MODULE 2 SESSION 3
GENOME COMPARATIVE GENOMICS

- What is it?:
- Genomic Variation/Comparative Genomics: WWWH?
- The Input
- The Methods
- The Output

Learning Outcomes
Session 1: Comparative Genomics

- Navigate through genomic resources to:
    - Retrieve Information on a specific gene (sequence, variants, orthologs…)
    - View and interpret genomic alignments

Comparative Genomics: a definition

- Comparative genomics is based on the fact that a genomic variation is happening in all organisms
- These changes affects several features in a genome (structure, organization, functions… that are embedded in this genome)
- The changes could help monitor evolution between organisms (species…)
- Comparative genomics deals with the process of comparing the sequences of **whole** or **parts (big chunks of DNA or even smaller portions like genes)** of **genomes**.
- Goal:
    - Identify **similarities and differences** between **features** in these genomes
    - Identify **evolutionary relationship** between organisms

Therefore: 2 keywords very important

- Comparative genomics: **features + evolution**

You could be trying to ***compare organisms that are relatively close to each other*** in ***terms of evolutionary*** and then you get out of this comparism a ***certain level of information***. You can also try ***comparing completely distant organism*** and you will get ***completely different information out of it***.
E.g. Comparing these meals:


Green Pepper


Red Pepper


Different Meal

GENOMIC VARIATION/COMPARATIVE GENOMICS: WWWH?

The **whole field of comparative genomics** is **based on** the **existence of genomic variation** or **certain level of genomic variation in all organisms**. **Before starting to talk about comparative genomics** and the **methods associated to this field** we first have to look at the ***genomic variation*** ***that we could*** **detect and could be** ***interesting to retrieve*** **using comparative genomics**.

Basic Questions about Genomic Variation:

Genomic variation : WWWH

- Why? Why will Genomic Variation be needed for an organism?
- When? When will these changes be beneficial for this organism?
- What? What are the changes that we could detect?
- How? How will we measure that?

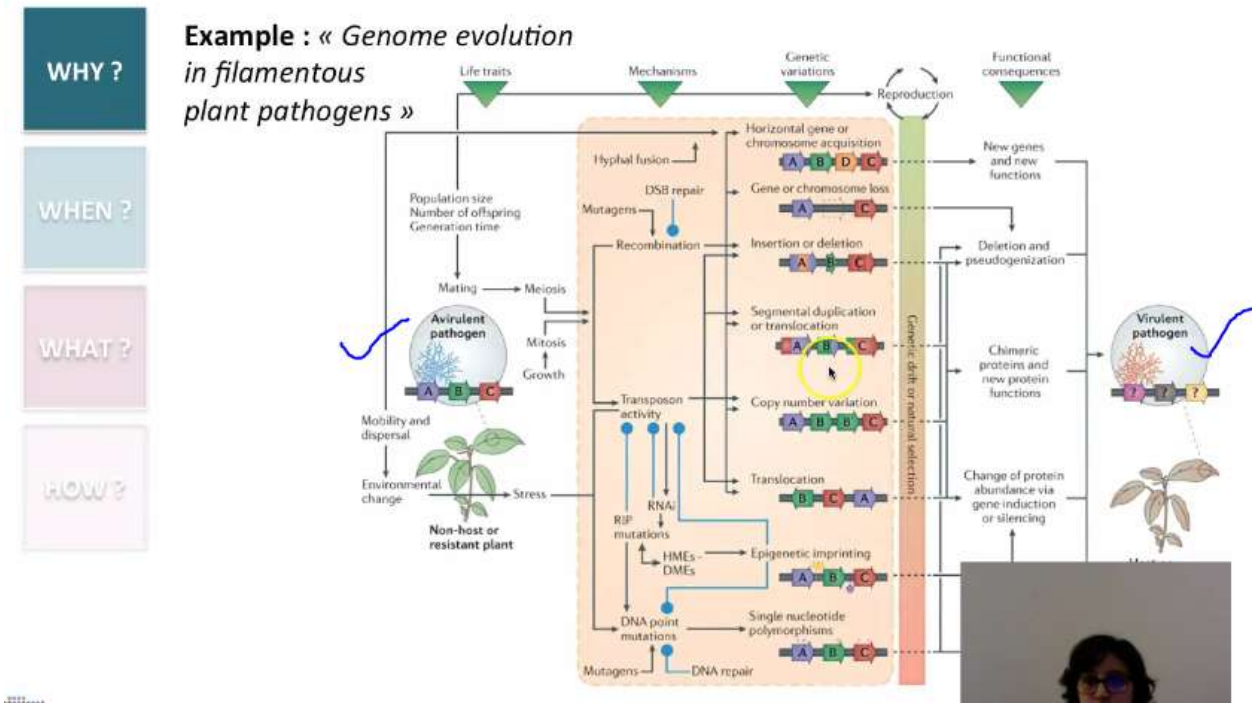
WHY? WHY WILL GENOMIC VARIATION BE NEEDED FOR AN ORGANISM?

Why would a genome evolve?

- Genomic plasticity allows an organism to:
    - Adapt to environmental changes
    - Find the best evolution changes
    - Acquire virulence genes, enhanced pathogenicity
    - Resistance to drugs
    - Increase survival chances of members of a population

E.g. A filamentous plant pathogens (life treat and mechanism affecting their genome and the genetic variations that are occurring and the functional consequences of this genetic variation) so different factors like **mating** etc or **different environmental induced factors** can affect the genome of the species, we are talking about the pathogen living in a plant that is resistance to that pathogen because the pathogen is **Avirulent Pathogen**, when it comes to mating or different kind of environmental factors and these environmental factors affects the genome of the **Avirulent Pathogen** then it could become a **Virulent Pathogen**.

Genomic variation : WWWH



So if this is an evolution that will be **beneficial** for that pathogen, then it could acquire it. And then this will conserved throughout its' evolution in terms of novel character that is acquired, so this pathogen will become in this case **fully pathogenic**.

When? When will these changes be beneficial for this organism?

Factors/events

- Gene transfer (for all organism)
- Environmental pressure for selection (beneficial or detrimental)
    - pH
    - temperature
    - host (changes)
    - pathogen (that is affecting something)
- A genetic variation could occur in response of **factors / events**

What? What are the changes that we could detect?

What could be affected?

- Overall genomic sequence (re-arrangements)
- DNA structure (could be modified)
- Regulatory elements
- Genes size, number, function, density (could be affected at this level
- Nucleotide composition (could be affected too)

How? How will we measure that?

How could this happen?

This could be due to…

- Large genetic structural variations (could be duplication, recombination… etc possibly events that will draw a different kind of diversion within a genome, this could giving rise to gene families through duplication events – with the duplication you could have maintenance of the function or not)
- Transposable elements (retrotransposons…) that could provide crossovers between two different sites, even if they are misaligned, this contributes to genomic evolution.
- Evolution of multigene families
- Evolution of genes with novel functions (could be acquired because a copy could be altered or not, you could maintain same function or have another function.
- Exon shuffling
- Tandem repeats modifications (basically, this is what we see with mini-satellite and micro-satellite) and several other kind of events that can lead to genetic variation

How to measure the changes in a genome?

- **Sequence Variation (comparison you can do) ← BASIC of Comparative Genomics**
    - Between 2 genomes (1 reference)
    - Between several genomes
- Other variations (structure, protein folding…) Won't be covered in this course work

Now using the Sequence Variation as the BASIC lets go over the questions WWWH

Why would we compare genomes?

- Identify evolutionary history (between the genomes)
- Highlight synteny (Usman A. Lamidi define synteny…?)
- Identify genomic rearrangements (large SV events…)
- Study convergent evolution for some organisms (e.g. viruses) Different viral lineages can affect the same host, and this can be use to study the convergent evolution of the factors that are commonly influencing the evolution of the viruses, this could be helpful to the viruses because that are often so different that you can only compare those that are closely related the lineages that are closes related.
- Understand disease outbreak
- Identify pathogenicity markers, drug targets

❖ **Why would we compare "portions" of genomes?**

- Comparing smaller portions of a genome allows to zoom into regions of genomic re-arrangements
- Could be genes:
    - Screen for functional genes **gain**
    - Screen for functional genes **loss**

- Gain of a **new function**
- **Exons** (length, number…)
- Conserved **pathways**
- **Coding / non-coding**

❖ **When do we need to use comparative genomics?**
- Establish genetic and evolutionary relationship between:
  - Entire organisms
  - Sequences

❖ **What we generally compare are <u>features</u> of 1 or more genomes to features of another <u>genome (reference)</u>**
❖ **A genome is <u>complex</u> and composed of different elements (regulatory, structural…)**
❖ **In fact, there are different types of DNA features that can be compared between 2 or more genomes:**
  - **DNA sequences** (small, large, coding/non-coding)
  - **Genes** (nature, order…)
  - Regulatory elements

Focusing on method to compare DNA and Genes will be classified into…

❖ **Could be classified in:**
  - Genome **Structure**
  - Genome **Function** (coding/non-coding
  - Genome **Evolution**

How do we do the above classification, basically we use….

❖ Comparative genomics uses **Sequence Alignment**
❖ Comparative genomics is based on **Phylogeny principle** that relies on several key issues
  - Several genomes are sequenced and available (one basic issue genome sequences have to be available and at highest possibly finished stages)
  - Homology between genes (similar functions)

In order to do that we will compare sequences and come out with an output that can be use….

❖ Use **complex** model genomes to **infer knowledge** (Annotation: function) to unknown or less complex genomes i.e. In order to **infer knowledge** from a **complex genome** to **unknown** or **less complex genome. NB: This is what we do when we do annotation of gene function.**

So this is done through…

❖ Algorithms/programs

❖ In vitro:

We could be screen for this kind of changes in terms of molecular biology using different types of methodology like…

- Fluorescence In Situ Hybridization **(FISH)**
- Spectral Karyotyping **(SKY)** and Multiplex-FISH **(M-FISH)**
- **CGH** - Comparative Genomic Hybridization

They are quite complex methods and they can be used for screen for **damages, inverted sequences** or **transposable elements** etc whatever you will like to compare between two genomes. **This is done in order to identify changes between the DC stages and normal stages.**

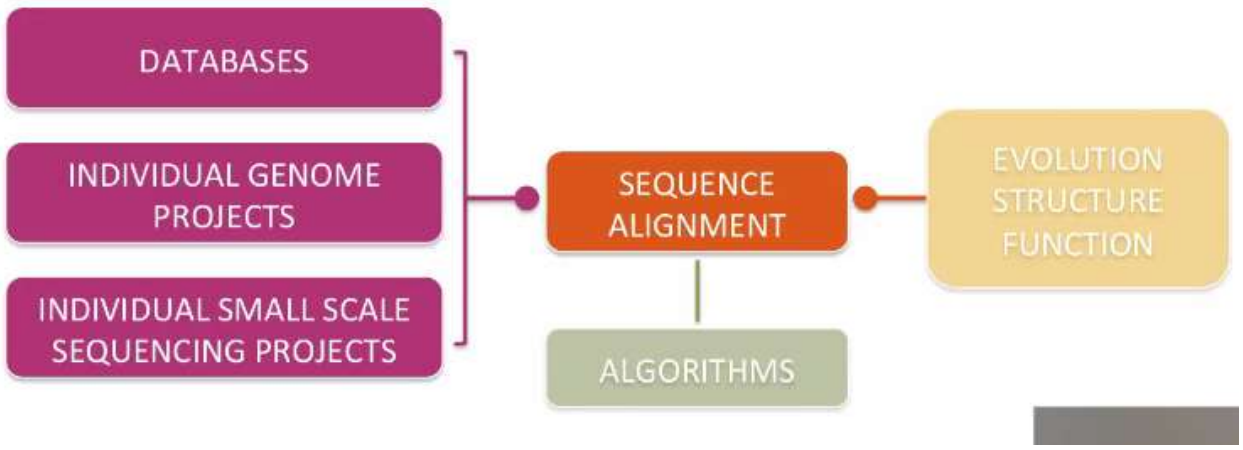
KEYWORDS IN COMPARATIVE GENOMICS

Collect sequences output from Databases OR sequence from individual genome projects or Individual small scale sequencing projects. Once you have collected the sequence of interest you have….

Align the sequences to other sequences from databases or your sequences among each other

Developed using specific **Algorithms**

What you will have at the end of the alignment that is rendered possible with the help of the algorithms are **insights** into the **genomic evolution**, the **structure** and **functions of genomes or portion of genome like genes as we said.**



THE INPUT: The sequences that you have to submit for query

❖ **Availability of genomes sequenced**

Different repositories like **NCBI** has genome available

Example of Site with **Article published for genomes**: **Nucleic Acids Research** and **Database**

User can query the resource for specific organism

1 of the Tools developed to highlight the possible exploitation of sequencing throughout the Database **ANITools** (Average Nucleotide Identity): A web tool for fast genome comparison within multiple bacterial strains (pair wise)

❖ **File Formats**

- **Different file formats may be accepted**
- **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) **in (NCBI) accepts:**
  - **Fasta sequence starts with > OR GenBank .gbk**
  - **Simple Sequence**
  - **Accession Number**
  - **Local file from disk (megablast)**
- **You can use Whole sequence or Subsequence**

Blast is a web resource that enables one to blast Nucleotide to nucleotide,

blastx for	**translated nucleotide to protein**
tblastn for	**protein to translated nucleotide**
Protein blast for	**protein to protein**

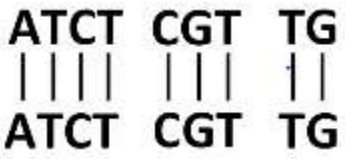
THE METHODS: The methods that you will apply to align your sequences

Sequence Alignment for DNA

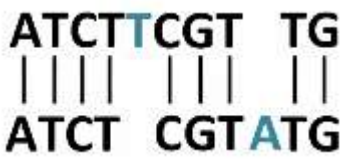
- DNA sequence alignment consists of aligning 2 DNA sequences in order to identify regions showing sequence similarity
- This highlights regions showing relationship in terms of:
  - **Evolution**
  - **Structure**
  - **Function**

Example: For simple sequences: Compare **ATCTTCGTTG** and **ATCTCGTATG**

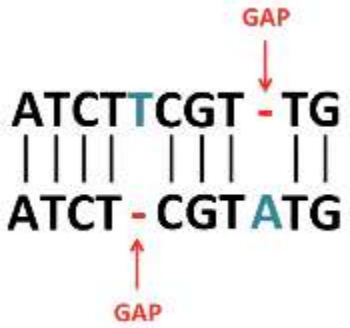
How will you perform a sequence alignment, finding the blocks that are exactly the same between the sequences?



We found 3 blocks that are same within the sequences



We have 2 extra with no correspondence



We will have to add a gap

You can interpret this kind of alignment? If you consider they share common ancesstor, in order for the two sequences to be aligned we have to introduce single point mutation to this alignment, such that a gap will represent an INSERTION or a DELETION in a nucleotide in one or both the sequences.

- **For sequences that are not as simple:**
  - **Long sequences**
  - **More complex sequences (divergent… containing different kind of variations)**
  - **Large number of sequences**

→ Need Algorithms!

TWO key approaches for sequence alignment

- **Global Alignment**
  - ~~Optimizes the alignment to span the full length of sequences that are aligned~~
- **Local Alignment**
  - ~~Optimises the alignment to take into account regions of the highest similarity between divergent sequences~~

- 2 key approaches

  - **Global Alignment**

  → Optimizes the alignment to span the full length of sequences that are aligned.

  ```
  ATCATTCGTTGACTGTG
  A---TT-G-TGAC--TG
  ```

  - **Local Alignment**

  → Optimizes the alignment to take into account regions of the highest similarity between divergent sequences.

  ```
  ATCATTCGTTGACTGTG
  ---ATT-G-TGACTG--
  ```

- **Algorithms efficiency and choice depends on the number of sequences to compare**
- **Pairwise Alignment (Hint: Not ideal from the Indian Video)**
  Sequence alignment of 2 sequences
  →Output: function, structure, evolutionary relationship
- **Multiple Sequence Alignment (MSA)**
  Sequence alignment of 3 or more sequences (same length)

→**Output:** **homology, evolutionary relationship**

PAIRWISE Alignment

- **A pairwise alignment is an optimized local or global alignment of 2 sequences**
  - **3 methods:**
    - **Dot matrix**
    - **Dynamic programming**
    - **Word-based**

NB: Efficiency can be reduced in low complexity regions (repetitive sequences…) can be evaluated by the **MUM** (Maximum Unique Match)

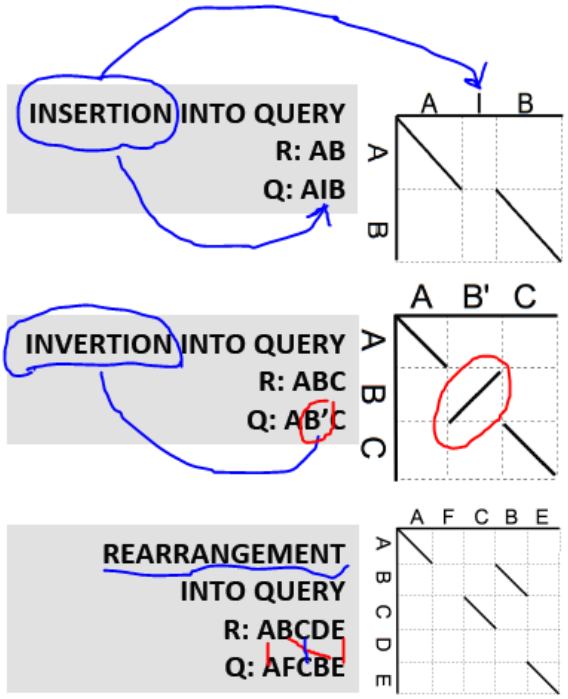
→ **Long MUM** sequences = more related sequences

Pairwise Alignment

- **Dot-matrix method**

  - 2 sequences (A and B) are aligned using a 2-dimensional matrix
  - Identity is shown with a dot
  - Diagonal shows high similarity

  → Dot plot of the sequence R against sequence Q



Adapted from http://mummer.sourceforge.net/manual/AlignmentTypes.pdf

Query on the column and reference on the row
\ (backward slash) shows high similarity
/ inversion

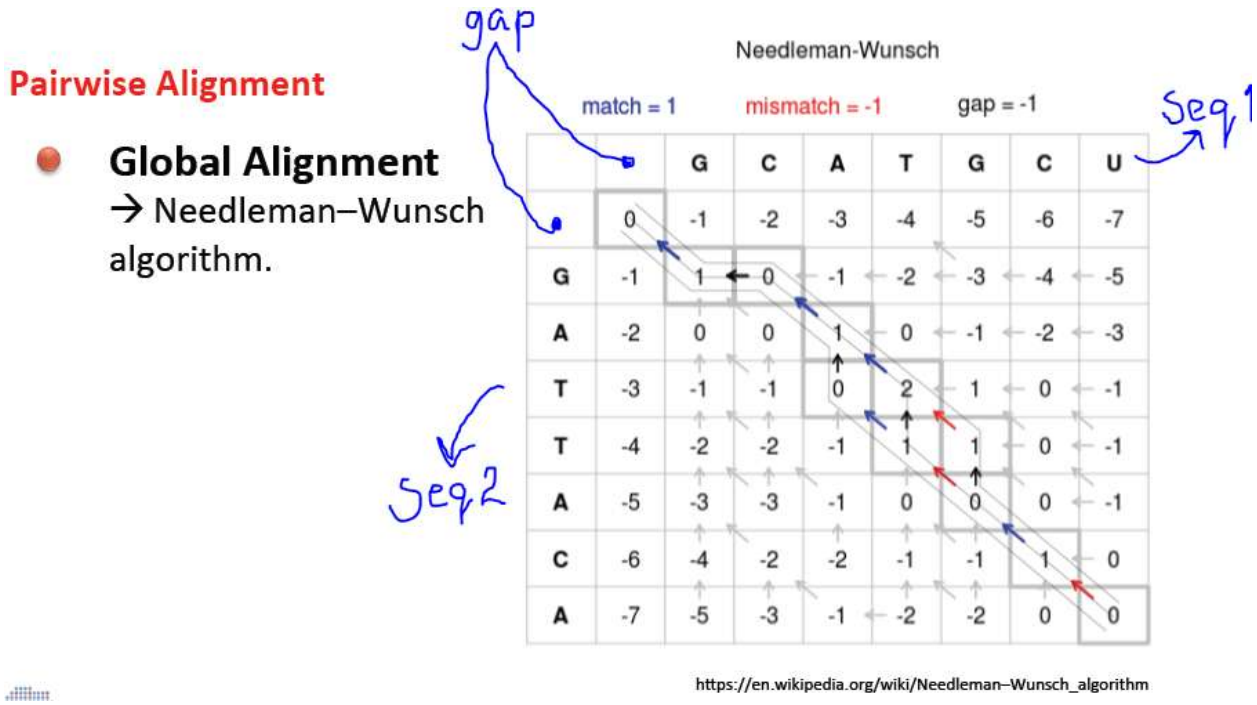
- **Dynamic Programming can**
  - Use a scoring matrix
  - Assign a match score (+), a mismatch score (-), and a gap penalty (-).
  - Use two different gap penalties for opening a gap and for extending a gap (gap opening >>> gap extension)

    → generally results in less gaps in an alignment and gaps are grouped together = more biological relevance

e.g. According to what you will like to highlight in the sequence alignment it is possible to assign different gap penalty for opening a gap and another for extending a gap if we take a **-10 for gap opening** and **-2 for gap extension**, then the number of gaps in the alignment is usually reduced in the final alignment. The residue and the gap are kept together and this is biologically more relevant, **therefore to achieve a biological relevance you will use a higher score for opening the gap.**

Different algorithms for Global and Local Alignment

- Dynamic Programming
  - Global Alignment
    - Needleman-Wunsch Algorithm
  - Local Alignment
    - Smith Waterman Algorithm



Alignment of two sequences 1 and 2

Assign **score to different position**, and **backtrack the alignment** to take into account the event that would have occurred and help you take the **most parsimonious** or **more likely events** that have resulted in the sequence been aligned this way.

Pairwise Alignment

- **Local Alignment**
  → Smith-Waterman algorithm.

$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 1 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 1 & 3 & 2 & 3 & 2 & 3 & 2 & 2 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 4 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 6 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 8 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 10 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{pmatrix}$$

$$T = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ G & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow & \uparrow \\ C & 0 & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow \end{pmatrix}$$

https://en.wikipedia.org/wiki/Smith–Waterman_algorithm

Determine similar regions between two strings of nucleotides. Instead of looking at the total sequence this algorithm is comparing segments of all possible length. **NB: Here we don't have negative scoring**. We then back track following the blue arrow or blue numbers until a cell with 0 score is reached, this yields a highest local alignment.

Pairwise Alignment

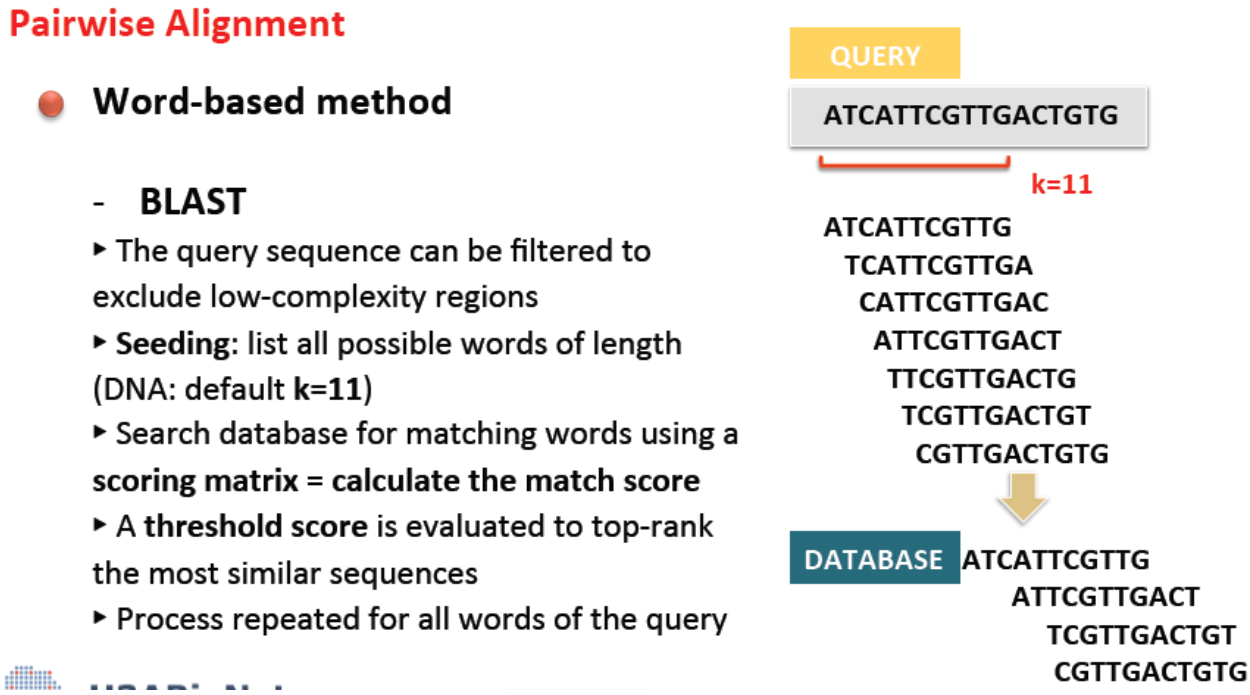
❖ **Word-Based Method**

- Optimal alignment not guaranteed, but efficient and faster than dynamic programming
- Useful for databases searches
- << words >> are small portions (length k) of the query sequence that are used to screen the database

➔ **Ex: BLAST** (heuristic approach i.e. ~~any method of problem solving that will **not guarantee optimal and perfect alignment** but that will be sufficient for your goal~~. The **idea** is to **cut the sequence** into **small portion** of **words** of **length k,** then use it to screen the database. ~~Useful when you have large scale database, you will get some bunch with not specific match~~.

- **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool)
  - Algorithm to compare a query sequence to a library or database of sequences
  - Allows to estimate identity with a certain confidence threshold
  - Popular in the scientific community (time efficiency…)
  - Useful when you want to know what species or organisms have a particular genes you are interested in other than the one you are extracting the sequence from
- **WORD-BASED METHOD**
  - **BLAST**
  - The query sequence can be filtered to exclude low-complexity regions
  - **Seeding:** list all possible words of length (DNA: default = **k 11**)

- Search database for matching words using a **scoring matrix = calculate the match score**
- A **threshold score** is evaluated to top-rank the most similar sequences
- Process repeated for all words of the query



Seeding all possible words of the length k by default is 11 as shown in the example above
Then **search** the **database for matching words** using a **scoring matrix**
Then **evaluate** the **threshold score** in order **to top rank the most similar sequences**
The **process** will be **repeated for all words of the query**
Best match is followed by an **extension in both directions**, **with scoring**
Extension continued only if the alignment is above the threshold
The contiguous alignment without gaps (now possible) and a higher score is the **HSP** (High Scoring Segment Pair)
Objective: From the HSP you can finally ascertain the Confident Align Sequence to your query

- BLAST

- A scoring matrix is used to evaluate the quality of the alignment
- A scoring matrix is a predefined substitution matrix (match =1, mismatch =0…) predefined you will decide the scores for match, mismatch or gap
- Ex: BLOSUM
- **Aim:** These allows you to judge the quality of your alignment

THE OUTPUT: The expected output
BLAST output

On the web page for BLAST, you can query using the **FASTA sequence** or enter the **Accession Number e.g. gi|343488507**, select the database, click and modify the algorithm parameter (check the bottom left) **for words size, gap costs type, match and mismatch values** etc.

Result Output

A list of sequences that have the best match to the query

It has different variety of format e.g. HTML, Plain Text, XML, (HTML by default for NCBI)

The 1st section is list of sequence with best match

The 2nd information is given in a table with matching sequence and several kind of matrix where you will find the **e-value**, **which is the probability that the alignment is found by chance. It takes into account the number of nucleotide in the reference, query and probability the alignment will be made by chance. NB: the lower the e-value the more interesting is the match.**

The 3rd section shows alignment details, the query and subject in the databases, the number of gaps etc

BLAST have different variant queries according to the type of **Query Sequence** (**Q**) and type of **sequence in the database** (**R**):

	Q		
BLASTN	Nucleic Acid	→	Nucleic Acid
BLASTX	Translated Nucleic Acid	→	Protein
TBLASTX	Translated Nucleic Acid	→	Translated Nucleic Acid
TBLASTN	Protein	→	Translated Nucleic Acid
BLASTP	Protein	→	Protein

MSA = MULTIPLE SEQUENCE ALIGNMENT

- Multiple Sequence Alignment have been developed to handle more than 2 sequences at a time
- Align all queried sequences to form a query group
- Allows to identify **conserved sequences** portions among a group of queried sequences that are:
    - Known to be evolutionary related
    - Of unknown/supposed evolutionary relationship → this multiple alignment helps to establish their relationships (phylogenetic trees)
    - **homology** that can be inferred

Multiple Sequence Alignment

- **Different methods:**
    - Dynamic programming
    - Progressive method
    - Iterative method: HMMs (Hidden Markov Models)

    → **Evaluation of Conservation across sequences**

❖ **Dynamic Programming**

- Optimized for 2 sequences, so computationally expensive here

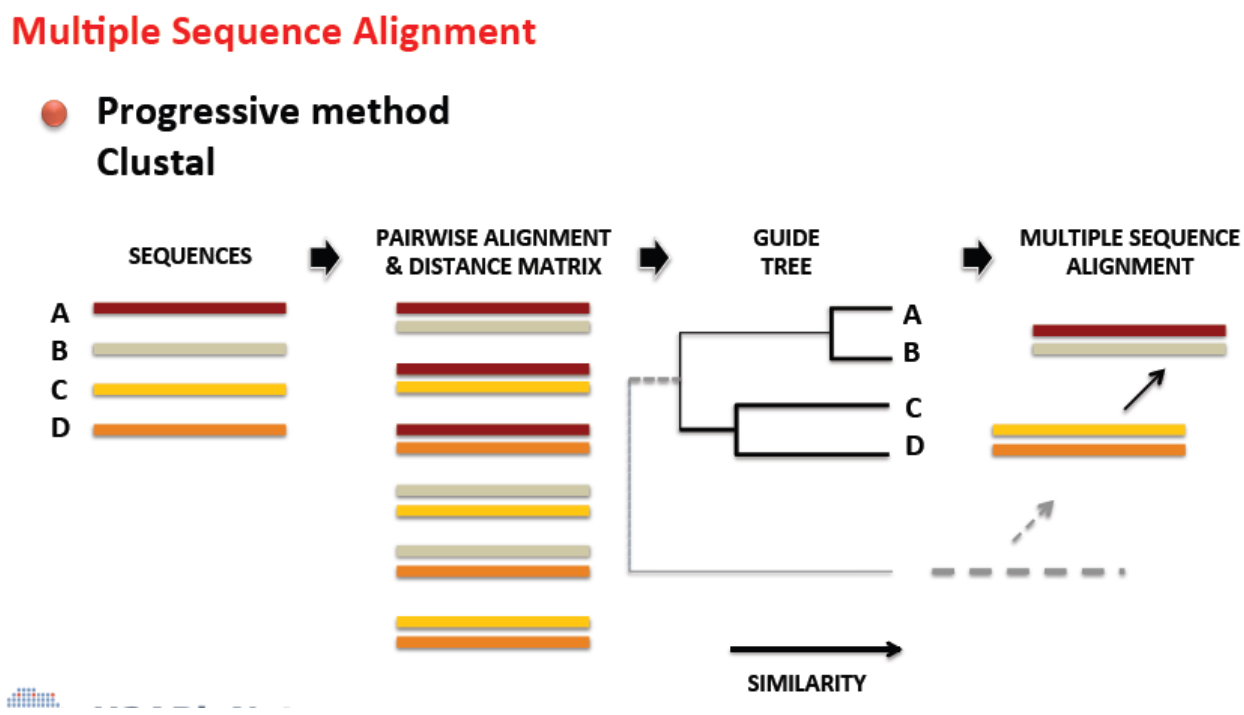
- Extends the sequence matrix from 2 sequences to the number of sequences in the query: basically alignment between pairs of sequences (constructing **n times** the sequence matrix formed from each two sequences until the whole number of sequences in the query is analyzed.
  → MSA

❖ **Progressive Method**

- Aligns sequences to identify the most similar ones
- Then progressively adds all other related sequences of the group

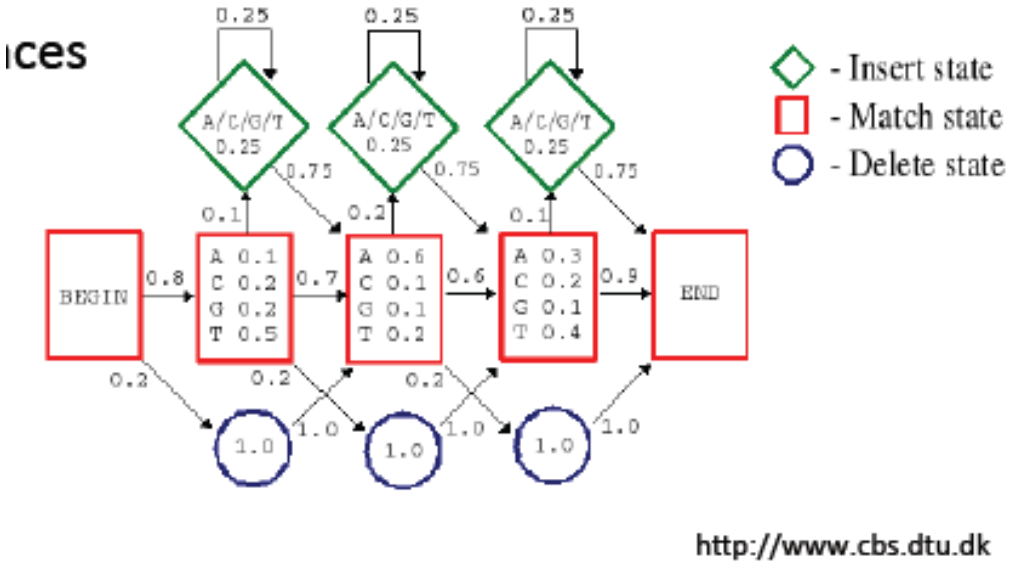
→ Clustal (clustal-Omega: medium-large alignments)
→ T-Coffee (small alignments)

Clustal basically **takes the sequence you have as input**, and **compute all possible pairwise alignment** and **distance matrix**, then it calculate **similarity score** using the **distance matrix**, the **score** is used to build a **GUIDE TREE.** This guide tree will be used to progressively align all the sequences you have in the query, this continues for any additional data or sequence you have until you finish all you have in your query. You add gaps during the alignment in order to optimize the sequence alignment after constructing the guide tree.



❖ **HMMs**

- Determine Probability scores for multiple sequence alignments
- The aligned sequences serves as a group, no need to previously order the sequences
- Can build an HMM profile
- Improved for more distant sequences



Is a type of MSA, it uses Vital-B Algorithm to determine the most likely path that could have produced the sequence, taking into account insert state, match state or deletion state. It is used to build an HTML profile. Widely used for protein as well as DNA, it is popular because it is really robust.

❖ **Conservation Scores**

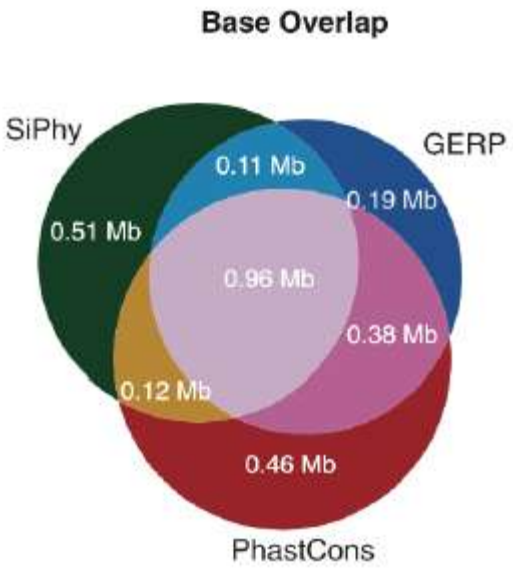
- Based on the fact that the highest conservation is maintained through evolution for the most important functions (promoters, essential enzymes, exons,…)
- Regulatory regions might be generally evolving << faster >>
- Multiple alignments → identify what **elements reject substitutions (substitutions occur in neutral DNA,** do not occur if an **element is functionally constrained).**

NB: The **higher the essentiality of a feature**, the **high probability it will be maintained** e.g. the case for **promoters essential enzymes, exons** etc whereas the **regulatory region evolve a bit more faster** e.g. **UTR regions** but we are talking about **millions of years of evolution here**.

NB: The **multiple alignments** will **help identify what elements would reject a substitution**, because **substitution occurs in neutral DNA** and **do not occur when an element** is **functionally constrained**. This evaluation is done through **conservation score.**

Conservation Scores … that is calculated by different tools, approaches and algorithm.
PhyloP (Phylogenetic p-values) Measures Base Conservation from non-coding regions
PhastCons (part of PHAST: Phylogenetic Analysis with Space/Time models) Measures Base Conservation based on HMM model
GERP (Genomic Evolutionary Rate Profiling) Measures Base Conservation to estimate the neutral evolution rate in genomes (**NB:** Good one is the GERP that measures base conservation through the mean number of substitution in each aligned genome, and this will help evaluate the neutral evolution rate) the higher the score the less variation there will be in the sequence.

SiPhy (Site-specific PHYlogenetic analysis) Models the pattern of substitution (based nucleotide substitution



The 3 methods in the chart are more-like complementary to each other.

WHOLE GENOME ALIGNMENT

- **MUMmer Ultra-fast alignment of large-scale DNA**
  http://mummer.sourceforge.net
  - Alignment of entire genomes (complete or draft)
  - **Maximal Unique Matcher:** Find the MUMs = subsequences that occur only once in both genomes compared and not extendable anymore
  - Suffix-tree based approach
- **WebACT Artemis Comparison Tool (ACT)**
  www.webact.org
  - Visualize the alignment of publically available prokaryotic genomes
- **Databases: Ensembl, VISTA**

Output of Comparative Genomics

Analyzing a genome structure means to analyze:

- **At the Genome Level:**
  - Base composition (%GC, codon bias, nucleotide distribution….)
  - Genome organization (Structural Variant (SV) events, genomic rearrangements…)
  - Sequence conservation (regulatory elements, repetitive regions…)
  - Synteny (conserved or not)
- **At the Gene Level**
  - Gene Order

Analyzing functions in a genome means to analyze:

- **At the non-coding sequence level**
  - Regulatory functions….
- **At the coding sequence level**

→Insights into functions

- Compare gene sequences
- Compare protein-cording portions

→How?

Different algorithms help identify portions of the genome coding for proteins

- *Ab initio* approaches
- Using homology…

In case of **Genomic Evolution**

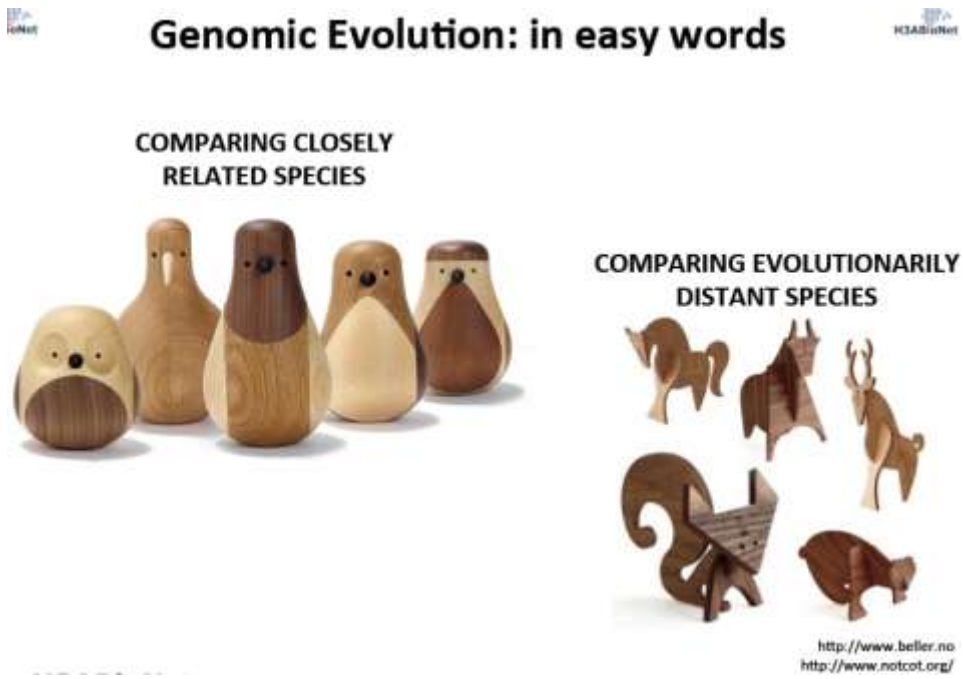
- From the multiple sequence alignments we can infer **homology** (Usman what is homology?) and estimate the evolution distance between sequences/organisms
- **Analysis could be:**
  - Based on entire Genome comparisons
  - Based on Gene comparisons

Based on entire genome comparison

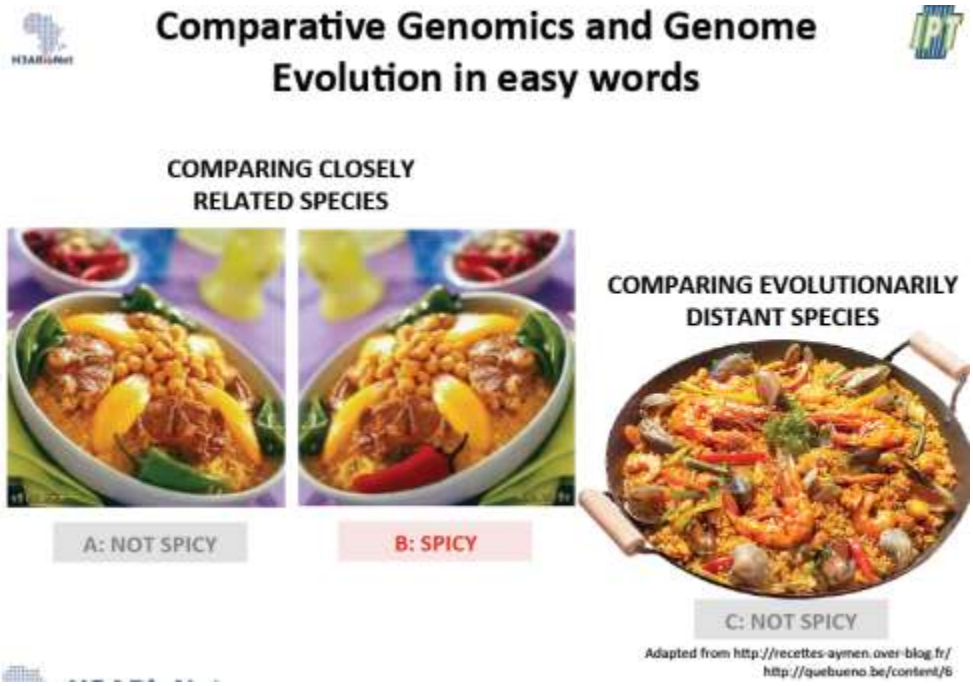
Phylogenetic relationship between organisms

Previous dogma: ***"Anything found to be true of E.coli must also be true of elephants" (Jacques Monod, 1954)*** ***NB:*** **This doesn't take into picture the whole phylogenetic context we know today, it will also depend on the expected output if you are comparing closely related species, or evolutionary distance species**

- Need to be related to its phylogenetic context
- Different outputs expected depending if:
  - Comparing closely related species
  - Comparing evolutionary distant species



Comparing closely related species what you will gain in terms of information (small changes and related to very peculiar characteristics of the genome) will be completely different from comparing distant species (mainly be seeing only small common parts of them) i.e. highlight only the differences.



You can easily say the only difference between the closely related species is the existence of the red pepper… Spicy (here you get small differences comparison)

If you are comparing the **A: Not Spicy** and **C: Not Spicy** dish we would be able to say it is just the pepper, there are so many differences between the dishes (here you get to compare different things and highlight one common features)

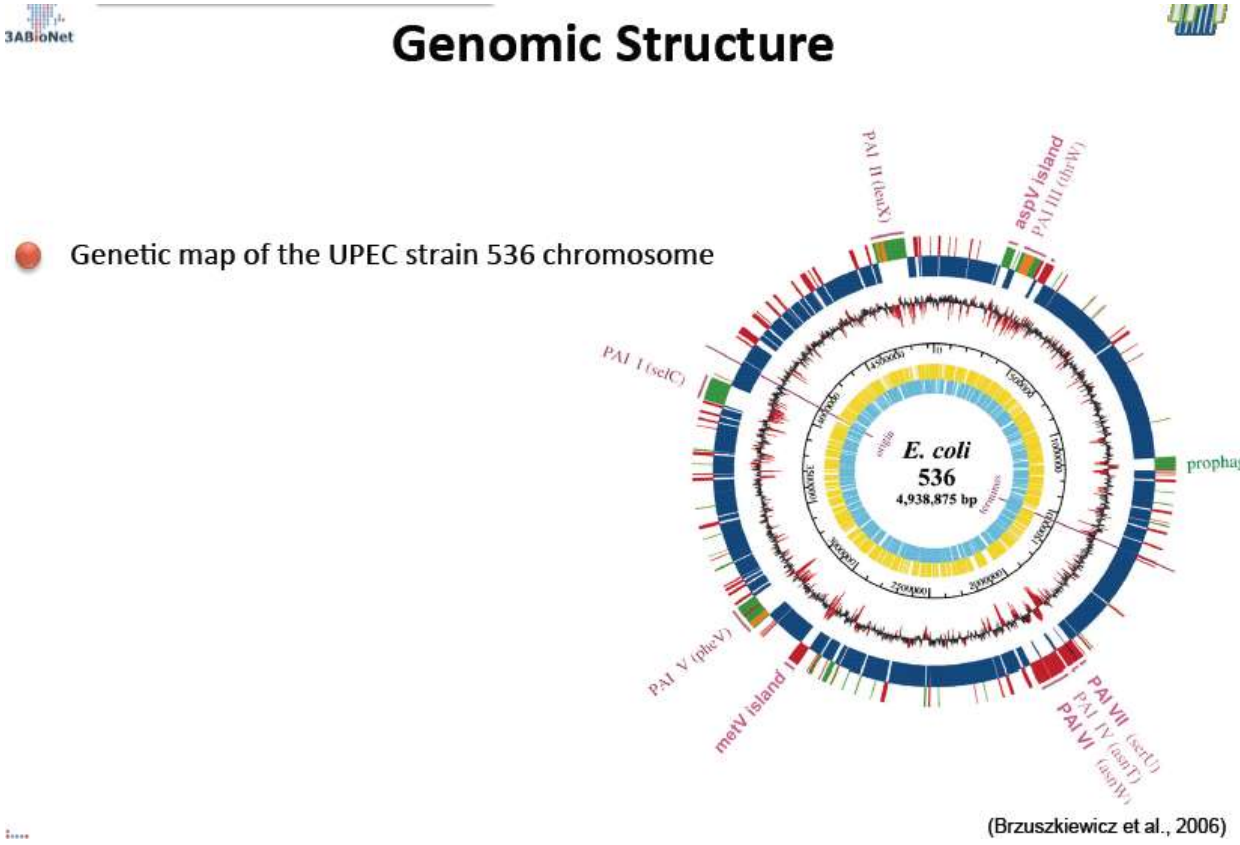
- Based on entire genome comparison

Examining the dynamics of closely related genomes helps to build therapeutic strategies for Ebola virus:

- Ebola virus largest outbreak (2014)
- Comparison of 100 available Ebola virus (Filiviridae) genomes to each other + to other viral genomes
- Floviridae are different from all other viral genomes
- Filovirus genomes: sequence diversity but proteins with similar functions and gene order

→ Ebola virus genomes very similar (sequence in general) but different in intergenic regions and genes of specific function = potential vaccine candidates.

Genomic Structure

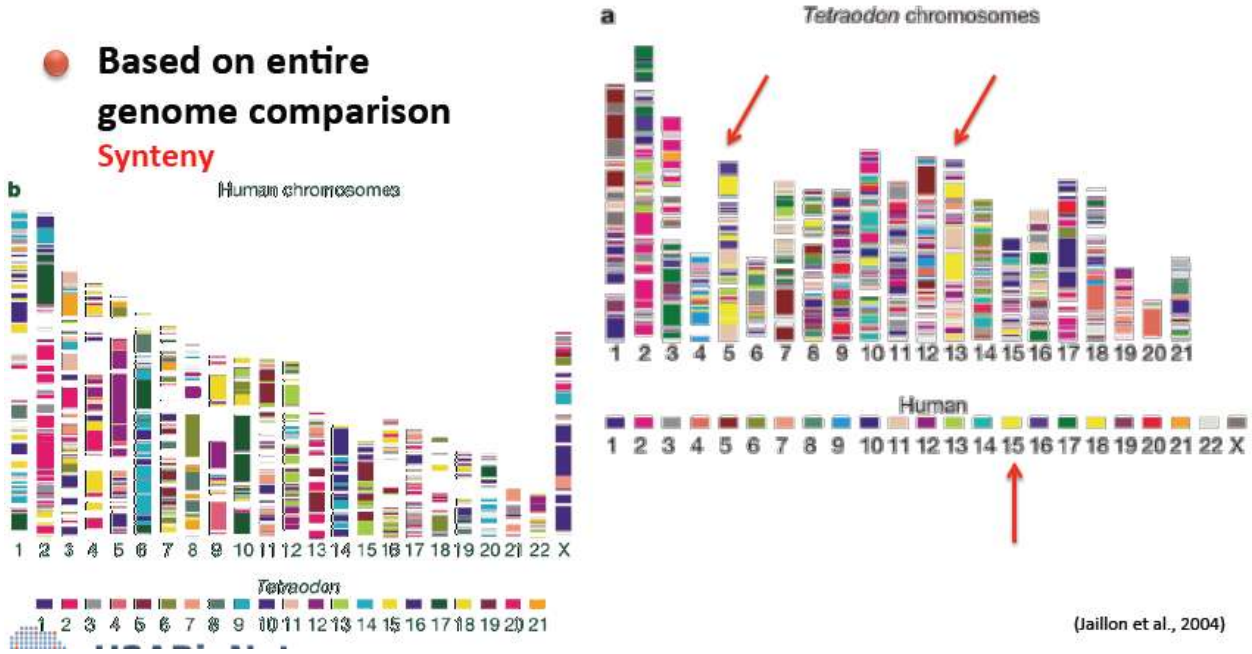


Based on entire genome comparison

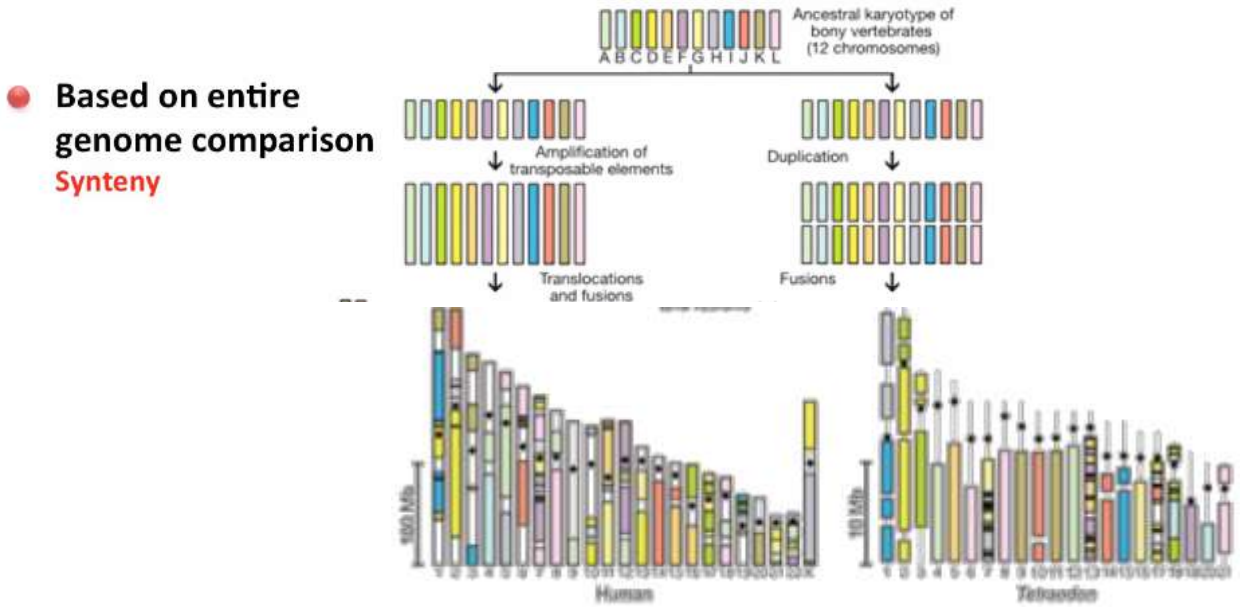
❖ **Synteny**

- Defined as the overall conservation of (gene/blocks) order in chromosomes between different genomes. Order maintained through…
- Evaluated in whole genomes, blocks could include large portions of genomes. Due to the fact…
- Recombination/crossing over affects groups of adjacent genes in a chromosome → **linkage group**

Example of Synteny in human genome



To observe synteny generally we color code the syntenic regions between two regions, and get the distribution of the colors of one color coded chromosome or an organism to chromosome of another organism. Using a fish Tetraodon chromosomes and human, the human genome is colored from 1 to 22 and X you will notice that chromosome 15 that is colored yellow, large block of this portion are maintained in chromosome 5 and 13 of that fish's chromosome. **NB:** So you can use the syteny blocks or linkage group in order to infer evolutionary relationship between two genomes or organisms.



