Quiz 1

- Complementary Base pairs A=T and C≡G
- What is alternative splicing
    - Intron Exon boundary bind different factor spliceosome join introns different
    - For the same mrna different length of mrna
- Bootstrap something with replacement
- Taking some sequencing and make another alignment
- Affinity matrix- binding of oligios to
- 64 possible genetic codes
- Template strand is read by RNA polymerase

Calculate length of introns, exons, 5', 3' UTR from Genbank coordinates

- mRNA (1...287,416...638,1537...1800)
- CDS (196...287,416...638,1537...1665)
- CDS runs start to stop codon in mRNA
- 1-195 is 5'UTR and 1666-1800 is 3' UTR
- UTR are transcribed but not translated and contribute to mRNA stability.
- 195 nucleotides of UTR 195-1+1
- 1800-1666+1
- 288-416 and 639-1536 is intron

Introns are additional sequence element that the

How many reading frames are on DNA? 3 on each strand

AGG·TGA·CAC·CGC·AAG·CCT·TAT·ATT·AGC

A·GGT·GAC·ACC·GCA·AGC·CTT·ATA·TTA·GC

-1 frame

AG·GTG·ACA·CCG·CAA·GCC·TTA·TAT·TAG·C

-2 frame

- Given a gene tree and species tree, understand how to reconcile the two.
- Evaluate the fit of sequence alignment data to a tree using parsimony.
- How many genes do genomes have? (order of magnitude) **20,000 protein coding genes**
- What is a bootstrap replicate?
- Contrast NGS (i.e. Illumina) with traditional Sanger sequencing.
- Review unix commands that were used repeatedly: pwd, cd, ls, less; review directory structure; review how a flag is used at the command line; review the concept of a loop in programming
- Be able to solve simple problems: Needleman-Wunsch, UPGMA, Nussinov

Computer stuff

Cloud computing is computational resources such as processors and hard disks are thought of as utilities to be rented from provider (e.g. AWS)

Unix command	Meaning

|	pipes the output from one command directly into the input of a second command.
Grep all gene1.gff>gene.gff	Content with the word all will be outputted to gene.gff
less	View content
For nucleotide in 1 3 6 9 12 >Echo $nucleotide >Done	(Loop) This will display 1 3 6 9 12 because they are the "nucleotide" in this case Echo is used as a variable to substitute and will display a line of text.
Sudo apt install genometool	Sudo-a program that allows users to run programs with security privileges of the superuser Apt- a program that handle the installation and removal of software on Ubuntu.

fasta file format

>gi|186681228|ref|YP_001864424.1| phycoerythrobilin:ferredoxin oxidoreductase

MNSERSDVTLYQPFLDYAIAYMRSRLDLEPYPIPTGFESNSAVVGKGKNQEEVVTTSYAFQTAKLRQIRA

Starts with a ">" for description line

What is Open reading frame

- It is the potential coding region or any chunk without the stop codon.
- The difference between ORF and CDS is that ORF no stop codon.

(-) in a sequence means gaps

Prokaryotic Gene Model	Eukaryotic Gene Model
- No intron - Operons - One transcript, many genes - Small genome, high gene density	- Introns are spliced out of mRNA - Mature mRNA - 5' cap Poly A tail (post transcriptional modification) - Alternative splicing- one gene many proteins - UTR- functionally important

What is difference in these two sequences?

mRNA (1…2, 6…9) CDS join (1…2, 6…9)

mRNA (9…8, 4…1) CDS join (9…8, 4…1)

These two strands are coding/ non-coding strands of each other.

Percent identity: $\frac{\#\ of\ identity\ residues}{\#\ of\ residues\ and\ gaps\ in\ the\ alignment} \times 100\%$

Percent similarity: $\frac{\#\ of\ similar\ residues}{\#\ of\ residues\ and\ gaps\ in\ the\ alignment} \times 100\%$

Rule: Almost always insert gaps in protein coding sequences in groups of 3. RESPECT READING FRAME

Higher percent identity can be achieved when aligning without considering reading frame, but it wouldn't make sense.

Gaps should be inserted in groups of 3 to maintain reading frame, which will only result in a slightly lower percent identity.

