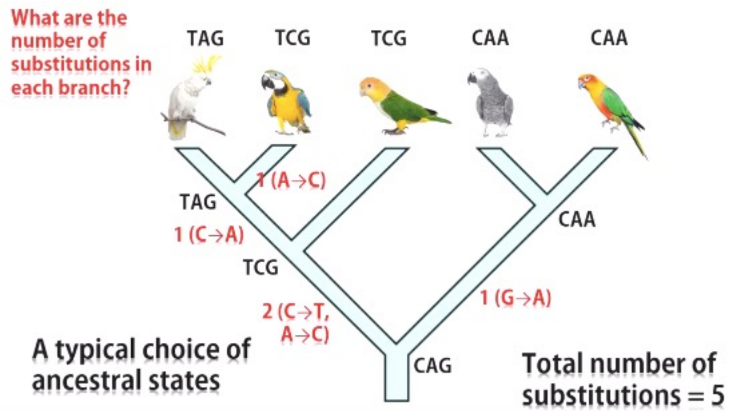


	Sites		
Species	1	2	3
	T	A	G
	C	A	A
	T	C	G
	T	C	G
	C	A	A

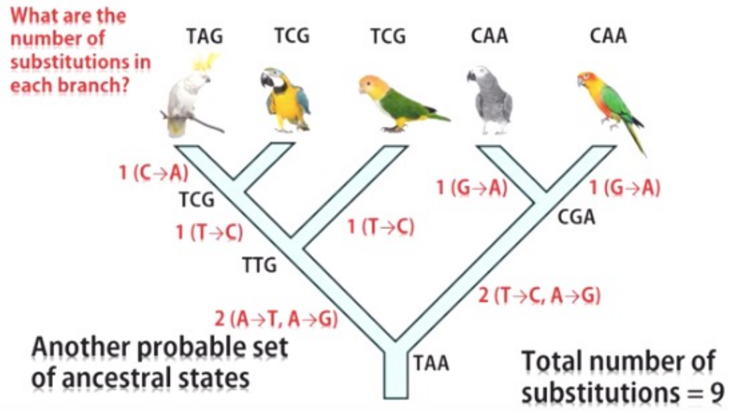
8

Calculate the number of substitutions steps on each branch?

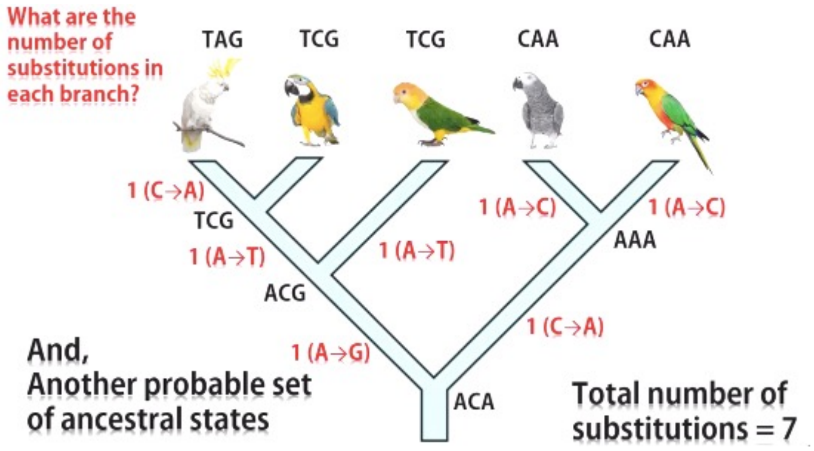
- So, this depends on the ancestral states.


What are the number of substitutions in each branch?
TAG
TCG
TCG
CAA
CAA
1 (A→C)
TAG
1 (C→A)
TCG
CAA
2 (C→T, A→C)
1 (G→A)
CAG
A typical choice of ancestral states
Total number of substitutions = 5

9


What are the number of substitutions in each branch?
TAG
TCG
TCG
CAA
CAA
1 (C→A)
TCG
1 (T→C)
TTG
1 (T→C)
1 (G→A)
1 (G→A)
CGA
2 (T→C, A→G)
2 (A→T, A→G)
TAA
Another probable set of ancestral states
Total number of substitutions = 9

10



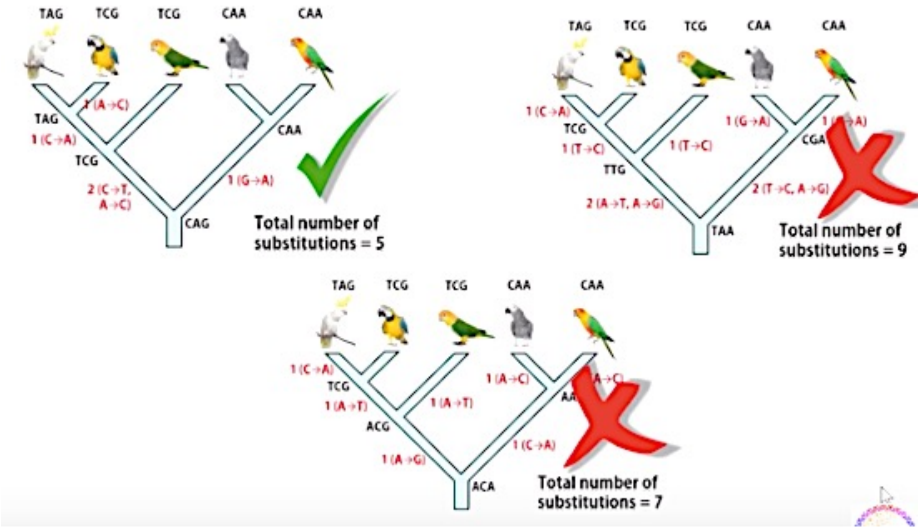
What are the number of substitutions in each branch?

And,
Another probable set of ancestral states

Total number of substitutions = 7

11

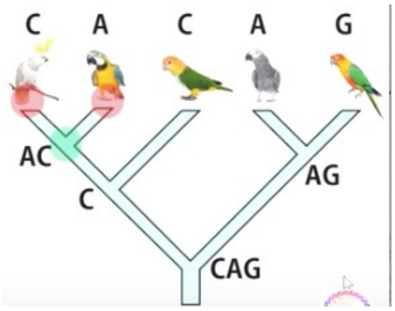
The maximum parsimony tree is the one with minimum total number of substitution.



12

Fitch’s algorithm: a computer work

- The maximum parsimony tree depends on the sequences of ancestral nodes.
- Fitch’s algorithm is used used to build ancestral nodes.
- To count the number of steps of a tree requires at given site:
- Construct a set of nucleotides that are observed there.
- Go down the tree.
- For each ancestral node (say A).
- Consider its two immediate descendent sets (say D1 and D2).
- Nucleotide set in ancestor A
- A=D1 U D2, if D1 and D2 are dis-join (U means Union)
- Else, A=D1 ∩ D2 means **Intersection**



https://www.youtube.com/watch?v=gLM5N__kPiU

13

Tutorial

- Make a Phylogenetic tree using NCBI?
- https://www.youtube.com/watch?v=ilE-SkJCFOc

14

Bioinformatics II:

Lecture 6: RNA structure and function (structure prediction)

Dr Manaf A Guma

University of Anbar- college of applied science-Heet.

Department of chemistry

1

What is the role of RNA in the life?

- The main cause of the divergence of species and the evolution is the RNA processes e.g splicing etc.
- So, it important to study such process on RNA and the structure and the function of the RNA
- **RNA structure**
  - Levels of organization
  - Energetics
- **RNA types & functions**
  1. Genomic information storage/transfer
  2. Structural
  3. Catalytic
  4. Regulatory

2

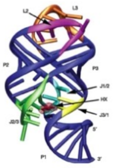
RNA structure

- **We do know in biology that:**
- **One sequence leads to ------A unique structure (usually).**
- **This is true for proteins and RNA molecules? how?**
- Some of RNA molecules are relatively unstructured (part of mRNA).
- But other RNA have structures such as tRNA ribosomes, enzymes ribo-switches.
- RNA molecules are often high complex and involved in all aspects of gene expression. See examples:

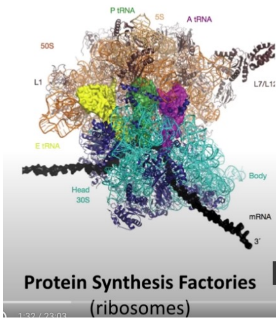
3

Some examples

Natural Biosensors
(G riboswitch)



- Some examples natural biosensors:
- 1- Its function is to sense the Metabolite in the environment.
- 2- catalytic RNA can cut and stitch RNA and DNA molecules.
- 3-Ribosome; the protein synthesis factor in your body
- See other examples:



Protein Synthesis Factories
(ribosomes)

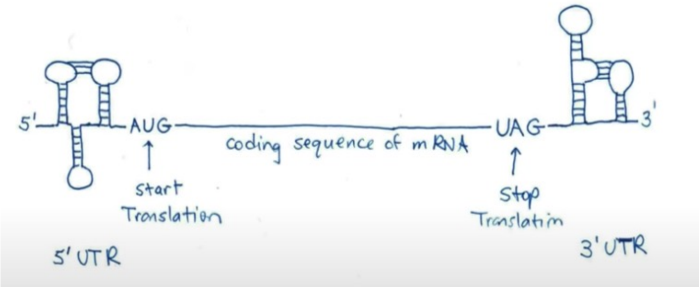
4

RNA types & functions

Types of RNAs	Primary Function(s)
mRNA - messenger	translation (protein synthesis) regulatory
rRNA - ribosomal	translation (protein synthesis) <catalytic>
t-RNA - transfer	translation (protein synthesis)
hnRNA - heterogeneous nuclear	precursors & intermediates of mature mRNAs & other RNAs
scRNA - small cytoplasmic	signal recognition particle (SRP) tRNA processing <catalytic>
snRNA - small nuclear snoRNA - small nucleolar	mRNA processing, poly A addition <catalytic> rRNA processing/maturation/methylation
regulatory RNAs (siRNA, miRNA, etc.)	regulation of transcription and translation, other??