Based on Individual gene Comparison

We can have…

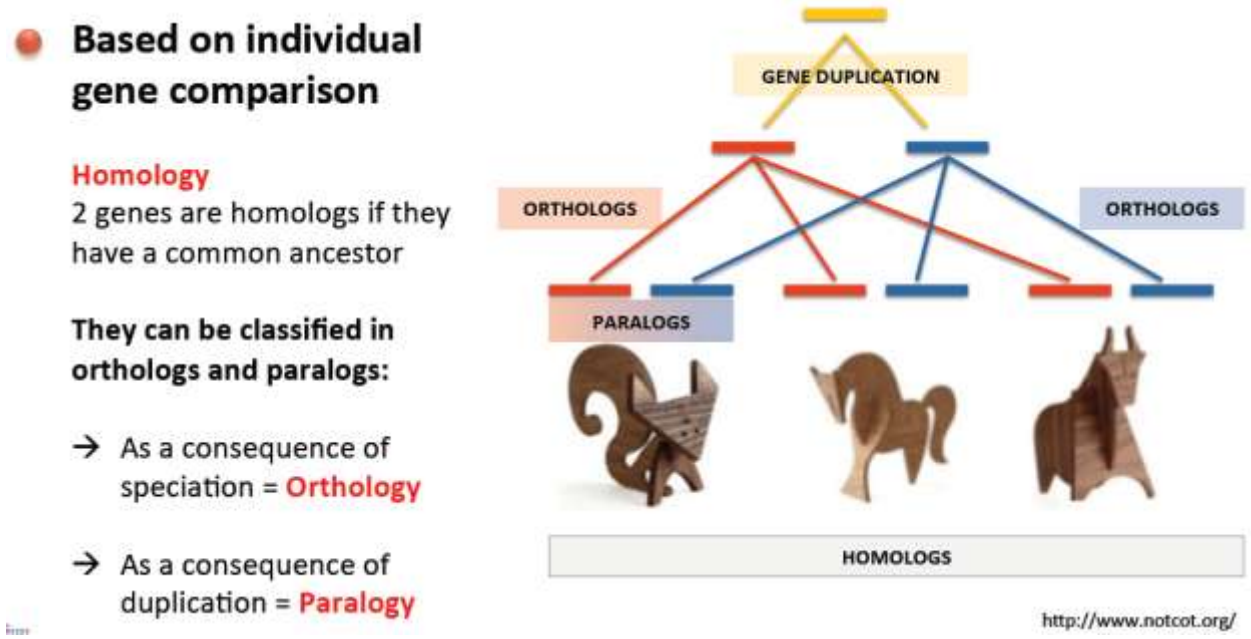
Homology between

*** 2 genes are homologs if they have a common ancestor**

They can be classified in **orthologs** and **paralogs**

As a consequence of speciation = **Orthology**

As a consequence of duplication = **Paralogy**



NB: Having common ancestor from the top "yellow" and gene duplication happens… the two genes at the 2nd level i.e. red and blue will follow two evolutionary path. Then if a sequence would originate or evolve as a consequence of speciation as genes in other species from Red to other reds at the bottom or from blue to other blues at the bottom, it is called Orthologs. Then if it is an event happening through duplication then it will be called Paralogs.

Orthology

Finding orthologs can be the first step in whole genome alignment

- BLAST Reciprocal Best Hit (best pairs of orthologs)
- OrthoMCL (possible predictions for several species)
- EnsemblCompara (precomputed data): orthology and paralogy predictions based on phylogenies
- eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups): OGs of proteins across different taxonomic levels

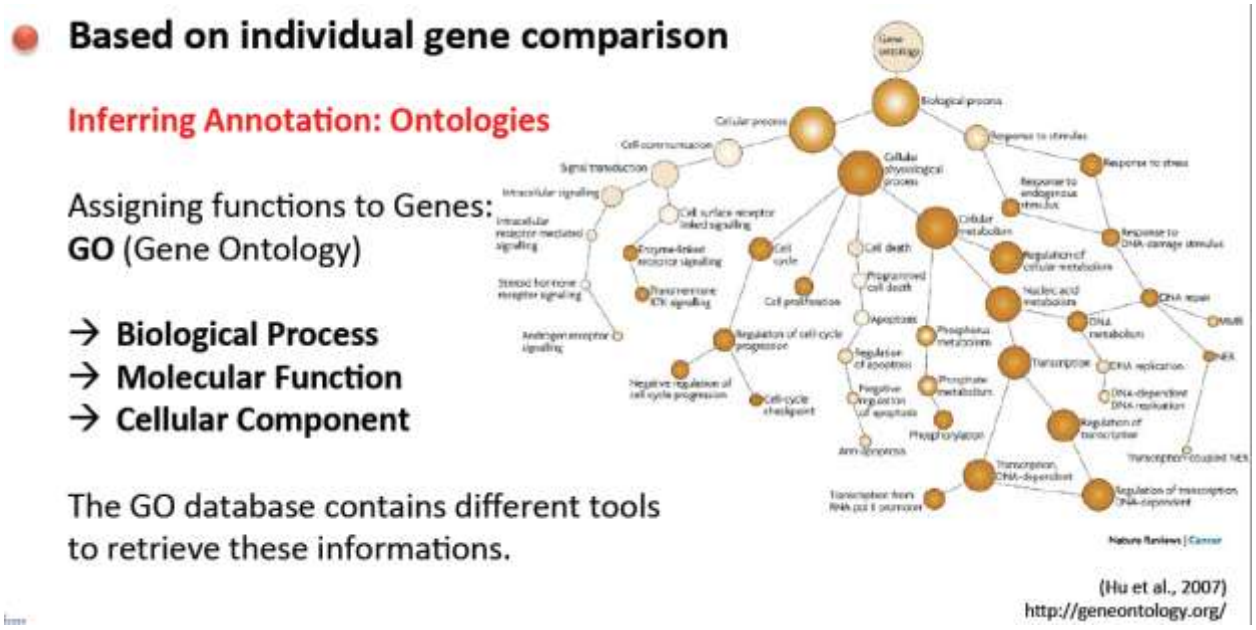
Inferring Annotation: Ontologies

Assigning functions to Genes: **GO** (Gene Ontology)

NB: Ontology is a way of categorizing a gene base on:

- **Biological Process** it is involved
- **Molecular Function** it has
- **Cellular Component** it is acting in

The GO Database contains different tools to retrieve this information shown as a tree below



Functional/Structural Predictions

Once you have predicted the function of a gene… you could have mu

❖ **Impact of a mutation on the function**

Analysis of the impact of **aa** substitution

→ Structural and/or Functional effect of single point mutations SNPs

- **PolyPhen-2** http://genetics.bwh.harvard.edu/pph2//...)
- **SIFT** http://sift.jcvi.org
- **VEP** http//www.ensembl.org/Homo_sapiens/Tools/VEP

❖ **Impact of the mutation on the Structure**

Impact on gene coding portions (gain/loss) or non-coding portions

- **VISTA**
  http://genome.lbl.gov/vista/index.shtml
  Collection of resources for comparative genomics
- **VISTA browsers** can be used to analyze pre-computer alignments or user generated or queried sequences
- VISTA servers
  - **mVISTA** (query sequences vs multi-species sequences)
  - **rVISTA** (identification or regulatory TF binding sites)
  - **gVISTA**(query sequences vs whole-genome assemblies)
  - **wgVISTA**(alignment of 10MB sequences (finished/draft): microbes…)
- **VISTA** tools: Reviewer region viewer to compare genomic intervals

Another Browsers include…

Ensembl Browser

http://www.ensembl.org

- Comparative analyses at the genome and gene levels
- Genome sequences compared using pair wise and multiple whole-genome alignments
- These alignments help to determine
  - Synteny
  - Sequence conservation scores
  - Gene homology relationships (Gene Trees)

Comparative Genomics

NB: Take Home Messages

- **Input / Output**
  - DNA Sequences (whole genome, or specific gene…)
  - Homology, similarity, evolutionary distance
- **Alignment** (Resources and Algorithms)
  - Whole genome: MUMmer…
  - Multiple genomes: MGA…
  - Multiple Sequence Alignment: Clustal…
  - Global/Local Sequence Alignment: BLAST…
- **Input / Output files**
  - Fasta/GenBank to alignment or Phylogenetic distances

MODULE 2 SESSION 4

Introduction to human variation

HUMAN VARIATION

Introduc0on to Bioinforma0cs online course: IBT_2016

Genomics | Dr Colleen J. Saunders

- Human genomes are ~99,5% similar across all individuals
- Varia6ons arise through muta6on
- Maintained by natural selec6on or neutrality
- Range from large karyotype differences to single bp changes
    - Karyotype = no. & appearance of chromosomes in nucleus
- Human Genome = 3,2 billion base pairs
    - Variants occur ~0,3-1 kb
    - 5-10 million variants within individual genome compared to another

Human genomes are very similar across all individuals

About 1 to .5 difference we see across each genome is due to variation that arise through mutation

This mutation established in a population by natural selection

Can range to quite large karyotype differences (diffs of number and chromosome within a nucleus) right down to vary single base pair difference in a sequence

The human consist of about 3,2billion base pairs

Between individuals there are 5 to 10million variance btw 1 individual and another when you look at the genome

Database for Human Genome Project

To sequence human genome

UK 10K	www.uk10k.org
Genome 10K	www.genome10k.soe.ucsc.edu/
1000 Genomes	www.1000genomes.org
Human Genome Project	www.genome.gov
Genomics England	www.genomicsengland.co.uk

HUMAN VARIATION

Most common

Single Nucleo0de Polymorphisms (SNPs)

- Most common form of varia6on
- Substitution of 1 single nucleotide for another (A,C,T,G)

Insertions and Deletions (INDELs)

- Small indels of 1-2bp
- Repetition of nucleotide patterns = Variable Number Tandem Repeats (VNTRs)

- Minisatellites (10-100bp)
- Microsatellites / Simple Tandem Repeats (2-6pb)

Structure Variation…

Copy Number Variations (CNVs)

- Deletion or duplication of larger regions of chromosome
- Gene dosage

Single Nucleotide Polymorphisms

- Single base change occurring at a frequency of >1% in 1 population
  - <1% = mutations / rare SNPs
  - Be careful – disease causing mutations may occur at higher frequencies!
- SNP's occur less frequently in exons
  - 50% exonic SNPs are non-synonymous
- Many SNP's do not confer any functional change ("silent")
- Others may affect
  - Protein AA sequence
  - Regulation/expression – mRNA stability
  - Phenotype by changing fractions

CODING SNPs:

Occur in coding regions of the gene (exons)

SYNONYMOUS SNPS: (Similar or the same) Result in No change in amino acid and may alter mRNA stability

NON-SYNONYMOUS/MISSENSE SNPS: Changes the amino acid sequences (codon) of the protein

NONSENSE SNPS: result in insertion of a stop codon which is not normally seen

INDELS SNPS: disrupt codon sequence (e.g. Shift or Frame Shift will disrupt how message are read in the RNA)

IUPAC Ambiguity codes for SNPS

IUPAC Code	Meaning	Complement
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H

B	C or G or T	V
N	G or A or T or C	N

IUPAC code are found in the sequence e.g. FASTA sequence e.g. seeing M means you probably have a SNP of A or C, N means G or A or T or C.

NON-CODING SNPs:

- Occur in regulatory/intronic/intergenic regions
- Many are relent
    - In terms of our current knowledge!
- May alter transcription
    - Located in promoter, silencer or enhancer regions
    - Transcription factor binding sites
- May alter mRNA stability and folding and thereby affect expression of protein

WORKING WITH VARIANT DATA

Two main ways include:…

1. HYPOTHESIS DRIVEN variant prioritization:

- Candidate gene association study
- "Candidate" variant of interest
- Small(er) numbers of variants

2. HYPOTHESIS FREE variant prioritization:

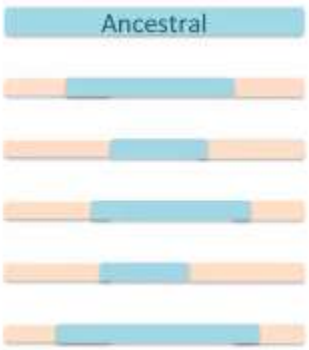
- NGS or GWAS studies (large number of variants)
- Investigate large numbers of variants

LINKAGE DISEQUILIBRIUM AND HAPLOTYPES

Population Genetics

- Within a population… Random mating & recombination should ensure mutations spread in the population
- Recombination events generate new arrangements for ancestral alleles
- Alleles at neighboring loci tend to co-segregate (tend to be inherited in same combination)
    - May affect ancestral combinations (haplotypes)

Given that there is an Ancestral Chromosome, within set of e.g. set of 5 individuals of a population we might see there are sequences or stretches of DNA with same combination of alleles across a different number of loci are inherited from the ancestral chromosome, what it means is that there isn't a lot of recombination happening within the stretch of DNA, so this combination are inherited together in blocks… because of a phenomenon known as: **Linkage Disequilibrium (LD)** = non-random association of alleles at different loci



Be careful to distinguish between LINKAGE and LD!

Linkage: is focused on 1 particular locus and recombination in last 2-3 generations (used in disease mapping studies, across or through a family)

Linkage Disequilibrium LD: is focused on particular alleles at a locus and recombination over a much longer period of time (many more generation than what is seen in one family)

MEASURING LD

There are 3 different measures:

1. ***D*** Measures the difference between the frequency at which alleles at different loci are inherited together, and the frequency at which we expect to observe those alleles together if they (aren't in LD meaning… if they) are in equilibrium
   - If D is significantly > 0, the loci are in LD
   - D increases with tighter linkage between the loci
2. ***D'*** is the absolute ratio of ***D*** and its maximum value, $\mathbf{D_{max}}$
   - D' = 1 indicates complete LD
   - Recombination over time causes the decay of D' towards 0
3. r^2 is the measurement of correlation between a pair of loci

Linkage Disequilibrium

- Varies throughout the human genome
  - Regions of high LD interspersed with regions of very low LD (the **low LD region** has lot of Recombination, so we call it **Recombination hotspot**)
- Understanding LD structure results in cost saving for association studies
  - Tight LD means known which allele occurs at 1 locus can tell us which allele occurs at the other locus
  - Basis of GWAS (Genome Wide Association Studies)
- High D' means that variants are good surrogates for each other (knowing there locus)
- D' estimates may be increased in small samples and if an allele at 1 locus is very rare

NB: Be careful of interpreting D' with LD

HAPLOTYPES & TAG SNPS

HAPLOTYPES: Blocks of sequence along a chromosome where no recombination occurs (**i.e. region of high LD**)

- Blocks of closely linked alleles that are inherited together
- All pairs of SNPs **within 1 haplotype block are in high LD**

TAG SNP: SNP that is representative of other SNPs in a haplotype block

- Can be used to infer the allele present at other loci within block
-

International HapMap Project

• catalogue common patterns of genetic variation in humans
• map of haplotype blocks and tag SNPs that identify the haplotypes
• No. of SNPs required to examine en6re genome (~10 million SNPs) to ~500 000 tag SNPs
    • Easier & cost-effec6ve to find disease associated genes & regions
• Common haplotypes occur in all popula6ons but at different frequencies
    • 270 individuals from CEPH, Han-Chinese, Japanese & Yoruba populations genotyped for 6 million SNPs
• Beware: by focusing on common variants, may miss rare disease associated variants

NB: Catalog Haplotypes that occur in frequency of higher than 5% in a population, are relatively common haplotypes and by focusing on this common variants we might sometimes miss rare diseases associated variance.

TO find TAG SNPs use the tool…

Tagger: tool for selection and evaluation of tag SNPs from genotype data
www.broadinstitute.org/mpg/tagger/
Tagger Server www.broadinstitute.org/mpg/tagger/server.html
You can find by **genomic region** or within you **genotype data**

Haploview

- Tool designed for haploytpe analysis
- LD and haplotype block analysis
- Estimate haplotype frequency in a population
- SNP and haplotype association tests
- Tagger tag SNP Selection
- Download phased genotype data from HapMap

VARIANT CALL FORMAT file

.vcf

The Variant Call Format Specification VCFv4.3 & BCFv2.2

- Text tile containing sequence variation data
- Meta-information lines (preceded by ##)
- Header line
- Variation data for a particular position in rows
- May include genotype information for samples
- BCF = binary compressed for large VCF (BCF is Not text like VCF)

Each variants information is started in a new row

META-INFORMATION:

• Starts with ##
• Key=value pairs
• Fileformat line is always required (line 1)
    • ##fileformat=VCFv4.3
• Other meta-data is optional (highly recommended!)
• Describe the information contained in the file
    • E.g. "FILTER" will describe quality filters applied to the data

HEADER LINE:

• Always contains the same fields in position 1-8:
    • CHROM
    • POS
    • ID
    • REF
    • ALT
    • QUAL
    • FILTER
    • INFO
    • FORMAT (If genotypes given)
    • SAMPLE ID's (If genotypes given)

NB: It will always be in that other

VARIANT INFORMATION:

• rows indicate variant information for chr. position
• **Missing data is indicated with a "."**
• **CHROM** = chromosome in reference genome (where the variant is found)
• **POS** = position of the variant in the reference genome
• **ID** = variant identifier (preferable dbSNP ID)
• **REF** = Reference base/s (allele)
• **ALT** – Alternate base/s (allele). Can be multiple.

• **QUAL** = Phred-scaled quality score for the variant call
• **FILTER** = Pass indicates this call has passed all filters

Different annotation might indicate different reasons why the variant hasn't passed the quality filters, those annotation will be described in the meta data at the top of the file

VARIANT INFORMATION:

• INFO = Additional informa6on
    • Multiple fields separated by ";"
    • Sub-fields listed in meta-data
    • AA = Ancestral allele
    • AC = Allele count in genotypes
    • AD = Read depths for each allele
    • BQ = base quality at this position
    • DP = combined depth across samples
    • Etc…

NB: In order to understand what that information is you need to refer to the Meta Data lines to see for example what AA or AC stands for.

VARIANT INFORMATION:

• If genotype data is reported:
• FORMAT = specifying type & order of genotype data for each sample
• Sub-fields separated by ":"
• GT = Genotype (| = phased; / = unphased)
• AD = per sample read depth for each allele
• DP = per sample read depth at this posi6on
• MQ = RMS mapping quality
• Followed by 1 data block per sample (**SAMPLE ID's**)

NB: Explanation of each sub-file will also be explained in the meta data

VARIANT PRIORITIZATION

NGS pipelines generate large .vcf files

– WGS experiment yields ~1-1,5 million variants per sample

– WES ~50 000

• How do we filter these to iden6fy those most likely to affect protein func6on or expression?

• How can those variants be further filtered to identify the one(s) likely to cause this disease that are good candidates for further investigation

Questions a biologist/clinician might ask?

• What is the frequency of the variant in the general popula6on?
  This specific population?

• What part of the gene is it in?

• Does it affect gene function?

• Is it in a gene known to be involved in the disease? A related disease/phenotype?

• in a genome region statistically implicated in the disease?

• involved with a func6on/pathway that coincides with the disease pathology?

• Etc …

In variant prioritization you look at…

Remove common variants	Variant level
Variants that change the amino acid	Variant level
Variants that have a functional effect	Variant level
SNPs in biologically plausible candidate genes	Gene Level

Step 1: Frequency Information using dbSNP maintained by NCBI

Summary of allele frequence across many different datasets including 1000genomes, HAPMAP, Human Genome Project, ExAC and ESP6500.

ExAC and ESP6500 both very good sources for earlier sequencing variance and have exomes of unrelated individuals where…

* The ESP6500 contains 6500 exomes from unrelated individuals and focus is heart, lung and blood disorders

* ExAC has 60706 exomes from unrelated individuals

Step 2: Genomic Context

Does it result in amino acid change?

dbSNP is a very good resource for this using the RsID for variance it will give you the Types and Allele information, sequence change information, Gene region function information etc

Ensembl will also help using the RsID

Step 3: Function Prediction

Different algorithm

Using **SIFT** to predict the function effect a variant might have (deleterious or tolerated)

- Coding variants only
- Variant with score <0.05 is predicted as deleterious

PolyPhen-2

Coding for non-synonymous SNPs only

Has two different algorithms:

- HVAR – for diagnostics of Mendelian disease
- HDIV – used when evaluating alleles in complex phenotypes

FATHMM

- Both Separate coding or non-coding algorithm score (for non-coding variants)
- Damaging (D) or tolerated (T)

RegulomeDB

- Non-coding variants
- Identifies DNA features and regulatory elements such as trancriptoon factor binding sites

SNPedia

- Is a very good source for predicting functional effects
- Wiki that investigate human genetics
- Variants centered information
- Links to peer-reviewed publications
- Cross-referenced to other databases

Step 3: Clinical Consequences

ClinVar (Looking At Clinical Consequences, if the variant has ever been implicated to other diseases)

- Links genomic variation to human health phenotype
- Levels of supporting evidence vary but indicate
- Content is not always curated

Cosmic (This are all curated)

- Catalogue of Somatic Mutations in Cancer
- Curated input from peer reviewed publications
- Genome wide screen data

Pharmacogenomics

- Search by gene or variant (gives information on data support on pharmacology)

Step 4: Gene Prioritization

Two Databases (**OMIM and dbGap in NCBI**)

- Gene centered information
- Relationships between genotype and phenotype

For MicroRNA genes

miRBase is a good place to start

- Searchable database of published miRNA sequences and annotation
- Includes link-outs to databases predicting gene target for the miRNA

KEGG Pathway Database

Manually curated pathway aps representing molecular interaction and reaction networks
Explore by pathways or by gene

Another way to prioritize gene is…

Gene knockout models
The Mouse Genome Database (MGD) and the Rat Genome Database (RGD)
Very often provide the missing link that solves 'strange' genetic disease cases

PART 1

NGS
What is DNA Sequencing
DNA sequencing technologies
Sequencing quality
Finished genome vs draft genome
DNA Sequencing method from Sanger to NGS
DNA-seq
NGS
Principles of DNA Sequencing: Pyrosequencing
Solexa
454 Technology
SOLID
Ion Torrent
PacBio
Nanopore
The four main advantages of NGS over classical Sanger sequencing: Speed, Cost, Sample Size, and Accuracy
DNA-Seq Protocol for Library Construction
Kits for DNAseq Library Prep
DNA-Seq Analysis Pipeline and Associated files

PART 2

Genome Browsing and Annotation
Genome browser: browse, search, retrieve and analyse
Genome browser multi species and species specific
Genome annotation: Structural annotation and Functional annotation
Genome annotation pipeline (genome assembly, annotation phase, computation phase, synthesis phase, manual curation
Genome assembly:

Overlapping reads assembled to form continuous consensus sequences = contigs
Contigs are linked together using sequencing read mate pairs = scaffolds

Gaps
Coverage
Feature alignment blast
Ab initio gene prediction
Evidence driven gene prediction
Manual curation
Annotation output

Challenges:

- NGS technology = shorter read lengths
- Assembly and annotation is difficult when no gene models exist
- Updating annotated data sets
- High quality, accurate annotation vs high throughput

Ensembl genome browser
Biomart

MODULE 3 Session 1

INTRODUCTION TO LINUX AND UNIX AND THE COMMAND LINE

- **Introduction to UNIX and Linux**
- **The UNIX files structure**
- **Creating directories and navigating through the file structure**
- **Some useful shortcuts and links**

What is UNIX

- **UNIX is an Operating System (OS) initially developed in the 1960**
- **There are many different versions of UNIX, that share common similarities**
- **The most popular varieties of UNIX are Solaris, Linux and MacOS**
- **UNIX systems have a graphical user interface (GUI) making easier the environment**

Why Linux

- **Linux is free and the most popular distributions are Ubuntu, Fedora/Red Hat, Mandriva etc**
- **Very stable system**
- **Fast developing OS (many developers)**
- **Most secure OS**
- **Best multi-user and multi tasking OS**
    - **Very popular as servers OS**
- **The world's fastest super computers run Linux**

Different Linux distributions are available at: http://distrowatch.com

NB: Ubuntu distro is easy and convenient to use for beginners

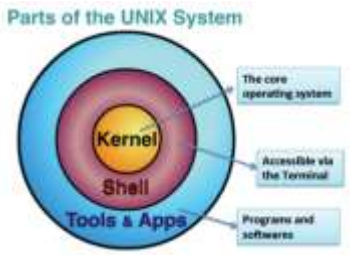
Questions that come over time…

LINUX VS UNIX

- **Linux is the open source clone of UNIX**
- **Because Linux is freely available it has a cost advantage against UNIX**
- **All command lines work the same on both**

Parts of the UNIX System

- **The core operating system (Kernel)**
- **Accessible via the terminal (Shell)**
- **Program and software (Tools and Application) Layers with lots of programs**



The Terminal

The shell is accessible via the terminal

Shell prompt

- **Username**
- **Machine name**

Even though the shell is a command line interface, the mouse is still handy (scroll, copy, paste etc)

UNIX File Structure

Organised in hierarchical structure like an inverted tree the tree grows downward from the root, each folder can contain sub-folders or classical files (text, image, excel etc files)

Major directories

/ Root directory of the entire file system

home Directory that contain all user home directories

dev Device files for different devices

bin All user command binary

etc All configuration files

Home directory and Working Directory

- **When you first log in on a UNIX system, the working directory is your home directory**
- **While working you will be associated to one directory called the working or the current directory**
- **An abbreviation of the working directory is displayed as part of the prompt on your terminal**
- **The command pwd gives the absolute path to the working directory**

What is a path or a pathname?

- **A path locates a given file in the system hierarchy**
- **An absolute path in the file system hierarchy for a given file or folder describes the parents all the way up to the root e.g. for file1 /home/Username/IBT2016/Genomics/file1**
- **A relative path describes the path to the starting from the current working directory**

e.g. from IBT2016 as current working directory file1 relative path = ./Genomics/file1

Refer to the parent and current directories

Every directory has two special sub-directory

. (dot): the current directory

.. (dot dot): the parent directory

Commands for manipulating directories

Command	Acronym & Output
mkdir	Make Directory (Creates a new directory)
rmdir	Remove Directory (Removes Directory)
pwd	Print Working Directory (Displays the absolute path of the current working directory)
cd	Change Directory (Allows moving from one directory to another)
ls	LIStirng (Lists a directory content)

pwd

pwd: print working directory

Displays the absolute path of your current location in the file system

Ls command

Ls: lists the content of the current directory by default

Commond structure: ls [OPTION] [dirname]

Some useful options

-l: shows sizes, modified data and time, file or folder name and owner of file and permissions

-a: list all files including hidden file starting with '.'

-lh: shows sizes in easier readable format

-R: recursively lists sub-directories

-lS: sorting by file sizes NB: Note the lower case l and UPPERCASE 'S'

Example ls –l gives:

- **File Permissions**
- **Owner**
- **Group**
- **Size(In bytes)**
- **Modification Time**
- **File Name**

Create a Directory

- **mkdir: makes a directory**
- **Command structure: mkdir dirname [path]**
- **mkdir dirname: would create a directory with the specified dirname**
- **The new created directory will automatically be created in your current working directory**
- **If you want to create it elsewhere you have to specify the path: mkdir dirname [path]**

About file names in Linux

- **No real distinction between the names of ordinary files and the names of directory files**
- **No two files in the same directory can have the same name**
- **Files in different directories can have the same name**
- **Linux is case-sensitive: Sanger, sanger and SANGER are different and would represent three distinct files**
- **In most cases, file extensions are optional (.txt, .exe, etc…)**

Change Directory Moving in the files system

cd change the working directory

Command structure: cd <path> #You can specify an absolute path starting from the root or a relative path starting from your current working directory

Username contain IBT2015 and IBT2015 contains Linux and Genomics

If you are in Linux and you wish to move to Username directory type:

cd/IBT2015/Username

OR **cd ../..**
OR **cd .. + cd ..**

From Linux and you wish to move to Genomics

cd /home/Username/IBT2015/Genomics
OR **cd ../Genomics**
OR **cd .. + cd Genomics**

Remove a Directory

- **rmdir works if there is no contents in the directory**
- **If the directory contains files or sub-directories, an error message will appear "Directory not empty"**
- **There is an option that will recursively remove a directory and its contents:**
  - **–r (which stands for recursive)**

How to get help for a command from the terminal?

man commandname displays manual pages for that specific command
commandname –h **OR** **commandname –help**
e.g. man rmdir #It gives you help on how to use rmdir

Commands Basic Structure
command[-options] [arguments]
Example:

ls –lh /home/Username/IBT2016
pwd
mkdir test1

Useful Shortcuts

cd ~ **Takes you to your home directory**
cd **Alone (nothing following) will change your location from the working directory to your home directory**

LINUX, SOME USEFUL COMMAND LINE KEY COMBINATIONS

Ctrl + A: remove the cursor to the beginning of the command line
Ctrl + C: end a running program and return to the prompt
Ctrl + D: logout from the current shell session: equivalent to exit
Tab: **auto-complete a file name**
Tab Tab: **displays command completion possibilities**
Ctrl + L: **clear the terminal i.e. Clear Screen**

MODULE 3 SESSION2

LINUX MANIPULATING FILES

Learning Objective

- Creating new files and using text editors
- Manipulating files and directories

- Viewing files contents
- Manipulating and extracting data from files
- Output redirection to a file
- Output redirection: combining different commands

Learning Outcome

- Learn how to create new file(s)
- Learn how to use text editors
- Use basic command to manipulate directors and files (view file content)
- Manipulate and extract data from files
- Redirect the output of a command
- Combine different commands

BASIC MANIPULATING FILE COMMANDS

touch is used to create, change and modify timestamps of a file

touch simply creates an empty file

touch command creates an empty (zero byte) new file using this

Structure: **touch filename**

Create more than one single file

touch filename1 filename2 filename3

touch command options

-a: change the access time only

-c: if the file does not exist, do not create it

-d: update the access and modification times

-m: change the modification time only

-r: use the access and modification times of tile

-t: create a file using a specified time

NB: man touch to see all this options above

TEXT EDITORS

nano: a simple and easy-to-use text editor

Is installed by default in Ubuntu and many other Linux distros

It's a **WYSIWYG editor: "what you see is what you get".** What you type directly goes into the ext input

Examples of other text editors include:

vim, emacs, gedit, Geany: excellent programs but do require some learning of the shortcuts, they are still working like the **nano**.

E.g. nano filename1 Same applies to using **pico filename1** Opens the content of filename1 or creates filename1 if it doesn't exist.

HINT: nano pathname it open an existing if a file has been created AND create new file in the specified path if the file does not exists

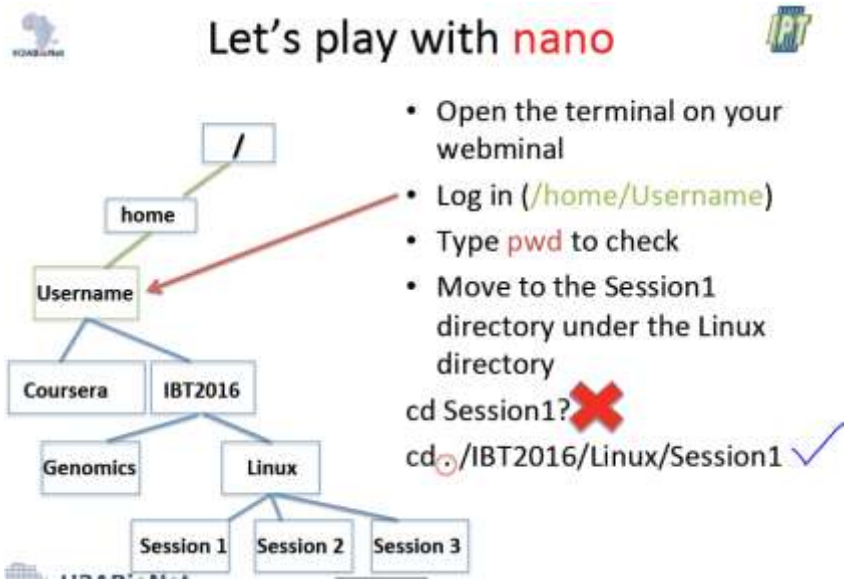
Getting started with nano

Nano window has:

Name of the program and **version number** at the header
The **filename** of the file you are editing also at the header
Shortcuts at the bottom

Ctrl + X to Exit
Ctrl + K to Cut
Ctrl + U to Paste
Ctrl + W to Search text string in a file
Ctrl + C to cancel mid-execution without destroying you buffer

Example: From Username move to Session1



Code: uslaad@ubuntu:~/Username$ **cd ./IBT2016/Linux/Session1**
OR uslaad@ubuntu:~/Username$ **cd ~/IBT2016/Linux/Session1**

Output: uslaad@ubuntu:~/Username/IBT2016/Linux/Session1$

Create file1 and file2 under the Session1 directory

uslaad@ubuntu:~/Username/IBT2016/Linux/Session1$ touch file1 file2

Check file content of file1

Code: nano file1

Type some text within the editor, hit enter to move to another line once you finish typing hit Ctrl + X
Save modified buffer answer yes i.e. Y
Ctrl + X to Exit the Nano Editor

Download files with wget

wget is a free utility for non-interactive download of files from the Web. It supports HTTP, HTTPS and FTP protocols.
Structure: wget <options> <fURL>

Creates a copy of the file in your machine in the specified path
Example: wget http://yyyy.org/files.fasta

Manipulating Files and Directories
Useful commands to use
Copy, Move and remove
NB: In linux file could be a directory, text or any other ordinary file
cp: copy files and directories
Structure: **cp <pathfrom> <path to>**
mv: move or rename files and directories
Structure: **mv <pathfrom> <path to>**
rm: remove files and directories
Structure: **rm pathname**

Copying command: cp
Simplest form: **cp file1 file2**
Copy the contents of file1 into file2. If file2 does not exit, it is created. Otherwise, file2 is silently overwritten with the contents of file1
cp filename dirpath
Make a copy of the file (or directory) into the specified destination directory

Other examples of cp:
Add the interactive mode with the option –i
cp –i file1 file2
Same as the previous one. However, if file2 exists, the user is notified before overwriting file2 with the content of file1
cp –R pathdir1 pathdir2
Copy the contents of the directory dir1. If directory dir2 does not exist, it is created. Otherwise, it creates a directory named dir1 within directory dir2

Copying Command: mv
Move command
The **mv** command moves or **renames** files and directories depending on how it is used
To rename a file: **mv filename1 filename2**
It automatically renames filename1 as filename2 if filename2 does not exists
If filename2 exists, its contents are silently replaced with the contents of filename1. To avoid overwriting, use the interactive mode:
To move a file (or a directory) to another directory: **mv file dirpath**
To move different files or many files (or a directory) to another directory: **mv file1 file2 file3 dirpath**
To move directory to another directory: **mv dir1 dir2**
If dir2 does not exist, then dir1 is renamed dir2. If dir2 exits, the directory dir1 is moved within directory dir2

The rm Command

The **rm** command deletes files and directories

To remove a file: **rm filename**

To remove many files **rm filename1 filename2**

Add the interactive mode to prompt user before deleting with –i **rm –I filename1 filename2**

Delete Directory or Remove directory with all their contents **rm –r dir1 dir2**

Example: Copy file1 from Session1 into Session2

Code: uslaad@ubuntu:~/Username/IBT2016/Linux/Session1$ cp file1 ../Session2

OR using the path for Session 1 and Session2

uslaad@ubuntu:~/Username/IBT2016/Linux/Session1$ cp /home/Username/IBT2016/Linux/Session1 /home/Username/IBT2016/Linux/Session3

NOTE: Be careful with rm!

- Linux does not have an undelete command
- Once you delete something with rm, it's gone!
- You can inflict terrific damage on your system with rm if you are not careful, particularly with wildcards
- Try this trick before using rm: construct your command using **ls** instead first

WILDCARDS

- Since the shell uses filenames so much, it provides special characters to help rapidly specifying groups if filenames
- A group of special characters are called wildcards allow selecting filenames based on pattern of characters

Wildcard	Meaning
*	Matches any characters
^	Matches any set of characters starting by the character following the sign
?	Matches any single character
[!characters]	Matches any character that is not a member of the set characters
[characters]	Matches any character that is a member of the set characters. The set of characters may also be expressing as a POSIX character class such as one of the following: **[:alnum:]** Alphanumeric characters **[:alpha:]** Alphabetic characters **[:digit:]** Numerals **[:upper:]** UPPERCASE alphabetic characters **[:lower:]** lowercase alphabetic characters

Wildcard	Meaning
a*	Any file name starting with **a**

*****	All possible filenames
^a	Any filename starting by **a**
A*.fasta	All filenames that begin with A and end with .fasta
????.vcf	Any filenames that contain exactly 4 characters and end with .vcf
[abc]*	Any filename that begins with "a" or "b" or "c" followed by any other characters
[[:upper:]]*	Any filename that begins with an UPPERCASE letter. This is an example of character class.

Viewing Files Contents

Basic Viewing file commands

Displaying whole content of a file or parts of it (default + options)

cat: view the content of a short file **Structure: cat <filename>**

cat("catenate") command displays the whole content of a text file

Structure: cat <options> filename

Useful for small file contents (<1 page), otherwise a continuous paging will be displaced until the file ends

For **<options>** see **man**

more: view the content of a long file, by portions **Structure: more <filename>**

you cannot scroll in more

more command displays a text file content, one pate at a time

Structure: more <options> <filename>

If the content is >1 page, it shows the percentage displayed and allows you to scroll down to view more (Enter: by lines, Space: by page…)

Will not allow you to go up

To **exit more**, type **q**

For **<options>** see **man**

less: view the last lines of a long file, **Structure: less <filename>**

less allows you to scroll up and down

less command displays a text file content, one screen at a time

Structure: less <options><filename>

Different from more: backward scrolling possible

Move a page down: **enter**, page **down key** or **space**

To **exit less** type **q**

To go to the end of the text file, type **g**

For **<options>** see **man**

Special Commands to basically view the beginning or end of a file

head: view the first lines of a long file, **Structure: head <filename>**

head command displays a text file content, by default: 10 first lines at a time

Structure: head <options> <filename>

Will **exit** automatically after the display

For **<options>** see **man**

Most common **<options>** is **–nX** where **X** is the the number of the **first lines** that you want to display

Code: uslaad@ubuntu:~/Username/IBT2016/Genomics$ head –n1 file1

Output: prints first line of file1 only

tail: view the last lines of a long file, **Structure: tail <filename>**

tail command displays a text file content, by default: 10 last lines at a time

Structure: tail <options> <filename>

Will exit automatically after the display

For **<options>** see **man**

Most common **<options>** is **–nX** where **X** is the number of the **last lines** that you want to display

Code: uslaad@ubuntu:~/Username/IBT2016/Genomics$ tail -n1 file1

Output: prints last line of file1 only

Some statistics about your file content: **wc** command

wc prints newline, word, and byte counts for each file (**count lines**)

Structure: wc <options> <filename>

-c: print the byte counts

-m: print the character counts

-l: print the newline counts

For more info about the different command use **man commandname** or **commandname –help**

Code: uslaad@ubuntu:~/Username/IBT2016/Genomics$ wc file1 #wc without argument

Output: 6 24 117 file1

Meaning 6 lines, 24 words and 117bytes in the file1

Code: uslaad@ubuntu:~/Username/IBT2016/Genomics$ wc -l file1

Output: 6 file1

Meaning it prints only the 6 lines in file1

BASIC COMMANDS FOR MANIPULATING FILES

How to manipulate files (default + options)

sort: reorder the content of a file "alphabetically"

Structure: sort <filename>

Sort outputs a sorted order of the file content based on the specified sort key (default: takes entire input) **Structure: sort <options> <filename>**

Default field separator: **Blank**

Sorted files are used as an input for sever other commands so sort is often used in combination to other commands

For **<options>** see **man**

Examples of sort:

Sort alphabetically (default option): sort <filename>

Sort numerically (sort numeric order): sort –n <filename>

Sort on a specific column (e.g. number 4): sort –k 4 <filename>

Sort based on a tab separator: sort –t $'\t' <filename>

…

uniq: removes duplicated lines

Structure: uniq <filename>

uniq outputs a file with no duplicated lines

uniq requires a **sorted files** as an input

Structure: uniq <options> <sorted_filename>

For **<options>** see **man**

Useful option is **–c** to output each line with its numbers of repeats

join: compare the contents of 2 files, outputs the common entries

Structure: join <filename1> <filename2>

Returns what they do have in common

join is used to compare 2 input files based on the entries in a common field (called "join field") and outputs a merged file

Join requires **sorted files** as an input

Lines with identical "join field" will be present **only once** in the output

Structure: join <options> <filename1> <filename2>

For **<options>** see **man**

e.g.

If file Gene1 contains: on each line

gene5

gene6

gene8

And file Gene2 contains: on each line

gene1

gene5

gene9

Code: $ join Gene1 Gene2
Output: gene5

diff: compare the contents of 2 files, outputs the difference
Structure: diff <finaname1> <filename2> OR Structure: diff <options> <finaname1> <filename2>
Diff is used to compare 2 input files and displays the different entries
Can be used to highlight differences between 2 versions of the same file
Default output: common lines not showed, only different lines are indicated and shows what has been **added (a), deleted (d) or changed (c)**
For **<options>** see **man**
Code: uslaad@ubuntu:~/Username/IBT2016/Genomics$ diff gene1 gene2
Output: # it provides the difference between the fines gene1 and gene2
0a1,2
> gene1
> gene2
2,3d3
< gene6
< gene7

BASIC COMMANDS FOR EXTRACTING DATA FROM FILE
How to extract data from files
(default + options)
grep: to search for the occurrence of a specific pattern (regular expression using the wildcards…) in a file
Structure: grep <pattern> <filename>
Grep ("global regular expression profile") is used to search for the occurrence of a specific pattern (regular expression…) in a file
Grep output the whole line containing that pattern
Structure: grep <options> <pattern> <filename>
For **<options>** see **man**
Example:
Extract lines containing the pattern xxx from a file:
grep xxx <filename>
Extract lines that do not contain pattern xxx from a file:
grep –v xxx <filename>
grep example
Let's consider a file named "ghandi.txt"
$ cat ghandi.txt
The difference between what we do
and what we are capable of doing
would suffice to solve
most of the world's problems
$ grep what ghandi.txt #Prints lines with the word 'what'

*The difference between **what** we do*
*and **what** we are capable of doing*
$ grep –v what ghandi.txt #Prints lines without the word 'what'
would suffice to solve
most of the world's problems

EXTRACTING DATA FROM FILES: CUT COMMAND

cut: is used to extract specific fields from a file

Structure: cut <options> <filename>

cut is used to extract specific fields from a file

For **<options>** see **man**

Important options are:

-d (field delimiter)

-f (field specifier)

Example:

extract ***fields 2 and 3*** from a file having ***'space'*** as a separator cut ***–d' ' –f2,3 <filename>***

Where: -d is the field delimiter option **' '** is the space and **–f2,3** is the fields 2 and 3

OUTPUT REDIRECTION TO A FILE

Be default, the standard output of any command will appear to the terminal screen

Redirection of the output result to a file is possible

This is particularly useful for big files

Structure: command options filename.in > filename.out

Parameter of command will work on filename.in and output will be written to filename.out

E.g. the example of join gene1 gene2 where the output displays gene5 on screen, you can redirect the output into another file.

Code: uslaad@ubuntu:~/Username/IBT2016/Genomics$ join gene1 gene2 > gene3

It automatically creates gene3 if it doesn't exist and output the join result to gene3

Try using nano gene3 to view the file content of gene3 in text editor i.e. the result of join gene1 gene2

Or cat gene3 to view the content of gene3 on screen.

Output redirection: combining different commands

Sometimes **you execute a command** and you willingly want **to use the output of the command as** an **input for another command.**

The **standard output** of any command will be **one unique output**

As seen previously, this output can be **printed** in the screen or **redirected to a file**

However, the **output** result of a command can also be **redirected to another command**

This is particularly useful when several operations are needed for a file, with no need to store the intermediate outputs

Combining several commands is done thanks to the use of **"|"** character i.e. pipe

Structure: command1 options1 filename1.in | command2 options2 > filename.out

This can be done for as many commands as needed, no limitation for the number of pipes you using in your code

When redirecting to another command, there is no need to specify the file. The second command will automatically take as input the output result of the previous command

Code: $ cut –d' ' –f2,3 Ghandi.txt | grep suffice > NewGhandi.txt

To check result **$ cat NewGhandi.txt**

Example:

Code: uslaad@ubuntu:~/Username/IBT2016/Genomics$ wc -l gene3 #Check how many lines are in gene3 file

Output: 2 gene3 #2 lines in filename gene3

You can use the command for join and pipe the output to be printed as number of the lines

Code: uslaad@ubuntu:~/Username/IBT2016/Genomics$ join gene1 gene2 | wc –l

Output: 2

PATHNAME

IF two folders are in the same folder e.g. Folder 2 and Folder 3 are inside Folder 1	
When you are in Folder1 and you wish to more into Folder2	**cd Folder2**
When you are in Folder1 and you wish to move into Folder3	**cd Folder3**
When you are in Folder2 and you wish to move to Folder 1	**cd ..**
When you are in Folder3 and you wish to move to Folder 1	**cd ..**
When you are in Folder2 and you wish to move to Folder3	**cd ../Folder3**
When you are in Folder3 and you wish to move to Folder 2	**cd ../Folder2**
Creating Folder4 into Folder3 while you are inside Folder2	**use –p for path**

mkdir –p ../Folder3/Folder4 It automatically adds Folder4 into Folder3

If Folder 4 is inside Folder3 and you wish to move from Folder4 to Folder2	**cd ../../Folder2**
Moving from Folder2 to Folder4	**cd ../Folder3/Folder4**

Linux Permissions, groups and process control

- **File and directory permissions**
- **Loops, variables and script generation to automate tasks**
- **Environment variables and why they are important**
- **How SSH onto to remote machine**

File and directories permissions

Linux is a multi-user OS, assigning access rights for confidentiality and permission in linux

Remember the ls –l examples previously used to list the content of a folder as follows:

When used as Code: uslaad@ubuntu:~/Username/Coursera$ ls –l

File permissions	**e.g.**	**drwxr-xr-x**
Owner	**e.g.**	**uslaad**
Group	**e.g.**	**uslaad**
Size in bytes	**e.g.**	**87**
Modification Time	**e.g.**	**18:17**
File Name	**e.g.**	**readme.txt**

Output:

-rw-rw-r-- 1 uslaad uslaad 3 Aug 16 10:57 file1

-rw-rw-r-- 1 uslaad uslaad 3 Aug 16 10:58 file2

Permissions are broken into 4 sections

Feature Type	User (Owner) Permissions (u)	GROUP Permissions (g)	Others Permissions (o)
-	rwx	r--	r--
"-" indicates a file "d" indicates directory "l" indicates a link	Read, write and execute permissions for the owner of the file	Read, write and execute permissions for members of the group owing the file	Read, write and execute permissions for other users

Access permissions on FILES

r indicates read permission: the permission to read and copy the file

w indicates write permission: the permission to change a file

x indicates execution permission, the permission to execute a file, where appropriate

Access permissions on DIRECTORIES

r indicates the permissions to list files in the directory

w indicates that users may delete files from the directory or move files into it

x indicates that the right to access files in a directory. This implies that you may read files in the directory provided you have read permission on the individual files.

NB: This means if you have x execute write on a given directory and you don't have read write on the files inside the directory, you won't be able to read the file.

chmod command

- **Used to change the permission of a file or a directory**
- **Syntax: chmod options permissions filename**
- **Only the owner of the file can use chmod to change the permissions**
- **Permissions define permissions for the owner, the group of users and anyone else (others)**
- **There are two ways to specify the permissions:**
  - **Symbols: alphanumeric characters**
  - **Octals: digits (0 to 7)**

Chmod options

Symbol Permissions	Meaning
u	User
g	Group
o	Other
a	All
R	Read
w	Write(and delete)
x	Execute(and access directory)
+	Add permission
-	Take away permission

Octal Permissions	
4 stands for "read"	
2 stands for "write"	
1 stands for "execute"	
0 stands for "no permission"	

chmod examples

chmod u=rwx,g=rx,o=r filename **in Symbol permission**

Same as

chmod 754 filename **in Octal permission**

The octal can be achieved from chmod u=rwx,g=rx,o=r filename

u as rwx				g as rx				o as r		
u = user				g = group				o = others		
r	w	x		r	w	x		r	w	x
4	2	1		4	0	1		4	0	0
7				5				4		

More examples

777: (rwxrwxrwx) No restrictions no permissions. Anybody may do anything

755: (rwxr-rx-x) The file's owner may read, write and execute the file. All others read and execute the file (common) for programs that are used by all users)

700: (rwx------) The file's owner has all the rights. Nobody else has any rights (private for the owner) It means it is a private file or directory to the owner, no one else can see the file
666: (rw-rw-rw-) All users may read and write the file (this type of file cannot be executed or executable e.g. Microsoft excel file you can read and write only
644: (rw-r--r--) The owner may read and write a file, while all others may only read the file (everybody may read, but only the owner may change)
600: (rw-------) The owner may read and write a file. All others have no rights

Environment Variables

Variables are areas of memory that can be used to store information and are referred to by a name
How to create variable: a line that contains the fname of the variable followed immediately by an equal sign ("=") NB: spaces are not allowed between the name of the variables
2 types of variables: shell variables and environment variables
Some variables are already set in your shell session
printenv: prints the values of all your environment variables

WHAT IS AN ENVIRONMENT VARIABLE

- **An environment variable is a dynamic "object" on a computer that stores a value which can be referenced by programs**
- **Environment variables help programs know what directory to install files in, where to store temporary files, where to find user profile settings etc**
- **Environment variables help to create and shape the environment of where a programs runs.**

Examples of Environment Variables (They are in UPPERCASE i.e. case sensitive in Linux)
HOME: the environmental value that shows the current user's home directory
PATH: the environmental variable which contains a colon-separated list of the directories that the system searches to find the executable program corresponding to a command issued by the user
PWD: always stores the value of your current working directory
e.g. after installing a program and execute the path, the system won't run the program, you need to update the path variable the system will recognize the command
NB: Environment variables are dynamic and could change

PATH
Add a new directory to your PATH environment variable using the export command i.e. when you installed the application

```
export PATH=/home/user/program      #This one will add the folders directory at the end of the folder it contain
```

Or

```
export PATH=/home/user/program:$PATH      #This one will put it in the beginning of the directory list, basically it will be the first directory, where the system will search for the command.
```

SHELL SCRIPTING

Basically means automating some tasks and executing different kind of tasks under a whole same script, you don't have to type command by command.