Quiz 2 Material

Indel is a mutation that results in the insertion or deletion of nucleotides into the genome

Synonymous mutation is a substitution in the coding region that does not change the protein sequence. CAU to CAC both His

Nonsynonymous mutation is a substitution in the coding region that DOES change the protein sequence. CAU to CAA His→ Gln

Polymorphism is a DNA sequence alteration observed in more than 1% of the population of a species. (alternative phenotype in population/ skin color)

Substitution matrix- considers the type of mismatch and how similar it is to the original (percent similarity)

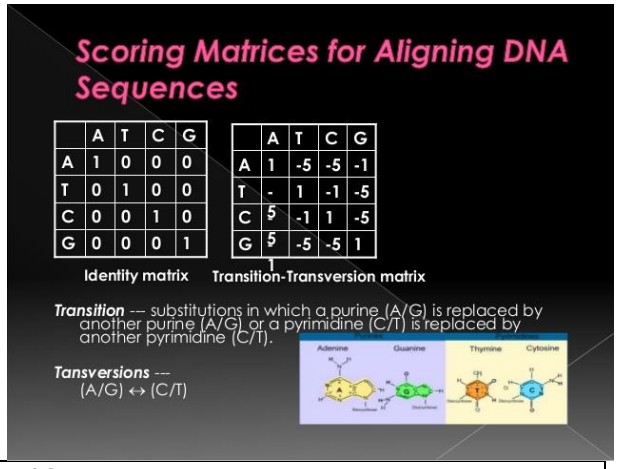
Identity matrix- only regards to if the alignment is a match or a mismatch (percent identity)

Scoring matrix- gives an optimal alignment from the score of pairwise alignment to determine if introduce gap or mismatch is optimal for that situation. You would use a substitution matrix as a reference for the scoring matrix.

(can all be applied to both amino acid and nucleotide sequences)

Examples of scoring matrix: PAM and BLOSUM

They can determine how similar certain amino acid are compare to others.



Transversion: purine to purine pyrimidine to pyrimidine

AG CT

PAM	BLOSUM
Global – optimal pairing of more than 2 species that share a common ancestor.	local alignment – optimal pairing of 2 subsequences with 2 sequences such that similarity score set threshold
	Negative number # of times often for that to be due to chance than common ancestry Positive number # of times often for that to be due to common ancestry than chance.

When would duplication occur relative to speciation?

Duplication 1 and then speciation 1

Why because you can tell that duplication is at the common ancestor since all the branches are the same top and bottom?

	Local Alignment (e.g. BLAST)	Global Alignment (e.g. MUSCLE)
Pairwise Alignment	Alignment of certain lengths	Alignment of the whole sequence
Multiple Alignment	e.g. t_coffee they align 2 or more sequences across their entire length	

Tblastn vs blastp

Tblastn is more likely to find distant relative because they translate the amino acid to different frames.

What information does BLAST give you?

Max Score	The higher the max score, the better the alignment between the hit and the query. This is based on the overall score of HSP between sequences.
Total score	By the sum of scores from all HSPs from the same database sequence.
Query	
Query coverage	The amount of query sentence, expressed as a percent, that overlaps the subject sequence
Max identity	The highest percent identity for a set of aligned segments to the same subject sequence.
e-value	The lower the e-value, the lower the chances of random alignment in a database of a size

How to count # of mismatches between query and subject when given the alignment?

Take total - # identical – gaps

Quiz 3 Material

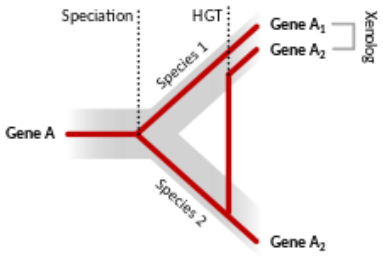
Lineage- a sequence of species each of which is considered to have evolved from its predecessor/linear descent

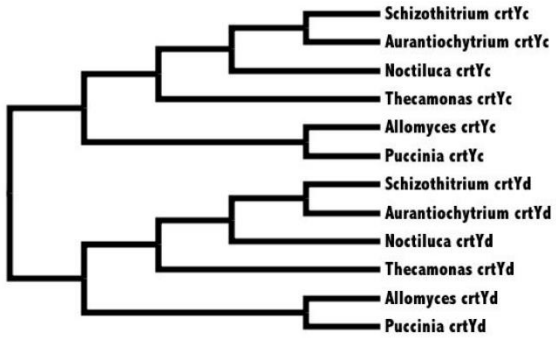
How do we decide which lineage are closely related to each other?