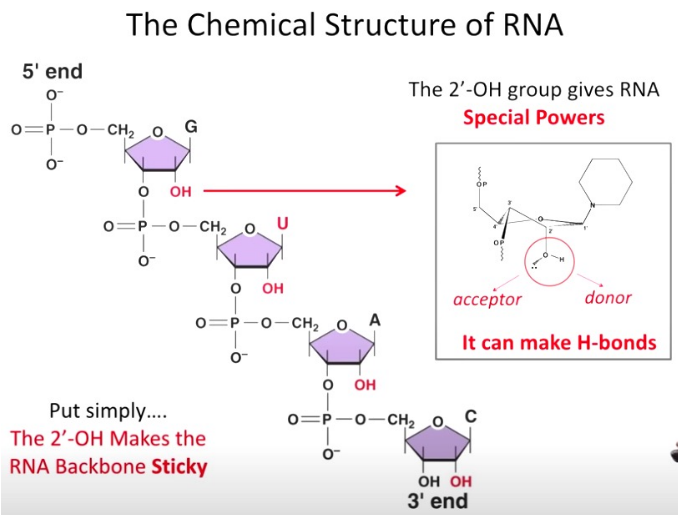
Structured and unstructured RNA

- Some mRNA contains structured regions:
- this is an example of a mRNA and a (middle part of mRNA molecule) contains cording regions: start and stop codons.



The chemical structure of RNA

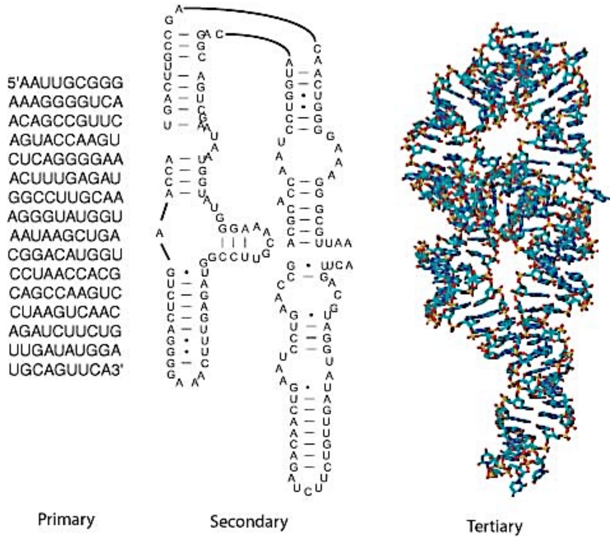
- It contains 2-OH group would you give Aaron a special power.
- 2-OH makes RNA backbone sticky.
- Also, sugars are not the flat because they are five membered rings.



Put simply....
The 2'-OH Makes the RNA Backbone Sticky

What are the types of bonds in the RNA structure? Covalent & non-covalent bonds in RNA

- Mainly there are three structures of RNA molecules:
- **1-Primary:**
- Covalent bonds
- **2-Secondar and 3-Tertiary both have:**
- Non-covalent bonds
  - H-bonds
  - (base-pairing)
  - Base stacking

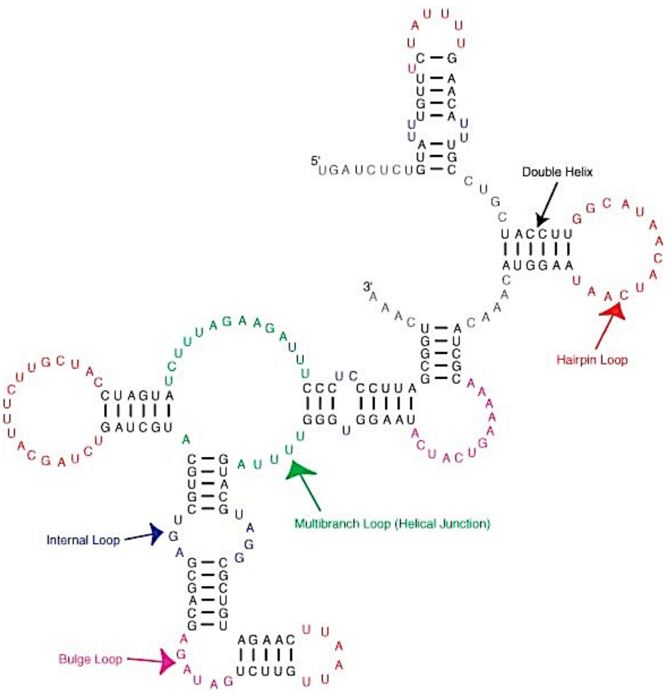

5'AAUUGCGGG
AAAGGGGUCA
ACAGCCGUUC
AGUACCAAGU
CUCAGGGGAA
ACUUUGAGAU
GGCCUUGCAA
AGGGUAUGGU
AAUAAGCUGA
CGGACAUGGU
CCUAACCACG
CAGCCAAGUC
CUAAGUCAAC
AGAUCUUCUG
UUGAUAUGGA
UGCAGUUCA3'

What are the Common structural motifs in RNA?

Helices

Loops

- Hairpin
- Internal
- Bulge
- Multibranch

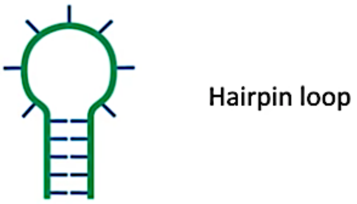


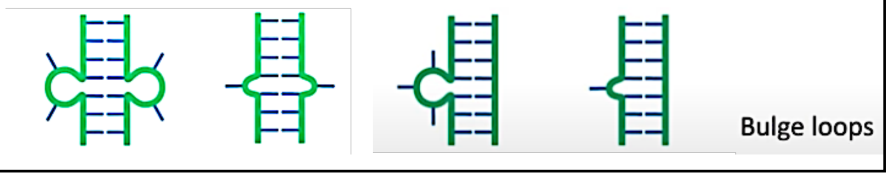
Some kinds of RNA secondary structure

- In RNA the bases do not stick like spines of cactus

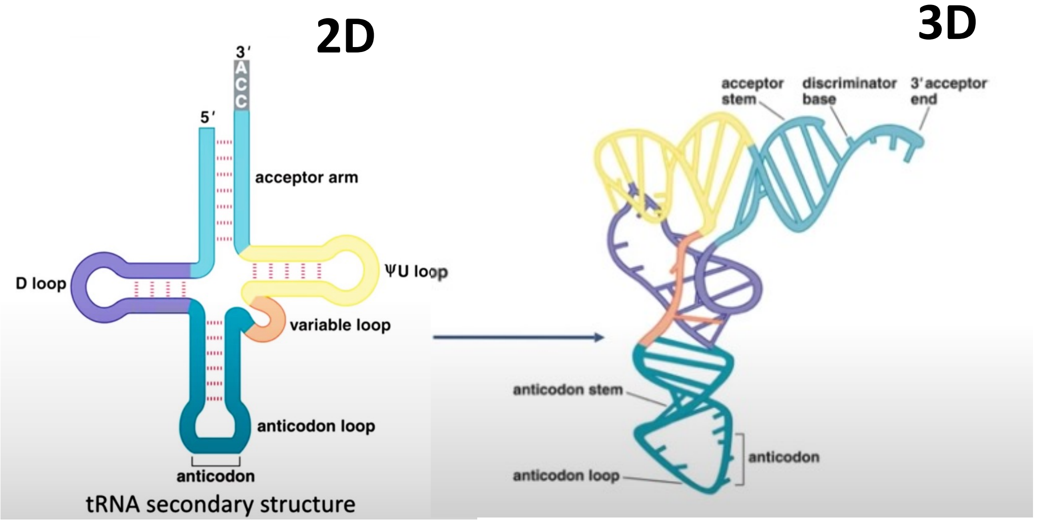
  But some are in a structure or dynamic they can be unpaired in the structure.

- Terminal leg of RNA hairpin have a specifi structure



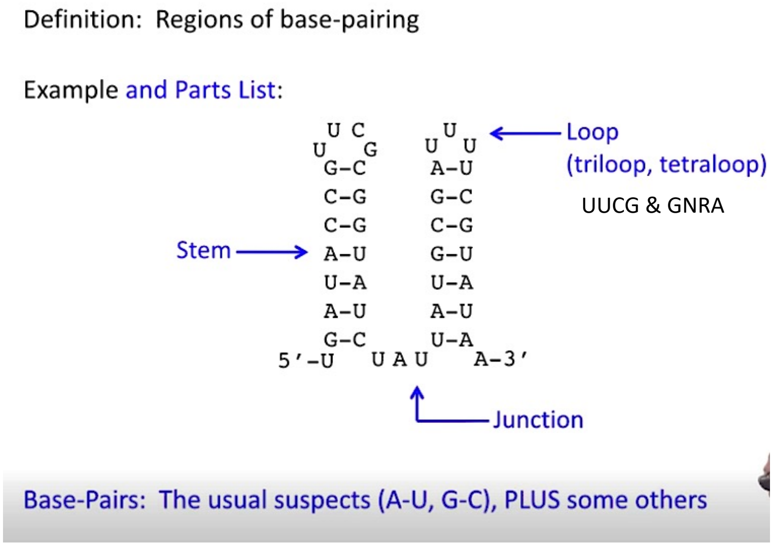


tRNA secondary structure



RNA secondary structure

- Regions of all in a base pairing as it shown in the figure



Definition: Regions of base-pairing

Example and Parts List:

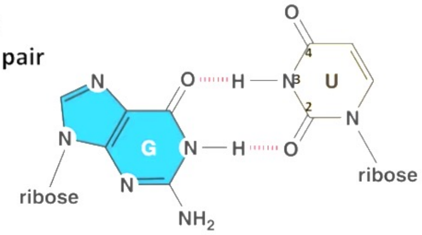
```
 U C      U
U   G    U U   <-- Loop
G-C      A-U       (triloop, tetraloop)
C-G      G-C        UUCG & GNRA
C-G      C-G
Stem --> A-U      G-U
U-A      U-A
A-U      A-U
G-C      U-A
5'-U   U A U   A-3'
          ^
          Junction
```