Commands Use in Scripting includes:
echo command

- **Syntax: echo options arguments**
- **Writes arguments to the standard output**
- **echo: just prints its command-line parameters to standard output**
- **If you redirect the result your arguments will be written into the file you are redirecting to**
- **Commonly used by the shall scripts to display results or ask the user to enter parameters for interaction**

Let's echo some stuff

- **echo Bioinformatics is great starting writing scripts**
- **If you want to jump to another line add \n and use the option –e**
  NB: to use a new line \n you have to use -e
- **echo –e "Bioinformatics is great \n starting writing scripts"**
- **Setting a variable: X=firstvariable**
- **echo X: prints X**
- **echo $X prints firstvariable (the value of the variable) Therefore you add the sign $ when you want to print the value of a variable**
- **echo '$X' -> $X #Single quote gives you the same string in the quote**
- **echo "$X" -> firstvariable #Double quote prints the values of the variable**

Print the result of a command

- **Asking the shell to substitute the results of a given command**
- **'command' or $(command)**
  Try echo 'pwd' #prints the character pwd SAME AS echo pwd #prints the strings pwd
  echo $PWD and echo $(pwd) #prints the current working directory path for pwd i.e. /home/uslaad
  NB: Take note of the UPPERCASE $PWD because $pwd outputs nothing so it has to be $(pwd)
  echo $HOME #prints the path for HOME directory i.e. /home/uslaad
  NB: don't use $(HOME) or $(home) it won't work

echo $PATH #gives the different folders where the system is going to look for executables
Output: /usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/local/games
The output above shows list of different directories where executions can be performed

WHAT IS A SHELL SCRIPT?

- **Short programs written in shell programming language useful to automate tasks under Linux OS**

- **A shell script is a file containing a series of commands**
- **Could be helpful to perform the same actions on many different files**
- **Shell scripts = scripting interpreter + command line interface to the system**
- **Echo is also commonly used to have a shell script display a message or instructions, such as Please enter Y or N in an interactive session with users.**

Let's start using the power scripting

1. **nano myfirstscript**
2. **Write the content of your script for example:**

CODE: Inside the file

#!/bin/bash ← Make use the shebang when using WEBINAL
clear ← For your Linux just start from this line
echo "Writing my first shell scripts" #Sentences with OR without " "
echo "Let the fun begin!!!"

Save the file and exit

3. **Run your script (using ./)**
   Run Code: ./myfirstscript
   Output: bash: ./myfirstscript: Permission denied
   Meaning there is no permission to execute

Check the permission with ls –l you will see the permission setting for the file myfirstscript
Output: -rw-rw-r-- 1 uslaad uslaad 94 Aug 18 11:32 myfirstscript
It shows am the owner of the file with the right to read and write but not execute

4. **Change the rights to make sure you have the right to execute using the chmod command**
   chmod u+x myfirstscript #u means permission for user of machine
   or
   chmod 744 myfirstscript

Recheck permission making use of ls –l

Output: -rwxr--r-- 1 uslaad uslaad 94 Aug 18 11:32 myfirstscript

5. **Run the command for the script ./myfirstscript**
   Output: Writing my first shell scripts
   Let the fun begin

THE SHEBANG

#!interpreter

#!/bin/bash

#! **Indicates the beginning of a script (the shell understands this)**

/bin/bash **The program interpreter: absolute path to an executable program (interpreted by the bash)**

A perl script could begin by #!/usr/bin/perl (absolute path)

If you don't know where your interpreter of shell script is… You can use the which command to locate the executable file associated with a given command

```
which perl → /usr/bin/perl
which bash → /bin/bash                    #Output from my own system: /bin/bash
```

Advice

- **Use variables in your scripts**
- **Make your scripts easier to maintain (you don't have to check through all variables)**
- **Reduce the amount of typing!**

Shell is much more powerful than that… IF STATEMENTS IN SHELL SCRIPTING

Syntax: if [conditional expression]

```
then
        commands
else
        commands
fi      #To end the if statement control
```

Example:

```
#!/bin/bash
echo Let's try some conditional tests
x = 'find*.fasta | wc – l'          #Output of find *.fasta will be input of the 2nd command wc - l
y = 2     #Another variable
echo "The current working directory contains $x fasta files" #Displaying value of x with $x
if [$x –gt $y] # if (($x > $y))     #Compare value of if $x is greater than (i.e. –gt) $y, the # at the middle followed by another control statement in any scripting language means interpreter wont interpret what is going on in the commands that precede the #. Therefore # is like a for comment and you can use either if[$x –gt $y] OR if(($x>$y)) is applicable. NB: -gt uses [ ] and the > uses ( ).

then
        echo There are many existing fasta files in this directory
else
        echo There are few fasta files in this directory here is the listing: 'ls *.fasta'
fi      #To end the if statement control
```

```
CODE:  #!/bin/bash
       echo Let try some conditional tests
       x=`find *.fasta| wc -l`          #NB the ` symbol is under the escape key on my keyboard
       echo The current working directory contains $x fasta files
       y=2
       if(($x>$y))
        then    #single space bar before the then
         echo There are many existing fasta files on this directory        #double space bar
       else
        echo There are very few fasta files on this directory: `ls *.fasta` #NB the ` symbol is not ‘’
       fi                #This closes the if command
OUTPUT:        Let try some conditional tests
               The current working directory contains 1 fasta files
               There are very few fasta files on this directory: new2.fasta
```

LOOPS IN SHELL SCRIPTING (for)

You can repeat a given treatment different times using the for loop

Syntax

```
for variable in values
do
        commands
done
```

Example: In relation to created file1 and file2 try this script, don’t forget permission… chmod

```
Code:  #!/bin/bash
       for x in file1 file2
       do
        head -n 3 $x
        echo operation completed on file: $x
       done
```

NB: The code above is just a way of doing a command on files and it prints a message to show that what has been completed or done.

```
Output:
       I will use 'cat' to view content of this file
       I am adding more lines
       Another line
       operation completed on file: file1
       This a way to learn
       Just trying out a script to enjoy the flow
       Interesting though
       operation completed on file: file2
```

Another powerful loop you can use in shell scripting is the WHILE LOOP

Loops in shell scripting (while)

Syntax:

```
While [condition]
Do
        Command1
        Command2
Done
```

Example:

```
CODE:  #!/bin/bash
       n=1
       while [ $n <= 5 ] #n should have an initial value
       do
        echo Welcome $n times
        n=$(( n+1 )) #increment $n
       done
```

OUTPUT:

OPERATORS SUPPORTED BY SHELL

Examples: Consider 2 variables a = 10 and b = 20

Operator	Description	Example
+	Addition – Adds values on either side of the operator	'expr $a + $b will give 30
-	Subtraction – Subtracts right hand operand from left hand operand	'expr $a - $b will give -10
*	Multiplication – Multiplies values on either side of the operator	'expr $a * $b will give 200
/	Division – Divides left hand operand by right hand operand	'expr $b / $a' will give 2
=	Assignment – Assign right operand in left operand	A=$b would assign value of b into a
==	Equality – Compares two numbers, if both are same then returns true	[$a == $b] would return false
!=	Not Equality – Compares two numbers, if both are different then returns true	[$a != $b] would return true
-eq	Checks if the value of two operands are equal or not, if yes then condition becomes true	[$a –eq $b] is not true
-ne	Checks if the value of two operands are equal or not, if values are not equal then condition becomes true	[$a –ne $b] is true
-gt	Checks if the value of left operand is greater than value of right operand, if yes then condition becomes true	[$a –gt $b] is not true
-lt	Checks if the value of left operand is less than the value of right operand, if yes then condition becomes true	[a$ -lt $b] is true.
-ge	Checks if the value of left operand is greater than or equal to the value of right operand, if yes then	[$a –ge $b] is not true.

	condition becomes true	
-le	Checks if the value of left operand is less than or equal to the value of right operand, if yes then condition becomes true	[$a –le $b] is true.
-eq	Checks if the value of two operands are equal or not, if yes then condition becomes true	

NB: All the operators should come between [ ] and don't forget the $ sign for variable calling

CONTROLLING TASKS

Commands to control processes i.e. the task you are running

- **ps** list the process running on the system
- **kill** send a signal to one or more processes (usually to "kill" a process) e.g. for loops
- **jobs** an alternate way of listing your own processes
- **bg** put a process in the background e.g. running the Bowtie or BWA or IGV that uses GUI, they all takes time, so to enable you have the prompt to write other commands, you can put such processes/job in the background run it later

Launching a background Job

Programs that takes time or open a new Graphical User Interface GUI

The prompt doesn't reappear after the program launched. The shell is waiting for the program to finish before control returns to you

- **Ctrl + Z: interrupts a program**

Or

- **You can put it in the background so that the prompt will return immediately**
- **Use the command name followed by & do do so**

NB: If you run the ps or job you will still see that your job is running but in the background

SSH INTO REMOTE MACHINE

SSH is basically a protocol that allows you to connect to remote machines

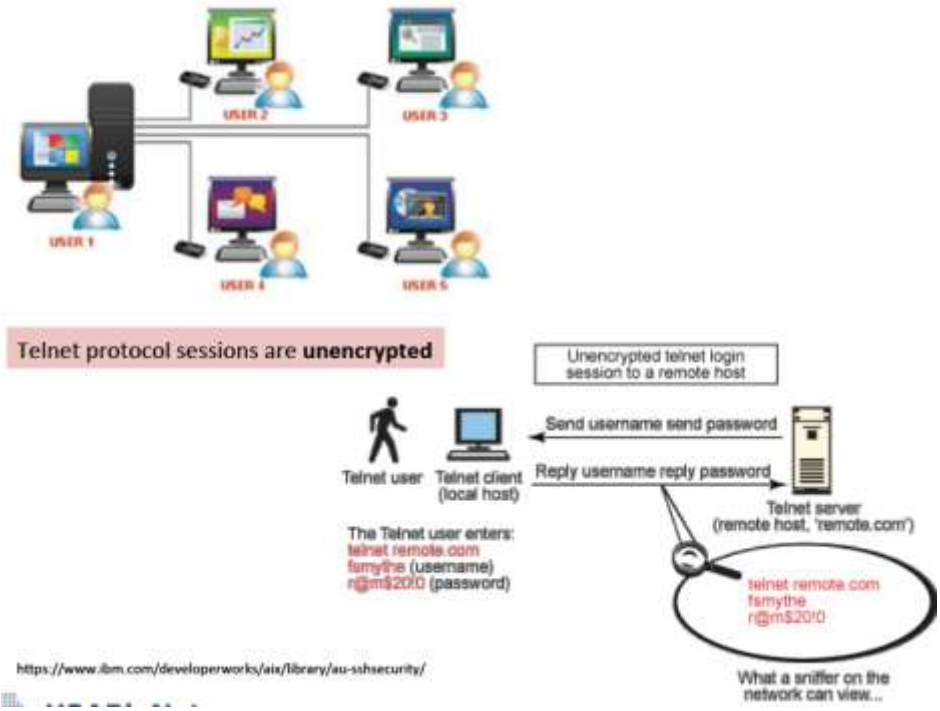
Linux is a multi-user machine

- **While your computer only has one keyboard and monitor, it can still be used by more than on user**
- **For example, if your computer is attached to a network, or the internet, remote users can log in via ssh (secure shell) and operate the computer**
- **Remote users can execute applications and have the output displayed on a remote computer**

WHAT IS SSH

- **SSH (secure Shell) is a protocol used to securely log onto remote systems (remote Linux machine and Unix-like servers)**
- **Basically like the telnet protocol, but telnet is unencrypted, using SSH to connect to a server the information is encrypted and secured i.e. SSH protocol sessions are encrypted.**

Telnet protocol sessions are unencrypted



- **ssh command is the tool used in Linux to connect via SSH protocol**
- **Syntax: ssh remoteusername@remotehost**
- **Remote host could be an IP address or domain name**
- **You will be asked to provide your password**
- **To exit and go back to your local session, use exit**

COPY FILES FROM OR TO A REMOTE MACHINE

scp: secure copy (from server to machine or from machine to server)

Syntax: scp pathfrom pathto

The difference: cp and scp: in scp, at least the source or the destination is in a remote machine

Example: uploading all the .txt files from your current working directory to a remote host

scp ./*.txt username@myhost.com:/home/username/folder

Meaning: **scp** secure copy

./ current working directory

***.txt** all text files

The long filepath as pathto

MODULE 4 SESSION 1

Learning Objectives

- To discuss the concepts of sequence alignment and its applications
- To explain the concepts of homology, identify, orthologues and paralogues
- Demonstrate basic sequence alignment, introducing concepts of point mutations, deletions and insertions
- To examine the different approaches to sequence alignment – **exhaustive vs heuristic**

Learning Outcomes

At the end of this module, participants should be able to:

- **Demonstrate basic applications of sequence alignment**
- **Explain the concepts of homology, identity, orthologues and paralogues**
- **Identify point mutations, deletions, insertions in aligned sequences and**
- **Differentiate between exhaustive and heuristic approach to sequence alignment**

Introduction

- **Comparison is one of the most effective methods use in making discovery in Science**
- **The most common comparative method use in molecular biology is alignment**

What is Sequence Alignment?

- **It is a way of arranging two or more sequences DNA, RNA or protein to identify regions of similarity**
- **Comparing two sequences – Pair wise alignment**
- **Comparing more than 2 sequences – Multiple sequence alignment**

Alignment Sequences

- **Sequence alignment of nucleotide or amino acid residues are typically represented as rows and columns**

Why Align Sequences?

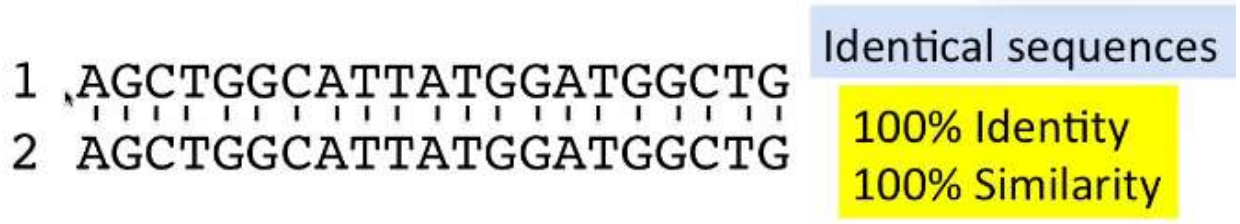
- **To predict function of gene or protein sequences**
- **To merge short strings of DNA or RNA sequences**
- **To automatically annotate newly sequenced genomes**
- **It also the first step in phylogenetic analysis**
- **Essential for molecular structure prediction**
- **To detect gene mutations – substitution, insertions/deletions**

Similarity and Identity – Nucleotide

- **Apparent likeness or resemblance between sequences**

Seq1 **ATCGGGAGCGAGCGACGATGCGC** **identical sequences**
Seq2 **ATCGGGAGCGAGCGACGATGCGC** **100% Similarity**

- **It is an objective repeatable criteria**

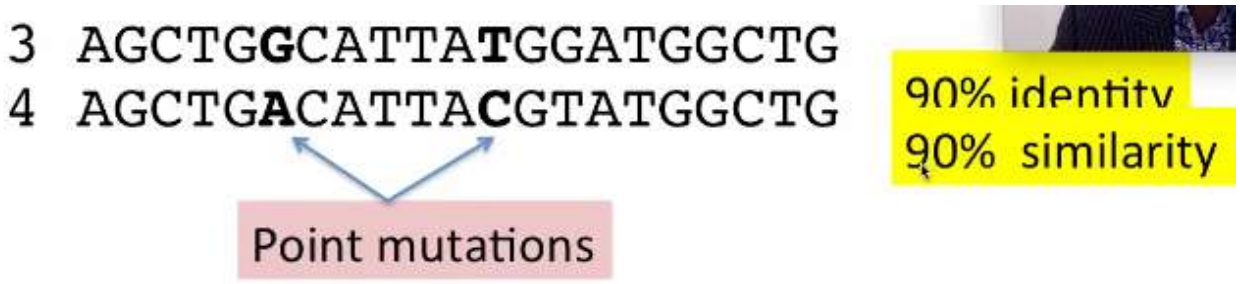


Comparing Two Sequences

Seq3 ATCGGGGGCGTGCGA

Seq4 ATCGGGAGCGAGCGA **Points of mutations in red (90% identity and similarity)**

- **Sequence similarity and sequence identity are synonymous for nucleotide sequences**



Percentage of Sequence Similarity/Identity

Two sequences of equal length, percentage of similarity or identity = $[2L/(L_y+L_z)] \times 100$

Where: **L** is the number of aligned residues with similar or identical characteristics

L_y is the total length of sequence y

L_z is the total length of sequence z

In a case where the two sequences are of different lengths (i.e. varying length)

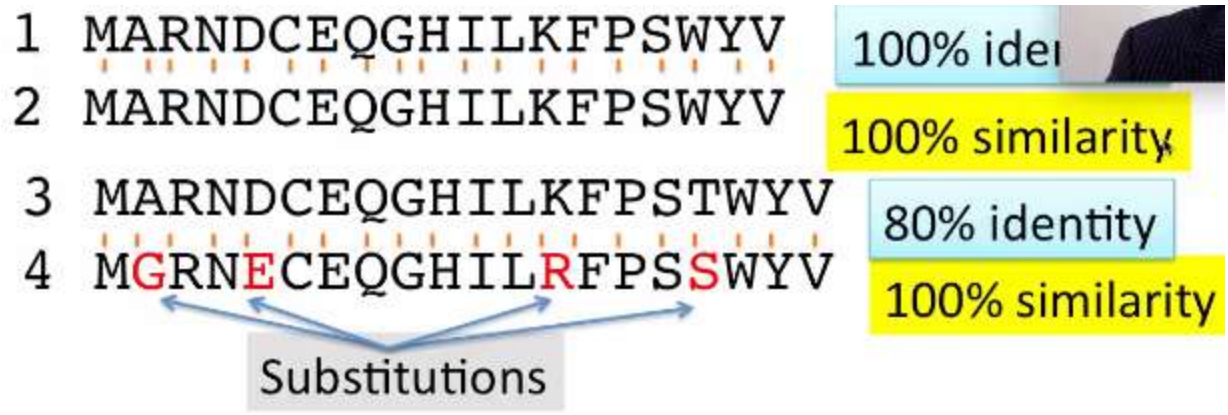
Two sequences of unequal length, percentage of similarity S or identity I

$$I(S) = (L_{i(s)}/L_y)100$$

Where: **$L_{i(s)}$** is the number of aligned residues with similar or identical characteristics

L_y is the length of the shorter of the two sequences

Similarity and Identity – Amino Acids

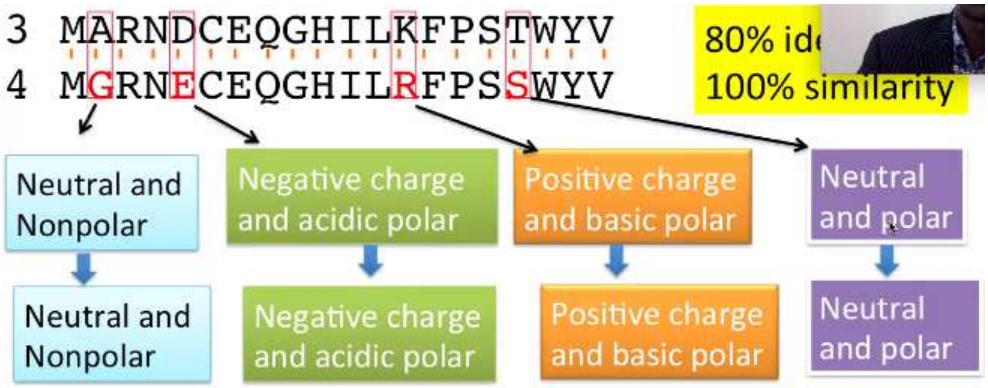


Mutation as a result of substitutions as shown above in amino acid

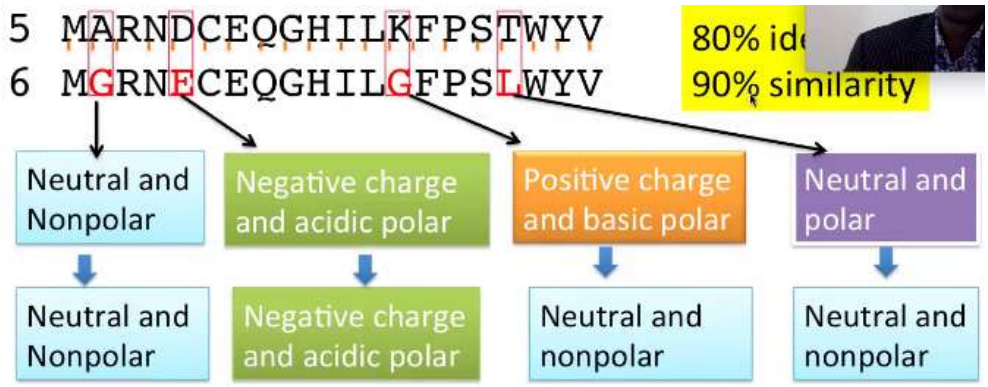
In terms of identity, sequence 1 and 2 are 100% identical and similar while sequence 3 and 4 are 80 identical and 100% similar.

NB: Sequence similarity and Sequence Identity are NOT Synonymous for amino acid residues

Some Amino acids are similar in characteristics even though they different amino acids… that is why they have 100% similarity

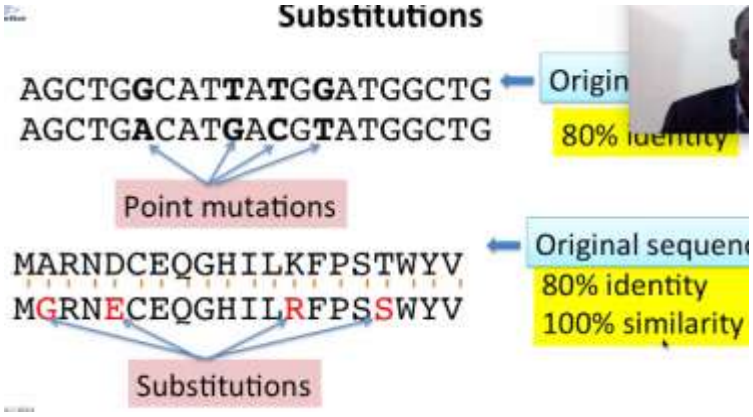


In this case, the amino acid differ in characteristics, to the similarity will be 90%



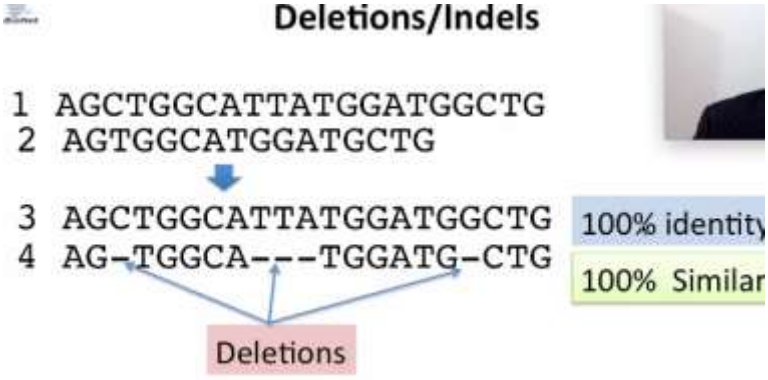
Looking at substitutions and point mutations in the nucleotide GTTG are substituted/mutated into AGCT (therefore the identity is 80%)

In the amino acid GERS are substituted still represents 100% similarity



Deletions/Indels

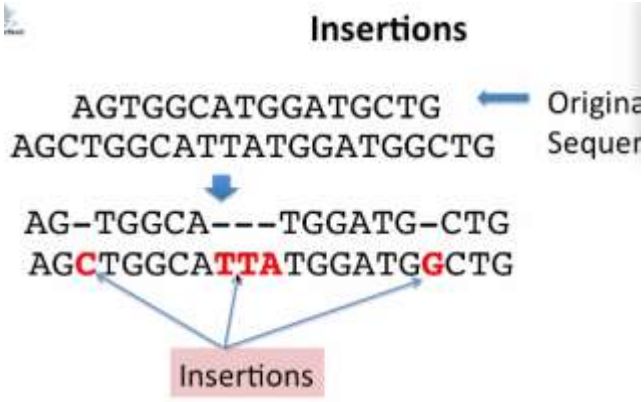
Given two sequences of different length it will be difficult to know what is happening except we align both



NB: Considering the upper sequences as the Original… the sequences above still maintain 100% identity and 100% similarity

Insertions

If the upper sequence below is taking as original, then we say there are insertions



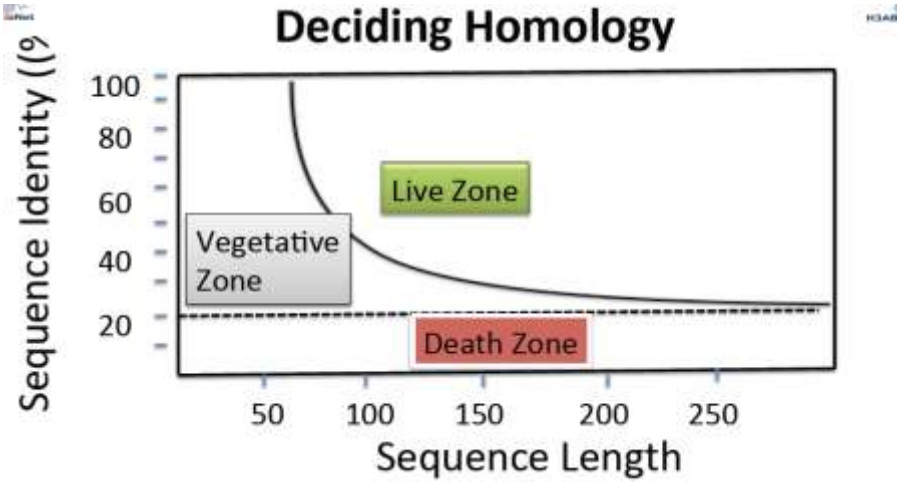
Sequence Homology

- **The two proteins, each having more than 100 one amino acids long (or 2 genes having more 100 nucleotides long) and share at least 25 percent identical amino acids residues or 70% identical nucleotides**

- **Two proteins exhibiting significant similarity, they are assumed to be evolutionary related**

Deciding Homology

In deciding homology there is a thin line between deciding either the two sequences are homology or not. In terms similarity we have 100%, 80% etc but in homology we don't have degree of homology



To decide whether two sequences are homologous, we look are the length of the two sequences, as the length of the sequences increases the percentage that you need to decide whether is it homologous, so at 25% it is taking as DEAD ZONE, but when it is less than 100% and percentage of identify is well beyond 25% then we say it is VEGETATIVE ZONE i.e. the border line. The SAVE ZONE or LIVE ZONE is when the two sequences exhibit identity more than 25% and having length more than 100 nucleotides or amino acid.

Homology

- **Conclusion drawn from two genes that shows/share common evolutionary history**
- **Sequence alignment methods try to model the molecular mechanisms by which sequences evolved**
- **Genes are either homologous or are not homologous, as there are no degrees for homology as there are for similarity.**
- **The changes that occur during divergence from the common ancestor can be categorized as substitutions, insertions and deletions**
- **Two segments of proteins can share ancestry because of either a speciation event or a duplication event**

Orthologues

- **Species diverges into two separate species, the copies of a single gene in the two resulting species are said to be orthologues**
- **Orthologues genes are genes in different species that originated by vertical descent from a single gene of the last common ancestor**
- **Orthologous sequences provide useful information in taxonomic classification and phylogenetic studies of organisms**
- **The pattern of genetic divergence can be used to trace the relatedness of organisms**

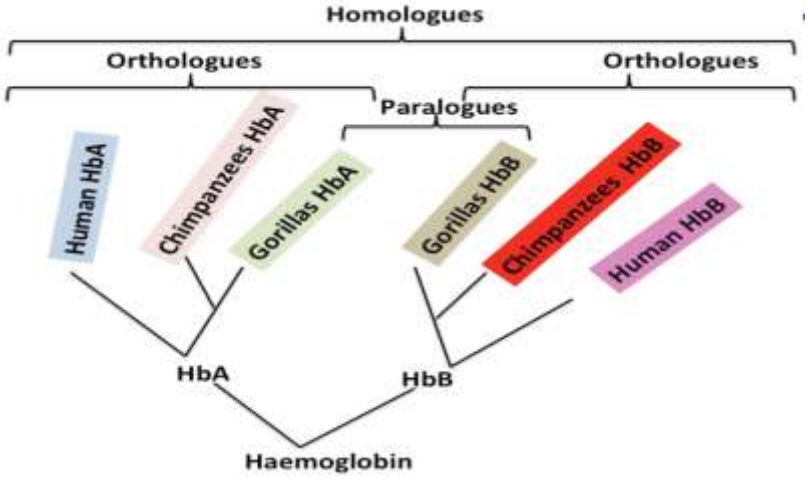
- **Two organisms that are very closely related are likely to display very similar DNA sequences**

Paralogues

- **When a gene in an organism is duplicated to occupy two different positions in the same genome, then the two copies are said to be paralogues**
- **Often paralogue genes belong to the same species**
- **Typically have the same or similar function but sometimes do not**
- **Paralogous sequences provide useful and dramatic insight into some of the way genomes evolve**
- **Four classes of hemoglobin (hemoglobin A, hemoglobin A2, hemoglobin B and hemoglobin F) are paralogous of each other**
- **While each of these proteins serves the same basic function of oxygen transport, they have diverged slightly in function: fetal hemoglobin (hemoglobin F) has a higher affinity for oxygen than adult hemoglobin**

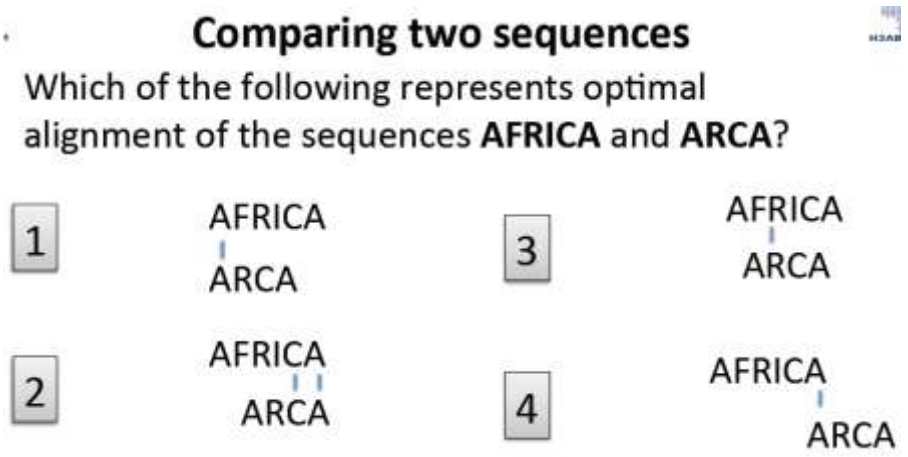
Illustrating these on a diagram

The ancestral "Haemoglobin" last common ancestral has only one haemoglobin, and this differentiated/diverted into different species HbA and HbB, they further split in Orthologues as same from one gene or paralogues in different but with same function



Basis of Sequence Alignment

- **The overall goal of pairwise sequence alignment is to find the best pairing of two sequences**
- **One sequence needs to be shifted relative to the other to find the position where maximum matches are occurring**



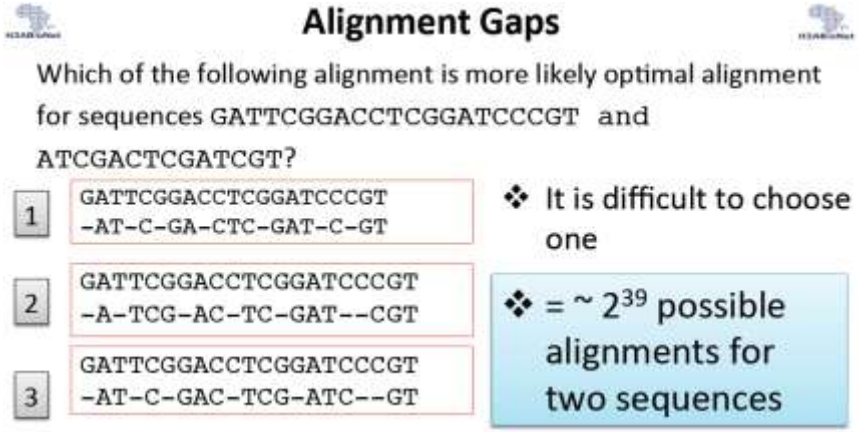
The optimal alignment is the best alignment between the sequences, which is better in Option 2 above.

Better Alignment

- **We insert gaps (a base-to-nothing match) using Global Alignment**

- **Maximizes the number of base-to base matches**
- **The order of bases in each of the sequence is maintained**
- **Gap-to-gap matched are not allowed**

It is difficult to choose one alignment as the optimal alignment for two sequences, in actual fact there are about approximately 239 possible alignments for the two sequences



Search for Optimal Alignment

- **It is desirable to allow some gaps to be introduced into an alignment to compensate for insertions and deletions but this should not be arbitrarily inserted**
- **There are number of strategies that have been proposed for panelizing gaps to obtaining optimal alignment**

Optimal Alignment

- Employ scoring scheme – to reward for matching, punish for mismatching and gaps
- Similarity of two sequences = high score
- Dissimilarity of two sequences = low score
- Account for substitutions, insertions or deletions

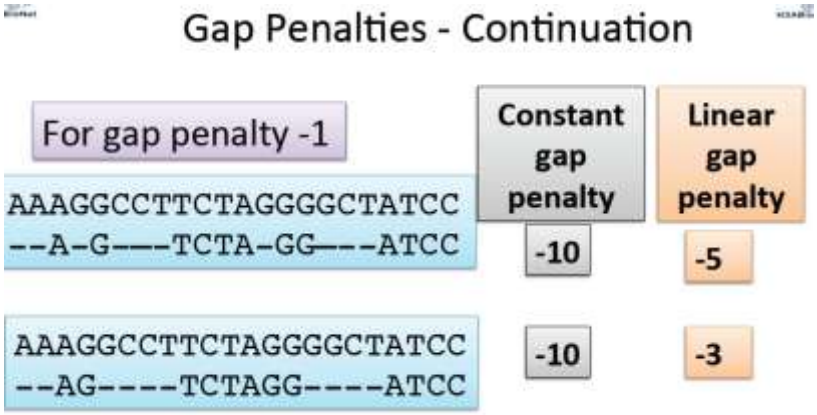
Gap Penalties

- Constant gap penalty, a penalty of "-a" for gap, irrespective of the length of the gap.
  For high score (to obtain high score)
  - The total number of gaps matters
  - Minimizing the number of gaps will improve the similarity score

Another approach is…

- Linear gap penalty, a penalty of "-a" per unit length of a gap
  - The penalty for one large gap is the same as for many small gaps

Example Constant Gap Penalty and Linear Gap penalty shown below



The gaps of the both sequences are the same, but they are inserted in different positions. Given that you award -1 for every gaps then Constant Gap penalty is -10 for both. For the Linear gap penalty it doesn't consider number gaps between the nucleotides, it will count the group of possible gaps as one therefore the upper has 5 while the lower has 3.

- Constant gap penalty and linear gap penalty do not consider whether the gap is opening or extending
- Gaps at the terminal regions are often treated with no penalty because in reality many true homologous sequences are of different lengths

Another approach of gap penalty is…

- Affine gap penalty is a differentia gap penalty introducing and extending gaps
  - For example, one may use a "-y" for gap opening penalty and "-z" for gap extension penalty

    Total gap penalty, G = y+(k-1)z

    Where k is the length of the gap

But…

- **Drawback, the penalty points (y and z) are arbitrarily chosen constants …it is not effective**

Affine Gap Penalty

- **It favors longer gaps over single gaps of the same total length**
- **Reflects more likely in biology, one gap of 15 is likely to occur in sequence than 15 small gaps of 1**

ALIGNMENT ALGORITHMS

- **Different optimization strategy are used in aligning similar residues**
- **Computer algorithm can then determine not just minimal edit distances but optimal alignment**
- **Types of alignment are based on one of these three methods:**
  - **The exhaustive method**
  - **The heuristic method**
  - **Dot matrix methods**

Exhaustive Algorithm

- **Produces list of all possible alignment between two sequences, score and select the alignment with best score (OPTIMAL ALIGNMENT)**
- **Dynamic programming (Needleman Wunsh, 1970)**
- **The number of possible global alignments between sequences of length L is $2^{2L}/\sqrt{(\prod L)}$**
- **For sequences of 21 bases/residues, there are ~2^{39} possible ways of aligning them**
- **Databases continued to increase at an exponential rate, making straight forward application of dynamic programming methods impractical for database searching**
- **NOT practical for multiple sequence alignment**

Heuristic Algorithm

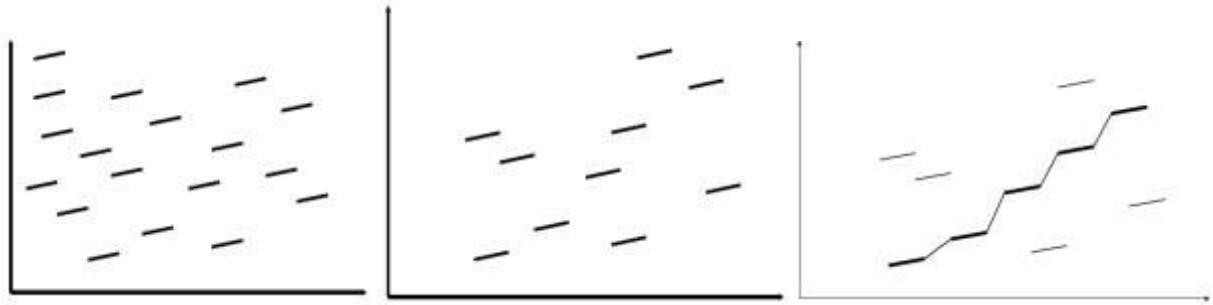
- **Heuristic method is based on the strategy of breaking a sequence up into short runs of consecutive letters called words**
- **Makes use of approximations to significantly speed up sequence comparisons**
- **Finds a solution close to the best one, fast and easily**
- **But with a small risk that true alignment can be missed**

Heuristic Algorithms

Other alignment methods that are based on heuristic algorithms include:

- **Progressive alignment**
- **Iterative alignment**
- **Block-based alignment**
- **Examples of heuristic algorithms – Clustal, Tcoffee, BLAST etc**

Block based alignment is basic form of the heuristic alignment method, where alignments are broken into small fragments of words, then compare the two sequence and find words that are common to each of them, and align them.



Broken Fragments | Match words on both sequences | Align the match words

Iterative Alignment method

When two sequences are aligned side by side, then Realign the initial sequences as well as adding new sequences to the growing multiple sequence alignment

Progressive Alignment

- Progressive alignment is a pairwise alignment algorithm that is used iteratively, first to align the most closely related pair of sequences, then the next most similar one to that pair and so on
- The rule "Once a gap, always a gap" based on the fact that the positions and lengths of gaps introduced between more similar pairs of sequences should not be affected by more distantly related ones.

Summary

- In molecular biology certain changes are more likely to occur naturally than others
- Amino acid substitutions tent to be conservative: the replacement of one amino acid by another with similar size of physiochemical properties is more likely to have occurred than its replacement by another amino acid with greater difference in their properties
- Alignment by eyeball we must define quantitative measures of sequence similarity and difference

Sequence alignment method is important technique for comparison of sequences

Apply computer algorithms to determine optimal alignment

MODULE 4 SESSION 2

Sequence Alignment Theory and Applications

Session: Pairwise Sequence Alignment

Learning Objectives

- Understand the difference between global and local pairwise alignment algorithms
- Understand the basic principles of pairwise alignments, scoring matrices and gap penalties

- **Understand the concepts of the dynamic programming approach for pairwise sequence alignment (global and local)**

Pairwise Sequence Alignment

- **Pairwise sequence alignment is the fundamental component of many bioinformatics applications**
- **It is extremely useful in structural, functional and evolutionary analyses of sequences**

Global Alignment

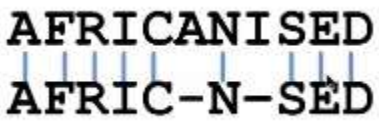
- **The two sequences to be aligned are assumed to be generally similar over their entire length**
- **Alignment is carried out from beginning to end of both sequences to find the best possible alignment across the entire length between the two sequences**
- **More applicable for aligning two closely related sequences of roughly the same length**

Disadvantage of Global Alignment

- **May not be able to generate optimal result for divergent sequences and sequences of variable lengths**
- **It fails to recognize highly similar local regions between the two sequences**

Example of Global Alignment:

Global Alignment of the following sequences AFRICANISED and AFRICNSED



Local Alignment

- **Finds local regions with the highest level of similarity between the two sequences and aligns these regions**
- **Used for aligning more divergent sequences with the goal of searching for conserved patterns in DNA or protein sequences**
- **The two sequences to be aligned can be of different lengths**
- **The paired subsequences may be surrounded by residues that are completely unrelated**
- **More appropriate for aligning divergent biological sequences containing modules, domains or motifs**

Example:

Local Alignment of the following sequences AFRICANISED and FRIC

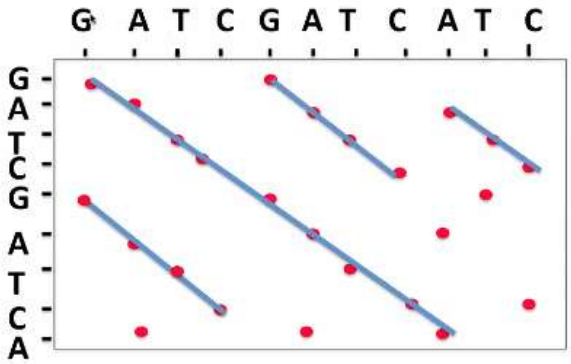
Alignment Algorithms

- **Both global and local are fundamentally similar**
- **Differ in the optimization strategy used in aligning similar residues**

Other methods used in alignment…

DOT – MATRIX REPRESENTATIONS

- The most basic sequence alignment method
- Use the sequences as the coordinates of a two-dimensional graph and then plot points of correspondence within its interior
- When two sequences are consistently matching over an extended region, the dots will merge to form a diagonal line segment



- Has the ability to reveal complex relationship involving multiple regions of local similarity
- Gives a direct visual statement of the relationship between two sequences and helps easy identification of the regions of greatest similarities

Advantages of Dot Matrix

- Easy identification of the regions of greatest similarities
- Identification of sequence repeat regions based on the presence of parallel diagonals
- Useful in identifying chromosomal repeats and in comparing gene order conservation between two closely related genomes

Disadvantages of Dot Matrix

- Lacks statistical rigor in assessing the quality of the alignment
- Restricted to pairwise alignment
- It is difficult for the method to scale up to multiple alignment

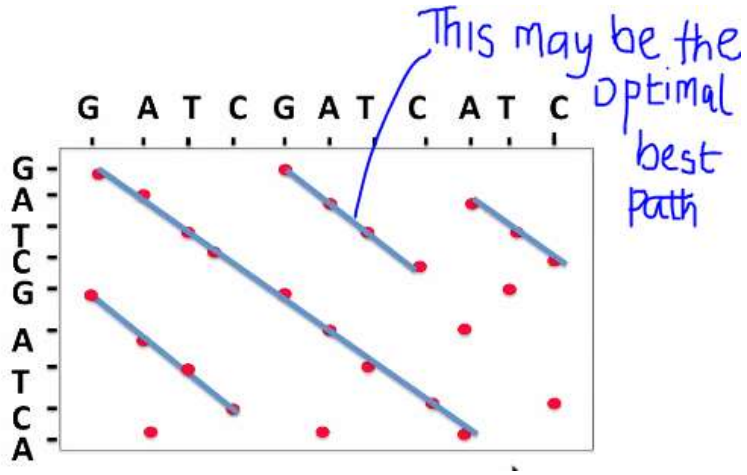
Optimal Alignment Methods

- To find the optimal path through a graph
- Sum the incremental contributions of each step along its route
- Positive incremental scores used for aligning identical residues, with negative scores used for substitutions and gaps
- Total score maximal will give us the best sequence alignment

One of the optimal alignments is called…

Needleman-Wunsch Algorithm

- **Is an application of a best-path strategy aimed to reduce the massive number of possibilities and still guarantee the best solution will be found (Needleman and Wunsch, 1970)**
- **Is an example of dynamic programming**
- **Any partial subpath that ends at a point along the true optimal path must itself be the optimal path leading up to that point**
- **The optimal alignment must extend from the beginning to end in both sequences**
- **It seeks global alignments**



There may be a long path from start to end, but there also may be a short length of optimal best path

How Does Dynamic Programming Work?