2 linages are more closely related to each other than to some other lineages if they share a MRCA.

Genes can share a common ancestor, or they can be related by species or function.

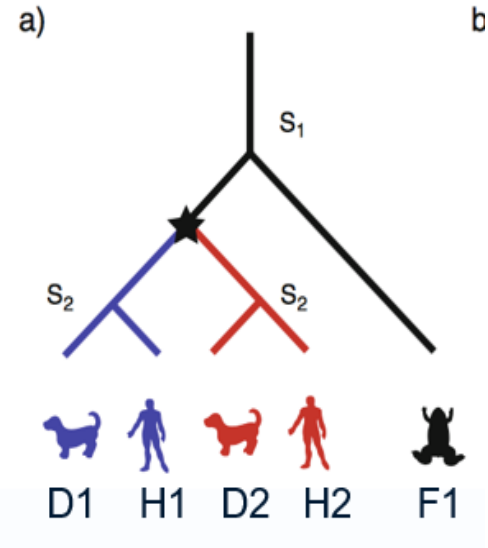
Gene lineage (phylogeny) are not the same as species lineage (speciestree.tre).

Orthologs	Genes sharing a common origin
Paralogs	Genes related by duplication
Co-orthologs	2 or more genes in a lineage that are both orthologous to one or more genes in another lineage due to lineage specific duplication
Out-paralogs	Paralogous genes: duplication **before** speciation
In-paralogs	Paralogous genes: duplication **after** speciation
Xenologs	Orthologs where a homologous gene is found in a distant lineage due to horizontal gene transfer event. 



When is duplication event relative to speciation?

Prior to speciation



Relationship between D1 and H1 **one to one paralog**

Relationship between D1 and H2 respect to speciation event leading to dogs and humans. **Out paralogs**

Relationship between D1 and H2 with respect to speciation event leading to frogs and mammals **inparalog**

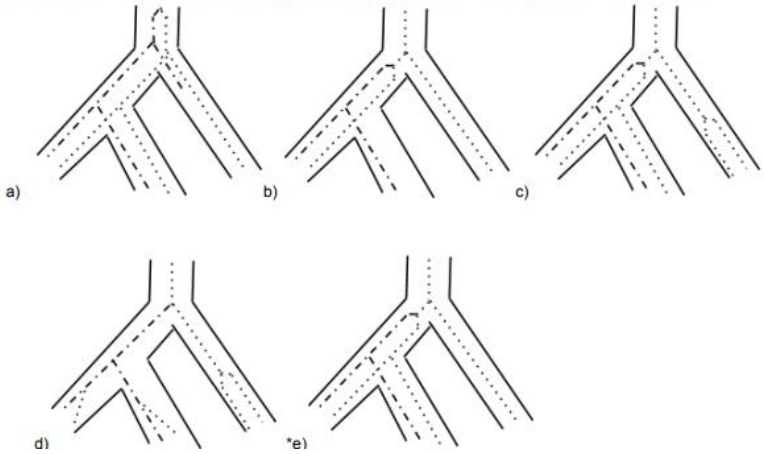
Relationship between D1 and H2 respect to F1 **co-orthologs**

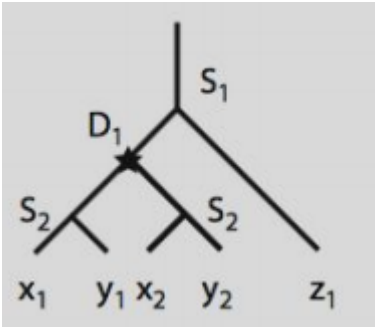
Convergent evolution is independent evolution of similar feature in species of different periods in timeline

Homoplasy is similarity that is not due to common ancestry but a result of independent evolution (convergence) can provide misleading evidence of phylogenetic relationships

Convergent evolution is due to divergent evolution from its common ancestors.