Base-Pairs: The usual suspects (A-U, G-C), PLUS some others

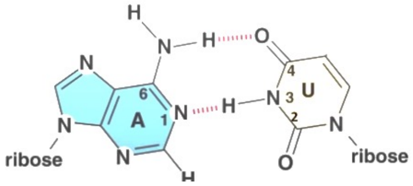
Additional base baring are common in RNA structure

Important example:

The G-U "wobble" pair

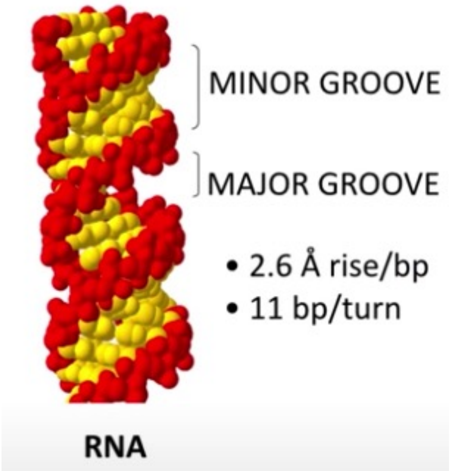


Compare the G-U pair with an A-U pair:

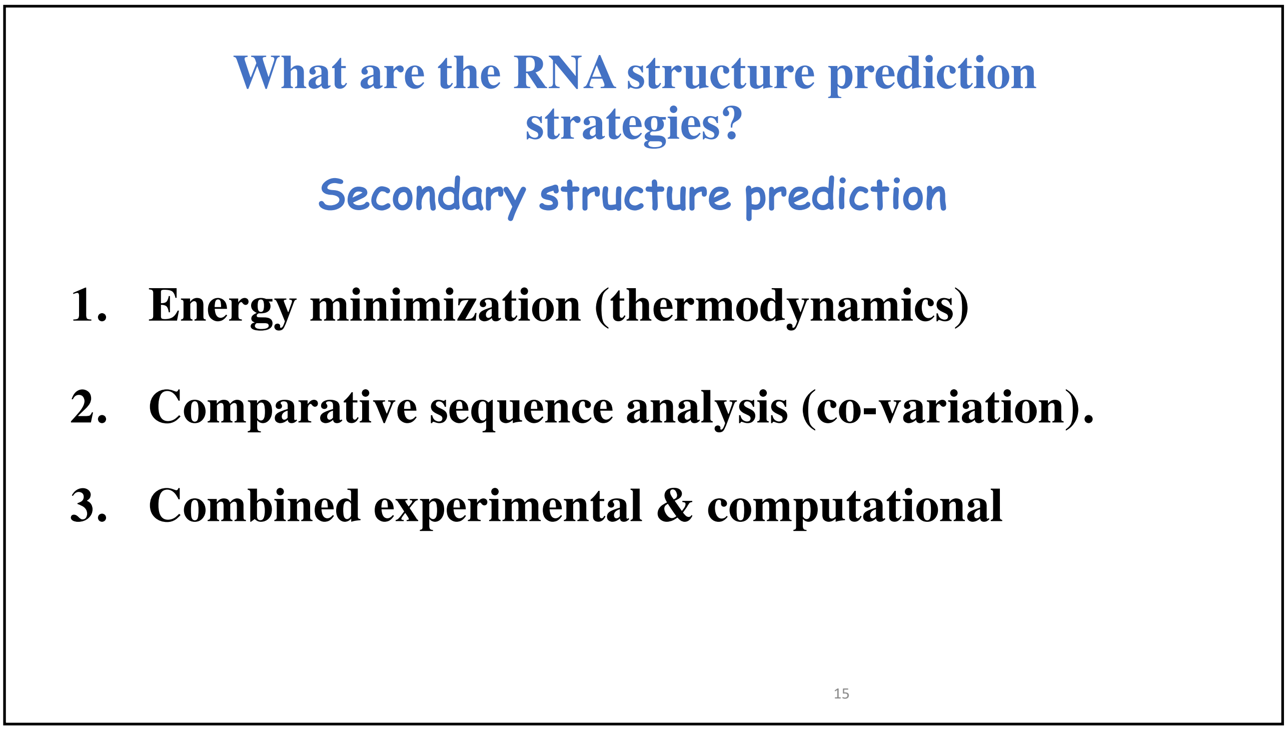


RNA Tertiary structure

- It is a specific arrangement of helices and of structure units in 3 dimension.
- RNA has multi-base interactions.
- It shows the minor and major grooves in the 3D.



- 2.6 Å rise/bp
- 11 bp/turn



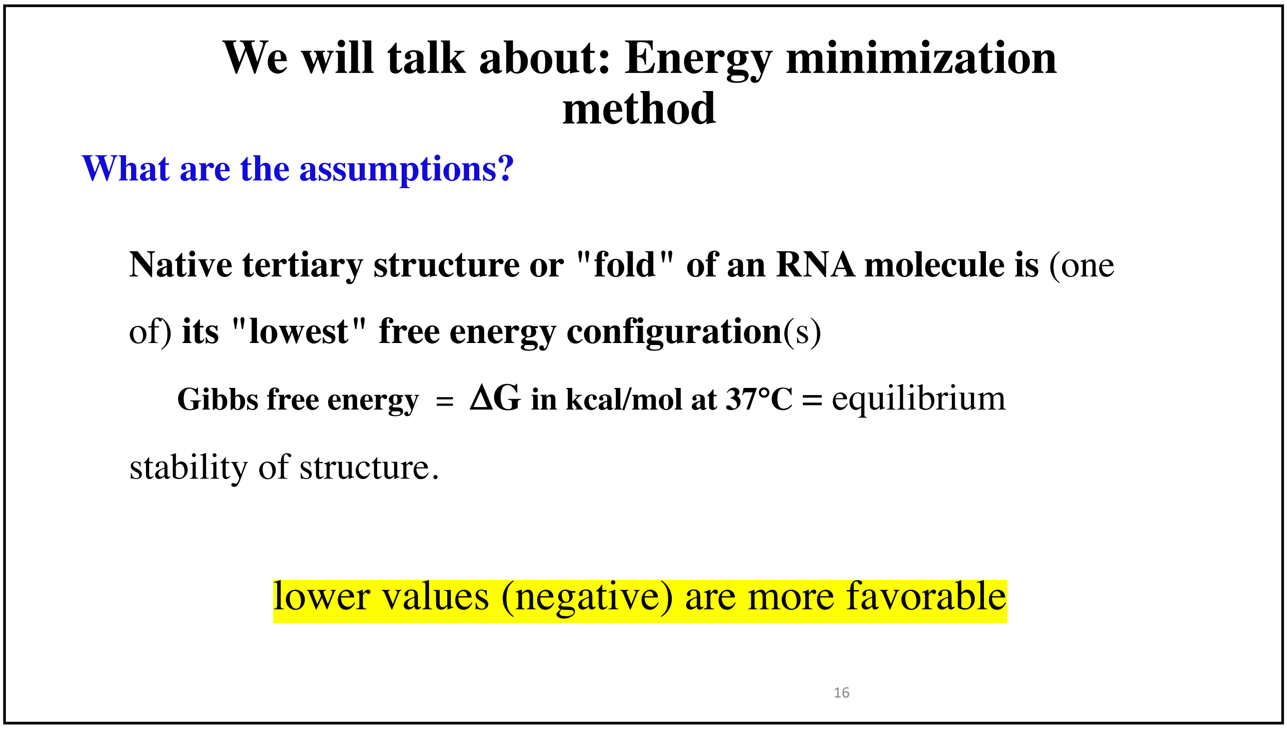
What are the RNA structure prediction strategies?

Secondary structure prediction

1. **Energy minimization (thermodynamics)**
2. **Comparative sequence analysis (co-variation).**
3. **Combined experimental & computational**

15

15



We will talk about: Energy minimization method

What are the assumptions?

Native tertiary structure or "fold" of an RNA molecule is (one of) **its "lowest" free energy configuration**(s)

Gibbs free energy = ΔG in kcal/mol at 37°C = equilibrium stability of structure.

lower values (negative) are more favorable

16

16

What is Gibbs Free energy (G)?

Gibbs Free energy (G) is formally defined in terms of state functions **enthalpy** & **entropy,** & state variable, **temperature**

$$\mathbf{G = H - TS}$$

$$\mathbf{\Delta G = \Delta H - T\Delta S \;\; (for\ constant\ temp)}$$

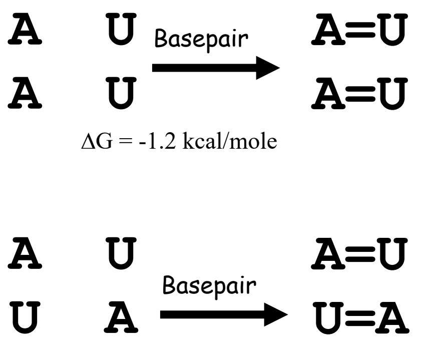
Enthalpy (H) = amount of heat absorbed by a system at constant pressure

Entropy (S) = measure of the amount of disorder or randomness in a system

Note = this is not the same as "entropy" in information theory, but is related

Free energy minimization

What are the rules?



ΔG = -1.2 kcal/mole

ΔG = -1.6 kcal/mole

What gives here?

Why 1.2 vs 1.6?

Because it is backward!

The best configuration when the free energy value les than 0

- https://www.youtube.com/watch?v=WCrlm18KQ48&t=26s
- https://www.youtube.com/watch?v=YB13BkjN8RA
- https://www.youtube.com/watch?v=encRU80nOHg

19

Bioinformatics II: Lecture 7: Protein Structure, Function and predication

Dr Manaf A Guma

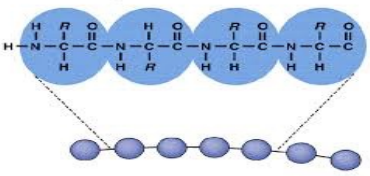
University Of Anbar- college of applied sciences-Heet.