- **Break the problem into smaller sub-problems**
- **Solve the smaller problems optimally**
- **Use the sub-problem solutions to construct an optimal solution for original problem**

Dynamic Algorithm

Consists of three steps:

- **Initialization of score matrix**
- **Calculation of scores and filling the trace-back matrix**
- **Deducing the alignment from the trace back matrix**

Example: Find the optimal alignment for the sequences AGATCAT and AGCGT, math +1, mismatch 0 and gap -1

Initialization Step:

- **Create Matrix with M+2 columns and N+2 rows. The second cell of the second row and the second column filled with 0.**

		A	G	A	T	C	A	T
	0	-1	-2	-3	-4	-5	-6	-7
A	-1							
G	-2							
C	-3							
G	-4							
T	-5							

Calculation of Scores

- **Horizontal movement is a gap, -1**
- **Diagonal movement is match, +1 or mismatch, 0**

		A	G
	0	-1	-2
A	-1		
G	-2		

		A
	0	-1
A	-1	1 or 0 -2 -2

There are three (3) ways of movement horizontally or vertically for gap, while diagonal is either a match of mismatch score. The highest sc ore from the three (3) values will retain the cell value

The highest scores are indicated as:

		A	G	A	T	C	A	T
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	-1 0 -3	-1 -4	-3 -5	-4 -6	-4 -7	-6 -8
G	-2	-1 -2 0	2 -1 -1	0 -2 1	-1 -3 0	-2 -4 -1	-3 -5 -2	-4 -6 -3
C	-3	-2 -3 -1 -4	0 -2 1	2 0 0	1 -1 1	1 -2 0	-1 -3 0	-2 -4 -1
G	-4	-3 -2 -5	0 -3 0	0 1 -1	2 0 0	1 0 1	1 -1 0	0 -2 0
T	-5	-4 -3 -6	-2 -1 -4	0 0 -2	2 1 -1	2 0 1	1 0 1	2 -1 0

Filling the trace back matrix

		A	G	A	T	C	A	T
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0	-1	-2	-3	-4	-5
G	-2	0	2	1	0	-1	-2	-3
C	-3	-1	1	2	1	1	0	-1
G	-4	-2	0	1	2	1	1	0
T	-5	-3	-1	0	2	2	1	2

TRACEBACK

- **Traceback is the process of deducing the best alignment**
- **Begins with the last cell filled with the score then move according to the traceback value written in the cell**
- **There are three possible moves – diagonally toward the top-left corner of the matrix, up of left**

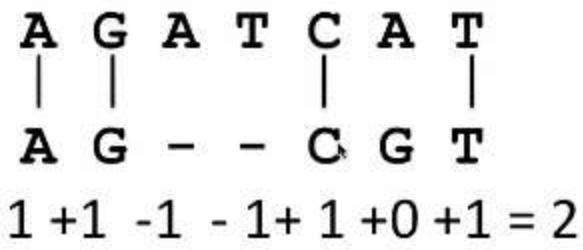
- **The traceback is completed when the first, top-left cell of the matrix is reached**

Traceback path

The optimal alignment score is calculated in the lower-right corner

		A	G	A	T	C	A	T
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0	-1	-2	-3	-4	-5
G	-2	0	2	1	0	-1	-2	-3
C	-3	-1	1	2	1	1	0	-1
G	-4	-2	0	1	2	1	1	0
T	-5	-3	-1	0	2	2	1	2

Optimal Global Alignment



In the image above, from the left side it shows: Diagonal, Diagonal, Diagonal, Horizontal, Horizontal, Diagonal and Diagonal
Optimal alignment score = 2

Advantages of Needleman-Wunsch Algorithm

- **Suitable for aligning two closely related sequences that are of the same length**
- **But two similar sequences of unequal lengths can be aligned without penalizing terminal gaps**

Disadvantages of Needleman-Wunsch Algorithm

- **Risk of missing the best local similarity**
- **Divergent sequences or sequences with different domain structures, the approach does not produce optimal alignment**

Semi-global Alignment

- **Look for the best alignment between a short and long sequence**
- **Internal and terminal gaps are scored differently as terminal gaps can arise from difference in sequence lengths**
- **Two modifications are made to the basic dynamic programming algorithm**
    - **No penalties for the initial gaps – the first row and column of the table are signed with zeros**
    - **No penalties at the end gaps – Vertical moves in the last column and horizontal moves in the last row have no gap penalty.**

Example: **Find the optimal alignment for the sequences AGATCAT and GAT, match +1, mismatch 0 and gap -1**

Initialization Step

- **Create Matrix with M + 2 columns and N + 2 rows**
- **The first row and column of the table are assigned with zeros**

		A	G	A	T	C	A	T
	0	0	0	0	0	0	0	0
G	0							
A	0							
T	0							

Calculation of Scores

Match = +1, Mismatch = 0 and Gap = -1

		A	G	A	T	C	A	T
	0	0	0	0	0	0	0	0
G	0	0	1	0	0	0	0	0
A	0	1	0	2	1	0	1	0
T	0	0	1	1	3	3	3	3

Traceback Path

The optimal alignment score is the one with the highest score

		A	G	A	T	C	A	T
	0	0	0	0	0	0	0	0
G	0	0	1	0	0	0	0	0
A	0	1	0	2	1	0	1	0
T	0	0	1	1	3	3	3	3

It will produce alignment in this form (Optimal Semi-global Alignment)

```
A G A T C A T
  | | |
- G A T - - -
```

0 +1 +1 + 1+0 +0 +0 = 3

Optimal Alignment score = 3

Smith-Waterman Algorithm

- **A simple modification to the basic strategy allows the optimal local alignment to be found**
- **The path for this alignment does not need to reach the edges of the search graph but may begin and end internally**
- **An alignment would be locally optimal if its score cannot be improved either by increasing or decreasing the extent of the alignment**
- **The cumulative score for a path will decrease in regions of poorly matching sequences**
- **When the score drops to zero, extension of path is terminated and a new one can begin**
- **There can be many individual paths bounded by regions or poorly matching sequence, the one with the highest score is reported as the optimal local alignment**

- **Optimal methods always report the best alignment that can be achieved, even if it has no biological meaning**
- **Semiglobal alignment is not sufficient for searching best match subsequences**
- **Because mismatching positions and gaps outside target subsequences produce favorable score**
- **In addition to the 3 basic (vertical, horizontal and diagonal) movements, there is a forth option**
    - **Score zero if all other movements have negative**
- **Initialized the first row and column of the table with zeros**
- **Fill in all partial scores**
- **Find the maximum partial score and work backward until a zero score is reached**

Example: Find the optimal alignment for the sequences AGATCAT and AGA, match +1, mismatch 0 and gap -1

- **Create Matrix with M+2 columns and N+2 rows**
- **The first row and column of the table are signed with zeros**

		A	G	A	T	C	A	T
	0	0	0	0	0	0	0	0
A	0							
G	0							
A	0							

- **Calculation of Scores**
    - **Match = +1, Mismatch = 0, Gap = -1 or zero**

		A	G	A	T	C	A	T
	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	1	0
G	0	0	2	1	1	0	0	1
A	0	1	1	3	2	1	1	0

- **Traceback Path**
    - **Find the maximum partial score and work backward until a zero score is reached**

		A	G	A	T	C	A	T
	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	1	0
G	0	0	2	1	1	0	0	1
A	0	1	1	3	2	1	1	0

Optimal Local Alignment

```
A G A T C A T
  | | |
A G A - - - -
1 +1 +1 + 0+0 +0 +0 = 3
```

Optimal alignment score = 3

Substitution Scores and Gap Penalties

- **Increase sensitivity to weak alignments**
- **It is well known that certain amino acids can substitute easily for one another in related proteins, because of their similar physiochemical properties**
- **PAM and BLOSUM are the two most commonly used matrices for aligning protein sequences**

Point Accepted Mutation (PAM)

- **The PAM matrices involve the use of evolutionary models and extrapolation of probability values from alignment of close homolog to more divergent ones**
- **If there were no selection for fitness, the frequencies of each possible substitution would be primarily influenced by the overall frequencies of the different amino acids (called the background frequencies)**
- **In related proteins, the observed substitution frequencies (called target frequencies)**

PAM (1)

- **Dayhoff and coworkers were the first to explicitly use a log-odds approach, in which the substitution scores in the matrix are proportional to the natural log of the ratio of target frequencies to background frequencies.**
- **One PAM is a unit of evolutionary divergence in which 1% of the amino acids have been changed**

PAM (2)

- **This does not imply that after 100 PAMs every amino acid will be different; some positions may change several times, perhaps even reverting to the original amino acid, whereas others may not change at all**
- **When aligning sequences that are highly divergent, best results are obtained at higher PAM values, such as PAM200 or PAM250**
- **Matrices constructed from lower PAM values can be used if the sequences have a greater**

BLOSUM Substitution Matrices

- **Make use of a different strategy for estimating the target frequencies**
- **Contains local multiple alignments ("blocks") involving distantly related sequences (as opposed to the closely related sequences used for PAM).**
- **The underlying data are derived from the BLOCKS database**
- **It is advantageous to have data generated by direct observation, rather than extrapolation**

BLOSUM (1)

- Similar to PAM there is a numbered series of BLOSUM matrices
- The number in this case refers to the maximum level of identity that sequences may have and still contribute independently to the model
- The BLOSUM62 matrix, sequences having at least 62% identity are merged into a single sequence
- Substitution matrices have been constructed using higher cutoffs (up to BLOSUM90) for comparing very similar sequences and lower cutoffs (down to BLOSUM30) for highly divergent sequences

Comparing PAM and BLOSUM

- The two result in the same scoring outcome
- Use differing methodologies
- The two can be compared but due to the very different method of obtaining this score, a PAM100 does not equal a BLOSUM100
- BLOSUM directly look at mutations in motifs of related sequences while PAM's extrapolate evolutionary information based on closely related sequences

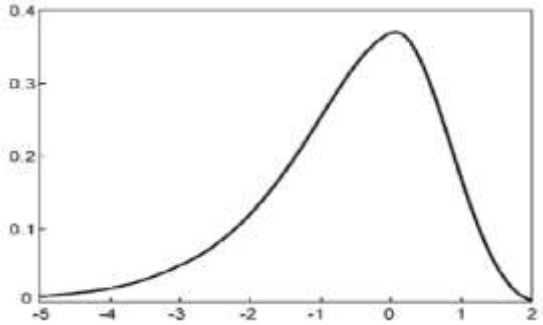
PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

Statistical Significance of Sequence Alignment

- When given a sequence alignment showing a certain degree of similarity, it is often important to ask whether the observed sequence alignment can occur by random chance or the alignment is indeed statistically sound
- The truly statistically significant sequence alignment will be able to provide evidence of homology between the sequences involved
- This problem requires a statistical test of the alignment scores of two unrelated sequence of the same length
- Calculating alignment scores of a large number of unrelated sequence pairs, a distribution model of the randomized sequence scores can be derived
- From the distribution, a statistical test can be performed based on the number of standard deviations from the average score

Gumble Distribution

Distribution of similarity scores assumes a peculiar shape that resembles a highly skewed normal distribution with a long tail on one side



- **The distribution matches the "Gumble extreme value distribution" $P=1-e^{-kmne-\lambda x}$**
- **Where m and n are the sequence lengths**
- **λ is a scaling factor for the scoring matrix used, and**
- **K is a constant that depends on the scoring matrix and gap penalty combination that is used**
- **X is the standard deviation of the distribution**

Evaluation of Pairwise Alignment

- **It is not known whether the Gumble distribution applies equally well to gapped alignments**
- **For all practical purposes, it is reasonable to assume that score for gapped alignments essentially fit the same distribution**
- **Sequence similarity value, by using the mathematical formula for the extreme distribution, the statistical significance can be accurately estimated**
- **The pool of alignment scores from the shuffled sequences is used to generate parameters for the extreme distribution**
- **The original alignment score is then compared against the distribution of random alignments to determine whether the score is beyond random chance**
- **If the score is located in the extreme margin of the distribution, that means that the alignment between the two sequences is unlikely due to random chance and is thus considered significant**
- **A *P-value* is given to indicate the probability that the original alignment is due to random chance**
- **A *P-value* resulting from the test provides a much more reliable indicator of possible homologous relationships than using percent identity values**

Interpretation of P-Value

- **It has been shown that if a P-value is smaller than 10^{-100}, it indicates an exact match between the two sequences**
- **The p-value is in the range of 10^{-50} to 10^{-100}, it is considered to be ea nearly identical match**
- **A P-value in the range of 10^{-5} to 10^{-50} is interpreted as sequences having clear homology**
- **A P-value in the range of 10^{-1} to 10^{-5} indicates possible distant homologues**
- **If P-value is larger than 10^{-1}, the two sequence may be randomly related**
- **However, the caveat is that sometimes truly related protein sequences may lack the statistical significance at the sequence level owing to fast divergence rates.**

MODULE 4 SESSION 3

SEQUENCE ALIGNMENT THEORY AND APPLICATIONS

BLAST

Learning Objectives and Outcomes

- **Understand the principles of the BLAST algorithms**
- **Understand the different BLAST algorithms, parameters and their applications**
- **Be able to adjust the sensitivity and specificity of BLAST searches**
- **Understand and evaluate BLAST results**

Database Similarity Searching

- **A main application of pairwise alignment is retrieving biological sequences in databases based on similarity**
- **This process involves submission of a query sequence with all individual sequences in a database**
- **Database similarity searching is pairwise alignment on a large scale**
- **Is one of the most effective ways to assign putative functions to newly determined sequences**

DNA/Protein Search

- **DNA is composed of 4 characters: A, G, C, T, at least 25% of the nucleotides of any 2 unrelated aligned sequences, would be identical**
- **Protein sequences is composed of 20 amino acids, the sensitivity of the comparison is improved**
- **It is accepted that convergence of proteins is rare, meaning high similarity between 2 proteins always means homology**
- **When comparing DNA sequences, we get significantly more random matches than we get with proteins**
- **The DNA databases are much larger, and grow faster than protein databases**
- **Bigger database means more random hits**
- **For DNA we usually use identity matrices, for protein more sensitive matrices like PAM and BLOSUM allow for better search results**

Database Search Requirements

- **Requirements for implementing algorithms for sequence database searching**
- **Sensitivity refers to the ability to find as many correct hits as possible**
- **It is measured by the extent of inclusion of correctly identified sequence members of the same family**
- **These correct hits are considered "true positives" in the database searching exercise**
- **The second criterion is selectivity, also called *Specificity***
- **Refers to the ability to exclude incorrect hits**
- **These incorrect hits are unrelated sequences mistakenly identified in database searching and are considered "false positives"**

- **The third criterion is *speed*, which is the time it takes to get results from database searches**
- **Depending on the size of the database, speed sometimes can be a primary concern**
- **Ideally, one wants to have the greatest sensitivity, selectivity and speed in database searchers**
- **However, satisfying all three requirements is difficult in reality**
- **What generally happens is that an increase in sensitivity is associated with decrease in selectivity**
- **A very inclusive search tends to include many false positives**
- **Similarly, an improvement in speed often comes at the cost of lowered sensitivity and selectivity**
- **A compromise between the three criteria often has to be made**

Heuristic Database Searching

- **Two major heuristic algorithms for performing database searches – BLAST and FASTA**
- **These methods are not guaranteed to find the optimal alignment or true homologous, but are 50-100 times faster than dynamic programming**
- **Programs can provide a reasonably good indication of sequence similarity by identifying similar sequence segments**
- **Both BLAST and FASTA use a heuristic word method for fast pairwise sequence alignment**
- **It works by finding short stretches of identical or nearly identical letters in two sequences**
- **These short stretches of characters are called *words***
- **The basic assumption is that two related sequences must have at least one word in common**

First identifying word matches, a longer alignment can be obtained by extending similarity regions form the words

BLAST

- BLAST – Basic Local Alignment Search Tool
- Is an approximate algorithm
- Finds the highest scoring local optimal alignment between query sequence and a database of sequences
- To discover all of the similar sequences in the database
- Create a statistical interpretation to enable the user distinguish a particular DNA or protein
- The sequence you give to blast is the query sequence
- Sequences similar to the query that blast returns are the hits or matches
- The database you search is the target database
- One innovation introduced in BLAST is the idea of neighborhood words
- Instead of requiring words to match exactly, a word hit is achieved if the word taken from the subject sequence has a score of at least T when a comparison is made using a substitution matrix to the word from the query
- This strategy allows the word size (W) to be kept high (for speed) without sacrificing sensitivity
- The first step is to create a list of words from the query sequence
- Each word is typically three residues for protein sequences and eleven residues for DNA sequences

1. Query sequence: AFRICANISED
2. Scan every 3 residues to be used in searching BLAST word database
3. Finds matches in the database

Query	CAN	CAN	CAN	CAN	CAN	…
Database	CTN	CAN	CWN	CAY	CCSA	…

4. Calculate sums of match scores based on BLOSUM62 matrix

Query	CAN	CAN	CAN	CAN	CAN	…
Database	CTN	CAN	CWT	CAY	CSA	…
Sum of score	12	18	6	12	6	

5. Find the database sequence corresponding to the best word match and extend alignment in both directions

6. Determine high scored segment above threshold

High Scoring segment pair, total score 42

7. The extension continues until the score of the alignment drops below a threshold due to mismatches

~~- Proteins = 22~~
~~- DNA = 20~~

Types of BLAST

- There are several variants of BLAST, each distinguished by the type of sequence (DNA or protein) of the query and database sequences
- BLASTP program compares a protein query to a protein database
- BLASTX compares a DNA query sequence to the protein database, which is useful for analyzing new sequence data and ESTs
- TBLASTN for a protein query against a nucleotide database
- TBLASTX takes DNA query and database sequences, translates them both, and compares them as protein sequences

Sequence Database Used with BLAST

- Non redundant – database provides comprehensive collection of both amino acid and nucleotide sequence data, with redundancy reduced by merging sequences that are completely identical
- SWISSPROT – The SWISS-PROT database
- 16S ribosomal RNA sequences
- Transcriptome short-gun assembly
- Human RefSeqGene sequences

Database Searching Artifacts

- A query sequence that contains repetitive elements is likely to produce many false and confounding database matches
- Likely artifacts would be finding hits to many proteins that seem to have no functional relationship to one another or hits to genomic sequences from many different chromosomes
- Both the query and the database are contaminated with foreign sequences from the same source, for instance, cloning vectors
- It is always good practice to practically evaluate database search results and be suspicious of artifacts when the data don't make sense
- A more proactive approach involves masking problematic sequences in the query before doing the search
- Both proteins and nucleic acids contain regions of biased composition, which can lead to confusing database search results

Low Complexity Regions (LCRs)

- In both protein and DAN sequences, there may be regions that contain highly repetitive residues

- **Estimates indicate that LCRs account for about 15% of the total protein sequences in public databases**
- **These elements in query sequences can cause spurious database matches and lead to artificially high alignment scores with unrelated sequences**
- **Alignment of LCR-containing sequences is problematic because they do not fit the model of residue by residue sequence conservation**
- **Methods for accessing the statistical significance of alignments are based on certain notions of randomness which LCRs do not obey**
- **The evolutionary, functional and structural properties of LCRs are not well understood**
- **LCRs arise by such mechanisms as:**
    - **Polymerase slippage**
    - **Biased nucleotide substitution**
    - **Unequal crossing over**
- **In proteins, LCRs are likely to exist structurally as non-globular regions**
- **Regions that have been defined physic-chemically as non-globular**

Masking of LCRs

- **To avoid the problem of high similarity scores owing to matching of LCRs that obscure the real similarities**
- **It is important to filter out the problematic regions in both the query and database sequences to improve the signal to noise ratio**
- **Masking… Two types of masking: hard masking and soft masking**
- **Hard Masking involves replacing LCR sequences with an ambiguity character such as N for…**
- **Nucleotide residues or X for amino acid residues**
- **The drawback is that matching scores with true homologs may be lowered because of shortened alignments**
- **Soft masking involves converting the problematic sequences to lower case letters, which are ignored in constructing the word dictionary**
- **But are used in word extension and optimization of alignments**

Blast Programs specific for each database

Blast Suite will indicate selected blast option on the website, paste sequences or enter accession number or browse the fasta file and upload

Provide a title if you wish

Align two or more sequences

There are options to exclude Modules with XM/XP which reference sequence or exclude environmental sample sequence

You can make header optimization like max target sequence e.g. 100, you can select expected threshold or size of the word (in nucleotide the default word size is 20), matches in a query range can also be set. You also have the opportunity to set the scoring parameters for match, mismatch and gaps.

You can do filtering and masking. E.g. filtering of Low complexity regions

BLAST OUTPUT

- **The graphical output includes colored horizontal bars and a diagonal in a two dimensional diagram showing the overall extent of matching between the two sequences**
- **The color coding of the horizontal bars corresponds to the ranking of similarities of the sequence hits**
- **Red = most related; Green and Blue = moderately related, Pink = highly related and Black = unrelated**
- **The lists of hits**
- **Database accession codes, name, description, general information about the hit**
- **Score in bits, the alignment score expressed in units of information**
- **Usually 30 bits are required for significance**
- **Expectation value E(), how many hits expect to find by chance with this score, when comparing this query to the database**
- **It is important to keep in mind that the E() value does not represent a measure of similarity between the two sequences**

❖ **Hit table is also given which list the various matches, the maximum score, total score, query cover (length of the match alignment), E value, the identity,. % of identity and accession numbers**

❖ **You also get a pairwise alignment between the query and the subject (score in bit, % of identity, % of similarity i.e. % of positives and you also get gaps.**

❖ **It is also possible to generate Lineage Report (or Taxonomy Report or Taxonomic Report of the result) indicating the different organism, number of hits, number of organism with description.**

❖ **It is also possible to construct Multiple Sequence Alignment from the blast output**

❖ **Also a Graphical report with black line at the top, showing the conserved domain, a click on the black line will redirect one to the conserve domains database providing more information about the conserved domain it matches with and also related domains in other organisms**

STATISTICAL SIGNIFICANCE

❖ **The significance scores help to distinguish evolutionary related sequences from unrelated ones**

❖ **The larger the database, the more unrelated sequence alignments there are**

❖ **E-value (E) indicates the probability that the resulting alignments from a database search are caused by random chance**

❖ **The significance scores help to distinguish evolutionary related sequences from unrelated ones**

❖ **The larger the database, the more unrelated sequence alignments there are**

❖ **E-value (E) indicates the probability that the resulting alignments from a database search are caused by random chance**

E-VALUE

E = M X N X P

- Where
  - M = the total number of residues in a database
  - N = the number of residues in the query sequence, and
  - P = the probability that an HSP alignment is a result of random chance

Example 1

- Find an E-value for aligning a query sequence of 100 residues to a database containing a total of 10^{15} residues results in a P-value for the ungapped HSP region in one of the database matches of 10^{-20}.

Answer:

Using E = m x n x p

- $E = 100 \times 10^{15} \times 10^{-20} = 10^{-3}$
- E = 1e-3 in BLAST output
- The E-value provides information about the likelihood that a given sequence match is purely by chance

SIGNIFICANCE OF E-VALUE

- The lower the E-value, the less likely the database match is a result of random chance
- E value less than 1e – 50 (or 1×10^{-50}), there should be an extremely high confidence that the database match is a result of homologous relationships
- E value between 0.01 and 1e-50, the match can be considered a result of homology
- E value between 0.01 and 10, the match is considered not significant, but may hint at a tentative remote homology relationship
- E value greater than 10, the sequences under consideration either unrelated or related by extremely distant relationships that fall below the limit of detection with the current method

WE MUST THINK! EEEEEEEEE!!!!!

- E = m x n x p
- E ∞ m, i.e. E-value is proportionally affected by the database size
- The problem is that as the database grows, the E-value for a given sequence match also increases

WEAKNESS OF E-VALUE

- The genuine evolutionary relationship between the two sequences remains constant
- The decrease in credibility of the sequence match as the database grows
- Means that one may "lose" previously detected homologs as the database enlarges
- Thus, an alternative to E-value calculation is needed

BIT SCORE

- Is another prominent statistical indicator used in addition to the E-value in a BLAST output
- The bit score measures sequence similarity independent of query sequence length and database size and is normalized based on the raw pairwise alignment score
- $S' = (\lambda \times S - \ln K)/ \ln 2$
- Where:

- λ is the Gumble distribution constant
- S is the raw alignment score, and
- K is a constant associated with the scoring matrix used

- Clearly, the bit score (S') is linearly related to the raw alignment score (S)
- Thus, the higher the bit score, the more highly significant the match is
- The bit score provides a constant statistical indicator for searching different databases of different sizes or for searching the same database at different times as the database enlarges

POSITION-SPECIFIC SCORING MATRICES (PSSM)

- In a standard substitution matrix, such as BLOSUM62 the substitution of one amino acid with another is associated with a single score
- An obvious simplification given that the same amino acid may have different conservation patterns in one context than another in accordance with differing roles in biological function
- Database searches can be tailored to find specific proteins families or domains through the use of substitution scores that reflect the substitution frequencies of each individual amino acid position in a domain
- In its simplest form, a PSSM consists of a set of 20 substitution scores at each position along the motif – one for each of the amino acids
- It is also possible to assign scores to insertions and deletions in a position-specific manner
- A commonly used software package, HMMER (Eddy et al., 1995), contains a set of related programs for constructing and using PSSMs
- Given a multiple alignment of several related proteins (e.g., one made using CLUSTAL W), the hmm build program may be used to calculate the position-specific scores and save it to a file (HMM file format)

HMM

- Using the HMMSEARCH program, the hmm file may be used as a query against a sequence database
- HMMPFAN is used to compare a single query sequence against a database of PSSMs (HMMs)
- The power of PSSMs in database searches can be further enhanced by iterative approaches in which the highest scoring matches in one search are incorporated into a PSSM used in successive searches

POSITION SPECIFIC ITERATED BLAST (PSI BLAST)

- Provides an automated facility for constructing, refining and searching PSSMs within the context of a single program
- Starting with a query sequence provided by the user, the process begins with a standard BLASTP search of a sequence database
- Highly significant alignments found in this search are then used to construct a PSSM on-the-fly
- Comparisons of the PSSM against the sequence database are performed using a variation for the word-based BLAST algorithm used for standard sequence comparisons
- The process continues until no new matches are found or a specified limit on number of iterations is reached

GAP-BLAST

- Use two hit approach
- Word can be followed by a second word within a certain gap threshold
- Matches are extended in both direction using a matrix in both directions until the score drops
- Weakness – performs dynamic programming at the end
- The best alignment may be lie outside of the range it has defined

Other Specialized BLASTS

- SmartBLAST, Primer-BLAST, GlobalAlign, CD-search, CDART, MOLE-BLAST, GEO, IgBLAST, VecScreen, Targeted Loci, Multiple Alignment, BioAssay etc

CONCLUSION

- Sequence alignment and database searching are performed tens of thousands of times per day by scientists around the world and represent critical techniques that all molecular biologists should be familiar with
- Described some of the fundamental concepts involved basic understanding of how the programs work so that parameters can be intelligently selected
- Aware of potential artifacts and know how to avoid them
- Important to apply the same powers of observation and critical evaluation that are used with any experimental method.

MODULE 5

Multiple Sequence Alignment

Session 1

Lec1: Building a Multiple Sequence Alignment

Learning Objectives

- Why Multiple Sequence Alignment is useful for scientists
- Identifying situations where multiple sequence alignments do not help
- Main criteria for building a multiple sequence alignment
- Main Applications of multiple sequence alignments
- What are the kinds of sequence you're looking for?
- Tips for naming sequences
- Tips for difficult MSA to interpret
- Comparing sequences you cannot align

In the coming lectures we will learn

1. Identifying the sequences you need to make a multiple sequence
2. Differences between Making a multiple sequence alignment with
   ClustalW (everybody uses it),
   MUSCLE (very fast)
   Tcoffee (accurate and combine sequences and structures)
   NB: Why you should use one tool over the other…

3. **Creating and comparing multiple sequence alignments with Comparing sequences you cannot align**

- **There are ways, multiple sequence alignments are to bioinformatics what Swiss knives are to MacGyver. Building multiple sequence alignments is far from an exact science**
- **In fact it's more than science, requiring that you use everything you know in bioinformatics and in biology"**
- **You can do many things with it, a lot of information can be derived but you need to know what information you are looking for.**
- **Don't work well for assembling the sequence pieces in a sequencing project**
- **If you want to turn an EST cluster into a gene sequence**
- **When the sequence you're interested in has no homologue (when you have no homologue for your gene) in any of the sequence databases (in this case you can use functional criteria and conducting a pattern search)**

BUILDING INFORMATIVE ALIGNMENTS

- **Gathering your sequences**
- **Compute a multiple sequence alignment**
- **Evaluate the quality of your alignment**
- **Interpreting your multiple sequence alignment**
- **Keep the sequences in save place for further analysis**

What we are looking for with MSA?

"The idea behind multiple alignment is to put amino acids or nucleotides in the same column because they're similar according to some criterion. You can use four major criteria for building a multiple alignment of sequences that all have different properties."

The criteria include:

1. **Structural Similarity**

Amino acids that play the same role in each structure are in the same column. Structure-superposition programs are the only ones that use this criterion. Means you looking for places with same amino acids, with structural meaning

2. **Evolutionary Similarity**

Amino acids or nucleotides related to the same amino (or nucleotide) in the common ancestor of all the sequences are put in the same column. No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it. (Homology for common ancestor)

3. **Functional Similarity**

Amino acids or nucleotides with the same function are in the same column. No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to

respect it – or you can edit your alignment manually. Looking for amino acids or nucleotides that have meaning i.e. functional meaning

4. Sequence Similarity

Amino acids in the same column are those that yield an alignment with maximum similarity. Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, their structural, evolutionary and functional similarities are equivalent to sequence similarity".

NB: Number 1, 2 and 3 above all depends on number 4 i.e. Sequence Similarity

Main Applications of Multiple Sequence Alignments

- Extrapolation

A good multiple alignments can help convince you that an uncharacterized sequence is really a member of a protein family. Alignments that include Swiss-Prot sequences are the most informative. Use the ExPASyBLAST server (at www.expasy.ch/tools/blast/) to gather and align them. For knowing if the sequence is a protein family.

- Phylogenetic Analysis

If you carefully choose the sequences you include in your analysis multiple alignment, you can reconstruct the history of these proteins. Use the Pasteur Phylip server at bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html

One of the tool called PHLIP is one the program you can use to build/make Phylogenetic analysis

- Pattern Identification

"By discovering very conserved positions, you can identify a identification region that is characteristic of a function. (in proteins or in nucleic acid sequences). Use the Weblogo server http://weblogo.berkeley.edu/logo.cgi". You can know what part of your sequence is characterized by functions.

- Domain Identification

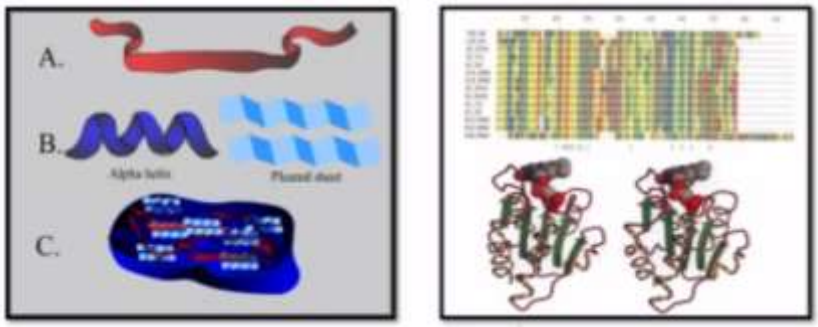
"It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain (PSSM). You can use this profile to scan databases for new members of the family. Use PROSITE (http://prosite.expasy.org/). For example you can find protein domain in a protein family.

- DNA regulatory elements

"You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potentially elements similar binding sites. Use the Gibbs Sampler to identify these sites: http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html"

- Structure Prediction

"A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for both proteins and RNA. Sometimes it can also help in the building of a 3-D model".



A good MSA will enable you extract good information about protein secondary structure, tertiary structure and build the 3-D model base on the MSA.

- **nsSNPS analysis**

"Various gene alleles often have different amino acids sequences. Multiple alignments can help you predict whether a Non-Synonymous Single-Nucleotide Polymorphism" is likely to be harmful or not. See the SIFT site for more details: http://sift.jcvi.org

- **PCR analysis**

"A good multiple alignment can help you identify the less degenerated portions of a protein family, in order to fish out new members by PCR (Polymerase Chain Reaction). If this is what you want to do, you can use the following site: blocks.fhcrc.org/codehop.html"

WHAT ARE THE KINDS OF SEQUENCES YOU'RE LOOKING FOR?

Always bear in mind that in evolution:

1 – Important amino acids (or nucleotides) are NOT allowed to mutate. For instance, active sites of enzymes are much conserved

2 – Less –important residues change more easily – sometimes randomly – and sometimes in order to adapt a function.

NB: When you look at a multiple sequence alignment and you find some parts of the multiple alignment that has many changes in that region, especially in the extreme edges, you have to know these are the less important region in the alignment. The most important region is the conserved region that is more stable.

TIPS OF NAMING SEQUENCES

- **Never use white spaces**
- **Do not use special symbols**
- **Never use names longer than 15 characters**
- **Never give the same name to two different sequences**

TIPS FOR DIFFICULT MSA TO INTERPRET

- **Remove Insertions/Deletions**
- **Redo MSA with the smaller set i.e. if you don't get any result**
- **Keep trimming to interpret i.e. to enable you get the alignment you are looking for**

NB: The above can spoil your alignment so you have to tackle them

ENHANCING ALIGNMENTS

- Remove Gaps
- Remove Extremities
- Keep Informative Blocks (You will build information on the informative blocks, depending on what you are looking for and what you want to do with multiple sequence alignment)

Try in your own time…

Searching Sequences on the ExPASy Server

Only to retrieve protein sequences in FASTA format

Example: Heat shock factor 1 (HSF1)

Choose http://www.uniprot.org/

SELECT THE SEQUENCES YOU WANT

This is the most delicate part of the process you can use the following guidelines

Select the top sequence

For a first analysis, you want to select ten sequences or fewer

Check it's similar to the query sequence – along its entire length

METHODS TO EXPORT YOUR SEQUENCES

- **FASTA: Generates a file that contains your sequences in FASTA format**
- **ClustalW, Tcoffee, and MAFFT: These are MSA packages running on the EMBnet server**
- **Reduce Redundancy: This option will extract the most meaningful sequences from your dataset**
- **PRATT: Will search for conserved motifs in your sequences without aligning them**

PRACTICAL

Try in your own time…

- **Go to https://www.expasy.org/proteomics**
- **Search for HSF1 (Heat Shock Factor 1)**
- **Click on (UniProtKB)**
- **Retrieve your protein sequences (e.g. Heat shock Factor1 "HSF1") from different organism**
- **This will take you to http://www.uniprot.org/uniprot/?query=HSF1&sort=score**
- **Select your organism (Human, Rat, Mouse, Arabidopsis, Chicken, Pig)**
- **Click Download (Download Selected) then (Go)**
- **Save it in FASTA format in one text file**
- **Align the sequences using Clustal Omega**
- **Checking the gene-based phylogenetics tree**
- **Add one more sequence NOT related sequence (Out Group)**
- **Checking the change on the gene-based phylogenetics tree**

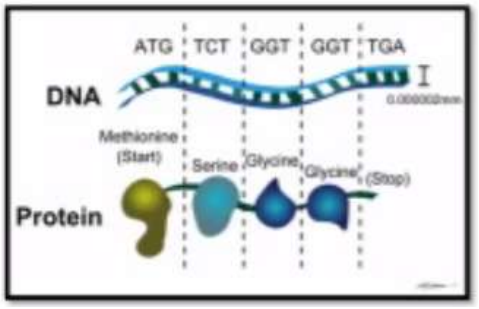
MSA CHOOSING THE RIGHT SEQUENCES WITH RIGHT SOFTWARE

Before you build your alignment, you must carefully choose the sequences you want to align. These sequences are members of the same protein family, and they all share a common ancestor. The family

is usually too large to be entirely included in your multiple alignment and picking the right sequences is an art.

A FEW GUIDELINES FOR SELECTING SEQUENCES

- **Use proteins whenever possible. Most Bioinformaticians prefer proteins. You can turn them back into DNA after doing the multiple alignment.**



- **If the sequences are non-coding sequences, you must use DNA**
- **Many Sequences Try to avoid using a lot sequences: Start with 10 – 15 sequences; avoid aligning more than 50 sequences except if you are looking for that**
- **Don't try to have very different sequences. Sequences that are less than 30 percent identical to more than half the other sequences in the set often cause troubles**
- **Identical sequences also they never help. Unless you have a very good reason to do so, avoid incorporating into your multiple alignment any sequence that's more than 90 percent identical to another sequence in the set**
- **Avoid having partial sequences. Multiple sequence alignment programs prefer sequences that are roughly the same length. Programs often have difficulties comparing items in a mixture of complete sequences and shorter fragments**
- **Try to avoid sequences that have repeated domains unless you don't have any other choice. Sequences with repeated domains cause trouble for most multiple alignment programs – especially if the number of domains in different. When this happens, you may be better off extracting the domains yourself with Dotlet or Lalign and making a multiple alignment of those segments**
- **If you still cannot generate a proper alignment from sequences that you know are related, you could use a local multiple alignment method, such as the Gibbs Sampler, or a pattern extraction motif such as Pratt. They can both help in better motif extraction.**

For carrying out a phylogenetic analysis on a set of coding DNA sequences, do the following

1. **Translate your DNA sequences into proteins**
2. **Perform the multiple alignments on the proteins**
3. **Thread them DNA back onto the protein multiple sequence alignment framework using pal2nal (coot.embl.de/pal2nal) or Protogene if you do not have the original DNA sequence (www.tcoffee.org)**

Choosing the Right Number of Sequences

- You should start with a relatively small number of sequences between 10 and 15 sequences would be suitable for most cases
- After you get something interesting happening with this small set, you can always increase its size
- It's hard to see any reason for generating multiple alignment with more than 50 sequences, unless you're interested in building some extensive phylogenetic tree

Why you should not use too much sequence to align?

- Computing big alignments is difficult
- Building big alignments is difficult
- Displaying big alignments is difficult
- Using big alignments is difficult
- Making accurate big alignments is difficult

Should you choose sequences that are very similar or very different?

- Make the right compromise between similarity and new information (So try to compromise by choosing sequences that are not very similar and not very different sequences)
- An alignment that only contains very similar sequences brings little information
- You can use it to extrapolate annotations, but you can't do phylogeny, structure prediction, function prediction, or any of the other useful applications that we mentioned before

GENERAL RULE – The general rule is that you want them to be as distantly related as possible – without requiring too many gaps in order to be properly aligned.

Two Things Multiple Sequence Alignment Programs really don't like are:

1. Sequences that are very different from every other sequence in the group
2. Sequences that need long insertions/deletions to be properly aligned

Gathering your sequences with online BLAST servers

- Characterized Sequences: Try to include sequences with good annotations and experimental information in your alignment because they bring information with them – and also allow feature propagation (Blasting this will get you more sequence on what you are looking for)
- Uncharacterized Sequences: including them in your multiple alignment is to distinguish between the conserved positions that cannot mutate and the other, less-important columns. They help in getting some contrast on your sequence of interest

NB: Sequences that are so similar to the query are probably homologous. We commonly refer to such sequences as hits or matches. (When blast is done, when there is similarity between our sequence and other sequences from the blast result, we simply try to get these sequences (hits or matches) to use them in building multiple sequence alignment because they are probably homologues)

Choosing the Right Method of Multiple Sequence Alignment

1. **ClustalOmega**

   The most commonly used multiple sequence alignment package, Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile (Markov Profile) techniques to generate alignments between three or more sequences.

2. **Tcoffee**

   One of the latest multiple sequence alignment packages that you can use. With Tcoffee, you can combine sequences and structures, evaluate an alignment, or merge several alternative multiple alignments into a single unified result.

3. **MUSCLE**

   One of the fastest and famous alignment methods around for aligning large set of sequences.

Aligning Sequences and Structures with Tcoffee

More accurate alignments at the cost of slightly longer running time

http://tcoffee.org.cat/

It can deal with and align DNA, RNA and Proteins

M-Coffee is another one that aligns DNA, RNA or Proteins and combining the output of other popular aligner, so it can combine DNA, RNA or Protein.

Regarding RNA Alignment…

R-Coffee aligns RNA sequences using predicted secondary structures

SARA-Coffee aligns RNA sequences using predicted tertiary structures (it is a new package)

Regarding Protein Alignment…

Expresso aligns proteins sequences using structural information, in this case it won't just use the sequence but also the embedded structure information with more accurate result

PSI-Coffee aligns just related proteins using homology extension, it may be small but accurate. NB: Just related proteins are used here

TM-Coffee is a very special that aligns trans-membrane protein using homology extension (also new)

Pro-Coffee aligns homologous promoters, now we are taking about the promoter region (i.e. regulatory element of the gene) and it is very special as it aligns promoter regions of genes

Accurate is another software that automatically compiles most accurate modules of DNA, RNA and Proteins (accurate meaning experimental)

Combine this tool that combines two more multiple sequence alignment into single one. It is very useful for lots of bioinformaticians and biologists.

EVALUATION

- **Give TCOFFEE a multiple alignment that you generated with your favorite method and you can ask Tcoffee to evaluate the quality of this multiple sequence alignment for you.**
- **You can use any of the most common formats (MSF, ALN, FASTA and PIR).**

Using…

CORE package in TCOFEE: it evaluates your alignment and output result

iRMSD-APDB Is another package in TCOFFEE that evaluates multiple sequence alignment using structural information

T-RMSD Is another package in TCOFFEE that allows five grained structural clustering of a given group of related protein domains, It also allows evaluation and cluster

Strike Is another package in TCOFFEE that evaluates proteins multiple sequence alignments using a single 3D structure, this one is very important software algorithm that is recently included in TCOFFEE.

ADVANCED (there are many advance depending on what you are looking for…)

T-Coffee Advanced can be used

T-Coffee results appears in colored keys, to evaluate the most good part of the alignment which is red, average will be yellow or orange and the worst or bad part of the alignment which is blue of green

CRUNCHING LARGE DATASETS WITH MUSCLE

Can be used when you have large datasets

MUSCLE is a newcomer in the multiple sequence alignment arena – but it is a remarkable efficient package for making fast, high-quality multiple sequence alignments. MUSCLE is ideal if you want to align several hundred sequences.

- **MUSCLE can also be found in the EBI website, under the Sequence Alignment Section**
- **MUSCLE has steps: 1. Alignments, 2. Result Summary, 3. Submission Details, 4. Submit Another Job**

PRACTICAL

Try in your own time…

4. **For your set of HSF1 FASTA sequences, what are the right TCOFFEE flavors to align them?**
5. **For your set of HSF1 aligned sequences, use the right TCOFFEE flavor to evaluate your alignment?**
6. **Calculate the timing for Aligning your sequence collection using MUSCUL, TCOFFEE and Clustal Omega and order them based on their speed?**

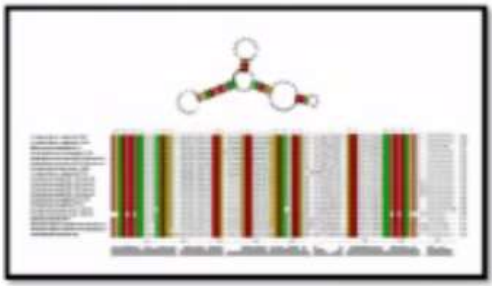
COLORING MSA FOR PUBLICATION

After successfully making your multiple sequence alignment, you will like to share it to the world, therefore it should be colored to convince people what kind of result you are showing

Comparing your multiple sequence alignment for publication

You want to show this alignment to your colleagues, you want to include it in publications

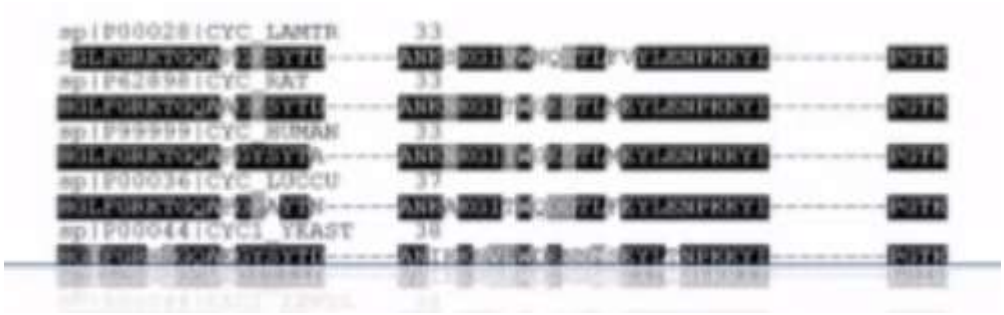
In short, you need a high-impact picture to convince people that your research is going well



You can make use of a very useful tool…

Using BOXSHADE (It is not colorful; it is just black and white, i.e. grayscale)

Boxshade is a utility that allows you to put some life into your alignment. It shades columns according to their level of conservation and produces files that you can easily manipulate for inclusion in reports or articles. In black and white with grayscales.



Just launch google.com search for BoxShade Server, you will find it easily in ExPASy. Then put your alignments there from different format e.g. Clustal, it accepts many formats like MSF, ClustalW etc.

NB: When pasting MSF or ClustalW files, please make sure that the pasted test starts with the header like of the alignment and contains no extra black lines at the bottom.

After putting your alignment, choose the alignment font size, format of the picture output e.g. RGF, whether or not to present the consensus sequence in the picture, you can also choose the input format. Fraction of sequence e.g. 0.5 means you want half of the sequence to be conserved before shading to occur. Finally click on Run BOXSHADE to get the result output.

Conservation doesn't necessarily mean identity in Boxshade. Similar residues, such as isoleucine and valine, also account for conservation.

Two types of shading exist:

- **Black: Identical amino acids or nucleotides**
- **Gray: Similar amino acids**

Some people may wish to have more colors in their publications to show different parts…

MView for colorful Multiple Sequence Alignment tool

JalView is more powerful tool

PRACTICAL

Try in your own time…

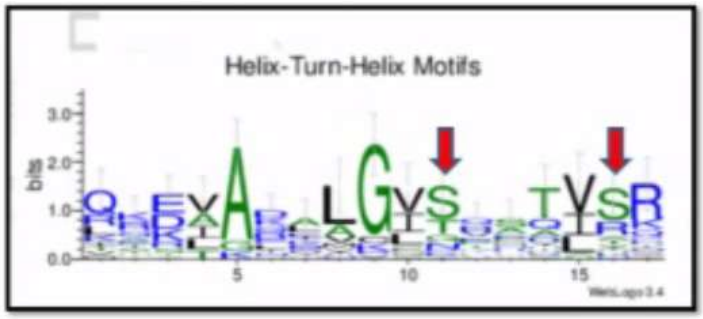
1. **Point your browser to www.ch.embnet.org/software/BOX_form.html**
2. **Choose RTF_new from the Output Format drop down menu**

3. Select the font size you want
4. Choose Add a Consensus Line with Letters from the Consensus Line drop down menu
5. Select the fraction of sequences you would like shaded
6. Select the format of the multiple sequence you want to use
7. Paste your multiple sequence alignment into the sequence window
8. Click the run Boxshade Button
9. Click the here is your output link, save to a local file then open the local file with MS-Word

INTERPRETING YOUR MULTIPLE SEQUENCE ALIGNMENT USING LOGOS

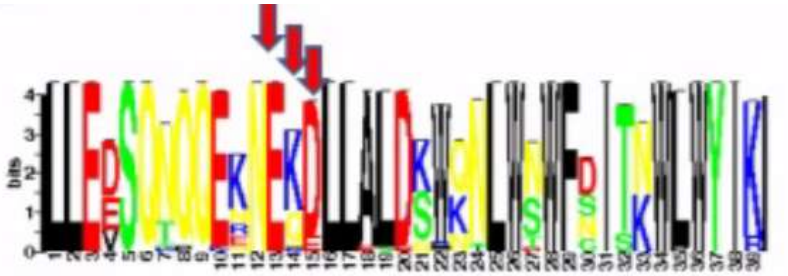
Logos is one good representation of your multiple sequence alignment

- Logos are a terrific way to generate high-impact pictures from MSA
- Logo Figure is a representation of the alignment
- Notice how the conserved amino acids (e.g. cysteines) stick out, indicating regions of potential biological importance



When looking at a sequence logo, you can consider the following elements:

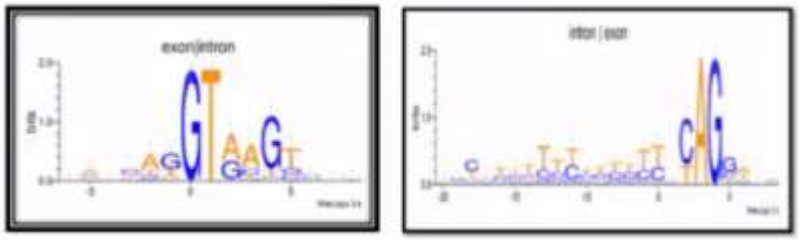
- Each position corresponds to a column in the multiple alignment
- The total height of a logo position depends on the degree of conservation in the corresponding multiple alignment column



The small or little height means it is less conserved

- Very conserved alignment columns give you high logo positions
- Positions that contain a very heterogeneous mixture of symbols yield low logo positions
- The size of each letter in a logo position depends on how frequent this letter is in the column
- The top letter is always the most frequent in the column
- Logos make sense only if you have a nice block with a few highly conserved positions surrounded by highly degenerated positions (you can extract more and better information from the logos)

- **There is a handy utility on the Web that identifies blocks within your multiple alignments and turns each of them into a logo**
  - **blocks.fhcrc.org/blocks/process_blocks.html**

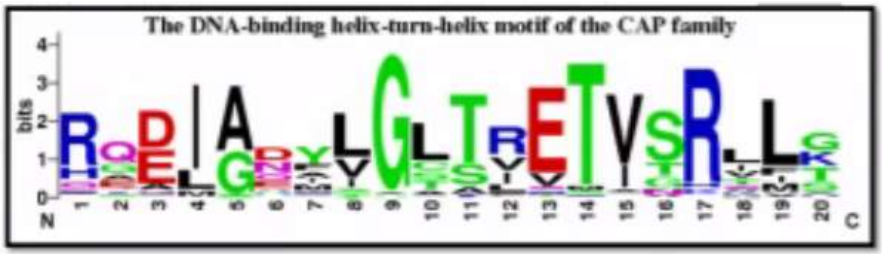


  - **WebLogo is a web-based application designed to make the generation of sequence logos easy and painless. Weblogo has been featured in over 4000 scientific publications**

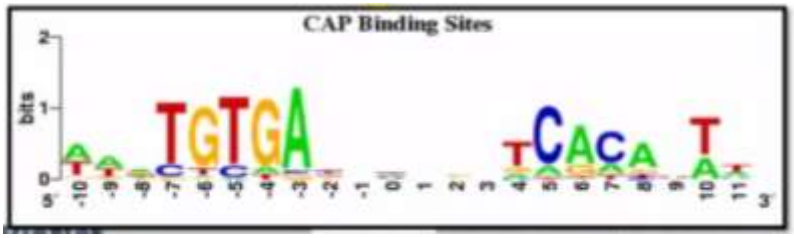
For example in…

CATABOLITE ACTIVATOR PROTEIN (CAP)

The helix-turn-helix motif from the CAP family of homodimeric DNA binding proteins. CAP (Catebolite Activator Protein, also known as CRP for cAMP Receptor Protein) is a transcription promoter that binds at more than 100 sites within the E.coli genome. Residues 1-7 form the first helix, 8-11 the turn and 12-20 form the DNA recognition helix. The glycine at position 9 appears to be critical in forming the turn. Positions 4, 8, 10, 15 and 19 are partially or completely buried, and therefore tend to be populated by hydrophobic amino acids, which are colored black. Positions 11-14, 17 and 20 interact directly with bases in the major groove and are critical to the sequence specific binding of the protein. The data for this logo consists of 100 sequences from the full Pfram alignment of this family (Accession number PF00325). A few sequences with rare insertions were removed for convenience.



The two DNA recognition helixes of the CAP dimer insert themselves into consecutive turns of the major groove. Several consequences can be observed in this CAP binding site logo. The logo is approximately palindromic, which provides two very similar recognition sites, one for each subunit of the dimer. However, the binding site is not perfectly symmetric, possible due to the inherent asymmetry of the operon promoter region. The displacement of the two parts of 11 base pairs, or approximately one full turn of the DNA helix. Additional interactions between the protein and the first and last two bases occur within the DNA minor groove, where it is difficult for the protein to distinguish A from T, or G from C\cite{Seeman76}. The data for this logo consists of 59 binding sites determined by DNA footprinting. *Robison, K. McGuire, A. M. Church, G. M. A. comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K12 genome. Journal of Molecular Biology (1998) 284, 241 - 254.*

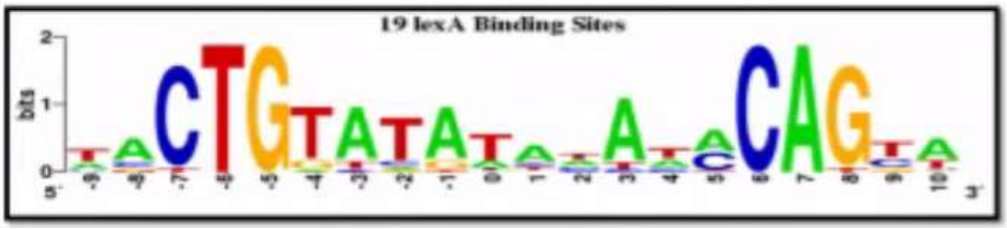


E. Coli Transcription Factor Binding Sites

The following logos (along with the CAP logo above) display a selection of E. Coli transcription factor binding sites determined by DNA footprinting. This data has been collated in the DPInteract database and has been used to search for additional binding sites within the E. coli genome.

- **LexA repressor is closely related to CAP, and has similar DNA protein interactions**
- **H-NS: Histone like, nucleoid-associated DNA-binding protein**
- **DNA biosynthesis initiation binding protein**
- **Arginine Repressor**

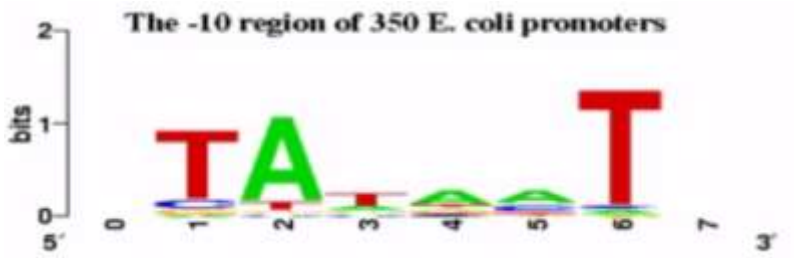
Robison, K. McGuire, A. M. Church, G. M. A. comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K12 genome. Journal of Molecular Biology (1998) 284, 241 - 254.



E. Coli Promoters (Transcription Start Signals)

In prokaryotes the DNA sequence just upstream of the transcription start point contains two important conserved regions. The first such region is centered at around 35bp upstream and is involved in the initial recognition of the gene by RNA polymerase.

The second region, sometimes referred to as the Pribnow box, is centered at about 10bp upstream. The typical separation between the -35 and -10 sites is 15-18bp. See baseflip: Strong Minor Groove Base Conservation in Sequence Logos implies DNA Distortion or Base Flipping during Replication and Transcription initiation for more information.



Other examples…

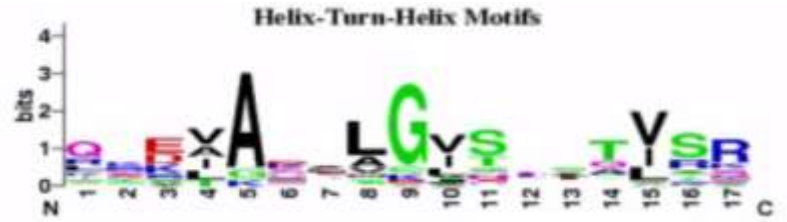
Globins

The end of the B helix through the beginning of the D helix of 34 globins

Prenyltransferases (motif A)

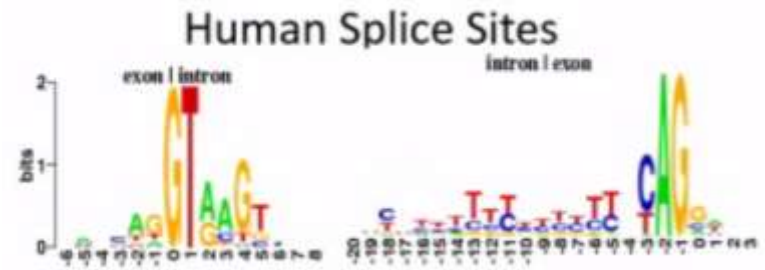
Here is an alignment found by the Gibbs sampling system. Both the identified site and some context are shown. Not those spaces are significant, so that the spaces included below (to aid identification of the site) will end up being considered amino acid positions.

HTH Proteins



Helix-Turn-Helix DNA binding motifs found by the Gibbs Sampling system. Compared to the CAP HTH logo there is much less sequence conservation within the DNA binding helix (11-17), as might be expected for a diverse sample of proteins.

Human Splice Sites



These logos show a small sample of Human intron-exon splice boundaries. Sequences of experimentally confirmed genes were extracted from EID: the Exon-Intron database. Additional discussion of the features in this logo can be found in the paper Features of spliceosome evolution…

- **Exon-Intron (Donor) Sites**
- **Edit Logo Intron-Exon (Acceptor) Sites**

Try in your own time…

PRACTICALS

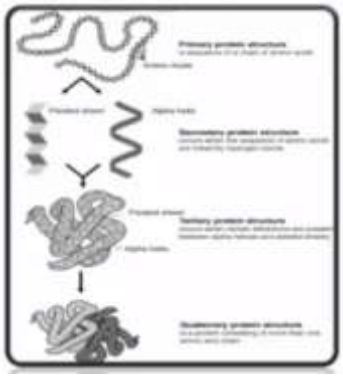
- **Interpret the conserved amino acids in your alignments using logos**

Identify the potential sites that have biological meanings try it with the HSF1

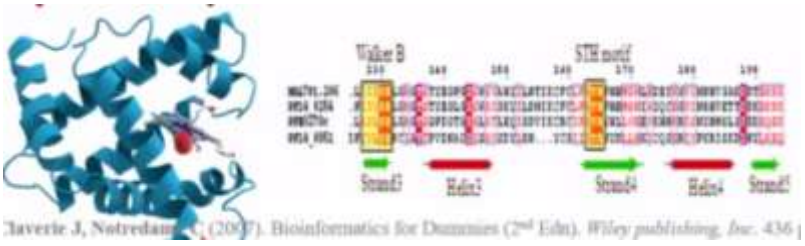
INTERPRETING YOUR MULTIPLE SEQUENCE ALIGNMENT

The interpretation of a multiple alignment depends very much on its appearance. Some tools on the Net can help you make sense of your multiple alignments by extracting blocks or singling out special positions.