4. Which tree represents the pattern in the gene tree displayed above, superimposed into the species tree?





why is it not a? D1 is after S1

b? shows that y1 or y2 is lost after duplication

c? no duplication at z1

D no duplication at z1 and just no

E is correct

How many duplication and losses are at A? 1 duplication 2 speciation 1 loss

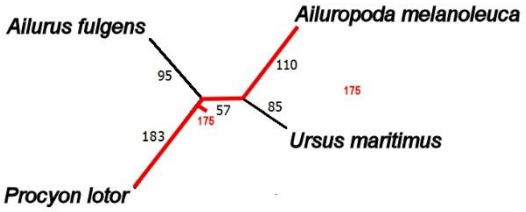
How to midpoint root

Find the longest length between two tips divide by two

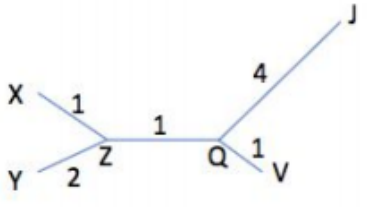
The root would be there

The answer would be c because 2+1+4=7

7/2=3.5



6. The tree at right is unrooted. Tips and internal nodes are labelled with letters. Branch lengths are indicated with numbers. Using midpoint rooting, where does the root go? a) Along the branch Z-X b) Along the branch Z-Y
c) Along the branch Z-Q d) Along the branch Q-V e) Along the branch Q-J
ab) At node Z ae) At node Q



How to bootstrap replicate?

Shuffling of sequences within columns and not rows

Why is it b?

Shuffling because 11355

A/D? The sequences came from nowhere

C? missing a column

Bootstrapping: Majority rule consensus of these three fundamental trees: Numbers indicate frequency of groups in the fundamental trees

7. The aligned sequences of four different species are as follows:

```
Addax CTAGG
Bat   TTCGT
Cow   CAACG
Dove  ATAGT
      12345
```

Which of the following may represent a single bootstrap replicate of the above alignment?

a)
```
Bat   TGCGT
Cow   CAAGG
Addax GTAGC
Dove  TTAGA
```

b)
```
Addax CCAGG
Bat   TTCTT
Cow   CCAGG
Dove  AAATT
```

c)
```
Addax CCGG
Cow   CCCG
Dove  AAGT
```

d)
```
Addax CTAGG
Addax CTAGG
Cow   CAACG
Dove  ATAGT
```

```
                  123456
Felis_silvestris  AACAAC
Felis_catus       AACCCC
Felis_margarita   ACCAAC
Felis_nigripes    CCACCA
Felis_chaus       CCAAAC
```

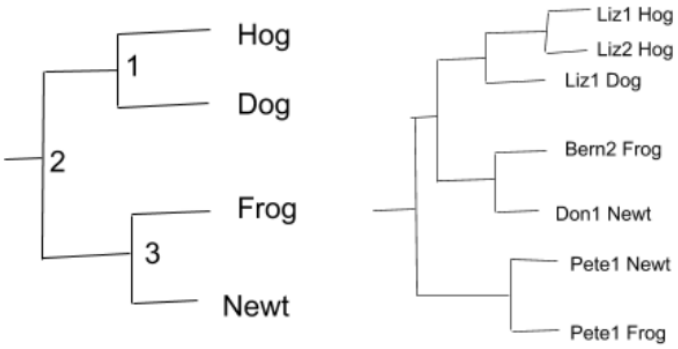
<- The real alignment.

```
                  134546
Felis_silvestris  ACAAAC
Felis_catus       ACCCCC
Felis_margarita   ACAAAC
Felis_nigripes    CACCCA
Felis_chaus       CAAAAC
```

Bootstrap replicate 1.

```
                  225566
Felis_silvestris  AAAACC
Felis_catus       AACCCC
Felis_margarita   CCAACC
Felis_nigripes    CCCCAA
Felis_chaus       CCAACC
```

Bootstrap replicate 2.



Note: Gene tree has genes named Liz1, Liz2, Don1, Bern2, and Pete1, species name follow the gene name.