Department of chemistry

1

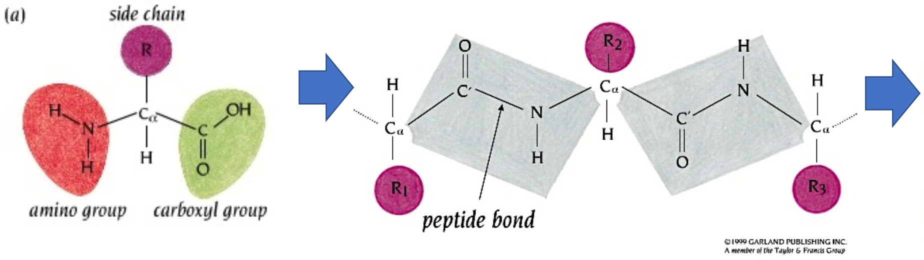
What is a protein?

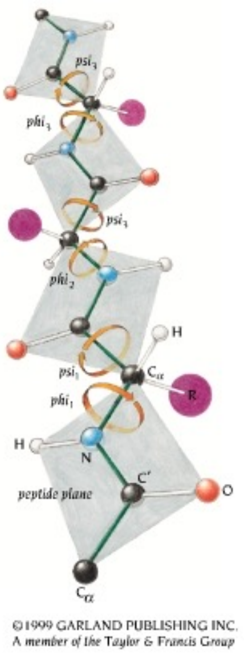
- Proteins is a biological molecules that our body is built of.
- It consists of sequence of amino acids connected as a polymer ‘’ multi mere’’ by peptide bonds
- It is folded in a specific 3D shape to do a specific function in the living cells.
- Proteins have different structures and functions based on the places where they exist in.

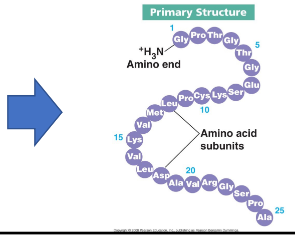


2

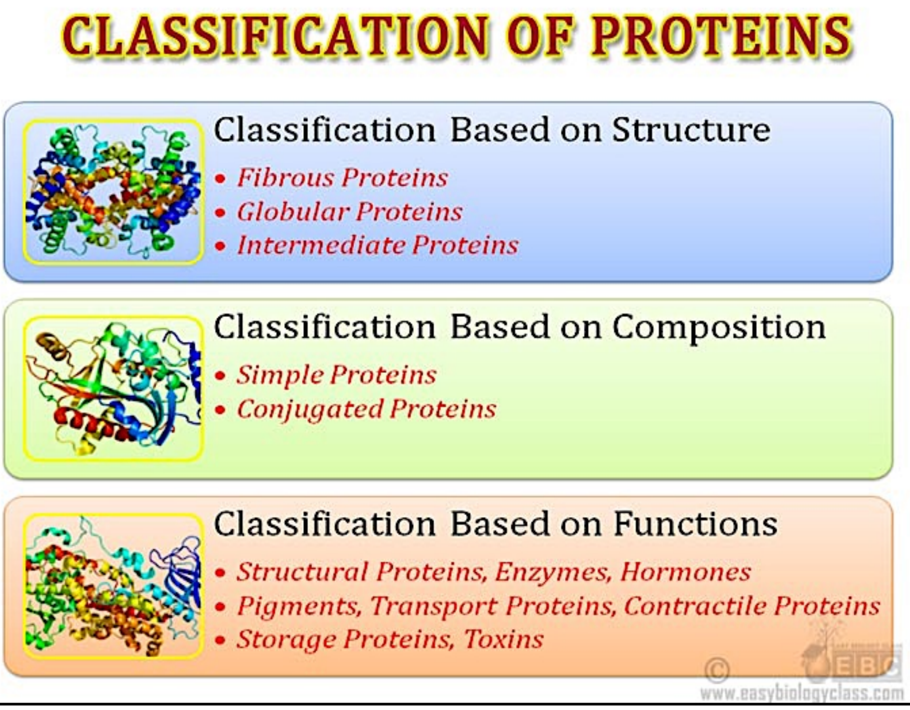
Protein: amino acid →peptide → polypeptide →protein




psi3
phi3
psi2
phi2
H
psi1
Cα
phi1
R
H
N
C′
O
peptide plane
Cα
©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group



3



CLASSIFICATION OF PROTEINS

Classification Based on Structure

- Fibrous Proteins
- Globular Proteins
- Intermediate Proteins

Classification Based on Composition

- Simple Proteins
- Conjugated Proteins

Classification Based on Functions

- Structural Proteins, Enzymes, Hormones
- Pigments, Transport Proteins, Contractile Proteins
- Storage Proteins, Toxins

4

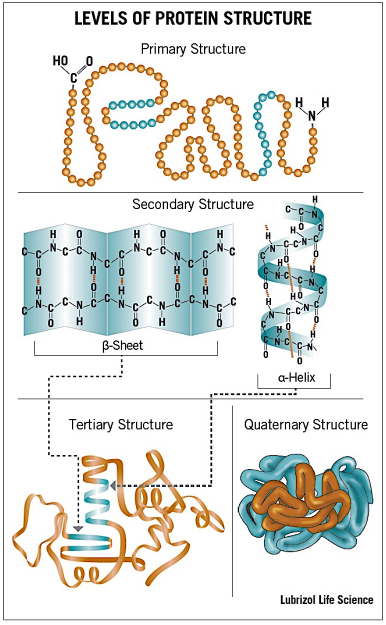
How can we study both: Protein Structure & Function of the protein?

- *Protein structure* - primarily determined by sequence.
- *Protein function* - primarily determined by structure.
- What can that (protein structure and function) be useful in bioinformatics?
- Most of amino acid sequences for specifics regions of a protein are similar to any other protein.
- So, by identifying protein sequences we can predict the structure and the function of a protein.

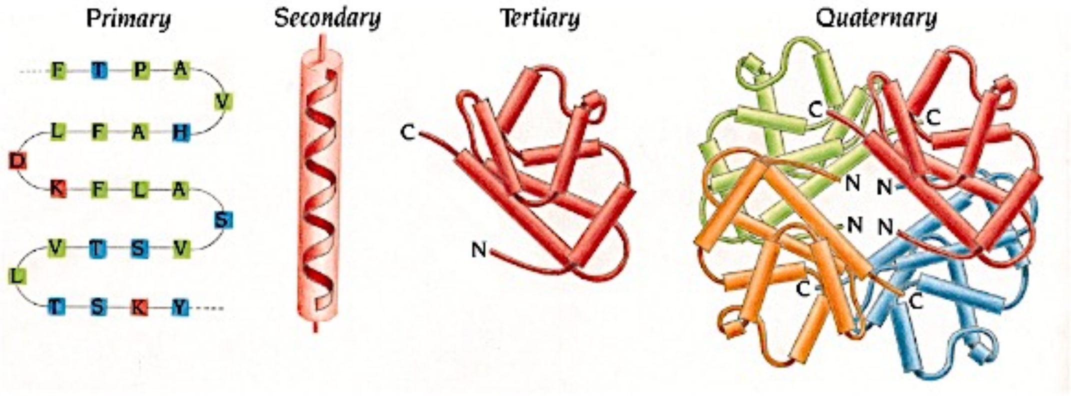
What are the different structures of a protein?

1. Primary structures (simple form)
2. Secondary structures (alpha helix and beta sheets) (3D)
3. Tartary structures
4. Quaternary structures.

They differ by the types of bonds that connect each others.



What are the basic Levels of Protein Structure?

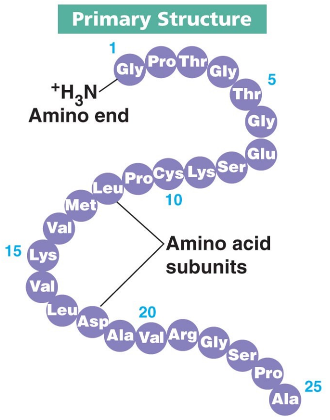


©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

7

What is the Primary structure of a protein?

- The Primary structures of protein are:
- Linear sequence of amino acids
- This linear sequence is referred to as a polypeptide chain. The amino acids in the **primary structure** are held together by covalent bonds.

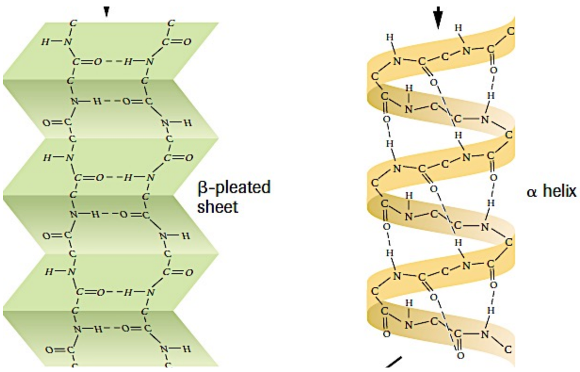


Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.

8

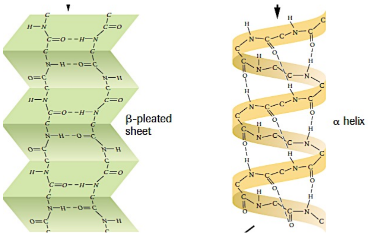
What is the secondary structure of protein?

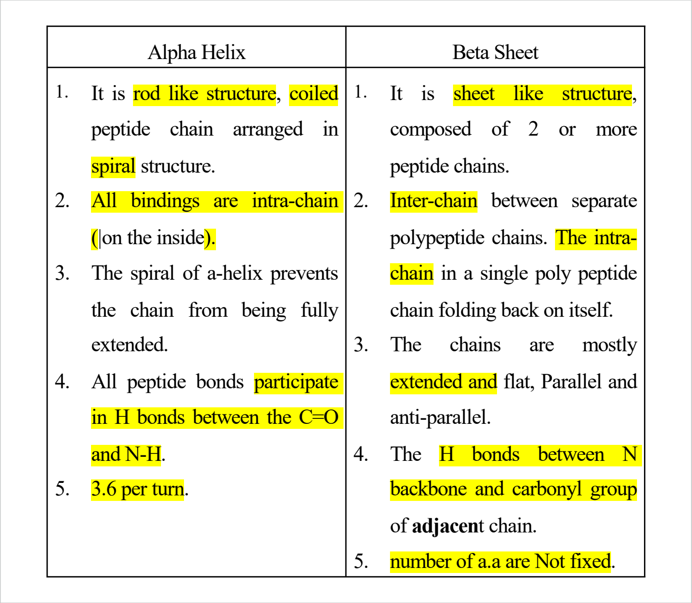
- The Secondary structures of protein are:
- Mostly consists of 2 types which are the α helix and the β pleated sheet.
- Other 2nd str:
- Both **structures** are held in shape by hydrogen bonds, which form between the carbonyl O of one amino acid and the amino H of another



9

What are basic differences between a-helix and b-sheet?