- **Interpreting an alignment is a bit of an art**
- **E-values (the scores that tell you how reliable your database search is)**
- **That means deciding whether your alignment is correct still involves some educated guesswork.**
- **DNA alignments are by far the most difficult to interpret, simply because you only have 4 letters are these letters are the ATGC**
- **If you're analyzing this type of sequence, you want a very high level of conservation, knowing that single conserved columns are likely to be meaningless.**
- **A DNA block is only informative when it contains several identical columns in a cluster of several colors**
- **Even with the DNA of closely related sequences, obtaining such an alignment is still difficult**
- **This is why most biologists prefer protein alignments**
- **The most convincing evaluative grid we have for a protein multiple alignment stems from out knowledge of protein structures. NB: You need to know about primary, secondary and tertiary structures of protein. The difference between them and what kind of amino acids (features) that involves in each structure; that will help to know which amino acid is conserved to give the structure.**

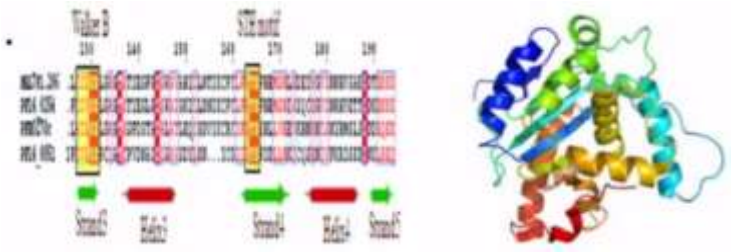


- **We know that structures contain surface loops that evolve rapidly. (Loops are softer portions of the protein that connect its more rigid portions).**

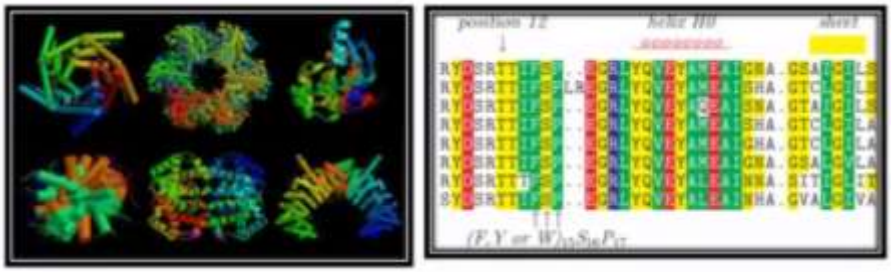


Javerie J, Notredam … (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc.* 436 |

- Protein structures also contain core regions that act as support walls for the protein. These support walls evolve less rapidly than the loops on the surface



- In your multiple alignment, to find a nice conservation you can expect to find nice, gap-free blocks that correspond to the core regions – and gap rich regions that corresponds to the loops



CABALISTIC SIGNS

The last line contains seemingly ClustalW, MUSCLE or Tcoffee alignment, cabalistic signs such as (*), (:), or (.).

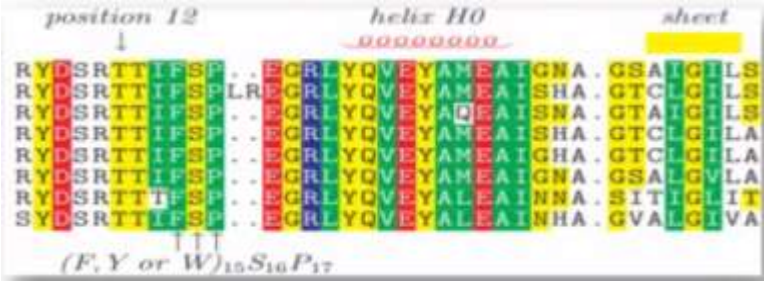
- (*) A star indicates an entirely conserved column
- (:) A colon indicates columns where all the residues have roughly the same size and the same hydropathy (Also a good indicator)
- (.) A period indicates columns where the size OR the hydropathy has been preserved in the course of evolution (less conserved)

The average good block is:

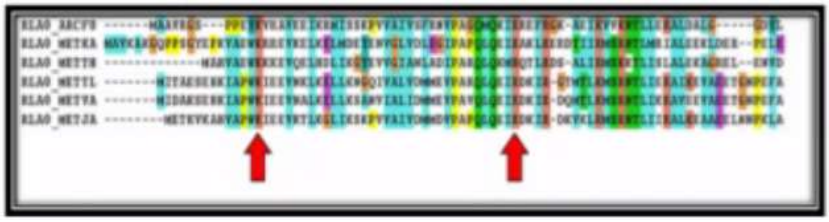
A unit at least 10-30 amino acids long, exhibiting at least one to three stars(*), a few more colons (:) close to the stars, and a several periods (.) scattered along the MSA result

So a block that has the above criteria is a good block

- **The magic thing about multiple sequence alignment is that 4 or 5 conserved positions over 50 amino acids can be enough to convince us that we're looking at a genuine signal. This is less than 10 percent identity!**
- **You have to remember that we require at least 25 percent identity to consider a pairwise alignment**



Conserved Columns in a multiple sequence alignment are meaningful only when the surrounding columns are not conserved



Therefore a conserved region is surrounded by non-conserved columns

Another criterion for a useful multiple alignment is knowing the type of amino acids you can expect to see conserved.

What type of amino acid you need to expect or see to assure it is conserved…

Amino acids aren't equal and they all have very characteristics patterns of mutation/conservation in a multiple sequence alignment. That is why soma amino acids can be replaced by other amino acids, and some amino acids cannot be replaced by others. If this replacement happens it will really make a big difference and big change in the characteristic it will exhibit, structural and functional of the protein (a problem for the protein).

W(tryptophans), F(phenylalanine), Y(tyrosine)

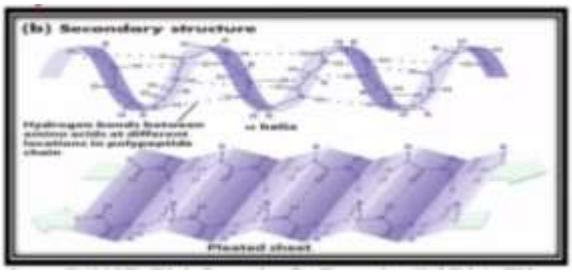
It is common to find conserved tryptophans. Tryptophan is a large hydrophobic residue that sits deep in the core of proteins. It plays an important role in their stability and it therefore difficult to mutate. When tryptophan mutates, it is usually replaced by another aromatic amino acid, such as phenylalanine or tyrosine.

Patterns of conserved aromatic amino acids constitute the most common signatures for recognizing protein domains.

Another case…

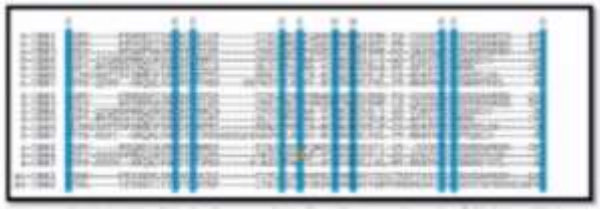
G(glycine), P(proline)

It is common to find conserved columns with a glycine or a proline in a multiple alignment. These two amino acids often coincide with the extremities of well-structured beta stands or alpha helices



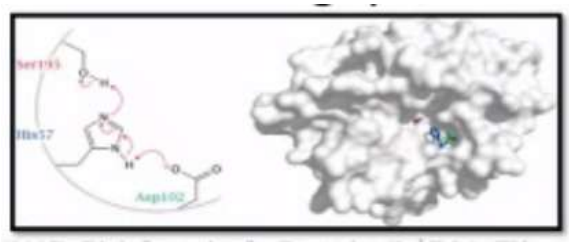
C(cysteines)

Cysteines are famous for making C-C (disulphide) bridges. Conserved columns of cysteines are rater common and usually indicate such bridges. Columns of conserved cysteines with a specific distance provide a useful signature for recognizing protein domains and folds.



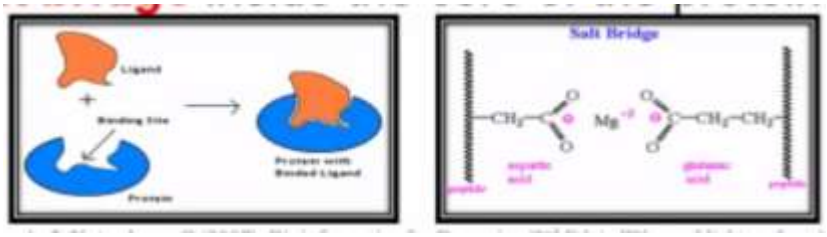
H(Histidine), S(Serine)

Histidine and serine are often involved in catalytic sites, especially those of proteases. Conserved histidine or a conserved serine are good candidates for being part of an active site.



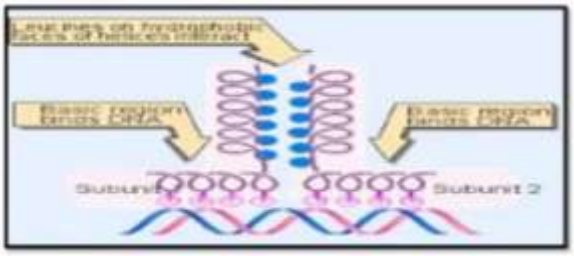
K(Lysine), R(Arginine), D(Aspartic Acid), E(Glutamic Acid)

These charged amino acids are often involved in ligand bindng. Highly conserved columns can also indicate a salt bridge inside the core of the protein.



L(Leucines)

Leucines are rarely very conserved unless they are involved in protein-protein interactions such as a leucine zipper.



JALVIEW

Installation

Jalview can make Multiple Sequence Alignment and many other things too

- **When you load sequence in Jalview, your sequence does not travel over the internet, it stays in your computer**
- **If you want to ensure that none of your data travels across the internet, choose the file to work offline option on your browser as soon as Jalview is loaded**
- **Do not load confidential sequences in Jalview before doing this. The web interface is NOT secure.**

Jalview is a multiple sequence alignment viewer, editor and analysis tool. Jalview is designed to be platform independent (running on Mac, MS Windows, Linux and any other platform that supports Java). The good news is that Jalview is free.

Jalview is capable of editing and analyzing large alignments (thousands of sequences) with minimal degradation in performance, and able to show multiple integrated views of the alignment and other data. Clustal Omega, ClustalW and MUSCLE are all embedded in it.

Jalview can read and write many common sequence formats including FASTA, Clustal, MSF(GCG) and PIR. It has many integrated views.

JALVIEW HISTORY

Jalview was initially developed in 1996 by Michele Clamp, James Cuff, Steve Searle and Geoff Barton at the University of Oxford and then the European Bioinformatics Institute

Development of Javliew 2 was made possible with eScience funding from the BBSRC8 in 2004, enabling Andrew Waterhouse and Jim Procter to re-engineer the original program to introduce contemporary developments in bioinformatics and take advantage of the latest web and Java technology.

When you open Jalview it gives an example of some sequences which can enable see you all the 5 different windows views of Jalview. In view menu you can choose to show or hide annotation. You can also check consensus regions, you can also draw phylogenetic tree in Jalview.

The example project that displays on startup, you can disable in preferences

Demonstrates key features of Jalview

Practical

Try in your own time…

Download and Install Jalview

Editing and Coloring Using Jalview

The learning objectives

1. **Import sequence or alignment**
    - **Drag/drop, Paste or URL**
2. **Decorate sequences with references and annotation from external databases**
3. **Create Alignment**
4. **Use built in shading, conservation analysis, tree and PCA capabilities to explore**
    - **Also use annotation and structure data if available**
5. **Select regions for refinement or further analysis**
6. **Import trees, annotation, etc created with other programs to explore further**
7. **Prepare annotated views for publication**

Learning Outcomes

- **Fetch DB references for your sequences**
- **Set Feature setting**
- **Change the view**
- **Using Jalview to Edit your Multiple Alignments**
- **Introducing and removing gaps**
- **Some Useful Features of Jalview**
- **Navigation in Jalview using Cursor mode**
- **Different ways to load your own sequences**
- **Insert your Alignment image into your document for Publication**
- **Jalview Project File**
- **Select, edit, exporting and save regions in an alignment**

Jalview projects store key data for a session

You can do the following in Jalview:

- **Alignments**
- **Annotation & Database IDs**
- **Structure & Jmol sessions**
- **Trees**
- **Display Settings**
- **Shareable archive of your results**
    - **New Jalview versions are tested for backwards compatibility**

Jalview has the ability to go online and fetch Sequences from different Databases using the web service menu of Jalview to select the database to use/fetch sequences from.

Another good thing in Jalview is sequence feature, from view menu you can select feature settings and click on what you are looking for and give it a color. You can give every feature in your protein sequence specific colors. You can color you Multiple Sequence Alignment in Jalview, very nice that you can include in your publication.

Click on Format and Click on Wrap, to have all the alignments in one square image.
You can also decide to view or not view the sequence ID

Why we use Jalview

- **Because Jalview is a multiple sequence aligning and editing multiple sequence alignment is so complicated, biologists have developed text editors that are specific for multiple sequence alignment.**
- **Sequence reordering is simple – Highlight the sequences to be moved then press the up or down arrow keys as appropriate.**
- **Simply you can press any sequence, if you wish to move a sequence up past several other sequences, it is often quicker to select the group past which you want to move it and then move the group.**

Exercise: Recording the alignment

Try in your own time…

- **Open an alignment (e.g. the PFAM domain PF03460). Select one sequence. Using the up and down arrow keys, alter its position in the alignment. Note that this will not work in cursor mode.**
- **Hold Ctrl and select two sequences separated by one or more un-selected sequences**
- **Note how multiple sequences are grouped together when they are re-ordered using the up and down arrow keys**

NB: Moving from one sequence to the other using the arrow keys

CLUSTALX COLOR SCHEME

Choosing a color scheme is really a matter of taste; most multiple sequence alignment specialists can spend hours explaining to you why they prefer a special color scheme to every other alternative in the galaxy.

What is the color scheme in this alignment?

To know more about the color schemes available in Jalview, read the excellent online documentation for this program at www.jalview.org/help.html; you will find an extensive section on color schemes there.

In Jalview, all the available color schemes are under the Color menu

The clustalX, initially designed by Tobby Gibson at the European Molecular Biology Laboratory, is probably one of the most popular

HIDING REGIONS

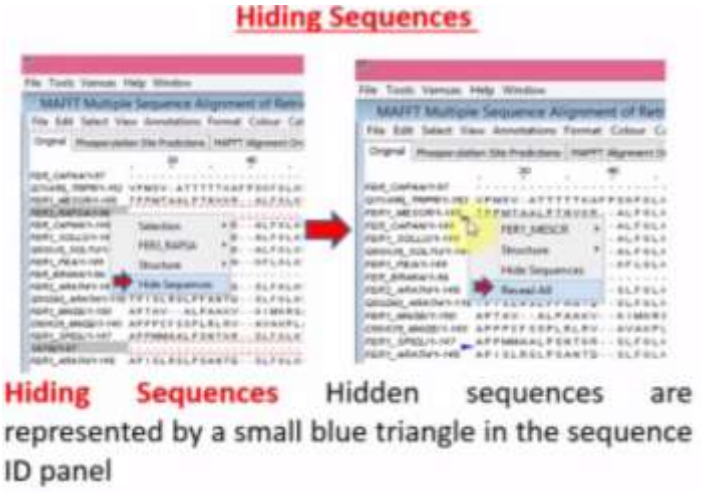
It is sometimes convenient to exclude some sequences or residues in the alignment without actually deleting them. You can hide some regions that you don't like to see;

Jalview allows sequences or alignment columns within a view to be hidden and this facility has been used to create the several different views in the example alignment file that is loaded when Jalview is first started.

E.g. Extremities of protein sequence you don't want to see and you like to show only the domain, you can select the extremities part with mouse and just hide it.

To hide a set of sequences:

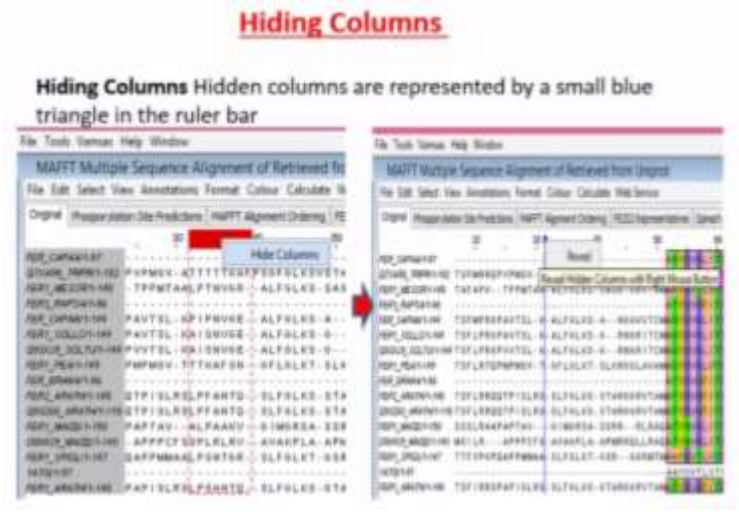
1. **Select them and right-click the mouse on the selected sequence IDs to bring up the context menu**
2. **Select Hide Sequences and the sequences will be concealed, with a small blue triangle indicating their position**
3. **To unhide (reveal) the sequences, right click on the triangle and select Reveal Sequences from the context menu.**



Similar Mechanism Applies to Columns (Hiding Columns)

Selected columns (indicated by a red marker) can be hidden and revealed in the same way via the context menu by right clicking on the ruler bar.

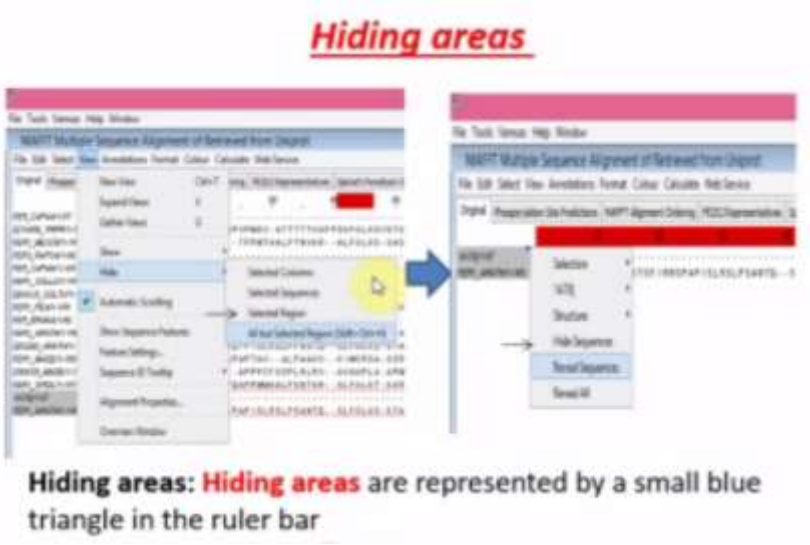
The hidden column selection is indicated by a small blue triangle in the ruler bar



It is often easier to select the region that you intend to work with, rather than the regions that you want to hide. Been this case, select the required region and use the

View → Hide → All but selected region menu entry OR

Press Shift + Ctrl + H to hide the unselected region.



REPRESENTING A GROUP WITH A SINGLE SEQUENCE

Instead of hiding a group completely, it is sometimes useful to work with just one representative sequence. The <Sequence ID> → Represent Group with <Sequence ID> option from the sequence ID pop-up menu enables this variant of the hidden groups function.

The remaining representative sequence can be visualized and manipulated like any other. However, any alignment edits that affect the sequence will also affect the whole sequence group.

NB: Select a group with a mouse, right click, click selection, click edit group, finally, click edit name and description of current group, you can name it anything you prefer.

INTRODUCING AND REMOVING GAPS

The alignment view provides an interactive editing interface, allowing gaps to be inserted or deleted to the left of any position in a sequence or sequence group

Alignment editing can only be performed whilst in keyboard editing mode (entered by pressing F2) or by clicking and dragging residues with the mouse when SHIFT or CTRL is held down (which differs from earlier versions of Jalview). NB: You can Cut, Copy, Edit Sequence and Change Case to UPPER, lower and Toggle Cases.

LOCKED EDITING

The Jalview alignment editing model is different to that used in other alignment editors. Because edits are restricted to the insertion and deletion of gaps to the left of a particular sequence position, editing has the effect of shifting the rest of the sequence(s) being edited down or up-stream with respect to the rest of the alignment.

The Edit → Pad Gaps option can be enabled to eliminate 'ragged edges' at the end of the alignment, but does not avoid the 'knock on' effect which is sometimes undesirable.

However, its effect can be limited by performing the edit within a selected region. In this case, gaps will only be removed or inserted within the selected region. Edits are similarly constrained when they occur adjacent to a hidden column.

NB: You cannot do this in MS Word or Notepad.

SOME USEFUL FEATURES OF JALVIEW

Click on Calculate → Auto-Calculate Consensus

Automatic update of the graph below the alignment; this graph looking like a city skyline – indicates the level of conservation within the alignment. If you set this option, the graph is updated automatically while you are editing.

Click Edit → Remove Redundancy

Makes sure that no pair of sequences is more than x percent similar Ctrl + D i.e. any redundant sequence will automatically be removed.

Click Calculate → Tree → Neighbor Joining Tree Using PID

Computes and displays a phylogenetic tree in graphic Neighbor joining format (you can choose your choice of algorithm to use i.e. Neighbor…) on which you can select sequences for group editing. i.e. Jalview use for making a tree.

Practice Hiding and Revealing Regions

Try in your own time…

- **Close all windows then open the PFAM accession PF03460. Select a contiguous set of sequences by clicking and dragging on the sequence ID panel. Right click on the selected sequence ID and select Hide sequences**
- **Right click on the blue triangle indicating hidden sequences and select Reveal Sequences. (If you have hidden all sequences then you will need to use the alignment window menu option View → Show → All Sequences)**
- **Repeat but using a non-contiguous set of sequences. Note that when multiple regions are hidden there are two options, Reveal Sequences and Reveal all.**
- **Repeat the above but hiding and revealing columns instead of sequences**
- **Select a region on the alignment, add in some additional columns on the selection and experiment with the "Hide all but selected region function'**
- **Select some sequences and pick one to represent the rest. Bring up the sequence ID pop-up menu for those sequences and select the Represent group with <Sequence ID> option. Use the pop-up option. Use the pop-up menu again to reveal the hidden sequences that you just picked a representative for.**

NAVIGATION IN CURSOR MODE

Points to Remember…

- **The primary function of Jalview is the editing and visualization of sequence alignments, and their interactive analysis, helps people who have many sequences**
- **It has many features like: Tree building, principal component analysis, physic-chemical property conservation and sequence consensus analysis.**
- **Web services enable Jalview to access remote alignment and secondary structure prediction programs, as well as to retrieve protein and nucleic acid sequences, alignments, protein structures and sequence annotation.**
- **Alignment views are dynamically linked with *Jmol* structure displays, a tree viewer and spatial cluster display 3D view, facilitating interactive exploration of the alignment's structure.**

- **The application provides its own *Jalview project file format* in order to store the current state of an alignment and analysis windows.**
- **Jalview also provides WYSIWIG7 style figure generation capabilities for the preparation of alignments for publication. NB: Style quality figure generation capability.**
- **Therefore, it helps in Sequences, alignments, trees, structures, features and alignment annotation may also be exchanged with the local file system**
- **Multiple visualizations of an alignment may be worked on simultaneously, and the user interface provides a comprehensive set of controls for coloring and layout**

DESKTOP JALVIEW vs ONLINE JALVIEW

- **Jalview Desktop, however, is much more powerful than the Online Jalview because it includes features like additional support for interaction with external web services and production of publication quality graphics.**

WHEN JALVIEW STARTS

When Jalview starts it will automatically load an example alignment from the Jalview site. This behavior can be changed in the Jalview Desktop preferences dialog opened from the Desktop's Tools → Preferences → Open File menu (check and uncheck Open File menu option)

JALVIEW NEWS RSS FEED

From time to time, important announcements are made available to users of the Jalview Desktop via the Jalview News reader. This window will open automatically when new news is available and can also be accessed via the Desktop's Tools → Show Jalview News menu entry

GETTING HELP

Jalview has comprehensive online help documentation. Select Help → Documentation from the main window menu and a new window will open. The appropriate topic can then be selected from the navigation panel on the left hand side. To search for a specific topic, click the search tab and enter keywords in the box which appears.

EMAIL LISTS

The Jalview Discussion list jalview-discuss@jalview.org provides a forum for Jalview users and developers to raise problems and exchange ideas – any problems, bugs, and requests for help should be raised here. The jalview-accounce@jalview.org list can also be subscribed to if you wish to be kept informed of new releases and developments. NB: Send message to both emails for any enquiries

LOADING YOUR OWN SEQUENCES

Jalview provides many ways to load your own sequences:

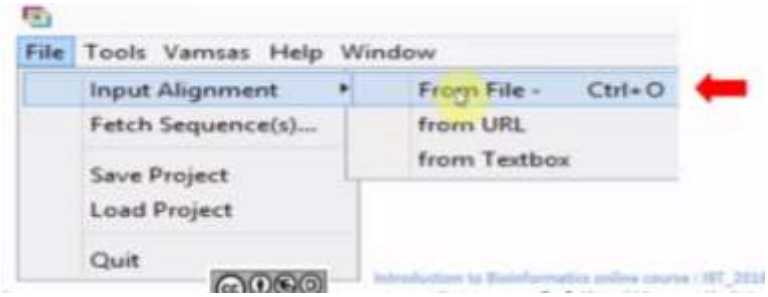
- **You can Drag and Drop**
- **From a File**
- **From a URL**
- **Cut and Paste**
- **From a Public Database**

Drag and Drop

- **In most operating systems you can just drag a file icon of any format from a file browser window and drop it on an open Jalview application window e.g. FASTA format etc**
- **Drag and drop also works when loading data from a URL – simply drag the link or URL from the address panel of your browser on to an alignment or the Jalview desktop background and Jalview will load data from the URL directly.**

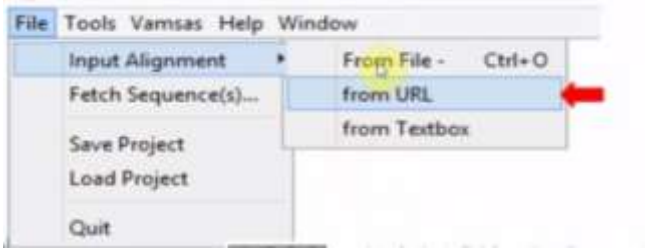
From a File

- **Jalview can read sequence alignments from a sequence alignment file. This is a text file, NOT a word processor document. Select File → Input Alignment → From File from the main menu. Remember to select the appropriate file type. Jalview can automatically identify some sequences file formats**



From a URL

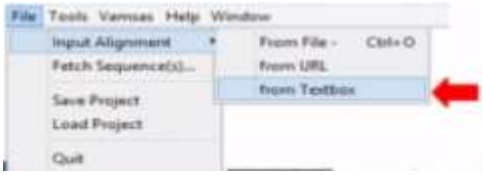
- **Jalview can read sequence alignments from a sequence alignment directly from a URL. Please not that the files must be in a sequence alignment format – an HTML alignment or graphics file cannot be read by Jalview. Select File → Input Alignment → From URL**
  Jalview will attempt to automatically discover the file format



Cut and Paste

Documents such as those produced by Microsoft Word cannot be readily understood by Jalview. The way to read sequences from these documents is to select the data from the document and copy it to the clipboard. There are two ways to do this:

1. **One is to right click on the background, and select the 'Paste to new alignment' option in the menu that appears**
2. **The other is to select File → Input Alignment → From Textbox from the main menu, and paste the sequences into the textbox window that will appear. In both cases, sequences should be in the right format**



From a Public Database

- **Jalview can retrieve sequences and sequence alignments from the public databases housed at the European Bioinformatics Institute, including Uniprot, Pfam, Rfam and PDB as well as any DAS sequence server registered at the configured DAS registry.**
- **Jalview's sequence fetching capabilities allow you to avoid having to manually locate and save sequences from a web page before loading them into Jalview. It also allows Jalview to gather additional metadata provided by the source such as annotation and database cross references. Select File → Fetch Sequences.**

  NB: If you know the accession number

Pressing the database selection button in the dialog box opens a new window showing all the database sources Jalview can access (grouped by the type of database). Once you've selected the appropriate database, hit (OK) close the database selection window and then enter one or several database IDs or accession numbers separated by a semicolon and press OK button.

(Jalview) will then attempt to retrieve them from the chosen database.

Exercise

Try on your own time…

Loading Sequences

- **Start Jalview then close all windows (if necessary) by selecting Window → Close All from the Desktop window**
- **Select File → Input Alignment → From URL from the Desktop and enter http://www.jalview.org/tutorial/alignment.fa in the box. Click OK and the alignment should load**

- Close all windows using the Window → Close All menu option from the Desktop. Point your web browser at the same URL (http://www.jalview.org/tutorial/alignment.fa) and save the file to your desktop
- Open this file in Jalview by selecting File → Input Alignment → From File from the main menu and selecting the file from your desktop. Click OK and load the alignment.
- Select Desktop → Window → Close All and drag the *alignment.fa* file from the desktop onto the Jalview window. The alignment should open. Try dragging onto an empty Jalview and onto an existing and observe the results. You can also try dragging the URL directly onto Jalview.

Exercise Continue

Try on your own time…

Loading Sequences

- Select File → Fetch Sequences from the desktop. Select the PFAM seed database and enter the accession number. PF03460. Click OK. An alignment of about 107 sequences should load
- Open http://www.jalview.org/tutorial/alignment.fa in a web browser. Note; if the URL is downloaded instead of opened in the browser, then locate the file in your download directory and open it in a text editor
- Select any copy the entire text to the clipboard (usually via the browser's Edit → Copy menu option)
- Ensure Jalview is running and Select File →Input Alignment → From Textbox
- Paste the clipboard into the large window using the Edit → Paste textbox menu option. Click New Window and the alignment will be loaded

NAVIGATION

The major features of the Jalview Desktop are illustrated in Figure. The alignment window is the primary window for editing and visualization, and can contain several independent views of the alignment being worked with. The other windows (Trees, Structures, PCS plots etc) are linked to a specific alignment view. Each area of the alignment window has a separate context many accessed by clicking the right mouse button.

Jalview has two navigation and editing modes:

1. Normal Mode: where editing and navigation is performed using the mouse
2. Cursor Mode: where editing and navigation are performed using the keyboard

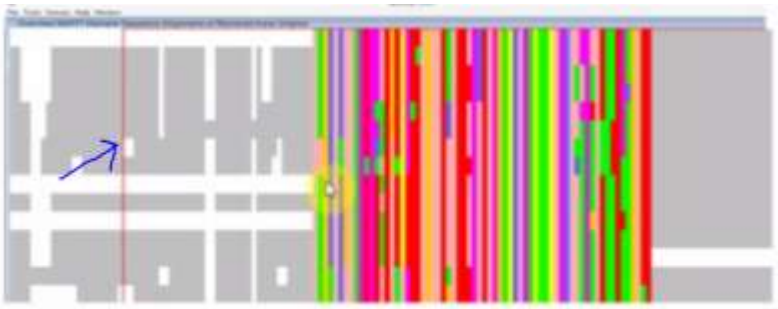
   The F2 key is used to switch between these two modes

You can move about the alignment by clicking and dragging the ruler scroll bar to move horizontally, or by clicking and dragging the alignment scroll bar to the right of the alignment to move vertically.

NB: If all the rows or columns in the alignment are displayed, the scroll bars will not be visible.

If you have a large alignment and will like to overview the whole thing in a small window… Each alignment view shown in the alignment window represents a window onto the visible regions of the alignment. This means that with anything more than a few residues or sequences, alignments can become difficult to visualize on the screen because only a small area can be shown at a time. It can help, especially when examining a large alignment, to have an overview of the whole alignment.

Select view → Overview Window from the Alignment Window menu bar



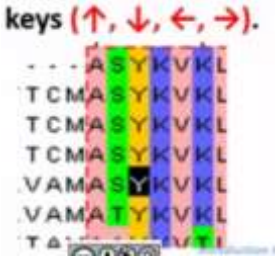
The red box in the overview window shows the current view in the alignment window
A percent identity histogram is plotted below the alignment overview
Shaded parts indicate rows and columns of the alignments that are hidden (in this case, a single row at the bottom of the alignment) you can navigate around the alignment by dragging the red box.

CLOSE WINDOWS

Alignment and analysis windows are closed by clicking on the usual 'close' icon (indicated by arrows on Mac OS X) if you want to close all the alignments and analysis windows at once, then use the Window → Close ALL
WARNING: Make sure you have saved your work because this cannot be undone.

Navigation in Cursor Mode

- **Cursor Mode navigation enables the experienced user to quickly and precisely navigate, select and edit parts of an alignment**
- **On pressing F2 to enter cursor mode the position of the cursor is indicated by a black background and white text.**
- **The cursor can be placed using the mouse or moved by pressing the arrows keys**



Rapid Movement to Specific Positions is accomplished as listed below:

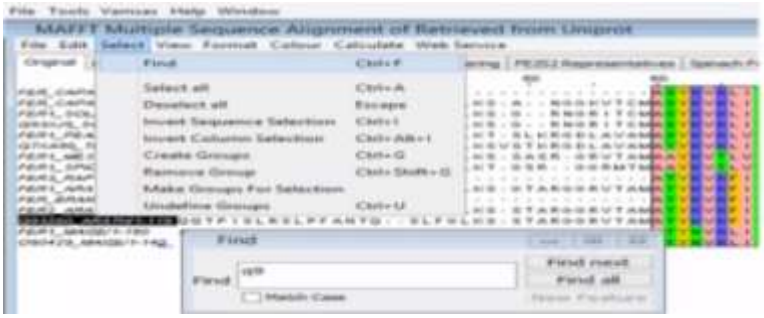
- **Jump to Sequence n: Type a number n then press S to move to sequence row n**
- **Jump to Column n: Type a number n then press C to move to column n in the alignment**
- **Jump to Residue n: Type a number n then press P to move to residue number n in the current sequence**

- **Jump to column m row n: Type the column number m, a comma, the row number n and press enter key i.e. Return key**

THE FIND DIALOG BOX

A further option for navigation is to use the Select → Find… function

- **This opens a dialog box into which can be entered regular expressions for searching sequences and sequence IDs, or sequence numbers**
- **Hitting the Find Next button will highlight the first (or next) occurrence of that pattern in the sequence ID panel or the alignment, and will adjust the view in order to display the highlighted region.**



NB: Very useful to find specific sequence from many sequences.

Exercise

Try on your own time…

- **Reload the example file by accessing the Desktop's File → Input Alignment → From URL dialog and clicking on the down arrow to retrieve the example file URL stored in its history (http://www.jalview.org/examples/exampleFile_2_7.jar)**
- **Scroll around the alignment using the alignment (vertical) and ruler (horizontal) scroll bars**
- **Find and open the Overview Window. Move around the alignment by clicking and dragging the red box in the overview window**
- **Look at the status bar as you move the mouse over the alignment. It should indicate information about the sequence and residue under the cursor**
- **Press F2 to enter Cursor mode. Use the arrow keys to move the cursor around the alignment**
    - **Move to sequence 7 by pressing 7S.**
    - **Move to column 18 by pressing 18C.**
    - **Move to residue 18 by pressing 18P.**
    - **Move to sequence 5 column 13 typing 13,5 [RETURN] i.e. enter key**

NOTE: that these can be two different positions if gaps are inserted into the sequence.

FROM CLUSTAL OMEGA TO JALVIEW

Connecting the output of Clustal Omega into Jalview

You can make an alignment with Clustal, TCoffee, Clustal-Omega etc you get the output and visualize it in Jalview.

Take the following steps:

- **Collect your sequences in FASTA format in one text file i.e. DNA or Protein FASTA file**
- **Login into Clustal Omega http://www.ebi.ac.uk/Tools/msa/clustalo/ and paste your sequences and choose the type of your input**
- **Download your alignment in FASTA Format**
- **Open Jalview and click on File → Input Alignment → From Textbox then take and upload your file**
- **Control your view e.g. Click on Format → Wrap (to see all sequences in front of you) etc**
- **Choose the best color for representing your sequence e.g. Click Color → Nucleotide (means for DNA every nucleotide has different color) i.e. Four bases Adenine, Guanine, Cytosine and Thymine all to have unique color**
- **For DNA also you can use Color → Purine/Pyrimidine (means that Purines like Adenines and Guanines will have same color and Pyrimidine like Cytosine and Thymine will have same color**
    - **Only the Color → Nucleotide and Color → Purine/Pyrimidine are for coloring DNA, others items in the Color menu are color schemes for proteins.**
- **Export your alignment to an image for publication after the above steps, make sure you have opened the file click Edit → Export Image → Click on the format e.g. HTML, EPS or PNG (recommended is PNG because it has high quality image).**
- **Save to a location on your system**
- **Insert your saved alignment image into your document for publication**

WRITING SEQUENCE ALIGNMENTS

MEMORY LIMITS

- **Jalview is a Java program. One unfortunate implication of this is that Jalview cannot dynamically request additional memory from the operating system.**
- **It is important, therefore, that you ensure that you have allocated enough memory to work with your data**
- **On most occasions, Jalview will warn you when you have tried to load an alignment that is too big to fit into memory (for instance, some of the PFAM alignments are very large). You can find out how much memory is available to Jalview with the desktop Click Tools → Show Memory Usage function.**

Result: Total Free Memory: 204.89 MB; Max Memory: 247.5 MB; 82.78%

Essence is to ensure that you have enough allocated memory to use.

Should you need to increase the amount of memory available to Jalview, full instructions are given in the built in documentation (Click on Help → Documentation) and on the JVM memory parameters page (http://www.jalview.org/ivmmemoryparams.html)

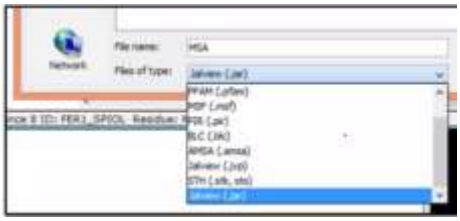
HOW TO SAVE YOUR ALIGNMENTS

Jalview allows the current sequence alignments to be saved to file from the alignment window menu select Click File → Save or Ctrl S or you may want to change the filename using SaveAs or change file format

Jalview will remember the last filename and format used to save (or load) the alignment, enabling you to quickly save the file during or after editing by using the File → Save entry.

Jalview offers several different formats in which an alignment can be saved. NB: The Jalview format (.jar) is the only one which will preserve the colors, groupings and similar information in the alignment that you have grabbed during your session.

The other formats like (.pfam, .msf, .pir, .blc, .amsa, .stk .sto) produce text files containing just the sequences with no visualization information. Unfortunately, only Jalview can read Jalview files.



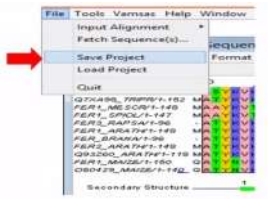
You can also save using Jalview Project with extension (.jvp) saving it as project to use it later on. Recommended: always save in .jar or .jvp

You can use the File → Output to Textbox menu option (select the format e.g. BLC, CLUSTAL, FASTA, MSF, PIR, PFAM, STH, AMSA) which allows the alignment to be copied and pasted into other documents or web servers. You may try to take the highest output from Jalview and open it in another program using the above step choosing the appropriate choice of the format for the new program, essence: is to transfer the data from Jalview to other programs.

JALVIEW PROJECTS

If you wish to save the complete Jalview session rather than just one alignment (e.g. because you have calculated trees or multiple different alignments) then your work should be saved as a Jalview Project file (.jvp). This will help keep all the Jalview information (alignments, trees, annotation and displayed structures) that can be used in next session

Very simple to save Jalview project: From the main menu select File → Save Project and a file save dialog box will appear. Loading a project will restore Jalview to exactly the view at which the files was saved, complete with all alignments, trees, annotation and displayed structures rendered appropriately.



Exercise

Try on your own time…

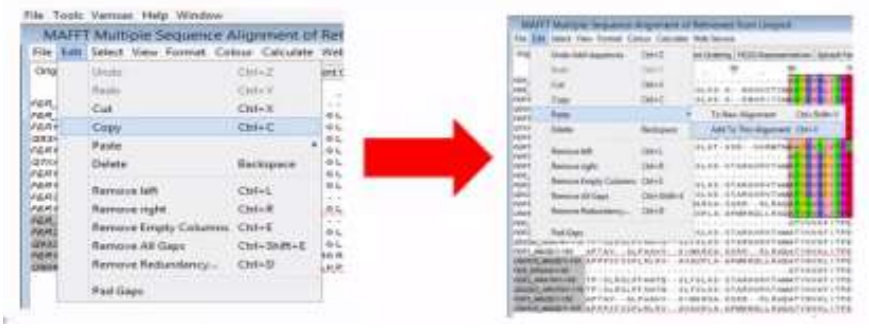
- **Start Jalview, close all windows and load the ferredoxin alignment from PFAM (PFAM seed accession number PF03460 (see Exercise 3).**
- **Select File → SaveAs from the alignment window menu. Choose a location into which to save the alignment and select a format**
    - **All formats except Jalview can be viewed in a normal text editor (e.g. Notepad) or in a web browser. Enter a filename and Click Save. Check this file by closing all windows and opening it with Jalview or by browsing it with your web browser**
- **Repeat the previous step trying different file formats**
- **Select File → Output to Textbox → Fasta You can select and copy this alignment to the clipboard using the textbox menu options Edit → Select All followed by Edit → Copy. The alignment can then be pasted into any application of choice e.g. a word processor or web form**
- **Ensure at least one alignment window is shown in Jalview. Open the overview window and scroll to any part of the alignment. Select File → Save Project from the main menu and save in a suitable place**
- **Close all windows and then load the project via the File → Load Project menu option. Note: how all the windows and positions are exactly as they were when they were saved.**

SELECTING AND EDITING SEQUENCES USING JALVIEW

Selecting parts of an alignment

Selections can be of arbitrary regions in an alignment, one or more complete columns, or one or more complete sequences

A Selected region can be copied and pasted as a new alignment using the: Edit → Copy and Edit → Paste → To New Alignment in the alignment window menu options

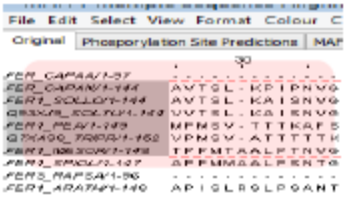


To clear deselect the selection press the ESC Escape Key OR Click Select Menu → Deselect All

SELECTING ARBITRARY REGIONS

To select part of an alignment, place the mouse at the top left corner of the region you wish to select. Press and hold the mouse button and drag the mouse to the bottom right corner of the chosen region then release the mouse button. A dashed red box appears around the selected region. Selecting a region in an alignment is shown below:

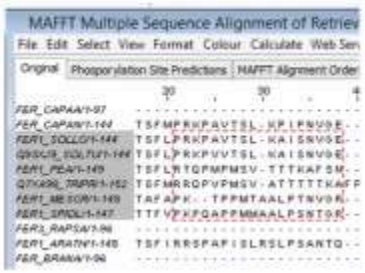
.Selecting a region in an alignment



You can now right click on the selected region to copy and paste in another alignment.

SELECTING COLUMNS (Same technique of mouse selection used in MS Word)

- **To select the same residues in all sequences, click and drag along the alignment ruler. This selects the entire height of the alignment.**
- **Ranges of positions can also be selected by clicking on the first position then holding down the SHIFT key whilst clicking the other end of the selection. i.e. Selecting Continuous regions.**
- **Discontinuous regions can be selected by holding down CTRL and clicking on positions to add to the column selection. NOTE: that each Ctrl selected columns are indicated by red highlighting in the ruler bar**

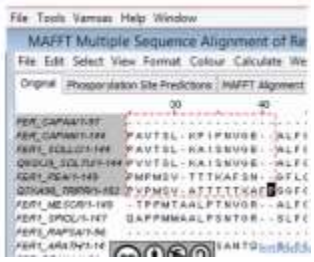


Selecting Sequences

To select multiple complete sequences, click and drag the mouse down the sequence ID panel. The same technique as used for columns above can be used with SHIFT – CLICK for continuous and CTRL-CLICK to select discontinuous ranges of sequences.

Making Selections in Cursor Mode

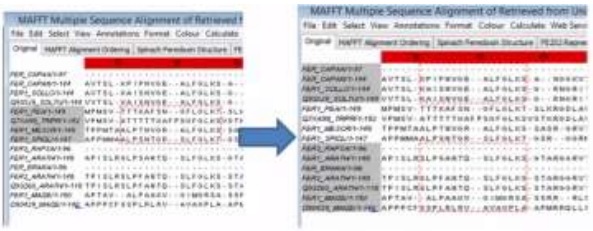
To define a selection in cursor mode (which is enabled by pressing F2 when the alignment window is selected), navigate to the top left corner of the proposed selection (using the mouse, the arrow keys, or the keystroke commands described). Pressing the Q key marks this as the corner. A red outline appears around the cursor. Navigate to the bottom right corner of the proposed selection and press the M key. This marks the bottom right corner of the selection. The selection can then be treated in the same way as if it had been created in normal mode i.e. Q for the start position, M for end position



INVERTING THE CURRENT SELECTION

The current sequence of column selection can be inverted, using Select → Invert Sequence/Column Selection in the alignment window i.e. when you want to hide other part, invert selection will select everything outside your main or original selection then you can hide the inverted selected region

Inverting the selection is useful when selecting large regions in an alignment, simply select the region that is to be kept unselected, and then invert the selection.



This may also be useful when hiding large regions in an alignment. Instead of selecting the columns and rows that are to be hidden, simply select the region that is to be kept visible, invert the selection, then select View → Hide → Selected Region

CREATING GROUPS

Selections are lost as soon as a different region is selected. Groups can be created which are labeled regions of the alignment

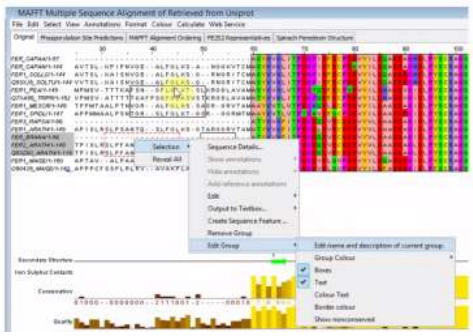
To create a group:

- **First select the region which is to comprise the group**
- **Then click the right mouse button on the selection to bring up a context menu**

Select (and Right Click) Selection → Edit Group → Edit name and description of current group

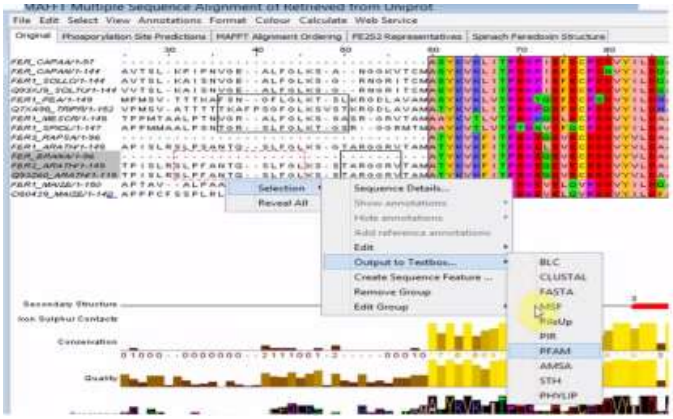
- **Name and description of current group**
- **Then enter a name for the group in the dialogue box which appears.**

By default the new group will have a box drawn around it. The appearance of the group can be changed. This group will stay defined even when the selection is removed.



EXPORTING THE CURRENT SELECTION

- The current selection can be copied to the clipboard (in PFAM format, CLUSTAL, PIR or any other formats which can be used in another program)
- It can also be output to a textbox using the output functions in the pop-up menu obtained by right clicking the current selection
- The textbox enables quick manual editing of the alignment prior to importing it into a new window (using the [NEW Window] button) or saving to a file with the File → SaveAs pull down menu option from the text box.



Exercise PRACTICAL

Try on your own time…

- Close all windows in Jalview and load the ferredoxin alignment (PFAM ID PF03460).
- Choose a residue and place the mouse cursor on it. Click and drag the mouse cursor to create a selection. As you drag, a red box will 'rubber band' out to show the extent of the selection. Release the mouse button and a red box should border the selected region
- Now press ESC key to clear the selection
    - Select one sequence by clicking on the ID panel. Note that the sequence ID takes on a highlighted background and a red box appears around the selected sequence
    - Now hold down SHIFT and click another sequence ID a few positions above or below. NOTE how the selection expands include all the sequences between the two positions on which you clicked
    - Now hold down CTRL key and click on several ID's both selected and unselected. NOTE how unselected IDs are individually added to the selection and previously selected IDs are individually deselected
- Repeat the step above but selecting columns by clicking on the ruler bar instead of selecting rows by clicking on the sequence ID
- Press F2 to enter Cursor mode. Navigate to column 59, row 1 by pressing 59,1 [RETURN] i.e. Enter key.
- Press Q to mark this position. Now navigate to column 65, row 8 by pressing 65,8 [RETURN]. Press M to complete the selection.
- Open the popup menu by right-clicking the selected region with the mouse
- Open the Selection → Group → Group Color menu and select 'Percentage Identity'

- This will turn the selected region into a group and color it accordingly
- Hold down CTRL and use the mouse to select and deselect sequences by clicking on their Sequence ID Label. Note how the group expands to include newly selected sequences and the 'Percentage Identity' coloring changes
- Use the mouse to click and drag the right hand edge of the selected group. Note again how the group resizes.

Right click on the text area to open the selection popup menu. Follow the menus and pick an output format from the Selection → Output to Textbox… submenu. Try manually editing the alignment and then press the NEW Window button to import the file into a new alignment window.

Good thing about Jalview is that you can use it with other programs via importing and exporting. The exchange of data between Jalview and other programs make it very powerful software.

MODULE 6

SESSION 1 – PROTEIN STRUCTURE BIOINFORMATICS

LEARNING OBJECTIVES

- To recap some basics of amino acids and proteins
- To study the different levels of protein structures
- To shed light on how protein structures are determined
- To learn about some relevant databases, file formats and file viewers

LEARNING OUTCOMES

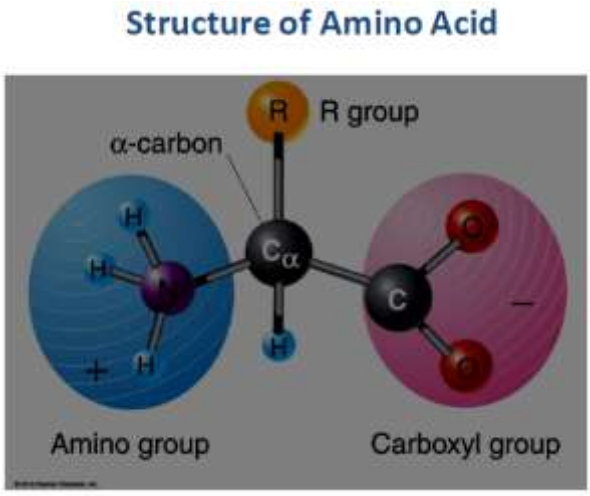
By the end of this session and practical, students are expected to be able to

- Explore some resources and tools in the PDB database
- Use some web-servers to predict Protein Structure

STRUCTURE OF AMINO ACID

Proteins can be considered as polymers of amino acids

Amino acid generally has central carbon called C alpha i.e. Cα



Amino acid composed of: α-Carbon, R group, Amino group, Hydrogen and Carboxyl group

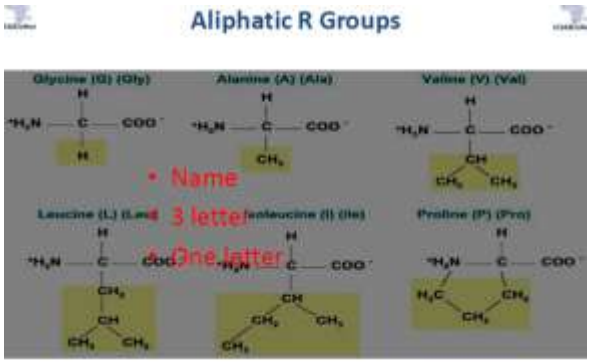
NB: All Amino group, Hydrogen and Carboxyl group are constant in each and every amino acid. R group is a variant in every amino acid means it is not same in every amino acid.