9)What is the relationship between Pete1 in Newt and Don1 in Newt?

a) xenologs b) one-to-one orthologs *c) paralogs d) one-to-many orthologs e) many-to-one orthologs

10. How many gene duplication events are required to fit the gene tree to the species tree in question 9?

a) 0 b) 1 *c) 2 d) 4 e) 7 ab) 14

duplication → speciation

paralogs

Answer

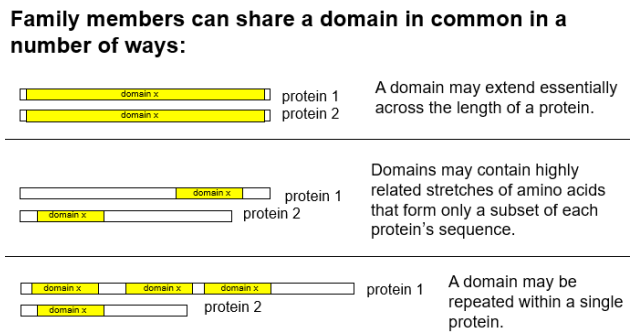
Quiz 4 Material

R is

- A data handling and storage facility
- A suite of operators for calculations on arrays, matrices
- A collection of tools for data analysis
- Graphical facilities for data analysis and display

Motif and domains

- a motif is ~3-20 residues, while a domain is much longer
- motif can be homologous or convergent, while domains are typically homologous
    - yes! Motif can be related by common ancestry or independent evolution but domain, spans for a longer range of residues, so are typically related by common ancestry
    - for example, protein phosphorylated site is 3 residues
    - [ST] – X – [RK]
    - Evolutionary convergence for common functions
- A domain may extend across the length of a protein
- A motif can be predicted by a computer, but a domain must be detected experimentally
    - NOT TRUE. Both can be detected by computer when aligning the family members in a global multiple alignment. <u>Both motifs and domains can then be identified as conserved regions of alignment</u>
    - By scanning the protein's sequence, it can find the high scoring position where it is most likely position of motif/ domain.



Why are some residues conserved over evolutionary time?

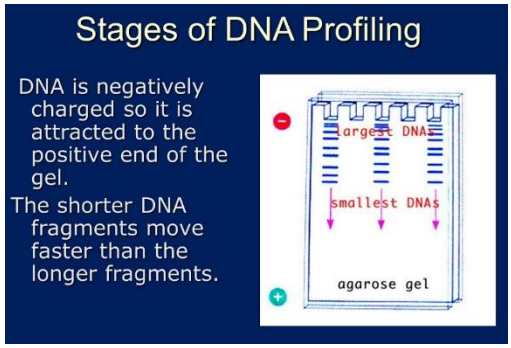
- Because mutations that alter the structure and function of a protein are subject to purifying natural selection
- The ones that changes won't make it so all the ones that passed on are the one that did not change.

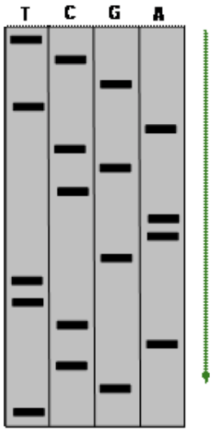
Regular expression- G-X-[WY] means G followed by any residue and either W or Y {P} means anything but P.

Sanger Sequencing

1) DNA is heated until the strands separate.
2) Cool of DNA to add primer on there
3) Separate annealed DNA to 4 tubes

4) DNA is added to all 4 tubes
5) All 4 dNTPs are added
6) One of each ddNTP is added to each of the 4 tubes
7) Polymerase attaches the dNTPs to template strand after primer until a ddNTP is base paired.
8) After pairing of ddNTP is added, it stops because it lacks a primer at the 3' carbon hydroxyl. No more interactions
9) You will now have DNA fragments of different lengths formed across all reaction vessel
10) Polyacrylamide gel electrophoresis is used to sequence DNA.





This should be the template strand

read for the template strand

5'→ 3' (read down up)

AGCAT…

3'→5' (read up down)

AGCATGCG…

Smallest fragments down

Next Generation Sequencing (Illumina)

Illumina can sequence the whole genome's DNA

4 basic steps

Sample prep- add adapter to fragment to bind to flow cell oligos

Cluster generation- each fragment is amplified

Flow cell with lawn with channel with 2 type of oligos

Hybridization polymerase create complement of hybridized molecule

It folds over to second type of oligo and it forms another two tethers to flow cells this results in all of them

Bridge amplification

Sequencing fluorescent nucleotide add to sequence based on template

These cluster are emitted by light # of cycle determine read.