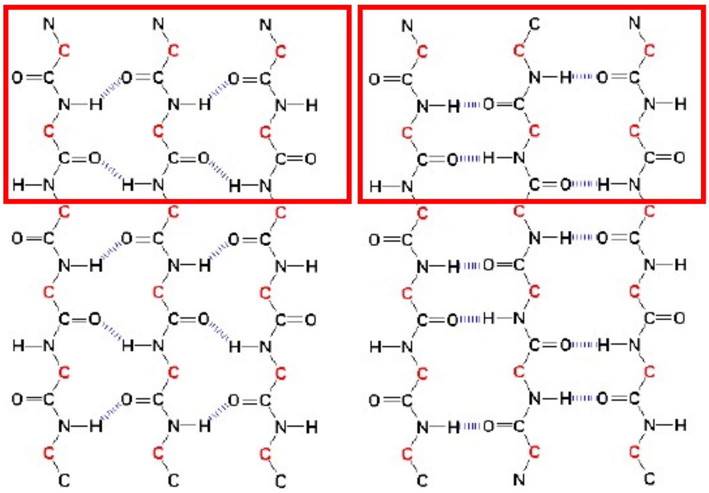




Alpha Helix	Beta Sheet
1. It is rod like structure, coiled peptide chain arranged in spiral structure. 2. All bindings are intra-chain (|on the inside). 3. The spiral of a-helix prevents the chain from being fully extended. 4. All peptide bonds participate in H bonds between the C=O and N-H. 5. 3.6 per turn.	1. It is sheet like structure, composed of 2 or more peptide chains. 2. Inter-chain between separate polypeptide chains. The intra-chain in a single poly peptide chain folding back on itself. 3. The chains are mostly extended and flat, Parallel and anti-parallel. 4. The H bonds between N backbone and carbonyl group of **adjacen**t chain. 5. number of a.a are Not fixed.

10

How many types can b-sheet form?



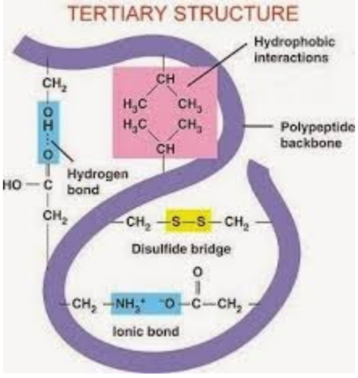
Parellel form

Anti-parellel

11

What are the tertiary & quaternary structures of protein?

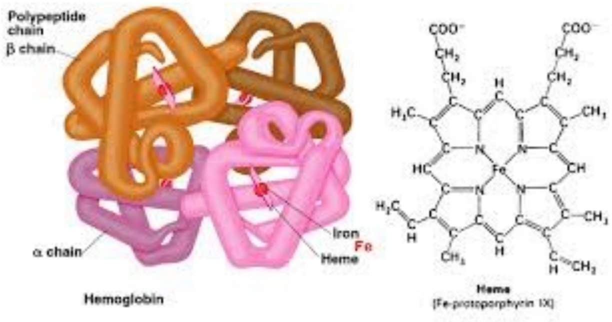
- Tertiary: It is 3D structures shape of protein.
- It has a single polypeptide chain "backbone" with one or more protein secondary structures that form the protein domain.
- Bonds?



12

What is the Quaternary structure of proteins?

- Quaternary:
- It is an arrangement of multiple folded protein subunits in a multi-subunit complex.
- It involves at least 2 polypeptides (domains).
- It can be a dimer, tetramer, homo or hetero protein.



what are the "Additional" Structural Levels of proteins?

1. **Coils:** type of 2nd structure that are not helices, sheets, or recognizable turns.
2. **Loops**: type of 2nd structure.
3. **Motifs**: combinations of 2nd structural elements
4. **Domains**: combinations of motifs.

- e.g: Globular proteins are built from recurring structural patterns
- Please look for their structure?

What are the types of protein structures based on their folding?

- Folding protein is a process by which a polypeptide chain folds to become a biologically active protein in its native 3D structure.
- Types?

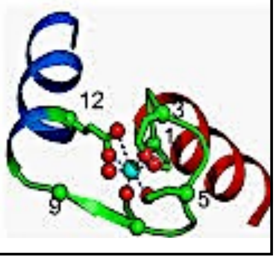
1. Folded proteins: they are typically stable.
2. Unfolded proteins: very little.
3. Partially folded.
4. *Intrinsically disordered proteins: unstable protein because they are* dynamic.

- Predicting protein structure and function can be very hard & *fun!*

15

What are Loops?

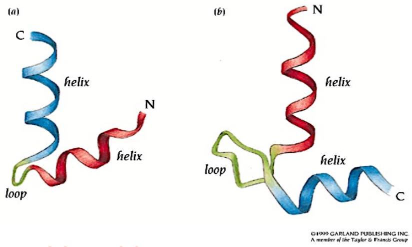
1. **Type of 2nd structure.**
2. **Connect helices and sheets.**
3. **They are usually located on surface of structure**
4. **Are more flexible and can adopt multiple conformations**
5. **Tend to have charged and polar amino acids**
6. **Usually involve active sites such as Calcium binding sites**
7. **Some fall into distinct structural families (e.g., hairpin loops, reverse turns)**



16

What are the most common structural motifs?

1. Helix-turn-helix e.g., DNA binding **H-T-H**
2. Helix-loop-helix **H-L-H** e.g., Calcium binding sites in troponin.
3. Coiled-coil: two coils twist over each other.



©1999 GARLAND PUBLISHING INC. A member of the Taylor & Francis Group

H-T-H H-L-H

17

Why do we study the protein structure in bioinformatics?

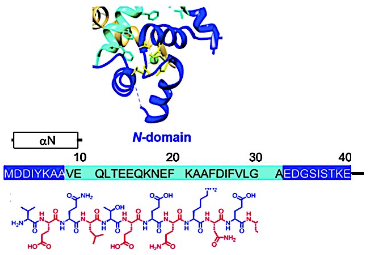
- To be able to predict the structure of the protein by simulation software's using the sequences of known proteins.
- Also, we can do docking, which is used to design drugs, that shows the ability of the protein to interact to the drug.
- Same sequences of a protein may have same function in other proteins.

18

How can we predict the 3D structure of a protein?

- Protein Structure Prediction or "Protein Folding" Problem

1. **shows the amino acid sequence of a protein.**
2. **predict its 3-dimensional structure (fold).**



The opposite of this is:

- "Inverse Folding" Problem
- **predict a protein fold.**
- **identify every amino acid sequence that can adopt its 3-D structure.**
- **Note: The are many websites to predict the 3D** Structure **of the protein based on the database.**

19

Protein structure databases, structural classification & visualization

Please search for types of proteins in the following webs.

1. PDB = Protein Data Bank http://www.rcsb.org/pdb/
2. (RISC) - several different structure viewers http://www.pdg.cnb.uam.es/cursos/Barcelona2002/pages/Farmac/CATH/index.html
3. MMDB = Molecular Modeling Database
4. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Structure

   (NCBI Entrez) - Cn3D viewer

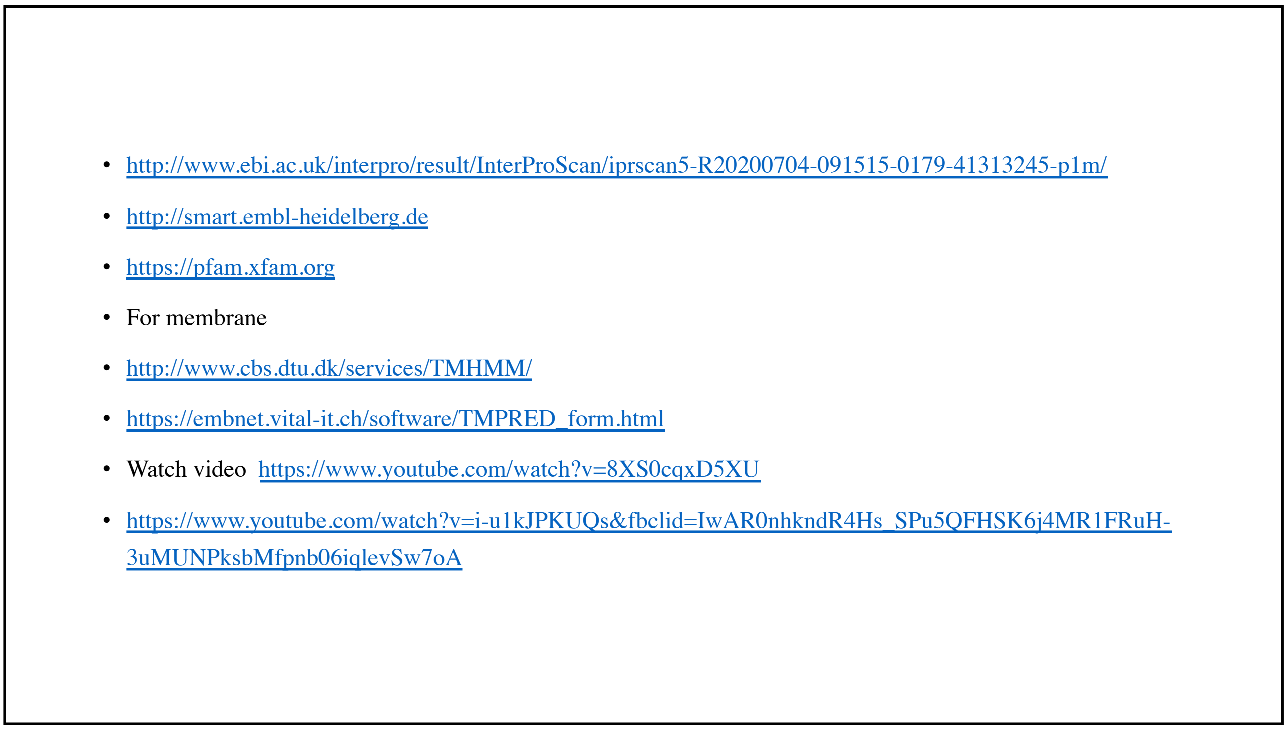
5. SCOP = Structural Classification of Proteins http://scop.mrc-lmb.cam.ac.uk

   Levels reflect both evolutionary and structural relationships

6. CATH = Classification by Class, Architecture, Topology and Homology http://www.cathdb.info

The next steps is to predict the protein structure, please follow…..