NB: There are generally 20 amino acids contributing to protein structures and these amino acids can be grouped based on many different characters.

ALIPHATIC R GROUPS

Six 6 R Groups are Aliphatic in nature:

1) **Glycine (G) (Gly)**
2) **Alanine (A) (Ala)**
3) **Valine (V) (Val)**
4) **Leucine (L) (Leu)**
5) **Isoleucine (I) (Ile)**
6) **Proline (P) (Pro)**

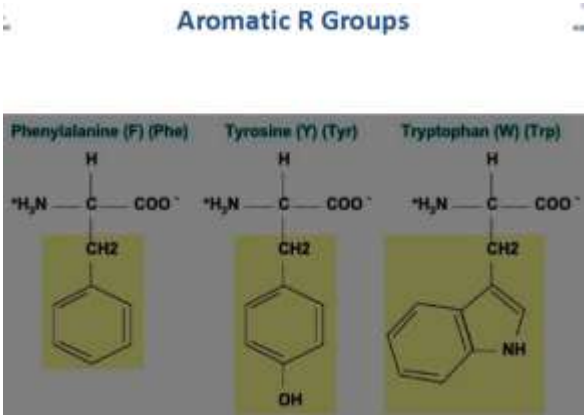


We can see that they can be identified by:

- **Name**
- **3 letter**
- **One letter**

AROMATIC R GROUPS

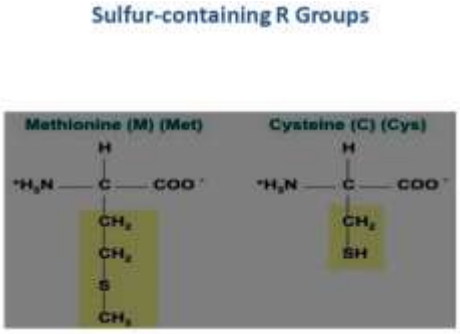
- **Phenylalanine (F) (Phe)**
- **Tyrosine (Y) (Tyr)**
- **Tryptophan (W) (Trp)**



SULFUR CONTAINING R GROUPS

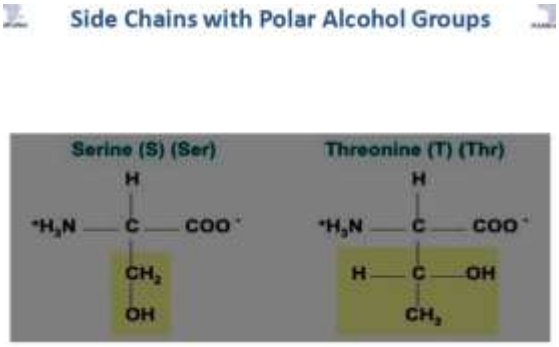
Some amino acids that contain sulfur

- **Methionine (M) (Met)**
- **Cysteine (C) (Cys)**



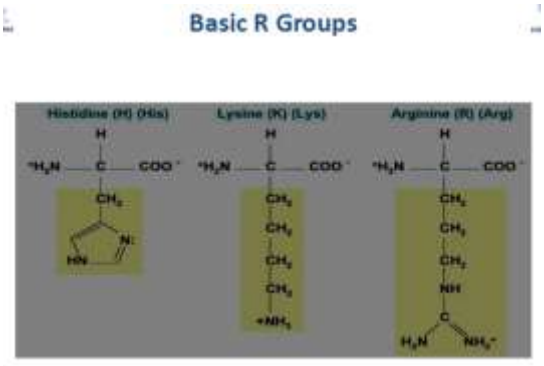
SIDE CHAINS WITH POLAR ALCOHOL GROUPS

- **Serine (S) (Ser)**
- **Threonine (T) (Thr)**



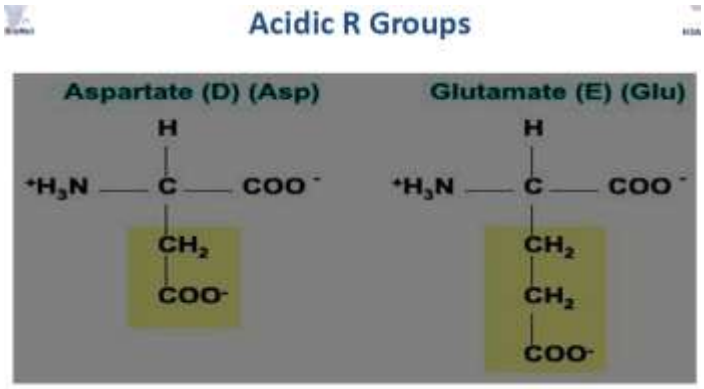
BASIC R GROUPS

- **Histidine (H) (His)**
- **Lysine (K) (Lys)**
- **Arginine (R) (Arg)**

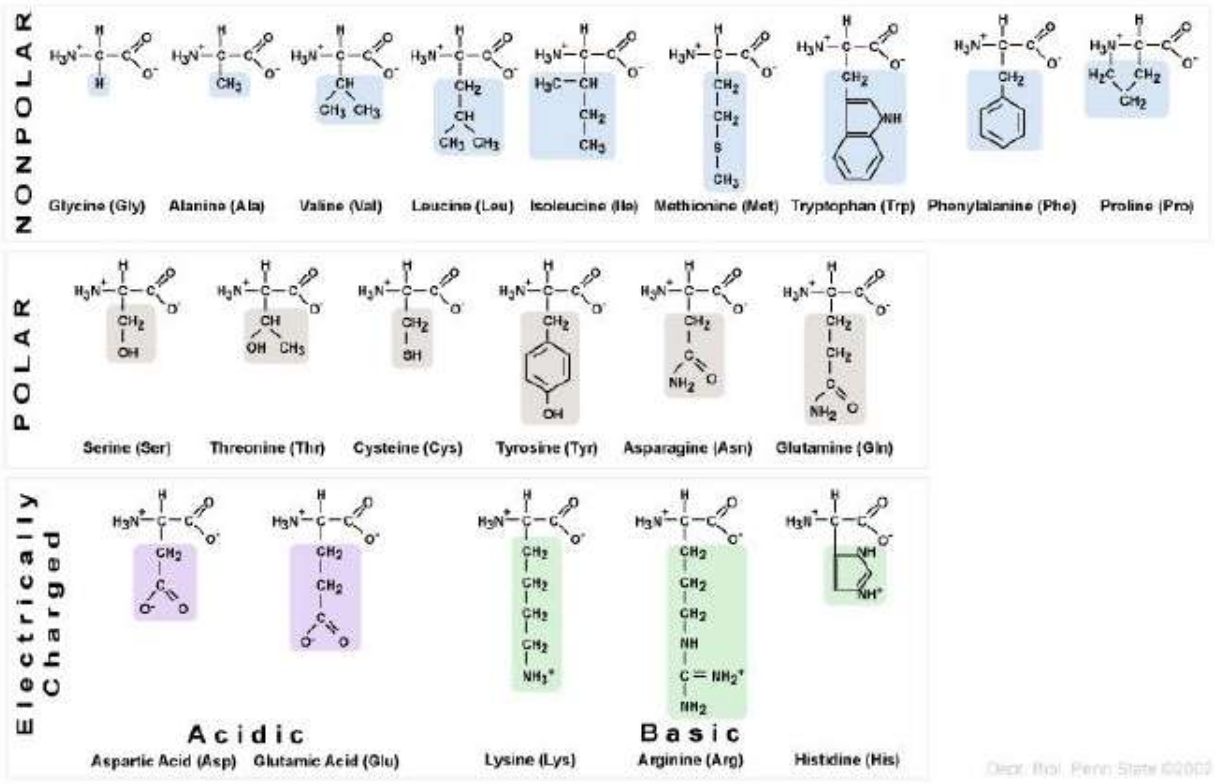


ACIDIC R GROUPS

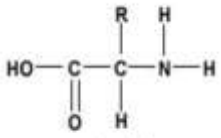
- **Aspartate (D) (Asp)**
- **Glutamate (E) (Glu)**

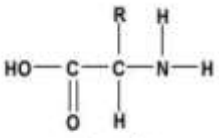


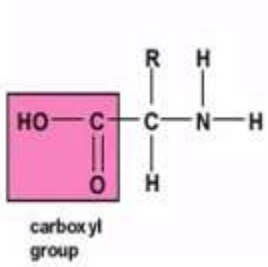
Some amino acids are POLAR while some are NONPOLAR

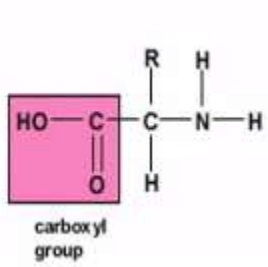


MOLECULAR INTERACTIONS BONDS AND PROTEIN STRUCTURES

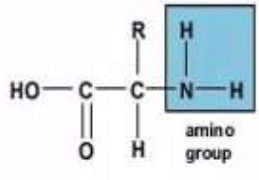


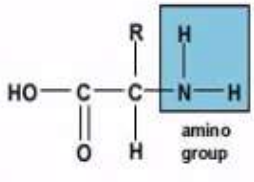






First amino acids shown in each, interacts with the…

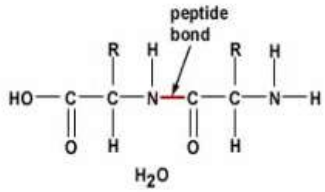




Second amino acid

As earlier said, protein is a polymer of amino acids

The two amino acids will bond and link at the H and HO hydration reaction makes it forming H_2O with the N – C peptide bond



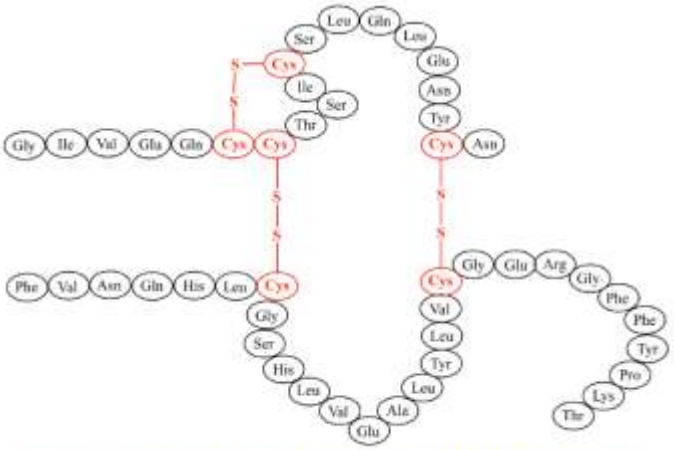
The above is a kind of form that is available in protein primary structures. It is called the intermolecular forces because it is happening inside the protein molecule

There are other intermolecular forces…

- **Dipole interactions**
- **Hydrogen bonds**
- **Van der Waals forces**
- **Hydrophobic Interactions**
- **Others**

NB: These kinds of intermolecular forces are much weaker compare to intermolecular forces

Special kind of intermolecular forces is the is compliable type that arise between sulfur groups in different amino acids S – S, we called disulfide bond

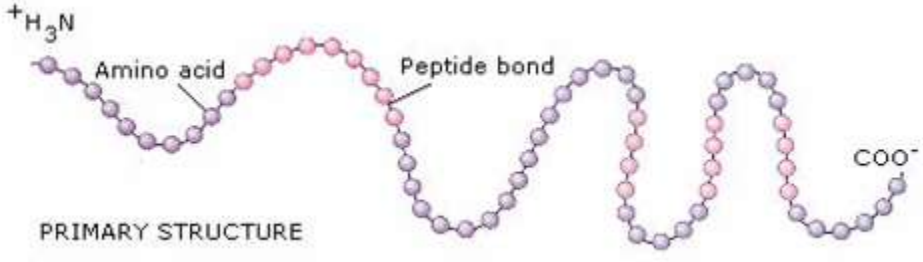


http://www.chem.ucla.edu/~harding/IGOC/D/disulfide_bridge.html

As shown the S – S is intermolecular force and it is a very strong type of intermolecular forces. It can happen between two different peptides and same peptide, as shown above the Cystine (Cys) bond to Cys from same peptide, or from different peptide.

INTERMOLECULAR FORCES

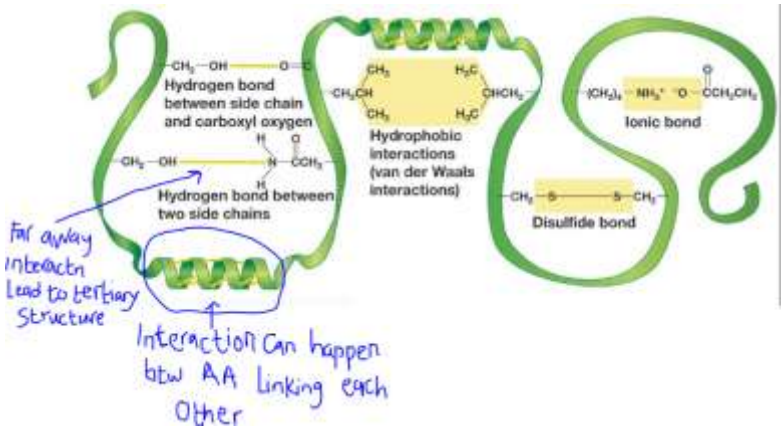
PRIMARY STRUCTURE OF INTERMOLECULAR FORCES



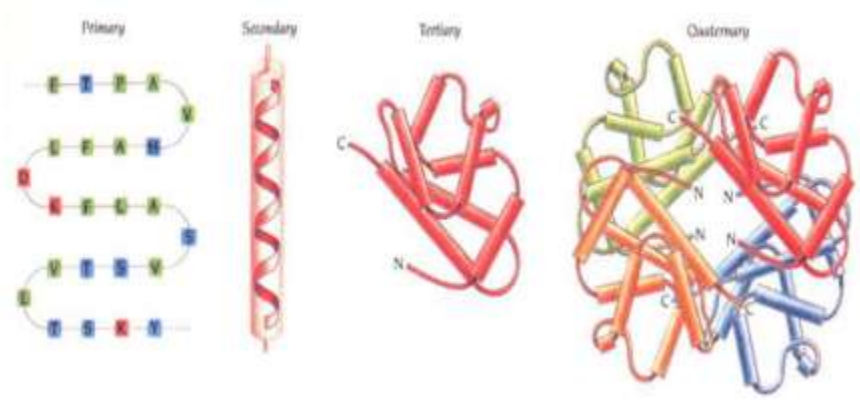
Intermolecular forces generally contributes to the secondary and tertiary structures.

The primary sequence of amino acid are called primary structure

When the amino acids began to interact with each other, the protein will adopt another form called the secondary structure



The coiled area in the image above shows the secondary structure due to the interaction between amino acids linking each other. When interactions began to happen between amino acids that are located far way from each other inside the primary structure we start to have tertiary structure.



Primary → Secondary→ Tertiary (far away AA)→ Quaternary

Sometime the tertiary structure of one peptide will interact with another peptide in its structure then it gives us the Quaternary Structure. The image above shows different peptides of green, orange, red and blue interacting.

STRUCTURE IS INSTRUCTED IN THE SEQUENCE

When it comes to protein, its structure is really what matters and structure is also instructed in the primary sequence.

Anfinsen's Dogma

Christian B. Anfinsen 916-1995, U.S. biochemist: Nobel Prize in Chemistry 1972

Proved that Proteins in their natural physiological environment will always be in the same shape and these shape is actually the most dynamic shape

Principles that Govern the Folding of Protein Chains

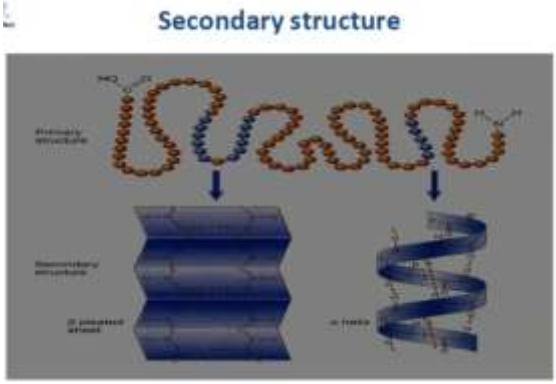
Science 20 Jul 1973 Vol. 181, Issue 4096, pp 223-230.

SECONDARY STRUCTURE

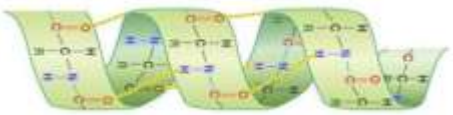
Intermolecular forces or intermolecular interactions between amino acids in the primary structure results in other forms or levels of structures like the secondary structure.

The two types of secondary structures are:

1. **Alpha Helices and**
2. **Beta Sheets**



α- helix



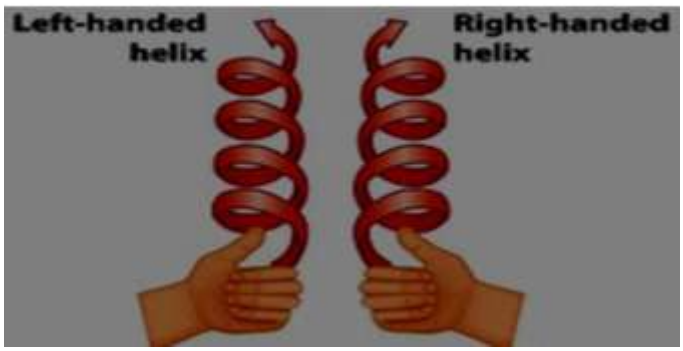
Alpha helix is a classic element of protein structure and non-alpha helix may contain 35 per… (Not too clear 1min video part2)

Hydrogen bonds are formed between Oxygen of the carbonyl group of each peptide and the hydrogen atom in the NH group of the peptide bond as well. Alpha helix arise when hydrogen bonds form between amino acids number I and amino I plus O. (not too clear 1.08min in the video part2)

Alfa helix can be:

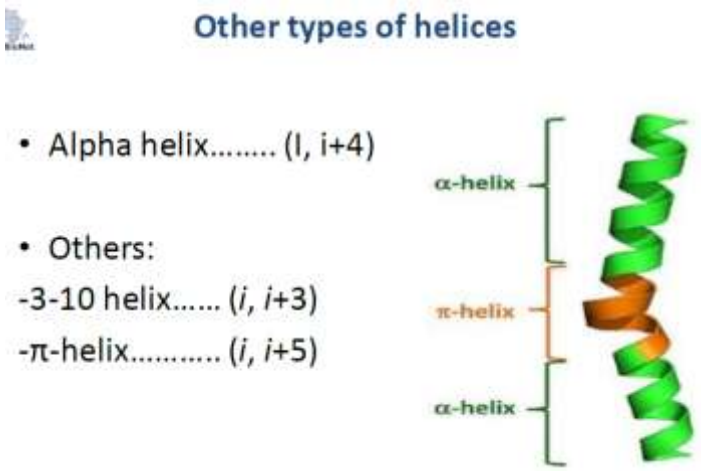
1. **Right handed helix or**
2. **Left handed helix**

α- helix

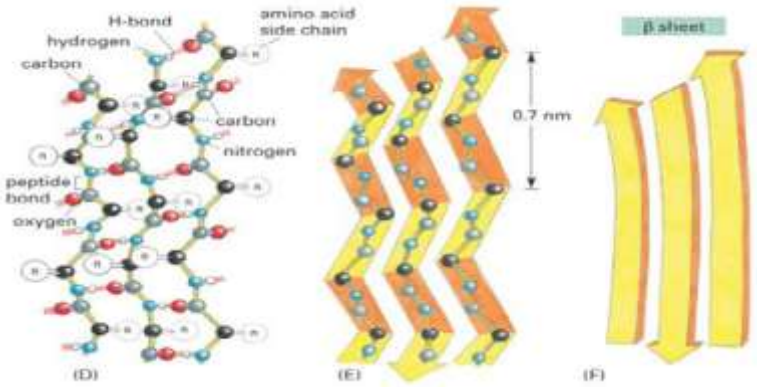


OTHER TYPES OF HELICES

- **Alpha helix … (i, i + 4)**
- **Others:**
  - **3 -10 helix… (i, i+3)**
  - **π-helix … (i, i+5): usually results in wide delta compared to the alpha helix**



BETA STRANDS (β-Strands)

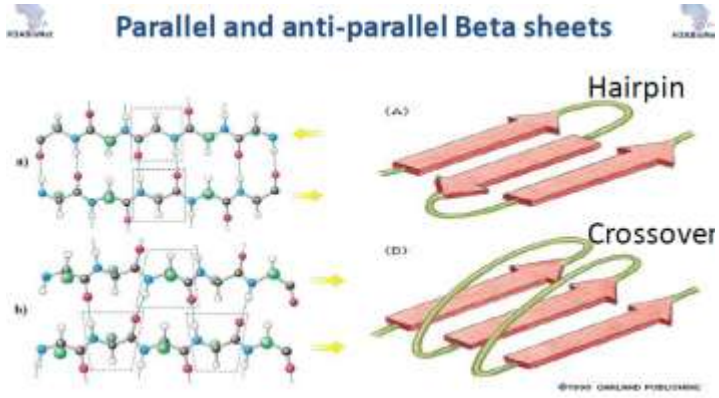


Beta strand is a strand is a stretch of polytide chain of same amino acid, the bonding in beta strand is between adjacent segments of the proteins. Here the hydrogen bonding is between the adjacent strands. The two terms Beta Sheets and Beta Strands are used interchangeably, but it is know that beta sheets consist of beta strands, in the image the beta sheet consist of 3 beta strands.

PARALLEL AND ANTI PARALLEL BETA SHEETS

Proteins are directionality meaning they go from N terminal to C terminal

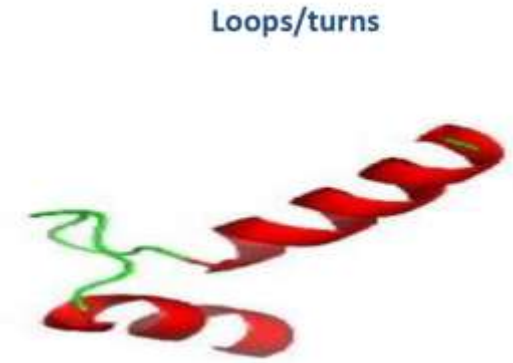
- **When the two strands in beta sheet goes opposite i.e. N-C and C-N they are called anti-parallel beta sheets**
- **When the two strands goes in same direction they are called Parallel beta sheets**
- **Beta strands are connected with regular or irregular shapes called Hairpin in the Anti-parallel beta sheets and it is called Crossover in the parallel beta sheets.**



LOOPS/TURNS

Loops are irregular secondary structures and usually connect regular secondary structures like alpha helices and beta sheets

Loops are diverse class that contains stains, random coils, strands etc all for connecting secondary structures

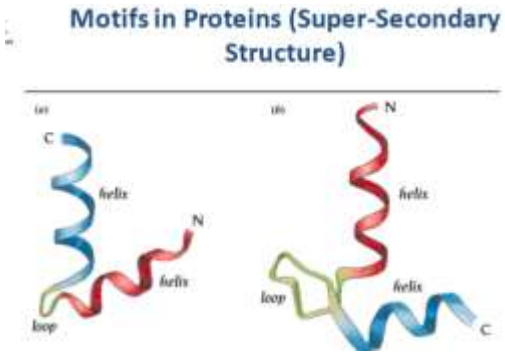


MOTIFS IN PROTEINS (SUPER-SECONDARY STRUCTURE)

As in DNA where there is motif, here also there is motif in proteins

A structural motif may not correlate with the sequence motif

Structural motifs are actually combinations of different secondary structures e.g. helix – loop – helix combination of helix and loop as shown in the image (a) and (b) below



PSI-Loop

Also contains alpha helices and beta sheets

- **Psi-loop**

DSSP (DICTIONARY OF PROTEIN SECONDARY STRUCTURE)

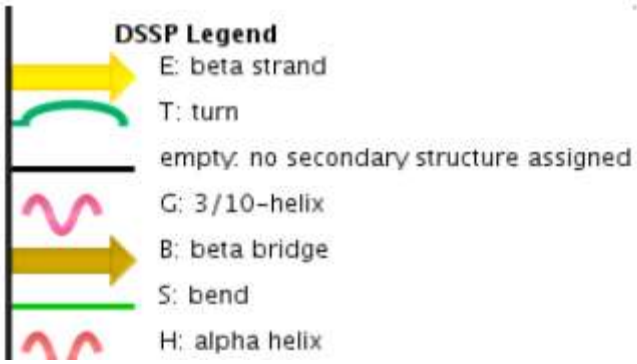
- **It basically defines criteria for secondary structure**
- **Programmed as a pattern-recognition process of hydrogen-bonded and geometrical features extracted from x-ray coordinates**

According to this dictionary

DSSP (Helix, Strand and Loops)

	Secondary Structure	Symbols
Helix	Alpha helix	G
	3-10 helix	H
	π-helix	I
Strand	Beta bridge	B
	Beta strand	E
Loops	Turns	T
	High curvature	S
	Space/no rule applies	C

DSSP (Dictionary of protein secondary structure)



E : **beta strand**
T: **turn**
Empty: **no secondary structure assigned**
G: **3/10-helix**
B: **beta bridge**
S: **bend**
H: **alpha helix**

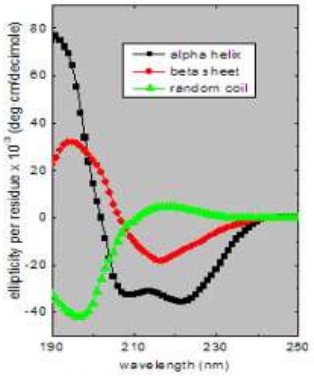
EXPERIMENTAL DETERMINATION OF SECONDARY STRUCTURE

There are many methods to determine secondary structure

- **Spectroscopy**
- **UV CD circular dichroism**
- **IR spectroscopy (Infra-red spectroscopy)**
- **NMR (nucleic magnetic … do not get it because she just spoke too fast here**

Most of these methods actually tells you how much of your protein is alpha helix and how much is better sheets they don't tell you what is the exact sequence or protein means, i.e. they don't tell about if specific sequence is alpha helix or beta sheets.

Showing Spectroscopy of different proteins



http://www.ap-lab.com/images/CD_STANDARDS.gif

SECONDARY STRUCTURE PREDICTION

- **Early/empirical methods:**
- **Probabilities and pre-computed residues preferences**
- **Chou-Fasman method (~60% accurate)**
- **CFSSP: Chou & Fasman Secondary Structure Prediction Server**
- **http://www.biogen.org/tools/chou-fasman**

For instance, calculate the helical propensity of residue type X

- **Pα(X) = frequency (X in helix)/frequency (X)**
- **Pα > 1 = favours helix (e.g., Pα(Glu) = 1.51) i.e. when probability of X is more than 1**
- **Pα < 1 = disfavours helix (e.g. Pα(Gly) = 0.57) i.e. when probability of X is less than 1**

Example:

- **Database of 2000 residues**
- **100 are Alanines**
- **500 residues are in a helix**
- **50 alanines are in a helix**
- **What is the propensity for Ala to be in a**
- **Helix? Is ALA a good helix**

Solution:

- **Pα(X) = frequency (X in helix)/frequency (X)**
- **Pα(Ala) = frequency(Ala, α)/frequency (Ala)**
- **Frequency (Ala, α) = 50/500 = 0.1**
- **Frequency (Ala) 100/2000 = 0.05**
- **Pα(Ala) = 0.1/0.05 = 2.0**
- **Ala is a good helix former**

Current, Machine Learning-Based methods employ information from multiple sequence alignment, information theory, and some machine learning algorithms like artificial neural network and Bayesian networks or a combination of those. E.g. PSIPRED http://bioing.cs.uk.ac.uk/psipred/

TERTIARY STRUCTURE

- **The tertiary structure is the final specific geometric shape that a protein assumes**
- **It is determined by a variety of bonding interactions between the "side chains" on the amino acids**
- **Bond involve: hydrogen bonding, salt bridges, disulfide bonds and non-polar hydrophobic interactions**

METHODS OF 3D STRUCTURE DETERMINATION

Information on 3D structure can be obtained by

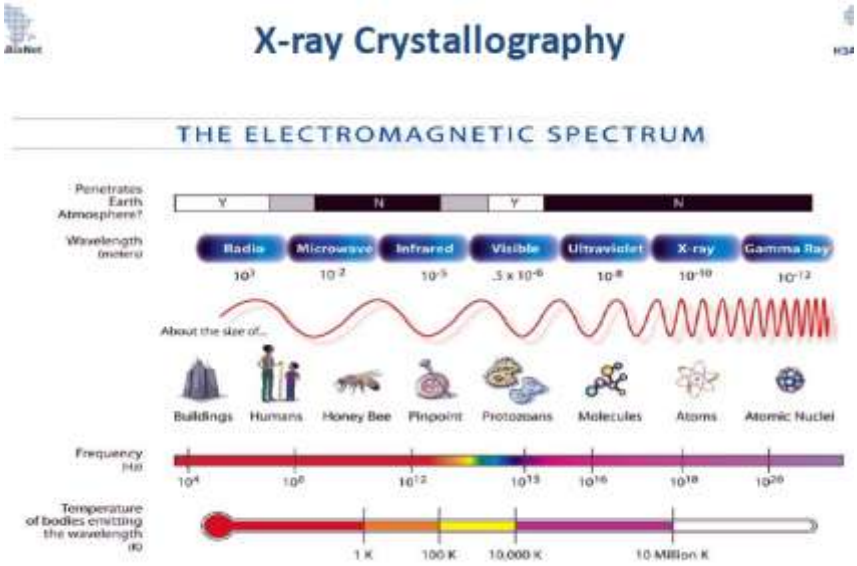
- **X-ray crystallography**
- **NMR spectroscopy or**
- **Cryo-electron microscopy**

Submitted by biologists and biochemists from around the world

Freely accessible on the internet via the websites of its member organizations

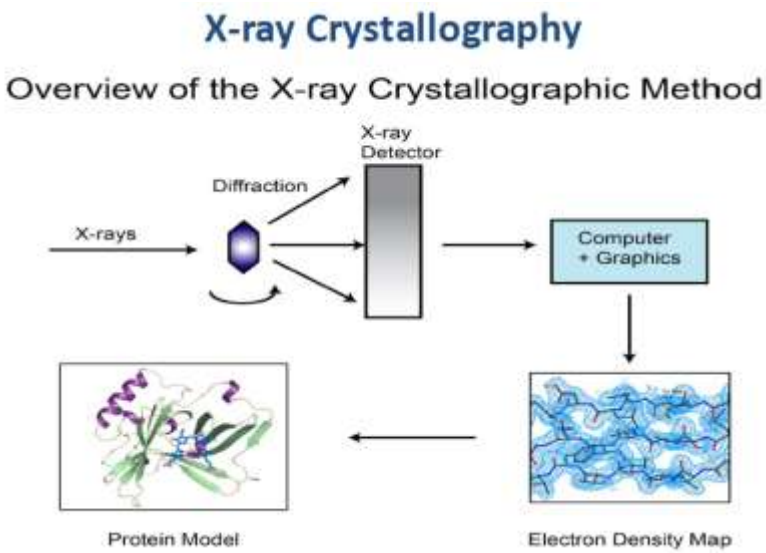
X-ray crystallography

The electromagnetic spectrum



To check positions of each atom in the protein, click on the 'Atoms' in the image above.

Overview of the X-ray crystallographic Method



X-rays emitted in the crystal, which will be diffracted and collected through the X-ray detector. Then using computer graphics you will have rough estimation of the location and position of each of the protein in the electron density map which is finally used for the protein model

X-RAY CRYSTALLOGRAPHY QUALITY PARAMETERS

When you look at the crystal structures of proteins, you will always here about some quality parameters like resolution and R-factor.

- According to the online dictionary of Crystallography the term resolution is used to describe the ability to distinguish between neighboring features in an electron density map
- R factor is one measure of model quality (The level of agreement between calculated and observed intensities) (0 to 0.6)
- Greater than > 0.5 is considered or poor quality

Example

X-ray crystallography

Resolution	Evaluation	Interpretation
1.2A	Excellent	Backbone and most side chains very clear. Some hydrogen may be resolved
2.5A	Good	Backbone and many side chains clear
3.5A	Ok!	Backbone and bulky side chains
5.0A	Poor!!!	Backbone mostly clear; side chains not clear

DATABASES

Databases for protein information include: PDB, Pfam, UniProt, Conserved Domains, Structural Classification of Proteins (SCOP), CATH, InterPro, Structural Biology Knowledge Base (with enough biological information) etc

www.wwPDB.org worldwide Protein Data Bank

RCSB PDB

- Repository of information about the 3D structures of large biological molecules
- Was established in 1971 at Brookhaven National Laboratory
- Research Collaboratory for Structural Bioinformatics (RCSB) became responsible for the management of the PDB in 1998.

There is yearly growth of total sequences and structure in the RCSB PDB database

PDB ID(s)

Structures in the PDB database have ID

- A 4-character ID e.g. 8CAT (alphanumeric)
- Unique, immutable identifier
- IDs are usually cited in papers
- The IDs can be used in searching other databases that recognized PDB IDs very well
- The IDs are automatically assigned and do not have meaning

DOMAINS

- **The domain is the basic building block of a protein structure**
- **Some proteins are composed of one domain, other have many domains**
- **This domains are usually:**
    - **1 – A spatially separated unit of the protein structure**
    - **2 – May have sequence and/or structural resemblance to another protein structure or domain**
    - **3 – May have a specific function associated with it**

Pfam

Is a database that information about protein families

- **Pfam 30.0 current version**
- **16306 entries (06.2016)**
- **Information about protein families that are represented in multiple sequences using (HMM) Hidden Markov Models**
- **Annotations**
- **Links to other databases: RCSB PDB, CATH, SCOP, Proteopedia etc.**
- **You can search using PDB IDs e.g. 8CAT**

CATH

Database that classifies proteins

The domains are classified within the CATH structural hierarchy:

- **Class (C) Level: Classification based on secondary structure content, i.e. all alpha, all beta, a mixture of alpha and beta, or little secondary structure**
- **Architecture (A) Level: The level based on arrangement in three-dimensional space**
- **Topology/Fold (T) Level: How the secondary structure elements are connected and arranged**
- **Homologous Super family (H) Level: Assignments are made if there is good evidence that the domains are related by evolution, i.e. they are homologous**

CATH

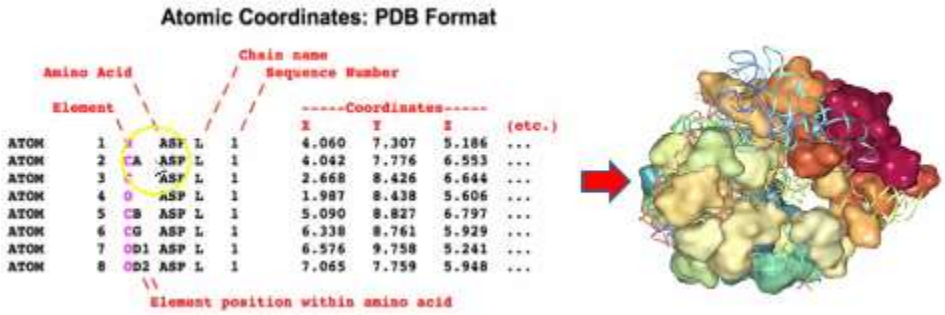
	CATH v4.1
PDB Release	01-01-2015
Domains	308999
Superfamilies	2737
Annotated PDBs	108378

Proteopedia

Wiki web resource whose pages have embedded three-dimensional structures surrounded by descriptive http://proteopedia.org/wiki/index.php/Main_Page It is beginners friendly site that you can visit to learn.

FILE FORMATS

- **The primary structure of proteins which is normally sequence file is represented in FASTA File**
- **Secondary Structure Files (FASTA-formatted file i.e. "ss.txt") i.e. FASTA file containing the sequence and some lines representing the secondary structures**
- **Secondary structures are usually represented in PDB entry files (namely PDB, PDBx/mmCIF, XML etc many of them). mmCIF is same as PDB but has no column limit can store large information**
- **The atoms and their locations as can be seen below from the Atomic Coordinates**



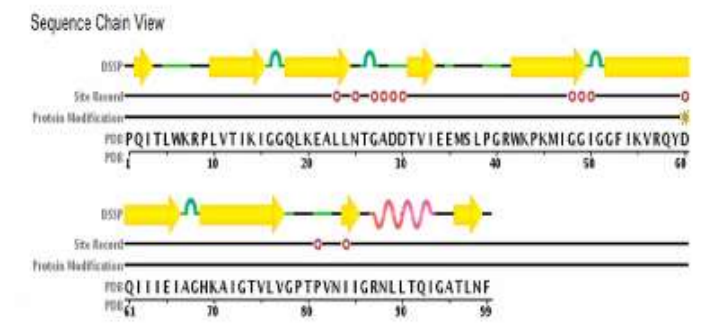
Atomic Coordinates of PDB Format shows the Element, Element Position within amino acid, Amino Acid name, Chain Name, Sequence Number, Coordinates X, Y and Z etc

- **Small Molecules Files can be in (PDB, CIF, SDF, etc…)**
- **Large Structures Represented in mmCIF/PDBx (containing >62 chains and/or 99999 ATOM records)**

FASTA-formatted file ("ss.txt")

- >101M:A:sequence
- MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTV LTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDI AAKYKELGYQG
- >101M:A:secstr
- HHHHHHHHHHHHHHGGGHHHHHHHHHHHHHH GGGGGG TTTTT SHHHHHH HHHHHHHHHHHHHHHHHHTTTT HHHHHHHHHHHHHTS HHHHHHHHHHHHHHHHHH GGG SHHHHHHHHHHHHHHHHHHHHHHHHTT >102L:A:sequenceMNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAAKSELDKAIGRNTNGVI TKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLQQKRW DEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYKNL
- >102L:A:secstr HHHHHHHHH EEEEEE TTS EEEETTEEEESSS TTTHHHHHHHHHHTS TTB HHHHHHHHHHHHHHHHHHHHH TTHHHHHHHS HHHHHHHHHHHHHHHHHHHHT HHHHHHHHTT HHHHHHHHHSSHHHHHSHHHHHHHHHHHHHSSSGGG

You can view the representation above in Sequence Chain View



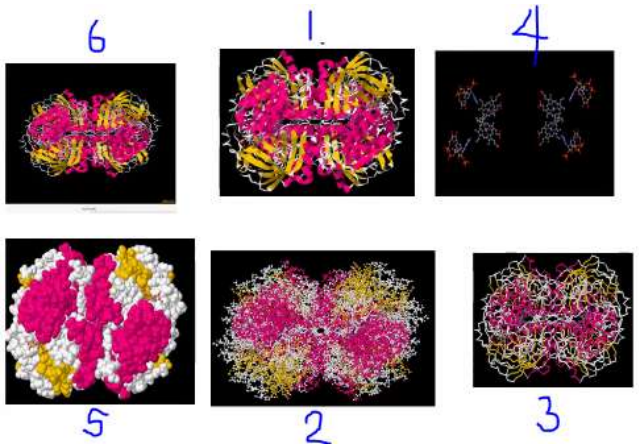
Molecular Graphics Software (This software are able to interpret and visualize PDB data)

- **Cn3D** http://www/ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
- **iCn3D** http://www.ncbi.nlm.nih.gov/Structure/icn3D/docs/icn3d_about.html
- **UCSF Chimera** http://www.cgl.ucsf.edu/chimera/index.html
- **Visual Molecular Dynamics (VMD)** http://www.ks.uluc.edu/Research/vmd
- **PyMOL** https://www.pymol.org/
- **Etc…**

MOLECULAR REPRESENTATION

There are many ways of representing; you will find them in the PDB database

1. **Cartoon Shape**
2. **…**
3. **Backbone Representation**
4. **Alignment Representation of**
5. **Space Form Representation**
6. **…**



STRUCTURAL BIOINFORMATICS

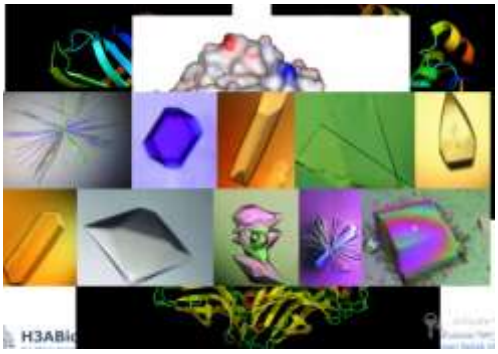
SOME QUESTIONS TO CLARIFY…

What do we mean by Structural Bioinformatics?

Structural Bioinformatics is a branch of bioinformatics that deals with structure of the biological macromolecules; DNA, RNA and Proteins (Deal-analysis, storage, visualization, prediction… etc)

Why protein Structure Bioinformatics?

- **Proteins are the building blocks of all cells;**
- **In the world of proteins; Structure = Function!?**
- **DNA encodes life… Yes! But proteins carry out life processes, replication, reproduction, defense etc.**
- **Study that will provide all the beautiful structures of proteins**



SESSION 2

COMPARATIVE PROTEIN STRUCTURE MODELING

INTRODUCTION

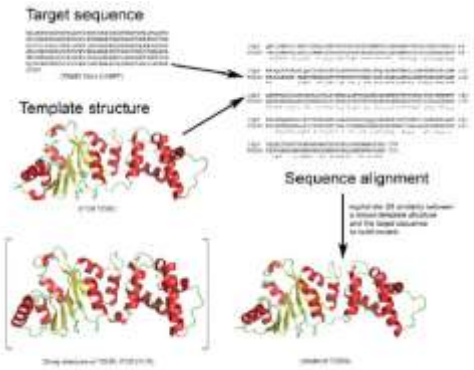
LEARNING OBJECTIVES

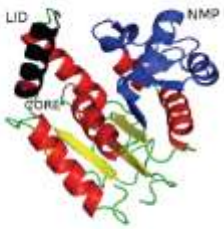
At the end of this lecture and practical you will know:

- **How to query the protein databank (PDB) and search for homologous protein structures (PSIPRED)**
- **How to model the 3D structure from a sequence and a known structure**
- **How to assess the quality of the 3D model**

NOMENCLATURE (Definition of some few terms…)

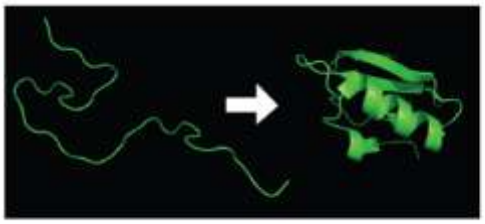
- **Homology: Sharing a common ancestor, may have similar or dissimilar functions**
- **Similarity: Score that quantifies the degree of relationship between two sequences**
- **Identity: Fraction of identical amino acids between two aligned sequences (case of similarity) i.e. physically counting the amount of residues and representing them as a fraction**
- **Target Sequence: Sequence corresponding to the protein to be modeled**
- **Template: 3D structure/s to be used during protein structure prediction (an experimentally solved structure, which shares homology to the target protein in this case)**
- **Model: Predicted 3D structure of the target sequence**

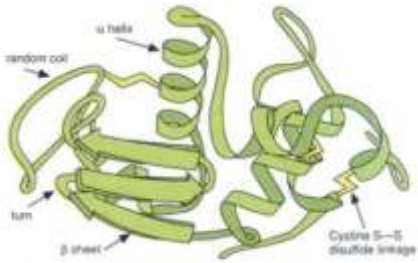




- **Fold: (Shape or confirmation of the protein sequence at domain level) Three dimensional conformation of a protein sequence (usually at domain level).**

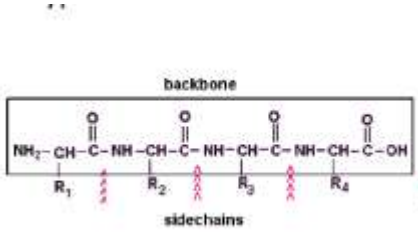
- **Domain:** Structurally globular part of a protein which may independently fold
- **Secondary Structure:** Regular sub-domain structures or patterns composed by alpha-helixes, beta-sheets and coils (or loops)

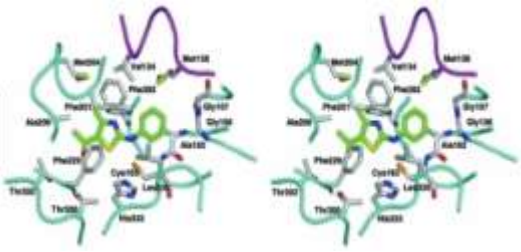




NB: Main difference between the secondary structure and the 3D structure of a protein is that the secondary structure contains no atomic 3D coordinate information.

- **Backbone:** Protein structure skeleton composed by the carbon, nitrogen and oxygen atoms
- **Side-Chain:** Specific atoms identifying each of the 20 residues types
- The main difference between protein lies in the different arrangement of this amino acid types





PROGRAMS, WEB SERVERS AND DATABASES

Programs (open source)

- **MODELLER:** (models 3D structures protein base on satisfaction of special restraints which are derived from alignment between the top sequence and template, these special restraints includes the distance, the dihedral angle obtained from the alignment. Implements most operations in comparative modeling
- **MODPIPE:** (Implements modeler and capable of calculating multiple or many comparative models from many sequences) Automatically calculates comparative models of many protein sequences

Web servers

- **SWISS-MODEL:** (it copies side chain directly from template to protein model) Automated homology modeling server
- **MODWEB:** (useful for comparative modeling) Provides a web interface to MODPIPE
- **MODLOOP:** (useful for modeling problematic loop regions within proteins) Models loops in protein structures
- **PSIPRED:** Threading (folding assignment) and secondary prediction (useful for forward assignment or threading and to predict secondary structures of proteins)

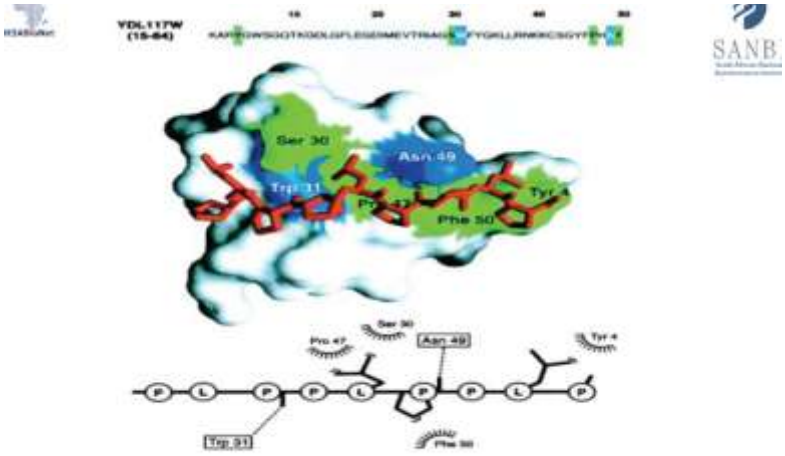
Databases

- **MODBASE: (contains predicted or theoretical models for sequences related to a known structure) Fold assignments, alignments models, model assessments for all sequences related to a known structure**
- **UNIPROT: contains thousand of protein sequences (comprehensive and free available database of protein sequences)**
- **PDB: repository for 3D structural data for large biological molecules that includes repository of X-ray, NMR and cryo-EM experimentally solved protein structures**

WHY IS IT USEFUL TO KNOW THE STRUCTURE OF A PROTEIN, NOT ONLY ITS SEQUENCE?

- **The biochemical function (activity) of a protein is defined by its interactions with other molecules**
- **The biological function is in large part a consequence of these interactions**
- **The 3D structure is more informative than sequence because interactions are determined by residues that are close in space but are frequently distant in sequence**
- **In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, structure is more conserved in evolution than sequences**
- **The net result is that patterns in space are frequently more recognizable than patterns in sequence**

For example in the image below, the model of yeast open reading frame protein shows residues Tyr 4, Phe 50 which are more closer in structure than in sequence, also residue Trp 31 and Asn 49 are closer in structure than in sequence. Most importantly about this protein or peptide complex is that the interactions here are determined by the binding sites rather than by its sequence.

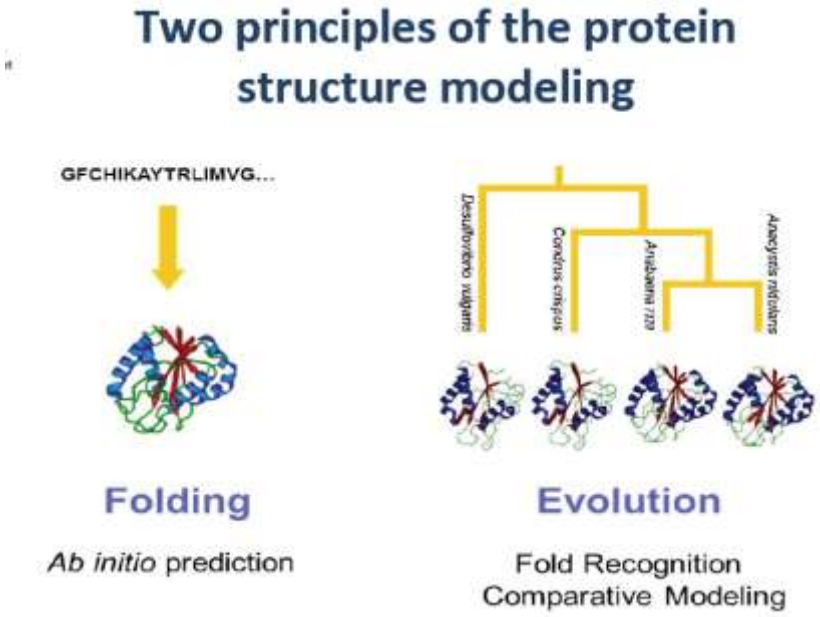


Modeling a yeast ORF SH3 domain-proline rich peptide complex interaction.
Interactions are determined by the structure of the binding site rather than its sequence

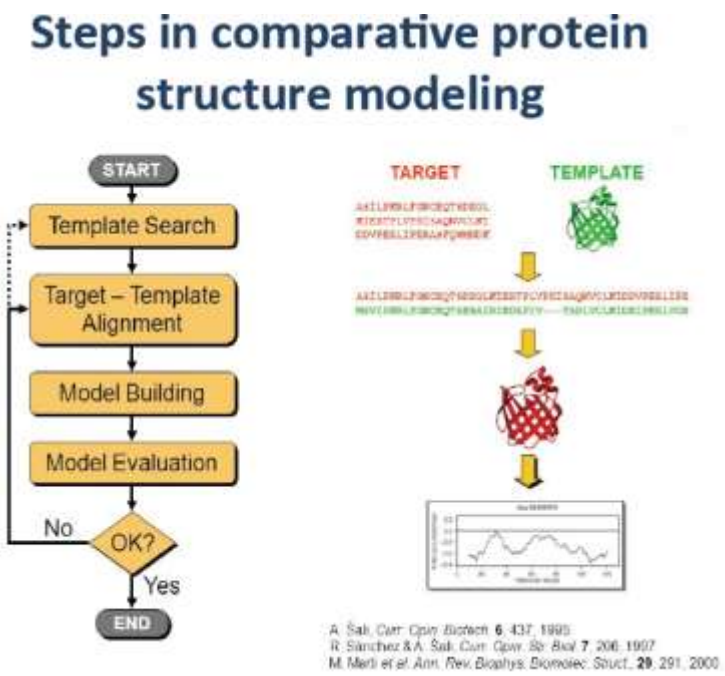
TWO PRINCIPLES OF THE PROTEIN STRUCTURE MODELING

1. **Folding: Uses Ab initio prediction. These methods are physics method and rely on atoms, electrostatic and vander-wal interactions. It performs only using the sequence and predicting the global energy minimum based on the folding of the protein structure**

2. **Evolution: Include Fold Recognition and Comparative modeling. Threading of fold recognition is structured base methods that are base on non-homologous protein structures. E.g. you take your protein sequence and align it, the secondary structure of your target sequence to a known structure to determine if there is any similarity in the folding pattern between this two sequences and structures. Comparative modeling relies on one or more related structure, identical at sequence level (here you use homology base methods).**



- **Steps in comparative protein structure modeling (starts with template identification, you take your target sequence and search for a known or a related structure within the protein database for example, create and generate an alignment between the two sequences (target and template). Then do the actual model building and afterwards model evaluation. If for some reasons the model does not satisfy the evaluation checks, then go back and repeat template identification or target – template alignment until you get a satisfy model.**



TEMPLATE SEARCH AND ALIGNMENT

How to find related structure to enable generation of alignment

- **Sequence-sequence search**
  - **BLAST (http://www.ncbi.nlm.nih.gov/blast/)**
- **Profile-Sequence search (more specific blast)**
  - **PSI-BLAST (http://www.ncbi.nlm.nih.gov/blast/)**
- **Profile-Profile search (implemented in modular)**
  - **pp_scan (mod9v15)**
- **Sequence-Structure Search**
  - **GenThreader (http://bioinf.cs.ulc.ac.uk/psipred/**

MODEL EVALUATION

We never know if our model is good or not, we grace by using varieties of potential energy functions, these functions have been derived from experimental observations using high resolution protein databank structures, base on the torsional or rotational angles solving the exposure and hydrogen bond geometry

- **Classical potential of mean force (is the first which tells us if a good model have a Z score greater than 0.7)**
  - **PROSA-WEB (https://prosa.services.came.sbg.ac.at)**
- **Protein conformation free energies (it looks at non-bonded interactions between atoms and a good model has very low energy score)**
  - **DFIRE (http://sparks.informatics.iupui.edu/yueyang/DFIRE/dDFIRE services**
- **Discrete optimized protein energy (use of potential energy function) (uses the dope score which is a statistical score, the lower the dope score of models, the closer the structure is to its native state and the more accurate prediction)**
  - **assess_dope (mod9v8)**

- **Stereochemistry (not necessary but useful) (assess the orientation of the file inside dihedral angles in the model and compare it to the protein with similar spherical shape and size)**
    - **PROCHECK (http://www.ebi.ac.uk/thronton-srv/software**

Some other useful… Model evaluation….