All strands

Quality score is based on the how good the color is (probably of an error in base calling) fastq format

Sanger	Short sequences, fast reaction time few hours

NGS (Illumina)	Long reaction time, 11 days, can sequence an entire genome with short sequences
Third generation sequencing (Nanopore)	High error rate, pore that reads the electrical charges of a sequence. LONG sequences can be read at one. You could take the average of the value for each sequence.

Read product is washed away
Data analysis

What can cause mismatches between Illumina read and reference read?
Error in read and polymorphism in sequenced DNA
Insertion/deletion

De novo assembly is sequencing novel genome where there is no reference.

Heatmap shows color gradients that correlate with the expression level of genes.
Intensity of the red color indicates high expression compared to a control and intensity of green color indicate low expression compare to control

Spotted Array to find expression of gene before and after change. You will find the fold change to see if it upor down regulated. mRNA is changed to cDNA using reverse transcriptase. There is a competition for each spot between normal and cancer cells.
Image visualization and why we take log2 (x)
Log2 will stabilize radiance
Unchanged expression will have ratio of 1 and value of 0
Up and downregulated genes will have range from (0-∞)

Transcriptome is defined as the identity and quantity of the entire population of RNA expressed from genome in a cell.
Proteome is the identity and quantity of the proteins expressed from the entire genome in a cell
Unlike the genome, the transcriptome and proteome can vary from one cell to another.
Transcription can be affected by genetic, environmental, and gene x environmental changes.

Affymetrix GeneChips- oligonucleotides are printed on gene chips and they are tested for mismatch.
This can be tested for mismatch base on the fluorescence intensity

Race to sequence the whole human genome

International Human Genome Sequencing Consortium (IHGSC) and Celera Genomics

IHGSC (Hierarchical sequencing)- generate and align large BAC or P1 clones (genetically mapping) → fragment and sequence a subset of the clones (subclone each genomic region

Celera (genome sequencing)-fragment and sequence entire genome. Focuses on sequencing DNA from randomly cloned fragments.

Quiz 5 Materials

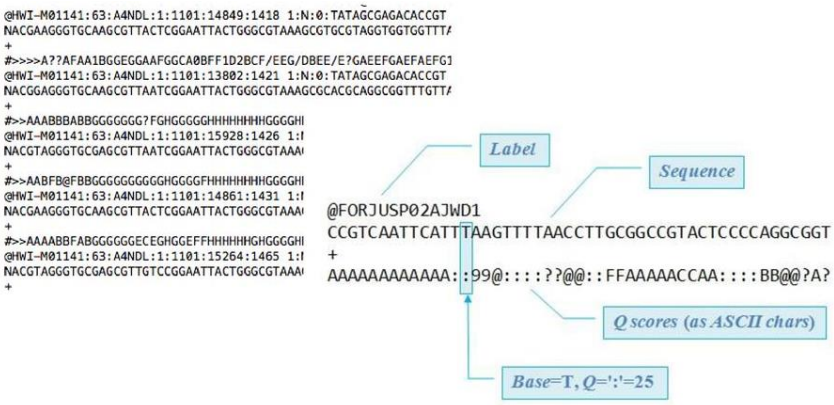
- How many rooted trees are possible from 4 taxa?
- 15 because you can get 3 different unrooted trees with 4 taxa. Each tree can have a 5 different rooting at each midpoint of the tree. 3 x 5= 15 rooted trees
- Parsimony- how many steps needs for each substitution
- RNA-seq can obtain what kind of info?
- Intron-Exon boundaries/ alt splicing/ expression level/ SNPs
- RNAseq uses transcriptomes

If it matches to first portion of read and breaks then matches with second portion of read, most likely there is an intron there.

FASTQ format

#Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

#Line 2 is the raw sequence letters.



#Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

#Line 4 encodes the quality values for the sequence in Line 2 and must contain the same number of symbols as letters in the sequence.

Cufflinks assembles transcriptomes from RNA-seq and quantify their expression

➔ Will get fastq format