20



- http://www.ebi.ac.uk/interpro/result/InterProScan/iprscan5-R20200704-091515-0179-41313245-p1m/
- http://smart.embl-heidelberg.de
- https://pfam.xfam.org
- For membrane
- http://www.cbs.dtu.dk/services/TMHMM/
- https://embnet.vital-it.ch/software/TMPRED_form.html
- Watch video https://www.youtube.com/watch?v=8XS0cqxD5XU
- https://www.youtube.com/watch?v=i-u1kJPKUQs&fbclid=IwAR0nhkndR4Hs_SPu5QFHSK6j4MR1FRuH-3uMUNPksbMfpnb06iqlevSw7oA

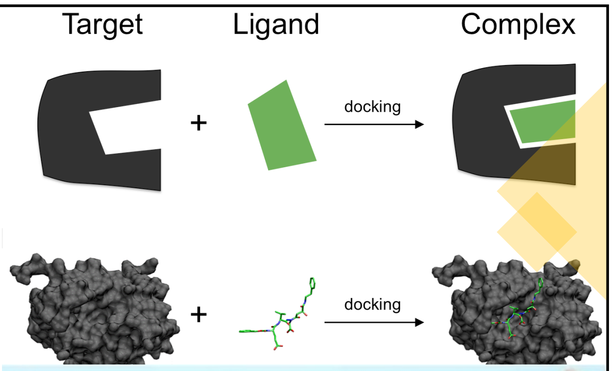
Bioinformatics II:
Lecture 8: Protein docking

Dr Manaf A Guma

University Of Anbar- college of applied sciences-Heet.

Department of chemistry

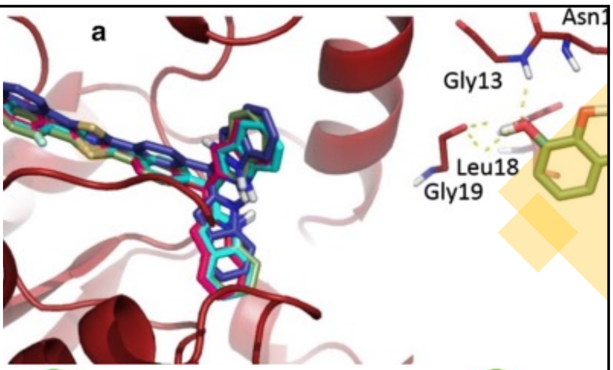
Basics Of Molecular Docking

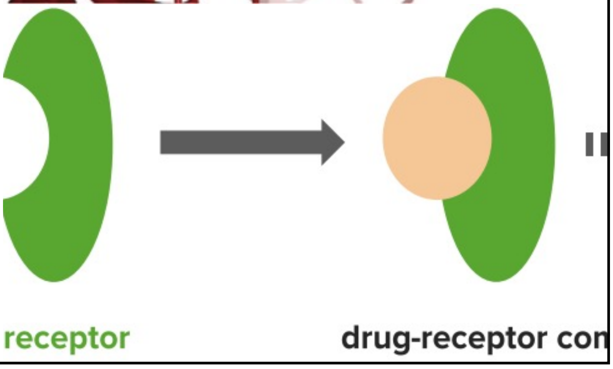


- Docking is a structure-based technique which attempts to find the “best” match, between two molecules.
- WHAT IS MOLECULAR DOCKING?
- In the field of molecular modelling, docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex.

Ligand-receptor Complex

- Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using for example scoring functions.

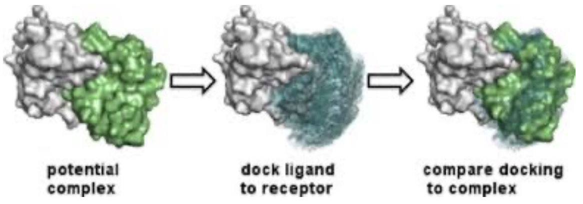


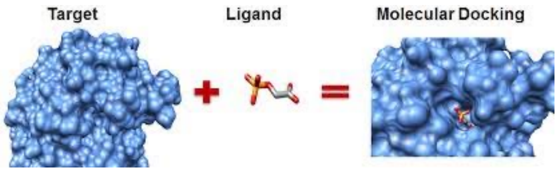


3

Types of docking?

- 1. Protein-protein docking:
- Both molecules usually considered rigid
- 2. Protein-ligand docking:
- Flexible ligand.
- Rigid-receptor





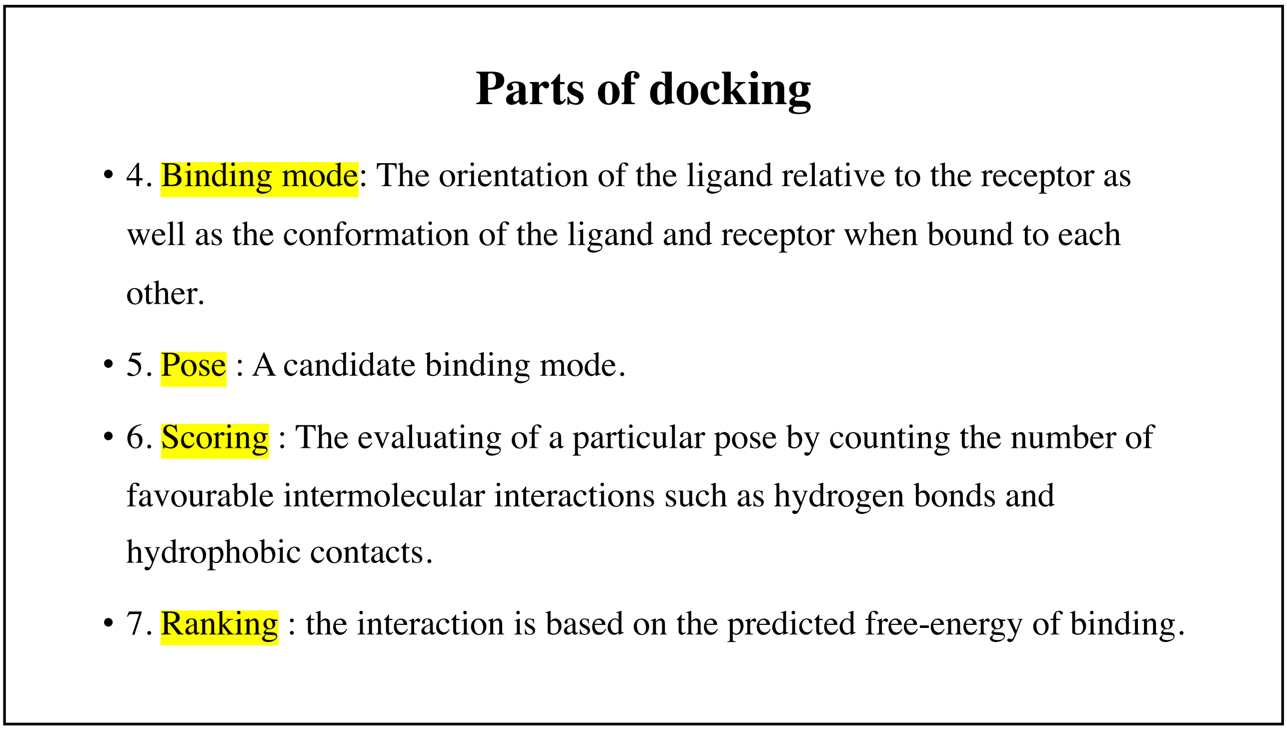
4

WHY is Docking Important?

- 1. Signal Transduction:
- The associations between biologically relevant molecules such as proteins, nucleic acids, carbohydrates, and lipids play a central role in signal transduction.
- 2. Drug-Designing and discovery:
- Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to in turn predict the affinity and activity of the small molecule.
- 3. The behaviour of diseases:
- Understanding the mechanism of interaction between proteins molecules which could explain the cause of diseases.