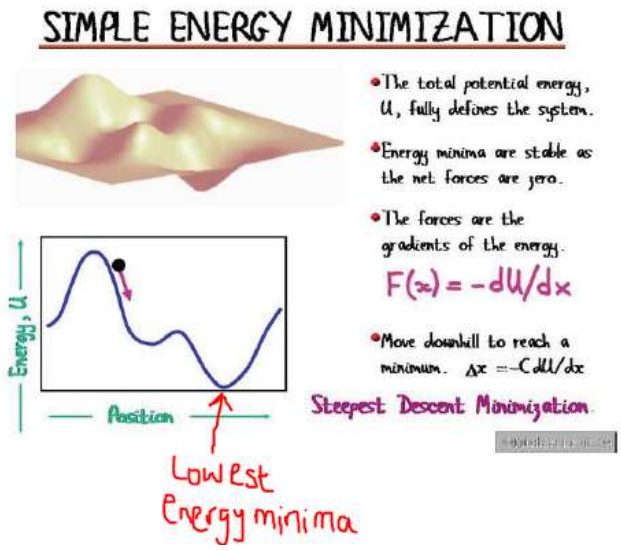
- **Anolea Atomic Mean Force Potential**
    - **ANOLEA (looks at regions of high energy in the model) and plotted as a line graph**
    - **http://swissmodel.expasy.org/workspace/>?func=tools_structureassessment1**
- **Composite scoring function for model quality estimation**
    - **QMEAN6 (most important score which is a comprehensive scoring function that includes rotational potential, distance dependent pairwise potential and solvent potential. A good model has a Qmeans6 = 0 suggesting there is no bad context in the model)**
    - **http://swissmodel.expasy.org/workspace/>?func=tools_structureassessment1**
- **Superposing model vs template (to measure deviation in the backbone atoms or all atoms**
    - **Root mean square deviation (RMSD) Align all atoms (PYMOL). If the RMSD is less than 1 this suggest homology between the structures and that the correct fold has been assigned to the model**

MODEL REFINEMENT

If we are fine with our model we can further refine the model using…

- **Energy Minimization (GROMACS or AMBER programs) to…**
    - **Relieve steric clashes between atoms**

      Simple energy minimization
    - ✓ **The total potential energy U, fully defines the system**
    - ✓ **Energy minima are stable as the net forces are zero**
    - ✓ **The forces are the gradients of the energy F(x) = -dU/dx**
    - ✓ **Move downhill to reach a minima DeltaX = -C dU/dx**
    - ✓ **Steepest Descent Minimization**

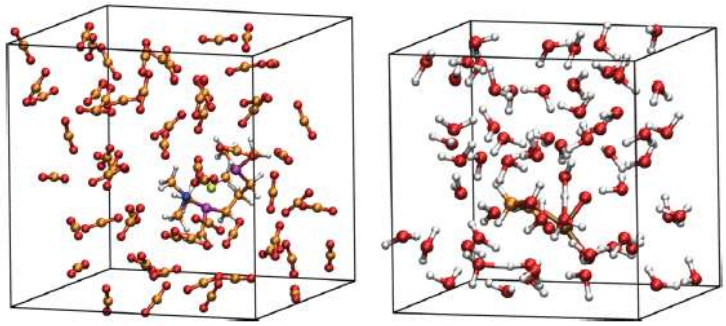


Energy minimization is like walking on mountain and you meet area where it is steep, then you move into a valley where there is low energy then you go up slightly and then down to the next energy value we want to calculate the lower energy minima.

Another method of Model Refinement…

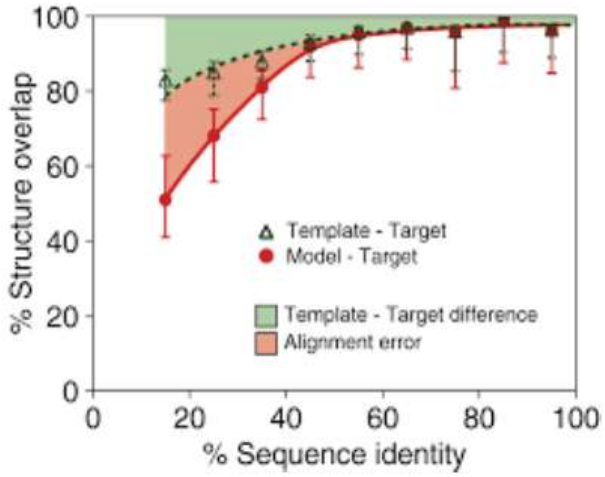
- **Molecular dynamic simulations**
  - **GROMACS and AMBER etc…**
  - **Move atoms toward the global energy minimum to find the near native conformation by applying heating, cooling and water molecules to solve Newton's laws of motion**

As can be shown in the two boxes below, atoms moving around to find the global energy minimum of this atom is where the least clashing is occurring.

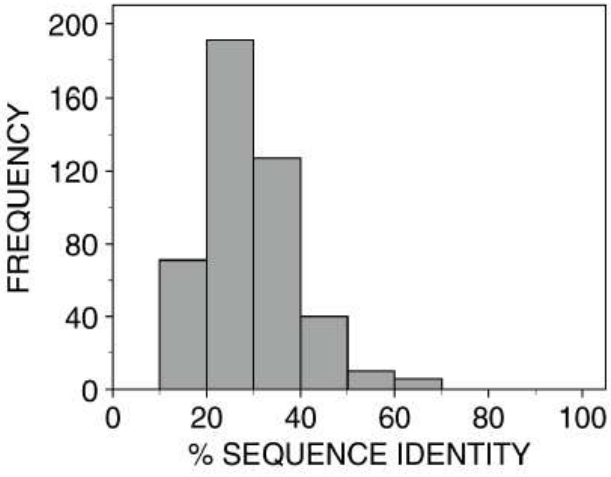


ACCURACY AND APPLICABILITY OF MODELS

Comparative modeling by satisfaction of spatial restraints types of errors and their impact on force potential



A) Percentage structural overlap vs sequence identity

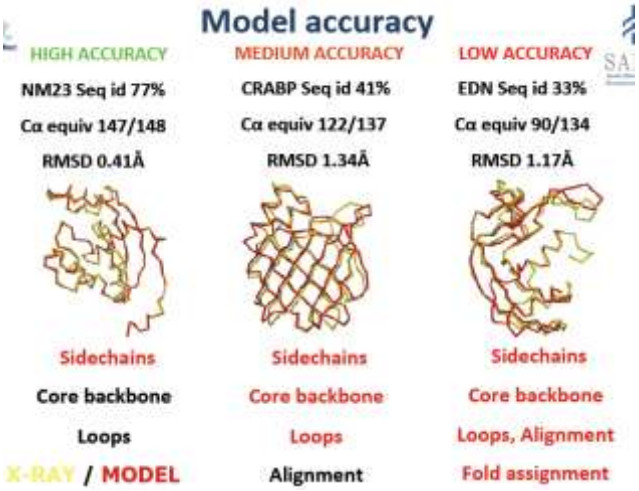


B) Frequency of errors vs % sequence identity

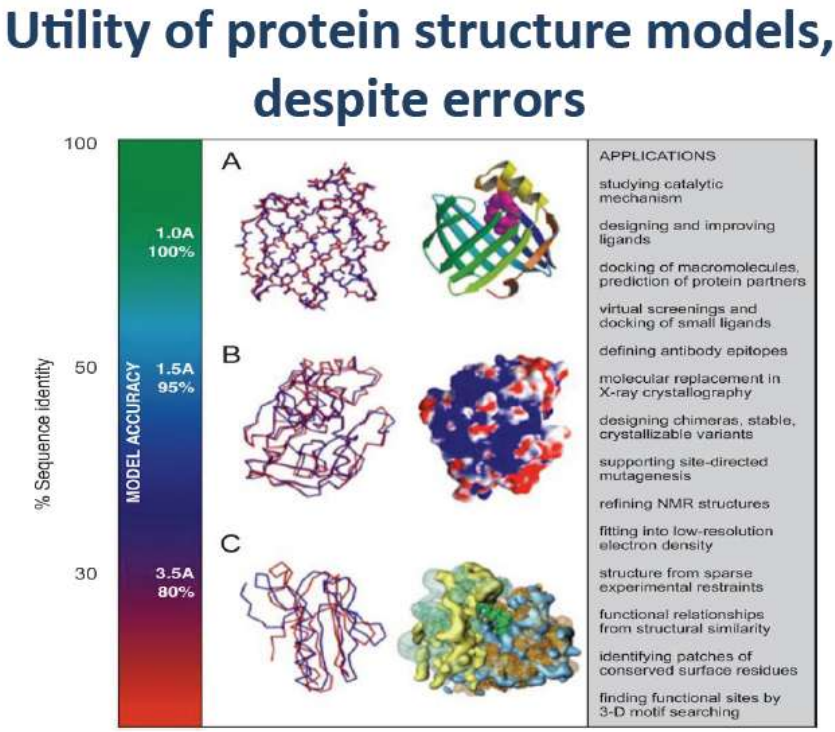
(a) Percentage structural overlap as sequence identity (b) Frequency of errors vs % sequence identity
A correlation between model accuracy and sequence identity, YES in Figure A the about of ethers decreases and reaches a Pluto, as the structural overlap increases and the amount of sequence identity.
The figure B side shows that the frequency of ethers drastically decreases as the sequence identity increases between your target and template

MODEL ACCURACY

High Accuracy	Medium Accuracy	Low Accuracy
NM23 Seq id 77%	CRABP Seq id 41%	EDN Seq id 33%
Cα equiv 147/148	Cα equiv 122/137	Cα equiv 90/134
RMSD 0.41A	RMSD 1.34A	RMSD 1.17A
Due to… correct	Due to incorrect….	Considerable amount…
Sidechains	Sidechains	Sidechains
Core backbone	Core backbone	Core backbone
Loops	Loops	Loops, Alignment
X-ray / MODEL	Alignment	Fold, Alignment



These are things to be aware of



With Ab initio models we can use this to establish evolutionary relationships

UTILITY OF PROTEIN STRUCTURE MODELS, DESPITE ERRORS

APPLICATIONS

A Model Accuracy 1.0A 100% % Sequence Similarity 100

- **Studying catalytic mechanism**
- **Designing and improving ligands**
- **Docking of macromolecules, prediction of protein partners**
- **Virtual screenings and docking of small ligands**

B Model Accuracy 1.5A 95% % Sequence Similarity 50

- **Defining antibody epitopes**
- **Molecular replacement in X-ray crystallography**
- **Designing chimeras, stable crystallizable variants**
- **Supporting site-directed mutagenesis**
- **Refining NMR structures fitting into low-resolution electron density**

C Model Accuracy 3.5A 85% % Sequence Similarity 30

- **Structure from sparse experimental restraints**
- **Functional relationships from structural similarity**
- **Identifying patches of conserved surface residues**
- **Finding functional sites by 3D motif searching**

MODULE 7

Evolution and Phylogenetic

Introduction: Historical perspective, mechanisms of molecular evolution and speciation

LEARNING OBJECTIVES

- **Understand the history of evolutionary thoughts, from 18th to 20th century**
- **Know about the modern concepts of evolution, species and speciation**
- **Understand the different mechanisms of evolution at molecular level**
- **Understand the mechanisms of speciation at population level**

LEARNING OUTCOMES

- **Be able to name a few figures of early evolution theory and identify their contribution**
- **Know what are the different mechanisms involved in evolution**
- **Know what are the different mechanisms leading to speciation**

Evolution and Phylogenetics

INTRODUCTION: Historical Perspective

Understanding the long way we have gone through since the beginning of 18th century.

Carl von Linne (Linnaeus), 1707 – 1778

- **Swedish zoologist and botanist (studies both animals and plants)**
- **Father of modern taxonomy (nomenclature and classification of organisms): kingdoms, classes, orders, genera, species, subspecies/varieties**
- **Proposed classification according to physical characteristics**
- **Formalized the binomial nomenclature used since then to names species (e.g. *Homo sapiens, Canis lupus, Oryza sativa*)**
- **Systema Naturae (1735-1758). Book publication title**
- **Pointed out the relatedness between humans and monkeys/apes**
- **Paved the way for the advent of the theory of evolution**

Linne's idea was to group in hierarchical way different organisms from bottom to top starting with classification in varieties of sub-species, then species, genera, orders, classes then kingdoms. Animal kingdom is a different kingdom from plant and bacteria. Sub-species and varieties define the lowest level of the hierarchical called clustering. Sub-specie term is used when talking about animals, varieties is used when talking about plants. Linne's classification was according to physical characteristics (scientific way of classifying beings, organisms to push forward base on observations of the physical characteristics).

Using two components where one component identify the genus and the second component identifies the species e.g. Homo sapiens for humans, Canis lupus for dogs, and Oryza sative for rice.

Jean-Baptiste de Lamarck (1744-1829)

- French soldier, then botanist, academic and naturalist
- Flore francaise (1779): First publication on plant
- Hisotire des animaux sans vertebres (1815)
- Coined the term "biology"
- Strong interest in what tells apart living systems from mineral/inorganic stuff ("force vitale")
- Influenced by 17th and 18th century concepts of mechanical/alchemical biology (vital force, fluids…)
- Influential (?) to Darwin's future work

Unify the science activities. Interested in what tears apart living and non-living systems, he said there is vital force that tears both apart. He believed in alchemy called pseudo-science today.

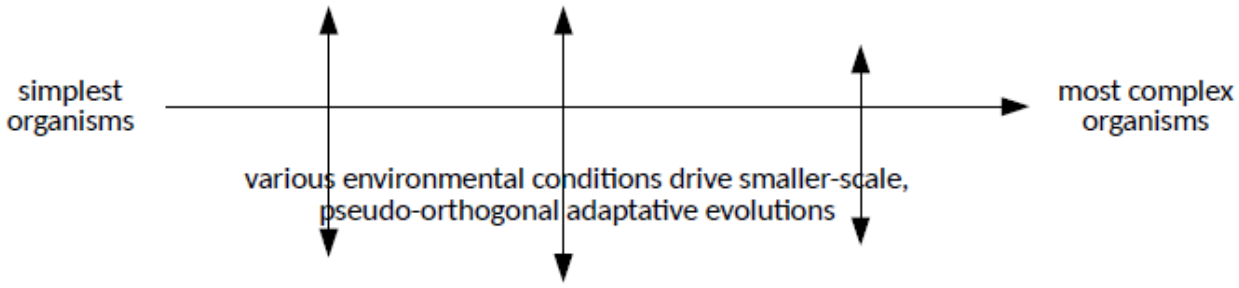
Lamarckian evolution: a theory

Lamarck viewed evolution as a process driven by 2 main forces:

1) Organisms evolve towards greater complexity from most basic organisms produced by spontaneous generation ("infusories")

2) a second driving force make them evolve and diversify to adapt to their various environments (e.g. through use/disuse of organs)

Simplest Organisms **Most complex organisms**



Various environmental conditions drive smaller scale pseudo-orthogonal adaptive evolutions

In the orthogonal driving force which meant to let organisms adapt to their various environment, Lamarck have a different way taking example of giraffe, where he said when you make use of a particular part of the body consistently it gets stronger from the body fluid, so that giraffe will have longer necks and will transfer such traits to their offspring. If you don't use this organ according to Lamarck, then the vital fluid will slowly lead to disability of the organ.

Charles Robert Darwin, 1809 – 1882

- British geologist, taxonomist, academic and naturalist
- Drew significant inspiration from his trip around the world on the HMS Beagle (1831 – 1836)
- The Voyage of the Beagle (1839)
- On the Origin of Species (1859): publishing evidence to support his new theory of evolution
- The Descent of Man (1871)
- Father of the term "natural selection"
- Backed the idea of evolution as descent with modification
- Made a strong case for common descent (a single ancestor to all beings)

Darwinian evolution: Natural Selection

The concept of natural selection is key to Darwin's theory of evolution: "As many more individuals of each species are born than can possible survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in a any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be naturally selected. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form". On the Origin of Species (1859)

Idea of survival of the fittest. Phenotypic i.e. observable characteristics are transmissible from parents to the offspring (one generation to the next).

Darwinian evolution: Transmutation of species

- Some species are extinct. We can see fossils of such species
- Species are not fixed
- Distinct but close species are differentially distributed e.g. over islands (observations on rheas, Galapagos mockingbirds and finches)
- Possible that "one species does change into another" (speculation in Darwin's Red Notebook)
- Variation in offspring "to adapt and alter the race to changing world"
- Character variation and fixation of changes in a population provide the mechanisms of speciation.

Darwin's "*B notebook*" on the "transmutation of species" (1837)

First phylogenetic tree first drawn by Darwin

In summary the fact that Linne was busy trying to understand how to classify individuals, Lamarck and Darwin were busy trying to understand how organisms evolved. Darwin wants to understand when is it that species appear?

Darwinian evolution: Common Descent

Darwin was the first to get the intuition that maybe all species derived from a single ancestor: "There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved". **On the Origin of Species (1859)**

Darwinian evolution: a few remarks

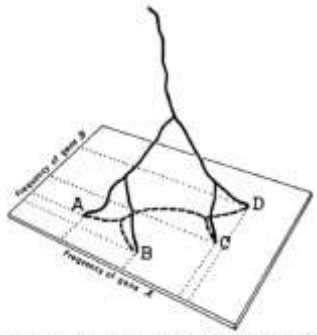
- **Struggle for life stems from the fact that more individuals come to life than can possibly survive long (competition for resources with survival of the fittest)**
- **It is absurd to talk of one animal being higher than another" → evolution does not yield "higher", "better" beings down a one-way alley. It is a bushy, omnidirectional process.**
- **Evolution has no fixed goal**
- **Human beings are animals (continuity of characters)**
- **Sexual selection plays a role in evolution**
- **Diseases also play a role (negative selection by death)**
- **Darwin's breakthrough opened the way for modern molecular evolutionary science**

20th Century: the advent of Computational Evolution

- **1950s: numerical clustering methods for classification and phylogenetic inference (Sokal & Sneath)**
- **1953: F. Sanger achieves first protein sequencing**
- **Same year: Watson & Crick publish the double-helix structure of DNA**
- **1955: Discovery of the DNA polymerase (Kornberg, Lehman & all)**
- **1960s: Phylogenies from gene frequencies, parsimony methods, likelihood methods (Edwards & Cavalli-Sforza)**
- **1960s & 1970s: Scientific computers and mini-computers become affordable (IBM 7040, PS/900…)**

Cavalli-Sforza & Edwards 1967

- **Studying gene frequencies among various populations**
- **Data: current gene frequencies for genes A and B in pop. A,B,C and D**
- **Problem: infer the evolutionary tree in space and time**
- **Necessary problem-solving step: make hypotheses about mode and speed of evolution (model)**
- **Authors go through a comparison of two approaches: minimum evolution and maximum likelihood**



An evolutionary tree and its projection onto the "now" plane

Tracing back the evolutionary history of the population, using gene frequencies; it is plotted on the image above to show areas A, B, C and D covers. Knowing what happens on the branches of the hypothetical tree (speed of evolution) minimum evolution and maximum likelihood.
The evolutionary tree is contains the idea of time (from ancient time to now). A and B can be said to be closer than to C and D. The same applies to C and D been more closer.

MECHANISMS OF MOLECULAR EVOLUTION & SPECIATION

Various mechanism drive evolution (evolution is driven by various factors, four important ones in terms of driving evolution include):

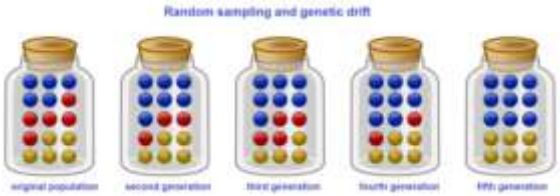
- **Genetic drift: random sampling from the gene pool of generation n to the genetic material of generation n+1 (rate of the random walk depends on the effective size of the population) Is the most important force, base on model where population size is constant and having n individuals in generation n and in generation n+1 you also have n+1 individuals who are offspring of individuals from generation n. This means that only some of the individuals in generation n have viable offspring that are living in generation n+1 and that are susceptible to give in turn their own legacy i.e. their own genetic materials to generation n+2. The randomness of the sampling is the force we called genetic drift i.e. genetic drift is associated to randomness. E.g. random selection, random mating etc**
- **Selection: processes of natural (adaptation to the environment or lack thereof) or sexual selection (bias in mating patterns in favor of individuals having certain positive features). Second force driving evolution is Selection, it can be Natural, Sexual or Artificial. Natural selection is what takes place where in order to adapt to the environment an individual will be positively selected, if that individual has the features that make him/her fit to the environment. E.g. if climate changes on planet surface from very cold to very hot, animals will much fur, long hair etc will be at the disadvantage because they are fit for cold environment. Sexual selection happens with bias e.g. female tend to mate with males who are strong and free from diseases, advantage for the offspring to benefit so mating here is not random. Artificial selection is driven by human beings e.g. selecting cows with much milk production in favor to produce lots of milk etc.**
- **Migration: geographic migration of groups of individuals to settle elsewhere, forming subpopulations. This population in turn will evolve into separate species. Adaptation of the new environment will also play in diverging this population from the original population.**

- **Mutation/recombination:** pseudo-random alteration of the genetic material during the creation of an offspring (inheritance with modification). Examples: nucleotide substitutions, chromosomal crossover, duplications and gene translocation etc. Mutation do not appear very often, inheritance of random genes in different locus.

Genetic Drift

By chance only, some individuals of generation n leave more offspring than others to generation n+1, hence a modification in allele frequencies (in any other genetic characteristics)

Random sampling and genetic drift



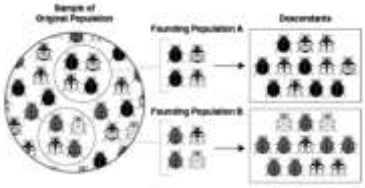
The left most jar represent the original population with 1/3 of blue, 1/3 of red and 1/3 of yellow. By just randomly picking some of these beads with replacement makes the second generation with same quantity of beads, subsequently doing same to create other jars. It shows that the red can simply disappear and if the last jar fifth generation without red is maintaining isolation, then you will never get the red beads; by genetic drift the outcome will most probably be having Fixation. Fixation is the phenomenon that happens when there is only one color ultimately in some generation. Genetic drift happens even without any selective pressure (positive or negative), there is Genetic drift.

Genetic Drift: Bottleneck Effect

There is one specific type of genetic drift that happens with bottleneck effect which occurs when a drastic decrease in population size occurs (e.g. massive hunting, ice age, severe environmental change), only a few individuals survive, hence a loss in genetic diversity. Few survivals after catastrophic reduction will then grow into a new population.

Genetic Drift: Founders Effect

When a small subgroup parts to found a new population; the new population has less genetic diversity than the original population.

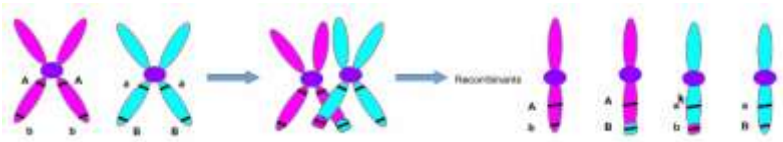


Small sample of Founders A and Founders B will have descendents will different characteristics because the genetic diversity (among of variance found in the genetic material) have reduced from the original population to populations A and B.

One of the example is the population in United States, in Ohio there is inbreeding of small individuals (Armish where people have more than 5 fingers or toes) the genetic diversity is smaller than any population else in the planet.

CHROMOSOMAL CROSSOVER DURING MEIOSIS

- **Key mechanism to ensure that organisms in generation n+1 are made of a unique makeup from the genetic material of their two parents (genetic diversity)**
- **Takes place during meiosis in diploid organisms using sexual reproduction, when haploid cells (gametes) are formed from diploid cells**
- **Genetic diversity is key to adaptation (plasticity of species)**



Humans have copy of 1 pair of chromosome from both parents (haploid cells) during meiosis arms overlap each other, two chromosomes exchange materials before been split into their arms. Resulting to genotype Ab and AB which wasn't in the parent AbAb and BaBa. In only one pair of chromosome and without chromosomal crossover then parent in this case will only be able to give birth to four genetically different children (4 genetically different children), because we have 4 different combinations without chromosomal crossover. Therefore with chromosomal crossover, you can have many more offspring (depending on the pool of haplotypes i.e. series of alleles found on one chromosome, you make a pool that is much richer.

Genetic diversity is key to adaptation because from population where you have four 4 different haplotypes, you have many chance of finding haplotypes that is positively selected by the environment than in a population where you have only two 2 haplotypes

WHAT IS A SPECIES?

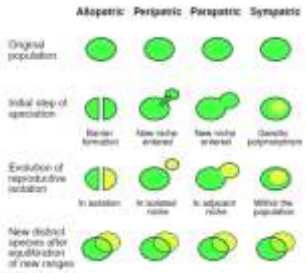
Historical definition by Ernst Meyr (1937), widely accepted:

- **A species is formed of a set of individuals who can interbreed to give viable and fertile offspring, while not being able to breed and give such offspring with members of other species. (Offspring been able to reach maturity and repeat same)**

When geographical isolation or artificial selection or traits provide significant phenotypic variation without preventing viable interbreeding, we talk about subspecies (animals) or varieties (plants).

Example: wolf (Canis lupus) and dog (Canis lupus familiaris or Canis familiaris (for dog that is domesticated)) do not (yet) form two distinct species. Because they do not form two distinct species, Wolf and Dog can interbreed and give viable offspring.

Different species can emerge from single one in a population



Four different patterns of speciation include: Allopatric, Peripatric, Parapatric and Sympatric

Allopatric speciation enters into play when there is formation of a strict geographical barrier splitting a population into two same populations, not allowing migration of individuals from one sub-specie to the other sub-specie. We will have the two populations evolving in isolation without any genetic flow (exchange of genetic material between some populations). The effect will be genetic differentiation, so their genetic materials will be different. Individuals from each side can no longer breed.

Peripatric speciation happens when a few individuals from the original population migrate to enter a niche, and the niche becomes geographically separated, separate isolated evolution where there is no inter-mating. The two populations will come to form different species.

Parapatric speciation is when there is no clear geographical separation between the two populations, but somehow one sub-population takes advantage of one feature in the environment that is not choose by the other population. E.g. where two populations of flowers bloom or flower different times in a year, it prevents gene flow between the two population because if the flowering period differ and they are in same environment they will not be able to exchange genetic material. They will be reproductively isolated from each other, making it possible for the two to form different species, here it means pollination of the both in a lab won't be successful.

Sympatric speciation here same they are in same geographic arena, there is natural appearance of genetic polymorphism, the small population will tend to mate with each other, because there is mechanisms of sexual selection in play that makes them favor themselves. E.g. population of dogs where the bigger sizes tend to mate each other because of their breed, German shepherd will mate other German shepherd. Here there is less inter-breeding.

All these patterns of speciation happens because at some point the gene flow between the two same population is reduced or highly unequal like in Peripatric.

A mechanism that plays against speciation is mechanism of gene flow, when there is migration between individuals belonging to one sub-specie towards one other, there is exchange of genetic materials, and this exchange between two same populations tends to counter this pattern of speciation, because if there is high or considerable gene flow between two populations they do not evolve in reproductive isolation, so there is divergence in gene content between two population tend to become negligible when there is high gene flow.

NB: Species defined as been entities or populations with limited to no gene flow between each other.

SESSION 2

Evolution & Phylogenetics

Traits, Phylogenies, Evolutionary models and divergence time between sequences

LEARNING OBJECTIVES

- **Understand the concepts and the vocabulary pertaining to phylogenetic trees**
- **Know from what type of biological data it is possible to build phylogenies**
- **Understand the concept of evolutionary divergence between sequences**
- **Understand what evolutionary models are and know how to use them**

LEARNING OUTCOMES

- **Be comfortable in discussing with a colleague about the parameters and details involved in running a phylogenetic analysis**
- **Be able to prepare a phylogenetic analysis (choice of the data and of an appropriate model of evolution), upstream of the phylogenetic inference process itself**

Traits, Taxa, Phylogenetic Trees

Some vocabulary

- **A phylogeny a.k.a phylogenetic tree, is a tree (connex acylic graph) whose edges represent direct evolutionary links and nodes represent past or present taxa**
- **A taxon (plural taxa) is a member of a taxonomic representation (species, virus strain, variety, identified subpopulation, etc)**
- **Terminal nodes (leaves of a tree) are the extant taxa, a.k.a OTU (Operational Taxonomic Unit) extant taxa are tools we can connect data to nowadays**
- **A trait or character is any biological feature that can be compared across taxa**
- **Traits can be qualitative/categorical variables (e.g. aligned nucleotides or amino acids i.e. non numeric) or quantitative, in which case they can be discrete, semi-continuous or continuous (e.g. number of repeats of a microsatellite because they differ in one specie to another, frequency of an allele, diameter of the skull etc)**

PHYLOGENETIC TREES

A mathematician's definition:

- **Tree topology: connex acyclic graph G = (V.E) i.e. Topology means the shape of the tree. Connex means you have only one component fully connected, there is path linking one node to every other node**
  - **V: set of vertices or nodes (e.g. species, virus strains, genes)**
  - **E: set of edges or branches (materializing evolution)**
- **Acyclicity and connexity impose that there is exactly one path between any two nodes of the three. Acyclic means there is no cycle in the graph**
- **The degree of a node is the number of edges it is connected to**
- **Branch lengths are not always present. It present, they correlate with the amount of time elapsed (branch length unit: expected number of character substitutions per site).**

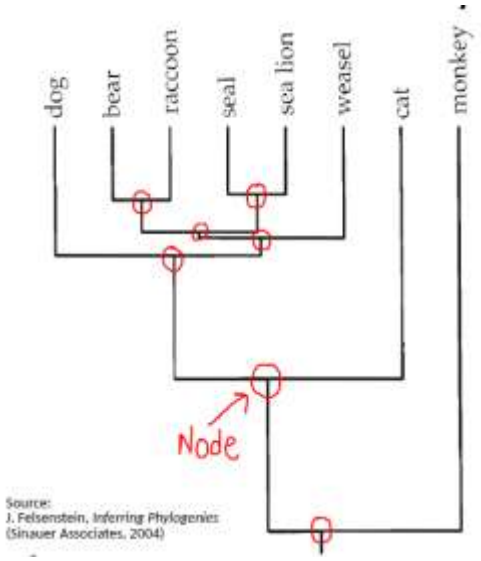
- **Tree is rooted if we know where is the most ancient node (evolutionary origin), unrooted otherwise**

Some more vocabulary and remarks

- **In a binary tree, all non-terminal nodes have 3 neighbors (or one parent node and two children in case the tree is rooted)**
- **A node with degree > 3 is called a multifurcation**
- **Multifurcations represent well geological "moments" of sudden adaptive radiation e.g. Cambrian explosion (lots of species develop, appear and develop, in short amount of geological time)**
- **Trees can be built on a single trait (e.g. one phenotypic characteristics) or (most commonly) on a set of characters (e.g. 1000 aligned sites). The tree will not be trustable because it is base one phenotypic characteristics. In order to increase the amount of phylogenetic signals we use set of characters not only one e.g. alignment made of 1000 sites will provide more accurate result than fewer size.**

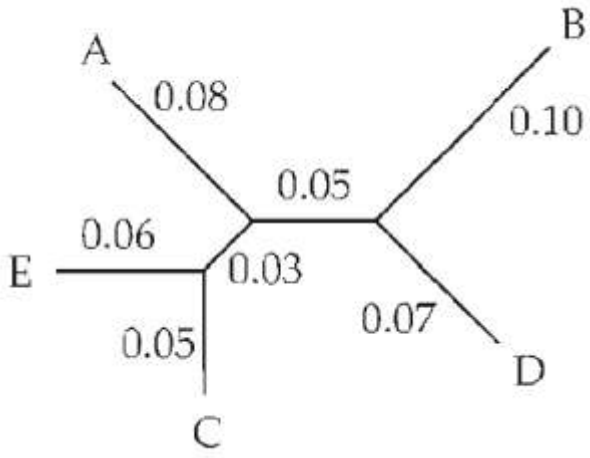
A rooted binary tree

- **In a rooted binary tree, every internal node (except the root) has one father and two descendants (sons)**
- **n taxa will give you**
  - **2n-1 vertices**
  - **2n-2 edges**



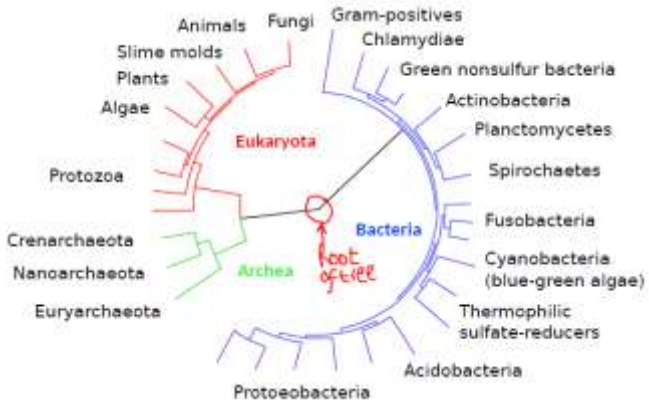
An unrooted binary tree

- **In an unrooted binary tree, every internal node has three neighbors**
- **n taxa will give:**
  - **2n-2 vertices**
  - **2n-3 edges**
  - **This tree has branch lengths**



Circular trees: A fancy representation

- **Another way of drawing tree, can have binary tree which in each node has 3 neighbors**
- **Linking the Eukaryote where humans belong, Archea and Bacteria**



The tree inference problem

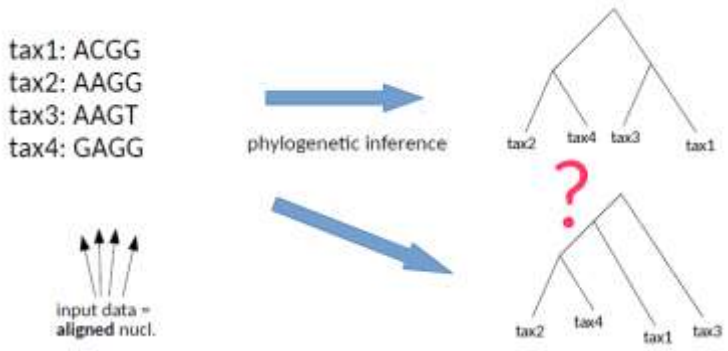
Problem: Assuming common descent, how to derive the "most probably correct" tree from the knowledge of the traits in the extant taxa (leaves)?

Tax1: ACGG

Tax2: AAGG

Tax3: AAGT

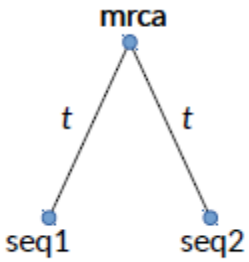
Tax4: GAGG



From the alignment of the sequence one can get the correct evolutionary tree. It is difficult to select the appropriate tree

Simpler Problem: Evolutionary distance between two sequences
Let us consider two aligned sequences and the elementary tree linking them:
Seq1: ACGGGTATTG
Seq2: ACGATTATTT

seq1: ACGGGTATTG
seq2: ACGATTATTT



We want to know the evolutionary time (=2t) separating seq1 and seq2 from their MRCA (Most Recent Common Ancestor).
Naïve distance = edit distance = 3/10 = 0.3 (Where nucleotide differs)

Shortcomings of the edit distance between two sequences

- **All mutations (transitions or transversions) yield same distance while from a biochemical standpoint, transitions (within purines, A ←→ G, or within pyrimidines, C ←→ T) are "less costly" than transversions.**
- **Temporal chains of successive mutations on the same site result in underestimation of the true evolutionary distance by the observed (edit) distance, e.g. A → G → T (edit distance 1, true distance 2) or A → G → A (edit distance 0, true distance 2). Silent mutation e.g. from root with A it can mutate to G and G in seq1 and seq2 respectively.**
- **We have no means to translate edit distances into actual biological time, and date the MRCA in Mya (= "Million years ago"): problem of calibration.**

MODELLING EVOLUTION

What is a model, in general?

- **Simplification of the reality based on current human understanding**
- **Must be simple enough to be computationally tractable (Complex model lead to protracted computation time and impossible to implement in programs)**
- **Must be complex enough to lead to measures, findings or simulations in approximate agreement with our observations (empirical truth). Goal of the model is to mimic the reality.**
- **Tradeoff between the complexity of a model and its tractability (a.k.a. "computability", which implies practical usefulness). Too simple model will not be realistic.**
- **Einstein: "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum experience" or "Everything should be made as simple as possible, but not simpler".**

What is a model of evolution? Specifically?

- **Deals with aligned sequences (nucleotides or amino acids) i.e. sequence must be aligned first else nothing can be done.**
- **All evolutionary models are stochastic: they predict probabilities of change, without yielding any certainty**
- **Models evolution in terms of character substitutions: enables to calculate e.g. $P_t(A \rightarrow C)$**
- **Is defined with a certain number of parameters to be estimated from some training data**
- **Essentially all models in use are Markovian (memory-less): the fate of a character depends only on its present state, not on its previous history of mutations**
- **Some models are time-reversible, some others are not**

Stochasticity

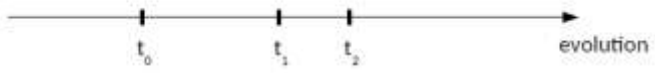
Stochastic (probabilistic) models are based on stochastic processes i.e. processes defining the trajectory of some random variables.

Examples of models:

- **Non stochastic: "nucleotides mutate once every thousand years according to the cycle: A→C, C→G, G→T, T→A"**
- **Stochastic: "nucleotide substitutions occur randomly at a constant rate α (expected number of substitutions per site per time unit), equal for all substitutions." (JC69)**

Markovian property

The evolutionary history of a character X doesn't further impact its future trajectory, only its present state matters to determine its future (memory-less property).



$$Pr(X(t_2)|\{X(t_0), X(t_1)\}) = Pr(X(t_2)|X(t_1))$$

timer: to, t1 and t2 into evolution

Probability of X at time t2 knowing the state of character X to and X t1. Meaning if you trying to predict character at t2 knowing the t1 you don't need any knowledge for anything before t1 i.e. t0. Implies probability of X t2 and state of X t1.

Stationarity

Stationary models accept limits for the frequencies of the characters, after the stochastic model has run for an "infinite" amount of time. For instance.

$$\lim_{t \to \infty} Pr_t(A \to C) = \pi_C$$

The limit of the probability of an A to become a C after an infinite amount of time, is called the stationary probability for character C. If a model is stationary it means that it converges, if you let the model run, with the infinite branch link you will see at the end of the branch link is distribution of possible nucleotides that do not depend on what was at the beginning.

Time-reversibility

- **Time-reversible models give no preference to evolution "in one direction" compared to "in the opposite direction", e.g. on some branch of a tree: they give no indication about the orientation of the time arrow.**

$$Pr(A)Pr_t(A \to C) = Pr(C)Pr_t(C \to A)$$

The probability of A mutating into C and probability of C mutating to become A

- **Consequence: time-reversible models cannot lead by themselves to a rooted tree. Unrooted trees are built first, and then rooted by some extra information (outgroup).**

Species that is far away from the rest of the species in the tree to determine the root of the tree.

Exchangeability Rates

When dealing with time-reversible models, on has:

$$\pi_A Pr_t(A \to C) = \pi_C Pr_t(C \to A)$$

And using instantaneous rates:

$$\pi_A q_{AC} = \pi_C q_{CA}$$

Substitutions of equality

So we can define symmetric exchangeabilities:

$$\frac{q_{AC}}{\pi_C} = \frac{q_{CA}}{\pi_A} = s_{AC} = s_{CA}$$

Exchangeabilities between characters only in the context of time-reversible models.

SIMPLEST EVOLUTIONARY MODEL ON NUCLEOTIDE DATA: JC69

JC69: Jukes and Cantor, 1969

- **Single constant rate for all substitutions: model has only one parameter $\alpha \in R^{-}$ (odd situation for $\alpha = 0$: no evolution) i.e. it should be a positive number**
- **Only substitutions, no insertions and deletions**
- **→ Model explains or creates only ungapped alignments**
- **→ Model can be trained on relatively small datasets**
- **- oversimplification of the reality: observations give credit to varying substitution rates depending on the position in the genome or in the Tree of Life (the fact transitions are more common than transversions, we should have higher rate for transitions than transversions, the Jukes and cantor works with one parameter).**
- **Advantage of JC69 is that you just need + very simple maths to calculate the probability of one sequence having evolved into another one**

JC69 instantaneous rate matrix

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

All none diagonal coefficients equals α

For instantaneous rate matrix, the coefficient on each line must sum up to zero that is why we have the -3α to balance the column into summing up to zero

$Q_{AC} = \alpha$

Matrix Q of instantaneous rates: over a very short duration dt, probability α*dt that e.g. a nucleotide A mutates into a C.

NB: As soon as you define what the value of α alpha is, you have defined what is your JC69 model i.e. only one parameter.

Dynamics of Nucleotide evolution under JC69

- **We consider a population of N nucleotides evolving independently from t = 0 under the JC69 model**
- **$n_A(t)$: count of adenines at time t**
- **Calculus leads to: $n_A(t) =$**

$$n_A(t) = \left[n_A(0) - \frac{N}{4}\right] e^{-4\alpha t} + \frac{N}{4}$$

  Formula above depends on the Count of A at initial time 0, total number of each nucleotides N/4 (and symmetric expressions for C, G and T)
- **JC69 is a stationary model: when $t \to \infty$, $n_A(t) \to N/4$. When t tends to infinity there is a limit distribution of nucleotides therefore $n_A(t)$ tends to N/4.**

From instantaneous rates to probabilities of change

- **We want to calculate from Q the probability of having two aligned nucleotides e.g. A and C separated by an evolutionary time t**
- **Fundamental relation: $Pr_t(A \rightarrow C) = [e^{Qt}]_{A,C}$**
- **Equilibrium probabilities for the JC69 model:**
  $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$

Estimation of divergence times under JC69

- **If p = proportion of observed differences between two nucleotide sequences (edit distance) and d is the expected number of mutations per site between the two sequences, we have:**

$$d = -\frac{3}{4}\log\left(1 - \frac{4}{3}p\right) \quad > p$$

- **In order to get to proper geological durations (millions of years), one needs to calibrate the rate of mutations (order of magnitude: 10^{-9} mutations per site and per year) using reliably dated data (e.g. fossils i.e. use of geographical/geological information).**

More Complex models for nucleotide data

- **JC69: is the simplest model of evolution for nucleotide data. Equilibrium frequencies all fixed to 0.25**
- **K2P (2 parameters): equal equilibrium frequencies but two exchangeabilities (one for transitions and one for transversions)**
- **F81 (4 parameters): one single exchangeability rate but four different equilibrium frequencies π_A, π_C, π_G and π_T**
- **F83/HKY (5 parameters): different equilibrium frequencies and two exchangeabilities (one for transitions and one for transversions)**
- **GTR (9 parameters): general time-reversible, most complex time-reversible model on nucleotides (is considered the most powerful)**

NB: Beware! Using the most complex model on a small dataset is not the best strategy because of over fitting in parameter estimation!

Substitution models for amino acid data

Matrices of amino acid substitution rates have been developed empirically:

- **JTT (Jones, Taylor, Thornton 1992): first matrix built from a large number of pairwise alignments from the Swissprot databank**
- **WAG (Whelan and Goldman 2001): derived from 3905 sequences in 182 protein families**
- **LG (Le & Gascuel 2008): estimated on 3,912 alignments from Pfam, comprising approximately 50,000 sequences and approximately 6.5 million residues overall**
- **mtREV: for mitochondrial protein data**

SESSION 3

Evolution AND Phylogenetics

Complexity of the search space, distance matrix methods, maximum parsimony

LEARNING OBJECTIVES

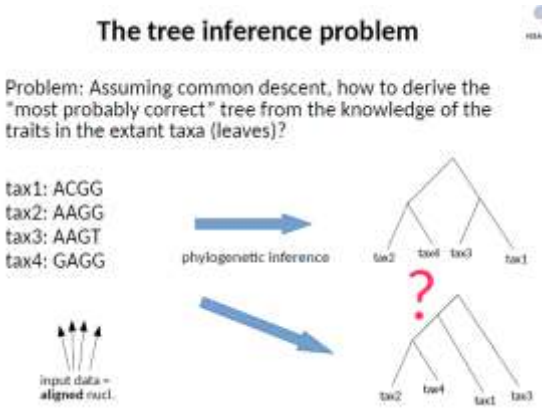
- **Understand the complexity of the space of all phylogenetic trees on n taxa**
- **Understand what are distance matrix methods for phylogenetic inference**
- **Understand what is the maximum parsimony paradigm and how it is used to infer trees**

LEARNING OUTCOMES

- **Be able to give reasonable estimates of the number of trees on n taxa, perhaps with the mathematical formula**
- **Be able to build "manually" the UPGMA tree (distance method) from a small distance matrix**
- **Be able to calculate the most parsimonious history on a given labeled tree**

Strategies in the quest for "the best" phylogenetic tree

Using the tree inference problem, assuming we have common descent meaning that it is possible to link all sequences via tree, apply the knowledge of the trait on the leaves i.e. tax1, tax2, tax3 and tax4 as shown in Session 2.



So many trees! Let's count them.

Fast growth of the number of binary tree topologies with n taxa:

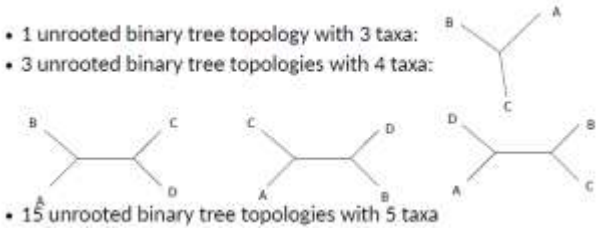
1 unrooted binary tree topology with 3 taxa:

3 unrooted binary tree topologies with 4 taxa:

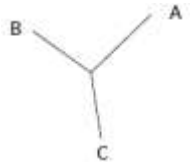
15 unrooted binary tree topologies with 5 taxa

(2n-5)!! = 3 x 5 x 7 x 9 x … x (2n-5) unrooted binary tree topologies with n taxa

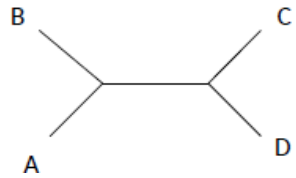
(2n-3) x (2n-5)!! = (2n-3)!! Rooted binary tree topologies with n taxa



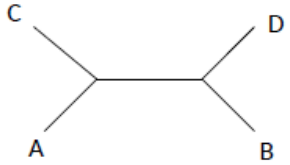
1 unrooted binary tree topology with 3 taxa: If you shuffle A, B and C and re-assign the label A, B and C to the different leaves, you will have same topology tree, just rotation in display or symmetry



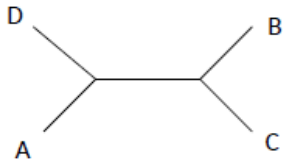
3 unrooted binary tree topologies with 4 taxa: If you have 4 taxa: To make sure we are exhaustive ensure the closest neighbor of A



Here closest to A is B
Swapping C and D still same

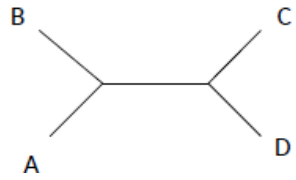


Here closest to A is C
Swapping D and B still same

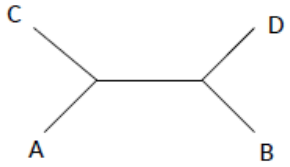


Here closest to A is D
Swapping B and C still same

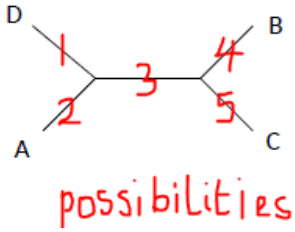
In unrooted If you have to introduce the fifth 5 taxa (you will check 1,2,3,4 and 5 possibility to graft the 5th taxa) total of 5 x 3 = 15 possibilities



Five possibilities here



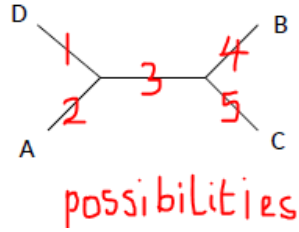
Five possibilities here



Five possibilities here

The formula !! double factorial means we skip every second integer i.e. (2n-5)!! = 3 x 5 x 7 x 9 x … x (here the even numbers are been left out).

In rooted topology differs unrooted, by singularizing the root there for there are five ways to singularize the root as shown below



i.e. (2n-3) x (2n-5)!! = (2n-3)!! Rooted binary tree topologies with n taxa. Therefore for the same number of taxa we have many more rooted binary topologies than unrooted binary tree topologies.

So many trees!