Parts of docking

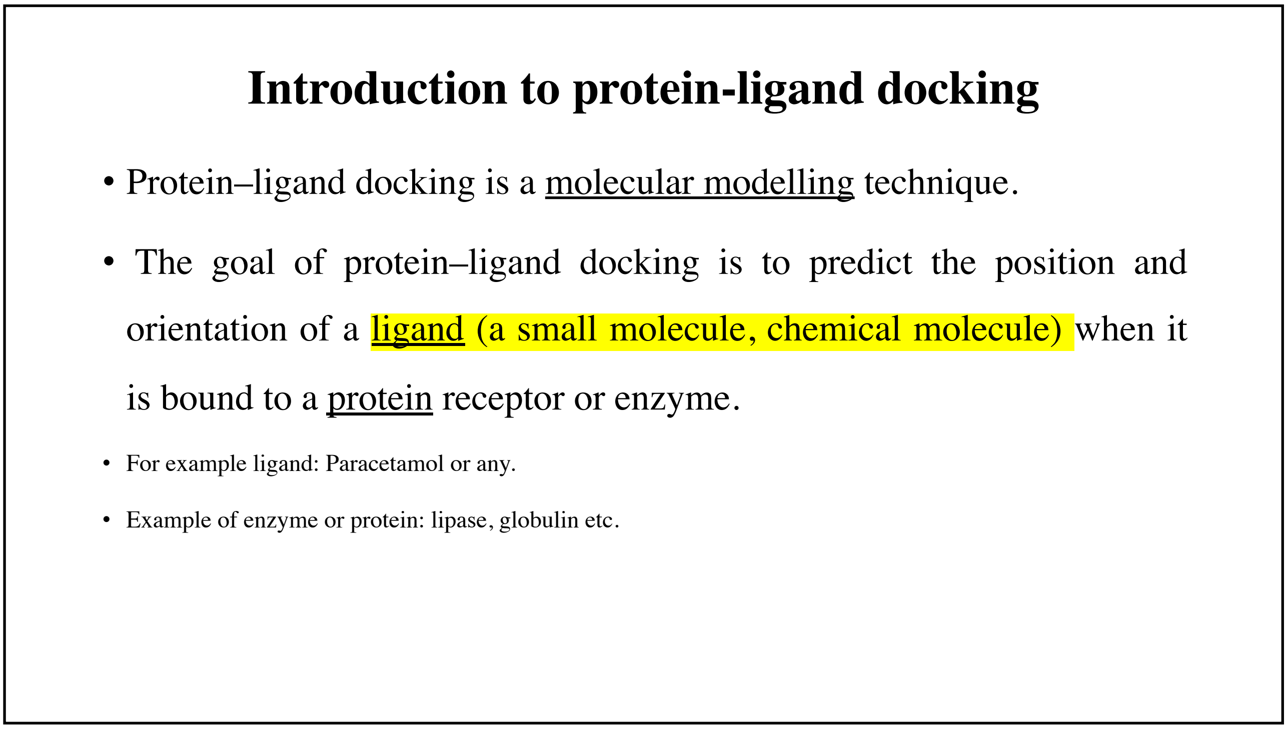
- 1. Receptor or host or lock: The "receiving" molecule, most commonly a protein or other biopolymer.
- 2. Ligand or guest or key: The complementary partner molecule which binds to the receptor. Ligands are most often small molecules but could also be another biopolymer.
- 3. Docking: Computational simulation of a candidate ligand binding to a receptor.



Parts of docking

- 4. Binding mode: The orientation of the ligand relative to the receptor as well as the conformation of the ligand and receptor when bound to each other.
- 5. Pose : A candidate binding mode.
- 6. Scoring : The evaluating of a particular pose by counting the number of favourable intermolecular interactions such as hydrogen bonds and hydrophobic contacts.
- 7. Ranking : the interaction is based on the predicted free-energy of binding.

7

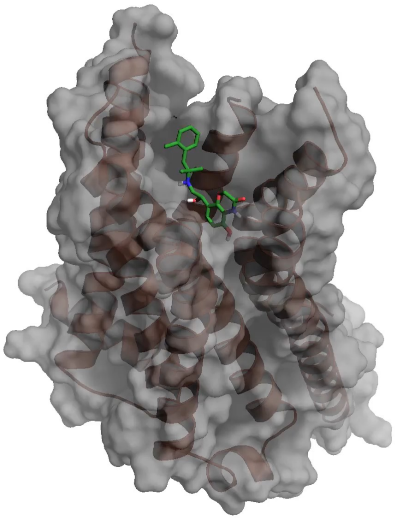


Introduction to protein-ligand docking

- Protein–ligand docking is a molecular modelling technique.
- The goal of protein–ligand docking is to predict the position and orientation of a ligand (a small molecule, chemical molecule) when it is bound to a protein receptor or enzyme.
- For example ligand: Paracetamol or any.
- Example of enzyme or protein: lipase, globulin etc.

8

Example

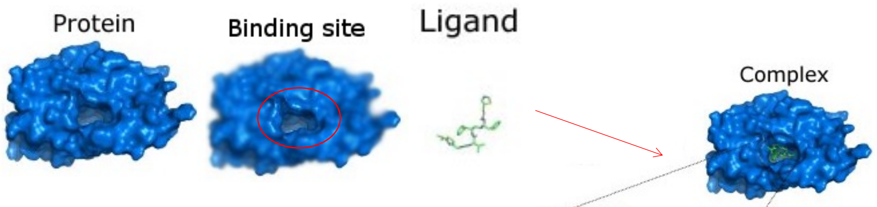


Computer-aided drug design (CADD)

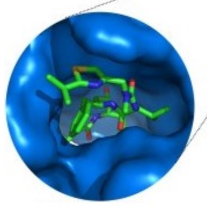
	Known ligand(s)	No known ligand
Known protein structure	**Structure-based drug design (SBDD)** Protein-ligand docking	***De novo* design**
Unknown protein structure	**Ligand-based drug design (LBDD)** *1 or more ligands* • Similarity searching *Several ligands* • Pharmacophore searching *Many ligands (20+)* • Quantitative Structure-Activity Relationships (QSAR)	**CADD of no use** Need experimental data of some sort

Protein-ligand docking

- A Structure-Based Drug Design (SBDD) method
  - “structure” means “using protein structure”
- Computational method that mimics the binding of a ligand to a protein
- **Given...**



- **Predicts...**
  - The **pose** of the molecule in the binding site
  - The binding affinity or a **score** representing the strength of binding

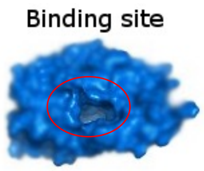


11

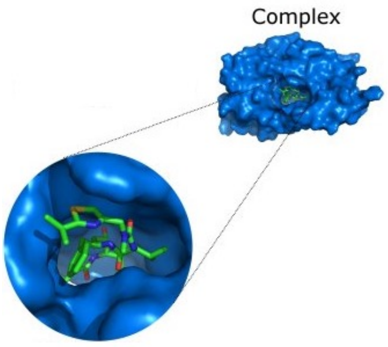
Pose vs. binding site

- **Binding site** (or “active site”)
  - the part of the protein where the ligand binds
  - generally a cavity on the protein surface

  We must have the crystal structure of the protein bound with a known inhibitor.



- **Pose** (or “binding mode”)
  - The *geometry* of the ligand in the binding site
  - Geometry = **location, orientation and conformation**
- *Protein-ligand docking is **not** about identifying the binding site*

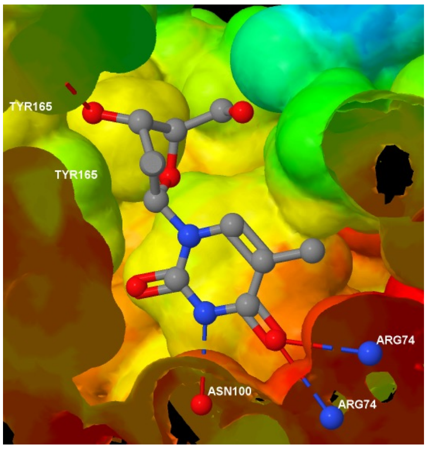


12

Uses of docking

- The main uses of protein-ligand docking are for
  - Virtual screening,
  - Pose prediction

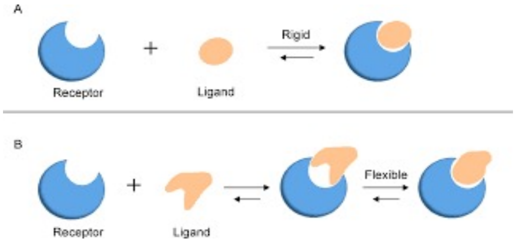
- **Pose prediction**
- If we know exactly where and how a known ligand binds...
  - We can see which parts are important for binding
  - We can suggest changes to improve affinity
  - Avoid changes that will 'clash' with the protein



13

2 Types of ligand-protein docking

- 2. Protein-ligand docking:
- Flexible ligand.
- Rigid-receptor

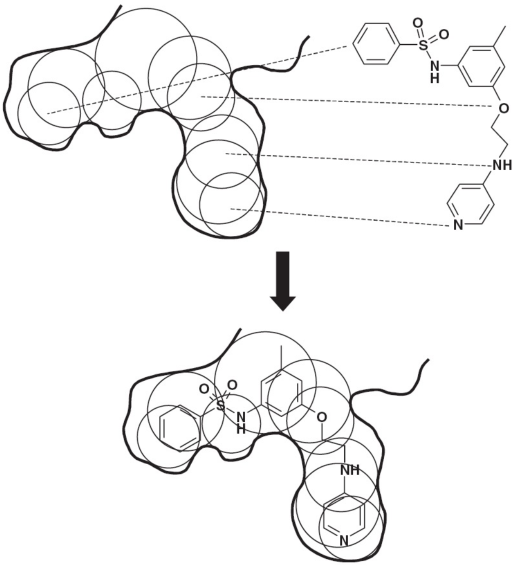


- We can classify the various search algorithms according to the degrees of freedom that they consider
- Rigid docking or flexible docking
  - With respect to the ligand structure

14

Rigid docking

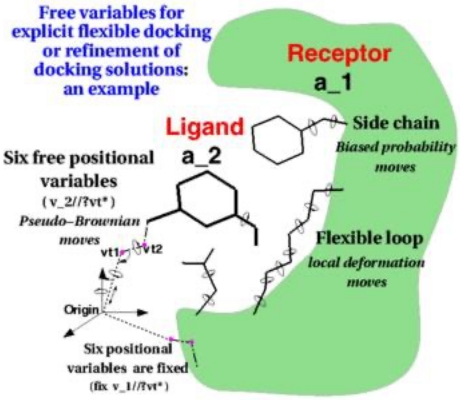
- The ligand is treated as a rigid structure during the docking.
- The DOCK algorithm developed by Kuntz and co-workers is generally considered one of the major advances in protein–ligand docking [Kuntz et al., *JMB*, **1982**, *161*, 269].
- Most docking software treats the protein as rigid



15

Flexible docking

- **Flexible docking:**
- is the most common form of docking today
  - Conformations of each molecule are generated on-the-fly by the search algorithm during the docking process
  - The algorithm can avoid considering conformations that do not fit


Free variables for explicit flexible docking or refinement of docking solutions: an example
Receptor
a_1
Ligand
a_2
Side chain
Biased probability moves
Six free positional variables
(v_2//?vt*)
Pseudo–Brownian moves
vt1
vt2
Flexible loop
local deformation moves
Origin
Six positional variables are fixed
(fix v_1//?vt*)