- **For n = 10 taxa, approx. 2 million unrooted binary trees**
- **For n = 20 taxa, approx. 2.2 x 10^{20} unrooted binary trees**
- **And then extra information in the branch lengths!**
- **Looking for "the right tree" is searching an infinite space**
- **One has to devise strategies, either "exact" (algorithms) or including some amount of "sub-optimal guessing" (heuristics) to perform tree inference. Heuristics do not guarantee that maximum likelihood tree will be found.**

Several families of method for phylogenetic tree inference

- **Distance matrix methods try and devise a tree from an input being a matrix of pairwise distances between taxa (contain only positive numbers). Pairwise distance is the distance between any two taxa in your dataset**
- **Minimum evolution (ME) a.k.a. maximum parsimony (MP) methods try and find the tree with minimal number of evolutionary events (character changes) along its branches. Try to explain what we see in the extent taxa, to check the most parsimonious**
- **Maximum Likelihood (ML) methods try and find the tree maximizing the probability of the data observed on the leaves, under a certain evolutionary model. Finding the tree that is most likely to have generated the data you see on the leaves.**
- **Bayesian methods try and associate trees with posterior probabilities (of the model, i.e. tree having considered the data and ultimately pick the tree maximizing that posterior probability. Using prior probability from taking count data of leaves on the tree, then pick the tool that maximize the posterior probability.**

DISTANCE MATRIX METHODS

Pros and Cons

- **Advantage: Can be used on virtually any data (traits), as long as one knows how to calculate pairwise distances from them (positive values and that grows when taxa is dissimilar)**
- **Advantage: distance matrix methods are fast, accommodate even large numbers of taxa (progressive alignment with a guide tree to build the alignment)**
- **Shortcoming: most often, calculated distances differ from patristic distances (sum of branch lengths to the path). Patristic distance is the length of the paths between e.g. i and j so it is not possible to match the same length in tree and original distance matrix.**
- **Shortcoming: loss of signal from the data, (complexity of a taxa summarized into its pairwise distances with others)**
- **Remark: some methods produce constrained trees (e.g. respecting the molecular clock).**

Most Common Distance Matrix Method: UPGMA

UPGMA (Sokal & Michener 1958): "Unweighted Pair Group Method using Arithmetic Averages"

- **Progressively grouping taxa into clusters, building the tree bottom-up (from leaves to root)**
- **The distance between two clusters is the average distance between pairs of sequences from each cluster**
- **Initially as many clusters as individual sequences. Place corresponding nodes at height 0.**
- **Iteratively choose the two clusters Ci and Cj with minimum distance dij, define a new cluster Ck containing all sequences from clusters Ci and Cj place the corresponding node at height dij/2 and link it with the two nodes representing Ci and Cj**
- **Continue until the root is reached, when all clusters are connected**

UPGMA at work: 1/3

Imagine that you are given this distances,

	A	B	C	D
A	-	9	7	3
B		-	2	6
C			-	5
D				-

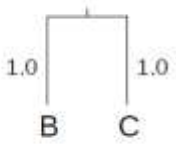
The matrix is symmetrical so representing the lower left triangle

	A	B	C	D
A	-	9	7	3
B	9	-	2	6
C	7	2	-	5
D	3	6	5	-

On the diagonals you have implicit zeros (0) because the distance between taxa and itself is zero.

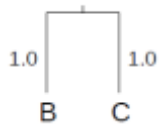
Now let's pick the smallest distance which is 2 i.e. distance between B and C

	A	B	C	D
A	-	9	7	3
B		-	2	6
C			-	5
D				-



- new cluster BC
- d(BC, A) = (9+7)/2 = 8
- d(BC,D) = (6+5)/2 = 5.5

First connect Taxa B and Taxa C



It will be natural to say they are connected to local root (most recent common ancestor) as shown above. The distance 1.0 and 1.0 on both side, shows that the patristic distance between B and C = 1.0 + 1.0 = 2. Therefore we have formed a new cluster of BC.

We need to calculate the distances between the new cluster of BC and all other taxa i.e. taxa A and D. The average distance between BC and A and that of BC and D.

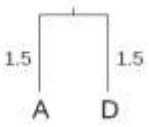
d(BC,A) = (9+7)/2 = 8 i.e. mean from distance from A to B = 9 and A to C =7 therefore (9+7)/2 = 8

d(BC,D) = (6+5)/2 = 5.5 i.e. mean from distance from B to D = 6 and C to D =5 therefore (6+5)/2 = 5.5

With the distances obtained we move to a matrix of size 4 to matrix of size 3

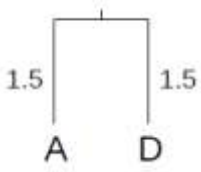
	A	BC	D
A	-	8	3
BC		-	5.5
D			-

	A	BC	D
A	-	8	3
BC		-	5.5
D			-



- new cluster AD
- d(BC, AD) = (8+5.5)/2 = 6.75

It is an iterative algorithm, it peaks another smallest distance in the matrix to group another cluster Therefore the distance between A and D is



Showing 1.5 + 1.5 = 3 as the patristic distance

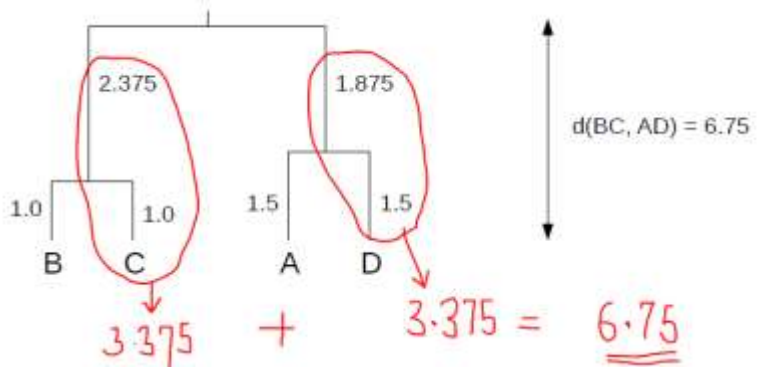
New cluster AD

The distance between the existing cluster BC and the new cluster AD

d(BC,AD) the mean distance of the two clusters, it is made up of four components i.e. distance from B to A, distance from B to D, then distance from C to A and distance from C to D. But because we have already calculated the cluster BC and A and BC and D then we can get the mean from the values shown below:

d(BC,AD) = (8+5.5)/2 = 6.75 (BUT that of the original matrix is 6 is it doesn't respect the distance)

Finally we will have our tree where we connect AD to BC, placing the root at height 6.75. Therefore each half will be half of 6.75 such that the distance between B and D or B and C will be equal to 6.75.



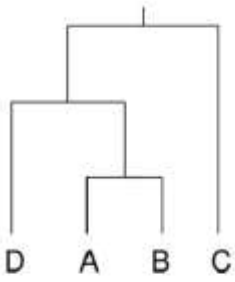
UPGMA Warning: Arithmetic Averages

- **The distance between two clusters is the average distance between pairs (i,j) of taxa with I in one cluster and j in the other**
- **For instance in the preparation of cluster ABC from clusters AB and C, we used a weighted average: If you have clusters with uneven distribution e.g. ABC and E where ABC was derived from AB and C, then to calculate the distance of ABC to E we use: The distance between AB and E with coefficient 2 and the distance between C and E with coefficient 1. Just because we have 2 taxa in d(AB,E) and 1 taxa in d(C,E):**

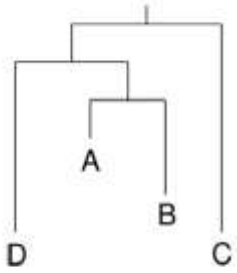
  d(ABC,E) = 2/3* d(AB,E) + 1/3 * d(C,E)

Properties of the UPGMA tree inference

- **Trees reconstructed by the UPGMA algorithm are ultrametric (same length from the root to any leaf). Assume that all taxa are extant so the distance is same.**
- **If the molecular clock hypothesis holds (constant mutation rate), then UPGMA reconstructs the true tree**



An ultrametric tree



A non-ultrametric tree

Additivity and the NJ Algorithm

In case the distance between any pair of leaves is equal to the length of the path between them, the tree is said to be additive.

- **UPGMA always yields an additive tree, but the final distances do not always equal the original distances from the distance matrix: if additivity holds but not unltrametricity, then UPGMA will not reconstruct the correct tree**
- **Saitou & Nei (1987) and Studier & Keppler (1988) developed a method that always reconstructs the correct tree if the set of distances is additive: the Neighbor-Joining (NJ) algorithm.**
- **NJ more complex than UPGMA, using corrected (modified) distances**
- **Most popular distance method, used e.g. to build quick guide trees for multiple sequence alignment algorithms**
- **Refined implementation: BioNJ (Gascuel 1997).**

Distance methods: A summary

- **Two main algorithms: UPGMA and NJ**
- **Methods suitable for a wide range of data: only need pairwise distances between taxa**
- **These algorithms are fast: (n-1) internal nodes created one by one, i.e. linear complexity (they are greedy by not undoing what has been done)**
- **Greedy algorithms: no costly exploration of the set of all tree topologies**

Rooting a tree with an outgroup

- **UPGMA yields rooted trees**
- **NJ and other methods to be seen later produce unrooted trees. How to root them?**
- **In case the molecular clock property holds, one can choose for the root the midpoint on the longest path between any two leaves (when you are sure there is no modification in the rate of evolution along the evolutionary history linking the sequences in your dataset, then it is acceptable to choose as the root, the midpoint on the longest path between any two leaves). Take the longest path between leaves i.e. Diameter of the tree and take the midpoint on this path which will be the inferred root in case the molecular clock property holds for your dataset.**
- **Otherwise, one can root using an outgroup (usually selected as far group i.e. outgroup e.g. choosing non-mammal when working on mammal taxa)**
- **An outgroup is an extra taxon for which we know its MRCA (Most Recent Common Ancestor) with any taxon in the tree clearly predate the MRCA of all other taxa in the tree**
- **The node where it connects with the rest of the tree is the inferred root (the node where the outgroup will connect to the tree is inferred root, you will add artificially 1 taxon in your dataset just for the purpose and you later remove it, mainly to keep track of the point where it connected to the rest of the tree is inferred root for your tree.**

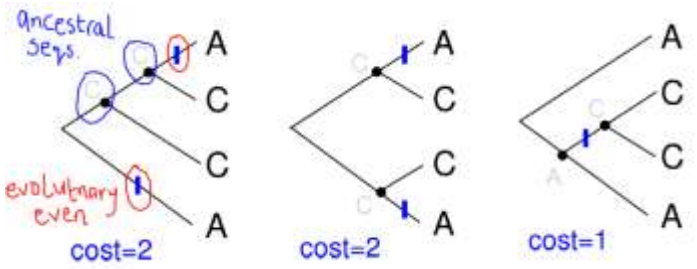
MAXIMUM PARSIMONY (MP) a.k.a Minimum Evolution (ME)

Minimum Evolution Principle

- **Minimum evolution principle goes for the evolutionary history accounting for (i.e. "explaining") the extant sequences with as few evolutionary events (characters, substitutions, deletions and insertions) as possible. (Best model should be made as simple as possible, meaning it needs to explain the reality of our experience of the real world, in this case the experience is what we see on the leaves, the extant taxa etc. The best model from the view point of minimum evolution is the model containing as few evolutionary events as possible).**
- **Minimum count of events: "Parsimony cost" or "Parsimony steps"**
- **Similar to trying and get a tree with smallest total branch length**
- **A principle (dogma), not a scientific truth (i.e. minimum evolution is a dogma not a theory, it is not that which needs to be proved, it is just a view point that is convenient because it enables us to build efficiently phylogenetic trees that are in most cases quite good approximation of the best trees we can construct so far.**
- **Beware! True number of substitutions ≥ greater than equal to inferred parsimony cost (think of the possible evolutionary history A → T → A). Zero cost in terms of parsimony A to A**
- **Be cautious with MP phylogenies on very divergent data. (Phylogeny on sequence that have greatly evolved from one to another, it is not a good idea to use maximum parsimony algorithm.**

Different Trees, Different Parsimony Costs

Given the same alignments, we have different costs along different trees



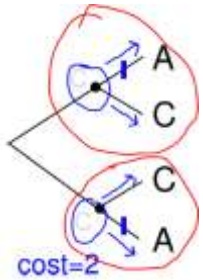
Cost=2 **Cost=2** **Cost=1**

Evolutionary even are marked in blue terms

Ancestral Sequences also marked in gray

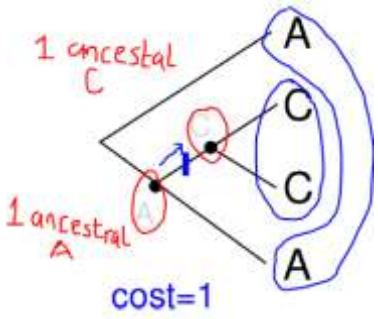
In figure 1: If you infer the history having an ancestral gray C and C in the figure, the parsimony cost for the 1st tree is 2.

In figure 2: Again having same alignment, if you infer with the tree linking, the cost is also 2, because A and C grouped together and another A and C grouped together



Meaning there has been some event of evolution along the blue arrowed path

Given the 3rd tree to link to different sequences, here there is parsimony cost of 1, because there is ancestral group C and C



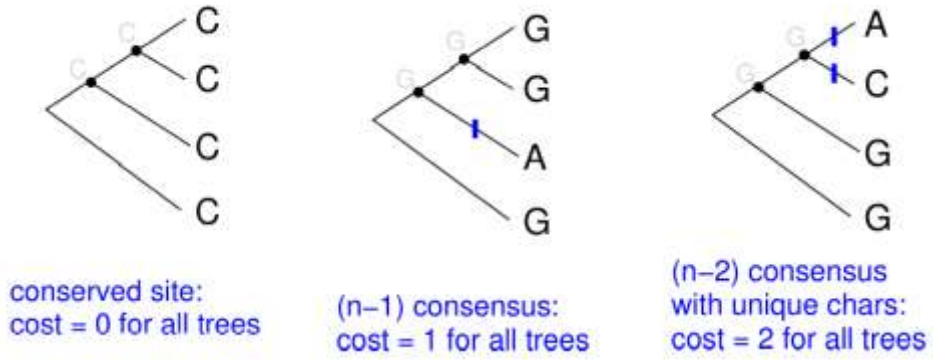
Here you have only 1 substitution taking place along the branch

Different sites "support" different trees

- **A site is a column in a multiple sequence alignment (one aligned character per taxa)**
- **Phylogeny built from a Multiple Sequence Alignment (MSA)**
  - **must take into account all sites**
- **Total parsimony cost = $\sum_{site1} \cos t_1$ ("Independent sites" paradigm) i.e. sum of the cost of the parsimony cost for the tree on each site.**

Some Sites "don't choose"

Some sites are uninformative from the viewpoint of MP: they are useless in deciding about the most parsimonious tree.



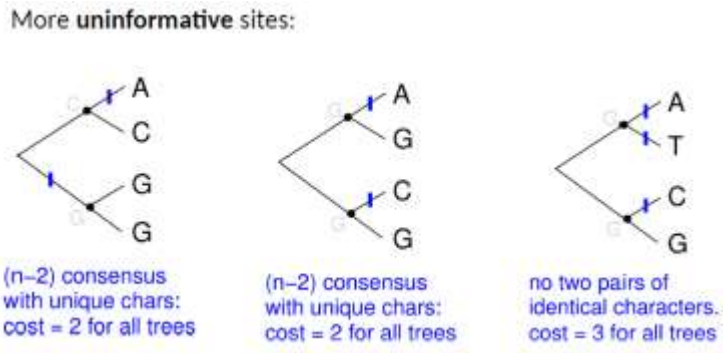
Conserved Site: cost = 0 for all trees (Because there is no observed mutation, the sequence is same) CCCC so the maximum parsimony cost = 0, all taxa has C on each site.

(n-1) consensus: cost = 1 for all trees (Consensus except 1. Here n-1 means only one 1 for all trees whatever tree drawn the maximum parsimony cost will be 1 e.g. GGAG

(n-2) consensus with unique chars: cost = 2 for all trees (Whatever the cost of the tree you will have a parsimony of 2 on the 2 marked blue marks on A and C in the figure above.

Some Sites "don't choose"

More uninformative sites:



(n-2) consensus with unique characters: cost = 2 for all trees
(n-2) consensus with unique characters: cost = 2 for all trees
No two pairs of identical characters: cost = 3 for all trees

Informative site for Maximum Parsimony (MP)

- **A site (column from an alignment) is informative for MP inference methods if and only if it contains at least two different characters in at least two copies each.**
    - **MP methods "ignore" potentially many sites**

Core to Maximum Parsimony analysis: Fitch's Algorithm

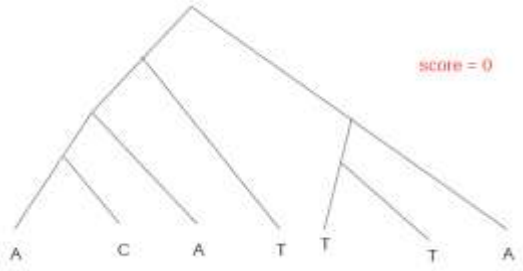
- **Published by Fitch in 1971**
- **Fast algorithm to calculate the parsimony score for any given site on any given tree**
- **Gives both the maximum parsimony score and a most parsimonious history (ancestral characters) in linear time**
- **Doesn't search the space of the different phylogenies: mere calculation on a given tree**

Fitch's Algorithm, step by step

- **Let s = 0 be the score counter initialization**
- **Let c be the set of "acceptable" characters on node k**
- **For leaves k, c_k is the singleton containing the character seen in the alignment**
- **Moving up from leaves to root (post-order traversal), when k is a node having sons i and j, let $c_k = c_i \cap c_j$ if that intersection is not empty, otherwise $c_k = c_i \cup c_j$ and s = s + 1**
- **One the root is reached, s is the total parsimony scores and the "acceptable sets" on the nodes define most parsimonious histories.**

Fitch's algorithm: example application

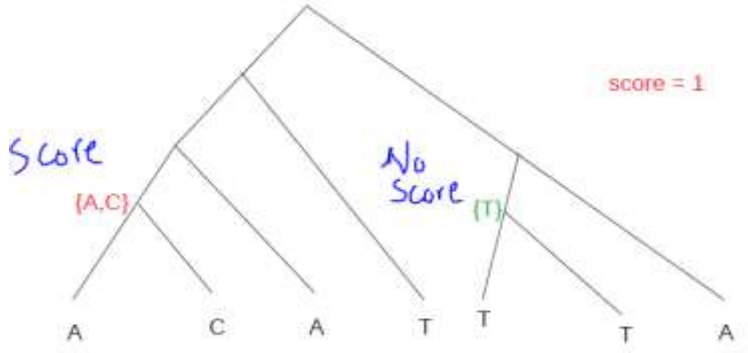
Imagine we have the following tree linking certain numbers of taxa



First initialize score = 0

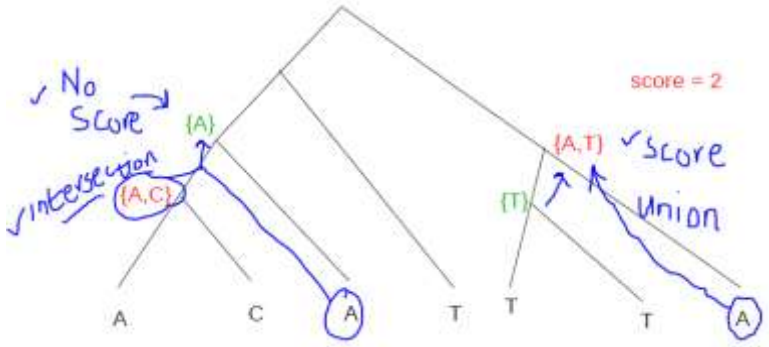
From bottom all the leaves ACATTTA, shows singleton set with 1 unique element (acceptable element)

If we move up… union of {A,C} is used to label the parent of A and C then we increment score = 1

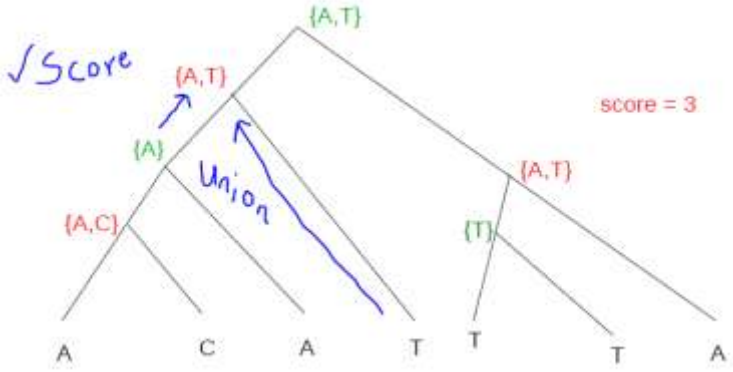


The node T and T gets labeled as {T} we don't modify the score because it is T and T

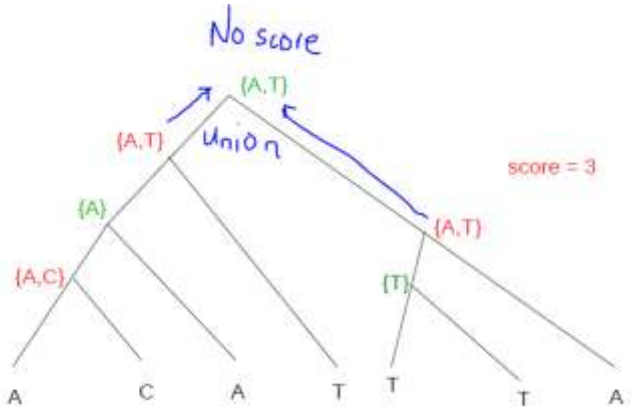
Moving up in the tree, is an iterative invariant, the node labeled {A,C} and {A} the intersection will be {A}, we then label the node above with {A} and we do not modify the score. On the right hand side {T} as one son and {A} as another son, therefore the parent cannot be labeled with intersection we have to use {A,T} and increment the score.



Moving up… we have singleton {A} and {T} then we use the union = {A,T} and it then increments the score.

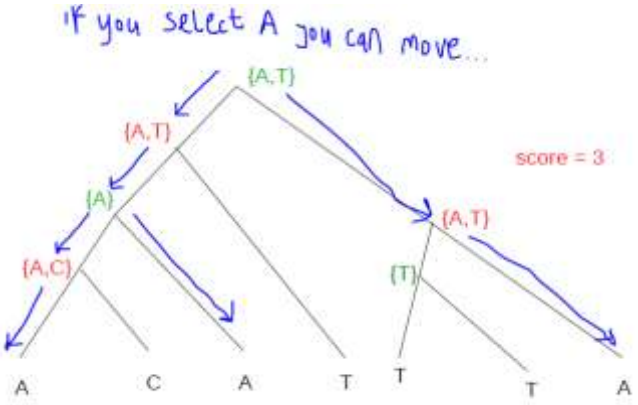


Finally we have {A,T} and {A,T} which are same implying {A,T} and we do not modify the score.

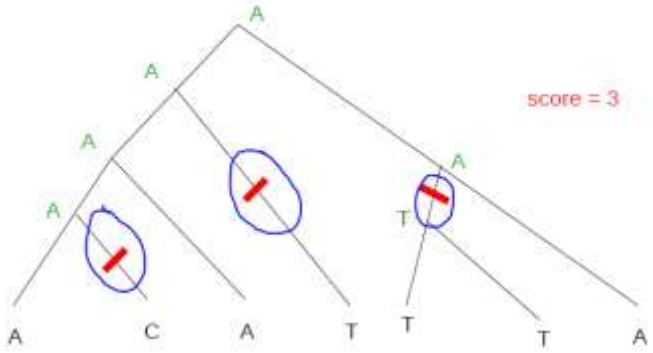


It gives one with the most parsimonious history, not necessarily unique, it is set of characters labeling the internal nodes, ancestral characters. These ancestral characters also define evolutionary events that are going to take place alongside the branches. Fitch algorithm also gives us very fast center length of the algorithm, start from the root pick at random one of the acceptable characters and then go down the tree whenever there is character available used for the parent pick it otherwise pick any other.

If you select A from the root, you can make the following moves



Fitch's algorithm: a most parsimonious history



From the labeling it showed we have evolutionary events that occur like character substitutions on branches where we have two nodes linked are not the same e.g. A and C means somewhere in the branch we have character substitution from A to C, then 2 different A and T same character substitution from A to T (it is possible it has multiple evolution in one branch).

Same calculations can be done for different sites just change the data sequence at the leaves.

Programme for the next session

- **Heuristics for an efficient travel in the space of trees**
- **The likelihood of a tree**
- **The maximum likelihood framework**
- **Assessing the quality of a tree: bootstrap method and local confidence statistics**

SESSION 4

Evolution and Phylogenetics

Heuristics based on tree alterations, maximum likelihood, Bayesian methods, statistical confidence measures

Learning Objectives

- **Know basic tree rearrangements widely used in the literature and in inference programs**
- **Know what is the likelihood of a tree**
- **Understand maximum likelihood methods**
- **Understand Bayesian methods**
- **Know about the bootstrap procedures and other techniques to assess the statistical significance of branches in a tree.**

Learning Outcome

- **Be able to run Maximum Likelihood analyses, understanding how it works**
- **Be able to understand the basic parameters of Bayesian inference methods**
- **Be able to interpret the supports on branches output by phylogenetic inference programs**

Strategies in the quest for "the best" phylogenetic tree browsing the space of topologies
The strategies use to alter a tree from one to another

Necessity for guided tree transformations

- **Tree inference problem is essentially an optimization problem: find the tree that maximizes/minimizes a certain criterion**
- **Remember: (2n-5)!! Unrooted binary tree topologies with n taxa**
    - **Looking for "the best tree", one cannot just try all of them and calculate e.g. the number of parsimony steps for each one (not feasible because of the huge number of tree)**
    - **Necessity to guide the search with a certain criterion or set of criteria, and to develop heuristics to decide which tree to try next (minimize the number of parsimony steps)**

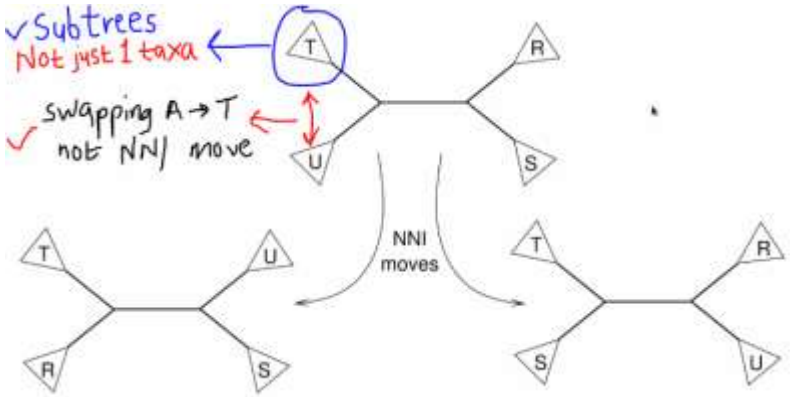
Pseudo-random Walk

- **Phylogenetic inference software implement pseudo-random walks in the space of tree topologies, trying different topologies (not completely at random)**
- **We go from one tree to the next with elementary tree alterations: NNI, SPR or TBI moves**
- **Iterative trial and error process: we try one tree calculate the corresponding parsimony cost, then try improve it with an elementary tree alteration (e.g. picking the move leading to largest improvement), calculate new cost etc.**
- **Pseudo-random process: try random alterations, conserve alteration if it is an improvement, otherwise drop it (backtracking) before attempting another alteration. (Iteratively improve). Trying to find the tree alteration that leads to maximum chain in moving towards your target.**

The moves…
NNI: Nearest Neighbor Interchange (is the simplest one, is a local transformation of a tree, transformation around one edge

NNI is a local rearrangement swapping two of the four sub-trees connected to a given branch

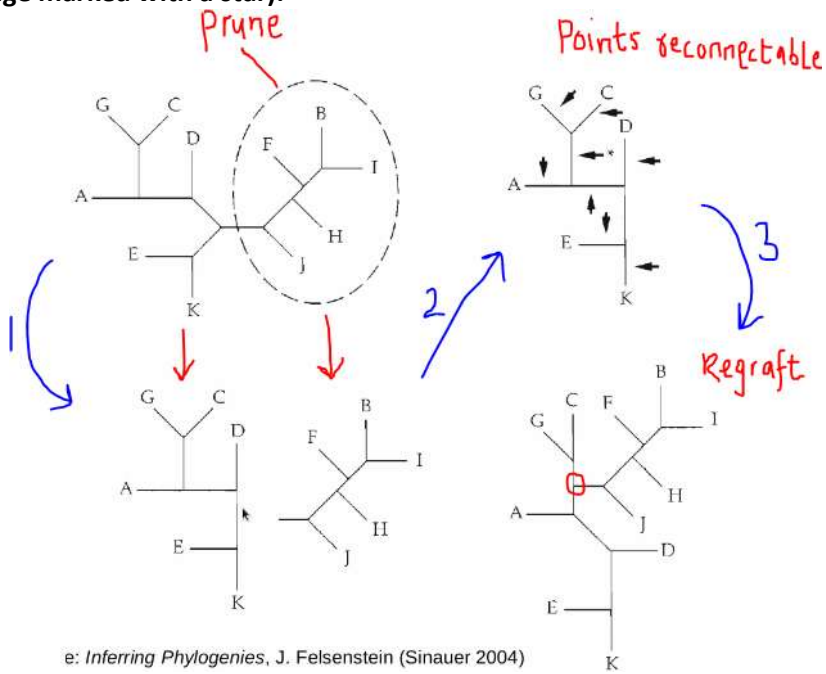
E.g. Unrooted binary tree with 4 sub trees joined together. The NNI will lead to one of the two possible topology called rearrangement around the edges, keeping the sub-trees T, U, R and S intact, but swapping them.



NB: Around one edge at a certain point there are only 2 NNI moves, because there are only 3 topologies preserving the 4 sub-trees around the edges. But in a given tree you have 2 x number of edges NNI moves (i.e. large number of NNI moves because you can focus on different branches)
NOTE: Therefore swapping T to U is not a NNI move, because you will be having same topology.

SPR: Subtree Pruning and Regrafting

SPR is a "less local" rearrangement pruning a subtree and regrafting it onto any of the branches of the tree (here, the edge marked with a star).

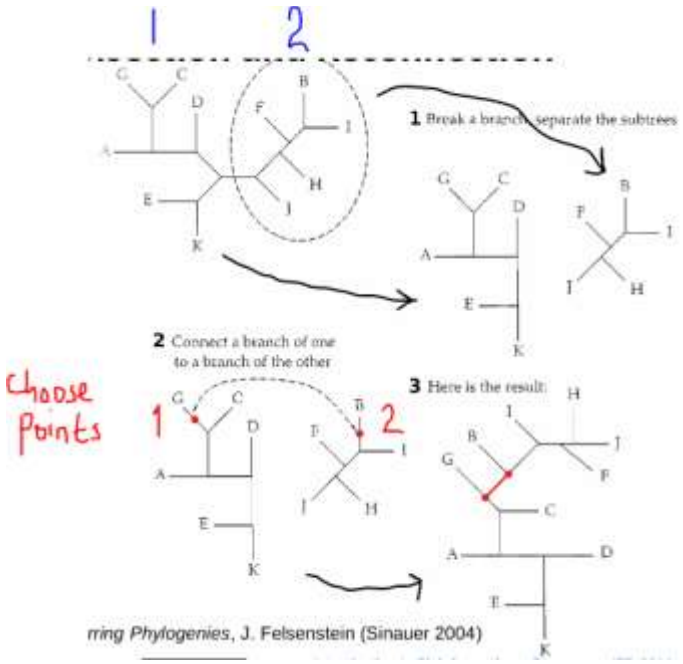


e: *Inferring Phylogenies*, J. Felsenstein (Sinauer 2004)

You can choose where you want to cut and you can also choose where to regraft it in the original tree. It preserves whatever is in the subtree, but it doesn't preserve all neighborhoods after the regrafting process, e.g. A was closer to G but after the above action it is not because of the branching.

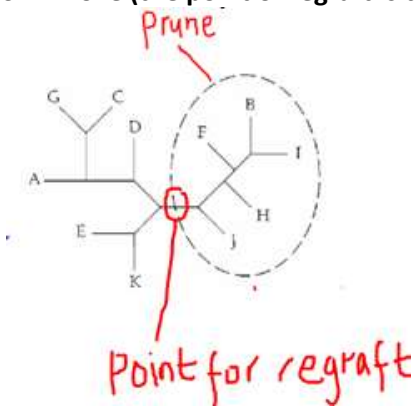
TBR: Tree Bisection and Reconnection

TBR is more involved a rearrangement bisecting a tree into two subtrees and reconnecting them by joining together any branch of the one tree with any branch of the other tree (here, marked in red).
SPR moves form a subset of TBR moves.
Two resulting subtrees, Here the view is a symmetrical view, we do not consider 1 sub tree cut and an original tree, here we have bisection from 1 tree we get 2 trees (i.e. 2 sub trees separated) equally symmetrically.
Choose 1 point in each from both tree and connect both with one branch, conserving the structure of binary tree.

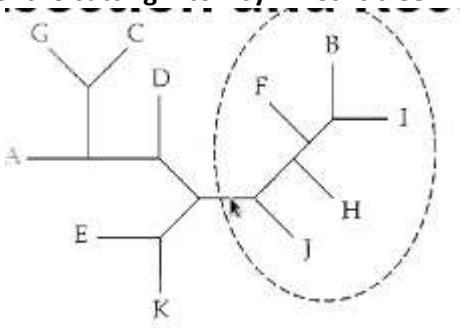


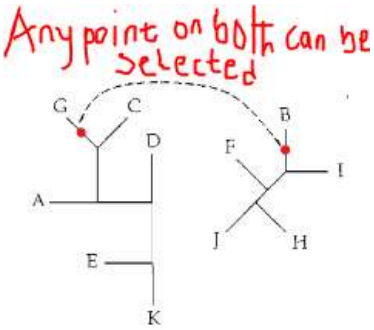
TBR moves are more generic than SPR moves (i.e. SPR moves forms a subset of TBR moves). The difference is that in SPR move you constrained the fact that you are going to regraft the subtree at the point where you cut the subset not anywhere else, the TBR is more generic that leads to more possible rearrangements you can choose 1 point at random in 1 tree and 1 point at random in the other tree.

SPR move (the point of regraft is the point you will move to any part of the original



TBR after the cutting into 2 symmetric tree





The likelihood of a tree

Maximum likelihood approaches for phylogeny reconstruction

Likelihood: definition

Given some observed data D, the likelihood (FR: "vraisemblance") of a model M is the probability that the observations originate from that (generative) model:

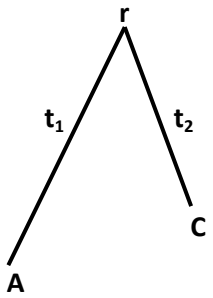
$$Lk(M) = Pr(D|M)$$

In phylogenetics, the D are the traits corresponding to the taxa on the leaves (e.g. alignment of nucleotide or amino acid residues) and the model M encapsulates:

- **A tree topology**
- **The corresponding branch lengths:**
- **An evolutionary substitution model (Matrix Q of instantaneous substitution rates) believed to be acting on the branches of the tree.**

NB: The model is a combination of topology branch length and substitution model.

Likelihood Calculations



$$Lk(T) = \sum_{i\in\{A,C,G,T\}} \Pr(r=i) \quad \Pr(r \xrightarrow{t_1} A) \quad \Pr(r \xrightarrow{t_2} C)$$

$$= \sum_{i\in\{A,C,G,T\}} \pi_i [e^{Qt_1}]_{(i,A)} [e^{Qt_2}]_{(i,C)} << 1$$

Based on this, recursive procedure enables us to calculate the likelihood for more complex trees, based on partial likelihood vectors calculated on the subtrees (Felsenstein's pruning algorithm).

Lk(T) is the likelihood

Two sequences aligned A and C, with a tree connecting them to a certain root r. The branch length t_1 and t_2 respectively. Likelihood of the model is the probability that we have A and C on the leaves knowing all the rest i.e. the rest is the model (topology, branch length, substitution model Q).

The calculation for the probability of having A and C at the leaves of the tree following the evolution on each branch, so a certain nucleotide will be at the r position e.g. i in the formula i can be {A,C,G,T} and it then transforms into A or C in a time t_1 or t_2 respectively.

Probability of transition Pr(r→A) in t_1 is used = matrix as $[e^{Qt1}]$, not the probability of substitution because you may have more than one substitution along the branches.

π_i is the background probability in Markov stationary model. If Q is the model of evolution each of the π_i will be equal to 1/4. In other models it will be different.

NB: When you sum all the vales from the formula above you will get the total probability to have A and C at the leaves of the tree, the model is the tree.

To calculate more complex tree you make use of the pruning algorithm for the likelihood.

Total Likelihood value: sites seen as independent

Lk(T) = $\prod_{sites} LK_s(T) = \prod_{sites} \Pr(s|T)$

And because likelihood values are usually very small, computer programs use their logarithms:

$$\log Lk(T) = \sum_{sites} \log Lk_s(T)$$

Typical log likelihood values for reasonably sized ML trees: -3000 for (small tree), -12000, -63000 for large trees etc. Product of likelihood for each site (can be significantly smaller than 1 simply because it is product of probability, i.e. probabilities are less than 1) therefore the product of the probability will tend to go towards 0 zero.

NB: The likelihood of an alignment is the product of the likelihood of the different sites.

Maximum Likelihood (ML) Framework

Under the Maximum Likelihood framework, one tries to find the tree with highest likelihood (optimization problem) i.e. the model M* such that:

Lk(M*) = $\max_M[LK(M)] = \max_M[\Pr(D \mid M)]$

This implies to try several trees, with heuristics including tree alterations (NNI, SPR and TBR).

- **No guarantee to find the best tree!**
- **Some popular ML software: PAML (Ziheng Yang), PhyML (Guindon/Gascuel). RAxML (Stematakis) is currently the most popular and so called best maximum likelihood reconstruction software.**

The idea is that you want to find the M* which is the best model that maximizes all possible likelihood of the model OR model that maximizes the probability for the data to be generated by the model, or the probability of the data given knowing the generality of the model.

No best tree, so for instance if you try 1000 moves without be able to improve the likelihood then you can say that is the best you can find.

Bayesian Inference

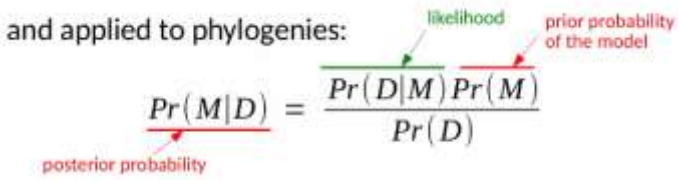
Baye's Theorem

Linking conditional, marginal and joint probabilities:
Pr(A,B) = Pr(A|B)Pr(B) = Pr(B|A)Pr(A)
Probability of A and B = the probability of having B i.e. Pr(B)* probability of A once we have B, knowing we have B i.e. Pr(A|B) called the conditional probability, meaning probability of A conditional on B. Because it is symmetrical so we can have it as Pr(A) * Pr(B|A)
So:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

Using the Baye's theorem and applied to phylogenies:

- **Posterior probability = Pr(M|D)**
- **Likelihood = Pr(D|M)**
- **Prior Probability of the model = Pr(M)**



Probability of M knowing D = (probability of D knowing M * probability of M)/ probability of D
i.e. Pr(M|D) Probability of M once we have seen D i.e. Data
E.g. from a dice rolled 100 times if 90 times appeared with 6 then that is the posterior probability from the data of 100 thrown dice.

Baye's theorem: denominator
Pr(D) (probability of the data) is a sum of joint probabilities over all models:
Pr(D) = Pr(D,M_1) + Pr(D,M_2) + Pr(D,M_3) + … (mutually exclusive events)

= $\sum_{M'} \Pr(D, M')$ = because of Baye's theorem = $\sum_{M'} \Pr(D \mid M') \Pr(M')$

But the space of all models (trees) is continuous:

Pr(D) = $\int_{M'} \Pr(D \mid M') \Pr(M')$

Bayes' Theorem: challenges arising

The final Baye's formula for phylogenetic inference is:

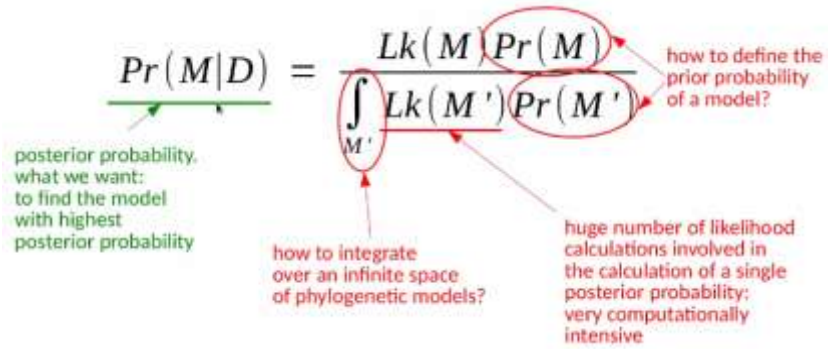
$$\Pr(M \mid D) = \frac{Lk(M)\Pr(M)}{\int_{M'} Lk(M')\Pr(M')}$$

Pr(M|D) is the posterior probability what we want to find the model with highest posterior probability

$\int_{M'}$ how to integrate over an infinite space of phylogenetic models?

Lk(M') huge number of likelihood calculations involved in the calculation of a single posterior probability; very computationally intensive

Pr(M) and **Pr(M')** how to define the prior probability of a model?



Several computational challenges arises

- The integration over an infinite space?
- After integration, how to calculate the likelihood of all the models?
- Definition of prior probability of the model?

Priors on Trees

Several strategies can be used to define prior probabilities on phylogenetic trees:

- Flat ("uninformative") priors: all trees have same prior (uniform distribution)
- Birth-death Markovian process of speciation
- Prior using an arbitrary distribution on branch lengths etc

The frequentist/ML viewpoint: no prior is fully satisfactory.

The Bayesian viewpoint: priors don't matter that much.

Sampling the space Trees

Remember the denominator $\int_{M'} Lk(M')\Pr(M')$? We just have to approximate the value.

Bayesian methods require to sample **intensively** the space of all phylogenetic models and calculate the corresponding likelihoods (i.e. high likelihoods).

Idea of the Markov Chain Monte Carlo (MCMC) methods: wander at random and long enough in the likelihood landscape, covering as best as possible the areas of "good-ish" likelihood, to get ultimately a reasonable image of the whole likelihood landscape.

An MCMC example: Metropolis-Hastings

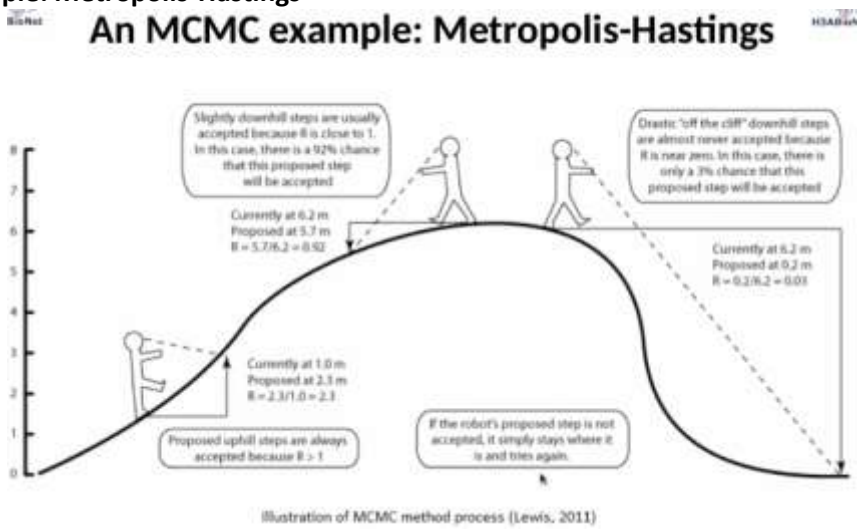


Illustration of MCMC method process (Lewis, 2011)

Metropolis-Hastings algorithm is most common algorithm belonging to the class MCMC algorithm. Imaging exploring the landscape moving from 1 tree to another, check a step using the altitude you will reach divide by your current altitude e.g. R = 2.3/1.0 = 2.3 it brings you a point higher than 1 then you conserve the step.

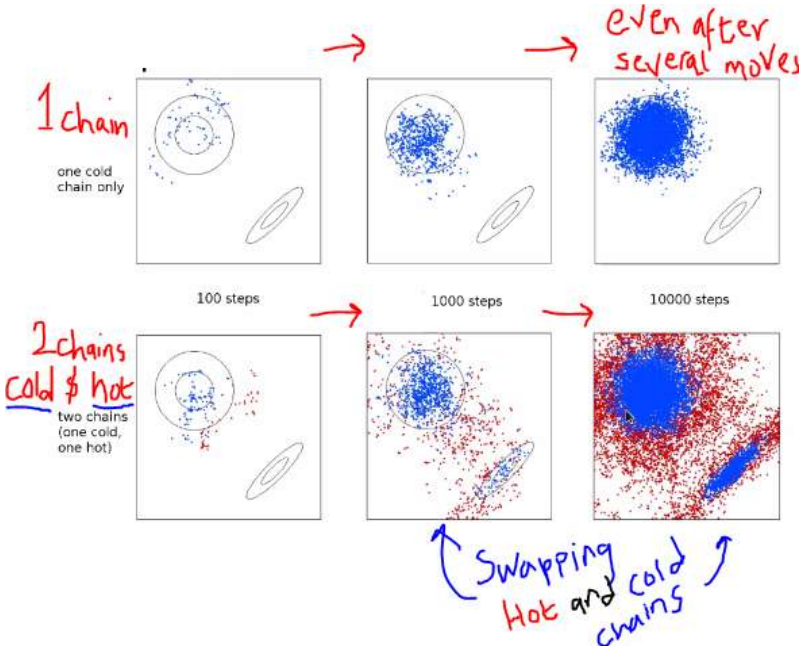
- **Proposed uphill steps are always accepted because R > 1. E.g. Currently at 1.0m and Proposed at 2.3m, therefore R = 2.3/1.0 = 2.3.**
- **Slightly downhill steps are usually accepted because R is close to 1. E.g. Currently at 6.2m and Proposed 5.7m, therefore, R = 5.7/6.2 = 0.92. In this case there is a 92% chance that this proposed step will be accepted.**
- **Drastic "off the cliff" downhill steps are almost never accepted because R is near zero. Currently at 6.2m and Proposed at 0.2m, therefore R = 0.2/6.2 = 0.03. In this case, there is only a 3% chance that this proposed step will be accepted.**

NB: It the robot's proposed step is not accepted, it simply stays where it is and tries again

Some MCMC Refinements

- **Multiple chains with different starting points (initial tree)**
- **Multiple chains with different expected leap length (cold/hot chains)**
- **Swapping a cold chain with a hot chain at certain random times (explore more thoroughly the whole space)**
- **Burnout sequence of n initial trees discarded**

NB: The adventurous chain is called the hot chain and the conservative chain is called the cold chain



The Markovian term in the name implies memory-less, meaning step proposed from one point doesn't depend on the work you have been doing before that point, it only depends on the point where you are.
Usually when you run a MCMC algorithm, you run it for a huge amount of time for the essence of obtaining a burnout sequence, discarding the irrelevant steps.

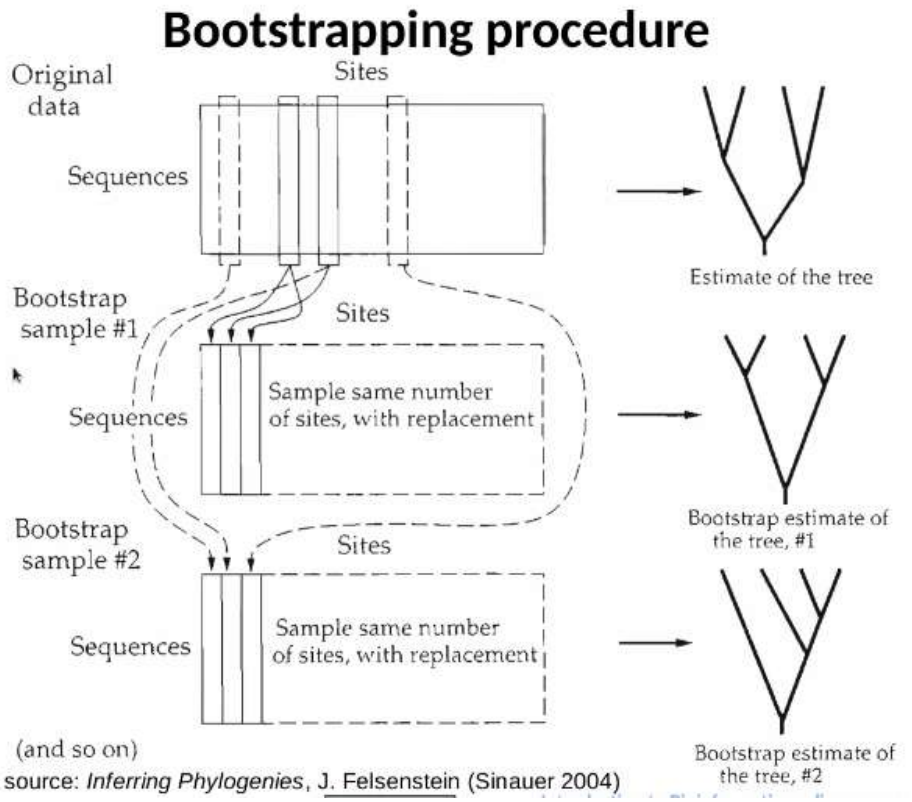
Bayesian Methods: Summary

- **Bayesian methods represent arguable the most elaborate way to infer phylogenies**
- **They aim at maximizing the posterior probability of a model rather than its likelihood**
- **They are very computationally intensive (common to have jobs running for weeks on relatively large datasets)**
- **Some famous software for Bayesian inference: PAML (Yang/Rannala), MrBayes (Huelsenbeck/Ronquist), PhyloBayes (Lartillot/Rodrigue), BEAST (Drummond/Rambaut). MrBayes and BEAST are the most famous software.**

Assessing confidence on the branches: bootstrap et al.

Bootstrapping Procedure

- **Idea: the specific "best tree" we got is a function of the alignment we fed the inference process with.**
- **Different sequence alignments on the same taxa would they lead to alternate trees? Probably… lead to different result and trees (maybe few difference in the branches)**
- **We can resample with replacement from the set of sites composing the input alignment and infer the "best tree" corresponding to that re-sampling (is the idea of bootstrapping)**
- **Perform this iteratively and then compare all the bootstrap trees you get with your original best tree.**
- **The statistical support of a branch is the proportion of bootstrap trees containing the split (split=branch). Sometimes they are given as percentages, fraction of the unit e.g. 99 means 99 of the 100 or 990 out of 1000.**



Original Data is alignment of certain number of taxa, over a certain number of sites. From the original data a phylogenetic inference program was run to get the best tree i.e. Estimate of the tree.

Bootstrap sample#1: Then constructing a bootstrap sample by sampling from the sites above with replacement to build a new alignment with same length as the original and same number of sites, you can select same sites several times i.e. duplicate sites, meaning some sites might not be selected at all. It is called pseudo alignment, it is a mock data made from the original data picking sites at random with replacements. You then run same phylogenetic inference program as you have run in the previous step above to get a tree, because the alignment are not same chances are that the tree is not going to exactly the same with the best tree initially estimated before.

Same is done in Bootstrap sample#2 so that when you have 1000 bootstrap samples, it implies you will have 1000 tree samples estimate to compare with the original tree. A branch that is found consistently in many or all bootstrap estimates will imply the branch is 100% stable.

Other Statistical Supports: Likelihood ratios

- **Other idea: a branch AB|CD is "certain" if the tree containing it has a much better likelihood than the trees obtained by NNI on that branch to include either AC|BD or AD|BC. Giving good confidence to a branch, split all branch in fact separating all four sub-trees AC|BD or AD|BC. If the likelihood is much better in the AB|CD than the two likelihoods AC|BD or AD|BC then you can say the branch AB|CD is quite sure.**
- **Alternatively: likelihood ratio between the phylogeny including the branch and the (non binary) phylogeny having a multifurcation there (branch of length 0 means the absence of the branch) → aLRT statistics (Anisimova & Gascuel, 2006).**