16

Böhm's empirical scoring function

- In general, scoring functions assume that the free energy of binding can be written as a linear sum of terms to reflect the various contributions to binding.
- Bohm's scoring function included contributions from hydrogen bonding, ionic interactions, lipophilic interactions and the loss of internal conformational freedom of the ligand.

$$\begin{aligned}\Delta G_{bind} = \Delta G_0 &+ \Delta G_{hb} \sum_{h-bonds} f(\Delta R, \Delta \alpha) \\ &+ \Delta G_{ionic} \sum_{\substack{\text{ionic} \\ \text{interactions}}} f(\Delta R, \Delta \alpha) \\ &+ \Delta G_{lipo} \left| A_{lipo} \right| + \Delta G_{rot} NROT\end{aligned}$$

17

ΔG_o calculation

- In analogy with any spontaneous process, protein–ligand **binding** occurs only when the change in Gibbs **free energy** (ΔG) of the system is **negative** when the system reaches an equilibrium state at constant pressure and temperature.
- The **binding free energy** can be **calculated** using the rate constants k_{on} and k_{off} as
- ΔG=G **bound**-G unbound =-kTln KeqC_0=-kTln $C_0 k_{on}/k_{off}$
- where K_{eq} is the **binding** equilibrium constant, C_0 is the reference concentration of 1 mol/L, k is Boltzmann's constant and T is the temperature in Kelvin

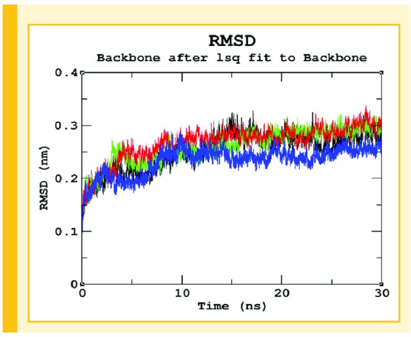
18

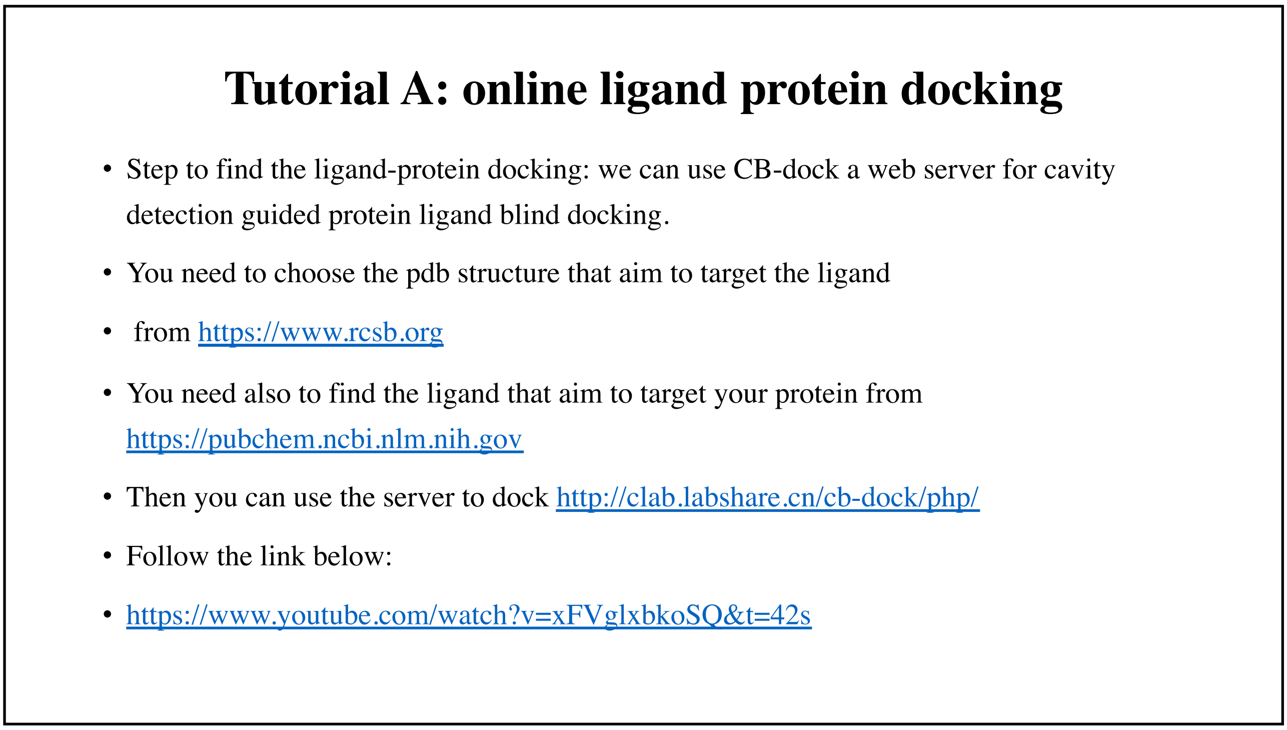
ΔG_o calculation reflect the various contributions to binding.

- The ΔG values on the right of the equation are all constants.
- **ΔG_o** is a contribution to the binding energy that does not directly depend on any specific interactions with the protein
- The **hydrogen bonding** and **ionic terms** are both dependent on the geometry of the interaction, with large deviations from ideal geometries (ideal distance R, ideal angle α) being penalized.
- The **lipophilic term** is proportional to the contact surface area (Alipo) between protein and ligand involving non-polar atoms.
- The **conformational entropy term** is the penalty associated with freezing internal rotations of the ligand. It is largely entropic in nature. Here the value is directly proportional to the number of rotatable bonds in the ligand (NROT).

Accuracy measured by RMSD

- Accuracy od docking is measured by RMSD (root mean squared deviation) compared to known crystal structures
  - RMSD = square root of the average of (the difference between a particular coordinate in the crystal and that coordinate in the pose)^2
  - Within 2.0Å RMSD considered cut-off for accuracy
  - More sophisticated measures have been proposed, but are not widely adopted

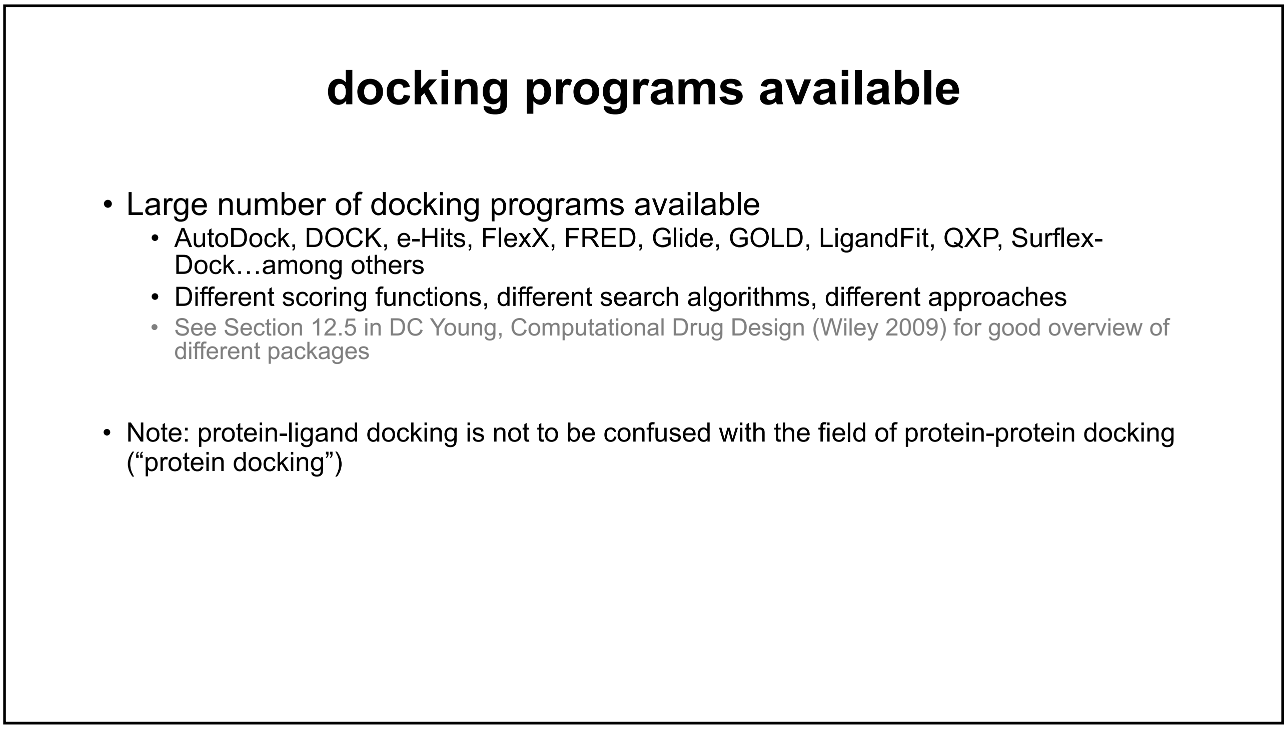




Tutorial A: online ligand protein docking

- Step to find the ligand-protein docking: we can use CB-dock a web server for cavity detection guided protein ligand blind docking.
- You need to choose the pdb structure that aim to target the ligand
- from https://www.rcsb.org
- You need also to find the ligand that aim to target your protein from https://pubchem.ncbi.nlm.nih.gov
- Then you can use the server to dock http://clab.labshare.cn/cb-dock/php/
- Follow the link below:
- https://www.youtube.com/watch?v=xFVglxbkoSQ&t=42s

21



docking programs available

- Large number of docking programs available
  - AutoDock, DOCK, e-Hits, FlexX, FRED, Glide, GOLD, LigandFit, QXP, Surflex-Dock…among others
  - Different scoring functions, different search algorithms, different approaches
  - See Section 12.5 in DC Young, Computational Drug Design (Wiley 2009) for good overview of different packages
- Note: protein-ligand docking is not to be confused with the field of protein-protein docking (“protein docking”)

22

Tutorial B: online protein- protein docking

- You can to the tutorial following the link below:
- https://www.youtube.com/watch?v=8-IPJqXYQ3Q

23